Electronic Supplementary Material (ESI) for Molecular BioSystems.

Supplementary Material

Dimension Conversion and Scaling of Disordered Protein Chains

Maodong Li[a], Tanlin Sun[a], Fan Jin[a], Daqi Yu[b], and Zhirong Liu[a,b,c]*

[a]Center for Quantitative Biology, Peking University, Beijing 100871, China

[b]College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

[c]Beijing National Laboratory for Molecular Sciences (BNLMS), Peking University, Beijing 100871, China

*Corresponding author.

E-mail address: LiuZhiRong@pku.edu.cn

Tel: +86-10-62752541

Fax: +86-10-62759595

Non-bonded interactions used in Coarse-grained models

In Eqns. (3-4), there are two characteristic parameters in the non-bonded interaction: the effective distance σ and the attraction strength λ. The excluded-volume effect is embedded in σ which is the only remained effect under $\lambda = 0$. There is an extra attractive effect when $\lambda > 0$, while there is an extra repulsive effect when $\lambda < 0$, see Fig. S1. The parameter λ not only accounts for the direct residue-residue interactions, but also reflects the indirect effect from residue-solvent interactions and is thus expected to change linearly with the denaturant concentration D in experiments. Non-bonded interactions were only considered when two beads i and j are separated sequentially by at least three residues as in common Gō Potential.

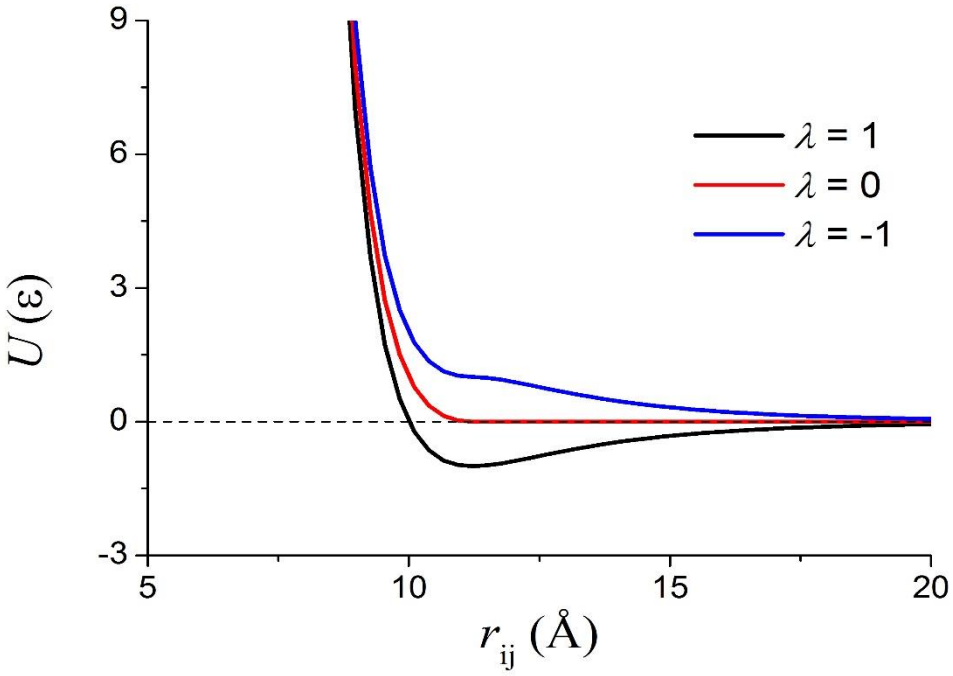

Figure S1 Non-bonded interactions in MD. Non-bonded interactions were only considered when two beads i and j are separated sequentially by at least three residues as in common Gō Potential as Eqns. (3-4). There is an extra attractive effect when $\lambda > 0$, while there is an extra repulsive effect when $\lambda < 0$.

$<R_g^2>^{1/2}$ depends linearly on $<R_{ee}^2>^{1/2}$ on different σ value

We found excellent linearity between $<R_{ee}^2>^{1/2}$ and $<R_g^2>^{1/2}$ on different σ. In Fig. S2, we found different σ resulted in the same slope of the $<R_{ee}^2>^{1/2} \sim <R_g^2>^{1/2}$ standard line. When it occurred to a real protein with native dimension, one can vary R_{gN} to modify only the intercept of the standard line.

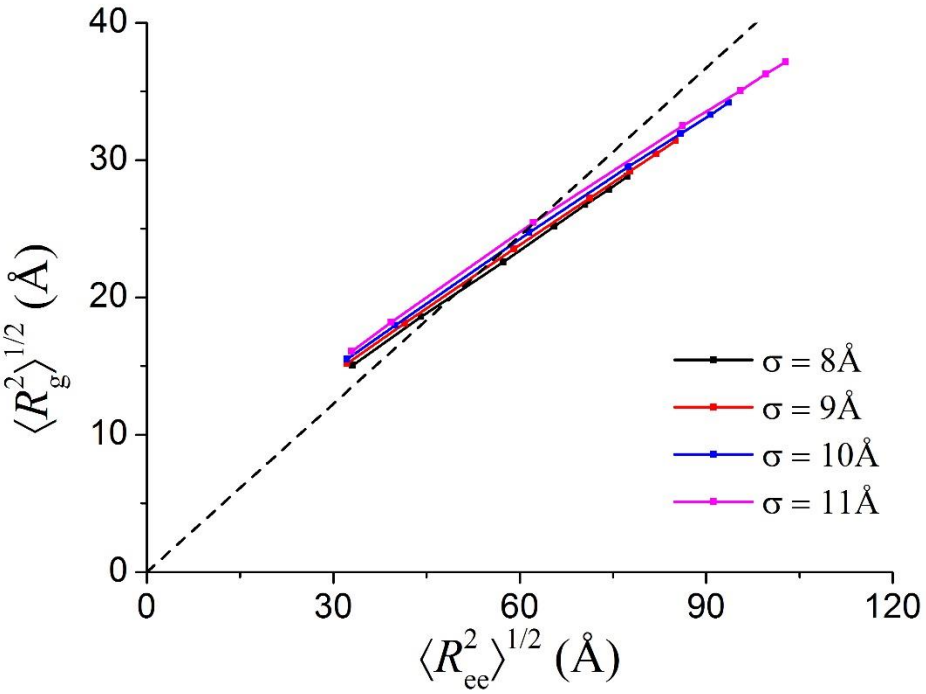

Figure S2 $<R_g^2>^{1/2}$ depends linearly on $<R_{ee}^2>^{1/2}$ on different σ value. Taking a protein chain with N = 60 as an example, $\langle R_g^2 \rangle^{1/2}$ and $\langle R_{ee}^2 \rangle^{1/2}$ of each data point was plotted with σ = 8, 9, 10, and 11 Å. The theoretical result for Gaussian chains [Eq. (2)] was represented as a dashed line.

Prediction of the full dimension based on partly labelled proteins

Using Eqns. (6-7), $\langle R_g^2\rangle^{1/2}$ values can be extracted from smFRET data. As the equation indicates, we should notice that this $\langle R_g^2\rangle^{1/2}$ values represent the dimension of proteins between two labelled dyes, mainly due to labelled length n. When comparing with other experimental data, like SAXS, the whole dimension $\langle R_g^2\rangle^{1/2}$ considering residues outside will be more direct.

Here, we also provide a conversion formula of $\langle R_g^2\rangle^{1/2} \sim \langle R_{ee}^2\rangle^{1/2}$ for the whole dimension as:

$$\left\langle R_g^2\right\rangle^{1/2}/\mathring{\mathrm{A}} = (0.047\ln n + 0.107)\left(\frac{N}{n}\right)\left\langle R_{ee}^2\right\rangle^{1/2}/\mathring{\mathrm{A}} + \left(\frac{1}{\sqrt{6}}\left(\frac{N}{n}\right)^{0.5} - (0.047\ln n + 0.107)\left(\frac{N}{n}\right)\right)(0.557n + 22.540)\left(\frac{N}{n}\right)^{0.2}, \quad \text{(S1)}$$

where some extra exponents are highlighted in red according to Eq. (7). The average error of $\langle R_g^2\rangle^{1/2}$ is about 3%, see in Fig. S3.

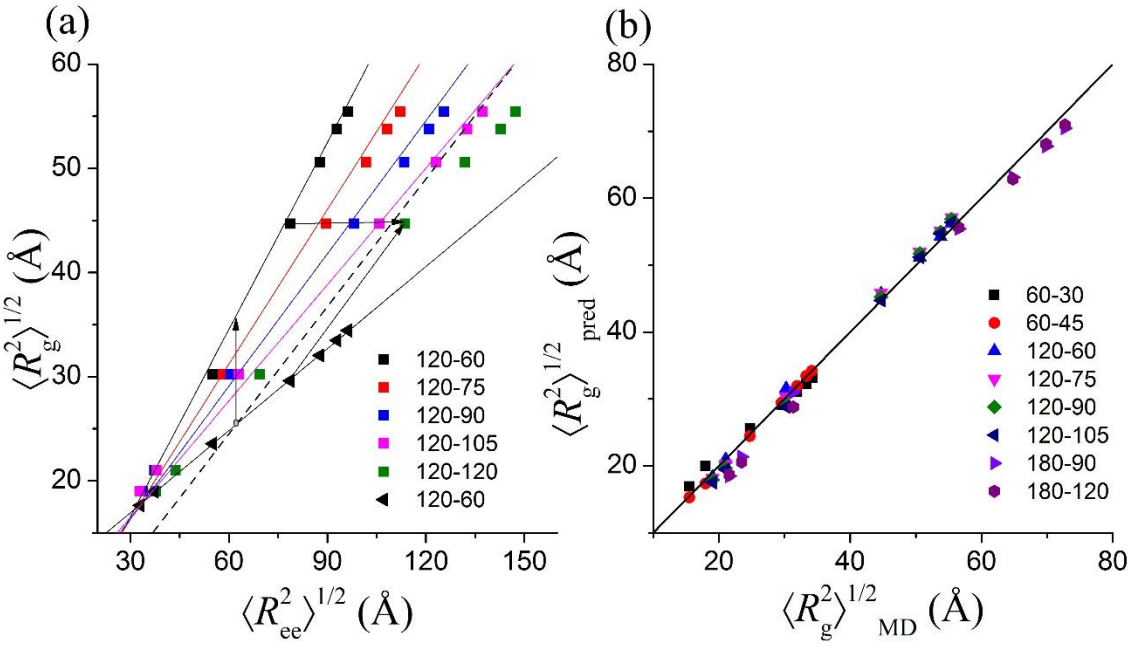

Figure S3 Prediction of the full dimension based on partly labelled proteins. (a) Different from Fig. 2, $\langle R_g^2\rangle^{1/2}$ of each data point was corrected to fully labelled reference. Both the slope and x_0 value were modified according to partly labelled results in triangles using Eq. (7). Linear results of Eq. (S1) were presented in lines. (b) The predicted $\langle R_g^2\rangle^{1/2}$ (converted from $\langle R_{ee}^2\rangle^{1/2}$ with the formula Eq. (S1)) was compared with the directly calculated ones in simulations.

Fitting results of our prediction comparing with Gaussian assumption

Fitting results of both training sets and testing sets are shown in Fig. S4. The max error of $\langle R_g^2 \rangle^{1/2}$ is below 3% and the average error is about 1%. Throughout the coil-globule transition, the average error of the Gaussian chain assumption is over 10%.

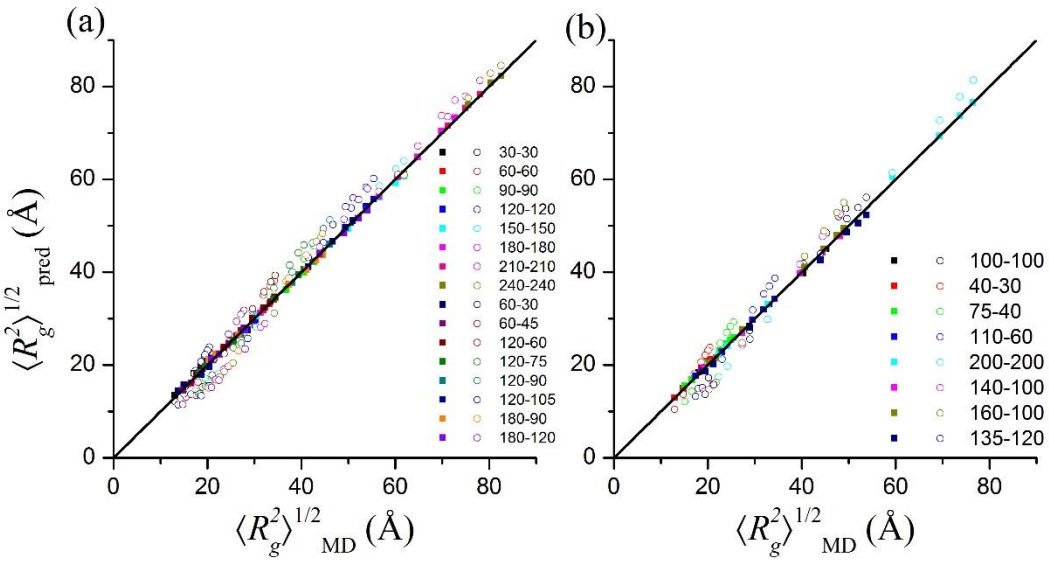

Figure S4 Fitting results of our prediction comparing with Gaussian assumption. Every solid square data point is extracted from a disordered protein chain, while open circle is calculated by Gaussian theory assumption: (a) training sets; (b) testing sets.

Dimension properties of θ-state check point

We examined the dimension properties for systems in which $\left\langle R_{\mathrm{g}}^2 \right\rangle^{1/2} = \frac{\left\langle R_{\mathrm{ee}}^2 \right\rangle^{1/2}}{\sqrt{6}}$ are satisfied.

From Fig. 2 and Eqns. (6-7), it was recognized that, the Gaussian approximation is valid at only one point for the $\langle R_{\mathrm{g}}^2\rangle^{1/2} \sim \langle R_{\mathrm{ee}}^2\rangle^{1/2}$ relationship for any specified (N, n) system, which gives $\left\langle R_{\mathrm{g}}^2 \right\rangle^{1/2} = \left(0.557N + 22.5\right)/\sqrt{6}$ for a full labelled chain. For a full labelled protein (N = n = 120), it gives $R_{\mathrm{g}\theta}$ = 36.48 Å. The corresponding λ value is about 0.15. The determined properties under λ = 0.15 were given in Fig. S5.

The R_{g} and R_{ee} distribution of the θ-state were described as:

$$P\left(R_g\right) = AR_g^6 \exp(-\frac{7R_g^2}{2R_{g\theta}^2}), \tag{S2}$$

$$P\left(R_{ee}\right) = AR_{ee}^2 \left(\frac{3}{2\pi Na^2}\right)^{3/2} \exp(-\frac{3R_{ee}^2}{2Na^2}). \tag{S3}$$

The only fitting parameter of R_{g} distribution $R_{\mathrm{g}\theta}$ = 37.9Å is close to predicted value 36.48 Å. But the distribution deviates from collapsed conformations. The only fitting parameter of R_{ee} distribution, Kuhn length a = 8.49 Å means the equivalent radius of each bead of C^{α} particle, larger than theoretical peptide distance 3.8 Å. This property is mainly due to the excluded volume effect. So instead of an ideal Gaussian chain, this critical protein is a Gaussian-like chain with excluded volume and a larger Kuhn length.

Taken excluded volume into account, the R_{g} distribution was described as:

$$P\left(R_g\right) = AR_g^6 \exp(-\frac{7R_g^2}{2R_{g\theta}^2} + \frac{N^2 b\sigma^3}{2R_g^3}\varepsilon - N(\frac{R_g^3}{Nb\sigma^3} - 1)\ln(1 - \frac{Nb\sigma^3}{R_g^3}) - N). \tag{S4}$$

The resulting fitting value of $R_{\mathrm{g}\theta}$ = 39.5Å is also close to predicted value 36.48 Å. The volume value $b\sigma^3$ = 38.9Å^3 is close to our previous work of 30.0 Å^3. The model interaction parameter ε = 1.14 also indicates the θ-state (ε is changeable with protein length, and values a little bigger than 1).

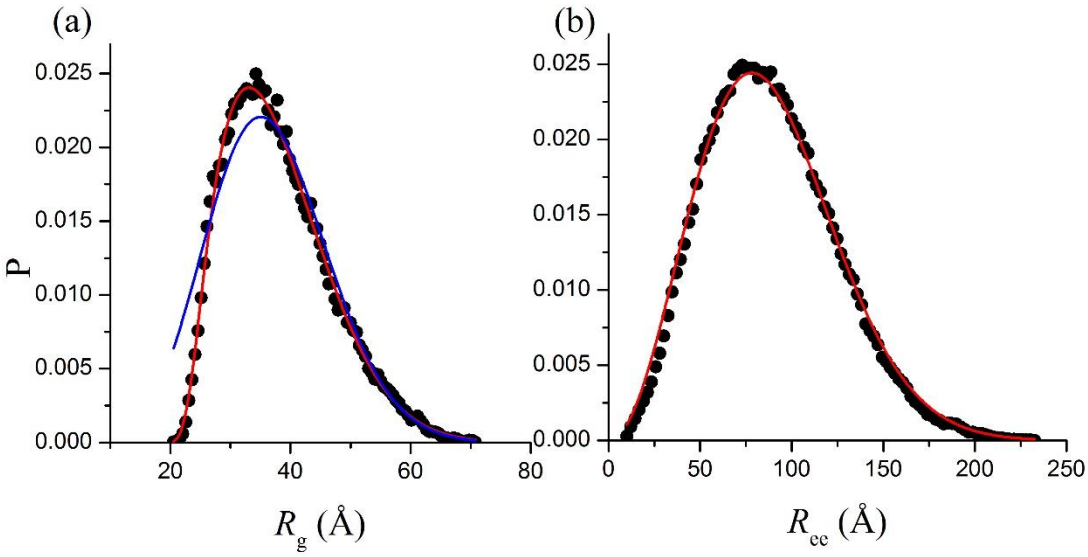

Figure S5 Dimension properties of θ-state check point ($N = n = 120$, $\lambda = 0.15$). The R_g and R_{ee} distribution of the θ-state were fitted by the Eqs. (S3-S4), R-squares were over 0.99. Fitting parameters are $R_{g\theta} = 39.5$Å, $b\sigma^3 = 38.9$Å^3, $\varepsilon = 1.14$, $a = 8.49$Å. The blue line showed the unsatisfactory Gaussian-fitting by the Eq. S2, with $R_{g\theta} = 37.9$Å.

Dimension prediction for an IDP

Using an IDP, the N-terminal domain of HIV Integrase IN (*N-n*: 60-57), as example, our prediction formula fitted much better than the Gaussian approximation according to MD results, see Fig. S6.

Our prediction formula, Eq. (7), was close to the exact linear fitting as dashed line shown in Fig. S6. It could be found that most of smFRET data located in the middle area of the prediction range.

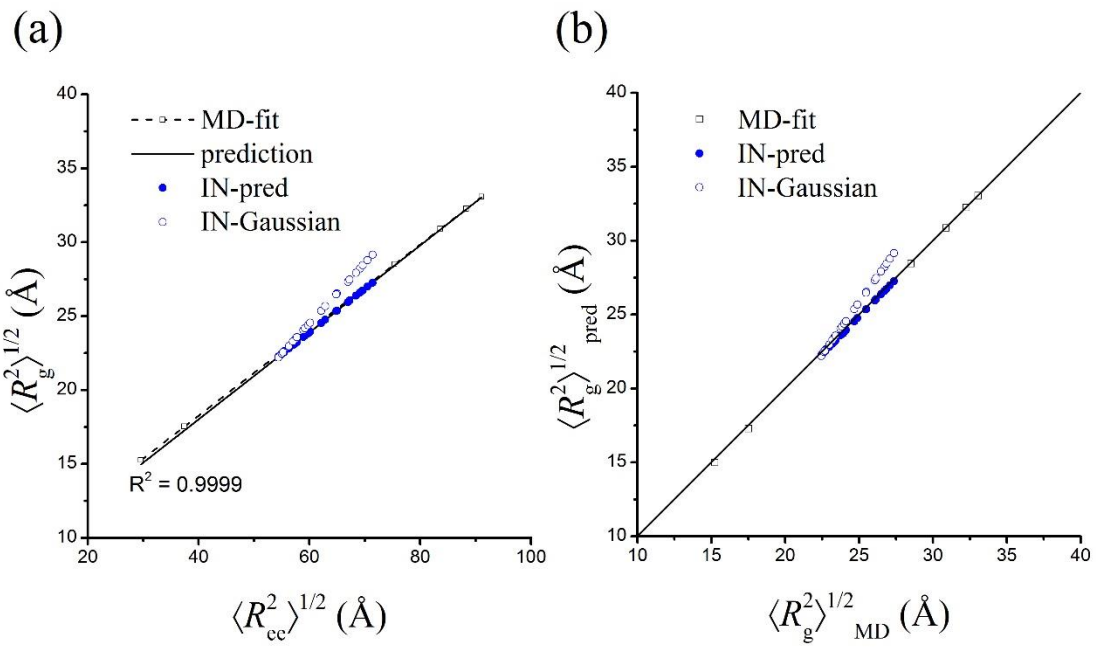

Figure S6 Dimension prediction. Results for protein IN (*N-n*: 60-57). Black open squares are taken from MD results. Blue solid circles are calculated from Eq. (7), and blue open circles are calculated by Gaussian approximation. (a) Black dashed line is fitted exactly according to MD results. Black solid line is predicted from Eq. (7) depending on protein length only. (b) Black solid line shows the prefect prediction as Fig. 4.

Modification of our formula in protein dimension calculation

Using our prediction formula of $\langle R_{ee}^2 \rangle^{1/2} \sim \langle R_g^2 \rangle^{1/2}$ to replace the Gaussian chain assumption of our previous Sanchez model, we could get a dimension modification on each smFRET system. In general, the variation of protein dimensions during denaturation is smaller than that we calculated before, see Fig. S7.

Using a linear fit with smFRET data in high denaturant concentration (D > 2M), we obtained an approximate value at D = 6 M. This result is used in further dimension analysis in Fig. 6.

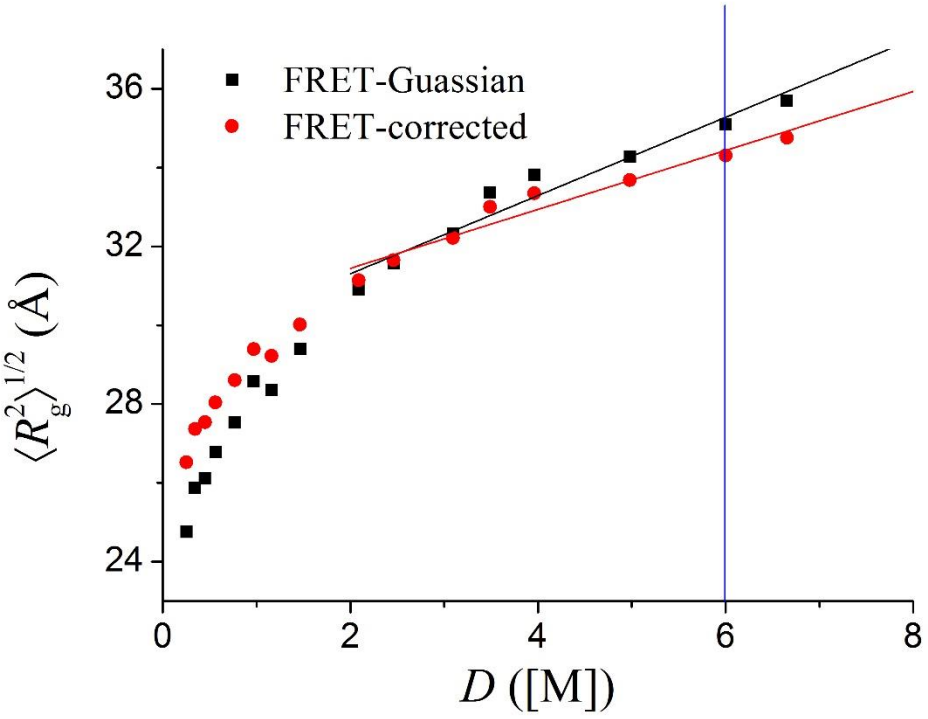

Figure S7 Dimension modification. Results for protein R17-93 (*N*-*n*: 116-94). Black squares are given by Gaussian chain assumption used in our previous work. Red circles are given by our prediction formula. Lines are linear fit with smFRET data in high denaturant concentration (D>2M) in corresponding colors.

Scaling exponent behavior: Not a S-curve

Different from the well-known S-curve as polymers, a novel picture of the scaling law was found with a steep valley before totally collapse in Fig. 6 (a). Similar results were also observed with the Sanchez chain model in Fig. S8.

Around the unexpected valley, protein chains with different lengths collapse into globule in different speed. Bigger proteins collapse faster than the smaller ones as attractive interaction increases, however the smaller ones become more sensitive to a stronger attraction.

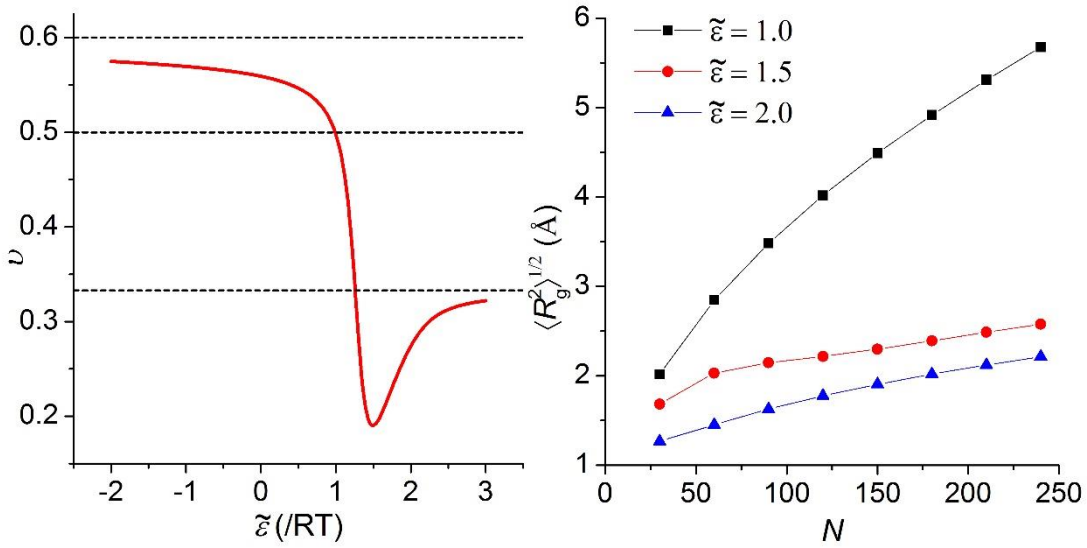

Figure S8 υ values calculated with the Sanchez chain model. υ was determined in a similar as Fig. 6. (Left) The scaling exponent υ as a function of the attraction strength $\tilde{\varepsilon}$ in Sanchez model. Dashed lines from top to bottom represent the result for ideal expanded state, θ-state and native state, respectively. (Right) $\langle R_g^2 \rangle^{1/2}$ as a function of N at a few representative $\tilde{\varepsilon}$ values.

Significant finite-size effect

Surprisingly, some determined υ values in Fig. 6 (a) exceed the theoretical critical boundary (1/3 ~ 3/5). It is actually a finite size effect since the scaling law was originally deduced for infinite chain length. For chains with finite N, accurate fitting to the logarithmic data gave a slope (υ) changing with N, while the theoretical critical boundary is approaching at large N (Fig. S9). This finite size effects are difficult to be fully eliminated under usual chain length, and may bring extra difficulty to related experimental studies.

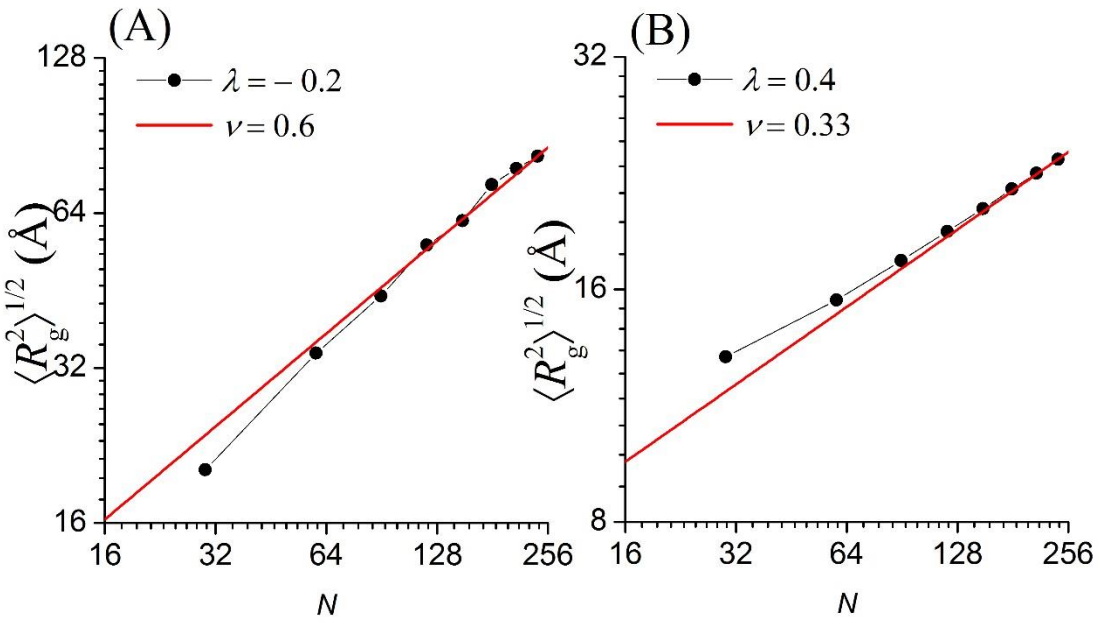

Figure S9 Significant finite-size effect. A logarithmic fitting of Eq. (1) according to the longest chain (N = 240) with a fixed scaling exponent (A) υ = 0.6 (choose λ = −0.2 as expanded state) and (B) υ = 0.33 (choose λ = 0.4 as collapse state).

The scaling law: total vs. individual

In Fig. S10, using a fixed parameter ρ = 3.4 Å, the scaling exponents of individual protein chain was calculated by Eq. (8). In the individual view, the upper limit is below 3/5 while the lower limit is above 1/3, locating in available area in definition. All the "S" curve met the critical exponent υ = 0.5 between λ = 0.1~0.2.

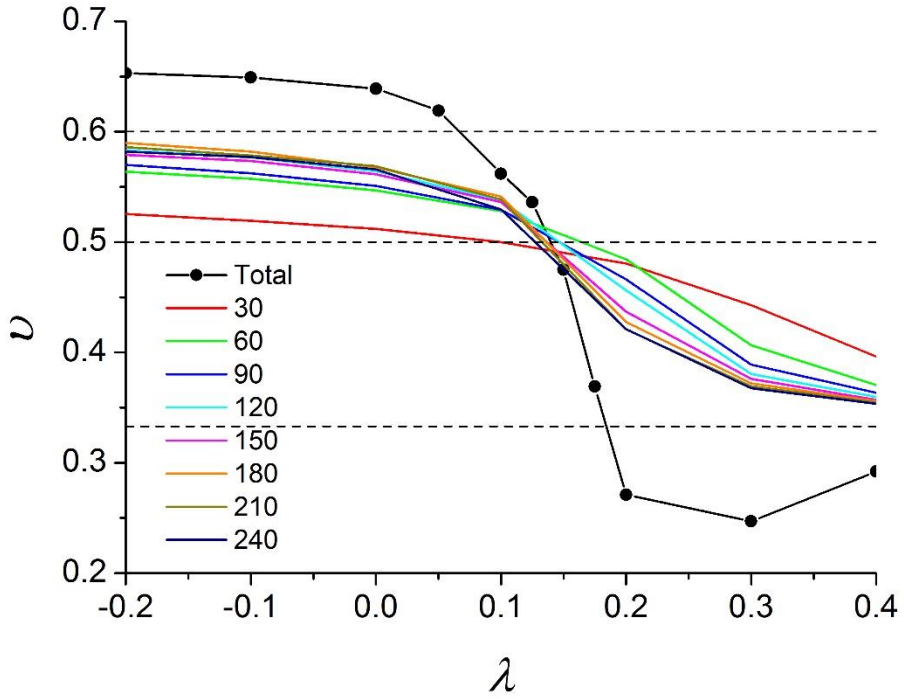

Figure S10 The scaling law of the total one and the individual one of each chain. The total results were calculated as Fig. 6, shown as black line with circles. The individual results were calculated by Eq. (8) with a fixed parameter ρ = 3.4 Å, shown as colorful lines.

Distribution analysis of $P(R_g)$ and $P(R_{ee})$

Though an excellent linearity existed between $\langle R_{ee}^2 \rangle^{1/2}$ and $\langle R_g^2 \rangle^{1/2}$, the behavior of the dimension distribution $P(R_g)$ and $P(R_{ee})$ were not so relevant as literatures shown. See Figs. S11-14, the shapes of distribution peaks vary as non-bonding interaction strength λ. In Fig. S14, the possible available value of dimensions is so narrow in strong attractive interaction that the relationship between $\langle R_{ee}^2 \rangle^{1/2}$ and $\langle R_g^2 \rangle^{1/2}$ was less reliable.

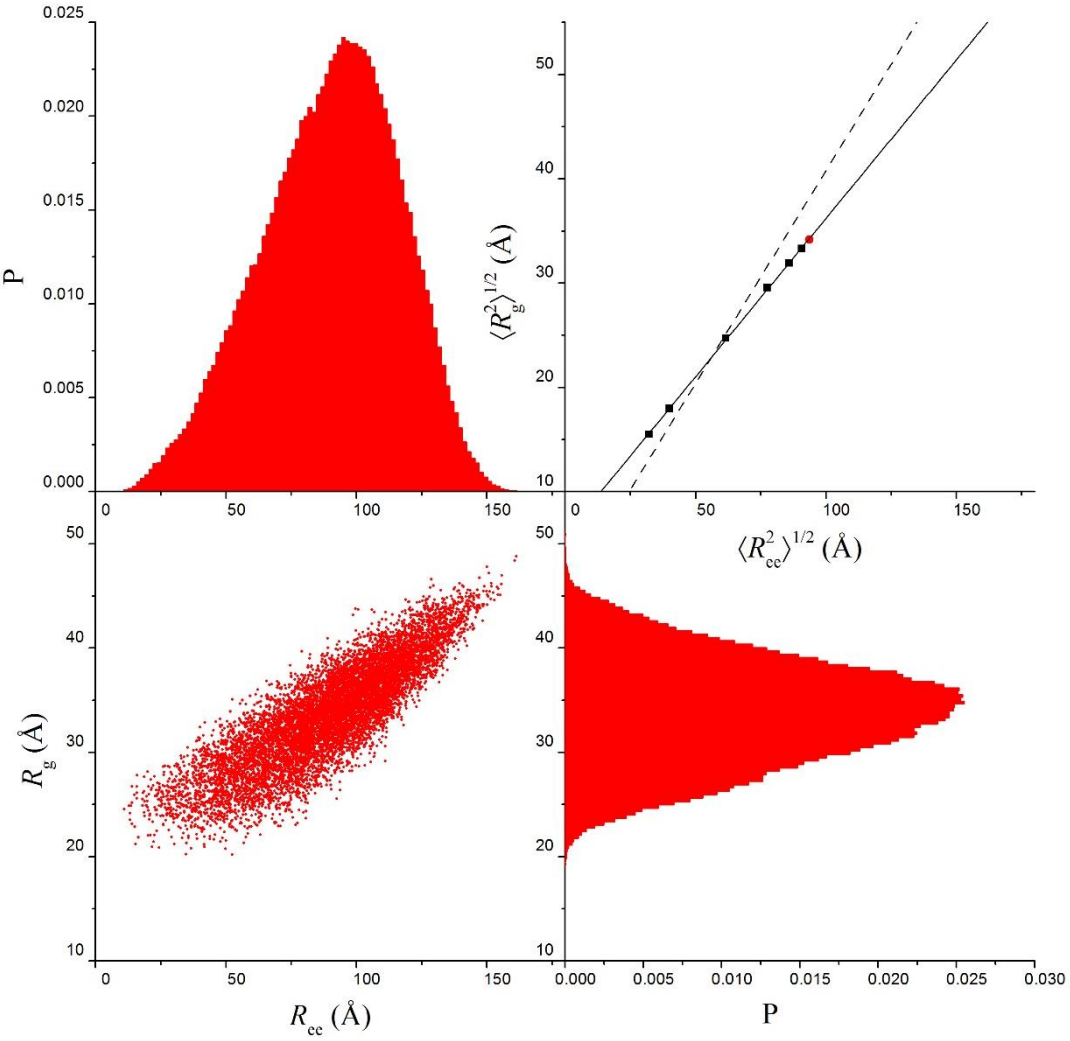

Figure S11 Using a fully labelled protein (N = n = 60) on non-bonding interaction strength λ = −0.2. The variation and correlation between $\langle R_{ee}^2 \rangle^{1/2}$ and $\langle R_g^2 \rangle^{1/2}$ is shown in Left-bottom column. The corresponding distributions of $\langle R_{ee}^2 \rangle^{1/2}$ and $\langle R_g^2 \rangle^{1/2}$ are shown in Left-top column and Right-bottom column, separately. The ensemble average property is shown in Right-top column. The Gaussian assumption is shown as dotted line. Every data point is extracted

from a MD trajectory, while the trajectory of $\lambda = -0.2$ in red. The standard line is shown as solid line.

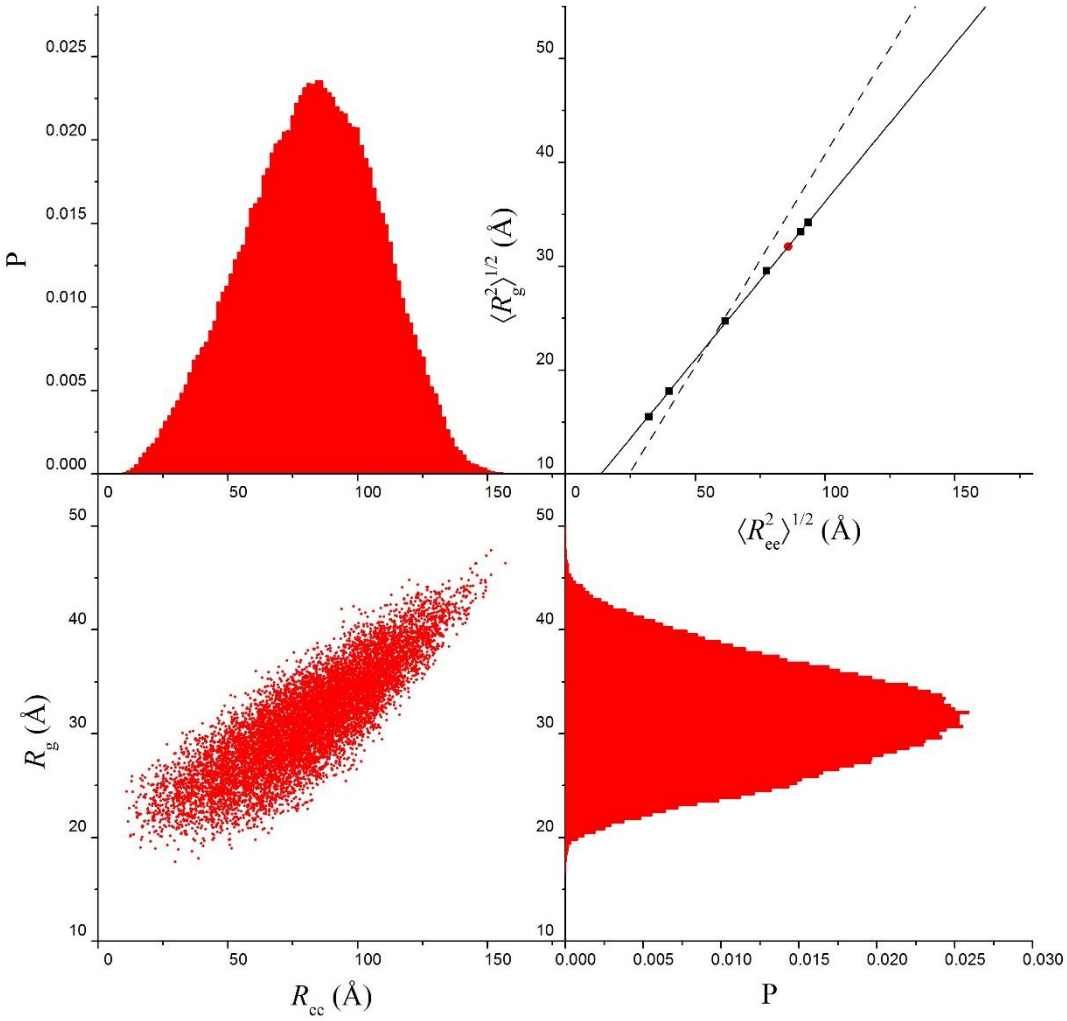

Figure S12 Using a fully labelled protein ($N = n = 60$) on non-bonding interaction strength $\lambda = 0.0$, plotting as Fig. S11.

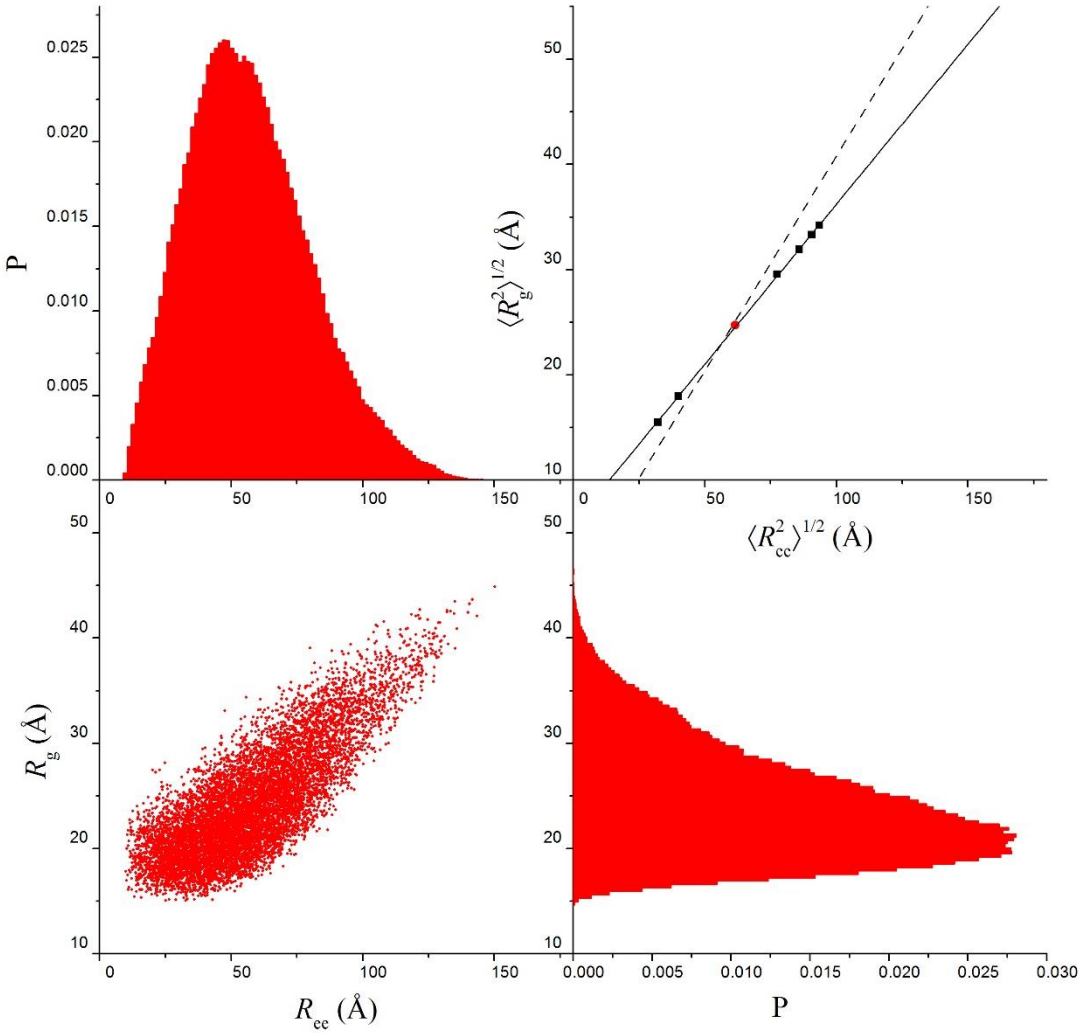

Figure S13 Using a fully labelled protein (N = n = 60) on non-bonding interaction strength λ = 0.2, plotting as Fig. S11.

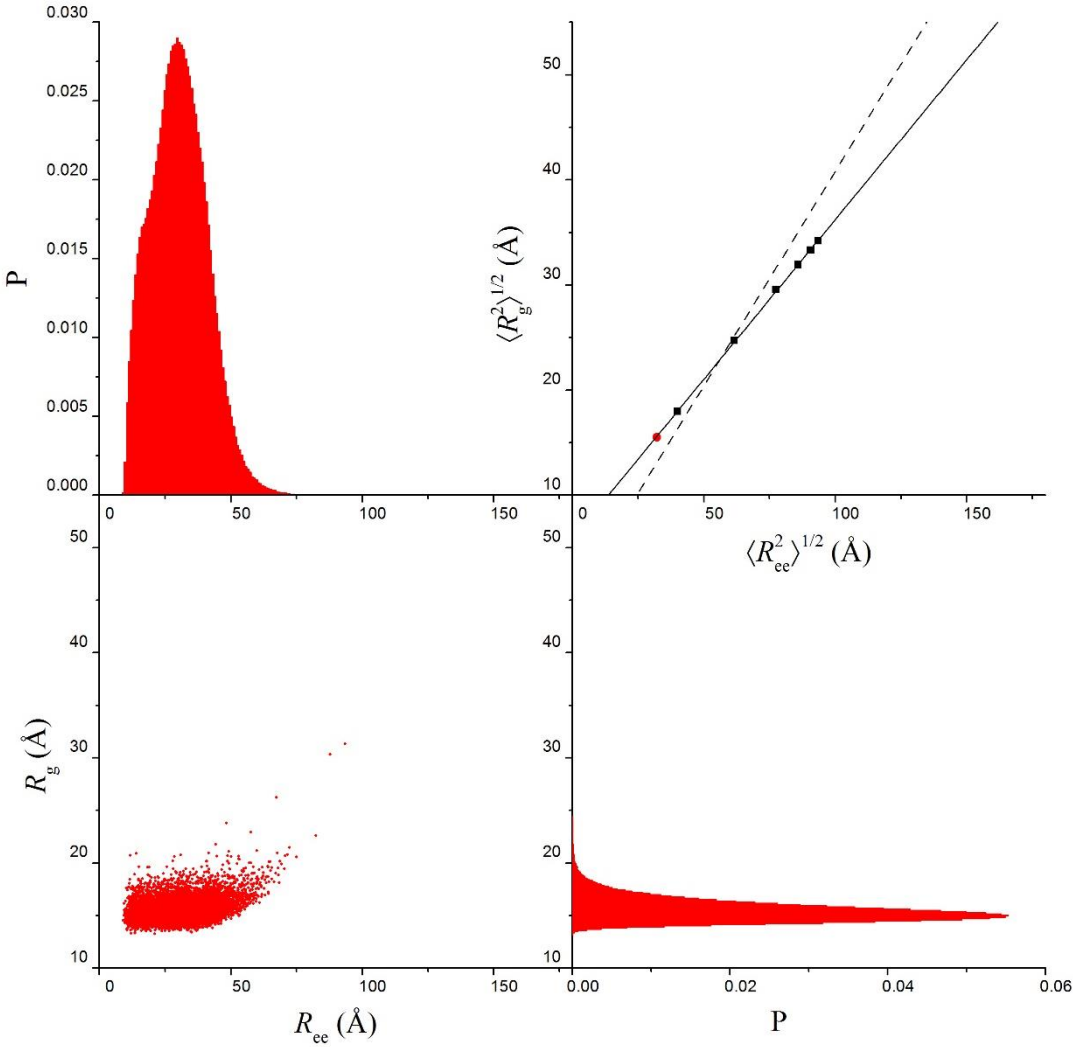

Figure S14 Using a fully labelled protein (N = n = 60) on non-bonding interaction strength λ = 0.4, plotting as Fig. S11.