Supporting Information

Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation

Utsab R. Shrestha[a], Puneet Juneja[b,1], Qiu Zhang[b], Viswanathan Gurumoorthy[c], Jose M. Borreguero[b], Volker Urban[b], Xiaolin Cheng[d], Sai Venkatesh Pingali[b], Jeremy C. Smith[a,e], Hugh M. O'Neill[b], and Loukas Petridis[a,e,2]

[a]UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory Oak Ridge, TN 37831; [b]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; [c]UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996; [d]Division of Medicinal Chemistry and Pharmacognosy, College of Pharmacy, The Ohio State University, Columbus, OH 43210; [e]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996

[1]Present Address: School of Medicine, Emory University, Atlanta, GA 30322.

[2]To whom correspondence may be addressed. E-mail: petridisl@ornl.gov.

Small-angle scattering experiments

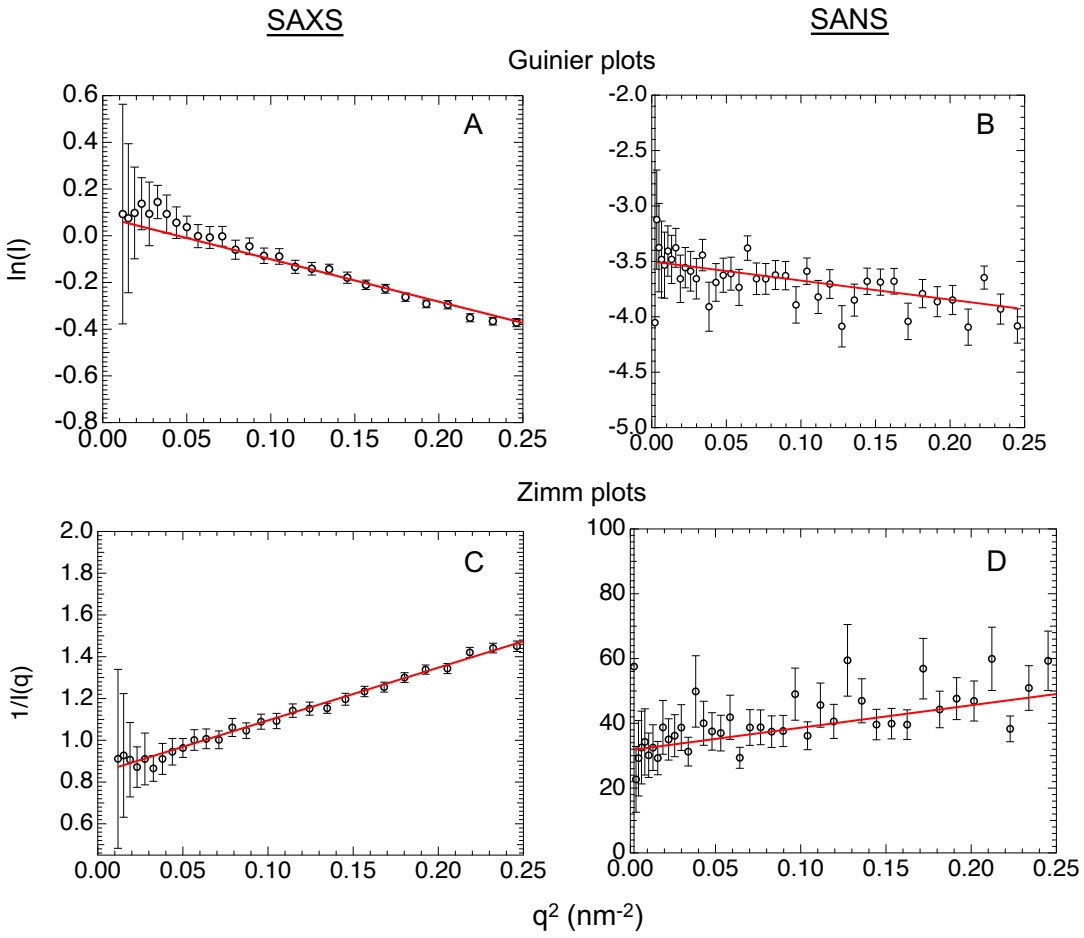

Figure S1: Guinier (A, B) and Zimm (C, D) plots of the SAXS (A, C) and SANS (B, D) experimental data. A linear behavior, indicated by the red line, at low q is consistent with scattering from monodisperse particles.

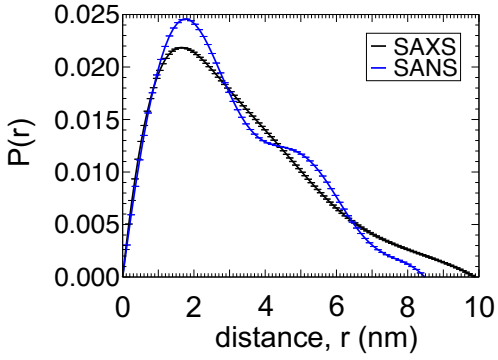

Figure S2. The pair distance distribution function *P(r)*, calculated as the indirect Fourier transform of experimental *I(q)*. *P(r)* is employed to determine the experimental R_g.

Standard MD simulations. Six independent standard MD simulations (five 1.2 μs each and one 3.6 μs, for a total of ~10 μs) were performed starting with different initial velocities using *Amber ff03ws+TIP4P/2005s*. We found that the standard MD simulation trajectories were not able to reproduce the experimental SAXS profile (Fig. S3).

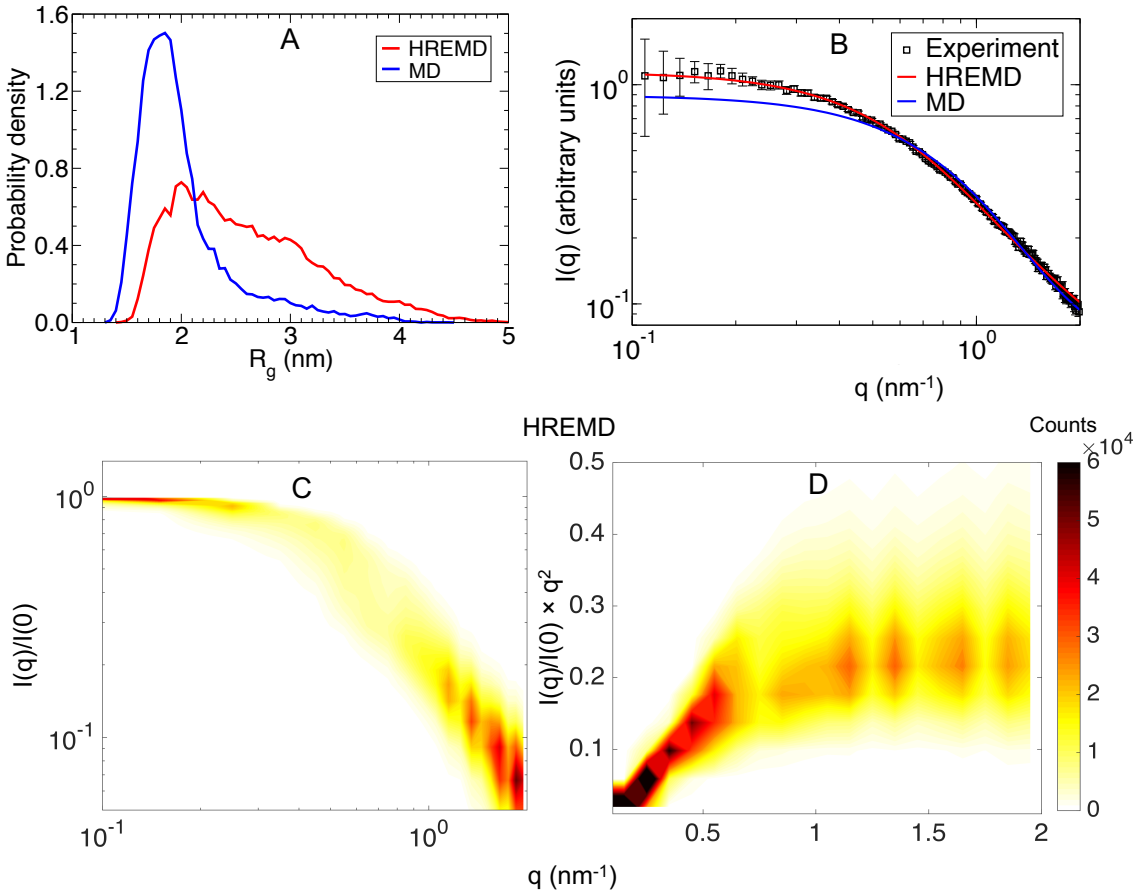

Figure S3. (A) Histograms of R_g obtained from the average of six independent standard MD (blue line) and the lowest rank replica of HREMD (red line) using *Amber ff03ws+TIP4P/2005s*. (B) SAXS profiles obtained from: experiment (black squares); the lowest rank HREMD simulation (red line); and ~10 μs standard MD (blue line) simulations. It is clear that the enhanced sampling provided by HREMD is required here to reproduce the experimental SAXS. Density maps of (C) theoretical HREMD-derived SAXS profiles, normalized at *I(0)=1,* and (D) corresponding Kratky plot.

HREMD simulation. The HREMD simulation was performed at 20 effective temperatures ranging from 300 K to 400 K, with the temperature of i^{th} replica defined by a geometric expression as,

$$T_{eff,i} = T_{eff,0}\, e^{\left[\frac{i}{(n-1)}\ln\left(\frac{T_{eff,max}}{T_{eff,0}}\right)\right]} \quad \text{..... (S1)}$$

with corresponding values of $\lambda_i = T_{eff,0}/T_{eff,i}$ ($\lambda_{max} = 1$ and $\lambda_{min} = 0.75$). The effective temperatures used here were: 300.00 K, 304.54 K, 309.14 K, 313.82 K, 318.56 K, 323.38 K, 328.27 K, 333.24 K, 338.27 K, 343.39 K, 348.58 K, 353.85 K, 359.21 K, 364.64 K, 370.15, 375.75 K, 381.43 K, 387.20 K, 393.06 K, 400.00 K.

The potential energy distribution between neighboring replicas shows good overlap (Fig. S4A), suggesting a sufficient number of replicas is employed and confirming that no phase transition takes place during the HREMD simulation. The exchanges of coordinates as a function of MD steps for replicas 0, 5, 10, 15 and 19 (exchanges were made every 400 MD steps or 0.8 ps) are shown in Fig. S4B, C, D, E and F respectively. These data illustrate a good sampling with frequent exchange of coordinates across all the neighboring replicas. The average acceptance probability at different replicas varies from 0.58 to 0.64.

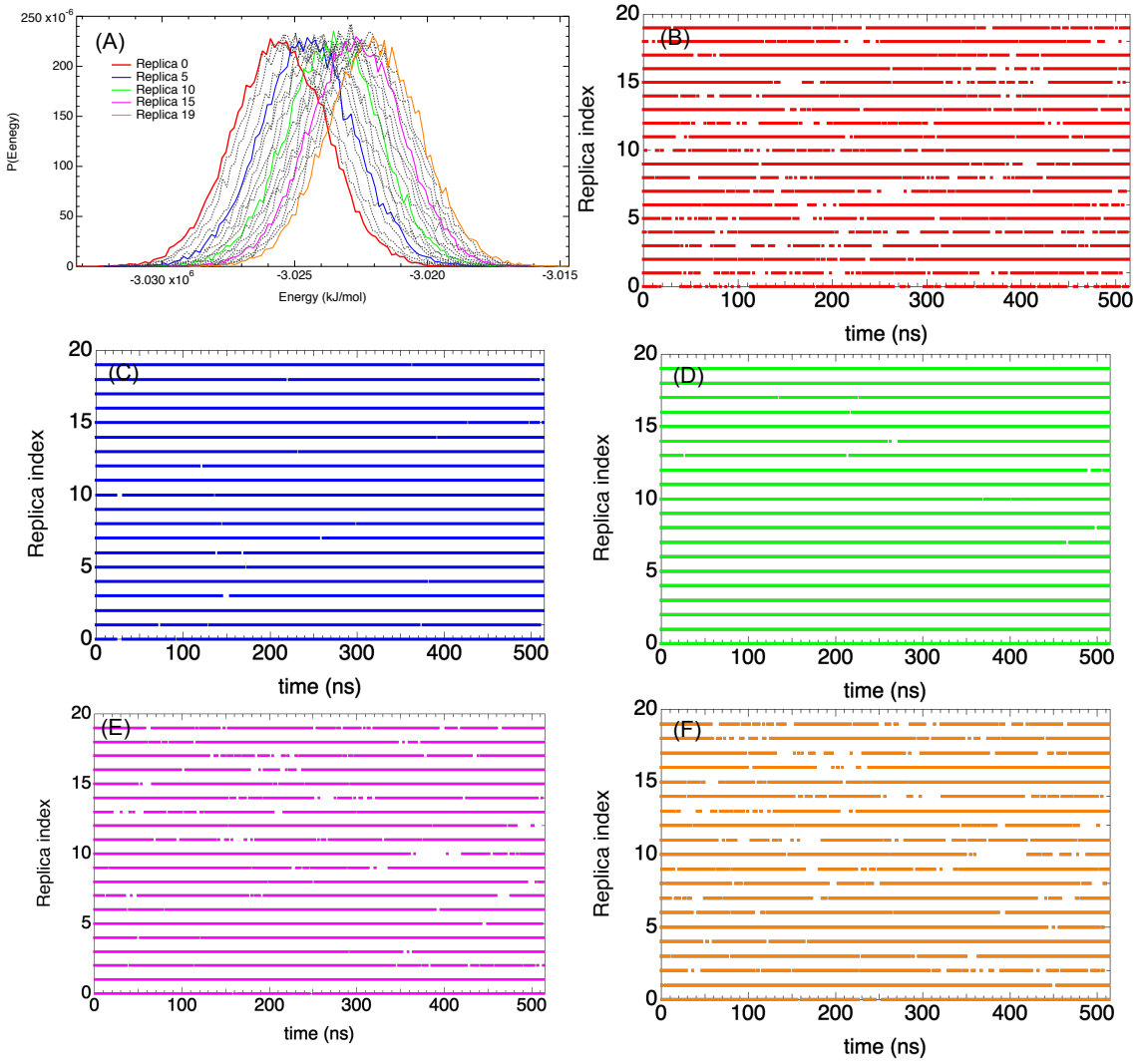

Figure S4. (A) Potential energy distributions of HREMD simulations for all the replicas. The solid red, blue, green, magenta and orange curves represent the replica 0, 5, 10, 15 and 19 respectively and the remaining replicas are shown by dashed gray curves. (B, C, D, E, F) The exchange of coordinates as a function of simulation time for five different replicas 0, 5, 10, 15 and 19.

Error analysis. The lowest rank trajectory from HREMD was divided into five blocks, each containing 10,277 frames. The frames in each block were either chosen randomly, or were consecutive in simulation time: 0-103 ns, 103-206 ns, 206-309 ns, 309-412 ns, 412-514 ns. Both sampling approaches yielded similar values of uncertainty. For all the

reported quantities, the mean value for each block, m_i (i=1 to 5), was first calculated. The reported error bars are the standard error of the mean of the $\{m_1, m_2, m_3, m_4, m_5\}$ distribution.

Error bar $= \sqrt{\frac{1}{n(n-1)} \sum_i^{n=5} (m_i - \overline{m})^2}$,

where $\overline{m}$ is the mean value and n=5 is the number of blocks used here.

Convergence of calculation of SAXS profiles from HREMD simulation. Variation in X-ray *I(q)* is quantified here by the cumulative χ^2 between experimental and theoretical SAXS (Fig. S5), defined in Eq 2. The χ^2 is found to decrease in the first ~300 ns and then vary little (Fig. S5F), indicating that this quantity has reached convergence.

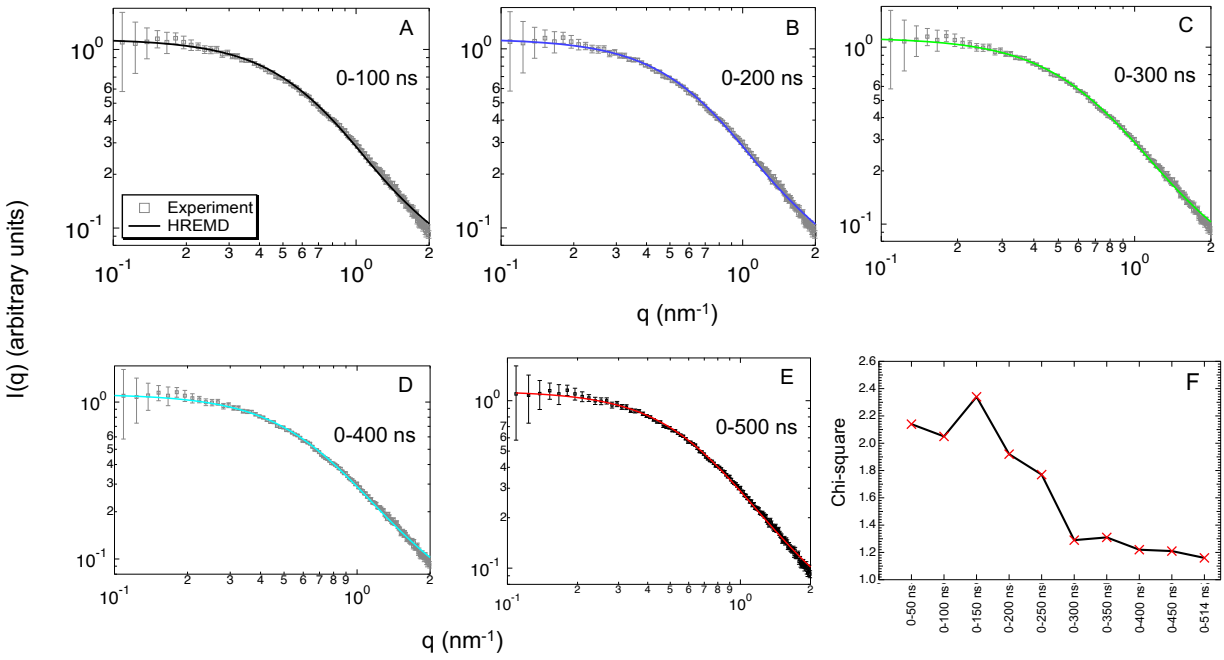

Figure S5: Comparison between the experimental SAXS profiles and those obtained from lowest rank HREMD trajectories of different length: 0-100 ns (A), 0-200 ns (B), 0-300 ns (C), 0-400 ns (D) and 0-514 ns (E). (F) Chi-square between experimental and theoretical SAXS intensities calculated from trajectories of varying length.

Convergence of calculation of SANS profiles from HREMD simulation. Variation in neutron *I(q)* is quantified here by the cumulative χ^2 between experimental and theoretical SAXS (Fig. S6), defined in Eq 2. The $\chi^2 < 1.4$ after ~400 ns (Fig. S6F), indicating good agreement with experiment.

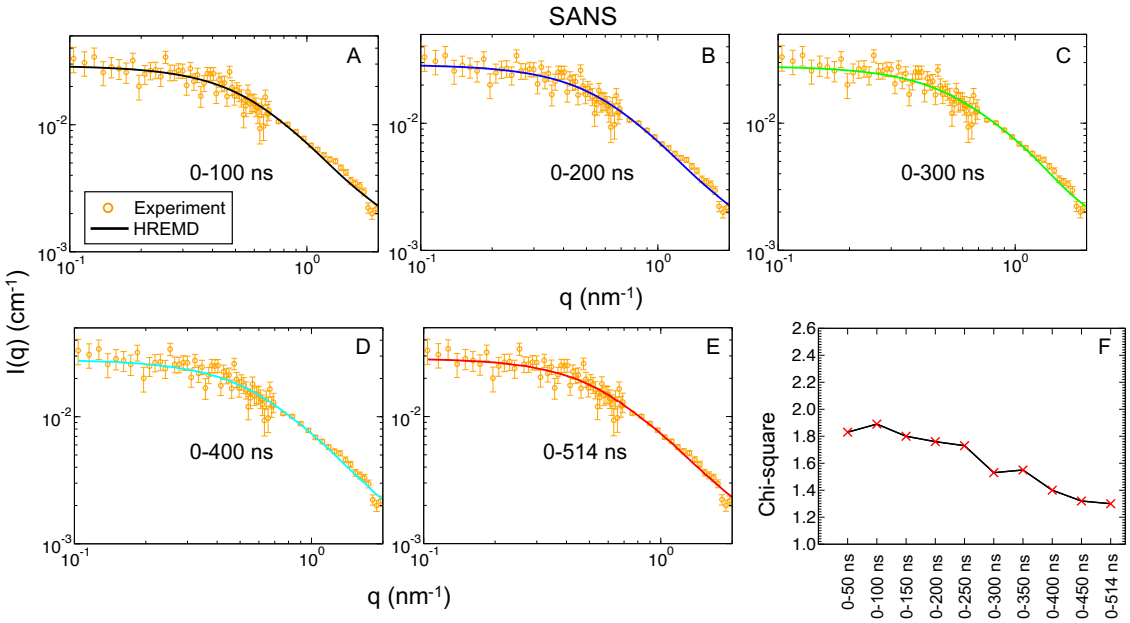

Figure S6: Comparison between the experimental SANS profiles and those obtained from HREMD trajectories of different length: 0-100 ns (A), 0-200 ns (B), 0-300 ns (C), 0-400 ns (D) and 0-514 ns (E). (F) Chi-square between experimental and theoretical SAXS intensities calculated from trajectories of varying length.

Convergence of calculation of NMR chemical shifts from HREMD. The NMR chemical shifts of backbone atoms are related to local structural properties of the protein, and are found here to converge in ~50 ns and ~250 ns for N^H and C^α respectively (Fig. S7), faster than the SAXS profile. The excellent agreement between NMR and HREMD is reflected by the regression coefficients $R^2 > 0.93$.

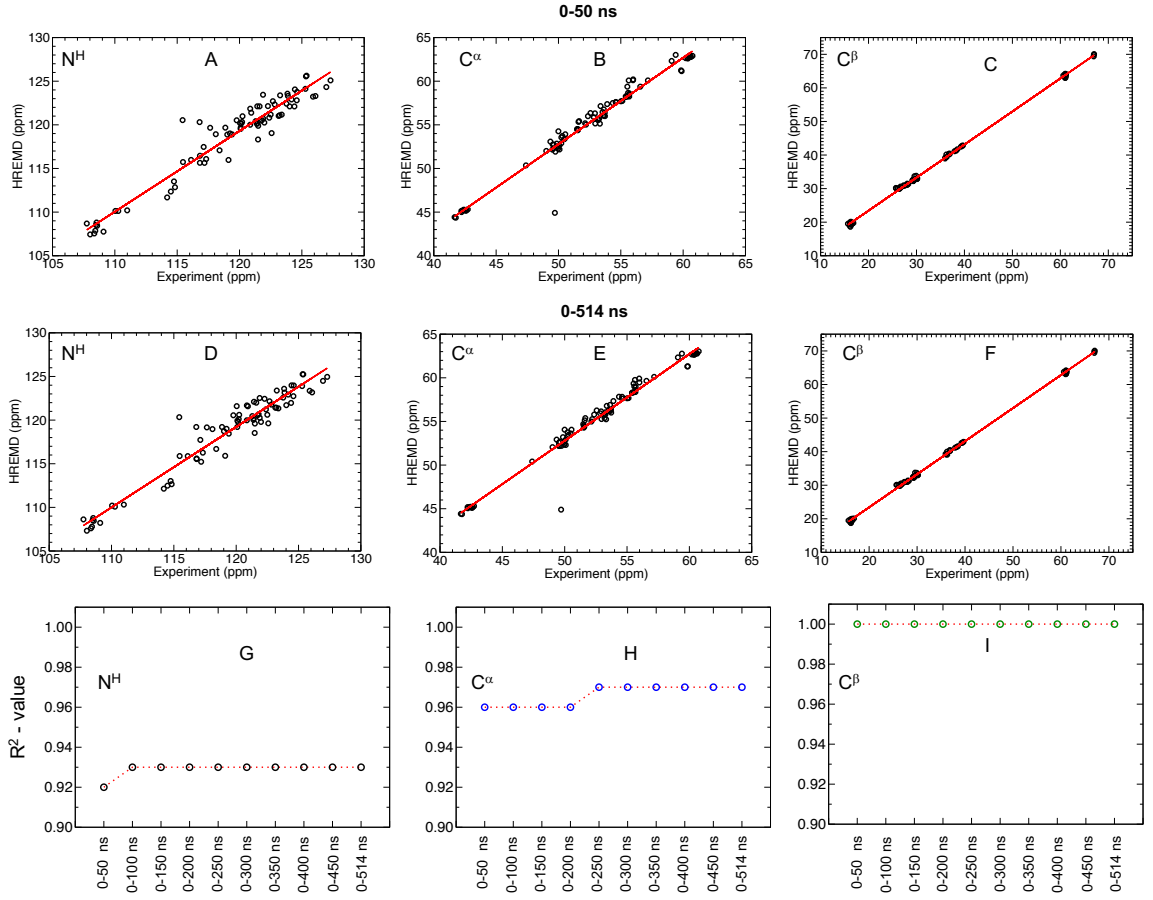

Figure S7. Comparison between experimental and HREMD NMR chemical shifts of backbone atoms, N^H, C^α and C^β of SH4UD, expressed in parts per million (ppm). The experimental data are taken from BMRB Entry 15563 (1). The theoretical data were calculated using SHIFTX2 (2) from the first 50 ns of the trajectory in (A, B, C) and from the entire trajectory in (D, E, F). The linear regression coefficient between experiment and theory is plotted as a function of simulation length in (G, H, I).

Convergence of HREMD free energy landscape. The distributions of three reaction coordinates of particular importance for IDPs (the R_g, asphericity and SASA, which describe the global protein configurations) are found to show fluctuations at short times that smooth out and not vary significantly after ~350 ns of HREMD (Fig. S8A,B,C). The associated potentials of mean force (PMF) also do not vary significantly after 350 ns (Fig. S8D,E,F).

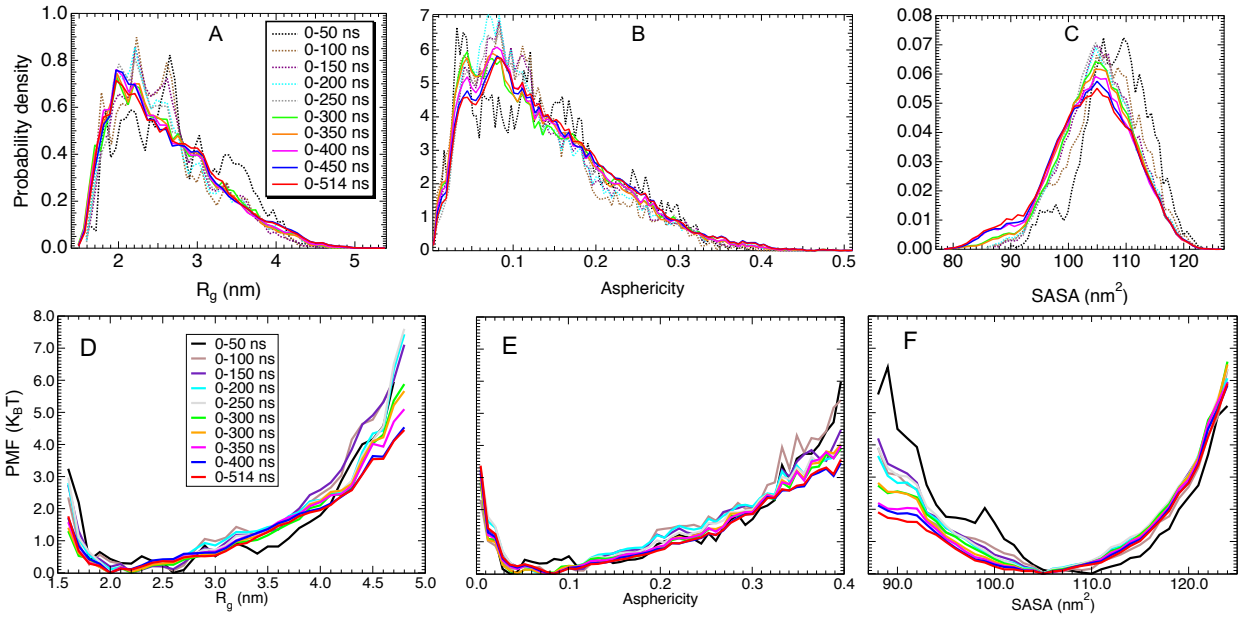

Figure S8. Probability distributions (A-C), and potentials of mean force (D-F) for three reaction coordinates: radius of gyration, asphericity and solvent accessible surface area.

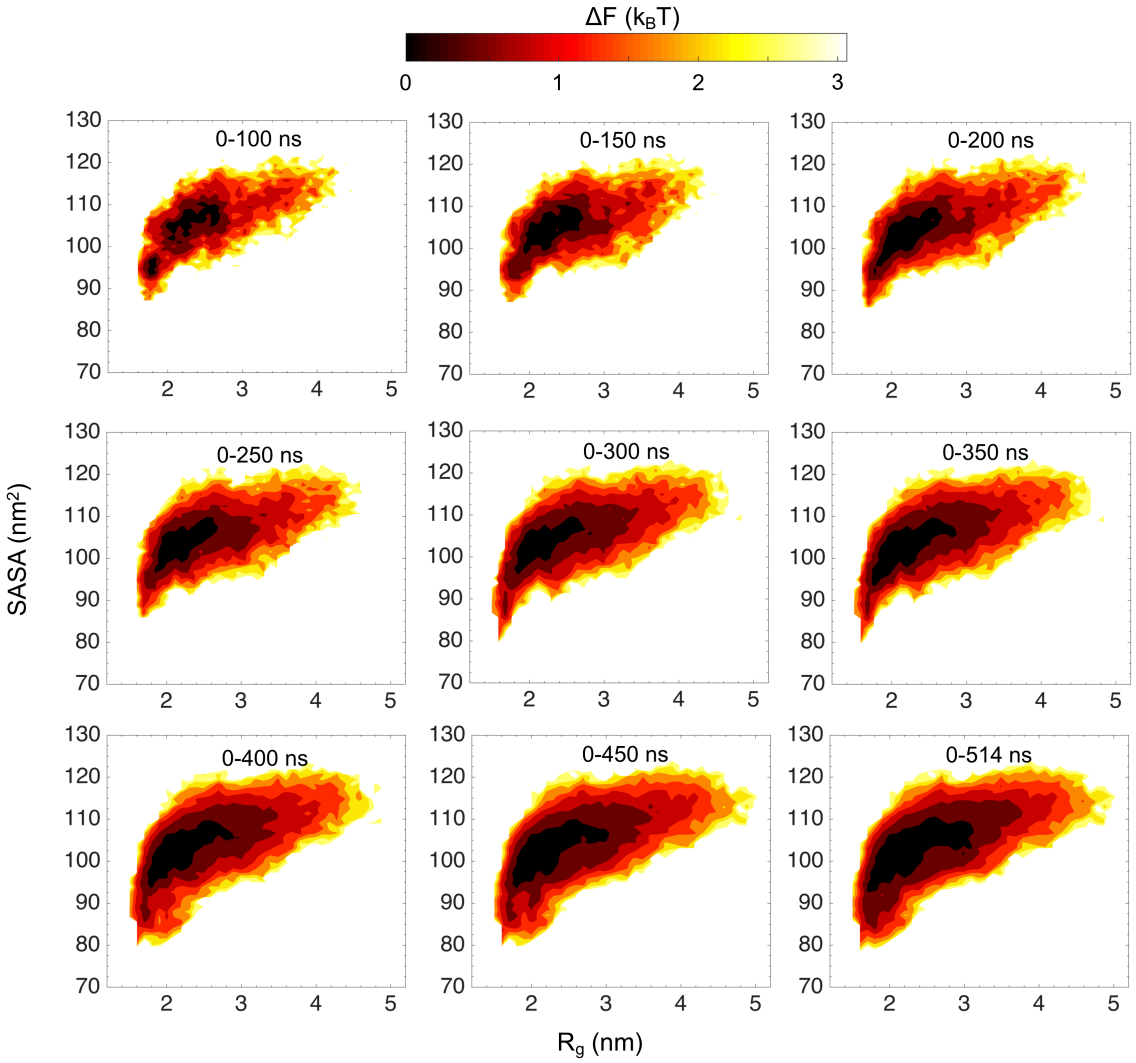

Figure S9. Free energy landscape projected on the radius of gyration and solvent accessible surface area as reaction coordinates, and plotted as a function of simulation length.

Pure coil structures in HREMD.

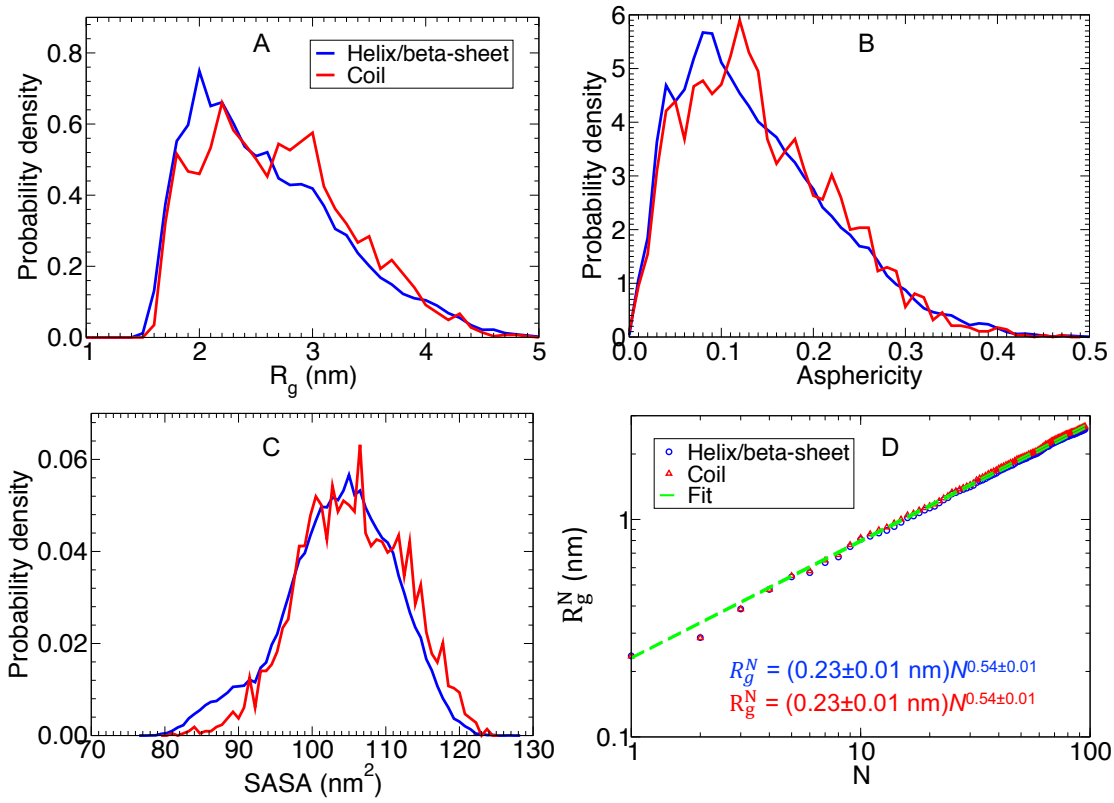

Figure S10. Analysis on lowest rank HREMD trajectory divided into two parts with ensemble of structures: (blue) having at least one helix or beta-sheet and (red) entirely coil. Histograms of (A) the radius of gyration, (B) asphericity and (C) solvent accessible surface area (SASA). (D) Radius of gyration (R_g^N) of a protein segment consisting of *N* residues. The fits of Eq. 1 to both ensembles (either entirely coil or at least 1 helix or beta-sheet) yield the same power-law behavior ($\nu = 0.54\pm0.01$).

Hydration shell of HEWL and SH4UD.

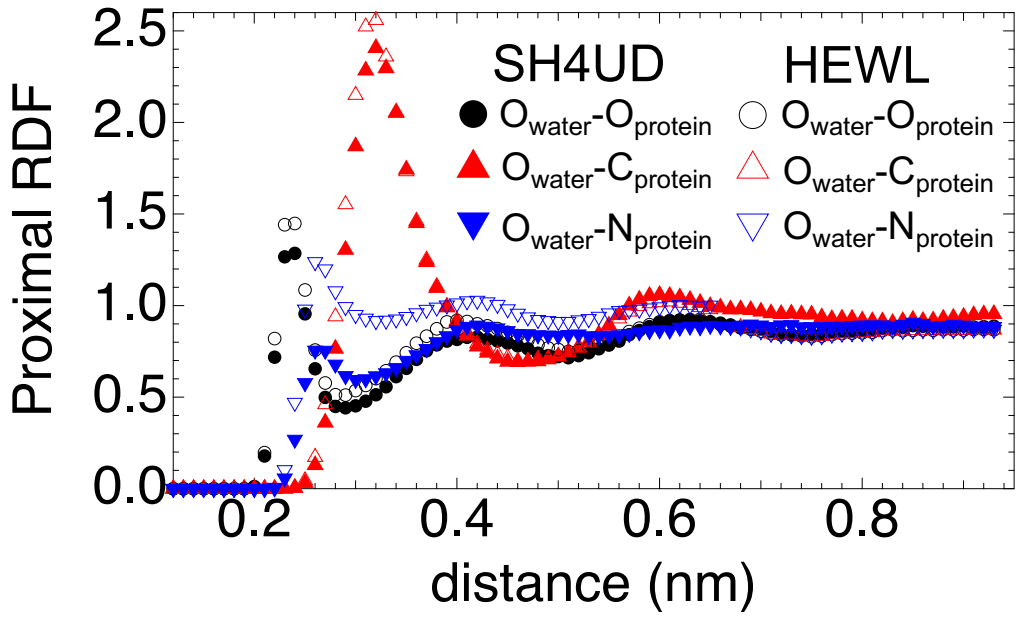

Figure S11. Decomposition of the proximal radial distribution function (pRDF) of Oxygen atom of water to contributions from protein Oxygen, O; Carbon, C; and Nitrogen, N atoms.

The surface residues in HEWL exposed to solvent are calculated using Swiss PDB viewer (https://spdbv.vital-it.ch/disclaim.html) and are listed in Table S1, whereas we assumed all the residues are exposed to solvent in SH4UD.

Table S1. Fraction of hydrophobic and polar residues on the surfaces of HEWL and SH4UD.

Protein	Fraction of surface residues	
	Hydrophobic	Polar/charged
HEWL	0.28	0.72
SH4UD	0.54	0.46

Table S2. Surface residues in HEWL exposed to solvent calculated from its crystal structure (PDB 1LYZ).

RESID	RESNAME	Property
1	LYS	+ charged
2	VAL	Hydrophobic
14	ARG	+ charged
19	ASN	Polar, neutral
21	ARG	+ charged
22	GLY	Hydrophobic
37	ASN	Polar, neutral
41	GLN	Polar, neutral
43	THR	Polar, neutral
44	ASN	Polar, neutral
45	ARG	+ charged
47	THR	Polar, neutral
48	ASP	- charged
62	TRP	Hydrophobic
67	GLY	Hydrophobic
68	ARG	+ charged
70	PRO	Hydrophobic
73	ARG	+ charged
75	LEU	Hydrophobic
77	ASN	Polar, neutral
81	SER	Polar, neutral
86	SER	Polar, neutral
87	ASP	- charged
93	ASN	Polar, neutral
97	LYS	+ charged
101	ASP	- charged
102	GLY	Hydrophobic
103	ASN	Polar, neutral
107	ALA	Hydrophobic
109	VAL	Hydrophobic
112	ARG	+ charged
113	ASN	Polar, neutral
114	ARG	+ charged
116	LYS	+ charged
117	GLY	Hydrophobic
119	ASP	- charged
121	GLN	Polar, neutral
125	ARG	+ charged
126	GLY	Hydrophobic
128	ARG	+ charged

Classification of SH4UD based on sequence .

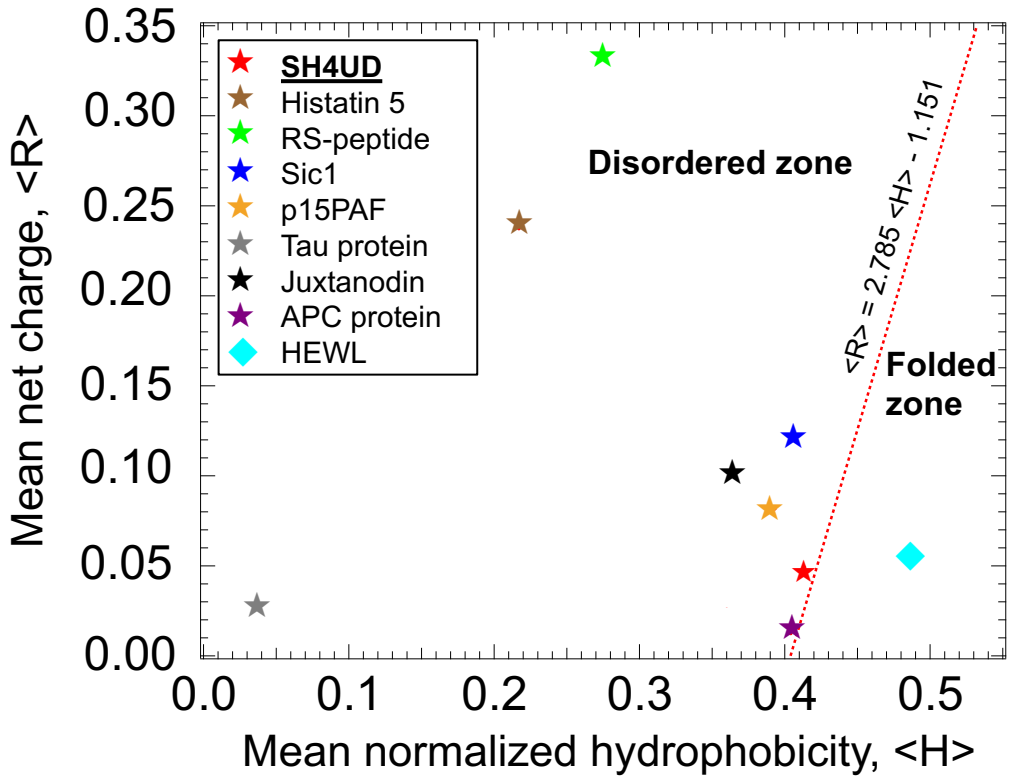

Figure S12. Charge-hydrophobicity phase diagram for 8 IDPs (Histatin 5, RS-peptide, Sic1, p15PAF, Tau protein, Juxtanodin, APC protein, and SH4UD) and a globular protein, hen egg white lysozyme (HEWL). The dashed red line represents the border between disordered and folded protein given by a relation, $\langle R \rangle = 2.785 \langle H \rangle - 1.151$, where <R> and <H> are the absolute mean net charge and mean normalized hydrophobicity of protein sequence (4-6). SH4UD falls within the disordered zone, but close the border line. Here, mean normalized hydrophobicity is calculated by Kyte and Doolittle approximation (7).

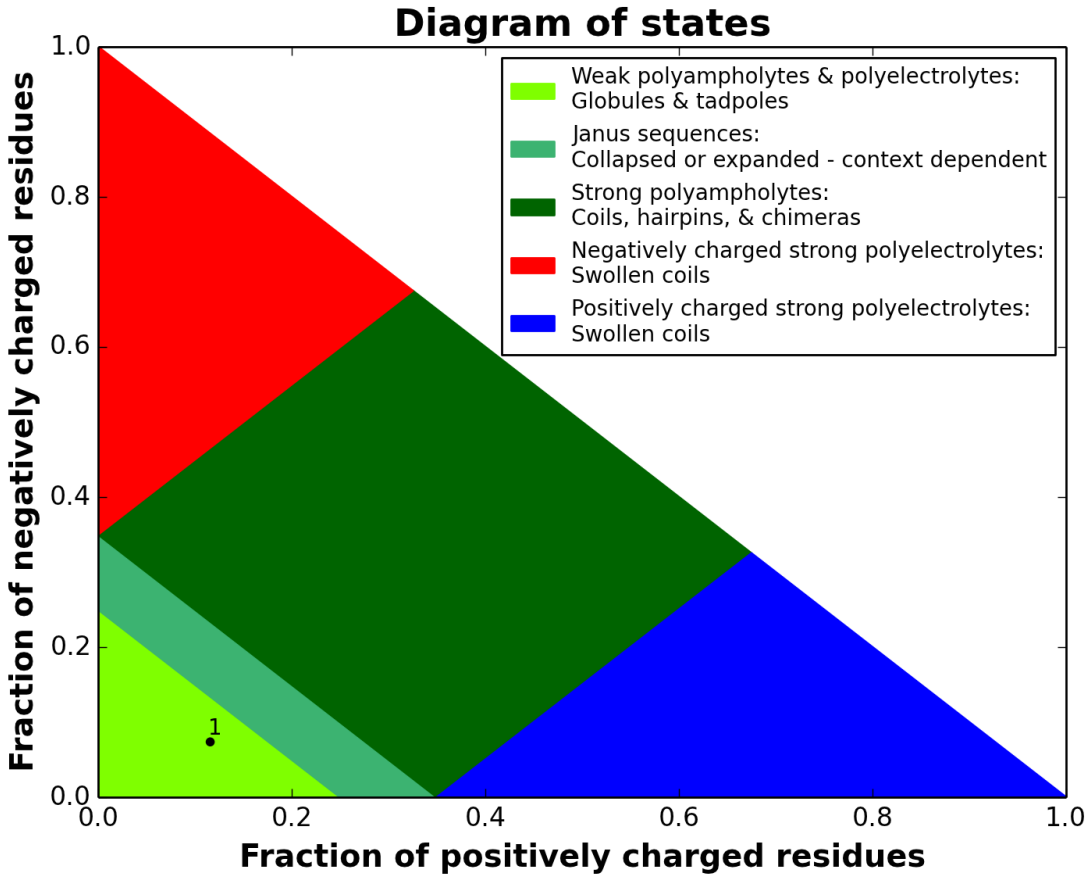

Figure S13. Prediction of the SH4UD conformation, labelled with a black dot, based on its sequence, using the CIDER server (8). IDP sequences are classified into distinct conformational classes based on their amino acid compositions.

REFERENCES

1. Pérez Y, Gairí M, Pons M, & Bernadó P (2009) Structural Characterization of the Natively Unfolded N-Terminal Domain of Human c-Src Kinase: Insights into the Role of Phosphorylation of the Unique Domain. *Journal of Molecular Biology* 391(1):136-148.
2. Han B, Liu Y, Ginzinger SW, & Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. *J Biomol NMR* 50(1):43-57.
3. Syakur MA, Khotimah BK, Rochman EMS, & Satoto BD (2018) Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Conference Series: Materials Science and Engineering* 336.
4. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 11(4):739-756.
5. Uversky VN (2011) Intrinsically disordered proteins from A to Z. *Int J Biochem Cell Biol* 43(8):1090-1103.
6. Uversky VN, Gillespie JR, & Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics* 41(3):415-427.
7. Kyte J & Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology* 157(1):105-132.
8. Holehouse AS, Das RK, Ahad JN, Richardson MO, & Pappu RV (2017) CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophys J* 112(1):16-21.