

Supporting information

Aromatic-Aromatic Interactions Drive Fold Switch of GA95 and GB95 with Three Residue Difference

Chen Chen^{a,b,1}, Zeting Zhang^{a,b,1 *}, Mojie Duan^{c *}, Qiong Wu^a, Minghui Yang^{a,b}, Ling Jiang^{a,b}, Maili Liu^{a,b},
Conggang Li^{a,b *}

Table of contents:

1. Materials and Methods	2
2. Supplementary Figures and Tables	4
3. References	24

1. Materials and Methods

1.1 Protein Expression and Purification. The DNA sequence encoding His-SUMO-GA95/GB95 was cloned into the vector pET21a and used as a template for targeted mutagenesis to obtain the truncated proteins with different sequence lengths, designed as His-SUMO-GA95-n and His-SUMO-GB95-n (n is the number of residues). Mutants were created using PCR mutagenesis and confirmed by DNA sequencing. All proteins were expressed in *E. coli* BL21 (DE3). The organisms were grown in an M9 medium containing 100 mg/mL ampicillin labeled with ^{15}N or 3FY (3-DL-tyrosine) and ^{15}N . Upon reaching an OD_{600} of 0.8 at 37 °C, protein expression was induced with 0.5 mM IPTG (Isopropyl β -D-1-thiogalactopyranoside). GA95 was induced at 20 °C for 16-24 hours, while GB95 was induced at 15 °C. Bacteria were harvested by high-speed centrifugation, resuspended in a solution of 50 mM Tris (pH 8.0), 500 mM NaCl, and 10 mM imidazole with an appropriate amount of protease inhibitors, and lysed under high pressure. The clarified bacterial lysate was collected by centrifugation. Fusion proteins were purified using a Ni^{2+} -NTA resin column, followed by desalting into a buffer containing 50 mM Tris (pH 7.2), and 150 mM NaCl. ULP1 protease was added at a 1:2 ratio for a 3-hour cleavage reaction at 15 °C. The His-SUMO fusion tags were separated from GA95 and GB95 using a Ni^{2+} -NTA resin column, and the effluent of the protein solution was collected to obtain the target proteins GA95 and GB95. The protein solution was concentrated from 0.03 to 0.2 mM (100 mM KH_2PO_4 , 200 mM KCl, pH 7.2) for NMR analysis. The protein solution was snap-frozen using liquid nitrogen and stored at -80 °C for later use.

1.2 NMR Spectroscopy. ^{15}N and $^{15}\text{N}/3\text{FY}$ -labeled protein samples were dissolved in the NMR buffer containing 100 mM KH_2PO_4 and 200 mM KCl (pH 7.2), supplemented with 5%-10% D_2O . The concentrations ranged from 0.03 to 0.2 mM, with total volumes were approximately 350-500 μL . The NMR spectra were acquired on a 600 MHz, 700 MHz, and 850 MHz spectrometer (Bruker) equipped with a z-axis-gradient ($^1\text{H}/^{13}\text{C}/^{15}\text{N}$) resonance cryoprobe, with the 600 MHz and 700 MHz additionally equipped with a ^{19}F cryoprobe. The NMR spectra of GA95, GB95, and their truncated or mutated variants were collected at 288 K, while the NMR spectrum of GA98 was collected at 279 K. (1) Acquisition parameters of 1D ^{19}F NMR spectra: the spectral width (SW) is set to 30 ppm; spectral center (O1P) is set to -135 ppm; the number of sampling points (TD) is 16384; the number of scans (NS) is adjusted appropriately according to the concentration of the protein samples, generally ranging from 512 to 4096 scans. (2) Acquisition parameters of 2D ^1H - ^{15}N HSQC spectra: the spectral width (SW) of ^1H is set to 16 ppm; the spectral center (O1P) of ^1H is set to 4.7 ppm; the spectral width (SW) of ^{15}N is set to 40 ppm; the spectral center (O1P) of ^{15}N is set to 117 ppm; the number of sampling points (TD) is set to 2048 \times 256; the number of scans (NS) was adjusted appropriately according to the concentration of protein samples, generally ranging from 16 to 64 scans. (3) All data were processed and analyzed using Sparky (Goddard and Kneller, University of California, San Francisco) and Topspin (Bruker BioSpin).

1.3 Circular Dichroism. Protein samples stored at -80 °C were thawed or the lyophilized powder was dissolved in a buffer of 100 mM KH_2PO_4 and 200 mM KCl (pH 7.2). The protein samples were diluted 10-fold to achieve a final concentration of approximately 10-20 μM . Circular dichroism (CD) spectra in the range of 190 to 260 nm were recorded using a circular dichroism spectrometer (Chirascan) at an experimental temperature of 288 K. Measurements were conducted in a quartz cuvette with a path length of 1 cm and a total volume of 500 μL volume. Each sample was measured three times and the averaged spectrum was obtained after subtracting the background spectrum of the buffer. The raw mdeg values were transformed in mean residue molar ellipticity ($\text{deg}\cdot\text{cm}^2\cdot\text{dmol}^{-1}$) using the following equation:

$$[\theta] = \frac{CD\ signal(deg) \cdot MRW}{concentration(g/L) \cdot l \cdot 10}$$

Where MRW is the mean residue weight and l is the path length of the CD cell in cm. The corrected CD spectral data were analyzed using input into the CDNN software for secondary structure analysis. In CDNN software, the CD signal type (Milli-Degrees) was selected, and the secondary structure content of the proteins was calculated based on the input data (molecular weight, protein concentration, and number of amino acids), including α -helix, β -sheet, and random coiling. Further data processing was performed using Origin software (OriginLab Corporation, Northampton, MA, USA).

1.4 MD simulation setup. All-atom MD simulations were performed to study the structure and folding of GA95-53 and GB95. The amber99SB-ILDN force fields¹ were employed for the proteins and the water molecules were modeled using the TIP3P model². The initial structures of simulations were built based on the experimental models (PDB id: 1TFO³ and 1FCC⁴, respectively). Proteins were solvated by water molecules in the cubic boxes with a volume of 8.5 nm³. To obtain the initial structures of GA95-53 and GB95 for REST2 simulations, high-temperature (500 K) simulations were performed. A two-step procedure was employed to minimize the systems. Firstly, the restraint force of 100 kcal/mol·Å² was exerted on the protein. The solvent and ion molecules were optimized by 2000 cycles of steepest descent and 2000 cycles of conjugate gradient minimizations. Then, the whole system was optimized by 2000 cycles of steepest descent and 2000 cycles of conjugate gradient minimizations by removing the restraints. Next, the minimized system was heated from 0 to 500 K step-by-step over a period of 20 ps and then equilibrated over 1000 ps in the NVT ensemble. The system was then equilibrated in the NPT ensemble (T = 500 K and P = 1 bar) for 100 ps. The temperature was controlled by the Langevin thermostat scheme⁵ with a collision frequency of 1.0 ps⁻¹. The particle mesh Ewald algorithm⁶ was used to handle the long-range electrostatic interactions under the periodic boundary condition, and a cutoff of 10.0 Å was used for the real space interactions. The SHAKE algorithm⁷ was used to constrain all covalent bonds involving hydrogen atoms, and the time step was set to 2 fs. At last, 100-ns production runs under high temperatures were performed.

1.5 REST2 simulations. Replica exchange with solute scaling (REST2)⁸ was employed to study the folding processes of GA95-53 and GB95. REST2 serves as an exceptionally efficient variation of replica exchange, significantly enhancing sampling in explicit solvent simulations of biomolecules. By selectively scaling the Hamiltonian for a designated "solute" segment of the system, REST efficiently targets tempering specifically to the degrees of freedom of interest, while excluding the rest of the system (i.e., the "solvent"). This approach reduces the need for a higher number of replicas to cover the same temperature span. In total, 16 replica REST2 simulations were utilized for GA95-53 and GB95, the scaling values for 16 replicas are set to be 1, 0.975, 0.949, 0.925, 0.9, 0.877, 0.855, 0.833, 0.811, 0.79, 0.77, 0.75, 1, respectively. The convergency analysis and transition probability analysis of the REST2 simulations are given in Fig. S17 and Fig. S18.

1.6 AlphaFold2 predictions. AF2 predictions were conducted in monomer mode. GA95 and GB95 sequences obtained from the PDB were employed as templates and adjusted for sequence truncation length and mutation sites. After the AF2 calculations, five mock structures were generated for each sequence, each accompanied by an associated pLDDT score. The model with the highest pLDDT score was selected as the representative conformation for that sequence. All sequences involved in the simulation were analyzed as described above.

2. Supplementary Figures and Tables

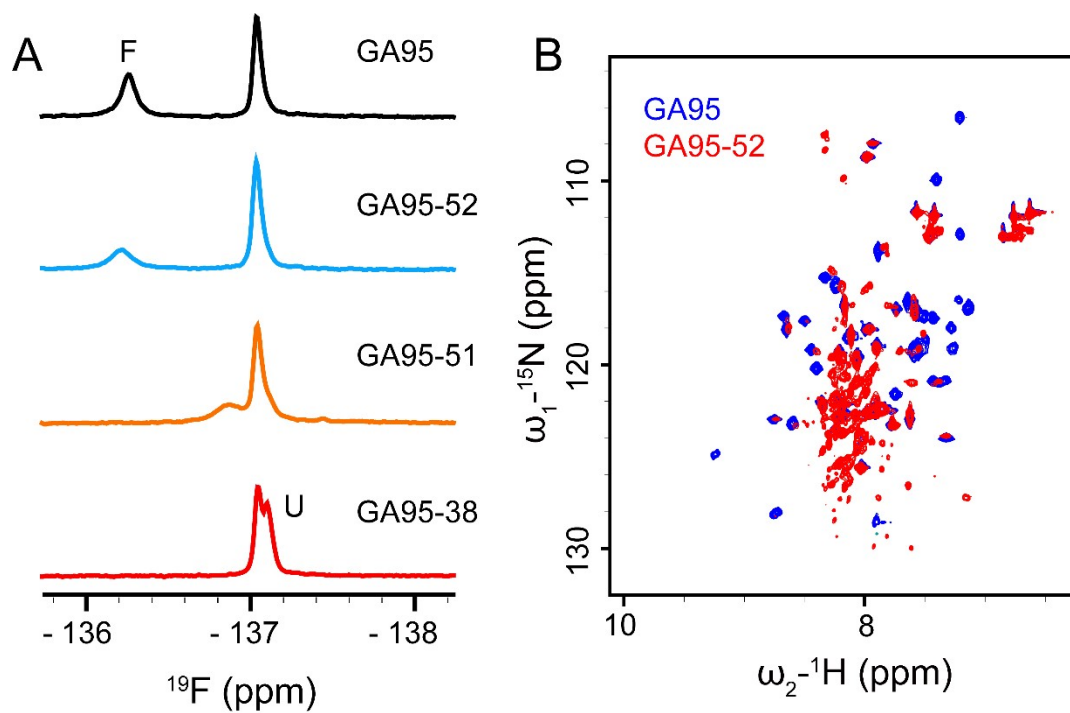


Fig. S1. One-dimensional ^{19}F NMR spectra and ^1H - ^{15}N HSQC spectra of GA95 and its truncated proteins. (A) In the one-dimensional ^{19}F NMR spectrum, GA95-38 is marked with red, GA95-51 is marked with orange, GA95-52 is marked with blue, and GA95 is marked with black. The folded (F) and unfolded (U) states of the protein are labeled. (B) Superposition of ^1H - ^{15}N HSQC spectra of GA95-52 and GA95. GA95 is marked in blue, GA95-52 is marked in red.

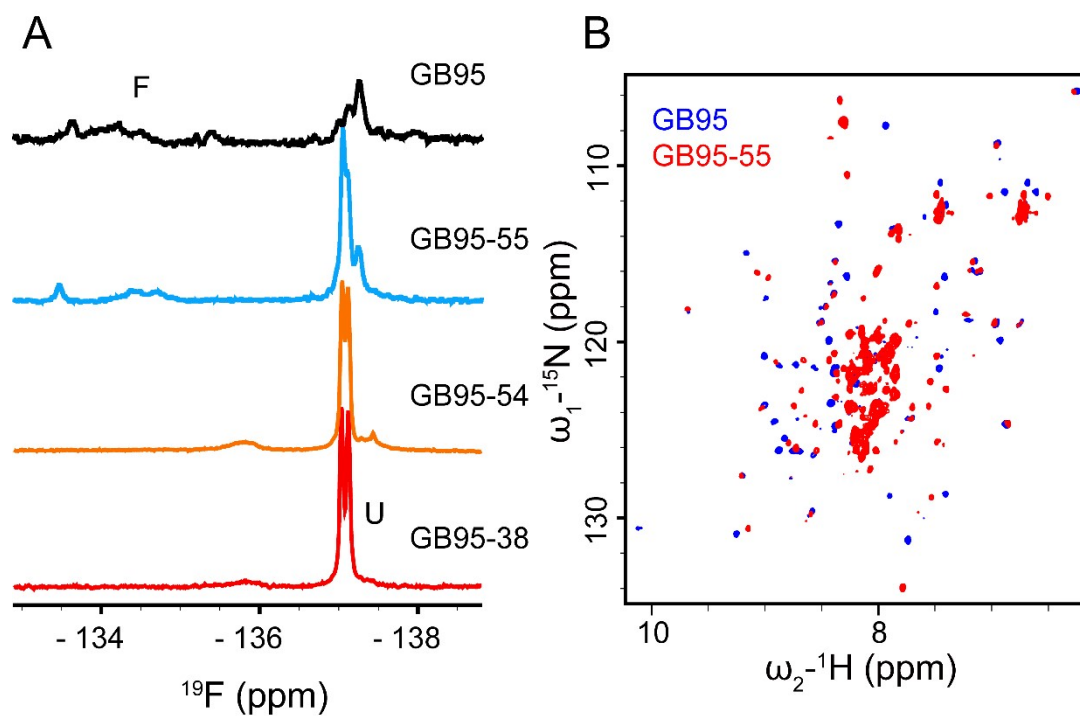


Fig. S2. One-dimensional ^{19}F NMR spectra and ^1H - ^{15}N HSQC spectra of GB95 and its truncated proteins. (A) In the one-dimensional ^{19}F NMR spectrum, GB95-38 is marked with red, GB95-54 is marked with orange, GB95-55 is marked with blue, and GB95 is marked with black. The folded (F) and unfolded (U) states of the protein are labeled. (B) Superposition of ^1H - ^{15}N HSQC spectra of GB95-55 and GB95. GB95 is marked in blue, GB95-55 is marked in red.

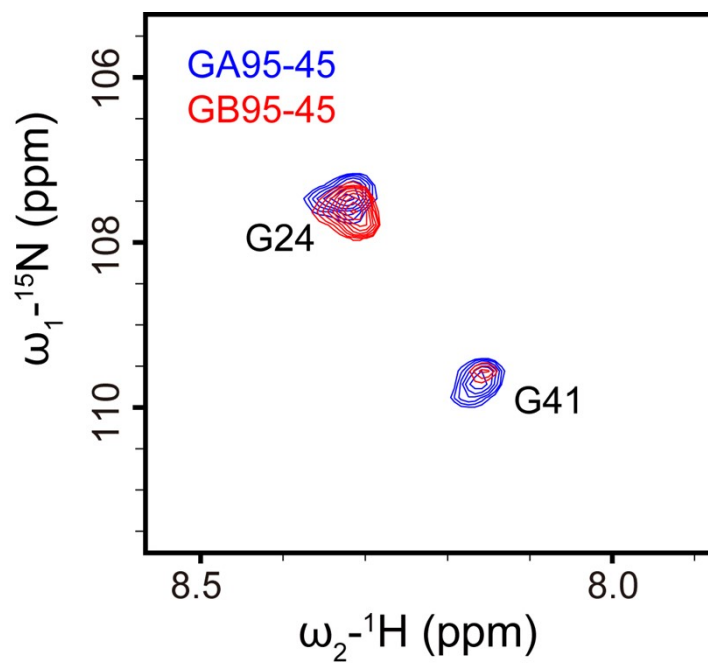


Fig. S3. Signal comparison of residues Gly24 and Gly41. The ^1H - ^{15}N HSQC spectra of GA95-45 (blue) and GB95-45 (red).

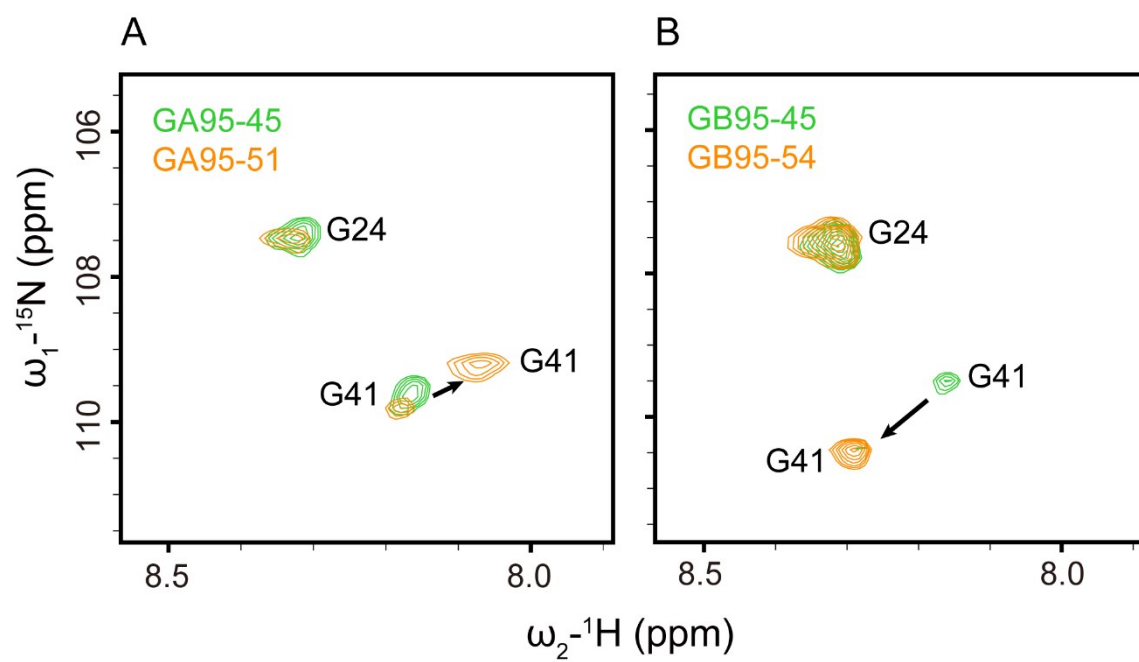


Fig. S4. Chemical shift changes in the signals of residues Gly24 and Gly41. (A) The ^1H - ^{15}N HSQC spectra of GA95-45 (green) and GA95-51 (orange); (B) The ^1H - ^{15}N HSQC spectra of GB95-45 (green) and GB95-54 (orange).

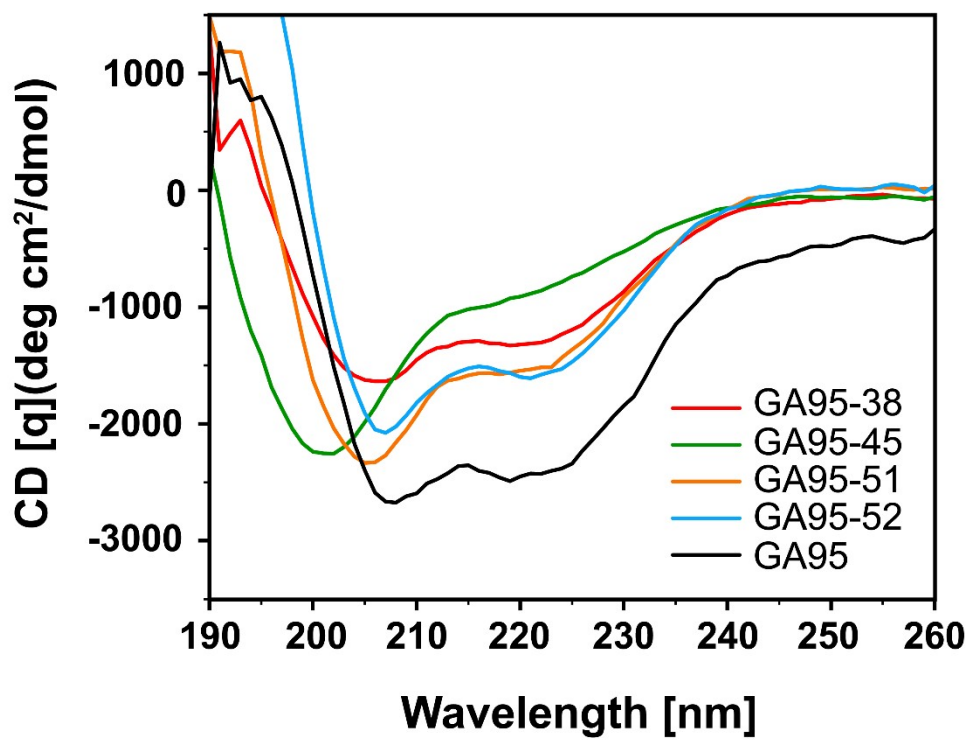


Fig. S5. The circular dichroism spectra of GA95 and its truncated proteins. GA95-38 is marked with red, GA95-45 is marked with green, GA95-51 is marked with orange, GA95-52 is marked with blue, and GA95 is marked with black.

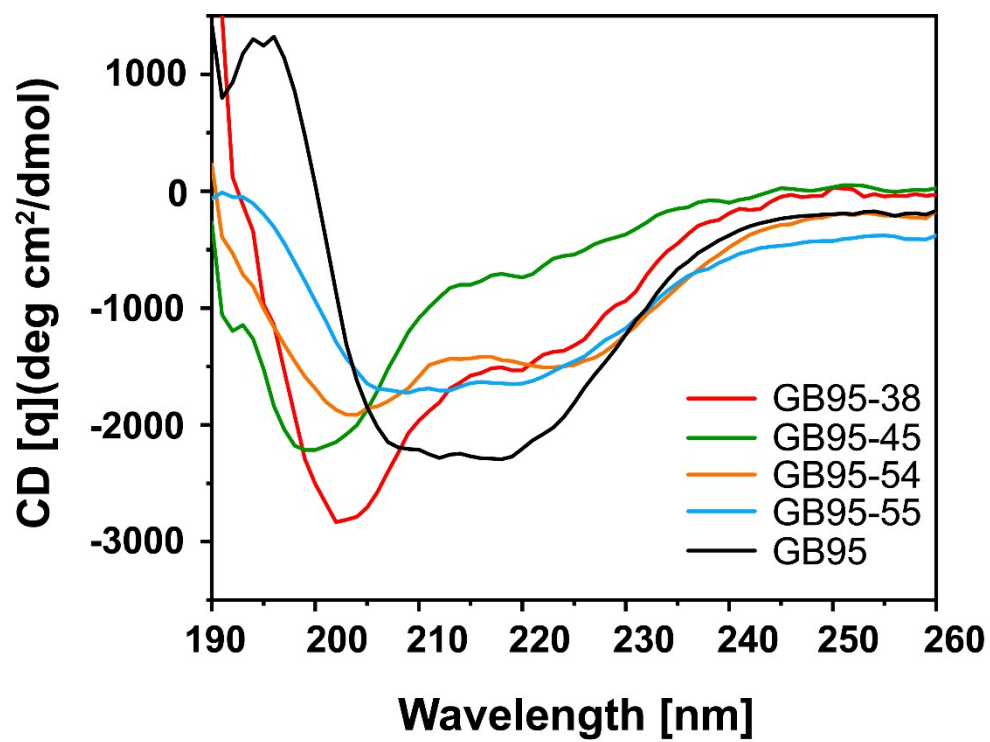


Fig. S6. The circular dichroism spectra of GB95 and its truncated proteins. GB95-38 is marked with red, GB95-45 is marked with green, GB95-54 is marked with orange, GB95-55 is marked with blue, and GB95 is marked with black.

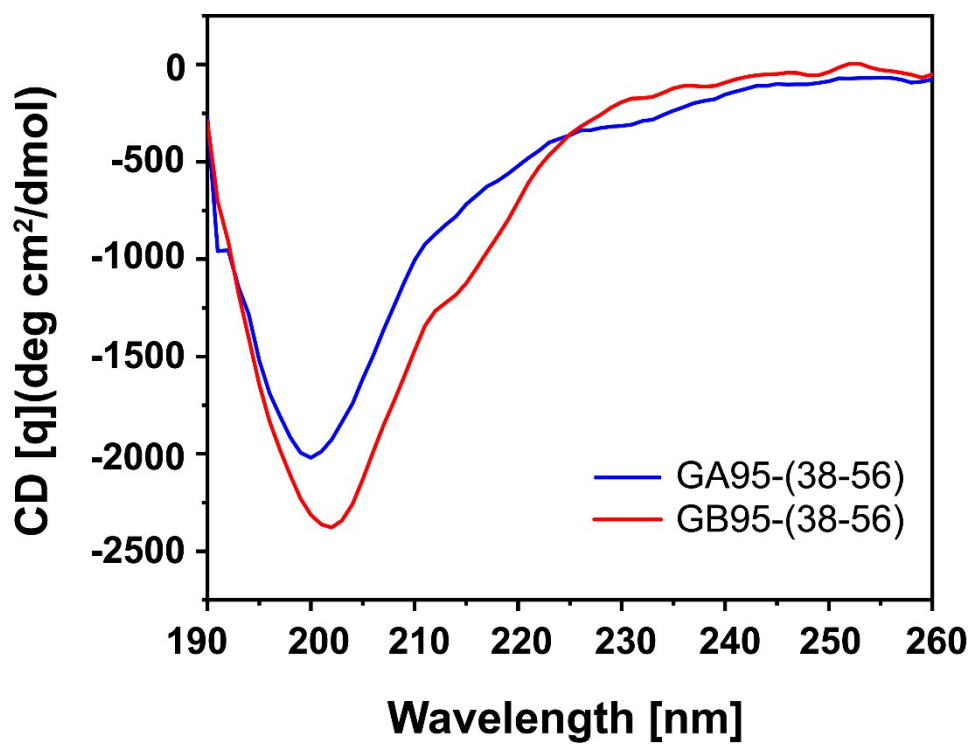


Fig. S7. The circular dichroic spectra of the C-terminal sequence (38-56) of GA95 and GB95. GA95-(38-56) is marked with blue, and GB95-(38-56) is marked with red.

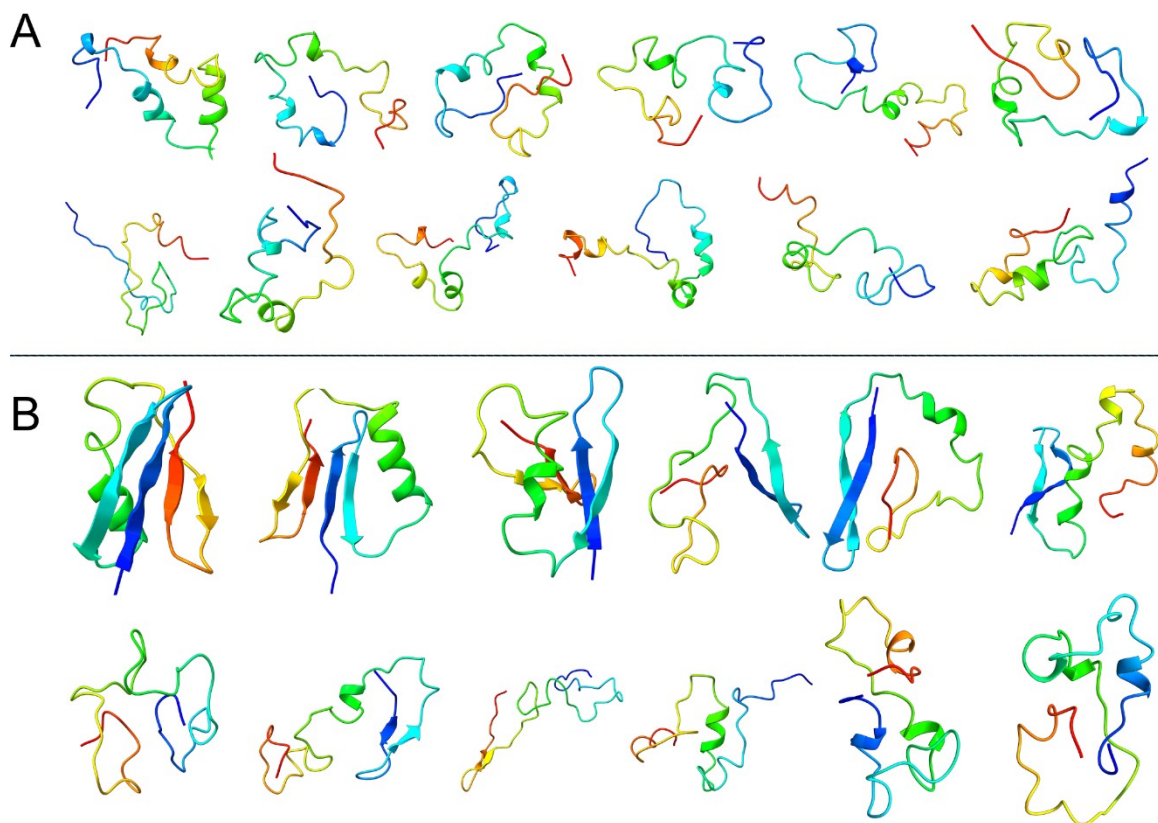


Fig. S8. The initial structures of REST2 simulations. (A) The initial GA95-53 conformations of 12 replicas of REST2 simulations. (B) The initial GB95 conformations of 12 replicas of REST2 simulations. The conformations were randomly selected from the 100-ns high-temperature simulation trajectories (under 500 K).

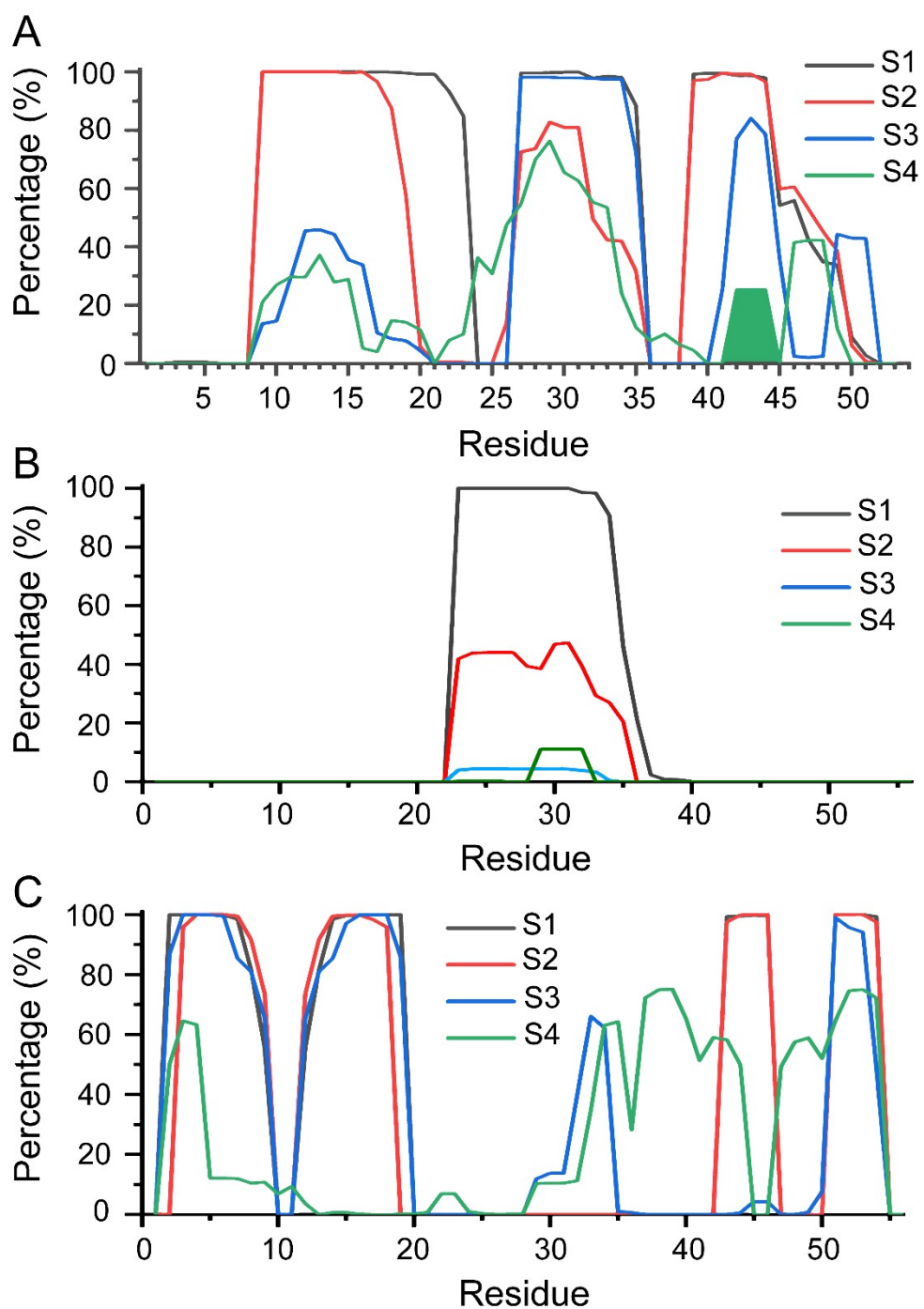
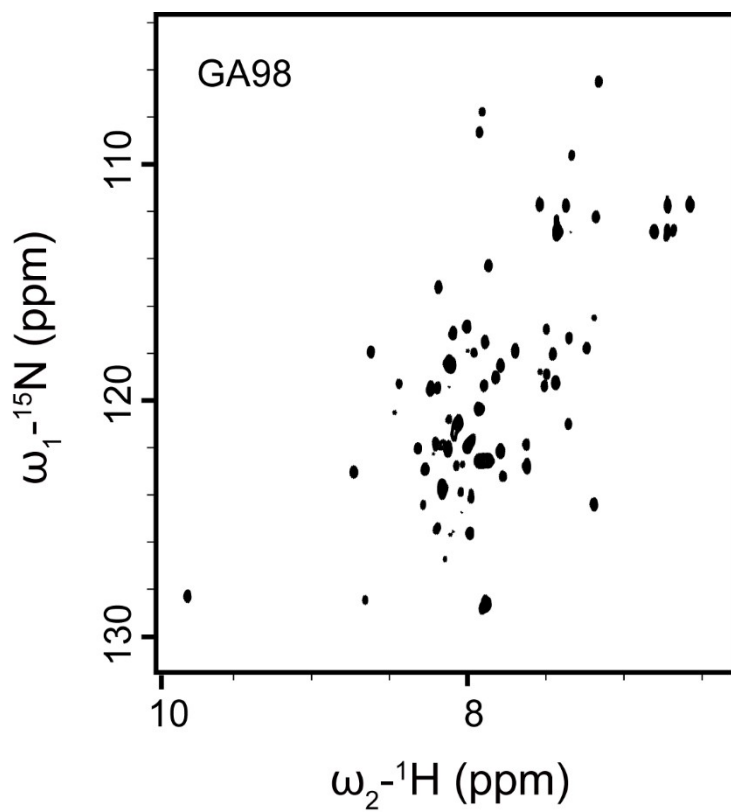


Fig. S9. The propensity of the residual secondary structure in different free energy states. (A) The propensity of the residual helical secondary structure of GA95-53; (B) The propensity of the residual helical secondary structure of GB95; (C) The propensity of the residual strand secondary structure of GB95.



GA98 TTYKLILNLKQAKEEAIKELVDAGTAEKYFKLIANAKTVEGVWTLKDEIKTFVTE
 GB98 TTYKLILNLKQAKEEAIKELVDAGTAEKYFKLIANAKTVEGVWTYKDEIKTFVTE

Fig. S10. The ^1H - ^{15}N HSQC Spectrum of GA98. The sequences of GA98 and GB98 are shown in the figure, with mutation site 45 colored in red.

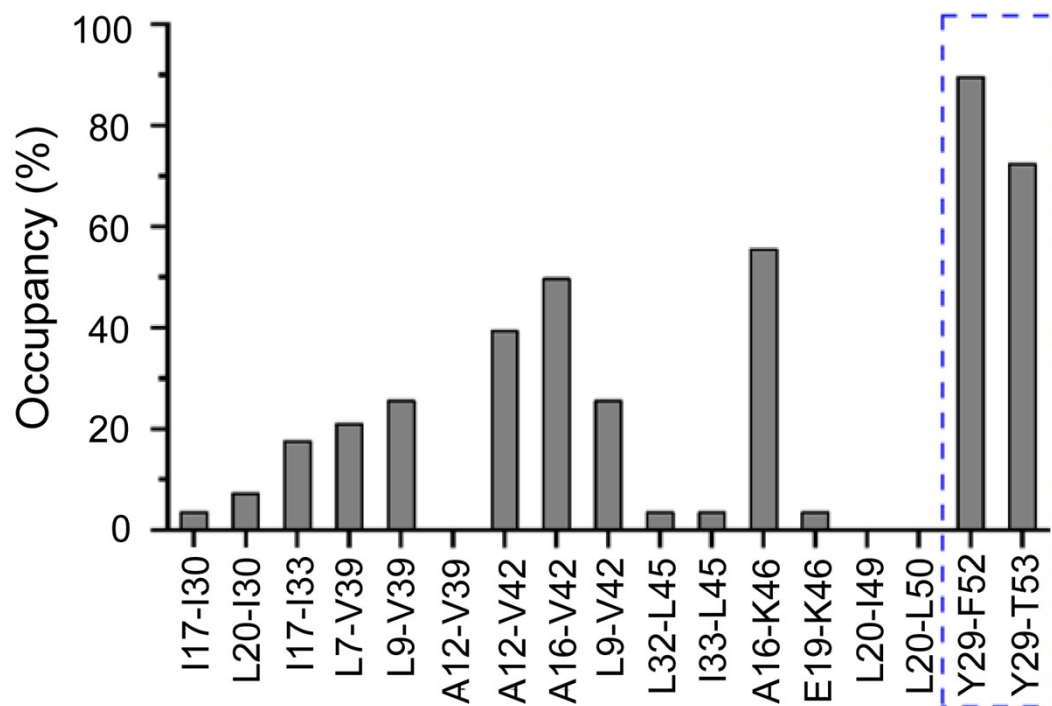


Fig. S11. The occupancy of important interactions in state T of the folding process of GA95-53. Y29 interacted with F52 and T53, respectively.

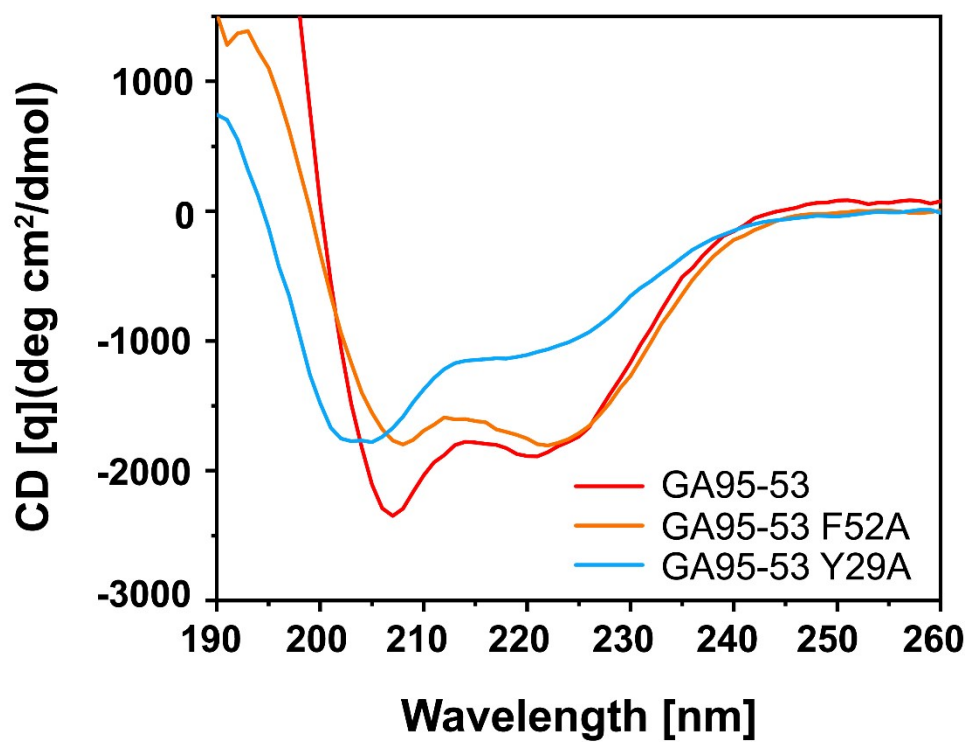


Fig. S12. The circular dichroic spectra of GA95-53 (red), GA95-53 F52A (orange), and GA95-53 Y29A (blue).

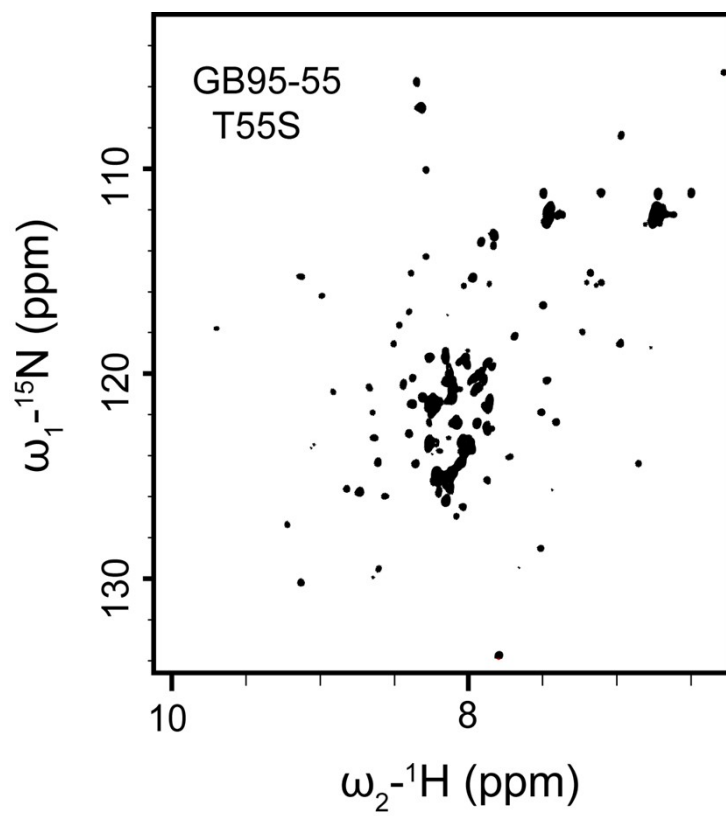


Fig. S13. The ^1H - ^{15}N HSQC spectrum of GB95-55 T55S.

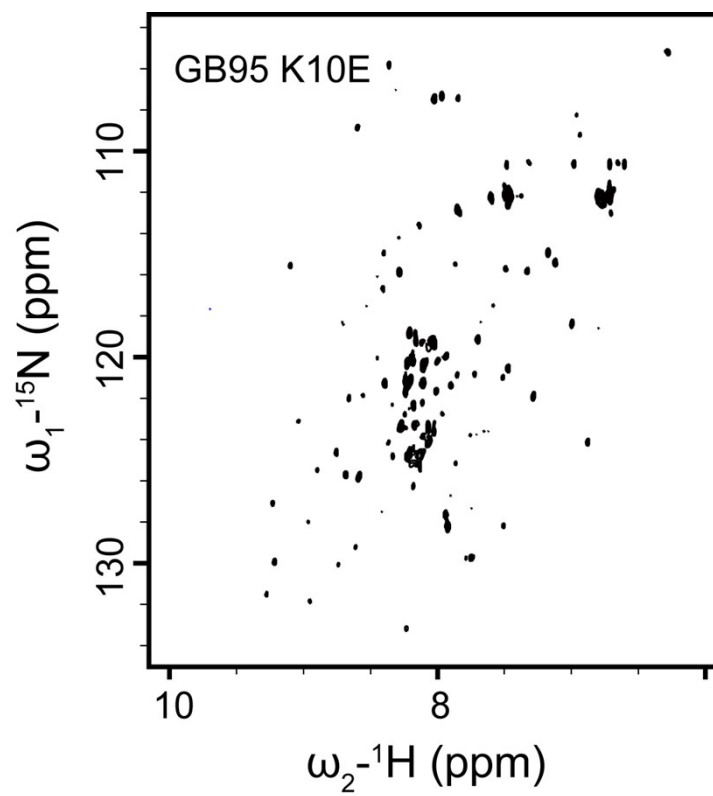


Fig. S14. The ^1H - ^{15}N HSQC spectrum of GB95 K10E.

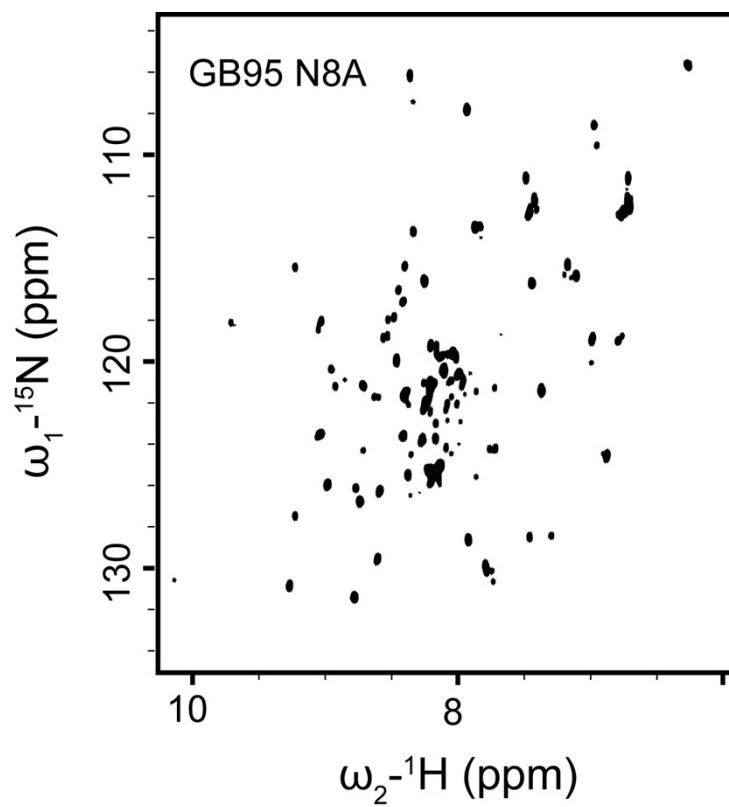


Fig. S15. The $^1\text{H}\text{-}^{15}\text{N}$ HSQC spectrum of GB95 N8A.

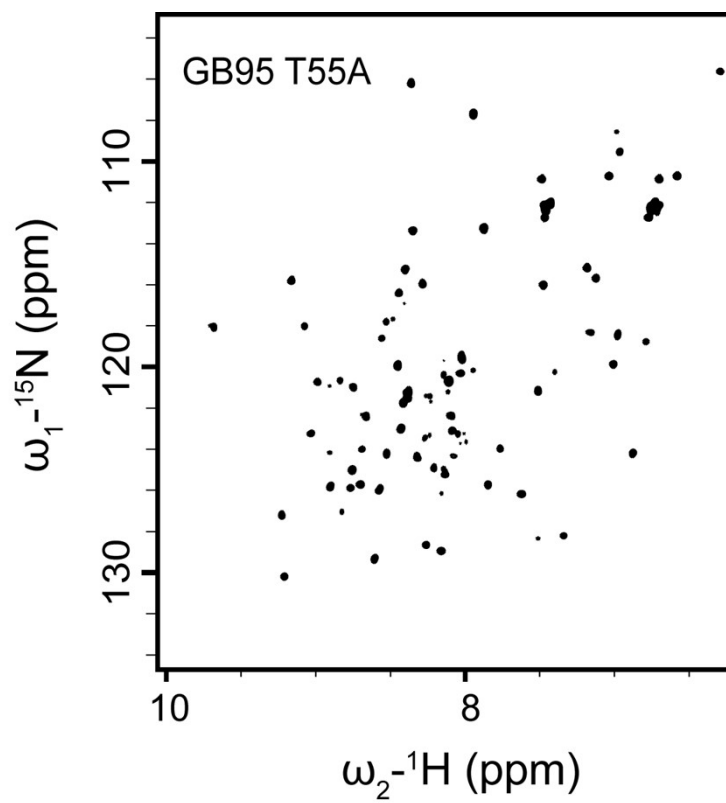


Fig. S16. The ^1H - ^{15}N HSQC spectrum of GB95 T55A.

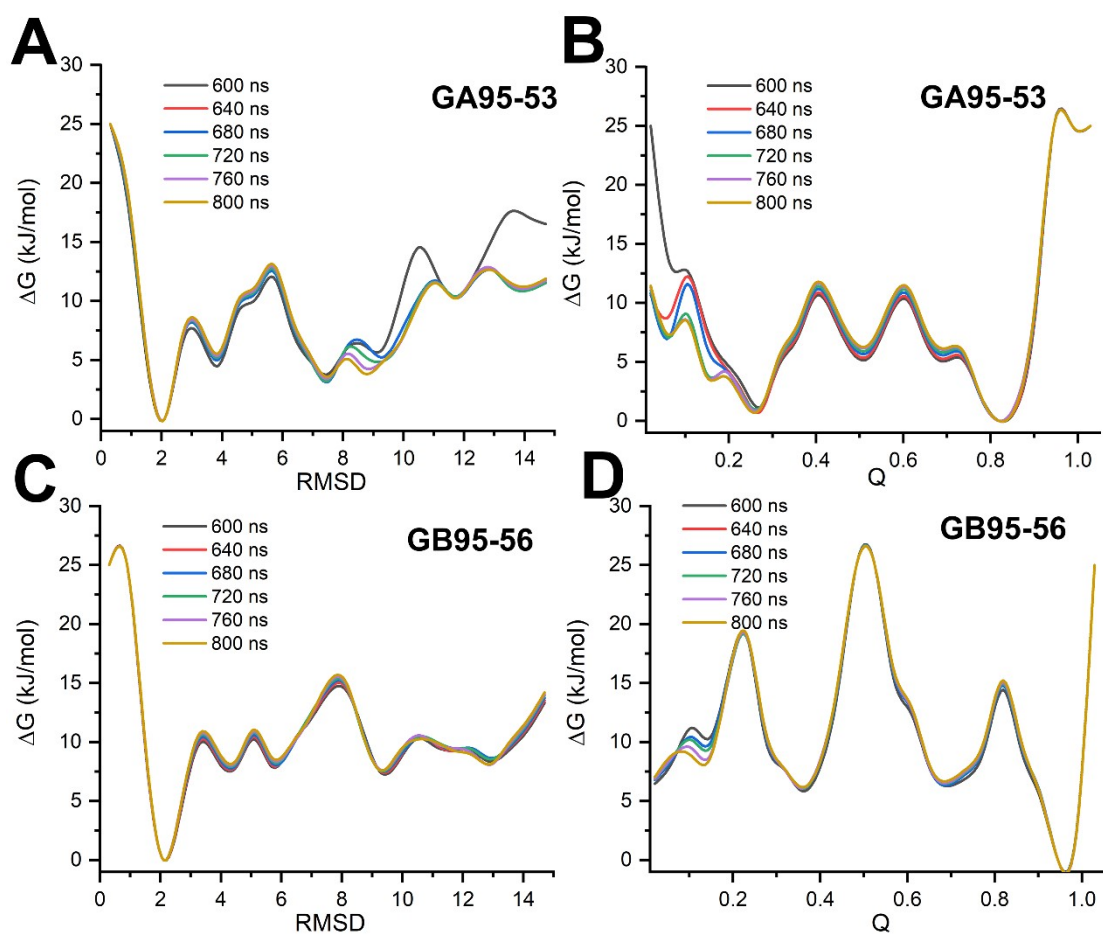


Fig. S17. The convergency analysis of REST2 simulations. (A) The free energy profiles as a function of root-mean-squared deviation (RMSD) of GA95-53. (B) The free energy profiles as a function of fraction of native contacts (Q) of GA95-53. (C) The free energy profiles as a function of root-mean-squared deviation (RMSD) of GB95-56. (D) The free energy profiles as a function of fraction of native contacts (Q) of GB95-56. The free energy profiles increased by 40ns in the last 200ns of simulations were calculated.

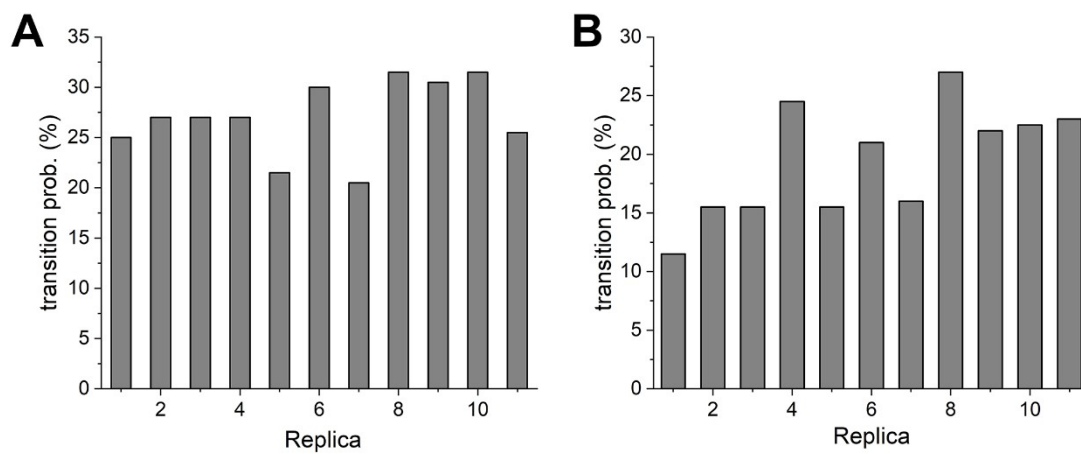


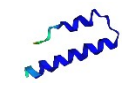

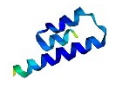
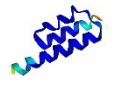



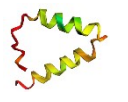
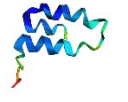
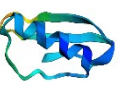


Fig. S18. The transporting probability between the replicas in the REST2 simulations. (A) GA95-53; (B) GB95-56.

Table S1. Three-dimensional models simulated by AlphaFold2 (AF2)

AF2	38	39	40	45	46	53
GA95-n						
GB95-n						






^apLDDT:  Very low (<50)  Low (60)  OK (70)  Confident (80)  Very high (>90)

Table S2. The results of proteins in AlphaFold2 (AF2) predictions or experiments

Protein	Experimental Result	AF2 Result	Concordance
GA95-38	Unfolded	α	No
GA95-45	Unfolded	2α	No
GA95-51	Unfolded	3α	No
GA95-52	Partially unfolded	3α	No
GA95-53	3α	3α	Yes
GA95-53 F52A	Unfolded	3α	No
GA95-53 Y29A	Unfolded	3α	No
GA95 L45I	3α	3α	Yes
GA95 I30F(GA98)	3α	3α	Yes
GB95-38	Unfolded	α	No
GB95-45	Unfolded	2α	No
GB95-54	Unfolded	$4\beta+\alpha$	No
GB95-55	Partially unfolded	$4\beta+\alpha$	No
GB95-55 Y29A	Partially unfolded	$4\beta+\alpha$	No
GB95-55 N8A	Unfolded	$4\beta+\alpha$	No
GB95-55 T55A	Unfolded	$4\beta+\alpha$	No
GB95-55 T55S	Partially unfolded	$4\beta+\alpha$	No
GB95 K10A	$4\beta+\alpha$	$4\beta+\alpha$	Yes
GB95 K10E	$4\beta+\alpha$	$4\beta+\alpha$	Yes
GB95 E56A	Partially unfolded	$4\beta+\alpha$	No
GB95 T55A	$4\beta+\alpha$	$4\beta+\alpha$	Yes
GB95 N8A	$4\beta+\alpha$	$4\beta+\alpha$	Yes
GB95 Y45I	Unfolded	$4\beta+\alpha$	No
GB95 Y45F	$4\beta+\alpha$	$4\beta+\alpha$	Yes
GB95 L32P	Partially unfolded	$4\beta+\alpha$	No
GB95 Y45L	Unfolded	$4\beta+\alpha$	No
GB95 F30I	Unfolded	$4\beta+\alpha$	No

^a The experimental results of GB95 Y45L and GB95 F30I were reported in the literature⁹.

3. References

- [1] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins*. 2006, **65**, 712-725.
- [2] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, *J. Chem. Phys.* 1983, **79**, 926-935.
- [3] S. Lejon, I.-M. Frick, L. Bjo"rck, M. Wikstro"m, S. Svensson, *J Biol Chem*. 2004, **279**, 42924-42928.
- [4] J. P. Derrick, D. B. Wigley, *Nature*. 1992, **359**, 752-754.
- [5] R. L. Davidchack, R. Hande, M. V. Tretyakov, *J Chem Phys*. 2009, **130**, 234101.
- [6] T. Darden, D. York, L. Pedersen, *J. Chem. Phys.* 1993, **98**, 10089-10092.
- [7] J.-P. Ryckaert, G. Ciccotti, H. J. C. Berendsen, *J Comput Phys*. 1977, **23**, 327-341.
- [8] L. Wang, R. A. Friesner, B. J. Berne, *J Phys Chem B*. 2011, **115**, 9431-9438.
- [9] P. A. Alexander, Y. H. Y. Chen, J. Orban, P. N. Bryan, *Proc Natl Acad Sci U S A*. 2009, **106**, 21149-21154.