Supplementary Information

Structural and Dynamic Insights into Redundant Function of YTHDF Proteins

Yaozong Li,[1,2,+] Rajiv K. Bedi,[1,+] Elena V. Moroz-Omori, [1,*] and Amedeo Caflisch[1,*]

[1]Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

[2]Department of Chemistry, Umeå University, SE-901 87 Umeå, Sweden

[+]These authors contributed equally

*To whom correspondence should be addressed. Tel: +41 44 635 5521; Fax: +41 44 635 6862; Email: caflisch@bioc.uzh.ch

Correspondence may also be addressed to Elena V. Moroz-Omori. Tel: +41 44 63 55587; Email: omori.elena@gmail.com

Yaozong Li's ORCid: 0000-0002-5796-2644

Rajiv Kumar Bedi's ORCid: 0000-0002-8193-9006

Elena V. Moroz-Omori's ORCid: 0000-0001-6485-3481

Amedeo Caflisch's ORCid: 0000-0002-2317-6792

Materials and methods

Cloning, expression, and purification

The plasmid containing full length human DF3 protein was a gift from Dr. Chuan He (Addgene ID: 70088), from which the YTH domain of the protein (residues 392-571) was PCR amplified using primers

DF3-Forward:
ATATCCATGGGACATCATCATCATCATCATAGTAGTGGAAGAGAAAATTTGTATTTTCAAGGACATCCCGTGCTGGAA

DF3-Reverse: GCGCCTCGAGTTATTCTTGACGCTTTTCATAATGTGCAAAGTCATCAAAGATTGAGGTGG

The amplified PCR product was inserted into pETDuet-1 vector between NcoI and XhoI sites.

The recombinant protein was purified to homogeneity in two chromatographic steps. The protein was overexpressed for 16 hours at 20°C in *Escherichia coli* BL21 (DE3) cells upon induction with 0.4 mM IPTG. The cells were harvested and resuspended in the lysis buffer containing 100 mM Tris-HCl at pH 8.0, 500 mM NaCl, and 10 mM imidazole. The cells were lysed by sonication, and the cell lysate was clarified by centrifugation at 48'000 g for one hour and loaded onto Ni-NTA affinity column (5 mL HisTrap FF from GE Healthcare). After extensive washing with the wash buffer containing 100 mM Tris-HCl at pH 8.0, 500 mM NaCl, and 50 mM imidazole, the target protein was eluted with elution buffer containing 100 mM Tris-HCl at pH 8.0, 500 mM NaCl, and 250 mM imidazole. The N-terminal hexahistidine-tag was removed by cleavage with tobacco etch virus (TEV) protease at 1:50 ratio. The excess imidazole was removed by overnight dialysis, and the sample was subjected to secondary subtractive Ni-NTA affinity chromatography step to remove the protease and uncleaved protein. Finally, the protein was subjected to a gel filtration step using Superdex 75 16/60 column in a buffer containing 20 mM Tris-HCl at pH 8 and 150 mM NaCl. The protein was concentrated to 10 mg/mL, flash-frozen in liquid nitrogen, and stored at -80 °C for future experiments.

X-ray crystallography

The purified YTH domain and GG(m^6A)CU (purchased form Dharmacon) were mixed together at 0.5 mM concentration each. Crystals were obtained in 25% PEG 3350, 100 mM Tris pH 8, and 200 mM NaCl.

Diffraction data were collected at the Swiss Light Source (Villigen, Switzerland) using the beamline X06DA (PXIII) and processed using XDS.[1] The structures were solved by molecular replacement using Phaser program[2] from the Phenix package.[3] The unliganded structure of YTHDF1 (PDB ID: 4RCI) was used as a search model. The model building and refinements were performed using COOT[4] and phenix.refine.[5] Data collection and refinement statistics are summarized below in Table S1.

PDB ID:	6ZOT
Data Collection	
Space group	P 4 21 2
Cell dimension a, b, c (Å)	100.75, 100.75, 72.88
Cell dimension α, β, γ (°)	90, 90, 90
Resolution (Å)	45.06 (2.7)
Unique reflections*	10844 (1696)
Completeness*	99.8 (98.2)
Redundancy*	16.9 (16.8)
R_{merge}*	58.4 (253.3)
CC (1/2)	99.7 (61.7)
I/σI	7.25 (1.4)

Refinement	
R_{work}/R_{free}	0.2153/0.2757
RMSD bond (Å)	0.011
RMSD angle (°)	1.444
B-factors P/L/W ($Å^2$)**	29.04/50.27/26.15
Ramanchandran Favored	91.36
Ramanchandran allowed	7.72
Ramanchandran Disallowed	0.92

*Statistics for the highest resolution shell is shown in parentheses.

** P/L/W indicate protein, ligand and ion, and water molecules, respectively.

Table S1: X-ray data collection and refinement statistics for the YTHDF3-RNA complex.

MD simulations

The CHARMM36 force field was used for the parameters of the protein.[6] The atomic coordinates of the YTH domain of DF1, DF2, and DF3 were extracted from the crystal structures 4RCJ,[7] 4RDO,[8] and 6ZOT, respectively. Chain A was kept in all structures and ribonucleotides (if any) were removed to build the apo systems. Crystal water molecules were kept as in the initial structure. For DF1, DF2, and DF3, the sequences 378~527, 409~547, and 405~554 were retained for representing the systems, respectively. The hydrogen atoms were added by the CHARMM program,[9] and the protonation states were determined at neutral pH conditions. The complex systems were solvated in a 70 Å of rhombic dodecahedron (RHDO) TIP3P water box to ensure at least 10 Å buffer space between the macromolecular atoms and the boundary of the water box. To neutralize the system and mimic the physiological conditions, Na^+ and Cl^- ions at 150 mM concentration were added to the solvated systems.

Each simulation system was initially minimized for 10'000 steps under a series of restraints and constraints on the protein and crystal water molecules to release their bad contacts and poor geometry. The minimized structure was heated to 300 K and equilibrated in NVT condition (constant volume and temperature). The system was further heated up to 600 K to enhance the sampling of ions and was cooled down back to 300 K after 100 ps. In this period, the heavy atoms of the protein were restrained with harmonic potential. Finally, the structure was further equilibrated in NPT condition (constant pressure and temperature) without any structural restraint. All the equilibration phases lasted for 5 nanosecond (ns) using the CHARMM program (version 42b2). Five production runs of 1 microseconds (µs) each and for each YTH domain were carried out in NPT conditions using the NAMD program (version 2.12).[10] The pressure was controlled by Nosé-Hoover Langevin piston method with 200 picosecond (ps) piston period and 100 ps piston decay time.[11] The temperature was maintained at 300 K using the Langevin thermostat with a 5 ps friction coefficient. The integration time step was set to 2 fs by constraining all the bonds involving hydrogen atoms by the SHAKE algorithm. Van der Waals energies were calculated using a switching function with a switching distance from 10 to 12 Å, and electrostatic interactions were evaluated using the particle mesh Ewald summation (PME) method.[12]

Five independent production runs with random initial velocities were carried out for a total of 5 µs for each of the three YTH domains. MD snapshots were saved every 20 ps along the MD trajectories for further analysis. RMSF and covariance matrix analysis, were performed with CHARMM routines. All statistical figures were plotted by MATLAB (Version 2018a MathWorks, Inc.) and structural figures were generated with the PyMOL graphic software (Version 2.2 Schrödinger, LLC.).

Analysis of MD trajectories

RMSF. The average structure along 5 μs MD trajectory was used as the reference for calculating RMSF values. The average structure was obtained by superposing all MD snapshots to backbone atoms of the crystal structure while excluding N- and C- terminals. Heavy atoms of backbone are involved in the RMSF calculation, and the RMSF values were mapped to the average structure (Figure 2 top).

Positional covariance analysis. The same average structure as that for the RMSF calculation was used for constructing the covariance matrix, which is used to reveal segments of the domain that move together. Each element of the symmetric matrix is the positional (*i.e.*, 3D spatial) correlation value between residues *i* and *j*, which is calculated by the equation $C_{ij} = \frac{\overline{(r_i - \bar{r}_i)(r_j - \bar{r}_j)}}{\sqrt{\overline{(r_i^2 - \bar{r}_i^2)(r_j^2 - \bar{r}_j^2)}}}$, where r_i is the vector of the coordinates of atom *i* and the overbar is the temporal average.[13]

The residue-averaged correlation value (shown in Figure 2 bottom) was calculated along each column *i* of the positional covariance matrix by the sum $\frac{1}{N_R}\sum_{j=1}^{N_R}|C_{ji}|$, where $N_R$is the number of residues.

Principal component analysis. The principal component analysis (PCA) was performed using a 5-μs trajectory for each of the three system, e.g., DF1, DF2, and DF3. First, the analysis was carried out by superimposing all the snapshots to the corresponding crystal structure and determining the average structure for each protein. Moreover, the covariance matrix of atomic fluctuations was determined against the average structure. Finally, the covariance matrix was diagonalized to determine eigenvectors and eigenvalues which correspond to the protein's vibrational modes (i.e., PCA modes) and frequencies, respectively. The three lowest-frequency motions are presented in Figure S6. In all superposition and construction of covariance matrix, we considered only the backbone heavy atoms of the rigid part of the proteins by ignoring flexible N- and C-terminals of the proteins.

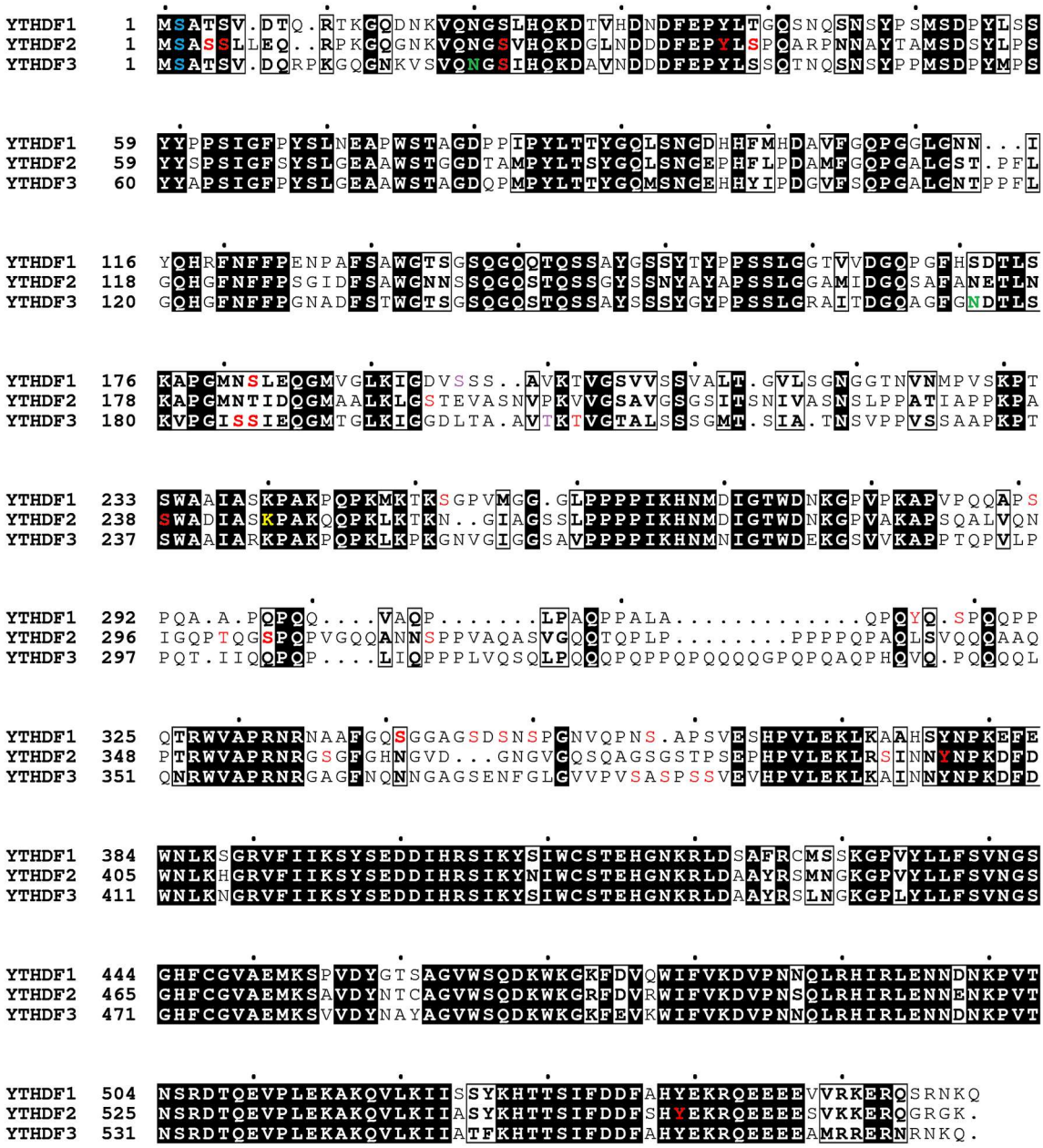

Figure S1. Figure legend: Sequence alignment of YTHDF1, YTHDF2 and YTHDF3 showing post translation modification sites. Post translation modification sites are color coded are as follow: Red = phosphorylation; Green = N-linked glycosylation; Purple = O-linked glycosylation; Blue = N-terminal acetylation; Yellow = N6-acetyllysine. Sequence alignment was performed by Clustal Omega,[14] figure was generated using ESPript 3.0,[15] and the post translation modification sites were retrieved from neXtProt.[16]

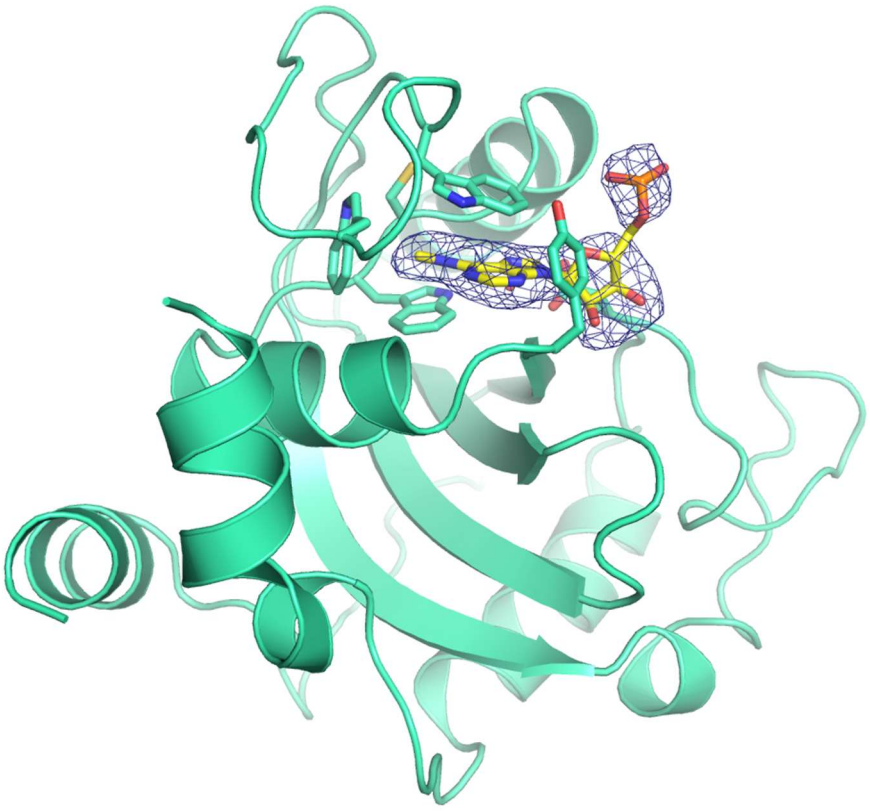

Figure S2. The weighted $2F_o$-F_c electron density map of m^6A (shown at 2 sigma level) in complex with YTH domain of DF3.

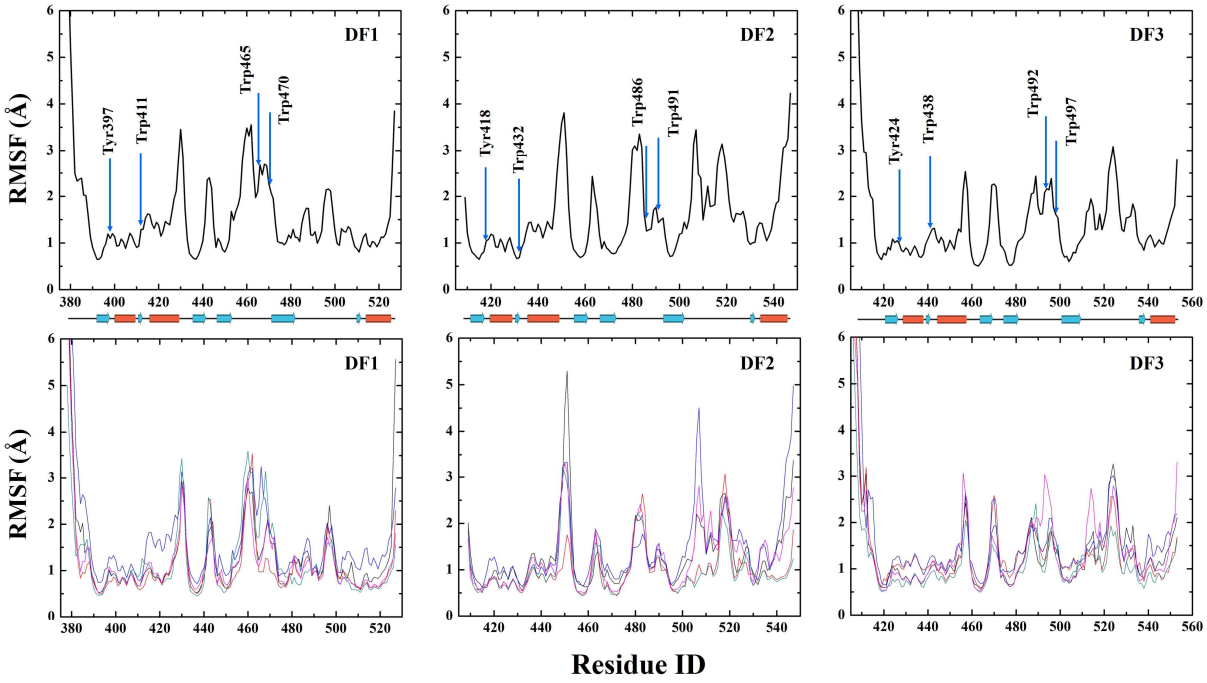

Figure S3. Root mean square fluctuation (RMSF) profiles along the sequence of the YTH domain of the three human DF proteins. Top panel: for each domain, the RMSF values were calculated using the average structure over the 5 μs as reference. For each residue, the non-hydrogen atoms of the backbone were used. The aromatic residues in the m^{6}A recognition pocket are highlighted (vertical arrows) and secondary structure elements are shown below the sequence (α-helices, red cylinders; β-strands, cyan arrows). Bottom panel: the RMSF values were calculated using the average structure over 1 μs for each trajectory as a reference. Five RMSF plots were generated for each system and shown in different colors. Notes: as all the RMSF plots were used different average structures as their references, an RMSF plot from the top panel is not a simple average of the corresponding five RMSF plots in the bottom panel.

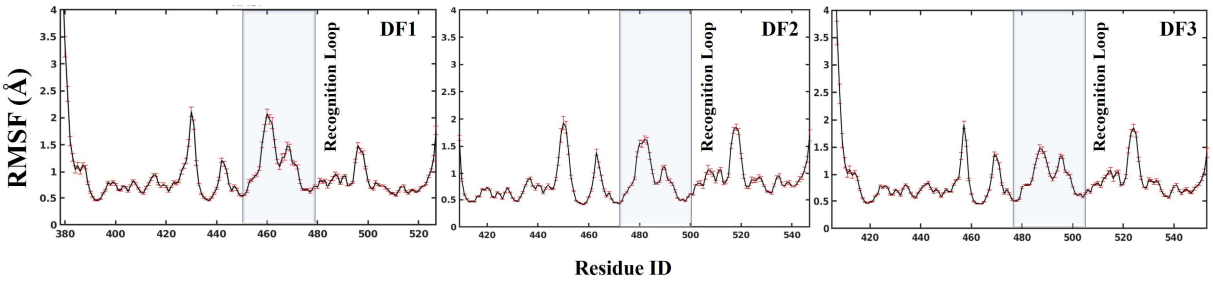

Figure S4. Root mean square fluctuation (RMSF) with standard error bars. One 5-μs trajectory for each system was truncated into 50 sub-trajectories (100 ns for each) and a corresponding RMSF plot was calculated for each of the 50 sub-trajectories. The final RMSF plot was obtained by doing the arithmetic mean of 50 sub-RMSF plots for each system and the associated standard error was calculated to show its uncertainty. Other details were the same as those in Figure S2. The averaged RMSF values are mostly smaller than those based on the entire 5-μs trajectory because of the averaging over relatively short trajectories. The region in gray indicates the recognition loop.

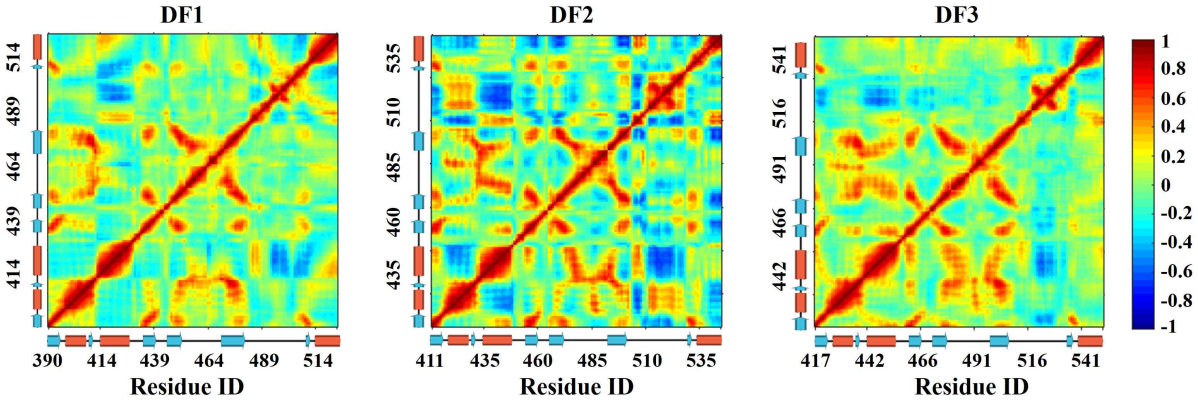

Figure S5. Covariance matrixes of positional displacement of the three DF proteins. For each residue, the non-hydrogen atoms of the backbone were used as for the RMSF calculations. The procedure for calculating the covariance matrix is described below in the subsection “Analysis of MD trajectories” of Materials and Methods. The numerical value for each pair of residues ranges from -1 (fully anticorrelated, blue), through 0 (uncorrelated, green), to 1 (correlated, red), and the matrix is symmetric.

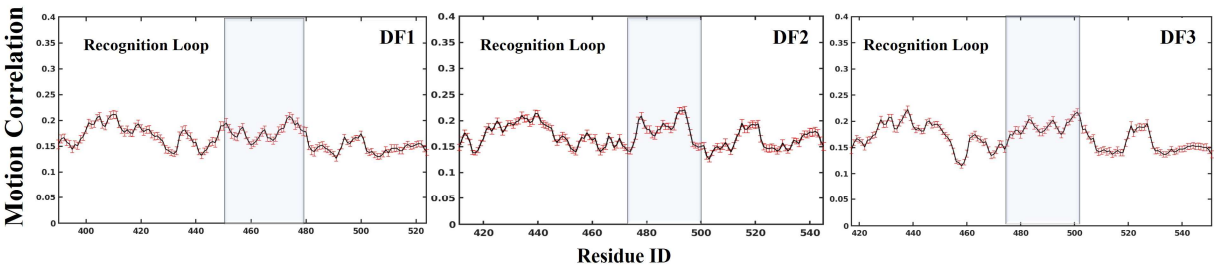

Figure S6. Positional covariance correlation with standard error bars. The basic procedure is the same as that in Figure 2. To estimate the uncertainty of the motion correlation, the 5-μs trajectory for each system was divided into 50 sub-trajectories. With each sub-trajectory, the residue-based covariance correlation was calculated. Finally, an average correlation plot was generated by doing the arithmetic mean of 50 sub-correlation plots and the associated standard error was calculated. The averaged correlation values are mostly smaller than those from the entire 5-μs trajectory because of the averaging over relatively short trajectories. The region in gray indicates the recognition loop.

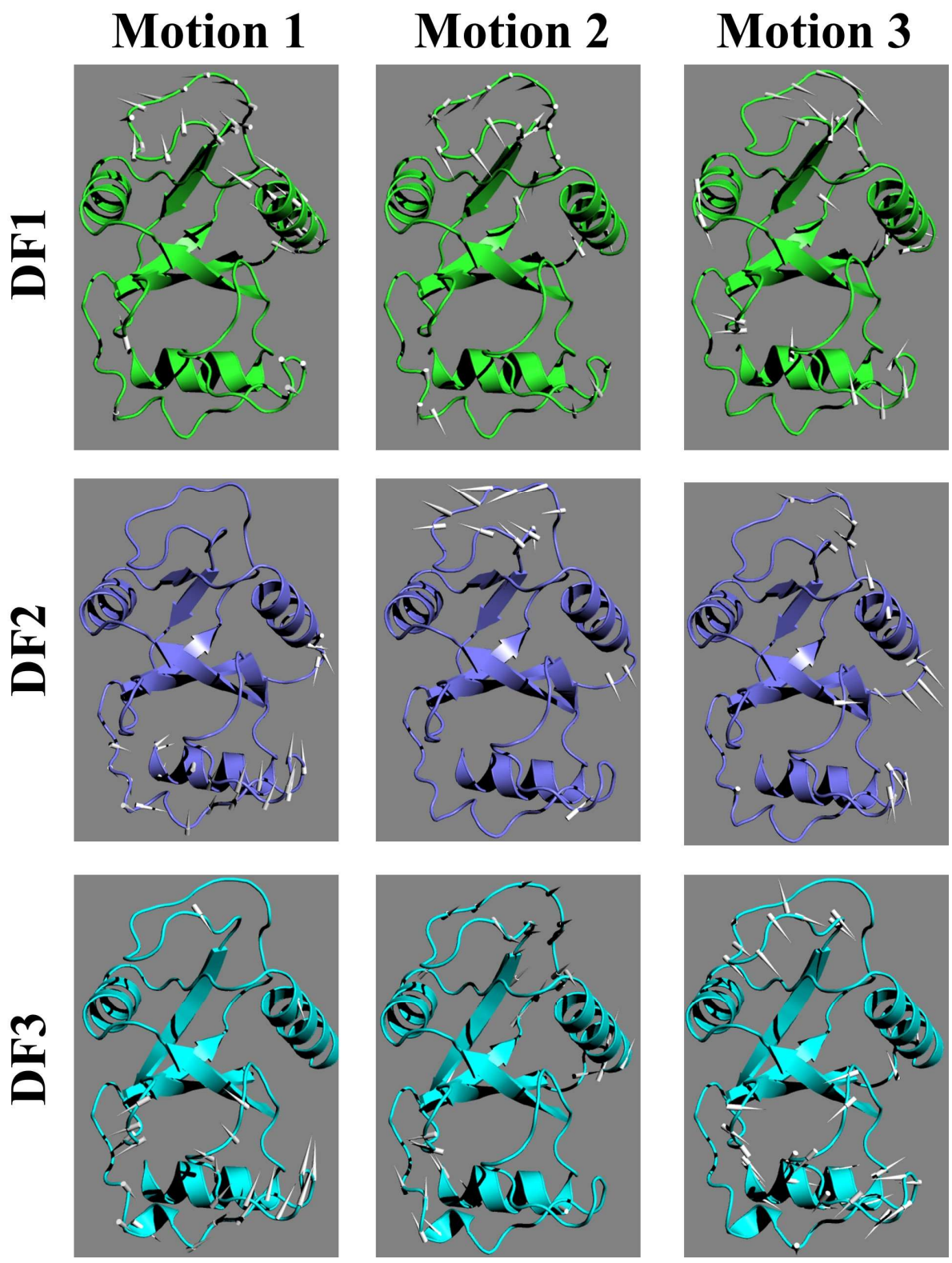

Figure S7. Principal component analysis (PCA). For each protein, the motion vectors 1 to 3 are arranged in ascending order by their mode frequencies. For each PCA mode, the average structure was kept fixed and the motion vector (arrows) was determined by the displacement between the average structure and the one utmost to average positions. Because the motion is quasi-harmonic, the directions indicated by the arrows are bi-directional, along which the proteins thermally vibrated around the average structure.

References

1. Kabsch, W. Integration, Scaling, Space-Group Assignment and Post-Refinement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 133–144 (2010).
2. McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. Phaser Crystallographic Software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
3. Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V. *et al.* Macromolecular Structure Determination Using X-rays, Neutrons and Electrons: Recent Developments in Phenix. *Acta Crystallogr. Sect. D Struct. Biol.* **75**, 861–877 (2019).
4. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and Development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).
5. Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. Towards Automated Crystallographic Structure Refinement with Phenix.refine. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 352–367 (2012).
6. Huang, J. & Mackerell, A. D. CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data. *J. Comput. Chem.* **34**, 2135–2145 (2013).
7. Xu, C., Liu, K., Ahmed, H., Loppnau, P., Schapira, M. & Min, J. Structural Basis for the Discriminative Recognition of N6-Methyladenosine RNA by the Human YT521-B Homology Domain Family of Proteins. *J. Biol. Chem.* **290**, 24902–24913 (2015).
8. Li, F., Zhao, D., Wu, J. & Shi, Y. Structure of the YTH Domain of Human YTHDF2 in Complex with an m6A Mononucleotide Reveals an Aromatic Cage for m6A Recognition. *Cell Res.* **24**, 1490–1492 (2014).
9. Brooks, B. R., Brooks, C. ., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W. *et al.* CHARMM: Molecular Dynamics Simulation Package. *J. Comput. Chem.* **30**, 1545–1614 (2009).
10. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L. & Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
11. Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **103**, 4613–4621 (1995).
12. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
13. Seeber, M., Felline, A., Raimondi, F., Muff, S., Friedman, R., Rao, F., Caflisch, A. & Fanelli, F. Wordom: A User-Friendly Program for the Analysis of Molecular Structures, Trajectories, and Free Energy Surfaces. *J. Comput. Chem.* **32**, 1183–1194 (2011).
14. Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D. & Lopez, R. The EMBL-EBI Search and Sequence Analysis Tools APIs in 2019. *Nucleic Acids Res.* **47**, W636–W641 (2019).
15. Robert, X. & Gouet, P. Deciphering Key Features in Protein Structures with the New ENDscript Server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
16. Zahn-Zabal, M., Michel, P.-A., Gateau, A., Nikitin, F., Schaeffer, M., Audot, E., Gaudet, P., Duek, P. D., Teixeira, D., Rech de Laval, V., Samarasinghe, K., Bairoch, A. & Lane, L. The neXtProt Knowledgebase in 2020: Data, Tools and Usability Improvements. *Nucleic Acids Res.* **48**, D328–D334 (2019).