Supporting Information

Yang et al. 10.1073/pnas.1000848107

SI Text

Materials and Methods. ***Crystallization and structure determination.*** Hanging drops were set up by mixing 1 μl of the CFI_m25-RNA complex with 1 μl of crystallization solution consisting of 100 mM sodium acetate pH 5.0, 200 mM NaH_2PO_4 pH 4.9 and 25% (w/v) PEG 8000. Crystals appeared after 7 days at 24 °C and grew to full size (150 × 150 × 50 μm) in approximately 3 weeks. Crystals were cryoprotected by equilibrating a hanging drop against a reservoir containing 2 M NaCl for 6 hours prior to flash cooling the crystals in liquid nitrogen. Complete X-ray diffraction datasets were recorded at 100 K at $\lambda = 1.5418$ Å on our laboratory MAR345 detector (MAR Research) mounted on a Rigaku RU-200 rotating anode generator equipped with Xenocs focusing mirrors. Data were processed using the XDS package (1) and scaled with XSCALE. Both CFI_m25-RNA complexes crystallized in the same space group, $P2_12_12_1$, and share very similar cell parameters ($a = 71.33$ $b = 72.08$, $c = 96.73$ Å *vs.* $a = 71.88$, $b = 73.88$, $c = 96.73$ Å). Two monomers are present in the asymmetric unit, which corresponds to a calculated solvent content of 55%.

The crystal structures were solved by molecular replacement using the program MOLREP (2) and the previously solved structure of CFI_m25 [PDB ID code 3BHO (3)] stripped of all nonprotein atoms as a search model. A solution was found only after the extended N-terminal segment (residues 20–33) was deleted from the model. Refinement consisted of iterative cycles of positional and B-factor refinement and was performed with PHENIX (4). Manual rebuilding was accomplished using COOT (5). The overall quality of the final model was evaluated using both PROCHECK (6) and MOLPROBITY (7). The asymmetric unit comprises two dimers, which present a dimer interface identical to that formed by the two monomers related by a 2 fold-crystallographic axis in the apo structures [3BAP (3), 2CL3 (8)]. A similar dimeric form in the asymmetric unit has been reported for the SO_4^{2-}-bound CFI_m25 crystal [2J8Q (8)]. After refinement, the extended N-terminal segment is only visible in one of the monomers (designated as molecule B), but not the other (molecule A). The model comprises residues 29–227 and 21–227 for molecule A and B, respectively, in the UGUAAA-bound model and residues 28–227 and 18–227 for molecule A and molecule B, respectively, in the UUGUAU-bound model. There are 2 glycerol molecules in the UGUAAA-bound complex and 3 in the UUGUAU model that originated from the protein storage buffer, which contains 10% (v/v) glycerol. All molecular structure figures were generated with PyMOL (9) and the electrostatic potential was calculated with Delphi (10).

Gel electrophoretic mobility shift assays. The ΔN21 deletion construct and the ΔN21 Glu55Ala, Arg63Ser, Glu81Ala, Phe103Ala, and Phe103Trp substitution variants were made using a QuikChange II XL mutagenesis kit (Stratagene) and confirmed by DNA sequencing. All protein variants were expressed and purified using the same protocol as for the full-length CFI_m25 (3).

α^{32}P-GTP-labeled RNAs containing the human PAPOLA upstream sequences (−56 to −39 relative to the poly(A) cleavage site (11) 5′-GGGUGUAAACAGAUGAUGUAU-3′), or sequence variants were prepared by T7 RNA polymerase transcription of synthetic oligonucleotides (three guanine residues were added to the 5′ end to ensure efficient transcription and labeling). In addition to the two 6 nt RNA used for crystallization, a 9 nt RNA (5′-UAUUUUGUA-3′) and a 21 nt RNA (5′-GGGUGUAAACAGAUGAUGUAU-3′) were purchased from Dharmacon (Lafayette, CO) and 5′ labeled with γ^{32}P-ATP-labeled using T4 kinase. A construct missing the first 21 residues, CFI_m25 ΔN21, was used for electrophoretic mobility shift assays (EMSA) rather than the full-length protein because RNA-protein complexes formed with full-length CFI_m25 were consistently retained in the well. The first 21 residues of CFI_m25 are not conserved among species (Fig. S1) and were not seen in the crystal structure, presumably because of disorder (3, 8). Moreover, the truncated CFI_m25ΔN21 protein retains the ability to bind RNA alone ($K_d = 645$ nM, Fig. S6) as well as in the context of the CFI_m 68 kDa subunit. RNA binding reactions contained 15 mM Tris-HCl pH 8.0, 40 mM potassium chloride, 5 mM DTT and 7.5% (v/v) glycerol, 0.75% polyvinyl alcohol (PVA) and 0.5 μg of *Escherichia coli* tRNA in a final volume of 20 μL. The final concentration of protein was 1 μM and RNA was 2 nM. The RNA binding reactions were incubated at 30 °C for 5 min and the protein-RNA complexes were resolved by electrophoresis on a nondenaturing 5% (80:1) polyacrylamide gel at 4 °C. RNA quantitation was done using a Personal Molecular Imager FX (Bio-Rad). We recorded the counts present in the RNA:protein complex (B) and free RNA (F) after subtracting the background. The RNA binding affinity of CFI_m25 was designated as the percentage of bound RNA (P), where $P = B/(B + F)$. The binding affinity for each protein variant and RNA mutation was plotted relative to CFI_m25ΔN21 and the wild type PAPOLA RNA.

1. Kabsch W (1993) Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. *J Appl Crystallogr* 26:795–800.
2. Vagin A & Teplyakov A (1997) MOLREP: An automated program for molecular replacement. *J Appl Crystallogr* 30:1022–1025.
3. Coseno M et al. (2008) Crystal structure of the 25 kDa subunit of human cleavage factor Im. *Nucleic Acids Res* 36(10):3474–3483.
4. Adams PD et al. (2002) PHENIX: Building new software for automated crystallographic structure determination. *Acta Crystallogr D* 58(Pt 11):1948–1954.
5. Emsley P, Lohkamp B, Scott W, & Cowtan K (2010) Features and Development of Coot. *Acta Crystallogr D* In press.
6. Laskowski RA, MacArthur MW, Moss DS, & Thornton JM (1993) PROCHECK: A program to check the stereochemical quality of protein structures. *J Appl Crystallogr* 26:283–291.
7. Davis IW et al. (2007) MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35(Web Server issue):W375–383.
8. Tresaugues L et al. (2008) The crystal structure of human cleavage and polyadenylation specific factor-5 reveals a dimeric Nudix protein with a conserved catalytic site. *Proteins* 73(4):1047–1052.
9. DeLano WL (2008) The PyMOL Molecular Graphics System. http://www.pymol.org (San Carlos, CA).
10. Nicholls A & Honig B (1991) A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *Journal of Comput Chem* 12(4):435–445.
11. Venkataraman K, Brown KM, & Gilmartin GM (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* 19(11):1315–1327.
12. Sickmier EA et al. (2006) Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Mol Cell* 23(1):49–59.

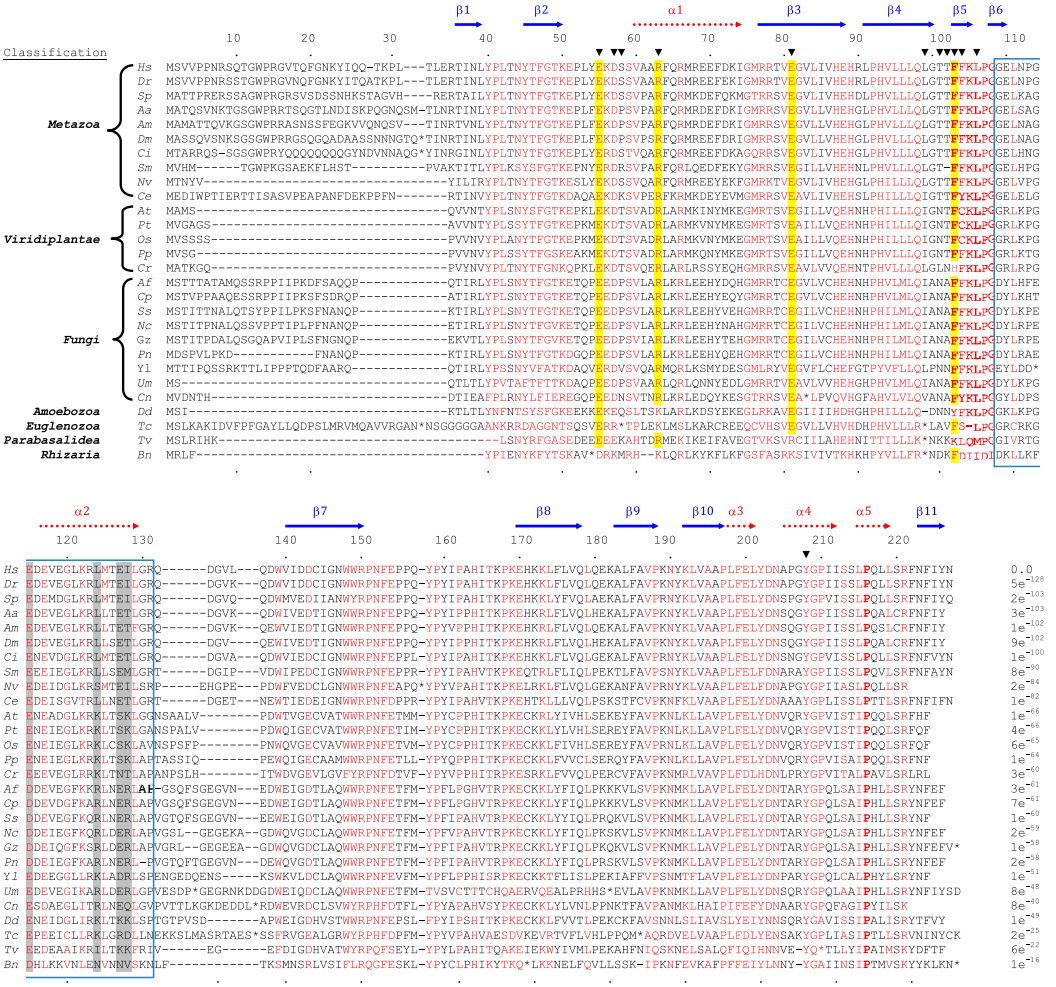

Fig. S1. Sequence alignment of CFI_m25 proteins. CFI_m25 protein sequences are listed according to phylogenetic classification and e values are shown on the right hand side. Secondary structure elements and amino acid position numbers for human CFI_m25 are depicted above the alignment. Residues of human CFI_m25 that contact the RNA in the crystal structures are denoted by inverted triangles above the alignment. Residues shown in red are at least 70% conserved and are bold if identical; residues in yellow were mutated as discussed in the text; gray shaded residues are the catalytic glutamate residues of the consensus Nudix box, which is delineated by a blue box. Gaps that were introduced for optimal alignment are denoted by dashes. Asterisks indicate the presence of two or more amino acids that were omitted from the alignment for clarity. Species abbreviations are as follows: *Hs, Homo sapiens; Dr, Danio rerio; Sp, Strongylocentrotus purpuratus; Aa, Aedes aegypti; Am, Apis mellifera; Dm, Drosophila melanogaster; Ci, Ciona intestinalis; Sm, Schistosoma mansoni; Nv, Nematostella vectensis; Ce, Caenorhabditis elegans; At, Arabidopsis thaliana; Pt, Populus trichocarpa; Os, Oryza sativa; Pp, Physcomitrella patens; Cr, Chlamydomonas reinhardtii; Af, Aspergillus flavus; Cp, Coccidioides posadasii ; Ss, Sclerotinia sclerotiorum; Nc, Neurospora crassa; Gz, Gibberella zeae; Pn, Phaeosphaeria nodorum; Yl, Yarrowia lipolytica; Um, Ustilago maydis; Cn, Cryptococcus neoformans; Dd, Dictyostelium discoideum; Tc, Trypanosoma cruzi; Tv, Trichomonas vaginalis; Bn, Bigelowiella natans*.

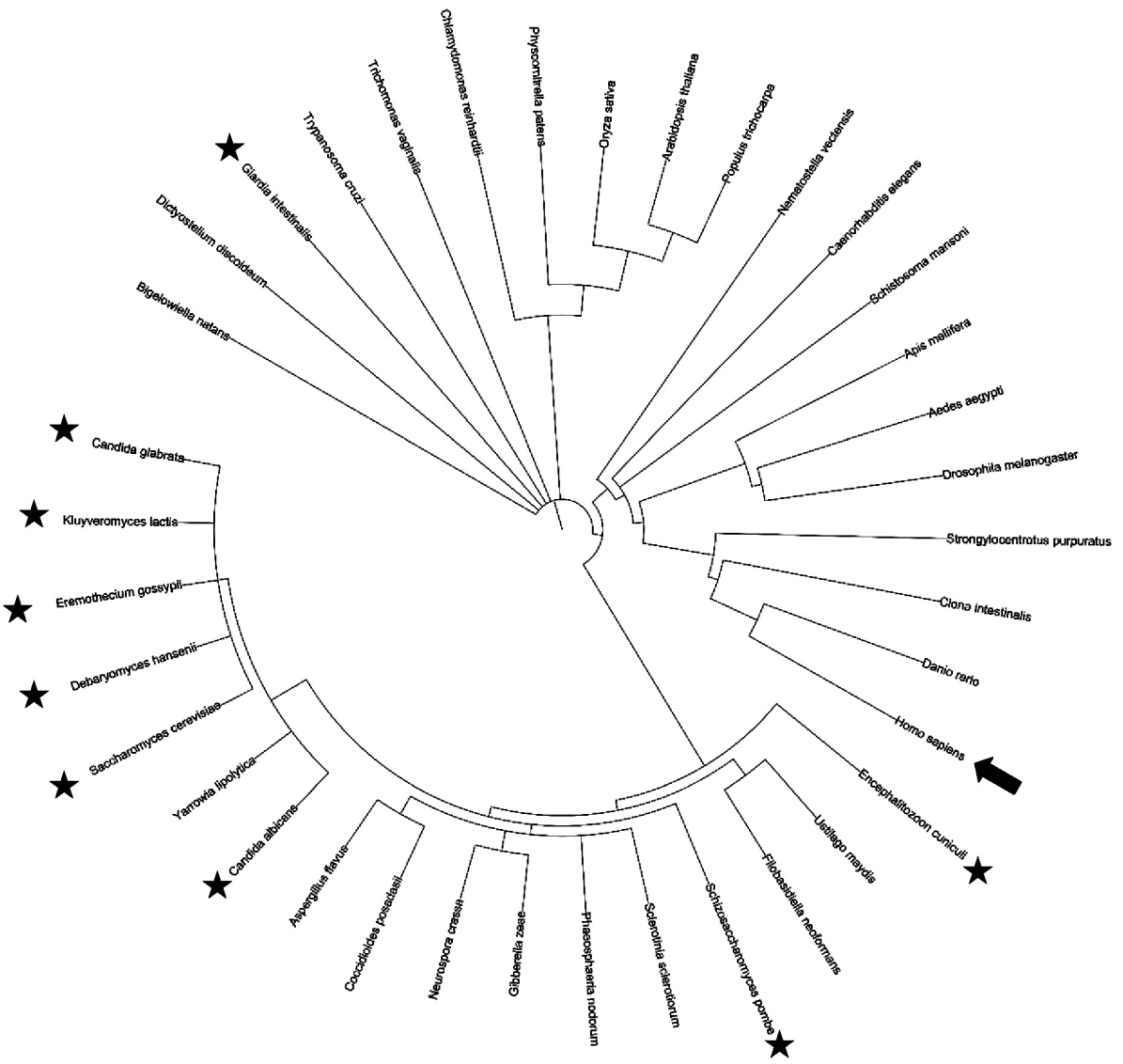

Fig. S2. Phylogenetic relationship among species presented in Fig. S1. *Homo sapiens* is indicated with the arrow. Additional species that lack a detectable CFI_m25 homolog are designated by stars. The tree was produced by iTOL (http://itol.embl.de)

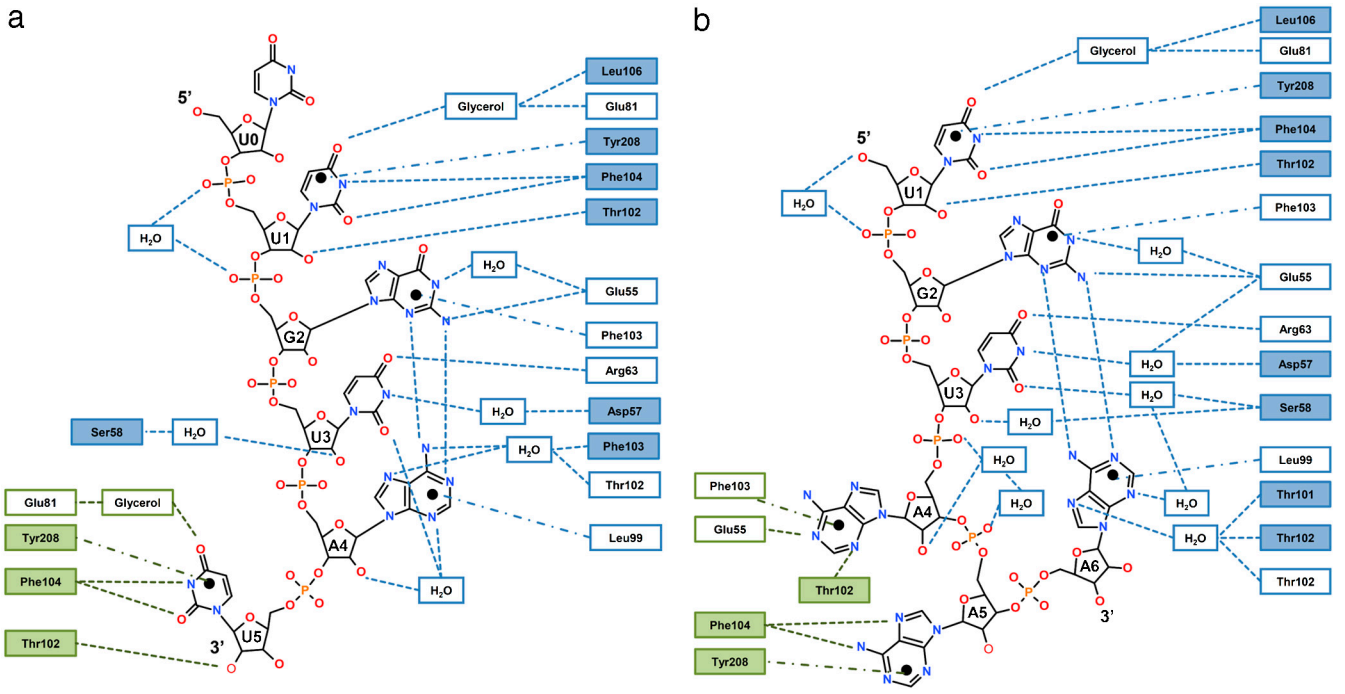

Fig. S3. Schematic diagram of interactions between CFI_m25 and (*A*) UUGUAU and (*B*) UGUAAA. Interactions with Mol A and Mol B_s are colored in blue and green, respectively. Side chain interactions are indicated by an open box and main chain interactions are represented with an filled box. Hydrogen bond interactions are represented with dashed lines whereas stacking interactions are represented by dashed lines ending with a filled circle. The representation of protein-RNA interactions was inspired by (12)

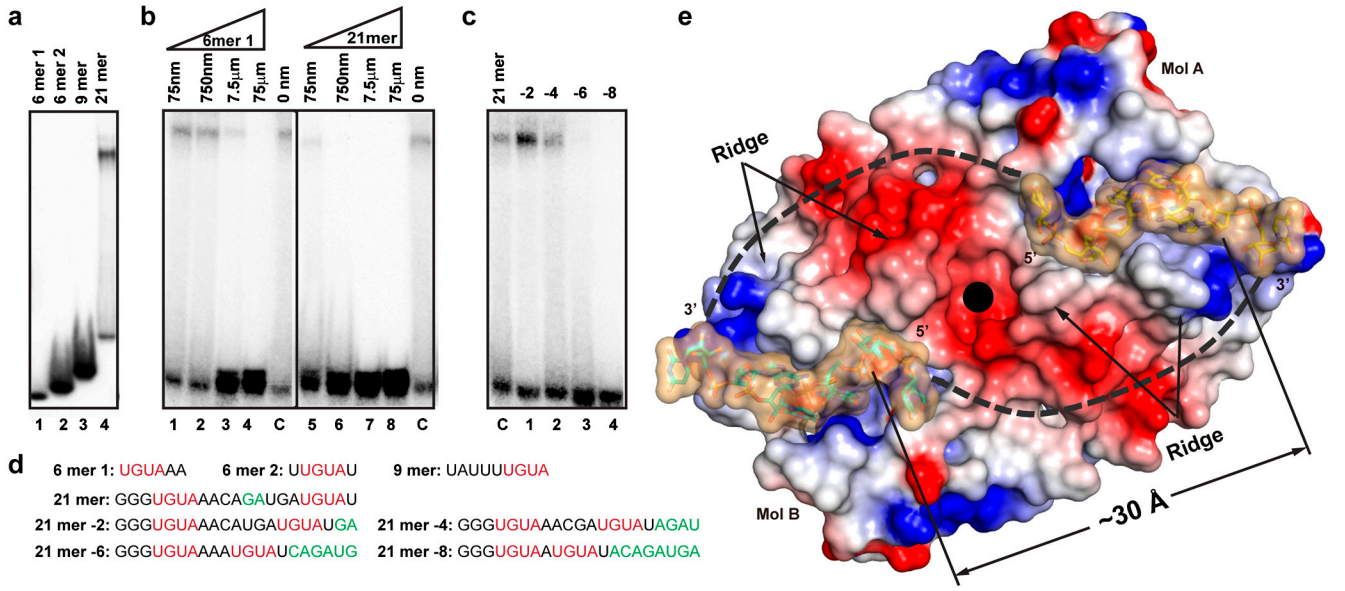

Fig. S4. The CFI_m25 homodimer binds two UGUA elements within a single RNA. (*A*) EMSA of RNAs containing either one or two UGUA elements. The sequence of each of the 5′ end-labeled RNA oligos is illustrated in (*D*). (*B*) Competition assay. EMSA was performed using 1 μM $CFI_m25\Delta N21$ and 2 nM α^{32}P-GTP-labeled 21 mer PAPOLA. Increasing amounts of the unlabeled UGUAAA oligo or the GGGUGUAAACAGAUGAUGUAU oligo were added to the binding reaction. (*C*) CFI_m25 binding of 21 mer PAPOLA RNA deletion variants by EMSA. Sequences of various lengths were deleted from the region between the two UGUA elements and appended to the 3′ end. (*D*) The sequence of the RNA variants used for EMSA. UGUA elements are highlighted in red, and internally deleted nucleotides appended to the 3′ end are highlighted in green. (*E*) Two potential paths traversed by an RNA containing two UGUA elements are illustrated by dashed lines. The estimated minimal length required to connect two CFI_m25-bound UGUA elements is ~30 Å, which is approximately the length spanned by a 5 nucleotide RNA. Arrows denote the ridge on the surface of the CFI_m25 homodimer that parallels the proposed RNA paths.

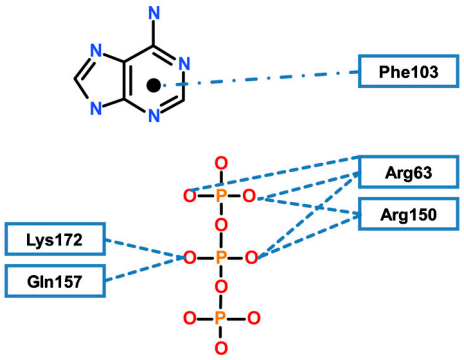

Fig. S5. Schematic diagram of interactions between CFI_m25 and Ap_4A. The figure is in the same orientation as in Fig. S3. Arg63 and Phe103 participate in the binding of Ap_4A as well as UGUA, suggesting that binding of these two substrates is mutually exclusive.

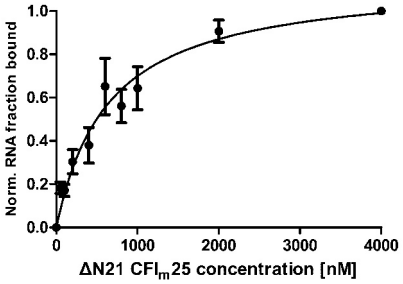

Fig. S6. Dissociation constant (K_d) determined by gel electrophoretic mobility shift assay. The truncated $CFI_m25\Delta N21$ and wild type α^{32}P-GTP-labeled RNAs containing human PAPOLA upstream sequences (−56 to −39 relative to the poly(A) cleavage site) were prepared as described in *Materials and Methods*. The gel electrophoretic mobility shift assay was carried out the in same manner as in *Materials and Methods*, except that the protein concentration was varied from 50 nM to 4,000 nM (50 nM, 100 nM, 200 nM, 400 nM, 600 nM, 800 nM, 1,000 nM, 2,000 nM, and 4,000 nM). RNA quantitation was done using a Personal Molecular Imager FX (Bio-Rad). The counts present in the RNA:protein complex (B) and free RNA (F) were recorded after subtracting the background. The RNA binding affinity of CFI_m25 was designated as the percentage of bound RNA (P), where $P = B/(B + F)$. The experiments were performed in triplicate and the binding affinity for the protein variants and RNA mutations were plotted relative to wild type PAPOLA RNA and $CFI_m25\Delta N21$ (4,000 nM). The averaged data were fit with a single hyperbola [$y = ax/(b + x)$] curve. The K_d is 645 ± 139.0 nM (number that follows the ± sign is a standard error).

Table S1. Data collection and refinement statistics

	CFI_m25-UGUAAA	CFI_m25-UUGUAU
PDB ID	3MDI	3MDG
Data collection		
Space group	$P2_12_12_1$	$P2_12_12_1$
Cell dimensions		
a, *b*, *c* (Å)	71.36, 72.04, 96.16	71.88, 73.88, 96.73
α, β, γ (°)	90, 90, 90	90, 90, 90
Resolution (Å)	20 − 2.07(2.15 − 2.07) *	20 − 2.22(2.35 − 2.22)
R_{merge} (%)	9.8(40.1)	11.7(35.3)
$I/\sigma I$	14.82(3.56)	9.3(3.83)
Completeness (%)	97.2(87.3)	93.8(83.7)
Redundancy	6.8(5.9)	3.7(3.4)
Refinement		
Resolution (Å)	19.93 − 2.07	19.68 − 2.22
No. reflections	58,072	48,585
R_{work}/R_{free} (%)	17.97/23.28 †	18.33/23.82 ‡
R.m.s deviations		
Bond lengths (Å)	0.006	0.007
Bond angles (°)	1.130	1.184
Wilson B-factors ($Å^2$)	31.08	39.90
B-factors ($Å^2$)		
Protein	26.02	35.29
Ligands	24.99	40.87
Water	28.97	36.00
Ramachandran plot statistics		
Most favored regions	96.45%	96.25%
Allowed regions	3.55%	3.75%
Disallowed regions	0%	0%

*Values in parentheses are for highest-resolution shell.

†R_{free} was calculated with 7% of the reflections not used in refinement.

‡R_{free} was calculated with 8% of the reflections not used in refinement.