Cell Research (2011) 21:1039-1051.

www.nature.com/cr

ORIGINAL ARTICLE

Structural basis of pre-mRNA recognition by the human cleavage factor I_m complex

Heng Li[1, 2], Shuilong Tong[1, 2], Xu Li[1, 2], Hui Shi[1, 2], Zheng Ying[1], Yongxiang Gao[1, 2], Honghua Ge[1, 2], Liwen Niu[1, 2], Maikun Teng[1, 2]

[1]Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China; [2]Key Laboratory of Structural Biology, Chinese Academy of Sciences, Hefei 230026, China

The cleavage factor I_m (CF I_m), consists of a 25 kDa subunit (CF I_m25) and one of three larger subunits (CF I_m59, CF I_m68, CF I_m72), and is an essential protein complex for pre-mRNA 3′-end cleavage and polyadenylation. It recognizes the upstream sequence of the poly(A) site in a sequence-dependent manner. Here we report the crystal structure of human CF I_m, comprising CF I_m25 and the RNA recognition motif domain of CF I_m68 (CF I_m68RRM), and the crystal structure of the CF I_m-RNA complex. These structures show that two CF I_m68RRM molecules bind to the CF I_m25 dimer via a novel RRM-protein interaction mode forming a heterotetramer. The RNA-bound structure shows that two UGUAA RNA sequences, with anti-parallel orientation, bind to one CF I_m25-CF I_m68RRM heterotetramer, providing structural basis for the mechanism by which CF I_m binds two UGUAA elements within one molecule of pre-mRNA simultaneously. Point mutation and kinetic analyses demonstrate that CF I_m68RRM can bind the immediately flanking upstream region of the UGUAA element, and CF I_m68RRM binding significantly increases the RNA-binding affinity of the complex, suggesting that CF I_m68 makes an essential contribution to pre-mRNA binding.

Keywords**:** cleavage factor I_m (CF I_m); pre-mRNA processing; poly(A) site recognition; RRM domain; RNA binding
Cell Research (2011) **21**:1039-1051. doi:10.1038/cr.2011.67; published online 12 April 2011

Introduction

Eukaryotic pre-mRNAs are synthesized and post-transcriptionally modified in the nucleus, before being exported into the cytoplasm to serve as templates for protein synthesis. The post-transcriptional modifications comprise 5′-end capping, splicing and 3′-end formation of the pre-mRNA. The maturation of the 3′-ends of most mRNA is catalyzed by multiple protein complexes, and requires the endonucleolytic cleavage of primary transcripts and the addition of poly(A) tails to the upstream cleavage products.

In mammals, the factors that are required for mRNA maturation *in vitro* include the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulatory factor (CstF), cleavage factors I_m and II_m (CF I_m and CF II_m), poly(A) polymerase (PAP) and nuclear poly(A) binding protein (PABN2). The cleavage reaction requires CPSF, CstF, CF I_m, CF II_m, and PAP. CPSF binds the highly conserved AAUAAA hexamer upstream of the cleavage site, and CstF binds the GU/U-rich sequence downstream of the cleavage site [1]. CPSF and CstF interact to form a stable complex before binding the pre-mRNA to recognize the two elements *in vivo* [2, 3]. CF I_m binds the pre-mRNA substrate in the vicinity of the poly(A) site concomitantly with CPSF. This stabilizes the binding between CPSF and the AAUAAA hexamer, facilitating pre-mRNA 3′-end processing complex assembly, and therefore enhances the rate and overall efficiency of poly(A) site cleavage *in vitro* [4-6]. Sequence-specific binding of CF I_m to pre-mRNA directs A(A/U) UAAA-independent poly(A) addition through interaction with the poly(A) polymerase and a CPSF subunit, hFip1 [7]. After cleavage, CPSF remains bound to the upstream

Correspondence: Maikun Teng[a], Liwen Niu[b]
[a]Tel: 86-551-3606314; Fax: 86-551-3601443
E-mail: mkteng@ustc.edu.cn
[b]E-mail: lwniu@ustc.edu.cn
Received 3 August 2010; revised 25 October 2010; accepted 17 December 2010; published online 12 April 2011

cleavage fragment, and recruits PAP onto the 3′-end of pre-mRNA. It also cooperates with PABN2 in the addition of a 250-nucleotide long poly(A) tail to the upstream cleavage fragment [8]. SELEX analysis has shown that CF I_m recognizes a five-nucleotide element, UGUAN (N = A > U ≥ C/G) with high affinity [7]. When added to partially purified 3′-end processing factors, recombinant CF I_m is sufficient to reconstitute poly(A) site cleavage activity *in vitro* (the CF I_m complex used in this study was a CF I_m25-CF I_m68 complex, as discussed further below) [5]. Repression of CF I_m activity by knocking down CF I_m25 does not affect HeLa cell viability, but increases the usage of the upstream poly(A) site, suggesting that CF I_m25 has an important role in poly(A) site selection [9].

CF I_m has been characterized as a heterodimer, consisting of a 25 kDa subunit (CF I_m25) and one of three larger subunits (CF I_m59, CF I_m68 or CF I_m72) [5]. CF I_m25, which is also known as NUDT21 or CPSF5, is a 227-amino acid polypeptide, which is highly conserved in metazoan, and which contains a nucleoside diphosphate linked to some other moiety, x (NUDIX) hydrolase domain (residues 79-203) [10]. CF I_m68, a member of the SR family of splicing factors, is a 551-amino acid polypeptide, which features an RNA recognition motif (RRM) domain at its N-terminal region, a central proline-rich region, and a C-terminal arginine/serine-rich (RS) domain. The RRM domain, which is also known as a RNA-binding domain (RBD) or ribonucleoprotein domain (RNP), is a motif found commonly in all organisms. It is characterized by an RNP1 consensus sequence (K/R-G-F/Y-G/A-F/Y-V/I/L-X-F/Y) and an RNP2 consensus sequence (V/I/L-F/Y-V/I/L-X-N/L) formed by aromatic and positively charged residues [11-13]. The RS domain is required for protein-protein interactions with other RS domains [1, 14].

In this study, we report the structure of CF I_m, comprising CF I_m25 (residues 34-227) and the RRM domain of CF I_m68 (CF I_m68RRM, residues 78-159), and the structure of a CF I_m25-CF I_m68RRM-RNA complex. The structural and mutational data reveals a novel RRM-protein binding mode, in which two CF I_m68RRM molecules bind to a CF I_m25 homodimer to form a heterotetramer. The structure of the CF I_m-RNA complex shows that two UGUAA RNA sequences, with anti-parallel orientation, bind simultaneously to one CF I_m25-CF I_m68RRM heterotetramer. Kinetic analyses demonstrate that the complex assembly increases RNA-binding affinity, and subsequent mutagenesis analyses reveal that CF I_m68 interacts with the immediately flanking upstream region of the UGUAA element via the L_3 loop of the RRM domain of CF I_m68.

Results

CF I_m68RRM is sufficient for CF I_m25 binding

In an *in vitro* binding assay, the N-terminal region of CF I_m68 (CF I_m68N, residues 1-226) has been shown to interact with CF I_m25 [15]. The central and C-terminal regions of CF I_m68 (residues 209-551) do not interact with CF I_m25 [15]. We have carried out detailed characterization of the region of CF I_m68 responsible for CF I_m25 binding. Pull-down assays showed that both CF I_m68N and CF I_m68RRM bind to GST-CF I_m25 (Supplementary information, Figure S1A). As the molecular weight of CF I_m68N is similar to that of GST alone, the GST-Rtt106p fusion was used as a negative control in binding studies. We also tested whether the N-terminal extension (RRMN, residues 1-80) or the C-terminal extension (RRMC, residues 160-226) of CF I_m68RRM interacts with CF I_m25. Immunoblot analysis showed neither His-MBP-RRMN nor His-MBP-RRMC binds to GST-CF I_m25 (Supplementary information, Figure S1B). These results demonstrate that the RRM domain of CF I_m68 is sufficient for CF I_m25 binding.

Overall structure of the CF I_m25-CF I_m68RRM complex

To better characterize CF I_m, we attempted to determine the structure of the CF I_m25-CF I_m68RRM complex using X-ray crystallography (Table 1). Crystals obtained from the full-length CF I_m25 in a complex with CF I_{m-}68RRM were not of an adequate quality to allow data collection. In the structure of apo-CF I_m25, residues 1-20 were not observed in the electron density map and residues 21-32 formed an extended loop structure [16]. To obtain crystals for high-resolution studies, a truncated CF I_m25 protein, with residues 1-33 removed, was constructed. A crystal, which diffracted to a resolution of 2.7 Å, was obtained and the structure was subsequently determined (see Materials and Methods for details). The CF I_m25-CF I_m68RRM complex was found to be a heterotetramer (approximate dimensions of 95 Å × 75 Å × 60 Å), with a pseudodyad passing through the heterotetramer, relating the pair of heterodimers (Figure 1). The heterotetrameric state of the CF I_m25-CF I_m68RRM complex was confirmed by size-exclusion chromatography, with the complex eluting with an apparent molecular weight of around 61.5 kDa (Supplementary information, Figure S2B and S2C). As the molecular weights of CF I_m25 and CF I_m68RRM are 24 and 10 kDa, respectively, this result suggests that the CF I_m25-CF I_m68RRM complex exists as a heterotetramer in solution. The heterotetrameric state is also consistent with previous reports that two subunits of CF I_m form a heterotetramer in solution [17].

Table 1 Data collection and refinement statistics

Data collection statistics		
	CF I_m25-CF I_m68RRM	CF I_m 25-CF I_m68RRM-RNA
Space group	C2	P21
Cell dimensions	a = 160.44 Å,	a = 104.78 Å,
	b = 105.69 Å,	b = 129.42 Å,
	c = 147.08 Å,	c = 111.16 Å,
	$\alpha = \gamma$ = 90.00°,	$\alpha = \gamma$ = 90.00°,
	β = 112.72°,	β = 94.01°,
Wavelength (Å)	1.000	1.000
Resolution (Å)	30.0-2.7 (2.77-2.70)[a]	50.0-2.9 (2.98-2.90)[a]
R_{merge} (%)[c]	8.3 (38.2)[a]	9.1 (22.2)[a]
I/σ (I)	14.7 (3.1)[a]	12.4 (3.5)[a]
Completeness (%)	98.9 (99.3)[a]	95.92 (74.42)[a]
Redundancy	3.4	3.0
Refinement statistics		
Resolution (Å)	15.0-2.7	50.0-2.9
Number of reflections	58226	59607
R_{work} / R_{free} (%)[c]	21.4/26.5	22.6/27.6
Number of atoms in protein	13257	17768
Number of water molecules	252	69
Rmsd bond (Å)	0.0105	0.006
Rmsd angle (°)	1.326	1.005
B-value (Å^2)	36.4	27.8
Most favoured (%)	95.5	97.3
Additional allowed (%)	4.4	2.7
Outlier (%)	0.1	

Values in parentheses correspond to the highest resolution shell.

[a] High-resolution shell is shown in parentheses.

[b] $R_{merge} = \Sigma|I_i-\langle I\rangle|/\Sigma|I|$, where I_i is the intensity of an individual reflection and <I> is the average intensity of that reflection.

[c] $R_{work} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$ for all reflections and $R_{free} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$, calculated on the 5% of data excluded from refinement.

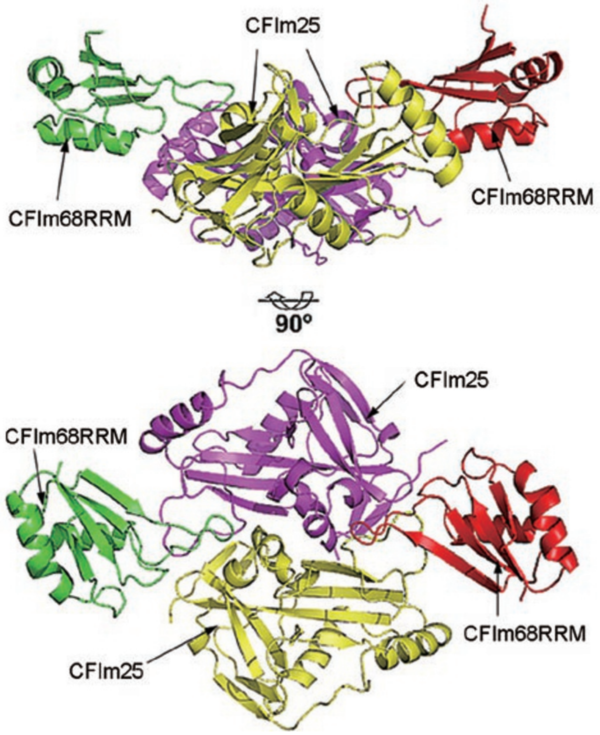


Figure 1 Overall structure of the CF I_m25-CF I_m68RRM complex. View of the CF I_m25-CF I_m68RRM complex in two orientations. The lower ribbon diagram of the complex is rotated by 90° around the horizontal axis relative to the upper one. A CF I_m25 dimer (yellow and magenta) binds two CF I_m68RRM molecules (green and red) forming a heterotetramer.

Structure of CF I_m68RRM

Although the sequence identity of CF I_m68RRM with other RRM domains is less than 30%, CF I_m68RRM adopts the classical compact αβ sandwich structures observed in other RRM domains, with a topology of $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ (Figure 2A). Residues 80-86 (β_1), 110-114 (β_2), 127-131 (β_3), 155-158 (β_4) constitute the four-stranded anti-parallel β-sheet arranged in the order of $\beta_4\beta_1\beta_3\beta_2$, from left to right, when facing the sheet, and two α helices, the α_1 helix (residues 94-104) and the α_2 helix (residues 134-146), pack against the β-sheet (Figure 2B). The RNP1 motif (residues 124-KGFALVGV-131) and the RNP2 motif (residues 83-LYIGNL-88) are located in the β_3 and β_1 strands, respectively. CF I_m68RRM has only 26% and 22% sequence identities with the RRM domain of CBP20 [18, 19] and the second RRM domain of sex-lethal protein (SXL-RRM2) [20], respectively. However, the program DALI [21] revealed that CF I_m68RRM is structurally similar to both CBP20 (*Z* score = 12.5) and SXL-RRM2 (*Z* score = 13.1). Superimposition of CBP20 and SXL-RRM2 with CF I_m68RRM showed RMS deviations at Cα positions of 1.01 Å with 52 residues and 1.30 Å with 57 residues, respectively, (Figure 2C). Tyr84 and Phe126 in RNP2 and RNP1 motifs of CFI$_m$68RRM are equivalent to Tyr43 and Phe83 in CBP20, respectively, and Tyr84 inCF I_m68RRM is equivalent to Tyr214 in SXL-RRM2.

Interface between CF I_m25 and CF I_m68RRM

Complex assembly between the CF I_m25 dimer and one molecule of CF I_m68RRM buries a surface area of 1 200 Å^2 (AREAIMOL [22]). The L_1 and L_3 loops of CF I_m68RRM interact with the L_{10} loop of CF I_m25 (Figure 3). This binding interface is strengthened by three hydrogen bonds and by more than 10 hydrophobic contacts. Tyr158, Tyr160 and His164 of CF I_m25 form three hydrogen bonds with Arg118, Asn120 and Glu116 of CF I_m68RRM, respectively. Tyr158 of CF I_m25 forms hydrophobic contacts with the backbone of Ala119, Asn120 and Gly121 of CF I_m68RRM, and Ala163 of CF I_m25 interacts hydrophobically with the backbone of Trp90 and Trp91 of CF I_m68RRM. His164 of CF I_m25 forms hydrophobic interactions with the backbone of Ser123 of CF I_m68RRM. Compared with the previously published

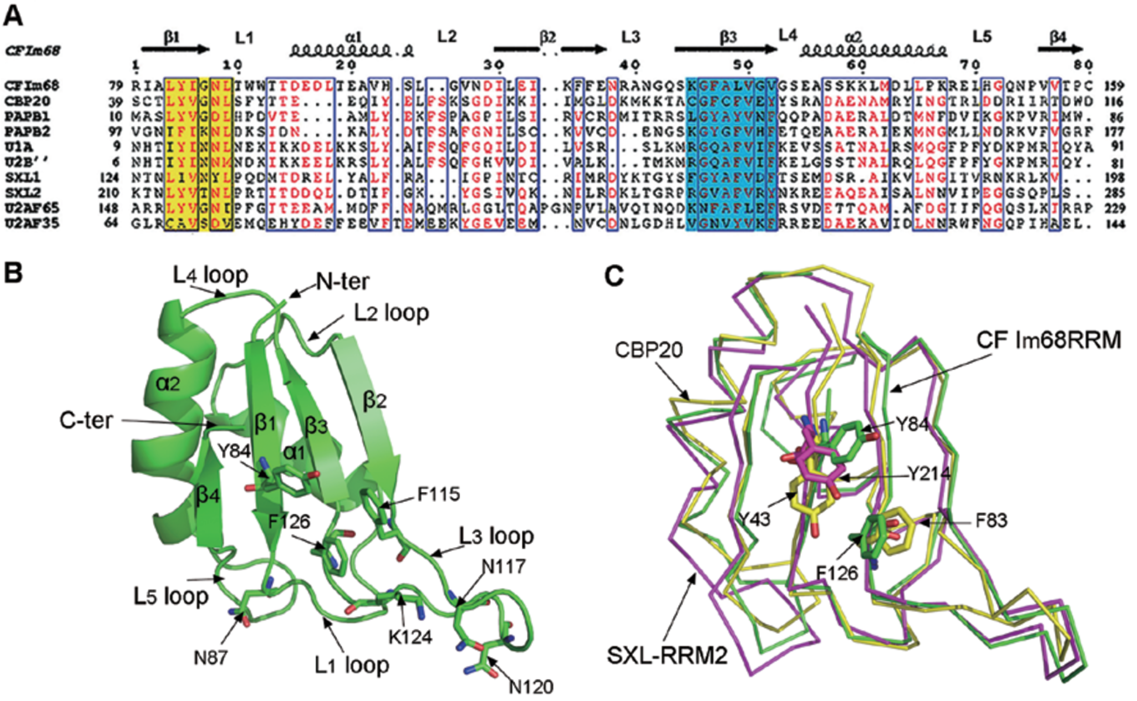

A
CFIm68
β1
L1
α1
L2
β2
L3
β3
L4
α2
L5
β4
1
10
20
30
40
50
60
70
80
CFIm68 79 RIALYIGNLTWWTTDEDLTEAVH.SL..GVNDILEI..KFFENRANGQSKGFALVGVGSEASSKKLMDLLPKRELHGQNPVVTPC 159
CBP20 39 SCTLYVGNLSFYTTE....EQIY.ELFSKSGDIKKI..IMGLDKMKKTACGFCFVEYYSRADAENAMRYINGTRLDDRIIRTDWD 116
PAPB1 10 MASLYVGDLHPDVTE....AMLY.EKFSPAGPILSI..RVCRDMITRRSLGYAYVNFQQPADAERALDTMNFDVIKGKPVRIMW. 86
PAPB2 97 VGNIFIKNLDKSIDN....KALY.DTFSAFGNILSC..KVVCD..ENGSKGYGFVHFETQEAAERAIEKMNGMLLNDRKVFVGRF 177
U1A 9 NHTIYINNLNEKIKKDELKKSLY.AIFSQFGQILDI..LVSR...SLKMRGQAFVIFKEVSSATNALRSMQGFPFYDKPMRIQYA 91
U2B'' 6 NHTIYINNLNDKIKKEELKRSLY.ALFSQFGHVVDI..VALK...TMKMRGQAFVIFKELGSSTNALRQLQGFPFYGKPMRIQY. 81
SXL1 124 NTNLIVNYLPQDMTDREL.YALF.RA...IGPINTC..RIMRDYKTGYSFGYAFVDFTSEMDSR.AIKVLNGITVRNKRLKV... 198
SXL2 210 KTNLYVTNLPRTITDDQL.DTIF.GK...YGSIVQK..NILRDKLTGRPRGVAFVRYNKREEAQEAISALNNVIPEGGSQPLS.. 285
U2AF65 148 ARRLYVGNIPFGITEEAM.MDFF.NAQMRLGGLTQAPGNPVLAVQINQDKNFAFLEFRSVDETTQAM.AFDGIIFQGQSLKIRRP 229
U2AF35 64 GLRCAVSDVEMQEHYDEFFEEVFTEMEEKYGEVEEM...NVCDNLGDHLVGNVYVKFRREEDAEKAVIDLNNRWFNGQPIHAEL. 144
B
L4 loop
N-ter
L2 loop
α2
β2
C-ter
β1
β3
Y84
α1
β4
F115
F126
L3 loop
N117
L5 loop
K124
N87
L1 loop
N120
C
CBP20
CF Im68RRM
Y84
Y214
Y43
F83
F126
SXL-RRM2

Figure 2 Structure of CF I_m68RRM. **(A)** Structure-based alignment of RRM domains. RNP1 and RNP2 are highlighted in cyan and yellow. CF I_m68RRM adopts the canonical αβ sandwich structure with a $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ topology. **(B)** Structure of CF I_m68RRM. Four β-strands constitute an anti-parallel β-sheet arranged in the order of $\beta_4\beta_1\beta_3\beta_2$, from left to right when facing the sheet, and two α helices pack against the β-sheet. The mutated residues in Figure 6C are indicated by ball-and-stick representation and highlighted. **(C)** Superimpositions of CF I_m68RRM with CBP20 and SXL-RRM2 (Pymol). CF I_m68RRM, CBP20 and SXL-RRM2 are colored in green, yellow and magenta, respectively. Tyr84 and Phe126 of CF I_m68RRM, Tyr43 and Phe83 of CBP20 and Tyr214 of SXL-RRM2 are indicated by ball-and-stick representation and are highlighted.

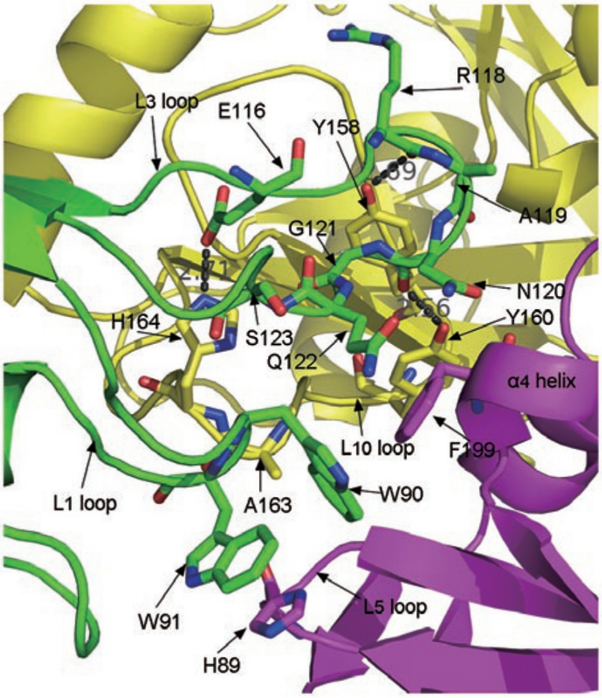


Figure 3 The interfaces between the CF I_m25 dimer and CF I_m68RRM. The L_{10} loop of CF I_m25 (yellow) interacts with L_1 and L_3 loops of CF I_m68RRM (green). The L_5 loop and α_4 helix of the other CF I_m25 molecule (magenta) of the dimer interacts with the L_1 loop of CF I_m68RRM. The residues involved in these contacts are indicated by ball-and-stick representation and are highlighted.

structure of apo-CF I_m25 [16], the side chain of His164 is rotated by 120° and reaches into a hydrophobic pocket, which is negatively charged at the top but is hydrophobic at the bottom, interacting with many residues of CF I_m68RRM (Supplementary information, Figure S3). These interactions fix the L_3 loop of CF I_m68RRM in an extended conformation. CF I_m68RRM also interacts with the second CF I_m25 molecule of the homodimer via hydrophobic contacts (Figure 3). The L_1 loop of CF I_m68RRM interacts with the L_5 loop and α_4 helix of CF I_m25. His89 of CF I_m25 interacts with Trp91 of CF I_m68RRM, and Phe199 of CF I_m25 interacts with Trp90, Asn120 and Gln122 of CF I_m68RRM.

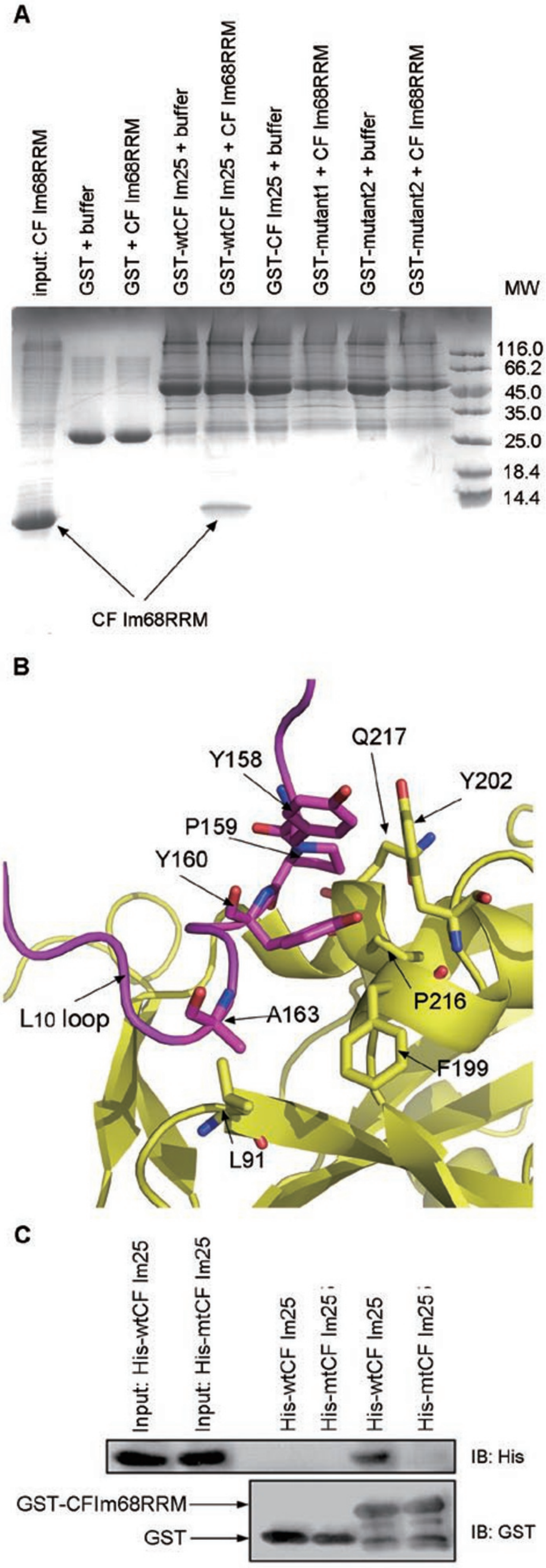
Mutational analysis supports structural model for RRM domain binding

To validate the two interfaces observed in the structure, two mutants of GST-tagged CF I_m25 were generated (GST-mutant1 and GST-mutant2). In GST-mutant1, Tyr158, Tyr160 and His164 were substituted and in GST-mutant2, His89 and Phe199 were substituted. Pull-down assays showed that CF I_m68RRM efficiently bound to GST-tagged wild-type CF I_m25 (GST-wtCF I_m25), whereas CF I_m68RRM did not bind to either the GST-mutant1 or the GST-mutant2 (Figure 4A). These results suggest that CF I_m68RRM interacts with both CF I_m25 molecules of the dimer, indicating the importance of CF I_m25 dimerization for CF I_m68RRM binding.

Although the L_{10} loop of CF I_m25 (residues 151-168) is a tortuous loop, it is well ordered in the apo-CF I_m25 structure. Superimposition of the CF I_m68RRM-bound CF I_m25 structure with the apo structure showed that the L_{10} loops of CF I_m25 adopt almost the same conformation. The conformation of the L_{10} loop of CF I_m25 could be important for CF I_m68RRM binding. Detailed structural analyses revealed that Leu91, Phe199, Tyr202, Pro216 and Gln217 of one CF I_m25 molecule of the dimer hydrophobically interacts with Tyr158, Ala163, Pro159 and Tyr160 of the L_{10} loop of the other CF I_m25 molecule and that these interactions may stabilize the L_{10} loop (Figure 4B). Mutagenesis and immunoblot analyses were performed to investigate this hypothesis. A mutant of CF I_m25 (mtCF I_m25) was generated, in which Leu91, Tyr202, Pro216 and Gln217 were replaced by alanines to eliminate the proposed stabilization of the L_{10} loop. Size-exclusion chromatography showed that mtCF I_m25 eluted similarly to the wild-type CF I_m25 (wtCF I_m25), indicating that these mutations do not affect dimerization (Supplementary information, Figure S2D). Immunoblot analysis showed that wtCF I_m25 bound efficiently to GST-CF I_m68RRM, whereas mtCF I_m25 did not (Figure 4C). Thus, substitution of four of the five residues (Leu91, Tyr202, Pro216 and Gln217) markedly impairs CF I_{m}-68RRM binding, suggesting that conformation of the L_{10} loop of CF I_m25 is possibly important for CF I_m68RRM binding.

Structure of the CF I_m25-CF I_m68RRM-UGUAA complex

We determined the crystal structure of the CF I_m25-CF I_m68RRM-UGUAA complex (Table 1). The UGUAA element binds to CF I_m25 in a positively charged cavity formed by the NUDIX domain (Figure 5A). As the electron density map for the chain S of the UGUAA sequences is the best (Figure 5B), the chain S is taken to describe the interactions between CF I_m25 and the UGUAA element. The RNA strand twists by about 90° after U3, flipping A4 out of the binding cavity, and twists back right after A4 (Figure 5A). The main chains of Tyr208 and Gly209 make hydrophobic contacts with the uracil base of U1, whereas Phe103 stabilizes the guanine base of G2 through a base-stacking interaction (Supplementary information, Figure S4A). The N3 and O2 of U1 form hydrogen bonds with the main chain carbonyl and amino



groups of Phe104, respectively, (Supplementary information, Figure S4A). The O4 and N3 of U3 form hydrogen bonds with the N2 group of the side chain of Arg63 and main chain carbonyl of Ser58, respectively (Supplementary information, Figure S4A). A5 interacts with Leu99 and Gly100 via hydrophobic contacts (Supplementary information, Figure S4B). The N6 and N1 of A5 form hydrogen bonds with the N3 group of G2 and O4* group of the sugar ring of U3, respectively, (Supplementary information, Figure S4B). A4 does not interact with CF I_m25.

Yang *et al*. [17] reported the structure of the CF I_m25-UGUAAA complex, which showed that the RNA hexamer was bound by one molecule of the CF I_m25 dimer and partially by an adjacent molecule in the crystal (Figure 5C). In addition, the authors found that the CF I_m25 dimer can bind two UGUA elements in solution. Consistent with this observation, our structure shows that two UGUAA elements bind simultaneously, in an antiparallel orientation, to one CF I_m25 dimer (Figure 5D). Superimposition of these two structures showed RMS deviations of Cα positions of 0.344 Å with 175 residues (Supplementary information, Figure S4C). The first three nucleotides (U1, G2 and U3) of UGUAA and UGUAAA adopt the same conformation and are recognized by CF I_m25 in a similar, but not identical manner. Arg63, Phe103, Phe104, Y208 and G209 interact with the first three nucleotides of both UGUAA and UGUAAA. In the CF I_m25-CF I_m68RRM-UGUAA complex, UGUAA interacts with Ser58, whereas in the CF I_m25-UGUAAA complex, UGUAAA interacts with Glu55, Asp57, Glu81, Thr102 and Leu106. The Glu81-U1 and Thr102-U1 interactions require a glycerol molecule. In our studies, no glycerol was added, and accordingly, no Glu81-U1 or Thr102-U1 interactions were observed. The conformation of the 3′-ends of UGUAA and UGUAAA is different. U3, A4, A5 and A6 of UGUAAA form a larger U shape, and A4 and A5 are located in the RNA-binding

Figure 4 The CF I_m25 dimerization is important for CF I_m68RRM binding. **(A)** Coomassie blue-stained protein SDS gel shows that GST-wtCF I_m25 can efficiently bind to CF I_m68RRM, but GST-mutant1 and GST-mutant2 does not, suggesting that CF I_m68RRM interacts with both CF I_m25 molecules of the dimer. **(B)** The other CF I_m25 molecule of the dimer hydrophobically interacts with the L_{10} loop. The L_{10} loop is shown in magenta and the other CF I_m25 molecule is shown in yellow. Residues involved in hydrophobic contacts are indicated by ball-and-stick representation and are highlighted. **(C)** Immunoblot analysis shows that the wild-type CF I_m25 (wtCF I_m25) efficiently binds to GST-CF I_m68RRM, whereas mtCF I_m25 does not. Neither wtCF I_m25 nor mtCF I_m25 binds to the GST tag.

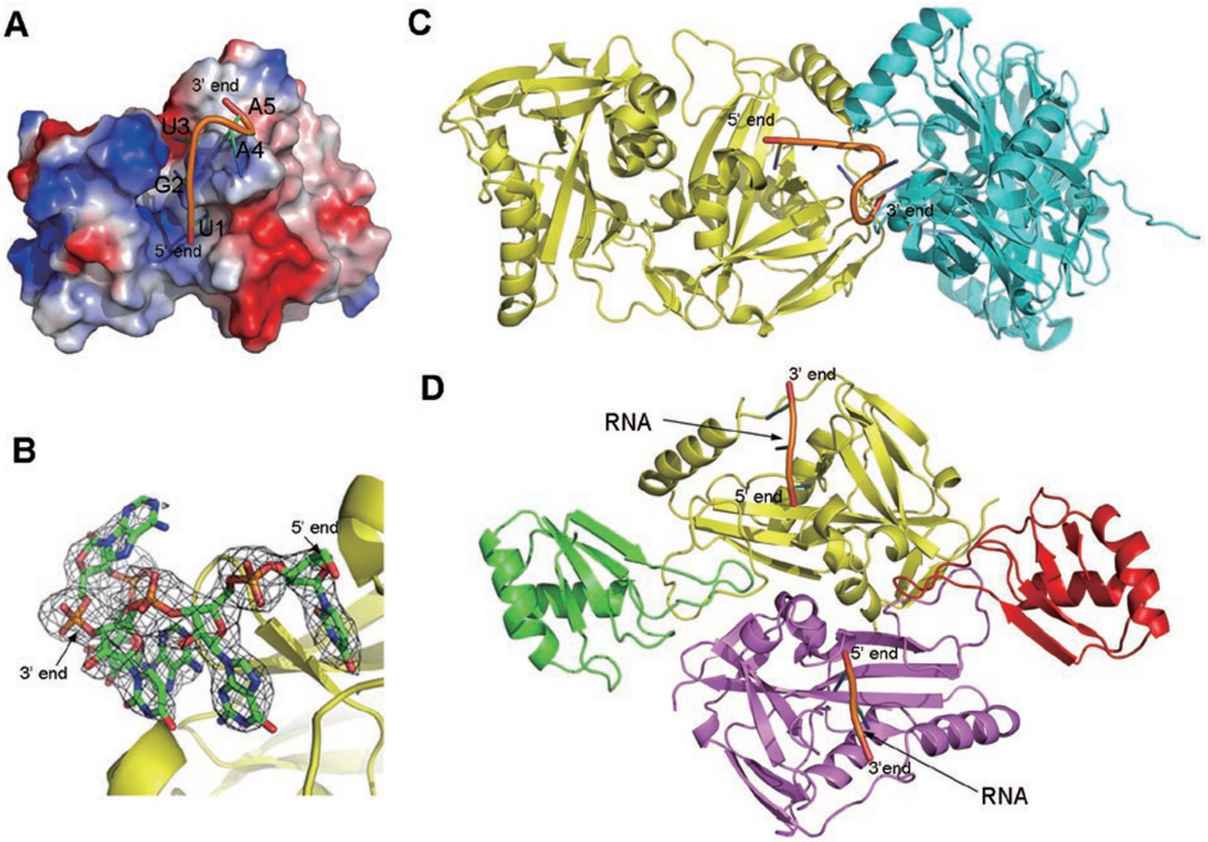


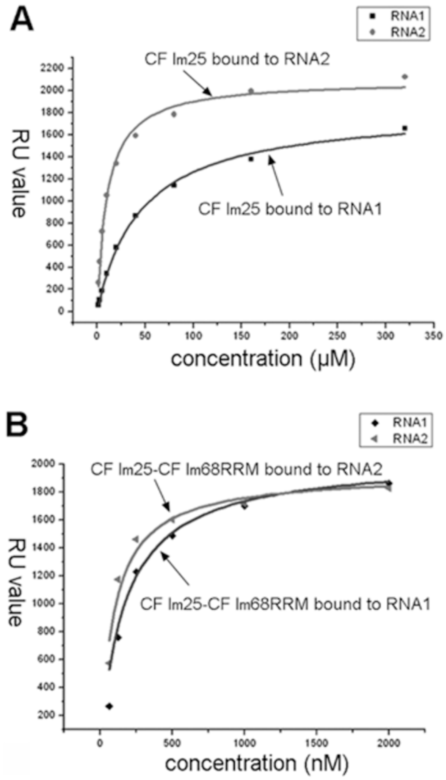
Figure 5 Structure of the CF I_m25-CF I_m68RRM-RNA complex. **(A)** The UGUAA element is located in a positively charged cavity of CF I_m25, with the exception of A4. Red represents positive charges and blue represents negative charges. **(B)** View of the UGUAA element covered with a $2F_o - F_c$ electron density map contoured at 0.9 σ. The UGUAA element is illustrated as a ball-and-stick scheme. **(C)** Structure of UGUAAA-CF I_m25 complexes (PDB entry 3MDI). The RNA hexamer, UGUAAA, is bound by one molecule of CF I_m25 dimer (yellow) and partially by an adjacent molecule (cyan) in the crystal. **(D)** View of the CF I_m25-CF I_m68RRM-UGUAA complex. Two RNA molecules with anti-parallel orientation bind to the same sites of the CF I_m25-CF I_m68RRM heterotetramer.

pocket of the adjacent dimer. In the CF I_m25-UGUAAA complex, Glu55 and Thr102 interact with A4, whereas in the CF I_m25-CF I_m68RRM-UGUAA complex, A4 does not interact with CF I_m25 and A5 interacts with Leu99 and Gly100.

CF I_m25 and CF I_m68RRM cooperate to bind RNA

Human *PAPOLA* pre-mRNA (encoding poly(A) polymerase α, PAPα) has a canonical poly(A) site and multiple copies of the UGUAA element upstream of the AAUAAA element. CF I_m binds to the UGUAA elements of *PAPOLA* pre-mRNA and promotes 3′-end processing [7]. To validate whether CF I_m assembly contributes to RNA binding, two RNA segments derived from human *PAPOLA* pre-mRNA, comprising the UGUAA element at the 3′-end were synthesized and used in surface plasmon resonance (SPR) assays. The two RNA segments are 5′-GCUAUUUUGUAAACA-3′ (RNA1) and 5′-CUAUUUUGUAA-3′ (RNA2). CF I_m25 bound to RNA1 and RNA2 with K_ds of 46 and 10 μM, respectively, whereas the CF I_m25-CF I_m68RRM complex bound with K_ds of 170 and 100 nM, respectively, (Figure 6A and 6B). Consistent with previous studies [15], CF I_m68RRM alone showed almost no binding to RNA (Supplementary information, Figure S5A). These results reveal that the CF I_m25-CF I_m68RRM complex assembly significantly enhances RNA-binding affinity.

Superimposition of the CF I_m25 dimer with the CF I_m25-CF I_m68RRM complex shows that CF I_m68RRM binding does not change the overall conformation of the CF I_m25 dimer, and Dettwiler *et al*. [15] have shown that CF I_m25 binding enables CF I_m68RRM to bind RNA. Thus, CF I_m68RRM may directly interact with RNA to enhance RNA-binding affinity. The anti-parallel β-sheet of the RRM domain is the major RNA-binding surface, and the highly conserved aromatic and positively charged



C

Site directed mutagenesis of CF I_m68RRM solvent exposed residues (in complex with CF I_m25)

mutants	Kd （nM）
Wt	173 ± 40
Y84A	136 ± 29
N87A	141 ± 25
F115A	158 ±25
N117A	279 ± 16
N120A	315 ± 59
K124A	341 ± 57

Figure 6 RNA binding analysis. Dissociation constants (K_ds) were derived using the 1:1 binding model. **(A)** CF I_m25 bound to RNA1 and RNA2 with K_ds of 46 and 10 μM, respectively. **(B)** The CF I_m25-CF I_m68RRM complex bound to RNA1 and RNA2 with K_ds of 170 and 100 nM, respectively. **(C)** Summary of the binding properties of site-directed mutants of CF I_m68RRM. Mutants of CF I_m68RRM were co-purified with CF I_m25.

residues from RNP1 and RNP2 motifs are generally involved in interactions with consecutive RNA bases. The β-sheet of CF I_m68RRM is exposed to solvent, suggesting that CF I_m68RRM may bind the RNA substrate via this region. To validate this hypothesis, mutagenesis analysis was performed. Point mutations of selected residues of CF I_m68RRM were generated and these mutants were co-purified with CF I_m25 for RNA-binding tests. As shown in Figure 6C, the mutants, N117A, N120A and K124A, showed reduced RNA-binding affinities, whereas the mutants, Y84A, N87A and F115A, showed similar RNA-binding affinities compared with the wild-type protein. We also generated a mutant in which the aromatic residue in the RNP1 motif, F126, was replaced by an alanine, but this mutant could not be obtained as a stable protein for kinetic analysis. Y84 and F115 are located in the β-sheet, indicating that, unusually for an RRM domain, the β-sheet of CF I_m68RRM is not involved in RNA recognition. N117, N120 and K124 are located in the L_3 loop of CF I_m68RRM, suggesting that the L_3 loop of CF I_m68RRM not only interacts with CF I_m25 but may also interact with RNA.

Discussion

Comparison with other RRM-protein structures

Biochemical and structural studies have revealed that RRM domains are involved in protein-protein interactions as well as in RNA recognition [23, 24]. Previous studies have shown that CF I_m68 interacts with CF I_m25 via the RRM domain and that the substitution of two residues within the RNP2 motif (G86V, N87V) abolished interaction with CF I_m25 [15]. We initially speculated that CF I_m68RRM might interact with CF I_m25 via its β-sheet. Interestingly, the structure of the CF I_m25-CF I_m68RRM complex shows that CF I_m68RRM interacts with CF I_m25 via its L_1 and L_3 loops, leaving the β-sheet exposed to solvent. To date, about 10 structures of RRM domain-protein/polypeptide complexes have been determined. The recognition mechanisms of proteins by RRM domains are very diverse, and no general mechanism has emerged. The RRM domain interacts with other proteins through its β-sheet [25-28], or α-helices [29, 30], or α-helices and L_4 loops [18, 19, 31]. To date, only one RRM-protein complex structure has been reported in which the L_3 loop is involved in interactions with other protein (the Mago-Y14-PYM complex [26]). Comparison of the CF I_m25-CF I_m68RRM complex with the Mago-Y14-PYM complex shows that CF I_m68RRM binds to CF I_m25 in a different manner. In the Mago-Y14-PYM complex, Y14 interacts with PYM through part of its L_3 loop and interacts with Mago through the entire β-sheet. In the CF I_m25-CF I_m68RRM complex, the β-sheet is not involved in protein recognition and the entire L_3 loop is involved in RRM-protein interaction. The novel structural information presented here demonstrates the diversity of protein recognition mechanisms, which underlie RRM domain binding.

CF I_m25 dimerization is crucial for UGUAA recognition and complex assembly

Recently, CF I_m has been shown to be a heterotetramer in solution [17]. Our structure confirms that CF I_m25 and CF I_m68RRM forms a heterotetramer, induced by CF I_m25 dimerization. Whether this dimerization of CF I_m25 is important for the activity of CF I_m is of interest, and our studies indicate that this dimerization of CF I_m25 is important for complex assembly. The structural and mutational data presented here show that CF I_m68RRM interacts with both molecules of the CF I_m25 dimer and that CF I_m25 dimerization may stabilize the conformation of the L_{10} loop of CF I_m25, which is important for CF I_m68RRM binding. CF I_m25 dimerization also enables the simultaneous binding of two UGUAA elements to the CF I_m25 dimer. Yang *et al*. [17] observed that the CF I_m25 dimer bound RNA containing two separated UGUAA elements with 100-fold higher affinity than RNA containing only one UGUAA element. The minimum distance between the two UGUA elements, which is required to observe this gain in affinity is five nucleotides. Unlike the structures of the UGUAAA- and UUGUAU-CF I_m25 complexes, which show CF I_m25 dimer binding with only one RNA molecule in the crystals [17], our structure shows that two UGUAA RNA sequences, with anti-parallel orientation, bind to one CF I_m25-CF I_m68RRM heterotetramer, providing structural evidence to support the simultaneous binding of two UGUAA elements to the CF I_m25 dimer. The anti-parallel positioning of the two UGUAA elements and the discontiguous RNA-binding surfaces confirm the importance of the separation by a certain number of bases of the two UGUAA elements.

CF I_m25 enables CF I_m68 to bind RNA through stabilization of the L_3 loop of CF I_m68

UV-crosslinking experiments showed that CF I_m68 binds to pre-mRNA substrate very weakly, but can efficiently bind to RNA upon complex formation with CF I_m25 [15]. Our studies therefore suggest that CF I_m25 may enable CF I_m68 to bind RNA. Although the β-sheet of CF I_m68RRM contains the conserved RNP1 and RNP2 motifs found in other RRM domains (Figure 2A) and is accessible for RNA, our studies show that it is not involved in RNA binding. Y84, N87 and F115 are located in the β_1 strand, L_1 loop and β_2 strand, respectively. The Y84A, N87A and F115A mutants showed similar RNA-binding affinities to the wildtype, suggesting that the β-sheet does not interact with RNA. This observation is supported by the recently determined structure of the CF I_m25-CF I_m59RRM complex (PDB entry 3N9U). The structure of the CF I_m25-CF I_m59RRM complex shows that the β-sheet is buried by a helix formed by the C-terminal extension of the RRM domain (Supplementary information, Figure S6). As CF I_m59 and CF I_m68 have highly homologous amino acid sequences, the β-sheet of CF I_m68RRM is possibly also buried by a helix and is inaccessible to RNA. Substitutions of N117, N120 and K124, which are located in the L_3 loop, reduce RNA-binding affinities, suggesting the L_3 loop of CF I_m68RRM may bind to RNA. This observation is particularly interesting. In SXL-RRM2 [20] and PABP RRM1 [32], the L_3 loops interact with RNA but not with protein, whereas in the Mago-Y14-PYM complex, the L_3 loop interacts only with protein. To our knowledge, the RRM domain of CF I_m68 is the first example of an RRM domain that binds to both protein and RNA via the L_3 loop. Without CF I_m25 binding, the L_3 loop of CF I_m68RRM is thought to be more flexible, which may affect RNA binding. Upon complex formation with CF I_m25, the L_3 loop of CF I_m68RRM is extended and accessible for RNA binding. CF I_m25 may enable CF I_m68 to bind RNA in this manner.

CF I_m68 makes an essential contribution to RNA binding

The CF I_m25-CF I_m68RRM complex binds to RNA1 and RNA2 much more efficiently than CF I_m25 alone (270-fold for RNA1 and 100-fold for RNA2), indicating that the large subunit, CF I_m68, makes an essential contribution to pre-mRNA recognition. Point mutation and kinetics analyses show that the L_3 loop of CF I_m68RRM may be involved in RNA binding. As the UGUAA element orientates opposite to CF I_m68RRM from the 5′ to 3′-end, the L_3 loop of CF I_m68RRM may bind the immediately flanking upstream region of the UGUAA element. Complex assembly places the L_3 loop of CF I_m68RRM and the NUDIX domain of CF I_m25 together, forming a large and continuous RNA-binding platform (Supplementary information, Figure S7A). We observed that CF I_m25 binds to RNA1 much more weakly than RNA2 (about fivefold). The program UNAFold [33] indicated that RNA1 folds into a hairpin structure (Supplementary information, Figure S7B). The U-A and G-C intramolecular interactions bury the UGUAA element, impairing the binding of CF I_m25. Pre-mRNA processing, such as splicing, is influenced by the secondary structure of the pre-mRNA [34]. Most proteins involved in splicing regulation recognize single-stranded, rather than base-paired RNA. The hairpin structure of RNA reduces the binding of CF I_m25 by about fivefold but has a much smaller impact on the CF I_m25-CF I_m68RRM complex (less than twofold), suggesting that CF I_m68 binding helps CF I_m25 to recognize the UGUAA element, besides promoting a higher binding affinity.

During 3′-end processing, CF I_m binds to pre-mRNA concomitantly with CPSF to stabilize the binding of CPSF to the AAUAAA hexamer at an early stage of processing complex assembly [5, 6]. It binds to the UGUAN element upstream of the poly(A) site in a sequence-dependent manner [4, 35]. The CF I_m25 dimer simultaneously binds to two UGUA elements within one molecule of pre-mRNA (termed as two UGUAA elements binding mode) [17]. In addition, CF I_m25 and CF I_m68 cooperates to bind the UGUAA element and the immediately upstream flanking region, respectively, for higher affinity (termed as synergistic binding mode). Bioinformatic analyses revealed that the two modes are both functionally important for pre-mRNA recognition by CF I_m (Table 2). Approximately 43.6% of human mRNAs (a total number of 44 563) contain A(A/U)UAAA elements immediately upstream of the poly(A) site and UGUAN elements upstream of A(A/U)UAAA and 29.3% contain multiple copies of UGUAN elements. We also analyzed the mRNAs from *Mus musculus*, *Danio rerio, Gallus gallus* and *Xenopus laevis* and found that the proportions are 42.8% and 28%, 38.6% and 28.9%, 27.8% and 19.7%, and 65.9% and 50.7%, respectively (Table 2). These results show that CF I_m binds to about half of all pre-mRNAs, of which the majority contain at least two copies of UGUAN elements and the minority contain only one copy of the UGUAN element. CF I_m may bind to pre-mRNAs containing at least two copies of UGUAN elements via both the two UGUA element-binding and the synergistic-binding modes. For the pre-mRNAs containing only one copy of the UGUAN element, which is the case for about 10% of all mRNAs, CF I_m binds the pre-mRNA via the synergistic binding mode to ensure efficient binding.

Materials and Methods

Protein expression and purification

The cDNAs of CF I_m25 (residues 34-227) and CF I_m68RRM (residues 78-159) were cloned into the vector pET22b. Cys159 of CF I_m68RRM was replaced by an alanine to prevent disulfide bond formation. GST or GST proteins were cloned into the pGEX-4T-2 vector. His-MBP-tag proteins were cloned into the modified pET32a vector, the products of which contain His-MBP-tags in the N-terminal of the recombinant proteins. Proteins were expressed in *Escherichia coli* Rosetta (DE3) (Novagen). Mutations were introduced using PCR by designing mutated residues in primers with the MutanBEST kit (TAKARA). All plasmids were confirmed by DNA sequencing. CF I_m25 and CF I_m68RRM were co-purified by Ni^{2+} ion-affinity chromatography, and further purified by Superdex-200 gel filtration and monoQ ion exchange chromatography (GE Healthcare). Finally, the CF I_m68-CF I_m68RRM complex was concentrated by centrifugal ultrafiltration (Millipore, 5 kDa cutoff) to approximately 8 mg/ml, as estimated using the BCA Protein Assay Kit (Pierce), in 2 mM Tris (pH 8.0), and 100 mM NaCl for crystallization assays.

Crystallization and data collection

The CF I_m25-CF I_m68RRM complex was crystallized at the concentration of approximately 8 mg/ml by hanging-drop vapor diffusion at 10 ºC with the reservoir solution containing 16% PEG3350, 5% dioxane and 0.1 M sodium citrate, pH 5.0. The 2.7 Å diffraction data was collected at beamline 3W1A of Beijing Synchrotron Radiation Facility (BSRF). The CF I_m25-CF I_m68RRM complex was incubated with twofold excess of UGUAA (TAKARA) on ice for 0.5 h, with a final concentration of approximately 6 mg/ml. The CF I_m25-CF I_m68RRM-UGUAA complex was crystallized by sitting-drop vapor diffusion at 10 °C with the reservoir solution containing 15% PEG3350, 9% dioxane and 0.1 M sodium citrate, pH 5.0. The 2.9 Å diffraction data was collected at beamline BL17U of Shanghai Synchrotron Radiation Facility (SSRF). The data were processed and scaled with the HKL2000 package [36].

Structure determination

The CF I_m25-CF I_m68RRM crystals belong to space group C2 with the cell parameters: a = 159.77 Å, b = 105.51 Å, c = 146.53 Å, and $\alpha = \gamma = 90.00°$, $\beta = 112.58°$. Each asymmetric unit contains three copies of a 2:2 CF I_m25-CF I_m68RRM complex. The phase was determined by molecular replacement using Molrep [37] and Phaser [38]. First, six CF I_m25 subunits were found using the structure of apo CF I_m25 (PDB entry 2cl3) as the search model and then fixed. Then six CF I_m68RRM subunits were found using the structure of RBMY protein RRM domain (PDB entry 2FY1) as the search model. The model completeness was carried out in COOT [39] and the refinement was performed by REFMAC5 [40] with non-crystallographic symmetry (NCS) restraints and CNS [41]. The NCS restraints were tight in earlier stages and completely re-

Table 2 Bioinformatics analyses of the UGUAN elements of mRNAs from multiple vertebrate organisms

Organisms	*Homo sapiens*	*Mus musculus*	Danio rerio	Gallus gallus	*Xenopus laevis*
Total number of mRNAs[a] (relative percentage)[b]	44563 (100.0%)	41910 (100.0%)	28699 (100%)	19131 (100%)	8549 (100%)
Number of mRNAs containing UGUAN upstream of A(A/U)UAAA[c]	19413 (43.6%)	17905 (42.8%)	11087 (38.6%)	5319 (27.8%)	5638 (65.9%)
Number of mRNAs containing multiple copies of UGUAN	13069 (29.3%)	11732 (28.0%)	8280 (28.9%)	3768 (19.7%)	4336 (50.7%)

[a]All the mRNA sequences were derived from the Refseq database of NCBI. Refseq database provides a multiple organism, non-redundant database of mRNA sequences.

[b]The total number of mRNA of each organism was set to 100%, respectively.

laxed in later stages. Water molecules were added to the model by inspection of $2F_o - F_c$ and $F_o - F_c$ density maps in the final stages. The final refined model has an R factor (R_{free}) of 21.4% (26.5%). The quality of model was checked using the program Molprobity [42]. The Ramachandran plot reveals that 95.5% of the residues are in the most favored region, with an additional 4.4% in the additionally allowed region. One residue is in the outlier region, which is Pro113. The CF I_m25-CF I_m68RRM-RNA crystals belong to space group P21 with the cell parameters: a = 104.78 Å, b = 129.42 Å, c =111.16 Å, and $\alpha = \gamma = 90.00°$, $\beta = 94.01°$. Each asymmetric unit contains four copies of CF I_m25-CF I_m68RRM heterotetramer. The phase was determined by molecular replacement with Phaser [38]. Seven RNA molecules can be traced in the electron density map within the cavities of the CF I_m25 subunits, whereas no electron density for RNA was observed within the same region of the remaining CF I_m25 subunit. Thus, seven RNA chains (termed chain Q, R, S T, U, V and W, respectively) were manually added with the guidance of $2F_o - F_c$ and $F_o - F_c$ density maps in Coot [39]. Chain Q contains U1 and the guanine base of G2, chain R contains all the five nucleotides with the exception of the adenine base of A4 and the phosphate group of A5, chain S contains all the five nucleotides, chain T contains the first three nucleotides and A5 with the exception of the phosphate group of A5, chain U contains the first three nucleotides and the adenine base of A5, chain V contains the first three nucleotides and chain W contains U1. The refinement was performed by REFMAC5 [40]. Finally, water molecules were added to the model by inspection of $2F_o - F_c$ and $F_o - F_c$ density maps. The translation-libration-screw (TLS) model was applied near the end of refinement. The final refined model has an R factor (R_{free}) of 22.6% (27.6%) and was validated using Molprobity [42]. The Ramachandran plot revealed that 97.3% of the residues are in the most favored region, with an additional 2.7% in the additionally allowed region. Structural figures were prepared with Pymol (http://www.pymol.org).

Glutathione S-transferase pull-down assays

Small-scale pull-down assays were performed. GST-proteins or GST tag (150 µg) from the soluble fraction of *E. coli* cell lysate was incubated with 75 µl of glutathione agarose beads (GE Healthcare) for 30 min at room temperature. After washing three times with binding buffer, beads were incubated with 150 µg of purified recombinant proteins for 1 h at 16 °C. After incubation, the beads were washed three times with binding buffer to remove unbound proteins. Bound proteins were eluted from the beads by boiling in SDS sample buffer and resolved on SDS-PAGE.

Immunoblot analysis

An aliquot containing 50 µg of protein from the soluble fraction of *E. coli* cell lysate expressing GST or GST-proteins was incubated with 20 µl of glutathione agarose beads (GE Healthcare) for 20 min at room temperature. After washing four times with 1× PBS, beads bound with GST or GST-proteins were incubated with 50 µg of the recombinant His-MBP-tagged or His-tagged proteins in 0.25 ml HNTG-buffer (20 mM Hepes-KOH, pH 7.5, 100 mM NaCl, 0.1% Triton X-100, and 10% glycerol) for 1 h at 4 °C. After incubation, the beads were washed six times with 1 ml HNTG buffer to remove unbound proteins. Bound proteins were eluted from the beads by boiling in SDS sample buffer and detected by immunoblot analysis.

Size-exclusion chromatography assay

The size-exclusion chromatography assays were performed with a Superdex 200 column (10/300 GL) (GE Healthcare). The protein sample or molecular mass standards were applied to the Superdex 200 column (10/300 GL) and eluted with 50 mM Tris, pH 8.0 and 200 mM NaCl. Standard proteins (GE Healthcare) were thyroglobulin (669.0 kDa), ferritin (440.0 kDa), albumin (69.0 kDa), ovalbumin (43.0 kDa), ribonuclease A (13.7 kDa). The void volume was determined with blue dextran (GE Healthcare).

Surface plasmon resonance measurement

All SPR studies were performed with a Biacore 3000 instrument (Biacore AB, Uppsala, Sweden). The RNA segment with the same sequence found upstream of the human *PAPOLA* pre-mRNA poly(A) site (5′-GCUAUUUUGUAAACA-3′ (−63 to −49) or 5′-CUAUUUUGUAA-3′ (−62 to −42)) was immobilized on a SA chip via a biotin at the 3′-end, and solutions containing wild-type proteins and mutants at different concentrations were passed over the chip at 10 µl/min and washed by 50 mM NaOH. All experiments were carried out at 25 °C in binding buffer (20 mM Tris, pH 8.0, 150 mM NaCl). The binding curves were fitted according to a one-site binding model using Origin software (http://www.originlab.com). The raw sensorgrams data obtained with different concentrations of proteins are provided as Supplementary information, Figure S5.

Accession codes

The coordinates and structure factors have been deposited with accession number 3P5T (for the CF I_m25complex-CF I_m68RRM complex) and 3P6Y (for the CF I_m25-CF I_m68RRM-UGUAA complex).

Acknowledgments

We thank the staff at BSRF beamline 3W1A and SSRF beamline BL17U for assistance with synchrotron data collection and Dr S Barabino (University of Milano-Bicocca) for plasmids, and Dr C Zhao and Dr J Zang (University of Science and Technology of China) for their valuable advice on experiments and manuscript. Financial support for this project was provided by research grants from the National Natural Science Foundation of China (30025012, 30900224 and 10979039), the Chinese Ministry of Science and Technology (2006CB806500, 2006CB910200 and 2006AA02A318), the Chinese Academy of Sciences (KSCX2-YW-R-60), the Chinese Ministry of Education (20070358025) and the Natural Science Foundation of Anhui Province (090413081). Heng Li was supported by the Graduate Innovation Foundation of USTC (KD2007034).

References

1 Zhao J, Hyman L, Moore C. Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* 1999; **63**:405-445.

2 Takagaki Y, Manley JL. Complex protein interactions within the human polyadenylation machinery identify a novel component. *Mol Cell Biol* 2000; **20**:1515-1525.

3 Gilmartin GM, Nevins JR. An ordered pathway of assembly

of components required for polyadenylation site recognition and processing. *Genes Dev* 1989; **3**:2180-2190.

4 Brown KM, Gilmartin GM. A mechanism for the regulation of pre-mRNA 3′ processing by human cleavage factor Im. *Mol Cell* 2003; **12**:1467-1476.

5 Rüegsegger U, Blank D, Keller W. Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Mol Cell* 1998; **1**:243-253.

6 Rüegsegger U, Beyer K, Keller W. Purification and characterization of human cleavage factor Im involved in the 3′ end processing of messenger RNA precursors. *J Biol Chem* 1996; **271**:6107-6113.

7 Venkataraman K, Brown KM, Gilmartin GM. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* 2005; **19**:1315-1327.

8 Bienroth S, Keller W, Wahle E. Assembly of a processive messenger RNA polyadenylation complex. *EMBO J* 1993; **12**:585-594.

9 Kubo T, Wada T, Yamaguchi Y, Shimizu A, Handa H. Knockdown of 25 kDa subunit of cleavage factor Im in Hela cells alters alternative polyadenylation within 3′-UTRs. *Nucleic Acids Res* 2006; **34**:6264-6271.

10 Bessman MJ, Frick DN, O'Handley SF. The MutT proteins or "Nudix" hydrolases, a family of versatile, widely distributed, "housecleaning" enzymes. *J Biol Chem* 1996; **271**:25059-25062.

11 Dreyfuss G, Swanson MS, Pinol-Roma S. Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends Biochem Sci* 1988; **13**:86-91.

12 Swanson MS, Nakagawa TY, LeVan K, Dreyfuss G. Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins. *Mol Cell Biol* 1987; **7**:1731-1739.

13 Adam SA, Nakagawa T, Swanson MS, Woodruff TK, Dreyfuss G. mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence. *Mol Cell Biol* 1986; **6**:2932-2943.

14 Smith CW, Valcarcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* 2000; **25**:381-388.

15 Dettwiler S, Aringhieri C, Cardinale S, Keller W, Barabino SM. Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. *J Biol Chem* 2004; **279**:35788-35797.

16 Coseno M, Martin G, Berger C, *et al.* Crystal structure of the 25 kDa subunit of human cleavage factor Im. *Nucleic Acids Res* 2008; **36**:3474-3483.

17 Yang Q, Gilmartin GM, Doublié S. Structural basis of UGUA recognition by the Nudix protein CFIm25 and implications for a regulatory role in mRNA 3′ processing. *Proc Natl Acad Sci USA* 2010; **107**:10062-10067.

18 Calero G, Wilson KF, Ly T, *et al.* Structural basis of m7GpppG binding to the nuclear cap-binding protein complex. *Nat Struct Biol* 2002; **9**:912-917.

19 Mazza C, Ohno M, Segref A, Mattaj IW, Cusack S. Crystal structure of the human nuclear cap binding complex. *Mol Cell* 2001; **8**:383-396.

20 Handa N, Nureki O, Kurimoto K, *et al.* Structural basis for recognition of the tra mRNA precursor by the sex-lethal protein. *Nature* 1999; **398**:579-585.

21 Holm L, Sander C. Mapping the protein universe. Science 1996; **273**:595-603.

22 CCP4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D Biol Crystallogr* 1994; **50(Pt 5)**:760-763.

23 Clery A, Blatter M, Allain FH. RNA recognition motifs: boring? Not quite. *Curr Opin Struct Biol* 2008; **18**:290-298.

24 Maris C, Dominguez C, Allain FH. The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J* 2005; **272**:2118-2131.

25 Kadlec J, Izaurralde E, Cusack S. The structural basis for the interaction between nonsense-mediated mRNA decay factors UPF2 and UPF3. *Nat Struct Mol Biol* 2004; **11**:330-337.

26 Bono F, Ebert J, Unterholzner L, *et al.* Molecular insights into the interaction of PYM with the Mago-Y14 core of the exon junction complex. *EMBO Rep* 2004; **5**:304-310.

27 Lau CK, Diem MD, Dreyfuss G, Van Duyne GD. Structure of the Y14-Magoh core of the exon junction complex. *Curr Biol* 2003; **13**:933-941.

28 Fribourg S, Gatfield D, Izaurralde E, Conti E. A novel mode of RBD-protein recognition in the Y14-Mago complex. *Nat Struct Biol* 2003; **10**:433-439.

29 Selenko P, Gregorovic G, Sprangers R, *et al.* Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol Cell* 2003; **11**:965-976.

30 Kielkopf CL, Rodionova NA, Green MR, Burley SK. A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell* 2001; **106**:595-605.

31 Price SR, Evans PR, Nagai K. Crystal structure of the spliceosomal U2B"-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* 1998; **394**:645-650.

32 Deo RC, Bonanno JB, Sonenberg N, Burley SK. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* 1999; **98**:835-845.

33 Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybriziation. *Methods Mol Biol* 2008; **453**:3-31.

34 Hiller M, Zhang Z, Backofen R, Stamm S. Pre-mRNA secondary structures influence exon recognition. *PLoS Genet* 2007; **3**:e204.

35 Tian B, Hu J, Zhang H, Lutz CS. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* 2005; **33**:201-212.

36 Otwinowski Z, Minor W. Processing of x-ray diffraction data collected in oscillation mode. *Methods Enzymol* 1997; **276**:307-326.

37 Vagin A, Teplyakov A. MOLREP: an automated program for molecular replacement. *J Appl Cryst* 1997; **30**:1022-1025.

38 McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. Phaser crystallographic software. *J Appl Cryst* 2007; **40**:658-674.

39 Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 2004; **60(Pt 12 Pt 1)**:2126-2132.

40 Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method.

Acta Crystallogr D Biol Crystallogr 1997; **53(Pt 3)**:240-255.

41 Brunger AT, Adams PD, Clore GM, *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* 1998; **54(Pt 5)**:905-921.

42 Davis IW, Leaver-Fay A, Chen VB, *et al.* MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 2007; **35**:W375-W383.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)