Supplemental Materials for

Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end processing endonuclease

Corey R. Mandel,[1] Syuzo Kaneko,[1,2] Hailong Zhang,[1,2] Damara Gebauer,[1] Vasupradha Vethantham,[1] James L. Manley,[1] Liang Tong[1]

[1]*Department of Biological Sciences, Columbia University, New York, NY 10027, USA.*

[2]*These authors contributed equally to this work.*

Correspondence and requests for materials should be addressed to L.T. (e-mail: ltong@ columbia.edu).

Supplemental Results and Discussion

Quality of refined structures. The refined structures have excellent agreement with the observed diffraction data and the expected geometric parameters (Supplemental Tables 2 and 3). The majority of the residues are located in the most favored region of the Ramachandran plot (Supplemental Tables 2 and 3). Residues Asp 40 (motif 1) and His 158 (motif 3) are in the disallowed region of the Ramachandran plot. Both residues have clearly defined electron density, and the motif 1 residue is in the disallowed region in the other structures as well, such as L1 metallo-β-lactamase [1] and RNase Z [2]. Several segments of CPSF-73 are disordered in these crystals (1-6, 113-121, 289-299 in the presence, and 1-8, 113-121, 184-188, 272-304 in the absence of zinc), as is the case in the structure of yeast CPSF-100 (269-275, 389-396, 423-625). The last missing segment in CPSF-100, covering about 200 residues, was removed by the fungal protease during crystallization.

Structural homologs of the β-CASP domain. The closest structural homolog of the β-CASP domain is the nucleotide binding fold (NBF, Supplemental Fig. 2), with a *Z* score of 6 from Dali [3], but the sequence identity among structurally-aligned residues is less than 10%. NBFs are found in ATP synthase and many other proteins [4]. However, there is a crucial difference between the β-CASP domain and the NBF. The NBF contains the Walker A motif for binding nucleotide phosphate groups in the loop connecting strands β1 and β2 (Supplemental Fig. 2) [5]. In contrast, the β-CASP domain lacks the β1 strand of the NBF (Supplemental Fig. 2), and the CPSF-73 proteins do not contain the Walker A motif. Therefore, the β-CASP domain appears to be a novel example of the NBF super fold, but is unlikely to bind nucleotides.

Structural comparison between CPSF-73 and CPSF-100. In the metallo-β-lactamase domain, residues following strand β13, including motif C, have a different conformation

in CPSF-100 (Fig. 1*d*, Supplemental Fig. 3). This places motif C too far from the expected positions of the zinc ions in CPSF-100 (Fig. 1*d*), and suggests that simply introducing the zinc ligands into CPSF-100 would be unlikely to confer zinc binding ability to this protein.

The β-CASP domain of CPSF-100 is much larger than the corresponding domain of CPSF-73, covering residues 218-661. It contains an extra, anti-parallel strand (βF) in the central β-sheet, and its βG-βH hairpin structure is much more extensive (Fig. 2*b*). However, residues 423-625 are not observed in the structure of this domain in CPSF-100. The majority of these residues are charged and hydrophilic in nature (Supplemental Fig. 4). This segment is probably highly flexible in structure and prone to proteolysis, consistent with its removal during crystallization [6].

Two zinc binding sites on the surface of CPSF-73. When zinc was included in the crystallization solution, we observed binding of two additional zinc ions. These were on the surface of CPSF-73 and far from the active site (Supplemental Fig. 6), one of which was also in the crystal packing interface. The individual domains of the two structures of CPSF-73 are highly similar to each other, with rms distance of about 0.35 Å, and the zinc ions in the active site have the same interactions with CPSF-73. There is a small change in the orientation (about 7° rotation) of the β-CASP domain relative to the metallo-β-lactamase domain in the two structures (Supplemental Fig. 6), as well as local differences near the binding sites of the two additional zinc ions (Supplemental Fig. 6). The structure of CPSF-73 in the presence of additional zinc is used in the descriptions, as more residues of the protein are ordered in this crystal.

Comparison of zinc binding modes. The binding modes of the zinc ions in CPSF-73 (Fig. 3*a*) are similar to that in RNase Z but distinct from those observed in canonical metallo-β-lactamases (Fig. 3*b*). The most striking difference is the fact that the zinc

ligand in motif 5 is located just after strand β13 in CPSF-73 (Fig. 1*c*) and RNase Z (Fig. 1*f*), whereas it is located just after strand β12 in L1 metallo-β-lactamase (Fig. 1*e*). In fact, canonical metallo-β-lactamases do not contain strand β13, as it belongs to the segment from the C-terminal region of CPSF-73 (Supplemental Fig. 1*a*). In addition, the Asp residue in motif 4 plays a structural role in L1 metallo-β-lactamase and does not coordinate the zinc ions (Fig. 3*b*), whereas it bridges the interactions between the two zinc ions in CPSF-73 (Fig. 3*a*) and RNase Z.

The β-CASP domain and access to the active site. A significant overall re-positioning of the β-CASP domain relative to the metallo-β-lactamase domain is unlikely, as the two domains share a rather extensive interface, with about 1000 $Å^2$ buried surface area and having ion pair, hydrogen-bonding and van der Waals interactions (Supplemental Fig. 7). Moreover, a similar domain organization is seen in the structure of yeast CPSF-100 (Fig. 1*b*). On the other hand, the structure shows that the βE-αG loop in the β-CASP domain, containing the highly conserved Gly 356-Tyr-X-X-X-Gly 361 motif, may provide a gate to the active site (Fig. 1*a* and Supplemental Fig. 7). This loop could be flexible in structure, as it has higher than average *B* values and the Tyr side chain has weak electron density. The two conserved Gly residues could act as hinges for this segment. A conformational change in this loop may be sufficient to allow access to the active site by the pre-mRNA substrate.

CPSF-73 endonuclease activity. Purified CPSF-73 and a mutant derivative were assayed for endonuclease activity using standard 3' processing conditions (Supplemental Fig. 9*a*). Although activity was weak, it was observed with multiple independent preparations, and it co-fractionated precisely with CPSF-73 during the final gel filtration purification step (results not shown). The cleavage observed was largely resistant to zinc-specific chelators (results not shown), which contrasts with authentic 3' cleavage in nuclear extract [7]. The resistance of purified CPSF-73 likely reflects the tight binding of

the two Zn atoms. The sensitivity of the authentic reaction could reflect sensitivity of another essential factor, e.g. CPSF-30, which contains multiple zinc finger motifs. Alternatively, it may be possible that the active site of CPSF-73 is more accessible in the authentic 3' processing complex.

Implication for Artemis. The structure of CPSF-73 has significant implications for the DNA nuclease Artemis. Artemis shares weak sequence homology with CPSF-73 (Supplemental Fig. 1*b*) and is expected to adopt a similar structure. However, Artemis is shorter by about 30 amino acids at the N-terminus as compared to CPSF-73 and CPSF-100 (Supplemental Fig. 1*b*). As a consequence, Artemis lacks motif 1 of the metallo-β-lactamase fold (Supplemental Fig. 1*a*), which plays an important structural role in stabilizing the conformation of residues in motif 2 (Supplemental Table 1 and Supplemental Fig. 8). Moreover, Artemis is missing the first three β-strands (β2 through β4) of the structure (Supplemental Fig. 1*b*), which are crucial for forming the hydrophobic core of the central β-sandwich of the metallo-β-lactamase domain (Fig. 1*c*). We note however that a sequence encoding Asp-Ser-Gly, conforming to motif 1 (Supplemental Fig. 1*b*), is found just upstream of the apparent initiation codon of the Artemis gene. Further studies are needed to clarify the structure and function of this protein.

Methods

The details of the crystallographic analysis for the structure determination of CPSF-100 will be presented elsewhere, including our serendipitous discovery that *in situ* proteolysis by a fungal protease is crucial for the crystallization of this protein [6].

Protein expression and purification. Residues 1-460 of human CPSF-73 was sub-cloned into the pET28a vector (Novagen) and over-expressed in *E. coli* at 20 °C. The

expression construct introduced a hexa-histidine tag at the N-terminus. The soluble protein was purified by nickel-agarose affinity chromatography and gel-filtration chromatography. The protein was concentrated to 10 mg/ml in a buffer containing 20 mM Tris (pH 8.5), 250 mM NaCl, 5% (v/v) glycerol, and 5 mM DTT. Yeast CPSF-100 (YDH1p, residues 1-720) was expressed and purified by nickel-agarose affinity chromatography, anion exchange and gel-filtration chromatography, and concentrated to 10 mg/ml in a buffer containing 20 mM Tris (pH 8.5), 250 mM NaBr, 5% (v/v) glycerol, and 10 mM DTT.

Protein crystallization. Crystals of human CPSF-73 were obtained at room temperature by the sitting-drop vapor diffusion method. The reservoir solution contained 100 mM MOPS (pH 6.5), 300 mM sodium sulfate, and 16% (w/v) PEG 3350. In an attempt to increase the occupancy of zinc ions, some crystals were grown from a reservoir solution that also contained 0.5 mM $ZnCl_2$. The crystals were cryo-protected by the introduction of 25% (v/v) ethylene glycol, and frozen in liquid propane for data collection at 100 K. The crystals belong to space group $P2_12_12_1$, with cell parameters of a=58.5 Å, b=83.2 Å, and c=104.9 Å. There is one molecule of CPSF-73 in the asymmetric unit. The selenomethionyl protein samples of CPSF-73 failed to crystallize.

Crystals of yeast CPSF-100 were obtained at 4 °C by the same protocol. It was discovered that a solution that had been infected by a fungus was crucial for the crystallization of this protein, the details of which will be described elsewhere [6]. The reservoir solution contained 20% (w/v) PEG3350 and 0.2 M ammonium citrate. The crystals belong to space group $I222$, with cell parameters of a=75.6 Å, b=122.8 Å, and c=126.9 Å. There is one molecule of CPSF-100 in the asymmetric unit. Two other crystal forms were also obtained (Supplemental Table 3). We could not produce the selenomethionyl sample of CPSF-100 due to lack of protein expression, in a variety of media and host strains.

Data collection and processing. X-ray diffraction data were collected on an ADSC CCD at the X4A beamline of Brookhaven National Laboratory. The diffraction images were processed and scaled with the HKL package [8]. The diffraction data were collected at the zinc absorption peak for crystals of CPSF-73, and at the gold absorption peak for a $KAu(CN)_2$ derivative of CPSF-100. The data processing statistics are summarized in Supplemental Tables 2 and 3.

Structure determination and refinement. The structure of CPSF-73 was determined by the single-wavelength anomalous diffraction method [9], using the anomalous signal of zinc. The zinc sites were located with SnB [10], and the reflection phases were calculated with Solve/Resolve [11], which also placed about 50% of the residues. The atomic model was built with the program O [12], and the structure refinement was carried out with CNS [13].

The structure of CPSF-100 was determined by the single-isomorphous replacement method, supplemented with anomalous diffraction. The location of the Au atom was determined from isomorphous as well as anomalous difference Patterson maps, with the program Patsol [14]. The reflections were phased with Solve/Resolve [11], and atomic model was built with the program O [12]. The details of the crystallization and structure determination of CPSF-100 will be presented elsewhere [6]. The statistics on the structure refinement are summarized in Table 1, and additional statistics can be found in Supplemental Tables 2 and 3.

Pre-mRNA 3'-end cleavage assay. 5' capped SV40 late (SVL) and adenovirus L3 pre-mRNA substrates were prepared as described [15]. To prepare 5' end labeled SVL pre-mRNA, unlabeled RNA substrate was treated with alkaline phosphatase followed by T4 polynucleotide kinase and [γ-^{32}P]ATP. To prepare 3' end labeled SVL pre-mRNA, unlabeled RNA substrate was treated with T4 RNA ligase and labeled pCp. CPSF-73 was

pre-incubated with 5 mM $CaCl_2$ at 37 °C for 30 min. Cleavage assays were carried out in reaction mixture (10 μl) containing ~1 ng labeled RNA substrates, 10 mM Hepes (pH 7.9), 10 % (v/v) glycerol, 50 mM KCl, 0.25 mM DTT, 0.25 mM PMSF, 0.125 mM EDTA, 50 μM $CaCl_2$, RNase inhibitor (2 unit), 500 ng BSA and indicated amounts of recombinant CPSF-73. Cleaved RNAs were isolated and fractionated on 6% urea PAGE. The data were analyzed by Phospho-Imager.

References

1. Ullah, J. H. et al. The crystal structure of the L1 metallo-b-lactamase from Stenotrophomonas maltophilia at 1.7A resolution. J. Mol. Biol. 284, 125-136 (1998).
2. de la Sierra-Gallay, I. L., Pellegrini, O. & Condon, C. Structural basis for substrate binding, cleavage and allostery in the tRNA maturase RNase Z. Nature 433, 657-661 (2005).
3. Holm, L. & Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233, 123-138 (1993).
4. Abrahams, J. P., Leslie, A. G., Lutter, R. & Walker, J. E. Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. Nature 370, 621-628 (1994).
5. Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the a- and b-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J. 1, 945-951 (1982).
6. Mandel, C. R., Gebauer, D., Zhang, H. & Tong, L. A serendipitous discovery that in situ proteolysis is required for the crystallization of yeast CPSF-100. Acta Cryst., submitted (2006).
7. Ryan, K., Calvo, O. & Manley, J. L. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. RNA 10, 565-573 (2004).
8. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. Method Enzymol. 276, 307-326 (1997).
9. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science 254, 51-58 (1991).
10. Weeks, C. M. & Miller, R. The design and implementation of SnB v2.0. J. Appl. Cryst. 32, 120-124 (1999).
11. Terwilliger, T. C. SOLVE and RESOLVE: Automated structure solution and density modification. Meth. Enzymol. 374, 22-37 (2003).

12. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. A47, 110-119 (1991).
13. Brunger, A. T. et al. Crystallography & NMR System: A new software suite for macromolecular structure determination. Acta Cryst. D54, 905-921 (1998).
14. Tong, L. & Rossmann, M. G. Patterson-map interpretation with noncrystallographic symmetry. J. Appl. Cryst. 26, 15-21 (1993).
15. Takagaki, Y., Ryner, L. C. & Manley, J. L. Separation and characterization of a Poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. Cell 52, 731-742 (1988).
16. Carson, M. Ribbon models of macromolecules. J. Mol. Graphics 5, 103-106 (1987).
17. Evans, S. V. SETOR: hardware lighted three-dimensional solid model representations of macromolecules. J. Mol. Graphics 11, 134-138 (1993).
18. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins 11, 281-296 (1991).

Table 1. Conserved motifs in the CPSF-73 family of proteins

Motif	Consensus Sequence	Location in the structure	Structural and functional role
1	DXG	Right after β4	Stabilize residues in motif 2—Asp residue is hydrogen-bonded to the main-chain amide and side chain hydroxyl of the (S/T) residue of motif 2. Gly residue is packed against the last two residues (DH) of motif 2.
2	(S/T)HXHXDH	β5-α2 loop	Zinc ligands. The Asp residue may also help the ionization of the bridging water into a hydroxide.
3	H	β9-β10 loop	Zinc ligand
4	D	Right after β11	Bridging ligand of the two zinc ions
5 (C)	H	Right after β13	Zinc ligand
A	D/E	At the end of β12	Hydrogen-bonded to the side chain of motif B and the main-chain amides of residues right after β11 and β12
B	H	Linker between β-CASP and metallo-β-lactamase domains	Hydrogen-bonded to oxygen atom on the phosphate group of the scissile nucleotide and the side chain of motif A. General acid for catalysis

Table 2. Summary of crystallographic information

	Human CPSF-73 with 2 zinc ions	Human CPSF-73 with 4 zinc ions	Yeast CPSF-100 (Ydh1p)
Data collection			
Space Group	$P2_12_12_1$	$P2_12_12_1$	$I222$
Cell dimensions			
a, *b*, *c* (Å)	58.8, 82.6, 103.7	58.5, 83.2, 104.9	75.6, 122.8, 126.9
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90
Resolution (Å)	2.1 (2.18–2.1)	2.1 (2.18–2.1)	2.5 (2.59–2.5)
R_{merge} (%)	4.8 (29.6)	4.9 (40.3)	8.7 (40.8)
I/σI	23.9 (4.1)	29.5 (3.4)	24.5 (5.6)
Completeness	97 (90)	97 (90)	94 (82)
Redundancy	4.0 (4.1)	5.0 (4.2)	11.3 (10.4)
Refinement			
Resolution (Å)	30–2.1	30–2.1	30–2.5
No. reflections	29,144	55,894[1]	19,531
R_{work}/R_{free}	23.1/27.8	22.4/25.2	21.2/29.0
No. atoms			
Protein	3,236	3,471	3,936
Ligand/ion	7	9	0
Water	162	247	148
B-factors			
Protein	45.1	46.4	51.0
Ligand/ion	44.7	63.9	–
Water	48.3	53.8	45.0
R.m.s. deviations			
Bond lengths (Å)	0.006	0.006	0.007
Bond angles (°)	1.3	1.3	1.4

1. The Friedel pairs are included as independent reflections.

One crystal was used for each structure.

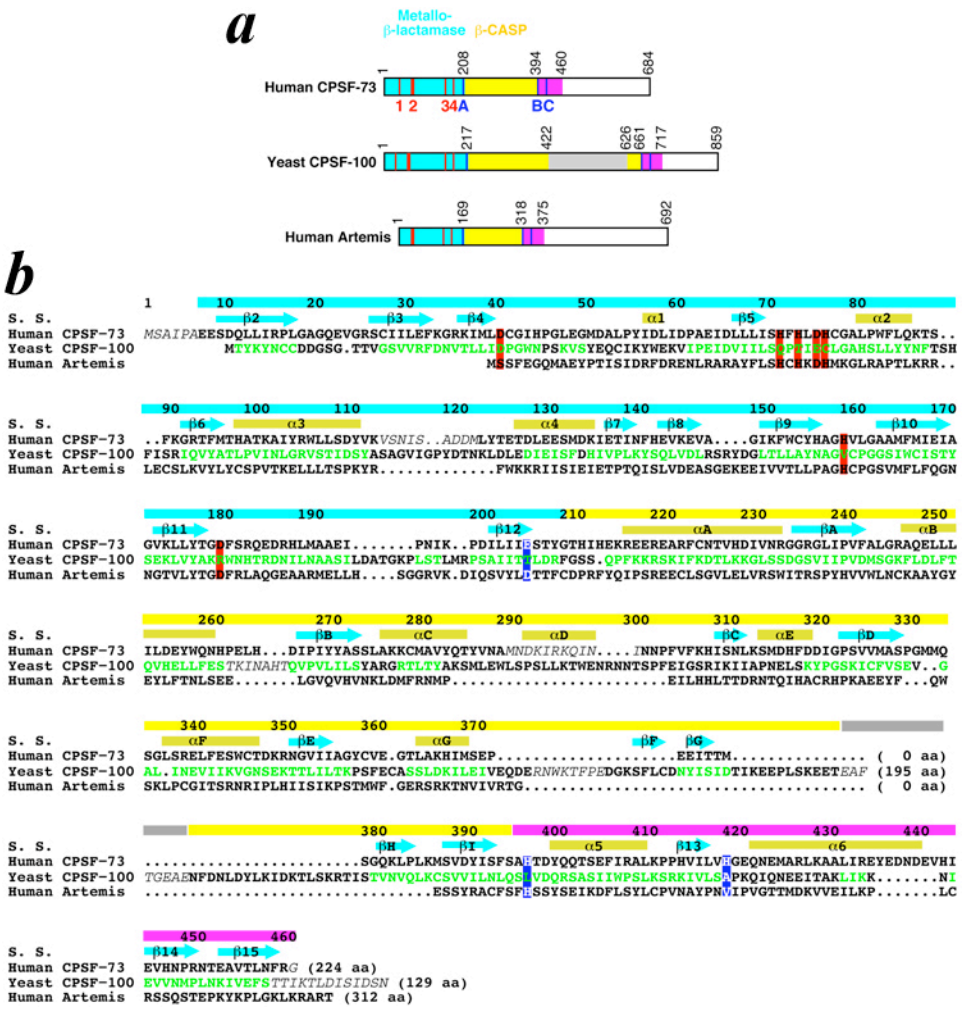

Fig. 1. Primary structures of CPSF-73 and CPSF-100. **(*a*)**. Domain organization of human CPSF-73, yeast CPSF-100, and human Artemis. The metallo-b-lactamase domain is indicated in cyan and magenta, and the b-CASP domain in yellow. The conserved sequence motifs are shown and labeled. (***b***). Amino acid sequence alignment of human CPSF-73, yeast CPSF-100, and human Artemis. The secondary structure elements are indicated (S.S.). Residues in the metallo-b-lactamase domain are indicated by the bars in cyan and magenta, and those in the β-CASP domain in yellow. Motifs 1-4 of the metallo-b-lactamase fold are highlighted in red, and motifs A-C in blue. Residues in CPSF-100 that are located within 3 Å of their equivalents in CPSF-73 are shown in green. Residues in italic are missing in the atomic models. The gray segment in CPSF-100 is highly hydrophilic and is missing from the current structure.

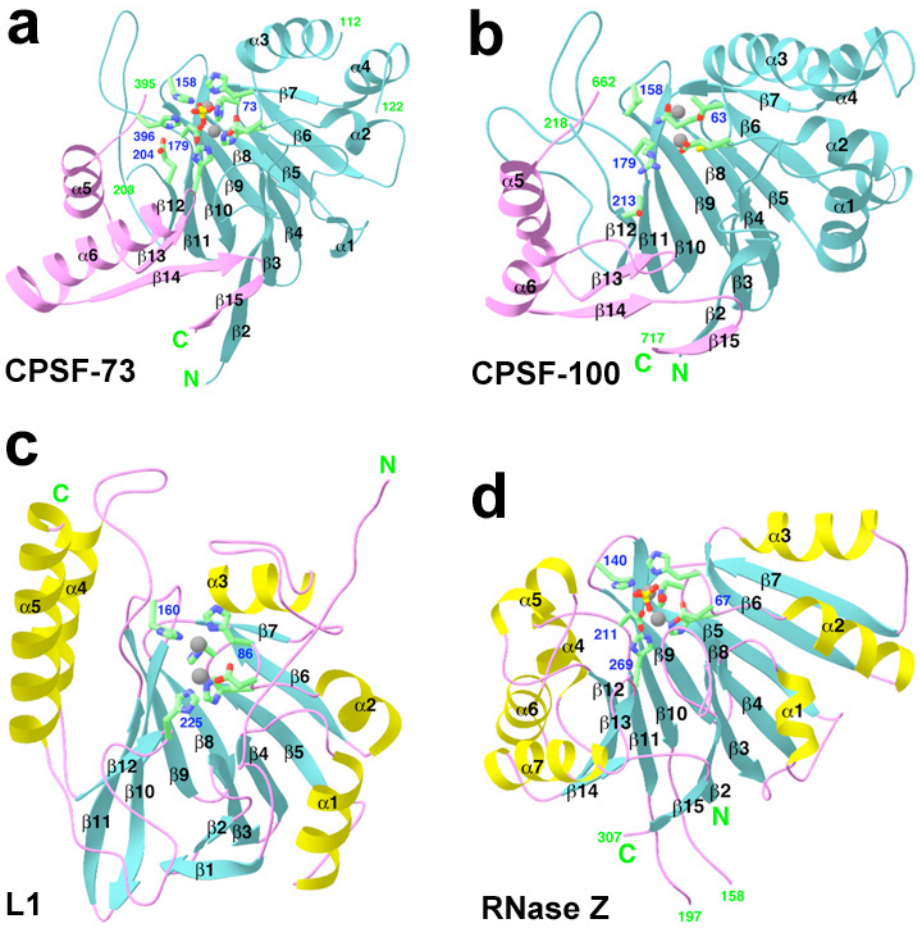

Fig. 2. Schematic drawing of the structures of **(a).** the metallo-β-lactamase domain of CPSF-73. **(b).** the metallo-β-lactamase domain of CPSF-100. **(c).** the L1 metallo-β-lactamase. **(d).** RNase Z.

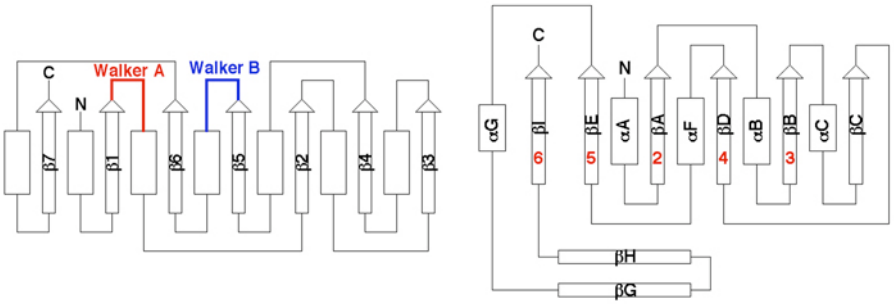

Fig. 3. Topology of the β-CASP domain and the NBF. (***left***). Topology diagram of the nucleotide binding fold (NBF). The locations of the Walker A and Walker B sequence motifs are indicated. (***right***). Topology diagram of the β-CASP domain of CPSF-73. The equivalent β-strands in the nucleotide binding fold are indicated in red.

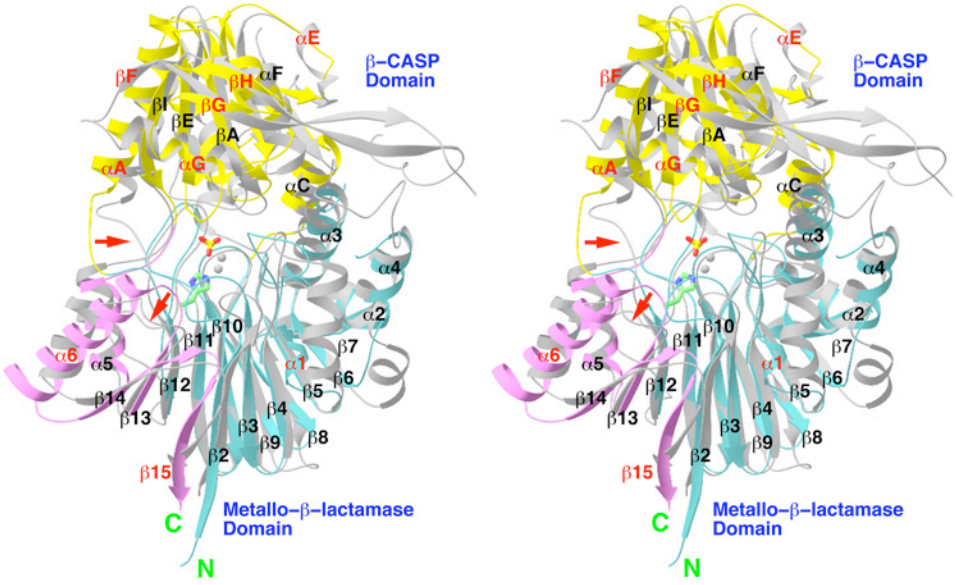

Fig. 4. Comparison of the structures of human CPSF-73 and yeast CPSF-100. The CPSF-100 structure is shown in gray. Regions of large structural differences are labeled in red. The red arrows point to two loops with large conformational differences between the two structures. Produced with Ribbons [16].

```
401  FLCDNYISID TIKEEPLSKE ETEAFKVQLK EKKRDRNKKI LLVKRESKKL  450
451  ANGNAIIDDT NGERAMRNQD ILVENVNGVP PIDHIMGGDE DDDEEEENDN  500
501  LLNLLKDNSE KSAAKKNTEV PVDIIIQPSA ASKHKMFPFN PAKIKKDDYG  550
551  TVVDFTMFLP DDSDNVNQNS RKRPLKDGAK TTSPVNEEDN KNEEEDGYNM  600
601  SDPISKRSKH RASRYSGFSG TGEAENFDNL DYLKIDKTLS KRTISTVNVQ  650
```

Fig. 5. Amino acid sequence for residues 401-650 of yeast CPSF-100 (Ydh1). Negatively charged residues (Asp and Glu) are shown in red, positively charged residues (Lys and Arg) in blue, and hydrophobic residues (Ser, Thr, Asn, Gln, His) in magenta.

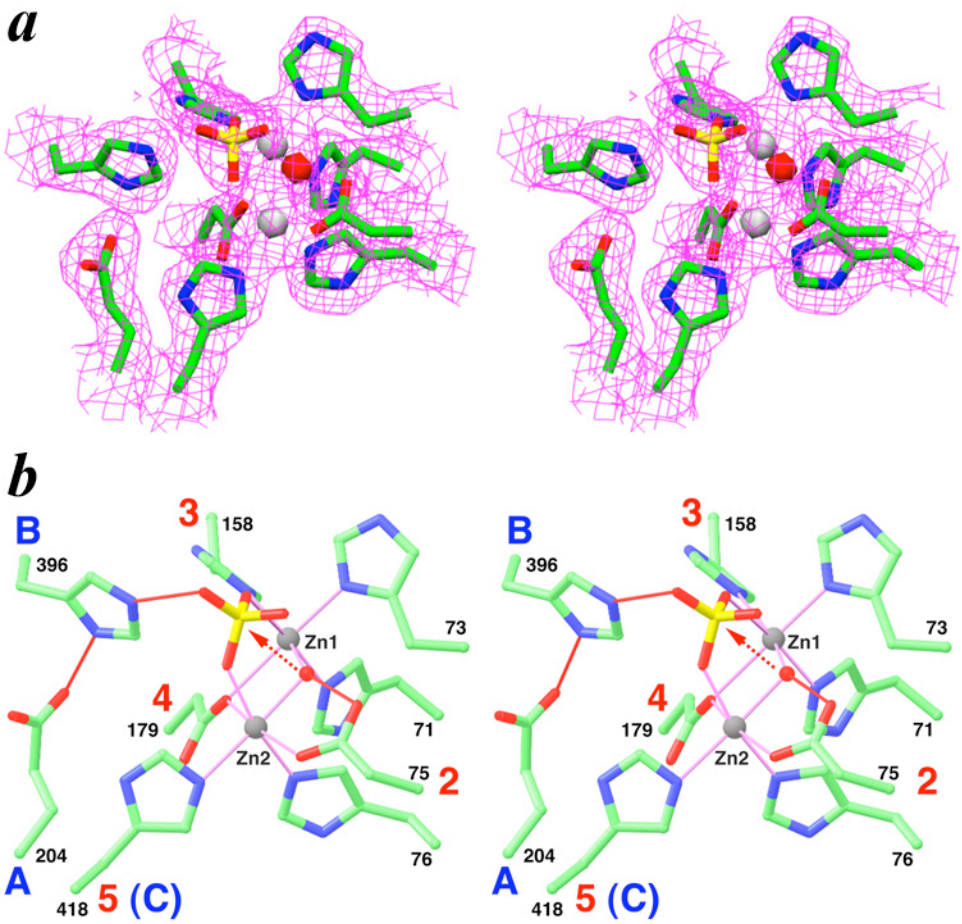

Fig. 6. (*a*). Final $2F_o–F_c$ electron density, at 2.1 Å resolution, for the zinc atoms (gray spheres), their ligands, sulfate ion, and the bridging hydroxide ion (red sphere) in the structure of CPSF-73. Produced with Setor [17]. **(*b*).** Schematic drawing in stereo of the zinc binding modes in the active site of CPSF-73. Produced with Ribbons [16].

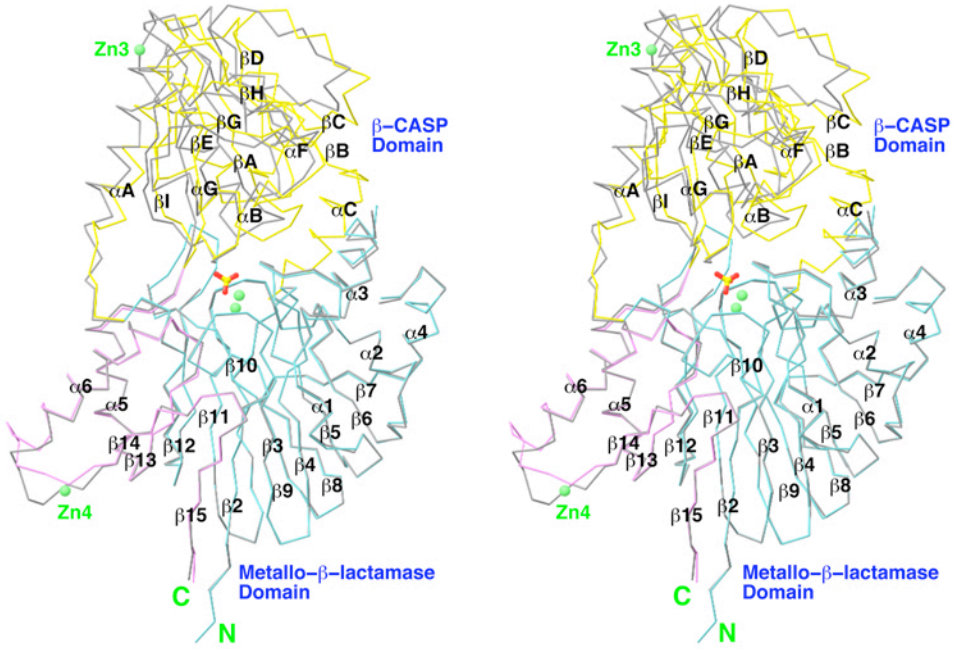

Fig. 7. Comparison of the structures of CPSF-73 grown in the presence (colored Cα trace, zinc ions in green) or absence (gray trace) of zinc ions in the crystallization solution. In the presence of zinc, two additional binding sites are observed, on the surface of the protein and far from the active site. Conformational differences are observed near these additional zinc binding sites between the two structures. Produced with Ribbons [16].

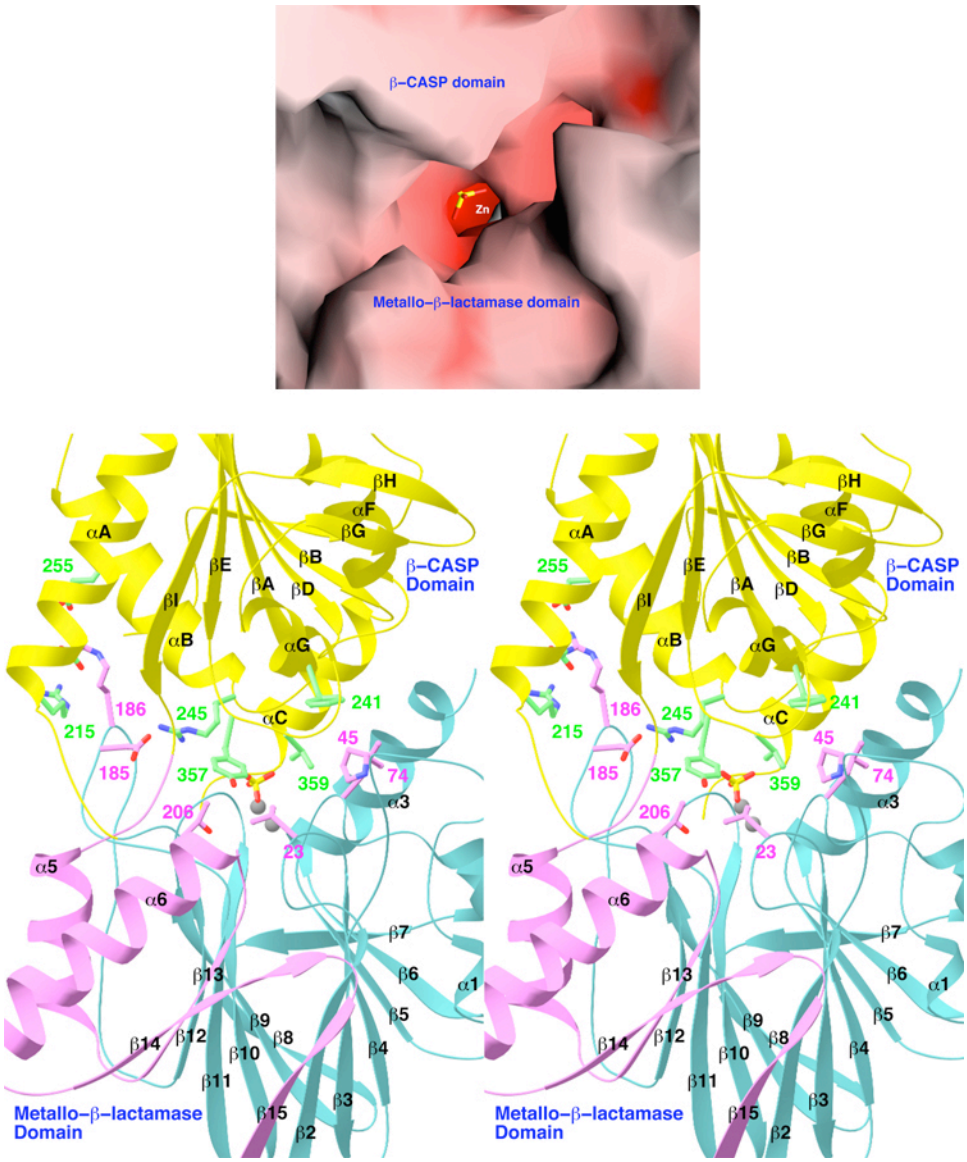

Fig. 8. (***Top***). Molecular surface of human CPSF-73 in the active site region. The sulfate group is shown as a stick model. The zinc ions are shown as gray spheres and labeled. Produced with Grasp [18]. **(*Bottom*).** The interface between metallo-β-lactamase and β-CASP domains. Representative residues in the interface are shown and labeled. The βE-αG loop, containing Tyr357, may be the gate for the active site. The zinc ligands are omitted for clarity. Produced with Ribbons [16].

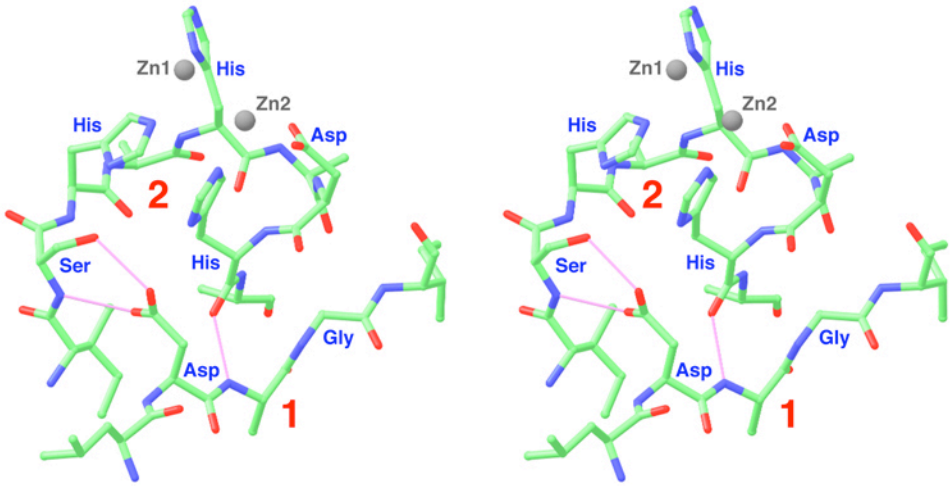

Fig. 9. Motif 1 (with the consensus sequence Asp-X-Gly) helps stabilize the conformation of residues in motif 2, as observed in the structure of CPSF-73. Hydrogen-bonding interactions are indicated by thin lines in magenta. Produced with Ribbons [16].

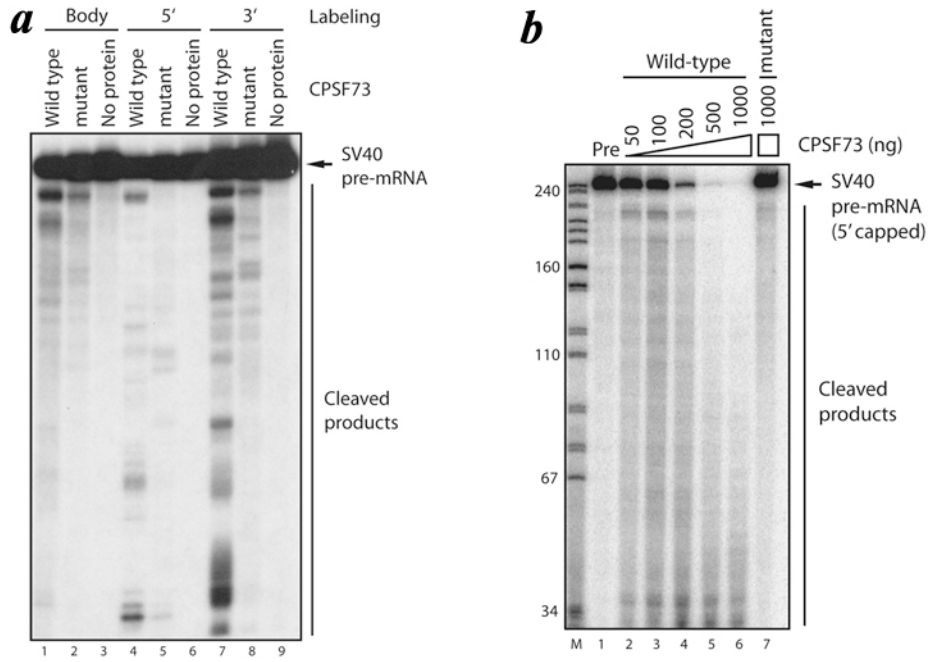

Fig. 10. (*a*). RNA cleavage assay of SV40 late polyadenylation site pre-mRNA. Standard 3'-end processing conditions were used, without preincubation with Ca^{2+}. **(*b*).** RNA cleavage by CPSF-73 is concentration dependent. Cleavage of 5' capped SVL pre-mRNA was performed with increasing amounts (50, 100, 200, 500, 1000 ng) of wild-type (lanes 2-6) or mutant (1000 ng, lane 7) CPSF-73.