The structure of the CstF-77 homodimer provides insights into CstF assembly

Pierre Legrand, Noël Pinaud, Lionel Minvielle-Sébastia and Sébastien Fribourg

INTRODUCTION

mRNA 3’ end maturation is part of a general scheme of pre-mRNA processing comprising 5’-capping and intron-splicing. All these maturation events are essential and tightly coupled and controlled for proper gene expression (1,2). mRNAs poly(A) tails are produced by cleavage and polyadenylation of the pre-mRNA molecule (3). Occurring co-transcriptionally, pre-mRNA 3’-end processing is critical for termination of transcription and mRNA export (1). As opposed to the striking divergence in intron-splicing between mammals and metazoa, cleavage and polyadenylation, the protein components of the pre-mRNA 3’-end processing complexes are quite well conserved from yeast to mammals. In metazoa, cleavage of the precursor requires the trimeric complex cleavage stimulation factor (CstF) and the cleavage and polyadenylation specificity factor CPSF. Both of them are crucial to identify during a preliminary step the precise sequence elements on the precursor where cleavage, and hence polyadenylation thereafter, would occur (4). Additional factors are then recruited, CF I m, CF II m, and the poly(A) polymerase PAP, to stabilize the initial interaction and trigger the processing. A network of physical interactions between subunits of the 3’-end processing machinery and the transcription apparatus has been partially drawn that could explain to some extent how processing, transcription termination and export can be regulated. Many interactions between the pre-mRNA 3’ end processing factors have been reported for the human, Drosophila and yeast systems.

CstF is a multimeric complex essential for the reaction to occur. In human and Drosophila, CstF is formed of CstF-50, CstF-64 and CstF-77 (4). The more likely yeast counterparts are respectively, Pfs2p, Rna15p and Rna14p (5,6). CstF-50 exhibits characteristic WD repeats which are involved in the assembly of multi-protein factors (7). CstF-64 bears an RRM-type RNA-binding domain required for the recognition of U/GU-rich elements located downstream of the poly(A) site (8–10). It plays a key role in the choice of the cleavage site and hence, in the efficiency of the reaction (11,12). CstF-77 is critical for the assembly of the complex, bridging both CstF64 and CstF-50. It is the prototypical Half-a-TPR-containing (HAT) protein as defined by Preker and Keller (13). Moreover, CstF-77 is located at the crossroads in the network of interactions with other 3’-end formation factors such as CPSF and CF II m. It connects CstF to CPSF-160 and hPcf11 (14). Mutations in Rna15p and Rna14p not only impair formation of the mRNA 3’-ends but also prevent RNA polymerase II to terminate properly (15,16). Export of the imperfect transcripts is affected and, as a consequence, they are subsequently degraded (17–21). A growing number of structural studies have shed light on how the catalytic reactions and the regulation may occur in this complex biological machinery (22–27). The structure of protein interacting domains and protein–RNA complexes have been also reported (26,28–30). Many basic questions are still open such as the exact subunit composition of some specific complexes. In this study, we report the crystal structure of the core subunit.
of the CstF complex at 2.0-Å resolution. CstF-77 is built around 11 HAT repeats that self-assemble to form a tight homodimer. The complex has an overall V-shape with large dimensions. Apart from the conserved dimerization interface, several other phylogenetically conserved areas appear at the surface of the complex that may well represent platforms for the association with other protein partners. Mapping experiments performed with the yeast orthologues of CstF-77 and CstF-64 allows the identification of the docking domain of Rna15p onto Rna14p.

**MATERIAL AND METHODS**

**Protein expression, purification and crystallization**

The full-length CstF-77 protein of *Encephalitozoon cuniculi* (31) was cloned into a modified pET-15b overexpression plasmid allowing the production of an N-terminally His-tagged fusion protein (32). Purification was carried out after cell lysis by centrifugation at 4 °C for 1 h at 50 000g. The supernatant was incubated in batch with an affinity resin (Talon) and the eluate was loaded on a Hi-Q-Sepharose (Pharmacia). The protein was concentrated to 30 mg/ml in 25 mM Tris-HCl pH 7.5 and 100 mM NaCl. Crystallization of the sample was carried out at room temperature using sitting-drop vapour diffusion by mixing 1 volume of protein solution with 1 volume of 10% PEG 2000 MME, 100 mM Tris–HCl pH 8.0 and 70 mM calcium acetate of reservoir solution (Nextal). Crystals were directly cryoprotected in a solution of 25% Methyl-2 Pentane-Diol, 10% PEG 2000% MME, 100 mM Tris–HCl pH 8.0 and 70 mM calcium acetate and flash-frozen in liquid nitrogen for data collection. Data were processed with XDS (33). Data collection and phasing statistics are shown in Table 1.

**Structure solution**

The crystal structure of full-length *E. cuniculi* CstF-77 (1–493) was solved to 2.55 Å using phases determined from a SAD (single anomalous dispersion) dataset on a crystal grown by macroseeding with selenomethionine-substituted protein. Thirty-six Se-sites were located using SHELXD (34) and phases were calculated with SHARP (35). An initial model was automatically built using Arp/Warp (36). This initial model was used as a template for molecular replacement against the best native dataset. The model was improved by manual docking of residues and missing portions of the molecules with Coot (37). Model refinement was achieved with REFMAC5 (38). The final model was refined to a resolution of 2.0 Å with a working and free R-values of 27.9 and 25.5%, respectively, and good stereochemistry (Table 1). Strikingly, the final model contains two monomers arranged into a non-crystallographic homodimer, in which short stretches of residues at the N- and C-terminal ends are missing (Figure 1). The final model consists of residues 12 to 465 with the exception of three short loops (62–65, 271–280 and 427–429). Chain B is less defined and consists of residues 12 to 454 with the exception of residues (60–68), (92–111), (131–149), (271–280) and (426–429).

**Surface conservation calculation**

Surface conservation has been calculated with Consurf server with a sequence alignment including *Homo sapiens, Encephalitozoon cuniculi, Drosophila melanogaster, Xenopus laevis, Caenorhabditis elegans, Saccharomyces cerevisiae* and *Arabidopsis thaliana* sequences (39).

**Pull-down assays**

The Rna14p constructs (1–677), (1–593) and (589–677) were amplified by PCR from yeast genomic DNA and cloned into the Ndel and BamHI site of a modified pET-15b vector allowing expression of a protein fused to a His-tag at its N-terminus. The full-length Rna15p protein was amplified from the yeast genome and cloned into the Ndel–XhoI a modified pET-28b vector. Co-expression assays were carried out by co-transformation of Rosetta cells. Cells were grown up to an OD600 of 0.6 and cooled down to 15°C. Overexpression was induced by an overnight incubation with 1 mM IPTG. Cells were harvested by centrifugation and sonicated. A crude extract sample was saved at this point and boiled in Laemmli buffer (T, total extract). After 10 min centrifugation at 4°C, 13 000 r.p.m., the supernatant was incubated for 30 min with His-tag affinity resin and washed three times with 50 mM Tris–HCl pH 7.5, 150 mM NaCl, 0.1% Triton X-100. The resin was boiled in Laemmli buffer and the samples were resolved by SDS–PAGE (B, bound). The proteins were transferred on a blot and analysed with polyclonal antibodies directed to Rna15p and Rna14p. Monoclonal antibodies were used to reveal His-tag fused proteins (Amersham, GE Healthcare).

---

### Table 1. Crystallographic data phasing and refinement statistics

<table>
<thead>
<tr>
<th></th>
<th>Native 1</th>
<th>Se-Met (peak)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space group</td>
<td>P2₁</td>
<td>P2₁</td>
</tr>
<tr>
<td>Cell dimensions</td>
<td>a, b, c (Å)</td>
<td>44.5 148.4 91.0</td>
</tr>
<tr>
<td>a, b, c (°)</td>
<td>90.0 92.2 90.0</td>
<td>90.0 91.0 90.0</td>
</tr>
<tr>
<td>X-ray source</td>
<td>ESRF-ID14-2</td>
<td>ESRF-BM-14</td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.9340</td>
<td>0.978 (peak)</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>20–2.0</td>
<td>25–2.55</td>
</tr>
<tr>
<td>Total measurements</td>
<td>291 014 (44 787)</td>
<td>290 548 (44 218)</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>77 953 (12 258)</td>
<td>74 414 (11 392)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>3.7 (3.6)</td>
<td>3.9 (3.8)</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>97.8 (95.1)</td>
<td>97.6 (95.0)</td>
</tr>
<tr>
<td>Rsym (%)</td>
<td>7.7 (38.6)</td>
<td>7.0 (34.3)</td>
</tr>
<tr>
<td>I/σ</td>
<td>15.17 (3.69)</td>
<td>17.8 (4.2)</td>
</tr>
<tr>
<td>FOM (acentric)</td>
<td>0.34</td>
<td></td>
</tr>
</tbody>
</table>

*Statistics for high resolution bin (2.0–2.12 Å) are indicated in parentheses.

†High resolution bin (2.55–2.7 Å) are indicated in parentheses.
RESULTS AND DISCUSSION

CstF-77 assembles into a homodimer

The structure of CstF-77 is entirely \( \alpha \)-helical and consists of 23 \( \alpha \)-helices arranged in pairs of anti-parallel \( \alpha \)-helices forming 11 HAT repeats as described by Preker and Keller (13). It can be divided into two domains, an N-terminal domain containing the first 4 HAT repeats (residues 12 to 151), and a middle domain containing HAT repeats 5 to 12 plus the C-terminal \( \alpha \)-helices (residues 162 to 465). Residues 466 to 493 are likely to form an independent domain not seen in our electron density map. Helix 8 provides \( \alpha \)-helix B of HAT repeat 4 and \( \alpha \)-helix A of HAT repeat 5. It links both domains forming a 145° kink (Figure 1a and Figure 3). The dimer has an overall V-shape with dimensions of 140-Å wide and 60-Å thick, each arm of the V measuring ~80-Å long (Figure 1b). The 110° angle between both arms is in good agreement with the one measured from electron microscopy pictures obtained with the yeast Rna14p–Rna15p corresponding complex and with the angle measured for the murine CstF-77 complex (40,41). The two CstF-77 monomers are oriented tail-to-tail and interact extensively through their middle domain to form a tight homodimer burying 4000 Å² of the surface area (Figure 1a and c). The interface between the monomers is provided by the C-terminal \( \alpha \)-helix (\( \alpha \)-helix 23) of each monomer interacting with HAT-repeat 11 on the one hand. On the other hand, the interaction is built up by HAT-repeats 9 to 11 from one monomer interacting with HAT-repeats 11 to 9 of the opposite monomer and shielded by a well-organized network of water molecules. Superimposition with the murine CstF-77 orthologue HAT-N and HAT-C domains shows limited differences between the two models (Figure 2). Three extra \( \alpha \)-helices defining 1.5 HAT repeat at the N-terminus are observed in the murine CstF-77 HAT-N domain in comparison to that of E. cuniculi (Figures 2a and 3)(41). Interestingly, the two last \( \alpha \)-helices observed in the murine and the E. cuniculi orthologues have similar orientations but structurally equivalent helices belong to opposite monomers (Figure 2b). In the murine CstF-77, these helices follow the curve formed by the HAT repeats whereas, in the E. cuniculi orthologue, the equivalent helices cross the concave surface defined by HAT repeats 8 to 11 to interact with HAT repeats 6 to 8 (Figure 1). In contrast to murine CstF-77, the prominent pocket observed on the concave surface of the homodimer is likely to be occluded by residues 426 to 430 that could not be placed in our model. Whether this reflects species-specific characteristics has to be tested.

Homodimerization of the Drosophila CstF-77 homologue, the Su(f) protein, has been proposed to account for the genetic complementation of lethal alleles of the \( su(f) \) gene with different domains of the Su(f) protein (42). Similarly, homodimerization of human CstF-77, as well as human CstF-50, were detected by \textit{in vitro} analysis in a CstF mapping study (43). In yeast, ultracentrifugation analyses of CF IA subunits Rna14p and Rna15p demonstrated a 1:1 stoichiometry of the complex. In addition, these data suggested that the Rna14p–Rna15p heterodimer self-associates via the Rna14p subunit to form a heterotetramer (40). Our model of CstF-77 provides the structural basis for the homodimerization of the protein and its conservation through evolution. Taken together,
these data strongly support the idea that CstF functions as a complex comprising two copies of each of its subunits.

**CstF-77 homodimer exposes conserved surface**

Orthologues of pre-mRNA 3'-end processing factors have been characterized from yeast to human. Functional complementation between Drosophila and human CstF-77 has been demonstrated, with the exception of the C-terminal domain (44). Therefore, the determinants of interaction are likely to be conserved as well. We performed sequence alignment for seven different CstF-77 homologues (Figure 4). Apart from the dimerization interface, two areas at the surface of the complex cluster a number of conserved residues (Figure 4a and b). The first area is located on HAT-repeats 9 to 11 and consists of charged residues (Figure 4a). The tail-to-tail orientation of the two monomers brings into close vicinity the equivalent areas of each molecule leading to a potentially unique extended and conserved surface. The second highly conserved area of the homodimer is located in the N-terminal domain (Figure 4b). The conserved residues of HAT-repeats 1 to 4 are located on the external portion of the dimer. Tyr53, Val57, Val70 and Phe73 cluster into this region and form a hydrophobic patch. Due to the V-shape of the complex and the location of the various conserved regions, the CstF-77 homodimer exposes four highly conserved areas provided by the neighbouring and equivalent areas of HAT repeats 9–11, and by the two independent N-terminal domains of the complex (Figure 4).

**Rna14p C-terminus mediates Rna15p interaction**

CstF-77 and its orthologues in yeast and Drosophila are central for CstF complex formation and for the interaction with CPSF. Indeed, Rna14p forms a tight complex with Rna15p and interacts with Pfs2p, Pcf11p and Nab4p/Hrp1p (5,6,40,45). The conserved exposed areas of CstF-77 are likely to provide platforms for the interaction with protein partners. On the basis of our crystal structure, we tested this assumption in pull-down experiments with various deletion constructs of Rna14p co-expressed with Rna15p in Escherichia coli (Figure 5a and b). As expected, full-length His-Rna14p could efficiently pull down Rna15p (Figure 5a, lane 4). However, deletion of the Rna14p C-terminal domain (residues 593 to 677) resulted in the loss of interaction with Rna15p (Figure 5a, lane 6). Co-precipitation of Rna15p with Rna14p C-terminal domain confirmed that this portion of the molecule is sufficient to establish an interaction between the two polypeptides (Figure 5a, lane 8). This domain is important for the function of Rna14p since the shortening of the protein by 16 amino acids at its C-terminus (stop codon at residue 633) observed in the yeast rna14-1 mutant, leads to a defect in 3' end pre-mRNA processing (46). Altogether, these data suggest that alteration of the interaction between Rna14p and Rna15p is the molecular basis for the loss-of-function phenotype observed in rna14-1 mutant strain.

In metazoans, similar interaction have been described for CstF-77 and CstF-64 (41,43). Interestingly, in Drosophila, a single mutation or insertion within the su(f) protein (su(f)R-9-18) impairs its function (42). The C-terminal portion of CstF-77 is highly conserved in metazoans but differs notably from the one in yeast (Figure 3). In summary, these data suggest that the C-terminal Pro-rich portion of CstF-77 and its homologues carries a similar function, even though it is not strictly conserved in sequence from yeast to human. Further analysis is required in order to determine whether this domain has a similar structure in yeast and metazoans.

**ACKNOWLEDGEMENTS**

We are grateful to Pr Vivarès, Université Blaise Pascal, Clermond-Ferrand for providing E. cuniculi GB-M1 genome. We would like to thank the staff of ESRF beamlines BM14 and ID14-2. We would like to thank A. Thompson and W. Shepard for critical reading of the manuscript. This work is supported by grants from the INSERM (Avenir program) and the ‘Association pour la Recherche contre le Cancer’ (to S.F.) and from the CNRS.
Figure 3. Sequence alignment CstF-77 homologues. The polypeptide sequence of CstF-77 from *E. cuniculi*, *H. sapiens*, *D. melanogaster*, *C. elegans*, *X. laevis*, *A. thalina* and *S. cerevisiae* were aligned with PipeAlign (47). Residues are coloured in red when >75% identity is reached. Secondary structures are shown for each monomer on top of the sequences. Dotted lines correspond to sequences for which no interpretable electron density was seen. HAT repeats are boxed in beige. The location of the mutated amino acid in *rna14-1*, *rna14-5* and *Drosophila Su(f)R-9-18* mutant are boxed in red.

and ‘La Fondation pour la Recherche Médicale/ Fondation BNP-Paribas’ (to L.M.-S.). Funding to pay the Open Access publication charges for this article was provided by Avenir program.

**Coordinates**

The atomic coordinates and structure factor have been deposited in the Protein Data Bank (accession code 2UY1).
Conflict of interest statement. None declared.

REFERENCES


Figure 4. CstF-77 homodimer exposes conserved surfaces for interaction with partners. (a) and (b) Residue conservation has been calculated with Consurf from a sequence alignment including CstF-77 homologues. Conservation ranges from white for non-conserved residues, to red for absolutely conserved residues. The two orientations of the complex are related by a 90° rotation in the y- and z-axis. The border of the dimer interface is shown as a solid grey line on both panels.

Figure 5. Mapping Rna14p–Rna15p interaction. (a) Pull-down experiments were performed between Rna15p and His-tagged Rna14p deletion mutants. Proteins were revealed with polyclonal antibodies to Rna15p and monoclonal antibodies directed against the His-tag. T, total extract; B, bound fraction. (b) Rna14p deletion constructs used in pull-down experiments. ‘+’ and ‘−’ corresponds to interaction and no interaction, respectively.


