

Crystal Structures of the *Pyrococcus abyssi* Sm Core and Its Complex with RNA

COMMON FEATURES OF RNA BINDING IN ARCHAEA AND EUKARYA*

Received for publication, July 30, 2002, and in revised form, October 25, 2002
Published, JBC Papers in Press, October 29, 2002, DOI 10.1074/jbc.M207685200

Stéphane Thorez‡, Claudine Mayer‡§, Claude Sauter¶, Stephen Weeks||, and Dietrich Suck**

From the Structural Biology Program, European Molecular Biology Laboratory, Meyerhofstrasse 1,
69117 Heidelberg, Germany

The Sm proteins are conserved in all three domains of life and are always associated with U-rich RNA sequences. Their proposed function is to mediate RNA-RNA interactions. We present here the crystal structures of *Pyrococcus abyssi* Sm protein (PA-Sm1) and its complex with a uridine heptamer. The overall structure of the protein complex, a heptameric ring with a central cavity, is similar to that proposed for the eukaryotic Sm core complex and found for other archaeal Sm proteins. RNA molecules bind to the protein at two different sites. They interact specifically inside the ring with three highly conserved residues, defining the uridine-binding pocket. In addition, nucleotides also interact on the surface formed by the N-terminal α -helix as well as a conserved aromatic residue in β -strand 2 of the PA-Sm1 protein. The mutation of this conserved aromatic residue shows the importance of this second site for the discrimination between RNA sequences. Given the high structural homology between archaeal and eukaryotic Sm proteins, the PA-Sm1-RNA complex provides a model for how the small nuclear RNA contacts the Sm proteins in the Sm core. In addition, it suggests how Sm proteins might exert their function as modulators of RNA-RNA interactions.

In addition to sharing a common ancestor with Eukarya according to ribosomal RNA sequence comparisons (1), archaeal genomes encode many components characteristic of eukaryotic ribonucleoprotein particles (RNPs).¹ Indeed, components of processes, specific to Eukarya like box C/D components implicated in ribosomal RNA modifications (2, 3) or splicing factors involved in the removal of introns from the pre-mRNAs (4, 5), are found within those genomes. Moreover, Archaea are

thought to be the precursor of the eukaryotic nucleus (6), which is the final location of all those RNPs.

Sm/Lsm proteins are focusing the attention because of their involvement in several eukaryotic processes such as mRNA splicing (7), histone maturation (8), telomere maintenance (9), or mRNA degradation (10, 11). The function of Sm/Lsm proteins first described in the literature is their role in splicing. Sm/Lsm proteins are involved in the biogenesis of the small nuclear RNPs (snRNPs) U1, U2, U4/U6, and U5, which are essential components of the spliceosome. In the absence of RNA, canonical Sm proteins exist as three independent sub-complexes containing the D₃B, D₁D₂, and EFG proteins, whereas the Lsm proteins are present as preformed heptameric complexes (12). The Sm complexes associate with cytoplasmic snRNAs in a step-specific order, which is regulated by the survival of motor neurons complex (13, 14). Their association onto the snRNA triggers hypermethylation of the N7-methylguanosine cap (15), which then forms part of the nuclear localization signal and allows snRNPs to return to the nucleus where they will perform their function (16, 17). In the case of the Lsm proteins, the target is U6, the only snRNA that is transcribed by polymerase III. Lsm proteins act at two levels on U6 snRNP. First, they facilitate its biogenesis as Sm proteins do with other U snRNAs, and second, they increase the recycling of the di-snRNP U4/U6 after splicing has occurred (18).

Eukaryotic Sm proteins are known to associate with snRNAs onto a sequence called the Sm site. It consists of the consensus sequence PuAU₄₋₆Gpu and is usually flanked by stem-loop structures necessary for proper functional binding (19). Nevertheless, the nonanucleotide AAUUUUUGA is sufficient for the formation of a proper Sm core complex as judged by its biochemical and structural characteristics (20). In case of the Lsm proteins, the interacting site is composed of a stretch of uridines located at the 3' end of U6 snRNA (12, 18, 21). Other Sm-containing complexes are known to interact with sequences similar to the consensus Sm site as in the case of the telomere and the U7 snRNP (22). In case of the Sm complex involved in mRNA degradation, the site has not been defined yet (10, 11).

A conserved bipartite domain of ~70 amino acid residues defines the Sm/Lsm protein family. Up to 30 distinct members compose the human family, whereas in Archaea it is limited to two subtypes called Sm1 or Sm2. Those types have been defined according to the phylogenetic tree obtained from sequence alignment with archaeal Sm protein sequences (5). *Pyrococcus* and *Halobacterium* species are the only Archaea where a single open reading frame is found. This gene encodes an Sm-related protein of type 1, the more abundant subtype in Archaea. Recently, the bacterial protein Hfq has been shown to belong to the Sm family despite the fact that it contains only the first part of the Sm signature sequence (23–25). Finally, the *Meth-*

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 1H64 and 1M8V) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (<http://www.rcsb.org/>).

‡ Both authors contributed equally to the work.

§ Supported by an EMBO long term fellowship. Present address: Université Pierre et Marie Curie, 4 Place Jussieu, 75252 Paris, France.

¶ Recipient of a Marie Curie Individual Fellowship (Improving the Human Potential Program, contract number: HPMF-2000-00434).

|| Present address: LifeSensors Inc., 271 Great Valley Parkway, Malvern, PA 19355.

** To whom correspondence should be addressed. Tel.: 49-6221-387307; Fax: 49-6221-387306; E-mail: suck@embl-heidelberg.de.

¹ The abbreviations used are: RNP, ribonucleoprotein particles; snRNA, small nuclear RNA; snRNP, small nuclear RNP; PA-Sm1, *P. abyssi* Sm protein; AF-Sm1/2, *A. fulgidus* Sm1/2 protein; Lsm, like-Sm.

TABLE I
 X-ray data and refinement statistics

r.m.s.d., root mean square deviation.

	PA-Sm1 native	PA-Sm1/U ₇	
		2.1 Å	2.6 Å
X-ray data			
Wavelength (Å)	0.933	0.915	0.915
Resolution (Å)	30.0–1.9	30.0–2.1	30.0–2.6
Total number of reflections	386,356	279,320	89,628
Number of unique reflections	156,396	74,641	37,042
R _{sym} (%) ^{a,b}	4.9 (14.7)	6.1 (11.8)	5.9 (18.4)
I/σ(I) ^a	27.7 (7.5)	35.0 (15.3)	29.1 (9.0)
Completeness (%) ^a	96.2 (95.2)	95.9 (80.1)	90.1 (57.1)
Refinement			
R _{crystal} (%)	23.7	22.3	21.0
R _{free} (%)	28.2	26.9	28.3
Number of proteins atoms	15,824	7924	7929
Number of RNA atoms	-	959	1121
Number of solvent atoms	1345	207	205
Metal ions		7 Ca ²⁺	7 Ca ²⁺
Average B value-all	38.1	30.7	38.7
Average B value-all protein	37.0	27.7	33.6
Average B value-all RNA		50.9	70.3
Average B value-all -solvent	51.2	39.6	40.9
r.m.s.d bond distance (Å)	0.007	0.008	0.013
r.m.s.d bond angle (°)	1.35	1.78	1.65
r.m.s.d dihedrals	25.0	25.4	25.1
r.m.s.d improper	0.78	1.83	1.26

^a Values in parentheses refer to statistics in the highest resolution shell (2.0–1.9, 2.21–2.10, and 2.74–2.6 for the native and 2.1 and 2.6 Å data sets, respectively).

^b $R_{\text{sym}} = \sum_{\text{hkl}} |I - \langle I \rangle| / \sum_{\text{hkl}} I$, where I is the observed intensity for a reflection of index hkl and $\langle I \rangle$ is the mean intensity.

anococcus jannashii genome does not contain any open reading frame encoding an Sm protein but does contain a protein related to the bacterial Hfq.

Crystal structures of several eukaryotic (26) and archaeal (27–29) Sm proteins have shown that the amino acid signature sequence defines a specific Sm-fold composed of a strongly bent 5-stranded β-sheet preceded by a short α-helix. The general architecture observed for the archaeal and predicted for the eukaryotic Sm complexes is a heptameric ring structure containing a highly positively charged cavity where RNA molecules bind. In the *Archaeoglobus fulgidus* Sm1 protein (AF-Sm1), three residues, namely His-37 and Asn-39 from the first part and Arg-63 from the second part of the Sm domain, form a uridine-specific binding pocket (29). These residues are highly conserved within the Sm/Lsm family, and UV cross-linking experiments show that Phe-37 of the human SmG protein (equivalent to His-37 in AF-Sm1) contacts the first uridine of the Sm site (30), suggesting that the RNA binding site has been conserved between Archaea and Eukaryotes. The bacterial Sm protein Hfq also exhibits RNA binding capacity. It binds oligo(A) (31) as well as uridine-rich sequences (25, 32), even though the second part of the Sm domain, which is involved in RNA binding in archaeal and eukaryotic Sm proteins, is not conserved.

Here we report the crystal structures of the *Pyrococcus abyssi* Sm protein (PA-Sm1) and its complex with a seven-nucleotide long oligo(U). The free protein forms heptameric rings, which associate to form dimer of heptamers in the crystal as well as in solution. The binding of RNA disrupts the dimer of heptamers but leads to only minor structural changes within the heptameric structure, indicating that the PA-Sm1 heptamer represents a rigid preformed RNA binding unit. Nucleotides contact the protein at two different sites: inside the internal cavity in the uridine-specific binding pocket and on the surface of the heptamer close to the N-terminal α-helix. The second site is shown to be important for the association of PA-Sm1 with non-symmetrical RNA. Our structure has allowed us to construct a model of the complex composed of the eukaryotic Sm proteins and a short sequence of the U1 snRNA.

The model suggests that the Sm core probably represents a platform for interactions between pre-mRNA and snRNA. The model may also help to understand the regulatory function of the individual Sm protein subunits within the Sm core because the binding to the U-rich sequence composing the Sm binding site can be distinguished from the binding to surrounding sequences.

EXPERIMENTAL PROCEDURES

Protein Purification

The *P. abyssi* Sm gene (PAB8160) was amplified by PCR from genomic DNA (Génoscope, Evry, France) and cloned into the expression vector pET24d (Novagen) or in a modified pET24 vector with an upstream sequence coding for a His₆ tag followed by a tobacco etch virus protease site. Overexpression of the protein was carried out in the *Escherichia coli* strain BL21-CodonPlus (DE3)-RIL (Stratagene). Cells were grown at 37 °C in LB medium supplemented with 100 μg/ml kanamycin for 3 h ($A_{600} \sim 0.6$), and the induction was triggered by adding 1 mM isopropyl-1-thio-β-D-galactopyranoside for 3 h. Cells were harvested by centrifugation. Bacteria were lysed using a French press in buffer A (50 mM NaH₂PO₄ (pH 7.5), 200 mM NaCl, 10 mM β-mercaptoethanol) in the presence of protease inhibitor mixture (Promega) and 10 μg/ml ribonuclease A (Sigma). His₆-PA-Sm1 was first separated from thermolabile host proteins by a heat shock at a temperature between 85 and 95 °C for 10 min and centrifugation (14,000 rpm, 30 min, 4 °C). The supernatant was loaded onto nickel-nitrilotriacetic acid beads (Qiagen), and the elution was carried out as recommended by the manufacturer. The pool containing PA-Sm1 was fractionated on a Superdex 75 gel-filtration column (Amersham Biosciences) in buffer A. Fractions containing the proteins were concentrated by ultrafiltration. Tobacco etch virus protease cleavage reaction was carried out overnight at 16 °C at an enzyme to substrate ratio of 1/50. Another gel-filtration step under the same conditions led to a >95% pure sample as judged from Coomassie Blue-stained SDS-page gel (data not shown). PA-Sm1 was concentrated to 18 mg/ml and stored at –20 °C in 20 mM Tris/Cl (pH 7.5), 200 mM NaCl. The uridine heptamer was purchased from Xeragon. The non-fused protein was purified in a similar way replacing the nickel column by ion-exchange chromatography (Resource Q, AP Biotech).

Gel Shift Assays

Gel retard experiments were done as described previously (29). Radioactively labeled RNA (~50 nM) was used in all assays. Protein concentrations varied between 11 and 55 μM. The mutation of tyrosine

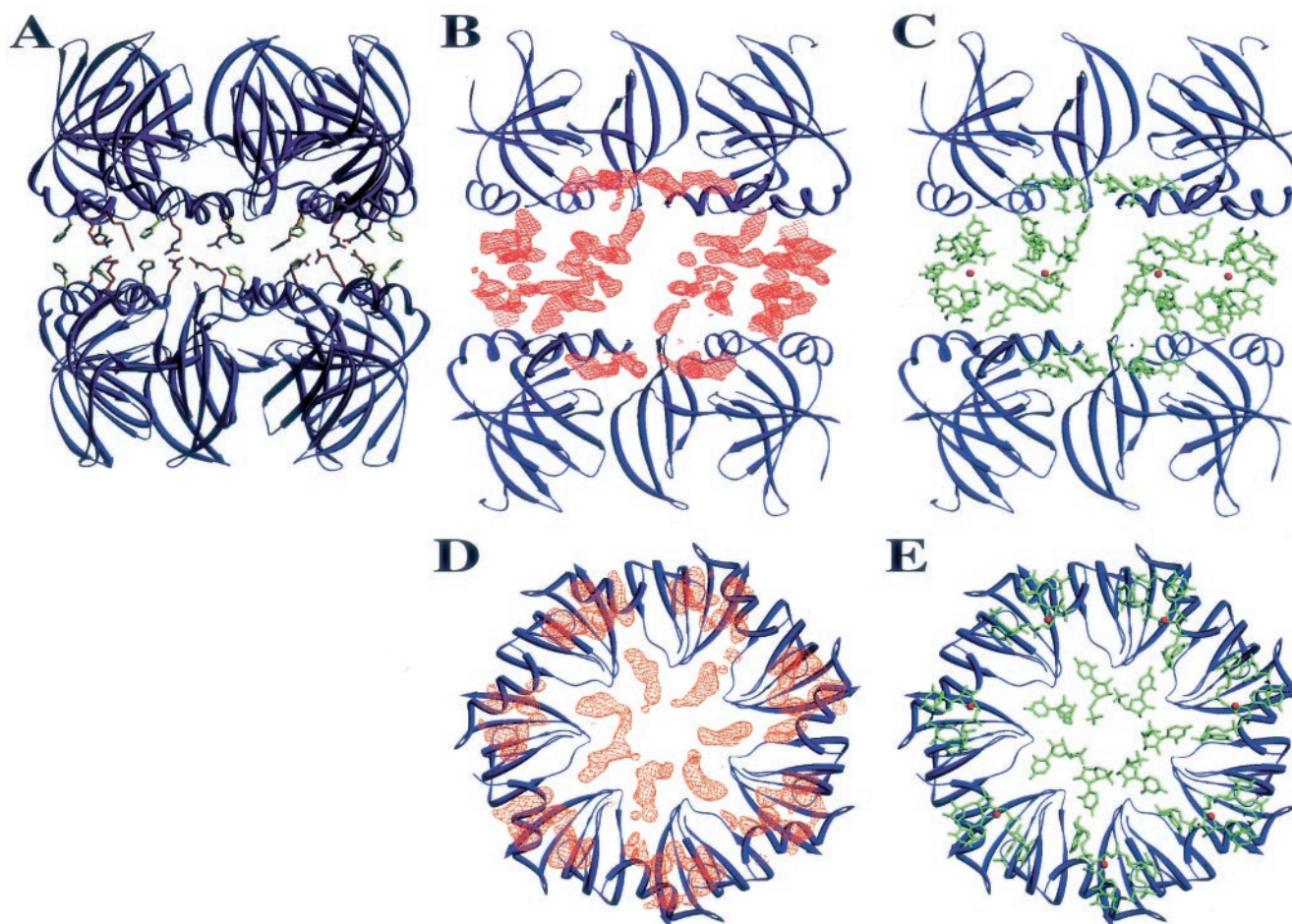


FIG. 1. PA-Sm1 forms a heptameric ring structure, free and in complex with RNA. A, heptamer-heptamer contacts observed in the crystal structure of the free protein. Heptamers interact via the Arg-4 and His-10 residues (highlighted in orange and yellow, respectively) from the N-terminal α -helix. B and D, PA-Sm1-U₇ $F_o - F_c$ difference density map calculated using two PA-Sm1 heptamers and contoured at 2.6 σ . Densities corresponding to the bound RNA are located between the two rings and within the central cavity. C and E, overall PA-Sm1-U₇ structure. RNA molecules bound to the external sites of the subunits connect the two rings, whereas at the internal uridine-binding pockets, only isolated nucleotides are visible. The calcium ions stabilizing the phosphate groups of the nucleotides in between the two rings are shown in red, protein molecules in ribbon representation are in blue, and RNA molecules are shown in green. Figs. 1–3 were prepared with the programs Setor and Ribbons (54, 55).

34 to valine was introduced using the QuikChange Mutagenesis Kit from Stratagene and according to the manufacturer's instructions. The oligonucleotides used were as follows: complementary strand, 5'-GGG-CAGGCTCATTGGAGTAGACATTCACCTGAATGTCG-3', and reverse strand, 5'-CGACATTCAGGTGAATGTCTACTCCAATGAGCCTGCC-3'. The vector was subsequently sequenced to confirm the point mutation. The same procedure was used to purify the mutant protein.

Crystallization and X-ray Data Collection

Crystals of PA-Sm1 alone were grown by vapor diffusion (hanging drops) from reservoirs containing 27–29% 2-methyl-2,4-methanediol, 150 mM magnesium acetate, 50 mM sodium cacodylate (pH 6.5) at 22 °C. Crystals were flash-frozen directly from the crystallization drops. Two native diffraction data sets at 2.6 and 1.9 Å were collected using an in-house source and the beamline ID14-2 at the European Synchrotron Radiation Facility (Grenoble, France), respectively. Data processing was done with XDS (33). The two data sets correspond to two crystal forms belonging to the space group P1 with cell parameters $a = 69.3$ Å, $b = 67.6$ Å, $c = 67.9$ Å, $\alpha = 117.9^\circ$, $\beta = 92.7^\circ$, and $\gamma = 105.4^\circ$ (crystal form I) and $a = 69.3$ Å, $b = 76.2$ Å, $c = 116.0$ Å, $\alpha = 90.2^\circ$, $\beta = 97.8^\circ$, and $\gamma = 107.5^\circ$ (crystal form II).

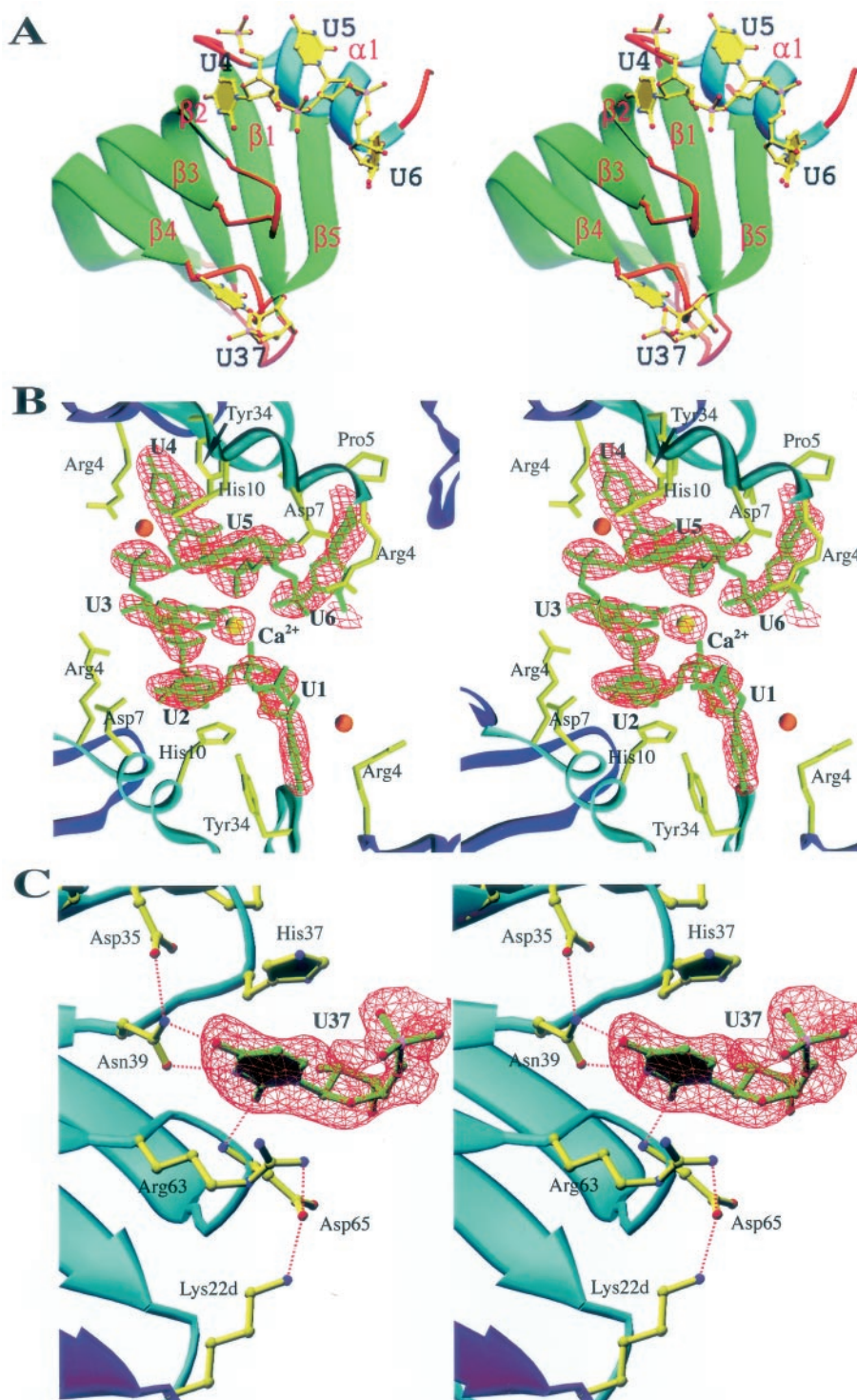
Crystals of PA-Sm1-U₇ were obtained for a monomer/RNA ratio of 1:1 with a protein concentration of 9 mg/ml in 10% polyethylene glycol 1000 molecular weight, 100 mM imidazole (pH 8.0), 250 mM calcium acetate at 4 °C. Drops of 2 μ l were prepared by mixing reservoir and protein/RNA solutions at a 1:1 ratio. They were immediately macroseeded with crystals from previous crystallizations. PA-Sm1-U₇ crystals were flash-frozen in 15% polyethylene glycol 1000, 100 mM imidazole (pH 8.0), 250 mM calcium acetate, and complete data sets were collected at 2.1 and

2.6 Å resolution under cryogenic conditions using the beamline ID29 at ESRF. Data were processed using the HKL package (34). The complex was crystallized in P1 with the following identical cell parameters for the two data sets: $a = 68.0$ Å, $b = 68.0$ Å, $c = 84.7$ Å, $\alpha = 100.0^\circ$, $\beta = 105.0^\circ$, and $\gamma = 110.0^\circ$. X-ray data statistics corresponding to the crystal form II of the native protein and the two data sets of the PA-Sm1-U₇ complex are given in Table I.

Structure Determination and Refinement

PA-Sm1 Structure—The protein structure was solved by molecular replacement using AMoRe (35) from the CCP4 package (36) from self-rotation and self-Patterson functions, revealing the presence of a non-crystallographic 7-fold axis ($\kappa = 52^\circ$) and a perpendicular 2-fold axis. Using the fact that a 7-fold-axis was found in the self-rotation function, a heptameric model was built using the coordinates of the crystallographic hexameric AF-Sm2 model (37).² The low resolution data set (corresponding to crystal form I) was first used to solve and understand the crystal packing of the structure and later was refined with CNS (38). The resulting heptameric model was used to analyze the high resolution data set (crystal form II). Two peaks in the rotation function clearly corresponded to the orientation of the two sets of head-to-head heptamers, and subsequently, the four heptamers could be successively positioned in the translation function. A final "fitting" step led to a correlation coefficient of 70% and an R -factor of 38%. After rigid body refinement, the structure was refined with CNS to an R -factor of 23.7%

² L. Moulinier, personal communication.

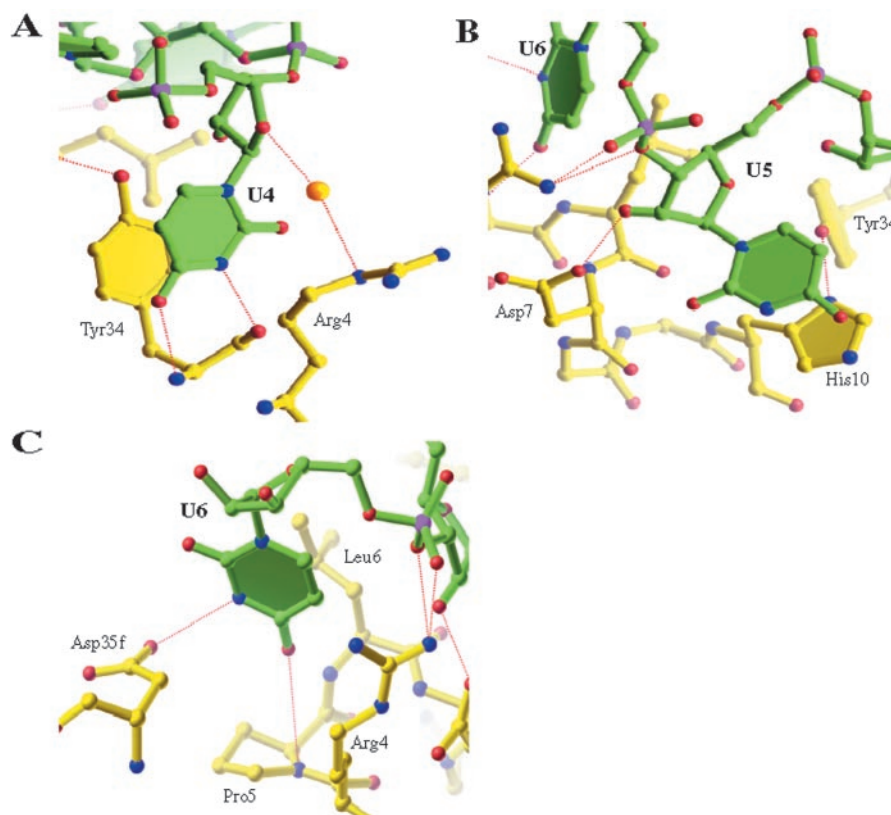


and R_{free} of 28.2% containing 28 copies of the monomer including residues 3–73 and 1345 water molecules.

PA-Sm1-U₇ Complex—The structure of a refined heptamer of PA-Sm1 was used as the search model for molecular replacement using AMoRe. Similarly, to the free protein, a 7-fold and a perpendicular 2-fold axis were observed in the self-rotation function of the PA-Sm1-U₇ complex. A solution consisting of 2 heptamers/unit cell was found with a correlation and an R -factor of 67 and 37%, respectively, and was refined using CNS. The electron density map calculated at this stage showed strong density in the solvent region between the two rings. The program COMA (39) was used to compute a correlation map taking into account the non-crystallographic symmetry operators. This clearly revealed the presence of RNA molecules that connected monomers from the two rings. It also facilitated the definition of an improved mask for

non-crystallographic symmetry averaging and bulk solvent correction. The model was adjusted by several cycles of model building using program O (40) followed by coordinate minimization and B -factor refinement. RNA molecules were introduced when the protein model was almost satisfying. Connections between nucleotides were built when they were clear in the density and stereochemically possible. The two independent data sets (Table I) revealed the same organization for the RNA strands connecting the protein rings. Nevertheless, the 2.6-Å data set was used in the last positional and B -factor refinement steps because it provided a much better definition for the bases interacting inside the central cavity of the rings. The final model at 2.6-Å resolution consists of 14 PA-Sm1 monomers, 205 water molecules, 7 calcium ions, and a total of 55 nucleotides, leading to an R -factor of 21.0% and an R_{free} of 28.3%.

FIG. 3. The RNA nucleotides composing the external binding site. *A*, binding site for U4 (identical for U1). The nucleotide position and conformation are well defined. Hydrogen bonds are indicated by red dotted lines. Stacking interactions with Tyr-34 stabilizes the uracil base. *B*, binding site for U5 (and U2). His-10 is kept in a stacking orientation by a hydrogen bond with Tyr-34. 2'-OH group of the ribose of U5 is hydrogen-bonded to Asp-7. *C*, binding site for U6. Asp-35 and Arg-4 from the neighboring PA-Sm1 subunit are interacting with the base. RNA and protein are depicted in ball and stick representation and colored in green and yellow, respectively.



Structures of PA-Sm1 free or in complex with the RNA show good stereochemistry as judged by the program Procheck (41). All amino acid residues have ϕ and ψ angles within the most favored and allowed regions of the Ramachandran plot. Refinement statistics are given in Table I. The coordinates have been deposited in the Protein Data Bank under accession codes 1H64 for the free protein and 1M8V for the protein/RNA complex).

Modeling of the Three-dimensional Structure of Eukaryotic Sm Proteins and RNA Binding

Sequences of human Sm proteins were obtained from the data base. A sequence alignment of human Sm proteins has been done based on the PA-Sm1 structure. The modeling of the three-dimensional structures of eukaryotic SmE, SmF, and SmG proteins was done with the program Whatif (42) based on sequence homology. In this respect, SmE and SmF were modeled according to the PA-Sm1 structure, and SmG was modeled according to the SmB structure. Construction of the eukaryotic heptamer was achieved by superimposing the eukaryotic Sm proteins onto the PA-Sm1 subunits using the LSQ command in program O and according to the organization of the Sm core (26, 43). The same procedure was followed to add the RNA molecule corresponding to the sequence 5' to the U1 snRNA Sm site, $^{123}\text{AUAAU}^{127}$. The first three nucleotides were superimposed to U4, U5, and U6. U^{127} was positioned as the uridine binding in the uridine-binding pocket. The nucleotide A^{126} is connecting the two binding sites. The backbone position of the nucleotides has been slightly adjusted to follow the surface of the eukaryotic Sm proteins. The electrostatic surface potentials of the PA-Sm1 and the eukaryotic heptamers were calculated using GRASP (44).

RESULTS AND DISCUSSION

Structure of the Free PA-Sm1—The protein forms a ring-like structure composed of seven monomers (Fig. 1A). The triclinic unit cell contains four heptamers, and the crystal packing is dominated by heptamer-heptamer interactions in a head-to-head orientation (whereby the head corresponds to the face containing the N-terminal α -helix). Interactions between the heptamers are essentially because of stacking between the Arg-4 and His-10 residues (from each of the 14 subunits) (Fig. 1A). The presence of Asp-7 in close proximity (around 3 Å) is essential in reducing the overall charge of the Arg-4 residues.

Several water molecules are also found within hydrogen-bonding distance from His-10, coordinating the NE2 position of the imidazole ring. Dimers of heptamers are also present in negatively stained electron microscopy micrographs (data not shown) and in solution as seen in gel filtration experiments, suggesting that the dimer of heptamer may be present in the cell. The dimerization of the heptamer of PA-Sm1 might reflect a functional need for the presence in close proximity of two binding sites.

Each monomer consists of an N-terminal α -helix followed by five strongly bent β -strands. The contacts between monomers are mainly hydrophobic with intersubunit β -sheet formation between β -strands 4 and 5 from adjacent subunits, resulting in a very stable structure even under denaturing conditions. The PA-Sm1 monomer structure can be closely superimposed with the other known archaeal or eukaryotic Sm structures, emphasizing the strong conservation of the Sm-fold. The root mean square deviation values for C α -trace superposition among PA-Sm1 and AF-Sm1 (Protein Data Bank code 1I4K), AF-Sm2 (Protein Data Bank code 1LJO), *Methanobacterium thermoautotrophicum* Sm1 (Protein Data Bank code 1I81), *Pyrobaculum aerophilum* Sm1 (Protein Data Bank code 1I8F), and the eukaryotic Sm structures (Protein Data Bank codes 1D3B and 1B34 for SmB/SmD3 and SmD1/SmD2, respectively) are all between 0.8 and 1.3 Å. It is interesting to note that despite the low sequence similarity between the human and the *Pyrococcus* Sm protein varying from only 18 to 35% (alignment done with the program DNAMAN and using the matrix Blosum250, Lynnon corporation, 2000), the fold is almost completely conserved including the N-terminal region. The closest homologue of PA-Sm1 is SmE (35% sequence homology), a protein known to be essential for viability in yeast (45). The similarity between SmE and PA-Sm1 is especially high in the N-terminal region and may indicate conserved structural features involved in RNA binding (see below).

P. abyssi is the first fully sequenced organism containing

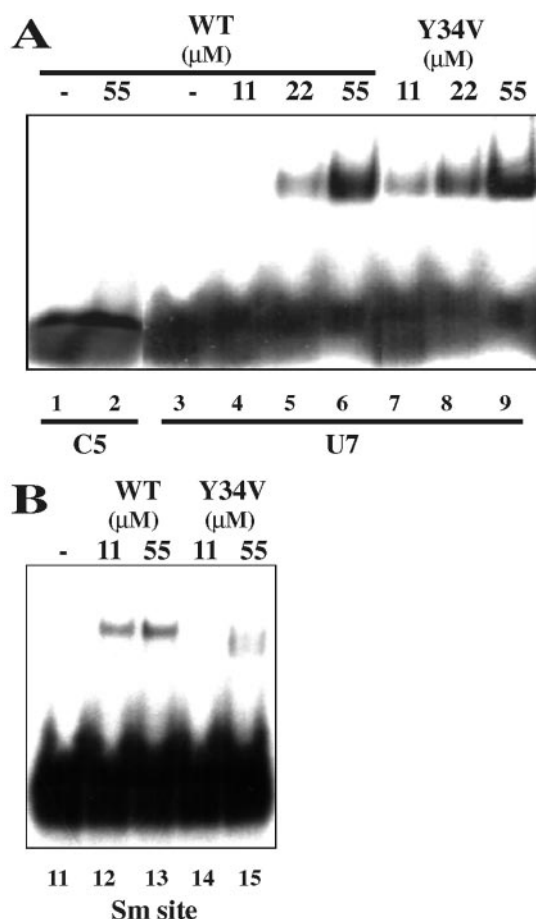


FIG. 4. **PA-Sm1 interacts with RNA *in vitro*.** A, RNA-binding properties of PA-Sm1. Wild type (WT) protein is binding specifically to U₇ but not to C₅ (compare lane 2 with 6). The Y34V mutant binds almost equally well to U₇ than WT protein (compare lanes 4–6 with 7–9). B, binding of the Sm consensus site is strongly decreased for the Y34V mutant (compare lanes 12 and 13 with 14 and 15). Protein concentrations are indicated in micromolars (μ M).

only one open reading frame encoding an Sm protein. Therefore, this heptameric organization is probably the biological unit in agreement with the seven-membered eukaryotic Sm protein ring model proposed by K. Nagai and co-workers (26). Nevertheless and in the absence of any *in vivo* data regarding the Sm complex(es) found in *Pyrococcus* species, we cannot rule out the possibility that the biological unit is not the heptamer but the dimer of heptamers.

Overall Structure of the RNA-bound Form—The PA-Sm1 protein was co-crystallized with a uridine heptamer (U₇). Crystals diffracted up to a 2.1-Å resolution, and two data sets were collected from two different crystals at 2.1- and 2.6-Å resolutions, respectively. The unit cell of the complex contains two heptamers instead of the four present in the RNA-free PA-Sm1 structure. The two heptamers are in a head-to-head orientation, but there are no direct contacts between the two rings (Fig. 1B). Instead, seven poly(U) strands are associated with two non-interacting heptamers (Fig. 1C). The affinity of the RNA molecules for its binding site on the individual monomers is strong enough to disrupt the association of the two protein rings. Interestingly, the stable conformation observed in the crystal structure of the PA-Sm1·U₇ complex consists of one RNA molecule associated with two monomers (Fig. 2B). This specific organization is probably because of the length and the sequence of the oligonucleotide, which allows binding to several sites at the same time (see below). Nevertheless, neither the

overall shape of the monomer nor the architecture of the heptamer changes significantly upon binding of the RNA. Changes mainly concern the side chains of residues involved either in RNA binding or crystal contacts. Residues, which were involved in the formation of dimer of PA-Sm1 heptamers, are now interacting with the RNA molecules (Fig. 2B) (see below). Accordingly, the α -traces of the heptamers in the two structures can be superimposed with an root mean square deviation of 0.56 Å. The very small changes in the protein structure upon RNA binding is probably because of the high specificity of the binding. Conservation of the protein structure is likely to reduce the entropic cost of coordinating the uridine base in the binding pocket (46).

Electron density corresponding to the bound RNA is present at two sites that are not connected to each other (Fig. 1, B and D). Because the difference map calculated with the 2.6-Å resolution data set was much more informative at one of these sites (the internal uridine-binding pocket, Fig. 1D), we subsequently used this data set to build the final model. Nucleotides were built into the difference density map contoured at 2.6 σ (Fig. 1, B and D). They were connected when the difference density indicated the presence of phosphate groups and when this was stereochemically possible. The final model contains 55 nucleotides (6 hexanucleotides, 1 pentanucleotide, and 14 mononucleotides) (Fig. 1, C and E). Fig. 2A shows the overall binding of the RNA to the monomer.

The Internal Site: a Specific Uridine-binding Pocket—The first or internal binding site is very similar to the previously reported uridine-binding pocket of the AF-Sm1 protein (29) and consists of residues His-37, Asn-39, and Arg-63 from the same monomer. Fig. 2C shows a typical difference electron density for this binding site. The uracil base forms stacking contacts with His-37 and Arg-63. The binding pocket is stabilized by a salt bridge between Arg-63 and Asp-65, which in turn forms an ionic interaction with Lys-22 (Fig. 2C). Similarly, in the AF-Sm1·U₅ complex, a highly specific hydrogen-bonding network involving the OD1 position of Asp-35, the OD1 and ND2 atoms of Asn-39, and N3 and O4 of the uridine base (Fig. 2C) renders this binding site specific for uridine. However, in contrast to the AF-Sm1·U₅ complex, we do not observe clear density connecting the uridines bound to neighboring binding pockets (see below).

The External RNA Binding Site—RNA molecules are located at the interface between the two heptamers. We are referring to this site as the external binding site. Seven oligo(U) strands connect the external sites of two monomers facing each other in the two heptamers (Fig. 2B). The binding sites on the two opposing monomers are identical and related by 2-fold non-crystallographic symmetry. Nucleotides belonging to the same chain are numbered U1–U6 with the exception of one case where only five uridines could be modeled and nucleotides were numbered U1–U5. Nucleotides bound to the external site display the usual C3'-endo conformation.

Residues Arg-4 from the N-terminal α -helix and Tyr-34 from β -strand 2 form the binding site of U1 (or U4) (Fig. 3A). The nucleotide is stacked between Tyr-34 and either the guanidinium group or the hydrophobic part of the Arg-4 side chain. In the latter case, a water molecule is bridging the NH1 atoms of Arg-4 and the O4' atom of U1 (Fig. 3A). Hydrogen bonds involving the N3 and O4 atoms of U1 and both the main chain carbonyl oxygen and amide nitrogen atoms of Tyr-34 discriminate against the binding of a cytidine at this position.

The following nucleotide U2 (or U5) is stacked on His-10, which is kept in a fixed orientation by a hydrogen bond with Tyr-34 (Fig. 3B). The binding of U2 (or U5) is further enhanced by a hydrogen bond between the side chain of Asp-7 and the

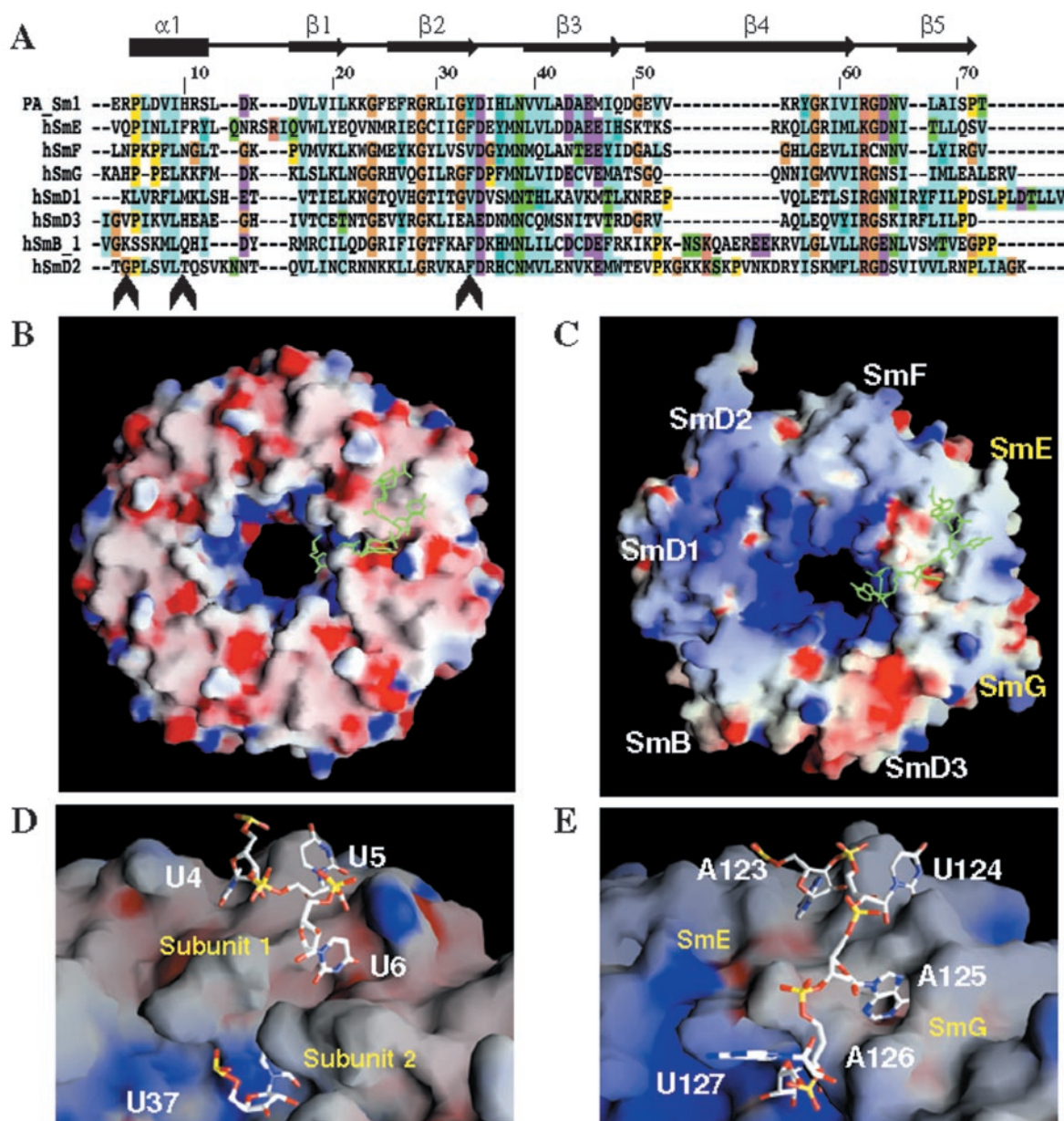


FIG. 5. Model for the binding of RNA in the eukaryotic Sm core based on the PA-Sm1·U₇ complex. **A**, structure-based sequence alignment of PA-Sm1 and the eukaryotic Sm core proteins. Secondary structure elements defined by the PA-Sm1 structure are shown on top. Modeling of the unknown eukaryotic structures was done according to this alignment. Arrowheads indicate amino acids involved in the external binding site. Accession numbers (GenBankTM) are: PA-Sm1, Q9V0Y8; hSmE, P08578; hSmF, Q15356; hSmG, Q15357; hSmD1, P13641; hSmD3, P43331; hSmB_1, P14678; and hSmD2, P43330. **B**, Overall view of the PA-Sm1 heptamer with the RNA bound. The cavity where the RNA binds is positively charged. The external surface is less charged but also accommodates specifically the RNA. **C**, the modeled eukaryotic Sm core complex (see “Experimental Procedures”). Subunit organization is as defined by Stark *et al.* (2001). The RNA pentamer, AUAAU, corresponds to nucleotides 123–127 of the human U1 snRNA. **D**, close-up view of the RNA binding sites observed in PA-Sm1. **E**, close-up view of the eukaryotic Sm core complexed to RNA. A¹²³ is positioned as U4 and is specifically recognized by the backbone of Phe-39 as well as stacking with its benzyl ring (numbering according to the PA-Sm1 sequence). U¹²⁴ is positioned as U5 and is stacking onto Phe-10. A¹²⁵ is positioned as U6 and is interacting with Asp-35 of SmG and Gln-4, Pro-5, and Ile-6 of SmE. U¹²⁷ is stacked between Tyr-37 and Lys-63 as well as hydrogen-bonded to Asn-39. A¹²⁶ is solvent-accessible and connects the nucleotides bound at the external and internal sites. The protein surfaces are color-coded according to their electrostatic potentials (red = −20 kT; blue = +20 kT). The RNA is colored in green in panels **B** and **C** and according to the atom type (red = oxygen; yellow = phosphate; white = carbon; blue = nitrogen) in panels **D** and **E**. Panels **B–E** have been produced with the program GRASP (44).

ribose 2′-OH group (Fig. 3*B*). Well defined densities are present for phosphate groups connecting nucleotides U1 and U2 or U4 and U5.

The nucleotide U3 connects two dinucleotides bound onto two external binding sites (Fig. 2*B*). Because these sites are identical, the poly(U) strand has two possible orientations, which only differ by the position of the connecting nucleotide. U3 actually breaks the 2-fold non-crystallographic symmetry relating the two external binding sites, and therefore, this

position shows a weaker electron density. Nevertheless, in most of the sites, it was clearly possible to build this nucleotide. In these cases, it is stacked between the uracil rings of U2 and U5.

Likewise, the nucleotide U6 displays significant density only in some monomers. This nucleotide has its phosphate stabilized by the amidino group of Arg-4 and its base by several hydrogen bonds (Fig. 3*C*).

Nucleotides U2, U3, U5, and U6 have one of their phosphate

oxygen at hydrogen bonding distance from a strong central density peak ($>4\sigma$ in the $2F_o - F_c$ map). The coordination indicates that this peak represents a divalent cation, presumably Ca^{2+} , which is the only divalent cation present in the crystallization buffer at a concentration of 250 mM. Moreover, this ion was essential for the crystallization process, presumably because it stabilizes the backbone conformation (Fig. 2B) and allows the RNA molecules to be associated to two heptamers at the same time.

We do not believe that the specific packing observed in the PA-Sm1-RNA complex composed of two heptamers interacting with seven oligonucleotides is reflecting the stoichiometry of the PA-Sm1-RNA complex found *in vivo*. However, we believe that the specific RNA sequence used in the crystallization procedure is able to bind to two different heptamers. Indeed, the external site is composed of three individual nucleotide binding sites, two of which display specificity for a uridine base, independent from its orientation. Therefore, the external site can interact with the 5' or the 3' end of the oligonucleotide, leading to the binding of two heptamers to the same RNA molecules. In this case, the biological unit is not likely to be a dimer of heptamers bound to one RNA molecule but rather a heptamer bound to one RNA.

Interrelation between the Two RNA Binding Sites—To better understand the function and relevance of the external RNA binding site and its relation to the internal site, we focused on Tyr-34, which is conserved in all the archaeal Sm1-type proteins and is involved in RNA binding in the PA-Sm1-U₇ complex. We mutated it to valine to maintain the hydrophobic character of the residue, to prevent the stabilization of His-10 in a stacking orientation with U2 or U5, and to remove the possibility of stacking with U1 or U4. The ability of PA-Sm1 wild type protein or the Y34V mutant to bind RNA has been analyzed by gel shift assays. The point mutation within the external binding site does not affect the binding to oligo(U) as seen in Fig. 4A, compare lanes 4–6 and 7–9. Indeed, the preferential binding site for oligo(U) is the internal binding site, indicating that the external binding site does not act as a recognition site for the oligonucleotide. On the other hand, complex formation between the eukaryotic Sm consensus RNA, AAUUUUUGG, and the wild type protein is strongly reduced with the mutant protein (Fig. 4B, lanes 12–13 and 14–15). This shows that a mutation in the external site almost abolishes the RNA-binding properties of PA-Sm1 for a non-symmetrical RNA without reducing the affinity of the PA-Sm1 heptamer for U-rich sequences. It suggests that the second binding site of the PA-Sm1 protein stabilizes additional nucleotides after specific recognition of the U-rich sequence by the uridine-binding pocket. Because the internal and the external sites are necessary for a proper binding to non-symmetrical RNA, the lack of density observed in the crystal between U6 and the uridine bound in the internal site is probably the result of a disorder of the nucleotide(s) connecting the two sites. Indeed, the phosphate group of the internal uridine is located almost at the height of the His-37 ring and is directed toward the external site (Fig. 2C). In line with this interpretation is the fact that single uridine nucleotides do not bind to AF-Sm2 (47). In case of the AF-Sm1-U₅ complex, the amino acids composing the external binding sites are blocked by crystal contacts leaving only the internal site, *i.e.* the uridine-binding pocket, available for binding of the RNA molecules. In this case, the increased length of the oligonucleotide and its simultaneous binding to two heptamers result in the observation of the external binding site and of an isolated base in the uridine-binding pocket.

The external binding site seen in our structure and shown by mutagenesis and gel shift experiments supports the idea that

interactions of Sm proteins with their RNA targets in Archaea as well as in Eukarya are not limited to the uridine-binding pocket. It suggests that the two sites observed in the crystal are probably necessary for the binding of PA-Sm1 to its *in vivo* target. Indeed, the mutation of tyrosine 34 hinders the association with the Sm site RNA. In addition, the heptameric organization might also determine the length of the U-rich sequence to which the oligomer binds. The *in vivo* RNA target of the PA-Sm1 protein has not yet been identified but it is likely that it will contain a uridine-rich stretch, which will bind to the internal site, whereas the upstream sequence would interact with the external site.

Homology with the Eukaryotic snRNP—The association of Sm proteins with the snRNAs plays a key role in the biogenesis as well as in the function of U snRNPs. The model proposed for human snRNPs by K. Nagai and co-workers (26) provided the first model for an Sm core. However, the association with the snRNA was not modeled.

Therefore, we decided to use the PA-Sm1-U₇ complex to model the association of eukaryotic Sm proteins with a short sequence of the U1 snRNA as described under "Experimental Procedures" (Fig. 5, B and C). A structure-based sequence alignment between PA-Sm1 and the Sm core proteins, namely SmD1, D2, D3, B, E, F, and G, shows that the external binding site of PA-Sm1 is conserved in the case of SmE and probably SmG. In SmE, Tyr-34 and His-10 become phenylalanines, and Pro-5 is conserved. Moreover, the organization of eukaryotic Sm proteins within the Sm core complex shows that the SmE protein directly contacts SmG, which can be cross-linked to the first uridine of the U1 snRNA Sm site (30). This putative binding site on SmE, which precedes SmG in a counter-clockwise orientation within the ring (43), suggests that the sequence directly upstream of the Sm site would bind to the SmE external site. The U1 snRNA sequence ¹²³AUAAU¹²⁷ was positioned as follows (Fig. 5, D and E): (i) the first three nucleotides bind to SmE, thereby protecting them from hydroxyl radical attack (48); (ii) U¹²⁷, the first uridine of the uridine stretch, interacts with the uridine-binding pocket of SmG in agreement with cross-linking studies (30, 49); and (iii) the nucleotide in between A¹²⁶ connects these two sites, because the distance between the O3' of A¹²⁵ and the phosphate of U¹²⁷ is $\sim 7.3\text{\AA}$. We do not see any connecting density in our structure between the external and the internal binding sites, most probably because of some disorder in the crystal. Nevertheless, the position of the connecting adenosine agrees with most of the biochemical studies done on the U1 and U5 snRNA Sm sites. Indeed, the orientation of A¹²⁶ upon binding of the snRNA to the Sm proteins would make its base solvent-accessible, triggering its reactivity to chemicals like dimethyl sulfate and/or diethylpyrocarbonate (N7-A > N7-G) (48, 50). It was also shown that this adenine was important for the stability of the snRNP but not for the binding of the Sm protein to U-rich RNA. This nucleotide has finally been mutated in the case of U5 snRNA without revealing an essential function (51).

The proposed model for the binding of the sequence 5' of the U1 snRNA Sm site and the model for the eukaryotic Sm core allows us to better understand the role of Sm proteins in facilitating RNA-RNA interactions. The electrostatic surface potential of the Sm core shows two distinct regions on the N-terminal side surface (Fig. 5C) (44). The area formed by SmE, SmG, SmD3, and most of SmB is globally neutral but contains specific RNA binding sites (on SmE and SmG) and would correspond to the snRNA binding site in agreement with the 10-Å model of the U1 snRNP (43) as well as with footprinting experiments (49). The surface composed of subunits D1, D2, and F is significantly more positively charged, allowing RNA

interactions based on unspecific electrostatic contacts. Because the U1 snRNP recognizes non-conserved sequences around the 5'-splice site (45) and the C-terminal tails of SmB, SmD1, and SmD3 have been shown to interact with the pre-mRNA (52), the Sm ring is probably the site for interaction between the pre-mRNA and the snRNA.

The different Sm complexes, which are associated with different RNA binding sites, are generally composed of different sets of Sm proteins (11, 12, 53). The present model gives a more precise view of the association between the SmE and SmG core proteins and the RNA for a specific case, the Sm core complex involved in splicing. But because these two proteins are also present in the other complexes, it is probable that those complexes will interact in a similar way with their respective RNA targets.

CONCLUSION

Because *P. abyssi* contains only one type of Sm protein, the free and RNA-complexed structures are providing the first model of a complete Sm core. It demonstrates that the uridine-binding pocket is the primary binding site of U-rich RNA in Archaea and most probably as well in Eukarya. The PA-Sm1-RNA complex also reveals a secondary RNA binding site located on the surface of the ring, which is involved in the binding of non-symmetrical RNA and may play a role in defining the length of the RNA sequence target. Thereby, it is of prime interest to identify the *in vivo* target for PA-Sm1 in order to understand the need of a heptamer or possibly of a dimer of heptamers in the function of modulating RNA-RNA interactions.

Based on this structure as well as on the available biochemical data, we propose a model of the eukaryotic Sm core proteins bound to a 5-mer RNA representing the sequence directly upstream of the U1 snRNA Sm site. According to this model, the SmE protein would serve as the binding site for the U1 snRNA leaving SmF, SmD2, and SmD1 free for unspecific interactions with the pre-mRNA. Therefore, this model suggests how the Sm proteins would achieve their function of modulating RNA-RNA interactions. The other Sm protein complexes with the exception of the complexes formed by the Lsm proteins are all containing the SmE and SmG proteins, which are the Sm proteins shown to have additional RNA interactions besides the uridine-binding pocket. We can now start to elucidate the regulatory function of the remaining Sm protein subunits forming Sm complexes with similar shape and fold but quite different localizations and targets.

Acknowledgments—We gratefully acknowledge the assistance of Andy Thompson (ESRF/EMBL outstation) and Klaus Scheffzek with data collection and processing. We thank Luc Moulinier for help with the construction of the PA-Sm1 search model, Bettina Boettcher for analyzing PA-Sm1 solutions by electron microscopy, and Elena Conti and Klaus Scheffzek for critically reading the paper.

REFERENCES

- Magrum, L. J., Luehrsen, K. R., and Woese, C. R. (1978) *J. Mol. Evol.* **11**, 1–8
- Kuhn, J. F., Tran, E. J., and Maxwell, E. S. (2002) *Nucleic Acids Res.* **30**, 931–941
- Mayer, C., Suck, D., and Poch, O. (2001) *Trends Biochem. Sci.* **26**, 143–144
- Smith, D. R., Doucette-Stamm, L. A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lum, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Reeve, J. N., et al. (1997) *J. Bacteriol.* **179**, 7135–7155
- Séraphin, B. (1995) *EMBO J.* **14**, 2089–2098
- Puhler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H. P., Lottspeich, F., Garrett, R. A., and Zillig, W. (1989) *Proc. Natl. Acad. Sci. U. S. A.* **86**, 4569–4573
- Padgett, R. A., Mount, S. M., Steitz, J. A., and Sharp, P. A. (1983) *Cell* **35**, 101–107
- Strub, K., and Birnstiel, M. L. (1986) *EMBO J.* **5**, 1675–1682
- Seto, A. G., Zaug, A. J., Sobel, S. G., Wolin, S. L., and Cech, T. R. (1999) *Nature* **401**, 177–180
- Bouveret, E., Rigaut, G., Shevchenko, A., Wilm, M., and Séraphin, B. (2000) *EMBO J.* **19**, 1661–1671
- He, W., and Parker, R. (2000) *Curr. Opin. Cell Biol.* **12**, 346–350
- Achsel, T., Brahm, H., Kastner, B., Bachi, A., Wilm, M., and Lührmann, R. (1999) *EMBO J.* **18**, 5789–5802
- Mattaj, I. W. (1998) *Curr. Biol.* **8**, 93–95
- Fischer, U., Liu, Q., and Dreyfuss, G. (1997) *Cell* **90**, 1023–1029
- Mattaj, I. W. (1986) *Cell* **46**, 905–911
- Hamm, J., Darzynkiewicz, E., Tahara, S. M., and Mattaj, I. W. (1990) *Cell* **62**, 569–577
- Palacios, I., Hetzer, M., Adam, S. A., and Mattaj, I. W. (1997) *EMBO J.* **16**, 6783–6792
- Mayes, A. E., Verdone, L., Legrain, P., and Beggs, J. D. (1999) *EMBO J.* **18**, 4321–4331
- Jarmolowski, A., and Mattaj, I. W. (1993) *EMBO J.* **12**, 223–232
- Raker, V. A., Hartmuth, K., Kastner, B., and Lührmann, R. (1999) *Mol. Cell. Biol.* **19**, 6554–6565
- Salgado-Garrido, J., Bragado-Nilsson, E., Kandels-Lewis, S., and Séraphin, B. (1999) *EMBO J.* **18**, 3451–3462
- Gilmartin, G. M., Schaufele, F., Schaffner, G., and Birnstiel, M. L. (1988) *Mol. Cell. Biol.* **8**, 1076–1084
- Moller, T., Franch, T., Hojrup, P., Keene, D. R., Bachinger, H. P., Brennan, R. G., and Valentin-Hansen, P. (2002) *Mol. Cell* **9**, 23–30
- Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C., and Storz, G. (2002) *Mol. Cell* **9**, 11–22
- Schumacher, M. A., Pearson, R. F., Moller, T., Valentin-Hansen, P., and Brennan, R. G. (2002) *EMBO J.* **21**, 3546–3556
- Kambach, C., Walke, S., Young, R., Avis, J. M., de la Fortelle, E., Raker, V. A., Lührmann, R., Li, J., and Nagai, K. (1999) *Cell* **96**, 375–387
- Collins, B. M., Harrop, S. J., Kornfeld, G. D., Dawes, I. W., Curmi, P. M., and Mabbitt, B. C. (2001) *J. Mol. Biol.* **309**, 915–923
- Mura, C., Cascio, D., Sawaya, M. R., and Eisenberg, D. S. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98**, 5532–5537
- Törö, I., Thore, S., Mayer, C., Basquin, J., Séraphin, B., and Suck, D. (2001) *EMBO J.* **20**, 2293–2303
- Urlaub, H., Raker, V. A., Kostka, S., and Lührmann, R. (2001) *EMBO J.* **20**, 187–196
- Hajnsdorf, E., and Regnier, P. (2000) *Proc. Natl. Acad. Sci. U. S. A.* **97**, 1501–1505
- Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. (2001) *Genes Dev.* **15**, 1637–1651
- Kabsch, W. (1993) *J. Appl. Crystallogr.* **26**, 795–800
- Otwinski, Z., and Minor, W. (1997) *Methods Enzymol.* **276**, 307–326
- Navaza, J. (1994) *Acta Crystallogr. Sec. A* **50**, 157–163
- Collaborative Computer Project 4 (1994) *Acta Crystallogr. Sec. D* **50**, 760–763
- Törö, I., Basquin, J., Teo-Dreher, H., and Suck, D. (2002) *J. Mol. Biol.* **320**, 129–142
- Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. Sec. D* **54**, 905–921
- Kleywegt, G. J., and Jones, T. A. (1999) *Acta Crystallogr. Sec. D* **55**, 941–944
- Jones, T. A., Zou, J. Y., Cowan, S. W., and Kjeldgaard (1991) *Acta Crystallogr. Sec. A* **47**, 110–119
- Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) *J. Appl. Crystallogr.* **26**, 283–291
- Vriend, G. (1990) *J. Mol. Graph.* **8**, 52–56
- Stark, H., Dube, P., Lührmann, R., and Kastner, B. (2001) *Nature* **409**, 539–542
- Nicholls, A., Sharp, K., and Honig, B. (1991) *Proteins* **11**, 281–296
- Bordonne, R., and Tarassov, I. (1996) *Gene (Amst.)* **176**, 111–117
- Bauer, C. B., Holden, H. M., Thoden, J. B., Smith, R., and Rayment, I. (2000) *J. Biol. Chem.* **275**, 38494–38499
- Achsel, T., Stark, H., and Lührmann, R. (2001) *Proc. Natl. Acad. Sci. U. S. A.* **98**, 3685–3689
- Hartmuth, K., Raker, V. A., Huber, J., Branlant, C., and Lührmann, R. (1999) *J. Mol. Biol.* **285**, 133–147
- Heinrichs, V., Hackl, W., and Lührmann, R. (1992) *J. Mol. Biol.* **227**, 15–28
- Krol, A., Westhof, E., Bach, M., Lührmann, R., Ebel, J. P., and Carbon, P. (1990) *Nucleic Acids Res.* **18**, 3803–3811
- Jones, M. H., and Guthrie, C. (1990) *EMBO J.* **9**, 2555–2561
- Zhang, D., Abovich, N., and Rosbash, M. (2001) *Mol. Cell* **7**, 319–329
- Pillai, R. S., Will, C. L., Lührmann, R., Schumperli, D., and Muller, B. (2001) *EMBO J.* **11**, 5470–5479
- Evans, S. V. (1993) *J. Mol. Graph.* **11**, 127–128, 134–138
- Carson, M. (1997) *Methods Enzymol.* **277**, 493–505