Acknowledgements

I have received an incredible amount of support during my graduate research, from personal support to educational guidance; with many individuals supplying both. Foremost I would like to thank my advisor, Cameron Mura, for his guidance, both personal and scientific. His availability, knowledge, and even criticisms were invaluable in guiding my scientific progress and development. I would also like to thank the members of my dissertation committee - Robert Bryant, Kevin Lehmann, and John Bushweller – for always reminding me about the longer goals of my projects.

I would like to thank all of my labmates from the Mura laboratory, and particularly Kimberly Stanek for feedback, Jennifer Patterson who collaborated with me on the *Thermatoga maritima* Hfq project (Chapters 3 and 4), and former members Kanishk Jain, an undergraduate in the laboratory now at UCLA, and H. Thien Nguyen, a technician now at UCSD, who helped me with *Pyrobaculum aerophilum* SmAP2. For my crystallography education I would like to thank Michael Sawaya and Duilio Cascio from UCLA, whose practical knowledge and infectious positive attitudes towards x-ray crystallography made it possible to complete the structure of *Pyrobaculum aerophilum* SmAP2. In addition, I would like to thank Linda Columbus and the Columbus laboratory members for their friendship, technical input, feedback, and general assistance.

Special thanks goes to my family, especially my parents for their unwavering support, and my brothers for keeping me humble. I would also like to thank my eclectic posse of friends who provided me with a relaxing outlet and keeping everything in perspective. My sincerest thanks for my wife Lindsay Randolph for her never ending love and support and continuing faith in my work.

Abstract

The Sm superfamily's central role in RNA processing and regulation, combined with their existence in all three domains, makes them a model system for exploring RNA processing evolution. Sm-mediated interaction between RNAs play vital roles in important pathways such as virulence, quorum sensing, cell death and aging, and mRNA splicing. The largest gap in our knowledge of the Sm superfamily is in the archaeal branch. Many archaeal systems can provide invaluable knowledge about their more complicated analogous eukaryal systems by supplying a simpler model to work with. Initial work on the Sm-like archaeal proteins (SmAPs) were crucial to our understanding of how Sm proteins oligomerize and bind RNA. Unfortunately, since this early work, the study of SmAPs has been limited, and SmAP *in vivo* functions are virtually unknown. Understanding these *in vivo* functions of SmAPs would allow us a better understanding of basic Sm protein function, provide a window into the evolution of the large eukaryal ribonucleoprotein complexes, and possibly link the evolution of bacterial Hfq and eukaryal Sm proteins.

The crenarchaea *Pyrobaculum aerophilum* is a deep-branching, hyperthermophile that encodes multiple SmAP paralogs. The two known *Pae* SmAP structures (SmAP1, SmAP3) illuminated Sm protein evolution and assembly, and implied that these homologs may represent an ancestral form of the complexes that developed into the extant heteromeric Sm assemblies of eukaryotes, such as those at the heart of the spliceosome. Our work on the final *Pae* SmAP, SmAP2, reveals that *Pae* SmAP2 oligomerizes as a unique octamer (unseen in previous SmAPs) in two rare space groups, and binds both A-rich and U-rich RNA reminiscent of the bacterial Hfq (*chaperone*). The crystal structure revealed that *Pae* SmAP2 lacks the conserved residues seen in the common U-rich and A-rich binding pockets of other Sm proteins, but does contain the aromatic (Tyr-42) necessary for lateral-rim binding. Further research is necessary to determine the specific binding mechanisms of *Pae* SmAP2···RNA binding, the *Pae* SmAP2 solution state, and determine the individual functions of the SmAP paralogs in *Pyrobaculum aerophilum*.

Many deep-branching bacteria share a high degree of similarity (genomically) with archaea, including the hyper-thermophilic *Thermotoga maritima*. *T. maritima* Hfq is an interesting homolog because of its simplicity (no C-terminal tail) and the aforementioned archaeal genome. The two studies reported here, one in archaea and one in bacteria, will help to illuminate the functions of ancient Sm proteins, supply a window into RNA processing in archaea, and the evolution of Sm proteins. In this study, we found that the putative Hfq homolog from the thermophillic *Thermotoga maritima* (*Tma*) heterologously co-purifies with U/C-rich nanoRNAs, binding with a nanomolar affinity. Identified nanoRNA sequences all contain a 5' monophosphate and a 3' hydroxyl and compete with U-rich sequences for the proximal face of Hfq. Data suggests that the position of cytosine within the sequence, rather than the absolute number of cytosines is the key factor in determining affinity. The crystal structure shows that, even under denaturing condition, a small amount of the heterologous nanoRNA remains bound. *Tma* Hfq forms a hexamer within the crystal, agreeing with previous studies on the functional form of Hfq in *Escherichia coli* (*Eco*) and other bacterial species. However, our studies of *Tma* Hfq suggest that an equilibrium exists between a homo-hexamer and a homo-dodecamer. Both oligomeric states are capable of binding poly-adenine and poly-uracil RNA with low nanomolar binding affinities, with poly-A and poly-U RNA preferentially binding to the dodecamer and hexamer, respectively. This leads to a shift in equilibrium between the states; poly-U shifting the equilibrium toward the hexameric state, and poly-A having no effect.

Table of Contents

List of Figures

Ch. 3

Ch. 4

List of Tables

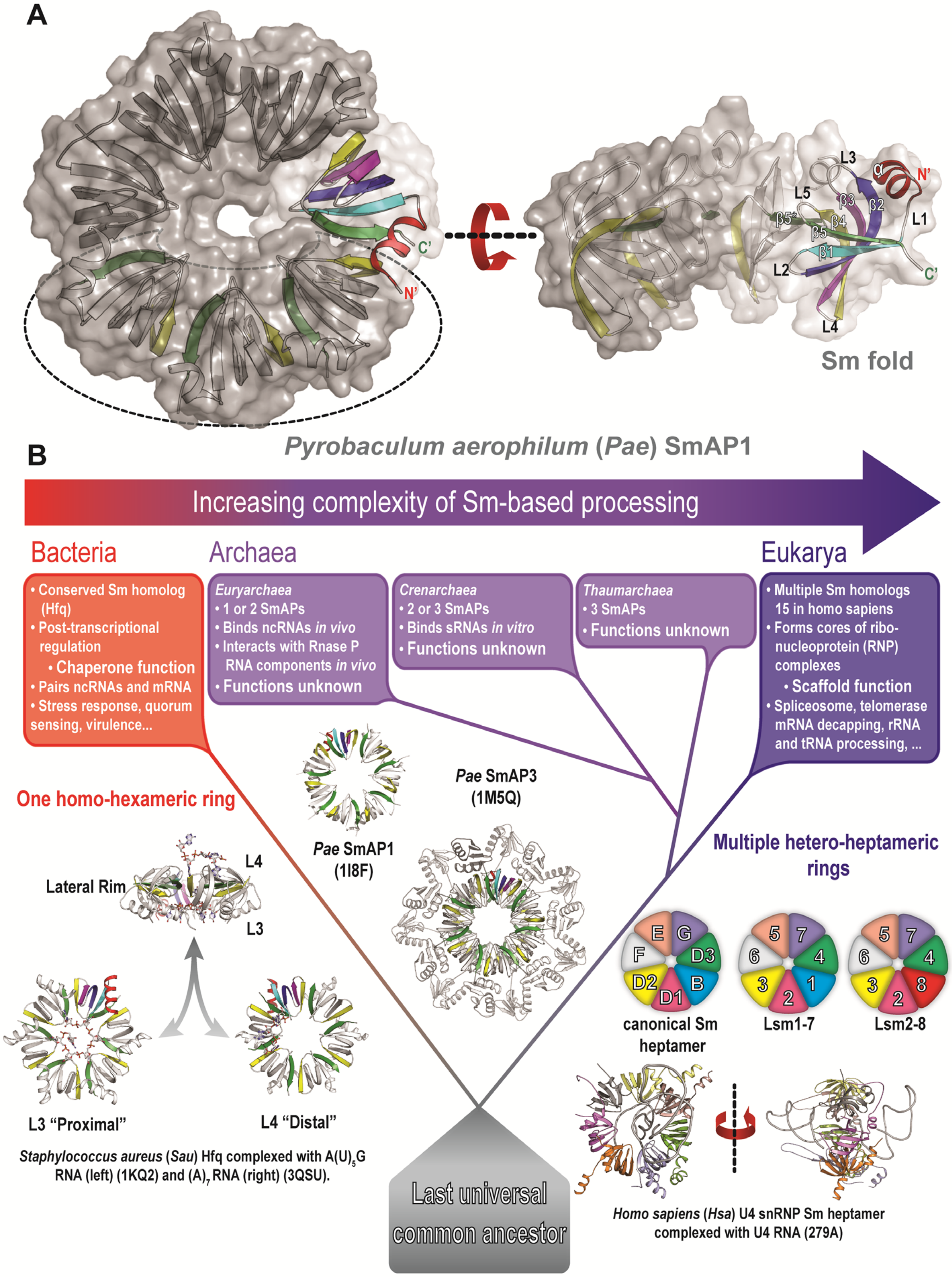
Table **Page**

Ch. 2

Ch. 3

Ch. 4

Overview of the Sm Superfamily



***Pyrobaculum aerophilum* (*Pae*) SmAP1**

Increasing complexity of Sm-based processing

Bacteria

- Conserved Sm homolog (Hfq)
- Post-transcriptional regulation
  - Chaperone function
- Pairs ncRNAs and mRNA
- Stress response, quorum sensing, virulence...

Archaea

Euryarchaea

- 1 or 2 SmAPs
- Binds ncRNAs *in vivo*
- Interacts with Rnase P RNA components *in vivo*
- Functions unknown

Crenarchaea

- 2 or 3 SmAPs
- Binds sRNAs *in vitro*
- Functions unknown

Thaumarchaea

- 3 SmAPs
- Functions unknown

Eukarya

- Multiple Sm homologs 15 in homo sapiens
- Forms cores of ribonucleoprotein (RNP) complexes
  - Scaffold function
- Spliceosome, telomerase mRNA decapping, rRNA and tRNA processing, ...

Staphylococcus aureus (*Sau*) Hfq complexed with A(U)$_5$G RNA (left) (1KQ2) and (A)$_7$ RNA (right) (3QSU).

Homo sapiens (*Hsa*) U4 snRNP Sm heptamer complexed with U4 RNA (279A)

Figure 1. Overview of the Sm superfamily. The Sm superfamily is found in all domains of life. (A) Sm proteins exhibit a conserved fold consisting of a small highly bent β-sheet preceded by a small α-helix which oligomerizes into toroid oligomers. Key features are Loop 3 (L3) and Loop 4 (L4) which differentiate the faces, commonly referred to as *proximal* (L3) or *distal* (L4) to the α-helix. (B) Sm proteins bind RNA and are involved in a variety of RNA related processes across all domains. A simple phylogenetic tree schematic shows the early development of the Sm family and represents the closer relationship between the archaea and eukarya branches.

Chapter 1: Sm proteins: An Ancient RNA-binding Superfamily

Peter Randolph

University of Virginia, Department of Chemistry, Charlottesville, VA 22904

1 Sm Protein Research: A Brief History

1.1 In Bacteria

Sm proteins were originally discovered in the late 1960s while researching the replication of the RNA bacteriophage Qβ in *Escherichia coli* [1]. Replication was found to depend on an *E. coli* protein dubbed host-factor I for replication of Qβ or 'Hfq' [2]. Over the next 30 years, Hfq was found to be a highly abundant, heat-resistant, stable homo-hexamer that binds both adenine-rich and uridine-rich single-stranded RNA (ssRNA) [3-6]. It was not until the 1990s that the *in vivo* functions of *E. coli* Hfq became apparent [7]. Hfq knockout (Δ*hfq*) studies in *E. coli* displayed a pleiotropic phenotype, including vulnerability to stress and an abnormal growth rate [7] (later, Δ*hfq* studies in pathogenic bacteria revealed a decreased virulence). The majority of the observed phenotypes were similar to those of bacteria with a disrupted *rpoS* gene, which encodes the stationary phase sigma factor of RNA polymerase σ^s [8,9]. Later studies elucidated the relationship by demonstrating that Hfq is a positive regulator of post-transcriptional *rpoS* expression. While *rpoS* regulation accounted for many of the phenotypes found for Δ*hfq* mutants, many phenotypes remained unclear suggesting additional targets for Hfq regulation [10]. Since then multiple Hfq targets including additional messenger RNAs (mRNAs) and regulatory small RNAs (sRNAs) have been identified. Hfq's main function role appears to be that of a *chaperone*, assisting sRNAs in binding their target mRNAs. Hfq's flexibility in binding various sRNAs and mRNAs allows it to serve as a central hub in a vast array of post-transcriptional regulatory pathways.

1.2 In Eukarya

In an unrelated line of inquiry dating to the late 1970s, the canonical Sm core proteins were discovered [11] as a group of small antigens involved in the autoimmune disease systemic lupus erythematosus [12,13]. Named for the original patient whom the extracts were taken, Smith [14], seven Sm proteins were found to associate with various ribonucleoprotein (RNP) complexes [15]. Canonical Sm core proteins were revealed to form the core of the uracil-rich small nuclear RNPs (U snRNPs) that are the major components of the intron removing spliceosome [16-18]. Continued work showed that Sm proteins held the catalytic small nuclear RNA (snRNA) components in place, allowing the various protein components to also bind, essentially serving as *scaffolding* for RNP biogenesis [19]. Sm proteins were ascertained to bind a uracil-rich region common to all U snRNAs. The first Sm structures solved in 1999 were a hetero-dimer (SmD1·SmD2) and a hetero-trimer (SmF·SmE·SmG) [20]. Biophysical studies revealed that canonical Sm core proteins did not form their full hetero-heptamer unless their target RNA was present [21]. Bioinformatic studies identified an additional Sm protein group in eukarya called the Sm-like (like Sm, Lsm) proteins [18,22]. Lsm proteins are more stable than Sm core proteins and, while Lsm proteins are also involved in splicing, they also have additional functions in various RNA processing pathways [23].

1.3 In Archaea

Bioinformatics also expanded our knowledge of the Sm family into the archaeal domain. Initially, Sm proteins were not expected to exist in the archaeal domain because of the absence of introns necessitating splicing in the archaeal genome. These Sm-like archaeal proteins (SmAPs) were observed to be homo-heptameric and much more stable than the canonical Sm core proteins making them ideal for structural studies. The first structure of a fully-formed Sm ring was released in 2001, with three separate SmAP structures published [24-26]. SmAPs were found to bind RNA and DNA but few functional experiments have been attempted [24-26] and thus much is still unknown about the function of this branch of the Sm family.

1.4 A Superfamily of Proteins

It was not until the early 2000s that structural studies identified the bacterial Hfq as belonging to the Sm family [27]. Weak sequence similarity in the N-terminal section had previously hinted at a relationship, but it was not until the first Hfq crystal structure was determined that Hfq was confirmed as the third branch of the Sm family [28-30]. With the three lines of inquiry described above merged, it became clear that the various Sm homologs were in fact a large superfamily of RNA binding proteins heavily involved in general RNA processing across all domains [31]. For an overview of the Sm superfamily see Figure 1.

2 Nomenclature

Nomenclature for the separate lines of research in the Sm superfamily (bacterial, eukaryal, archaeal) became fairly ingrained before unification. In the present work, the bacterial Sm protein will be referred to by its commonly used name, ‘Hfq’. Many archaeal Sm proteins were originally referred to as either Sm1 or Sm2, which is commonly confused with the Sm1 and Sm2 motif, so all Sm archaeal proteins will be denoted ‘SmAPs’ (SmAP1, SmAP2, *etc.*). The eukaryal Sm proteins, discovered and named first, will herein be called the canonical Sm core proteins, and other eukaryal Sm proteins will follow the common nomenclature of Like-Sm (Lsm) proteins. The use of ‘canonical Sm core proteins’ will differentiate these particular homologs from the more generic term ‘Sm proteins’ in reference to elements common to the entire superfamily.

3 Evolution of the Sm Superfamily

Sm proteins are likely one of the most ancient protein superfamilies [32], and they may give insight into the transition from the primordial RNA world to the contemporary protein-based world. The diversity of the Sm superfamily appears to have arisen from a series of gene duplication events and divergence of these copies, leading to the development of multiple hetero-oligomers in eukarya from the homo-oligomers in bacteria and eukarya. However, horizontal gene transfer also may have played a role in the increasing number of Sm genes [33]. Progressing from bacteria to archaea to eukarya, increase in the number of Sm paralogs steadily increases. Bacteria generally encode one conserved Hfq, though recent studies have identified a second *hfq* gene in some bacteria (dubbed *hfq2*) [34,35]. Many archaea encode multiple SmAP paralogs (2 or 3), however no hetero-oligomers have been seen. Because of the seven-fold symmetry in eukaryal Sm/Lsm and SmAPs, a full seven paralogs must arise before a hetero-heptamer is possible. Converting from homo-oligomers to hetero-oligomers is a relatively frequent event in the evolution of prokarya to eukarya (*e.g.* exosome, ubiqituin-like proteins, type II chaperones [36-38]). Typically, this is thought to be the result of gene duplication followed by sequence divergence ("diversification-duplication" model) leading to unique paralogs which can oligomerize into a complex similar to the original. While this seems a waste of genetic resources, the homomer to heteromer transition allows functions to become more specific and results in tighter regulation of such complexes necessary with specific sequence of steps to generate the oligomer [39].

A key step in the evolution of Sm proteins is the introduction of non-self-splicing introns, necessitating a spliceosome to recover function. It is speculated that spliceosomal introns arose when self-splicing type II introns began to diverge, and some lost their ability to form the necessary structures/sequence for self-splicing, or became early snRNA genes [40-42]. This theory is supported by the similarity in the internal base-pairing of self-splicing type II introns and the pre-mRNA···snRNA base-pairing in the Spliceosomal complex [43]. The U6 snRNA shares a particularly close sequence similarity to the catalytic domain of the self-splicing group II intron [44]. From these observations, it is proposed that one ancient Sm gene was duplicated multiple times, resulting in seven paralogs which diverged; this, in turn, led to functional specialization of each of the components, culminating in a hetero-heptamer. Eukaryal Lsm proteins are considered the earliest forming Sm hetero-oligomers because of their involvement in mechanisms which arose before splicing (see section 1.2) and conservation of specific introns in the encoding genes of the Lsms. Lsm6 has been predicted as the most likely 'original' Sm protein which all eukaryal Sm's originated from based on its relatively equal homology with the others; however, this is difficult to confirm [33]. A second round of gene duplication likely generated the canonical Sm core proteins from the Lsm proteins, as each Sm protein has a greater sequence similarity to a corresponding Lsm protein than any of the other Sm core (Lsm1/SmB, Lsm2/SmD1, Lsm3/SmD2, Lsm4/SmD3, Lsm5/SmE, Lsm6/SmF and Lsm7/SmG) [18,32,33]. As the complexity of the splicing regulation and spliceosome itself increased, the Lsm proteins had multiple functional restraints resulting from its role in various functions. Gene duplication of the Lsm proteins removed the restraint from the canonical Sm core proteins and allowed them to become more specialized until they were solely used in splicing. Additional Sm proteins arose after the last eukaryal common ancestor (LECA). Lsm10 and Lsm11 are encoded by animals, fungi, and amoeba but absent from plants; and SmN is only present in mammals [45-47].

4 The Sm Fold and RNA-binding Properties

4.1 Monomers and Oligomers

Crystal structures show that the Sm protomer is highly conserved between domains of life, even with low sequence similarity between bacteria and eukarya/archaea (Figure 2B). Two Sm

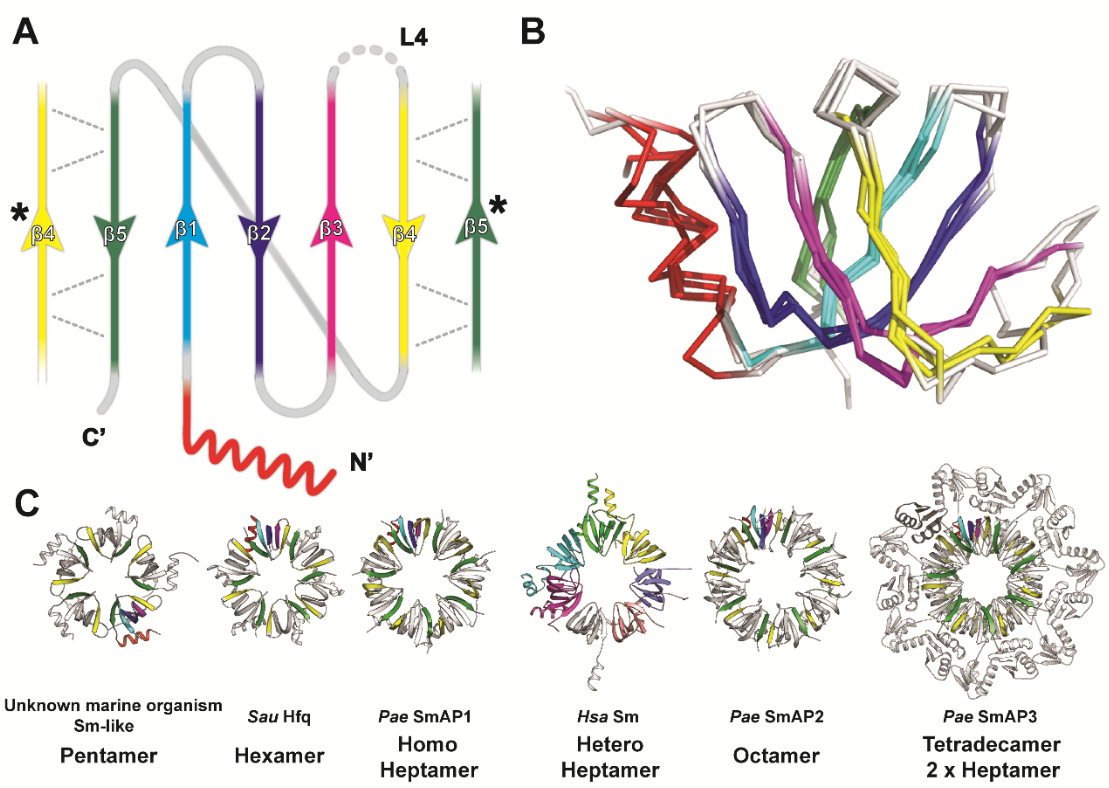


Figure 2. Sm topology, conservation, and oligomeric plasticity. (A) Topology of the Sm fold. The Sm fold consists of a small five stranded β-barrel preceded by a small α-helix. The monomer-monomer interface for oligomer formation is between the β4 and β5 strands. Loop L4 is variable, and can be quite extended in eukarya and archaea, but shortened in bacterial hfq. (B) Backbone overlay (C_α) of Sm monomers from bacteria (*S. aureus* Hfq), archaea (*P. aerophilum* SmAP1 and SmAP2), and eukarya (*H. sapiens* SmD3) demonstrate the conserved nature of the Sm fold across the domains. (C) While the Sm fold is highly conserved at the level of the monomer, the Sm family shows immense plasticity in its oligomerization, existing as a pentamer (3BY7), hexamer (1KQ1), heptamer (1I8F), octamer, and tetradecamer (1M5Q).

monomers from even distantly related species will usually have between 1-2 Å backbone root-mean-square deviations (RMSDs). The Sm fold is a highly conserved ~70 residue structure consisting of a small highly bent, five-stranded β-barrel preceded by an α-helix (Figure 1A, Figure 2A) which forms into a toroid or ring multimer (Figure 1A, Figure 2C). Loop 3 is on the α-helix face which is commonly known as the *proximal* or L3 face, while Loop 4 is on the *distal* or L4 face Figure 1A). Highly conserved glycines located in β-strands 2, 3, and 4 permit the structural variability required for Sm proteins to form the bent fold [48]. Two conserved motifs, Sm 1 (~32 residues) and Sm 2 (~14 residues), are recognized (though there is some dispute on whether they should be considered distinct motifs) and separated by a non-conserved linker [48,49]. The Sm1 motif consists of β-strands 1, 2, and 3 and is the most conserved section between domains. The Sm2 motif contain β-strands 4 and 5, and form the monomer···monomer interface of the multimer [50]. The Sm2 motif varies between bacterial Hfq and the eukaryal Sm/Lsm proteins, possibly leading to the difference in oligomeric form. Loop 4 (L4) contains a variable region that can be lengthy (up to 25 residues) in archaeal and eukaryal Sm proteins but is extremely short in the bacterial Hfq. In addition to the Sm fold, many Sm proteins are known to have a C-terminal domain which can vary drastically in both electrostatics and size between orthologs, ranging from a few residues (e.g. *Aquifex aeolicus*) to over three times the length of the core Sm fold itself (e.g. *Moraxella catarhallis*) [51].

While the Sm fold is highly conserved (Figure 2B), the Sm family shows immense plasticity in its oligomer formation (Figure 2C). Most Sm proteins have been found as either hexamers (bacterial Hfq) or heptamers (eukaryal Sm proteins and SmAPs); a pentamer and an octamer structure have been found (in the latter case a dimer combined with four-fold crystallographic symmetry could mean the octamer is an artifact) [24-26,52-55]. Sm proteins rings are usually ≈ 60 to 80 Å in diameter, 35 Å in height, with a variable-width pore. Hfq forms highly stable, heat resistant homo-hexamers, while the eukaryal canonical Sm core proteins (SmB, SmD1, SmD2, SmF, SmE, SmG, SmD3) form dimers or trimers until bound to their target RNA (formation is assisted by chaperone proteins *in vivo*, including SMN; but is spontaneous *in vitro* when the hetero-protomers are incubated with corresponding U snRNA [20,56,57]). The eukaryal Lsm proteins and SmAPs are intermediates between Hfq and the canonical Sm core proteins, with the eukaryal Lsm proteins forming stable hetero-heptamers in two forms, one containing Lsm1 through Lsm7 (Lsm1-7); and the other exchanging Lsm1 for Lsm8 (Lsm2-8). Most SmAPs oligomerize as highly thermostable homo-heptamers, though one has been found to transition between a hexamer and a heptamer in a pH and RNA-dependent manner [58]. *Pyrobaculum aerophium* SmAP3 was identified and structurally characterized with a large stable C-terminal domain which interacts with the same domain of another SmAP3, forming a tetradecamer that consists of two heptamers face-to-face surrounded by interlocking C-terminal domains (Figure 2B) [59]. Differences in oligomeric variation and the RNA dependence of oligomerization, result in distinct binding profiles between the bacterial Hfq and the eukaryal Sm core proteins.

4.2 Eukaryal Sm/Lsm···RNA Interactions

The canonical Sm core proteins bind to the uridine-rich Sm sites of U snRNAs. U snRNAs thread around and through the pore of the Sm ring [57,60,61]. A recent structure of the U1 snRNP demonstrates the specific binding pockets necessary to guide the U1 RNA through the pore, as shown in Figure 3 [62]. The first nucleotides of the Sm binding site that contact the Sm ring are Adenine-125, which contacts SmD2 and SmF, followed by Adenine-126 which binds to SmE. Uracil-127, Uracil-128, and Uracil-129 comprise a uracil patch that binds SmG, SmD3, and SmB, respectively. Steric clashes prevent Guanine-130 from fitting in the binding pocket and instead it lies above Uracil-124, while its purine base contacts SmB. The nucleotide corresponding to this Guanine-130 on the U4 snRNA Sm site is a uracil, which fits into the uracil binding pocket in SmD1, expanding the uracil-rich patch. Uracil-131 binds in a pocket on SmD2 and Guanine-132 begins the descent into the pore, making several interactions with the ribose phosphate backbone of neighboring nucleotides, and stacking with residues on Loop 2 of SmD2 and SmD1. Guanine-132, Guanine-133, and Uracil-134 traverse the pore, exiting on the Loop 4 face. In summary, the Sm binding site of U snRNAs begins at SmE and goes in a counter-clockwise manner (viewed from above, *proximal* to the α-helix face) around the pore with the snRNA binding SmG, SmD3, SmB, SmD1, SmD2 (in that order) and exiting on the far face at SmF [56]. Each nucleotide on the Sm binding site of the U snRNAs is specific for each Sm protomer. While there are currently no structures detailing binding of U6 snRNA to Lsm 2-8, based on homology, it is expected that U6 snRNA binds to Lsm5 first, then counter-clockwise to Lsm7, Lsm4, Lsm8, Lsm2, Lsm3, exiting at Lsm6.

4.3 Bacterial Hfq···RNA Interactions

Both Hfq and eukaryal Sm/Lsm are known to bind uracil-rich single-stranded RNA (ssRNA) near the pore on the face *proximal* to the α-helix (L3 face). The eukaryotic heptamer has a wider pore than the hexameric Hfq, which allows the ssRNA to thread through the pore, rather than simply running along the face as in Hfq. Because ssRNA cannot traverse the Hfqpore, this Sm has distinct binding faces. Hfq has at least four distinct surfaces that interact with macromolecules: *i)* L3 face, *ii)* L4 face, *iii)* lateral rim, and *iv)* the C-terminal domain (Figure 4A) [55,63]. In addition to the binding of U-rich RNA to the *proximal* (L3) face, adenine-rich sequences bind on the *distal* (L4) face, and non-specific UA-rich RNA have recently been found to bind on the *lateral* rim (Figure 4B).

Uracil-rich binding on the L3 face is constrained to a ring directly around the pore, with a one-to-one nucleotide to monomer ratio [64,65]. The uracil-binding site is formed by amino acid residues of two adjacent monomers (Figure 4C). Uracil stacks with a conserved aromatic residue, (Phe-42 in *E. coli*) while the O2 and O4 atoms of uracil hydrogen bond with residues from the

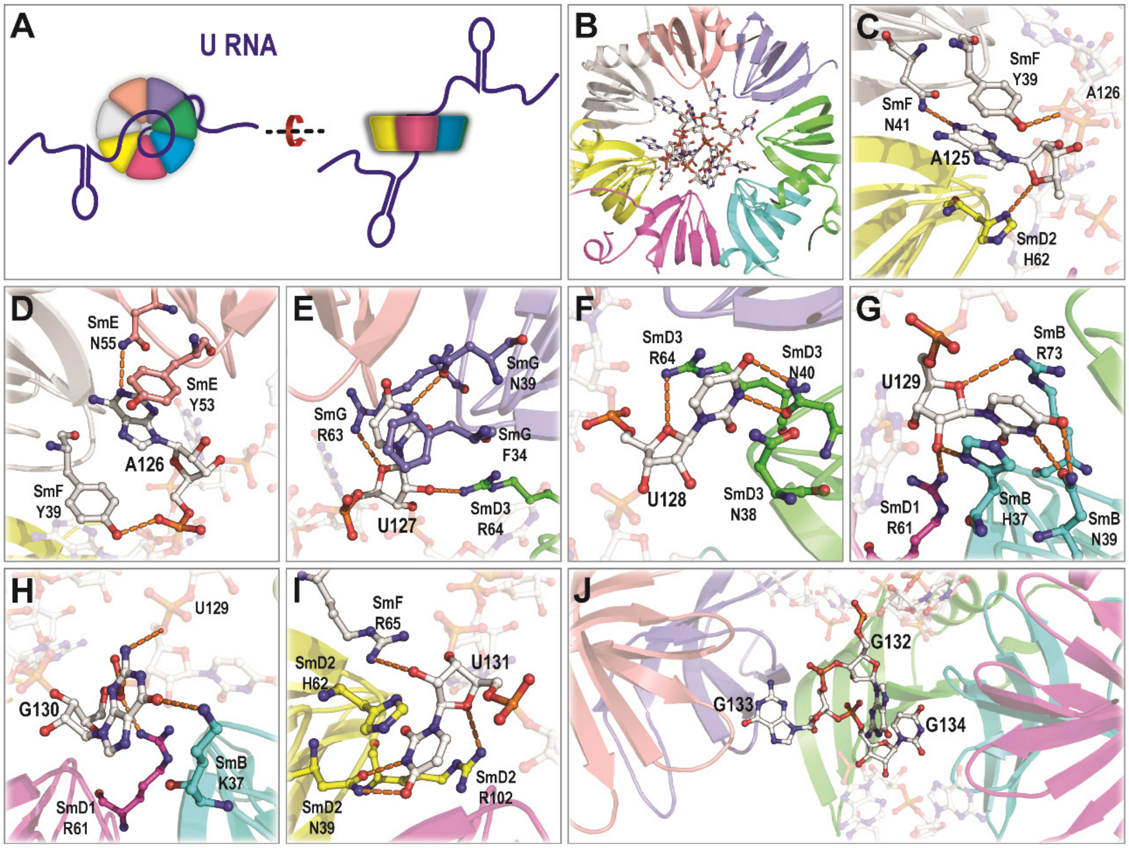


Figure 3. Canonical Sm core protein···RNA interactions. (A) Canonical Sm core proteins bind U RNA around the L3 face with the U RNA contacting each subunit as the RNA threads through the pore and exits on the L4 face. The nucleotide binding pockets are many times located between adjacent monomers, contacting both. (B) The U1 snRNP core demonstrates the specific canonical Sm core protein···RNA interactions necessary to guide the RNA through the pore. (C) Adenine-125 with SmD2 and SmF, (D) Adenine-126 and SmE, (E) Uracil-127 and SmG, (F) Uracil-128 and SmD3, (G) Uracil-129 and SmB, (H) Guanine-130 is positioned above the pore contacting SmD1 as the RNA begins to enter the pore, with (I) Uridine-131 completing the spiral around the pore, returning to SmF. (J) Guanine-132, 133, and 134 traverse the pore exiting on the L4 face.

adjacent subunit (His-47), supplying uracil specificity (Figure 4D) [27,55,64-67]. On the *distal* or L4 face, Hfq binds to an adenine rich repeating sequence of Adenine-Purine-Any Nucleotide (ARN-repeat) (Figure 4E) [68]. The binding on the L4 face is farther from the pore and has a three-to-one nucleotide to monomer ratio. Interestingly, for the L4 face, in the N-site the nucleotide is oriented with the base-pairing edge exposed to the solvent, though currently no base pairing has been observed (Figure 4H). The A-site is located between adjacent monomers of Hfq. *E. coli* crystal structure shows hydrophobic contact between the adenine and Leu-32 in addition to hydrogen bonds formed between adenine and Gln-52, Leu-32, and Gln-33 (Figure 4F). R-site binding is highly dependent on a conserved aromatic residue (Tyr-45 in *E. coli*) which stacks with the base, though additional hydrophobic interactions from other residues (Thr-61 and Gln-52) contribute (Figure 4G) [68-71].

While most structural studies have focused on Gram-negative bacteria, the structure of the Gram-positive *Bacillus subtillus* Hfq bound with $(AG)_3A$ RNA, determined in 2011, shows notable differences in the A-rich binding site (Figure 4I) [72]. *B. subtillus* has a much lower affinity for A_{18} RNA than *E. coli* but a higher affinity for AG-repeat RNA. In the crystal structure of *B. subtillus* Hfq bound to $(AG)_3A$ RNA there appears to be only two binding pockets per protomer instead of three. The A-site is the same as the R-site for Gram-negative, with the adenine stacking between two highly conserved phenylalanines (Phe-24 and Phe-29), with additional contribution from hydrogen bonds between the adenine and Ser-60 and Thr-61 (Figure 4J). The guanine site appears to only form a hydrogen bond with Arg-32, which is highly conserved in Gram-positive bacteria (Figure 4K) [72].

The adenine-rich and uridine-rich binding surfaces, described above, correspond closely to elements of mRNA and sRNA, respectively. Many sRNAs contain a uracil-rich stretch at the 3’ end, which is a common feature for RNAs produced via Rho-independent termination pathways [73]. mRNAs usually have an adenine-rich region in their 3’ untranslated region (UTR) [74,75].

The lateral rim-binding mode recently observed in *E. coli* Hfq consists of a conserved aromatic (Phe-39 in *E. coli*) and a positive patch of residues (Arg16, Arg17, and Arg19) on the edge of the L3 face of each monomer [76,77]. This lateral rim has been shown to be necessary for many biological functions, including *RybB* sRNA binding [78]. The lateral rim appears to play a role in binding and stabilization by contacting UA-repeat regions of either mRNA or sRNA. Very recent work has shown that the lateral rim can serve as a transition junction for sRNA···mRNA pairing, allowing either the sRNA or mRNA to bring base pairing elements into proximity [78]. Others have suggested that the lateral rim plays a role in Hfq cycling, allowing a site for RNA association and dissociation [79].

4.4 SmAP···RNA Interactions and Homology

SmAPs are closer in sequence and structure to the eukaryotic Sm/Lsm proteins than the bacterial Hfq. Structural alignments have shown *Pae* SmAP1 monomer and dimer structures to be almost identical to the SmD1-SmD2 and SmD3-SmB heterodimers and most previous SmAP structures have been heptamers, though they form stable homo-oligomers, similar to the bacterial Hfq. SmAPs have been shown to bind RNA and one of the first SmAP structures shows binding to uridine on the L3 face similar to eukaryal Sm and Hfq, but whether RNAs bound to SmAPs then thread through the pore (eukaryal Sm) or just stay along the face (bacterial Hfq) is unknown [24-26,80]. Crystal structures from *Archaeoglobus fulgidus* bound to U_3 RNA, *Methanobacterium thermautotrophicum* bound to single UMP nucleotides, and *Pyrococcus abyssi* bound to UMP nucleotides detail a conserved uracil binding pocket on the L3 face of SmAPs, analogous with eukaryal Sm/Lsm and bacterial Hfq pockets (Figure 5A, 5B, 5C). The ligands bind in a conserved crevice near the pore, with the uracil base intercalated between conserved arginine and histidine residues (Figure 5C, 5D, 5E)[24,80,81].

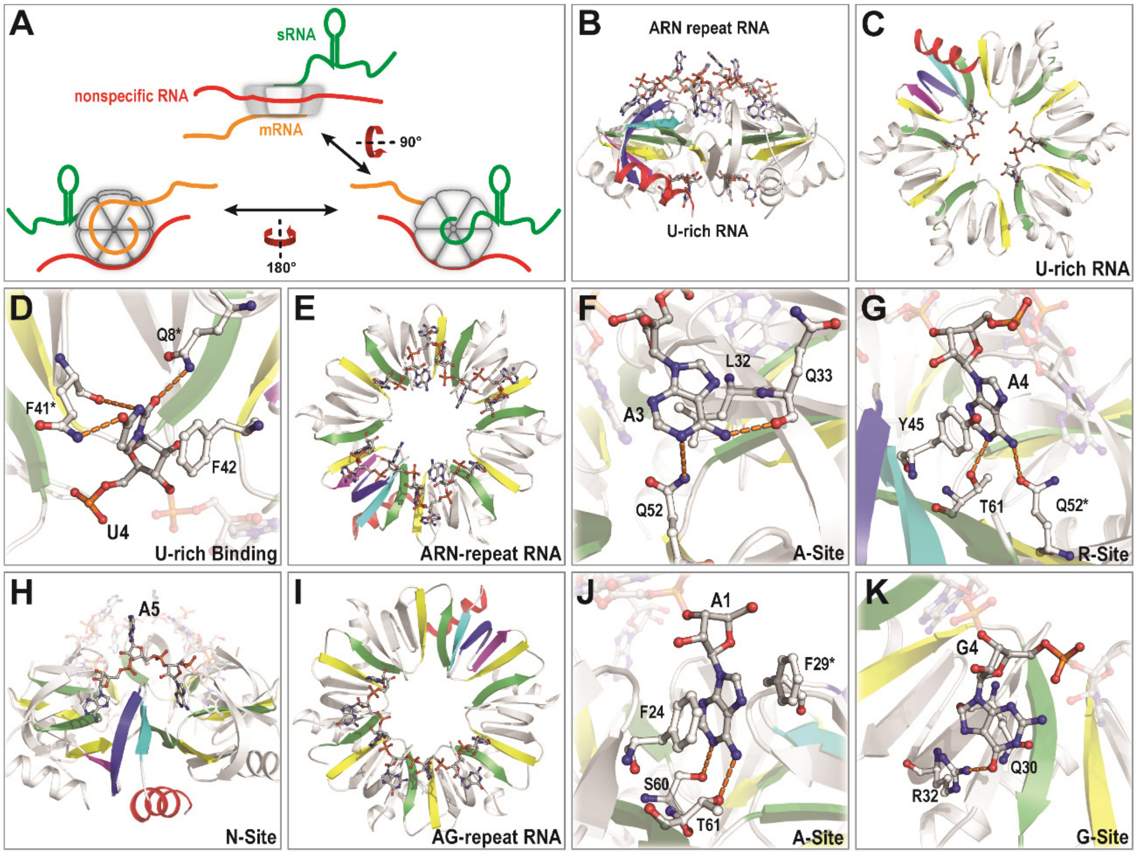


Figure 4. Hfq···RNA interactions. (A) The L3 face of Hfq binds uridine-rich RNA analogous to the eukaryal Sm proteins. The RNA cannot thread through the pore, allwing the L4 face to function independently, binding adenine-rich RNA. The lateral rim has recently been shown to bind UA-rich RNA. (B) Hfq can bind RNA simultaneously on both faces. (C) The U-rich binding site is similar to the eukaryal Sm site, (D) and involves a conserved aromatic residue (F42 in *E. coli*) with additional hydrogen bonding for specificity. (E) Hfq from Gram-negative bacteria displays an ARN-repeat binding motif on the L4 face. (F) The A-site is near the edge of the monomer, while the (G) R-site is between monomers, stacking with a conserved aromatic residue (Y42 in *E. coli*). (H) The N-site is above the protein and lacks specificity. (I) Hfq from Gram-positive bacteria bind an AG-repeat on the L4 face. (J) The A-site is the same as the R-site for gram-negative but with additional hydrogen bonding from S60 supplying A-specificity. The (K) G-site is located at the same location as the A site in ARN-motif, but does not have the specificity.

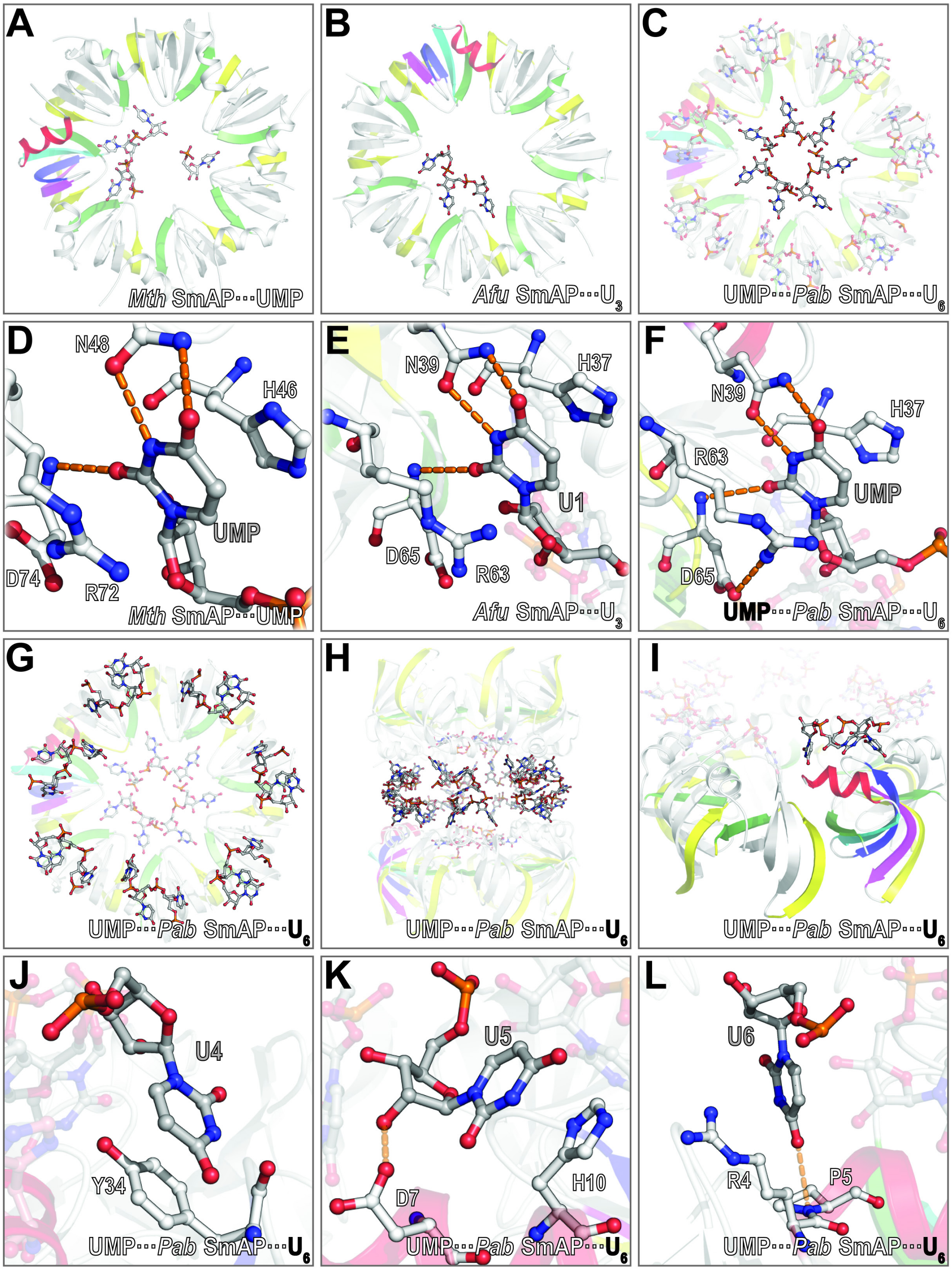

A
Mth SmAP···UMP
B
Afu SmAP···U$_3$
C
UMP···Pab SmAP···U$_6$
D
N48
H46
UMP
D74
R72
Mth SmAP···UMP
E
N39
H37
U1
D65
R63
Afu SmAP···U$_3$
F
N39
H37
R63
UMP
D65
UMP···Pab SmAP···U$_6$
G
UMP···Pab SmAP···U$_6$
H
UMP···Pab SmAP···U$_6$
I
UMP···Pab SmAP···U$_6$
J
U4
Y34
UMP···Pab SmAP···U$_6$
K
U5
D7
H10
UMP···Pab SmAP···U$_6$
L
U6
R4
P5
UMP···Pab SmAP···U$_6$

Figure 5. SmAP···RNA interactions. Only a few structures are available demonstrating SmAP binding to RNA. All show conserved U-rich binding on the L3 face in the pore. Structures from (A) *Methanobacterium therautotrophicum* bound to single UMP nucleotides, (B) *Archaeoglobus fulgidus* bound to U_3 RNA, and (C) *Pyrococcus abyssi* bound to UMP detail a conserved uracil binding pockets on the L3 face of SmAPs, analogous with the eukaryal Sm/Lsm and bacterial Hfq pockets. (C, D, E) The ligands bind in a conserved crevice near the pore, with the uracil base intercalated between conserved arginine and histidine residues. (G) *P. abyssi* SmAP was also co-crystallized with U_6 RNA, which binds on the lateral rim close to the L3 face with the RNA (H) bridging two *P. abyssi* SmAPs. (I) U_6 runs across the lateral rim, with (J) U-4 π-stacking Y34, in the same location as the conserved aromatic for Hfq's lateral binding site (K) U-5 π-stacking H19, and (L) U-6 hydrogen bonding P5.

P. abyssi SmAP structure is also bound to a small U_6 RNA strand on the lateral-rim binding surface previously mentioned for Hfq (Figure 5G, 5H, 5I). The U_6 strand bridges two *P. abyssi* SmAPs, with three nucleotides bound to each SmAP. Interestingly, this means that the RNA extends in different directions on each SmAP (5' to 3' on one SmAP and 3' to 5' on the other) and the U_6 RNA does not make contact with any residues farther down the outer rim, where the arginine-patch in Hfq is located. The lateral rim-binding site in *P. abyssi* consists of U-3/U-4 π-stacking with Tyr-34 (equivalent position as Phe-39 in *E. coli*), U-2/U-5 π-stacking with His-10 located on the α-helix, and U-1/U-6 hydrogen bonding the backbone amide of Pro-5 (Figure 5J, 5K, 5L). Whether this lateral binding-rim is conserved in SmAPs is unknown. Many other details on SmAP binding are lacking, including whether it binds adenine-rich RNA, or whether the uracil-rich RNA extends through the pore.

5 Physiological Roles of the Sm Family

5.1 Non-coding RNA Biology

The biological roles of the Sm family can be closely linked with the roles of noncoding RNAs (ncRNAs) within the three domains of life. The RNA world model for early life holds that information/genetic storage and functional components entirely devised of RNA (no DNA or proteins) [82-84]. Over time the introduction of proteins and DNA allowed more diverse/expanded functionality and more stable information/genetic storage, respectively. RNA was initially considered to have been relegated to an intermediate role, with its main function taking the information from DNA (mRNA) and assisting in the creation of functional proteins (rRNA and tRNA); but over the last 30 years it has become apparent that RNA plays many vital roles in the cell beyond direct involvement in protein production. While many catalytic RNAs were discovered previously (snRNAs, ribosome, *etc*), in the early 1990s regulatory RNAs were first discovered in the eukarya, demonstrating that even small RNA strands play an important role in cell function

[85]. Eukaryal non-coding RNAs are involved in many crucial pathways: RNA processing/regulation (H/ACA box and C/D box small nucleolar RNAs (snoRNAs)), splicing (Uridine-rich small nuclear RNAs (U snRNAs)), DNA replication (Y RNAs), genome defense (Piwi-interaction RNAs (piRNAs)), chromosome structure (telomerase RNA), and gene regulation (microRNAs (miRNAs), small interfering RNA (siRNA)) [see review 86 for citations]. In higher eukarya it is estimated that >50% of genes are regulated by small non-coding RNAs [87].

Bacterial regulatory RNAs were found in the early 1980s in non-chromosomal genetic elements and were dubbed small regulatory RNAs (sRNAs) or noncoding RNAs (ncRNAs) [88,89]. Bacterial sRNAs are analogous in function to eukaryal miRNAs but vary significantly in length (between 50 and 250 nucleotides) and structure (can contain multiple stem-loops), while all miRNAs are processed to be about 22 nucleotides. sRNAs have been found to bind and directly affect specific proteins, but the majority target specific mRNAs in an antisense manner. Since their discovery, over a 100 sRNAs have been identified in bacteria, many active in high impact areas such as drug resistance and virulence [90,91]. mRNA targeting sRNAs in bacteria fall into two categories, *cis*- or *trans*-encoded. *Cis*-encoded sRNAs are encoded on the antisense strand of the target gene and will have perfect base-pairing complementary with their target mRNA, while *trans*-encoded are in an intergenic region and have imperfect pairing. Conserved 5’ regions in homologous sRNAs are commonly referred to as ‘seed’ regions, and serve to recognize the mRNA target [92].

In the early 2010s, regulatory RNAs were bioinformatically predicted and identified in archaea, though less is known about these than in bacteria and eukarya [93]. Small guide RNAs (related to and named after snoRNAs) were the first ncRNAs found in archaea and are involved in the modification of RNA nucleotides on ribosomal RNA [94,95]. A number of *cis*-encoded and *trans*-encoded sRNAs have also been identified, but their functions are currently unknown; they may be involved in gene regulation similar to the bacterial sRNAs [96]. Some of the sRNAs discovered in archaea contain short regions of open reading frames (ORFs), seemingly combining the functions of both mRNA and regulatory RNA [97].

5.2 The Bacterial Hfq

5.2.1 A Global Post-transcriptional Regulator

The perception of Hfq as the core component of a global post-transcriptional regulatory network arises from its role in *chaperoning* imperfect base-pairing interactions between *trans*-encoded sRNAs and their mRNA targets. The imperfect base-pairing of *trans*-encoded sRNAs mean that many of them do not bind their target mRNA without Hfq present. As mentioned previously, these sRNAs are heterogeneous in size and structure, and are involved in diverse processes including quorum sensing, stress response, virulence. Hfq···sRNA interactions can have a variety of downstream effects: e.g. Hfq can inhibit (down-regulate) translation by assisting sRNA in base-pairing mRNA near the ribosomal binding site (RBS) preventing ribosomal binding and translation (Figure 6A) [98-100]. Conversely, if the mRNA forms a secondary structure that blocks the RBS, Hfq can promote translation by unmasking the RBS. sRNA, assisted by Hfq, binds near or on the secondary structure blocking the RBS, melting the blocking secondary structure and freeing the mRNA for translation (Figure 6B) [101,102]. Hfq plays a role in the degradation of RNAs either stabilizing sRNA, by protecting from RNases (including RNase E) (Figure 6C), or promoting degradation by assisting sRNA and mRNA binding, in turn signaling for degradation by RNase E (Figure 6D) [103,104]. Hfq is involved in the poly(A) polymerase (PAP) degradation pathway by binding mRNA and recruiting PAP which polyadenylates the mRNA targeting it for exoribonuclease (Exo) 3’-to-5’ degradation (Figure 6E) [105]. The different modes of Hfq action are dependent on the specific sRNA···mRNA pairing.

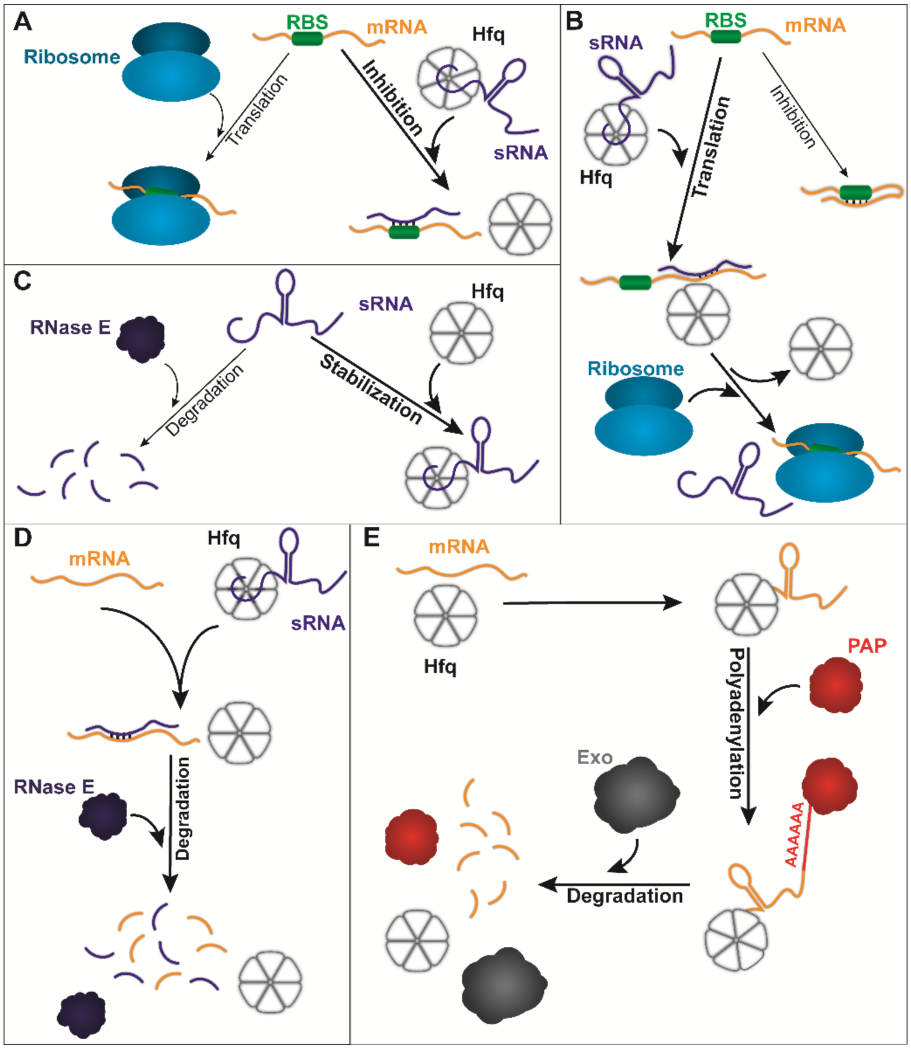


Figure 6. Hfq is an RNA chaperone, guiding interactions between various RNAs. While the *chaperoning* function is generic, the specific RNA targets determine function. (A) Hfq can assist sRNA in binding mRNA, blocking the ribosomal binding site (RBS) leading to suppression of expression. (B) Hfq and sRNA can relieve an internal structural element that is blocking the RBS on an mRNA, leading to upregulation of expression. (C) Hfq can stabilize and protect RNA from degradation, or (D) can promote degradation by recruiting RNase E. (E) Hfq may stimulate polyadenylation though interactions with poly(A)polymerase (PAP) leading to degradation.

Initial *in vitro* analysis of Hfq function considered the ternary sRNA···Hfq···mRNA complex as the biological regulatory unit [29,74,106]. However, the biologically stable unit has been revised because of the excess amount of RNA available compared to Hfq [79,107]. *In vivo* there is constant cycling of RNA on and off of the Hfq, and Hfq can be considered as more of a 'catalyst', facilitating the formation of sRNA···mRNA complexes, and being cycled back for further rounds of pairing [108]. As mentioned previously, Hfq can bind both sRNA and mRNA simultaneously on either face (or lateral rim), this close contact of mRNA and sRNA reduces the entropic cost for pairwise encounters between RNAs. sRNAs are not limited to an individual mRNA target, some sRNAs have been found to target multiple mRNAs, and similarly, multiple sRNAs will target the same mRNA. It is predicted in some bacteria that 40% or more of genes are regulated by sRNA (both *trans*- and *cis*-acting) [109]. In addition to their interactions with RNA, Hfq is also predicted to interact with various proteins. The ribonuclease RNase E is a prime candidate for interaction with Hfq [110,111], and is thought to interact directly with Hfq, instead of indirectly through a Hfq···RNA···protein bridge. This line of enquiry into Hfq···protein interactions is ongoing.

5.2.3 Hfq-based Responses to Environmental Conditions

As mentioned previously, many of the initial phenotypes associated with Hfq were a result of their role in regulating subunits of RNA polymerase known as sigma (σ) factors. σ factors are responsible for direct responses to stimuli by focusing gene expression towards specific regulons. σ^{70} is known as the primary σ factor, and is responsible for transcribing most of the genes in a cell [112]. Alternative σ factors allow bacteria to quickly react to environmental factors by targeting specific regulons [113,114]. Regulation of *rpoS* mRNA, which encodes the stationary phase or stress σ^{s} factor, is one of the most studied Hfq-dependent σ factors [8,115]. Depending on environmental conditions, multiple Hfq-dependent sRNAs can interact with *rpoS*; *i)* DsrA in response to temperature stress, RprA for envelope stress, and OxyS for oxidative stress [100-

102,116-118]. DsrA and OxyS down-regulate translation of *rpoS* while RprA up-regulates translation [119]. Hfq-dependent regulation of *rpoS* (σ^s) and *rpoE* (σ^e, extracytoplasmic response to misfolding of proteins in the cell envelope) has been detailed in *Salmonella enterica*, *Pseudomonas aeruginosa*, and *Klebseialla pneumonia* [8,120,121]. A significant percentage of the genes regulated by Hfq in *P. aeruginosa* and *K. pneumonia* belong to either the *rpoS* pathway (66% in *P. aeruginosa*, 17.3% in *K. pneumonia*) or *rpoE* pathway (19.5% *K. pneumonia*) [122,123]. Within bacterial species lacking *rpoS* and *rpoE* (e.g. *Francisella tularensis*) Δ*hfq* still has a deleterious pleiotropic affect. *F. tularensis* Hfq was found to regulate 6% of all *F. tularensis* genes, and a Δ*hfq* mutation caused almost identical pleiotropic affect as Δ*hfq* mutant studies in *rpoS* and *rpoE* encoding bacteria [124]. In many bacteria Hfq also plays a role in regulating iron metabolism. A conserved sRNA known as *rhyB* regulates iron metabolism, depends on Hfq for its activity [125,126]. In *E. coli* when environmental iron is low, suppression of Fur (global iron uptake regulator) is removed, and *rhyB* expression is induced, which represses translation of ferrous proteins allowing the levels of intracellular iron to increase [125,127].

5.2.4 Hfq and Bacterial Pathogenicity

Bacteria can fall into three general categories, *i)* primary pathogens cause disease simply as a result of their presence/activity, *ii)* opportunistic pathogens can cause disease only in a host with a compromised immune system, and *iii)* non-pathogenic bacteria do not result in disease states [128]. A combination of factors can lead to a disease state including the current state of the host, quantity of infecting bacteria, localization and route of entry, and virulence factors expressed by the bacteria to evade the hosts immunosuppression. Hfq, while not present in all pathogenic bacteria, is considered a virulence factor because of the phenotypes associated with Δ*hfq* mutants, including the previously mentioned growth restraints. To date, Δ*hfq* mutants of 34 different pathogenic species have elucidated Hfq's role in virulence. Hfq plays a role in many virulence pathways including biofilm formation, quorum sensing, polysaccharide biosynthesis, secretion systems, and outer membrane protein expression.

Biofilm formation is an important pathogenic trait for many bacteria as biofilms provide protection from immunosuppression [129]. Biofilms are organized colonies of bacteria clustered together to form micro-colonies which secrete an extracellular matrix attaching the colony to a surface. The extracellular matrix consists of a variety of secreted macromolecules including DNA, proteins, and polysaccharides (referred to as exopolysaccharides or EPS). Hfq-dependent sRNAs are required for biofilm formation in a number of pathogens (both primary and opportunistic) including *E. coli*, *Stenotrophomonas maltophilia*, *S. enterica*, *Proteus mirabilis*, and multiple *Burkholderia* species [130-134]. While EPSs have a variety of functions, they are studied mainly for their role in forming biofilms, and protective outer capsules. The *Burkholderia cepacia* EPS known as cepacian plays a role in phagocytosis evasion[135]. *B. cepacia* Δ*hfq* mutants showed an almost 50% decrease in cepacian production. *K. pneumoniae* Hfq modulates the production of capsular polysaccharide most likely through downregulation of the transcriptional activators of the capsular genes *rmpA* and *rcsA* [34]. Another class of polysaccharides important for bacterial virulence are lipopolysaccharides (LPSs). LPSs are composed of a polysaccharide chain with a lipid base. LPSs are well-studied virulence factors, in fact they were originally identified as causing a host immune response in the late 1890s by Dr. Richard Pfeiffer [136]. *S. enterica* and *E. coli* Δ*hfq* mutants show an increase in lipopolysaccharide core completion, suggesting Hfq is involved in suppression of this system possibly an effect of the Hfq-dependent sRNA *mgrR* [131,137].

Hfq has recently been identified as a factor in the biogenesis and localization of outer membrane proteins (OMPs). OMPs are highly regulated because of their roles maintaining membrane integrity and their roles as gatekeepers of the cell. OMPs are often the targets of immunosuppression because OMPs are the main environmentally exposed proteins, similar to LPSs [138]. Δ*hfq* strains of *M. catarrhalis, B. militensis*, and *S. enterica* Typhimurium exhibit upregulated OMPs [51,137,139,140], increasing their susceptibility to immunosuppression. Genomic studies have identified a variety of sRNAs that down-regulate OMP, including many previously shown to require Hfq for function (*MicA*, *MicF*, *MicC*, *MicM*, and *CyaR*) [119,131,141-143]. A number of other virulence factors have Hfq-dependent regulation including: *Bordatella pertussis* adenylate cyclase toxin, pertussis toxin, and filamentous hemagglutinin, *Haemophilus ducreyi* LspB and LspA2, autotransporter DsrA, *P. aeruginosa* quorum sensing regulator and exotoxin, *Vibrio parahaemolyticus* hemolysin, and *Yersinia enterocolitica* enterotoxin [144-148].

5.2.5 Hfq Expression

While Hfq regulation of gene expression itself is an ever growing field of research, very little is known about the regulation of the *hfq* gene. *E. coli* studies suggest that Hfq suppresses its own expression [7]. Hfq binds to the 5' UTR region of the *hfq* mRNA, exposing an RNase E binding site encouraging degradation [149]. Recent studies have shown that Hfq expression is dependent on a number of regulators beyond itself (e.g. *LetA*, and σ^s) [150]. In many bacteria, Hfq expression is dependent on the growth stage. *P. aeruginosa* Hfq levels are 200% greater in the stationary phase than the exponential phase [123]. This is not universal though, there is no change in Hfq abundance at different growth phases in *N. meningitides* [151]. Interestingly, recent bioinformatic studies have found multiple putative *hfq* genes in some bacteria. *B. cepacia* appears to encode two Hfq paralogs, *hfq1* and *hfq2*, while *Bacillus anthracis* appears to have three Hfq paralogs, two encoded on the chromosome and one on a plasmid [34,152]. The *B. cepacia* Hfqs are phase dependent, with one Hfq having a higher abundance in the exponential phase, the other in the stationary phase. In *B. anthracis* one of the chromosomal encoded Hfqs and the plasmid Hfq have traditional sRNA *chaperone* roles, with growth phase dependent expression. The third Hfq's (chromosomal encoded) function is unknown, but it does not appear to play a role as a *chaperone* or have growth phase dependent expression [152]. Hfq is not equally dispersed throughout the cell, instead Hfq appears to localize in subcellular pockets. Hfq is found throughout the cytoplasm, but has a much higher abundance in nucleoid, in fact is one of the most abundant proteins in the nucleoid [153,154]. In addition, electron microscopy has revealed a high concentration of Hfq near the outer membrane, though the reason is still unknown [155].

5.3 Eukaryal Sm/Lsm Proteins Function as Structural Scaffolds

5.3.1 Splicing Overview

Compared to bacteria, eukarya encode a variety of Sm homologs that form hetero-rings with different localizations and functions, though all of their functions are based off their roles as *scaffolds*. The majority of studies in eukarya have focused on the seven canonical Sm core proteins: SmB, SmD1, SmD2, SmD3, SmE, SmF, and SmG. SmB has a splice variant called SmB' and a neuronal variant called SmN. Canonical Sm core proteins along with the U small nucleolar (snRNAs) are the major components of the large (4.8MD) macromolecular machines known as the major and minor spliceosome [17,156]. Spliceosomes excise non-coding introns from precursor mRNA (pre-mRNA) and splice the resultant exons together to form mature mRNA for translation [157,158]. Most human genes contain multiple introns that need to be removed, and many mRNAs can be spliced in *alternative* manners, combining exon components of the same mRNA in different orders resulting in multiple protein [157,158]. Alternative splicing is one of the major ways that the relatively small genome of some organisms can display such variety of phenotypes [159,160]. There are two main classes of U snRNPs, those with the canonical Sm core proteins, and those utilizing Lsm proteins. Spliceosomal U snRNPs containing the canonical Sm core heptamer are U1, U2, U4, U4atac, U5, U11, and U12, and Lsm2-8 forms the core of the U6 and U6atac snRNPs [161].

5.3.2 RNA Splicing

While the splicing mechanism is a fairly simple chemical reaction consisting of two transesterifications, eukarya employ an immense amount of regulation and energy (ATP) to ensure accuracy. The basic reaction is as follows: *i)* the 2'OH of branch point within the intron attacks the first nucleotide of the 5' splice site, forming a loop or lariat structure. *ii)* The 3'OH of the now released 5' exon attacks the last nucleotide of the intron at 3' splice site, sealing the gap and releasing the intron [41]. To control the reaction, eukarya engage a stepwise ATP-dependent assembly of the spliceosome components. *i)* U1 and U2 snRNPs identify and bind to the 5' and 3' splice site, respectively, by base pairing of the U1 and U2 snRNAs to the mRNA forming initial complex E. *ii)* U1 and U2 snRNP are recruited to form the pre-spliceosome complex A, *iii)* which is followed by pre-assembled U4-U6/U5 tri-snRNP, complex B. *iv)* U1 and U4 snRNPs are released (complex B*) and *v)* the first catalytic step occurs freeing one exon end (complex C). *vi)* ATP-dependent rearrangement of complex C causes the second catalytic step, removing the intron and bringing the exons together. *vii)* The excised intron is now released, along with the U2 and U6/U5 snRNPs which re-enter the splicing cycle [156].

The minor spliceosome is functionally analogous to the major spliceosome, but deals with a specific type of rare introns (<0.5% in humans) called U12-introns (major spliceosomal introns are referred to as U2-type) [162]. Both the major and minor spliceosome contain U5 snRNP, but the minor has analogous substitutions for the other snRNP components: U1 is replaced with U11, U2 with U12, U4 with U4atac, and U6 with U6atac [163]. U12-introns are typically found in genes related to information processing pathways (transcription, RNA processing, DNA replication/repair, etc.) but have been identified in other processes (vesicular transport, ion channels, etc.). Interestingly, U12-introns are absent in genes related to energy metabolism and biosynthesis [164].

Because splicing is such a crucial step in the generation of viable mRNAs, any issues with splicing regulation results in serious diseases including retinitis pigmentosa, chronic lymphocytic leukemia, and myelodysplasia [165-167]. Full disruption of the Sm core proteins leads to a complete loss of splicing, which makes the organism non-viable. The minor spliceosome plays a role in development, and defects can lead to developmental disorders such as Peutz-Jeghers syndrome, spondyloepiphyseal dysplasia tarda, and Taybi-Linder syndrome or microcephalic osteodysplastic primordial dwarfism type I (TALS/MOPD1) [168].

5.3.3 Biogenesis of the snRNP Core

Genesis of a U snRNP core is a complicated process involving a host of chaperone proteins, and includes exportation from the nucleus into the cytoplasm then importation back. Most snRNAs are generated in the nucleus by a process similar, but not identical, to mRNA biogenesis (both utilize RNA polymerase II (PolII), but have different promoters and post-transcription processing) [169,170]. Co-transcriptional processing includes capping of the 5' end and cleavage of the 3' end by a large multi-subunit factor called the integrator complex [171]. 3'-end processing is a highly regulated and complicated system that is dependent on three features: snRNA-specific promoter, a *cis*-acting 3'-box element downstream of the cleavage site, and a variety of *trans*-acting factors that interact with the C-terminal domain of polymerase II [172]. Even though snRNAs function in the nucleus, after initial processing precursor snRNAs are exported into the cytoplasm through Cajal bodies [173,174]). In the cytoplasm pre-snRNAs undergo additional processing and are bound by canonical Sm core proteins to form the early snRNP core. Only after snRNA forms the snRNP core with the canonical Sm core proteins are they imported back into the nucleus. It is speculated that requiring an alternate location for final processing prevents immature snRNA from interacting with precursor mRNA substrate [156].

U snRNA binding canonical Sm core partners (Figure 7) involves a host of chaperone proteins. In the cytoplasm, canonical Sm core proteins do not form a typical Sm ring after translation, instead are dimers (SmD1·SmD2, SmD3·SmB) and trimers (SmF·SmE·SmG) [20,57,175]. The canonical Sm core proteins are recruited by the protein Arginine N-methyltransferase 5 (PRMT5) complex, which methylates the C-terminal arginines in SmB, SmD1, and SmD3 [176] (Figure 7A). After methylation, canonical Sm core proteins oligomerize with pICln, an Sm-dimer mimic protein, to form an initial pseudo-Sm ring complex called the 6S complex (Figure 7B) [177,178]. The 6S complex contains SmD1·SmD2, SmF·SmE·SmG and pICln replacing SmD3·SmB [179]. The 6S complex is recruited into the SMN complex with proteins SMN and GEMIN2 forming the 8S complex (Figure 8C). Even though it is named the SMN complex, research now suggests that GEMIN2 is actually the primary architect of the Sm core snRNP, rather than SMN. GEMIN2 binds across multiple canonical Sm core proteins, and can hold a subset of the Sm proteins in a horseshoe shape without SMN [180,181]. Concomitant with 8S complex formation, the mature U snRNA is recognized by a separate GEMIN protein called GEMIN5 and another SMN which are thought to recognize the Sm site and 3’ stem-loop of the U snRNA [182]. SMN·U snRNA·GEMIN5 associates with the 8S complex, expelling pICln. At the same time, the SmB·SmD3 dimer is recruited and the Sm ring is closed around the U snRNA [178].

Much of the mechanism of how this rearrangement works is unknown, even though the specific elements involved have been identified. SMN remains complexed with the now fully formed U snRNP core which is imported back into the nucleus (after 5’ 2,2,7-trimethylguanosine (TMG) capping [183,184] and 3’ trimming by exonuclease of the U snRNA). Within the nucleus, the U snRNP core releases SMN and localizes in Cajal bodies for final processing and assembly [185,186]. The complicated, highly-regulated U snRNP biogenesis pathway stands in stark contrast to the spontaneous assembly of Hfq and Lsm proteins. Interestingly, if the correct components are present, U snRNP cores will assemble spontaneously *in vitro*. The detailed, ATP-dependent [187],

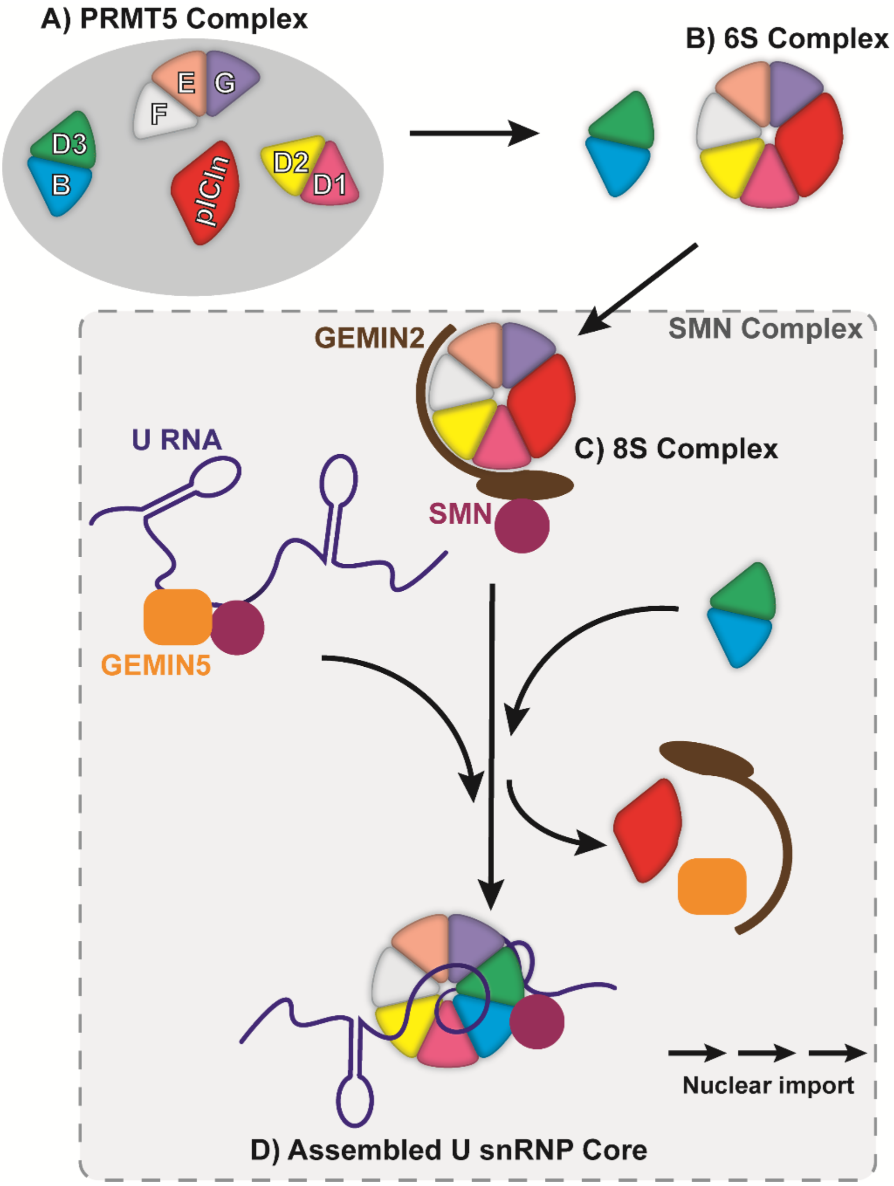


Figure 7. U snRNP core formation. U snRNPs are composed of two main elements, the canonical Sm core proteins and the U snRNA. (A) In the PRMT5 complex, the canonical Sm core proteins form a pseudo Sm-ring with pICln substituting the SmD2·SmD1 dimer, referred to as the (B) 6S complex. (C) The 6S complex is recruited into the SMN complex with the addition of GEMIN2 and SMN to form the8S complex. U snRNA with GEMIN5 and an additional SMN protein reacts with the 8S complex, causing a rearrangement forming (D) the full U snRNP core is, which is then imported into the nucleus.

PRMT5 and SMN requiring, stepwise *in vivo* assembly is thought to prevent canonical Sm core proteins from forming on non-target RNAs and to prevent kinetically trapped intermediates [179,182,188].

Spinal muscular atrophy (SMA) arises from a defection in SMN translation. Normally, there are two SMN genes, *SMN1* which encodes a viable SMN protein, and *SMN2* which differs by a single nucleotide mutation from *SMN1*. Production of SMN from *SMN2* necessitates alternative splicing, with only a fraction becoming viable [189]. In SMA, homozygous deletion of *SMN1* causes a large drop in the abundance of SMN (if both *SMN1* and *SMN2* are impaired than the organism is not viable [190]). Why the drop in SMN abundance leads to muscle atrophy is not currently understood, but splicing defects are seen especially in the late stages of the disease, most likely from the limited amount of Sm core···U snRNA complexes being formed [191-193].

U6 snRNA deviates significantly from the other U snRNAs by being transcribed by RNA polymerase III and containing a γ-monomethyl cap [194]. In addition, U6 snRNA is the only snRNA to never leave the nucleus [195]. The Lsm2-8 ring is assembled in the cytoplasm in the absence of U snRNA, and is transported into the nucleus to bind U6 snRNA and form the U6 snRNP core without the necessity for chaperone proteins. The simplicity of U6 snRNP formation compared to the other U snRNPs could be the result of the different locations of the Sm/Lsm binding sites, which are near the 3’ termini of U6 snRNA (and thus more accessible), but are located in the middle of the other U snRNAs [168].

5.3.4 Other Roles of Sm Proteins in RNA Processing

While splicing is the most well-studied function of eukaryal Sm proteins, they are involved in a variety of other RNPs. Additionally, some U snRNPs have alternative functions. U1 snRNP was found to protect premature cleavage/polyadenylation by binding to mRNA and preventing the binding of the cleavage and polyadenylation specificity factor (CPSF), which cleaves mRNA at the 3' end. The U6 snRNP appears to also assist in degradation of pre-mRNAs and mRNAs in the nucleus [196]. Lsm2-8 can form the U8 snRNP with U8 snRNA. U8 snRNP appears to process ribosomal and transfer RNA [197,198].When U6 snRNA is complexed with Lsm1-7 instead of Lsm2-8, it localizes in the cytoplasm instead of the nucleus, and is involved in degrading mRNA ribonucleoprotein complexes, assisting in the turnover of mRNA [199,200]. Additional Sm rings in eukarya include the Sm10/Sm11 ring, where Lsm10 and Lsm11 replace SmD1 and SmD2, respectively. Lsm11 is a unique Sm protein with an extended C-terminal domain. The Sm10/Sm11 ring binds with U7 snRNA, forming the U7 snRNP, which mediates processing of the 3' UTR stem-loop of the histone mRNA in the nucleus [201]. A number of two domain Sm/methyl transferases (Lsm12-16) have been discovered [202,203]. Currently the specific function of them is unknown, although some studies have suggested a role in mRNA translational control (Lsm13, Lsm14, Lsm15) or formation of P-bodies (Lsm16) [204-207].

5.4 The Enigma of SmAP Function

While Sm-like Archaeal Proteins or SmAPs, were the first Sm proteins for which atomic-resolution structures of the full intact oligomeric ring were resolved [24-26], knowledge of their biology lags behind the other domains. Currently, there is very little information on the physiological role of SmAPs, and no answer to the key questions of whether SmAPs function as *chaperones* or *scaffolding*, or what type of macromolecules (RNA, protein), SmAPs interact with *in vivo*. Since the function of Sm proteins in bacteria and eukarya is dependent on their interactions with regulatory RNAs and RNA processing, a closer look at the known RNA processing of archaea may give insight into SmAP function.

5.4.1 Archaeal Non-coding RNAs

Though they lack the large macromolecular complexes in eukarya, archaea share many features with eukarya in terms of their information processing pathways [208]. One of the major differences between eukarya and archaea information processing as it relates to Sm function, is the lack of pre-mRNA introns that require splicing, removing the need for a spliceosome homolog. However, both eukarya and bacteria, archaea do contain introns in transfer RNA (tRNA) and ribosomal RNA (rRNA). Interestingly, archaea contain higher percentages of introns per tRNA than the other domains. Commonly, 15% of archaeal tRNAs contains intros, but in Thermoproteales, the percentage can rise to 70% [209,210]. tRNA introns can vary in length between 16 and 44 nucleotides, and some tRNAs can contain multiple introns needing to be spliced. At the highest end of the spectrum, 87% of *Pyrobaculum calidifontis* tRNAs contain introns, with half of them containing more than one intron [209]. Most tRNA introns in archaea and eukarya require a homologous splicing endonuclease protein complex (RNase P for the 5’ end and tRNase Z for the 3’ end) which recognize a conserved bulge-helix-bulge structure motif (BHB) [211,212]. Archaeal tRNAs can require additional type of processing, such as split tRNA, which are *trans*-spliced from different loci, and permuted tRNA, where the 3’ half is upstream of the 5’ half [213,214]. As previously mentioned (Section 3), the exosome is another RNA processing pathway conserved between archaea and eukarya, and demonstrates the same evolutionary progression as Sm proteins, from simple homo-oligomers to complicated hetero-oligomers [37]. Exosomes are responsible for degrading RNA including mRNA, rRNA, and sRNA.

RNA-seq high-throughput sequencing has identified a growing repertoire of noncoding RNAs in archaea. C/D box sRNAs (related to small nucleolar RNAs (snoRNAs) in eukarya) were the first identified sRNAs in archaea [94], and were unexpected because archaea lack a nucleolous. Both archaeal and eukaryal C/D box snoRNAs guide methylation of the ribosome. Following the discovery of C/D box snRNAs came H/ACA sRNAs, which are also components of ribosome processing, and CRISPR RNAs (crRNAs) which defend against foreign DNA [95,215,216].

Regulatory sRNAs were found in archaea shortly after the discovery of archaeal snoRNAs. sRNAs were found in multiple archaea species including *M. janaschii*, *Pyrococcus furiosus*, *A. fulgidus*, *Sulfolobus solfataricus*, and *Haloferax volcanii* [217-222]. While many of the targets of archaeal sRNAs are unknown, the current theory is that they are analogous to bacterial sRNAs or eukaryal miRNAs, though compared to bacterial sRNAs, archaeal sRNAs are not highly conserved, even within the same genus [223].

5.4.2 SmAPs: RNA Chaperones or RNP Scaffolds?

SmAPs are encoded in almost all archaeal genomes sequenced. Euryarchaea encode the least number of SmAPs, with either one or two; Crenarchaea are known to encode two to three, and Thaumarchaea usually encode three. While most SmAPs have a higher sequence homology and more similar overall structure (heptamer) to the eukaryal Sm/Lsm, they form stable homo-oligomers similar to Hfqs. In addition, the Euryarchaeota *M. jannaschii* encodes an Hfq-like hexameric SmAP [81] and *A. fulgidus* encodes both a heptameric SmAP and a SmAP that can transition from a heptamer to a hexamer in a pH and substrate dependent manner [48,224]. *M. jannaschii* Hfq-like SmAP exhibits the conserved Sm-fold, but when compared to *E. coli* Hfq, *M. jannaschii* has an acidic charge on the L4 face, which could prevent adenine-rich binding [81]. *M. jannaschii* Hfq-like SmAP recovers function in Δ*hfq* mutant strains of *E. coli* and *S. enterica* [81,225]. *M. jannaschii* Hfq-like SmAP stabilized bacterial sRNAs preventing degradation (Figure 6C), assisted in sRNA mediated mRNA degradation (Figure 6D), and forms a ternary mRNA···SmAP···sRNA complex *in vitro* [81,225]. The *M jannaschii* ternary complex differs significantly from the traditional Hfq ternary mRNA···Hfq···sRNA complexes by having the mRNA and sRNA compete for binding rather than using separate binding surfaces [81]. Hfq-like SmAPs with an N-terminal C_2H_2 zinc-finger domain have now been found in both *Thermococcus* plasmids and *Methanococcal* plasmids [226]. The function of these Hfq-like SmAPs has yet to be determined, but the addition of the zinc-finger domain is intriguing as the combination of an RNA-binding domain and DNA-binding domain could have significant implications for DNA-RNA processing.

As mentioned previously, much of the information processing (including RNA metabolism) in archaea is more closely related to eukarya than bacteria [208]. As would be expected, crystal structures of SmAPs showed features similar to eukaryal Sm proteins [24-26]. The first SmAP structures revealed SmAPs to be heptamers and to have the elongated Loop 4 reminiscent of eukaryal Sm/Lsm compared to the shortened bacterial Hfq. The elongated Loop 4 could play an important role in function as it would appear to occlude the adenine-rich binding site on Hfq. Multiple SmAPs have been co-crystallized with uridine-rich RNA or UMPs which bind in the same conserved pore region as Hfq and eukaryal Sm proteins [24,80,81]. A *P. abyssi* structure co-crystallized with U_6 RNA and UMP reveals an additional binding site near the lateral rim, similar to the bacterial lateral rim binding site [81]. For SmAPs to function as *chaperones* and not *scaffolds* they would require more than one binding site to bring RNAs in proximity, which could be the function of the lateral binding site. One of the few functional studies on SmAP has shown that an *H. volcanii* SmAP knockout mutant (Δ*smap*) shows a similar pleiotropic phenotype as an Δ*hfq* mutant [227]. However, a follow-up paper showed that much of this phenotype was caused by the deletion of a section of the promoter gene *rpl37R*, which overlapped with the *smap* gene [228]. In the same study, *H. volcanii* SmAP co-immunoprecipitated with a variety of protein and non-coding RNAs [227]. The proteins identified are similar to expected partners for bacterial Hfq and eukaryal Sm proteins (ribosomal proteins, elongation factors, ribonucleases, *etc*.), though whether the interaction is direct or indirect is unknown. The functions of most of the sRNAs is unknown, though they appear to be similar to regulatory sRNA targets in bacteria, thus hinting at a *chaperone* role for SmAPs [96,218,227,229].

Eukaryal Sm proteins are involved in biogenesis of RNAs and RNA processing pathways. Many of these pathways are conserved in archaea but are not shared with bacteria. In the early

2000s a study on SmAP from *A. fulgidus* showed that the RNase P catalytic RNA (both pre- and post- processing) co-immunoprecipitated with SmAP *in vivo* [24]. The interaction of *A. fulgidus* SmAP with RNase P RNA suggests a role similar to the eukaryal Sm/Lsm as they function in large macromolecular complexes involved in RNA biogenesis, except with tRNA. Both the canonical Sm core proteins and Hfq have been shown to interact with tRNA [188,230], though whether they have a direct function on maturation or processing of this class of RNAs is unknown.

Many proteins which are co-expressed demonstrate functions in similar pathways, and conserved, co-expressed gene neighbors can often supply an understanding of a protein's function. The most conserved SmAP appears to be co-expressed with the *l37e* gene, which encodes the zinc-finger protein L37e [25]. *Haloarcula marismortui* L37e was shown to bind a conserved adenine-rich section in 50s rRNA. In the previously mentioned knockout and co-immunoprecipitation studies in *H. volcanii*, L37e is co-transcribed with SmAP but did not appear in the co-immunoprecipitation, meaning any interaction is most likely indirect [231]. The close genomic association of L37e suggests that SmAPs could have a role in rRNA processing similar to that of the Lsm proteins (see section 5.3.4).

6 Deep-branching Sm Proteins

Sm proteins provide a window into the world of RNA processing and metabolism in all domains of life. While this ancient protein has been extensively studied in bacteria and eukarya, few studies have examined the function of the Sm-like Archaeal Proteins (SmAPs). The motivation for studying SmAPs is two-fold, *i)* because of their close homology with eukarya, examination of RNA processing in archaea, specifically Sm proteins, could lead to better understanding of the evolution of the large RNPs in eukarya, and *ii)* SmAPs offer a more accessible and simpler context than the more complex eukaryal Sm proteins. In addition, SmAPs could also offer evolutionary insight into the transition from the bacterial Hfq to the eukaryal Sm/Lsm, as indicated for example by the interaction of SmAPs with sRNAs similar to the bacterial regulatory sRNAs. The euryarchaeaote *Pyrobaculum aerophilum* is an organism of interest, because *P. aerophilum* encodes three SmAP paralogs. Each *P. aerophilum* SmAP contains features that make them distinct, including being highly charged (*P. aerophilum* SmAP2) or an extended C-terminal domain (*P. aerophilum* SmAP3) [59]. Many deep-branching bacteria share a high degree of similarity (genomically) with archaea, including the hyper-thermophilic *Thermotoga maritima*. *T. maritima* Hfq is an interesting homolog because of its simplicity (no C-terminal tail) and the aforementioned archaeal genome. The two studies reported here, one in archaea and one in bacteria, will help to illuminate the functions of ancient Sm proteins, supply a window into RNA processing in archaea, and the evolution of Sm proteins.

7 References

1. de Fernandez MTF, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. *Nature*. 219(5154), 588–590 (1968).

2. de Fernandez MTF, Hayward WS, August JT. Bacterial proteins required for replication of phage Q ribonucleic acid. Pruification and properties of host factor I, a ribonucleic acid-binding protein. *Journal of Biological Chemistry*. 247(3), 824–831 (1972).

3. . Nucleotide sequence specific interaction of host factor I with bacteriophage Qβ RNA. *FEBS Lett*. 43, 20–22 (1974).

4. . Site-specific interaction of Qβ host factor and ribosomal protein S1 with Qβ and R17 bacteriophage RNAs. *J Biol Chem*. 251, 1902–1912 (1976).

5. . Interaction of Escherichia coli host factor protein with oligoriboadenylates. *Biochemistry*. 19, 6138–6146 (1980).

6. Carmichael GG, Weber K, Niveleau A, Wahba AJ. The host factor required for RNA phage Qbeta RNA replication in vitro. Intracellular location, quantitation, and purification by polyadenylate-cellulose chromatography. *Journal of Biological Chemistry*. 250(10), 3607–3612 (1975).

7. . Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in Escherichia coli K-12. *Mol Microbiol*. 13, 35–49 (1994).

8. Brown L, Elliott T. Efficient translation of the RpoS sigma factor in Salmonella typhimurium requires host factor I, an RNA-binding protein encoded by the hfq gene. *J Bacteriol*. 178(13), 3763–3770 (1996).

9. . The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for rpoS translation in Escherichia coli. *Genes Dev*. 10, 1143–1151 (1996).

10. Muffler A, Traulsen DD, Fischer D, Lange R, Hengge-Aronis R. The RNA-binding protein HF-I plays a global regulatory role which is largely, but not exclusively, due to its role in expression of the sigmaS subunit of RNA polymerase in Escherichia coli. *J Bacteriol*. 179(1), 297–300 (1997).

11. Lerner MR, Steitz JA. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci USA*. 76(11), 5495–5499 (1979).

12. Tan EM, Kunkel HG. Characteristics of a soluble nuclear antigen precipitating with sera of patients with systemic lupus erythematosus. *J Immunol*. 96(3), 464–471 (1966).

13. Tsokos GC. In the beginning was Sm. *J Immunol*. 176(3), 1295–1296 (2006).

14. Reeves WH, Narain S, Satoh M. Henry Kunkel, Stephanie Smith, clinical immunology, and split genes. *Lupus*. 12(3), 213–217 (2003).

15. Lerner MR, Boyle JA, Hardin JA, Steitz JA. Two novel classes of small ribonucleoproteins detected by antibodies associated with lupus erythematosus. *Science*. 211(4480), 400–402 (1981).

16. Séraphin B. Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J*. 14(9), 2089–2098 (1995).

17. Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA. Are snRNPs involved in splicing? *Nature*. 283(5743), 220–224 (1980).

18. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J*. 18(12), 3451–3462 (1999).

19. Good MC, Zalatan JG, Lim WA. Scaffold proteins: hubs for controlling the flow of cellular information. *Science*. 332(6030), 680–686 (2011).

20. Kambach C, Walke S, Young R, *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell*. 96(3), 375–387 (1999).

21. Will CL, Lührmann R. Protein functions in pre-mRNA splicing. *Curr Opin Cell Biol*. 9(3), 320–328 (1997).

22. Séraphin B. Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.* 14(9), 2089–2098 (1995).

23. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. *Int Rev Cell Mol Biol*. 272, 149–189 (2009).

24. Törö I, Thore S, Mayer C, Basquin J, Séraphin B, Suck D. RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex. *EMBO J.* 20(9), 2293–2303 (2001).

25. Collins BM, Harrop SJ, Kornfeld GD, Dawes IW, Curmi PM, Mabbutt BC. Crystal structure of a heptameric Sm-like protein complex from archaea: implications for the structure and evolution of snRNPs. *J Mol Biol*. 309(4), 915–923 (2001).

26. Mura C, Cascio D, Sawaya MR, Eisenberg DS. The crystal structure of a heptameric archaeal Sm protein: Implications for the eukaryotic snRNP core. *Proc Natl Acad Sci USA*. 98(10), 5532–5537 (2001).

27. Schumacher MA, Pearson RF, Møller T, Valentin-Hansen P, Brennan RG. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. *EMBO J.* 21(13), 3546–3556 (2002).

28. Møller T, Franch T, Højrup P, *et al.* Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell*. 9(1), 23–30 (2002).

29. Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G. The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol Cell*. 9(1), 11–22 (2002).

30. Arluison V, Derreumaux P, Allemand F, Folichon M, Hajnsdorf E, Régnier P. Structural Modelling of the Sm-like Protein Hfq from Escherichia coli. *J Mol Biol*. 320(4), 705–712 (2002).

31. Dayhoff MO. The origin and evolution of protein superfamilies. *Fed Proc*. 35(10), 2132–2138 (1976).

32. Scofield DG, Lynch M. Evolutionary diversification of the Sm family of RNA-associated proteins. *Mol Biol Evol*. 25(11), 2255–2267 (2008).

33. Veretnik S, Wills C, Youkharibache P, Valas RE, Bourne PE. Sm/Lsm genes provide a glimpse into the early evolution of the spliceosome. *PLoS Comput Biol*. 5(3), e1000315 (2009).

34. Sousa SA, Ramos CG, Moreira LM, Leitão JH. The hfq gene is required for stress resistance and full virulence of Burkholderia cepacia to the nematode Caenorhabditis elegans. *Microbiology (Reading, Engl)*. 156(Pt 3), 896–908 (2009).

35. Ramos CG, Grilo AM, Feliciano JR, Leitão JH. The second RNA chaperone, Hfq2, is also required for survival under stress and full virulence of Burkholderia cenocepacia J2315. *J Bacteriol*. 193(7), 1515–1526 (2011).

36. Archibald JM, Logsdon JM, Doolittle WF. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. *Mol Biol Evol*. 17(10), 1456–1466 (2000).

37. Lin-Chao S, Chiou N-T, Schuster G. The PNPase, exosome and RNA helicases as the building components of evolutionarily-conserved RNA degradation machines. *J Biomed Sci*. 14(4), 523–532 (2007).

38. Hochstrasser M. Evolution and function of ubiquitin-like protein-conjugation systems. *Nat Cell Biol*. 2(8), E153–7 (2000).

39. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res*. 33(14), 4626–4638 (2005).

40. Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. *Nature*. 440(7080), 41–45 (2006).

41. Sabath I, Skrajna A, Yang X-C, Dadlez M, Marzluff WF, Dominski Z. 3'-End processing of histone pre-mRNAs in Drosophila: U7 snRNP is associated with FLASH and polyadenylation factors. *RNA*. 19(12), 1726–1744 (2013).

42. Cech TR. Self-splicing RNA: implications for evolution. *Int Rev Cytol*. 93, 3–22 (1985).

43. Robart AR, Zimmerly S. Group II intron retroelements: function and diversity. *Cytogenet Genome Res*. 110(1-4), 589–597 (2005).

44. Seetharaman M, Eldho NV, Padgett RA, Dayie KT. Structure of a self-splicing group II intron catalytic effector domain 5: parallels with spliceosomal U6 RNA. *RNA*. 12(2), 235–247 (2006).

45. Pillai RS, Grimmler M, Meister G, Will CL, Lührmann R, Fischer U. Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes & Development*. 17(18), 2321–2333 (2003).

46. McAllister G, Amara SG, Lerner MR. Tissue-specific expression and cDNA cloning of small nuclear ribonucleoprotein-associated polypeptide N. *Proc Natl Acad Sci USA*. 85(14), 5296–5300 (1988).

47. Pillai RS, Will CL, Lührmann R, Schümperli D, Müller B. Purified U7 snRNPs lack the Sm proteins D1 and D2 but contain Lsm10, a new 14 kDa Sm D1-like protein. *EMBO J*. 20(19), 5470–5479 (2001).

48. Mura C, Randolph PS, Patterson J, Cozen AE. Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. *RNA Biol*. 10(4), 636–651 (2013).

49. . snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein–protein interactions. *EMBO J*. 14, 2076–2088 (1995).

50. Hermann H, Fabrizio P, Raker VA, *et al.* snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. *EMBO J*. 14(9), 2076–2088 (1995).

51. Attia AS, Sedillo JL, Wang W, *et al.* Moraxella catarrhalis expresses an unusual Hfq protein. *Infect Immun*. 76(6), 2520–2530 (2008).

52. Sauter C, Basquin J, Suck D. Sm-like proteins in Eubacteria: the crystal structure of the Hfq protein from Escherichia coli. *Nucleic Acids Res*. 31(14), 4091–4098 (2003).

53. Naidoo N, Harrop SJ, Sobti M, *et al.* Crystal structure of Lsm3 octamer from Saccharomyces cerevisiae: implications for Lsm ring organisation and recruitment. *J Mol Biol*. 377(5), 1357–1371 (2008).

54. Das D, Kozbial P, Axelrod HL, *et al.* Crystal structure of a novel Sm-like protein of putative cyanophage origin at 2.60 A resolution. *Proteins*. 75(2), 296–307 (2009).

55. Sauer E, Weichenrieder O. Structural basis for RNA 3'-end recognition by Hfq. *Proc Natl Acad Sci USA*. 108(32), 13065–13070 (2011).

56. Raker VA, Plessel G, Lührmann R. The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro. *EMBO J*. 15(9), 2256–2269 (1996).

57. Leung AKW, Nagai K, Li J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. *Nature*. 473(7348), 536–539 (2011).

58. Kilic T, Sanglier S, Van Dorsselaer A, Suck D. Oligomerization behavior of the archaeal Sm2-type protein from Archaeoglobus fulgidus. *Protein Sci*. 15(10), 2310–2317 (2006).

59. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D. Structure and assembly of an augmented Sm-like archaeal protein 14-mer. *Proc Natl Acad Sci USA*. 100(8), 4539–4544 (2003).

60. Krummel DAP, Oubridge C, Leung AKW, Li J, Nagai K. Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*. 458(7237), 475–480 (2009).

61. Krummel DAP, Nagai K, Oubridge C. Structure of spliceosomal ribonucleoproteins. *F1000 Biol Rep*. 2 (2010).

62. Kondo Y, Oubridge C, van Roon A-MM, Nagai K. Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5' splice site recognition. *Elife*. 4 (2015).

63. Sauer E, Schmidt S, Weichenrieder O. Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. *Proc Natl Acad Sci USA*. 109(24), 9396–9401 (2012).

64. Kovach AR, Hoff KE, Canty JT, Orans J, Brennan RG. Recognition of U-rich RNA by Hfq from the Gram-positive pathogen Listeria monocytogenes. *RNA*. 20(10), 1548–1559 (2014).

65. Murina VN, Nikulin AD. RNA-binding Sm-like proteins of bacteria and archaea. similarity and difference in structure and function. *Biochemistry Mosc*. 76(13), 1434–1449 (2012).

66. Adamson DN, Lim HN. Essential requirements for robust signaling in Hfq dependent small RNA networks. *PLoS Comput Biol*. 7(8), e1002138 (2011).

67. Wang W, Wang L, Zou Y, *et al.* Cooperation of Escherichia coli Hfq hexamers in DsrA binding. *Genes & Development*. 25(19), 2106–2117 (2011).

68. . Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. *Proc Natl Acad Sci USA*. 106, 19292–19297 (2009).

69. Wang W, Wang L, Wu J, Gong Q, Shi Y. Hfq-bridged ternary complex is important for translation activation of rpoS by DsrA. *Nucleic Acids Res*. 41(11), 5938–5948 (2013).

70. Hämmerle H, Beich-Frandsen M, Rajkowitsch L, Carugo O, Djinović-Carugo K, Bläsi U. Structural and biochemical studies on ATP binding and hydrolysis by the Escherichia coli RNA chaperone Hfq. *PLoS ONE*. 7(11), e50892 (2012).

71. Mutyam SK, Mura C, Marco S, Sukhodolets MV. Sm-like protein Hfq: location of the ATP-binding site and the effect of ATP on Hfq-- RNA complexes. *Protein Sci*. 16(9), 1830–1841 (2007).

72. Someya T, Baba S, Fujimoto M, Kawai G, Kumasaka T, Nakamura K. Crystal structure of Hfq from Bacillus subtilis in complex with SELEX-derived RNA aptamer: insight into RNA-binding properties of bacterial Hfq. *Nucleic Acids Res*. 40(4), 1856–1867 (2011).

73. Lesnik EA, Sampath R, Levene HB, Henderson TJ, McNeil JA, Ecker DJ. Prediction of rho-independent transcriptional terminators in Escherichia coli. *Nucleic Acids Res*. 29(17), 3583–3594 (2001).

74. Geissmann TA, Touati D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*. 23(2), 396–405 (2004).

75. Brescia CC, Mikulecky PJ, Feig AL, Sledjeski DD. Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure. *RNA*. 9(1), 33–43 (2003).

76. Panja S, Schu DJ, Woodson SA. Conserved arginines on the rim of Hfq catalyze base pair formation and exchange. *Nucleic Acids Res*. 41(15), 7536–7546 (2013).

77. Murina V, Lekontseva N, Nikulin A. Hfq binds ribonucleotides in three different RNA-binding sites. *Acta Crystallogr D Biol Crystallogr*. 69(Pt 8), 1504–1513 (2013).

78. Schu DJ, Zhang A, Gottesman S, Storz G. Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. *EMBO J*. 34(20), 2557–2573 (2015).

79. Wagner EGH. Cycling of RNAs on Hfq. *RNA Biol*. 10(4), 619–626 (2013).

80. Mura C, Kozhukhovsky A, Gingery M, Phillips M, Eisenberg D. The oligomerization and ligand-binding properties of Sm-like archaeal proteins (SmAPs). *Protein Sci*. 12(4), 832–847 (2003).

81. Nielsen JS, Bøggild A, Andersen CBF, *et al.* An Hfq-like protein in archaea: crystal structure and functional characterization of the Sm protein from Methanococcus jannaschii. *RNA*. 13(12), 2213–2223 (2007).

82. Robertson MP, Joyce GF. The origins of the RNA world. *Cold Spring Harb Perspect Biol*. 4(5), a003608 (2012).

83. Benner SA, Kim H-J, Yang Z. Setting the stage: the history, chemistry, and geobiology behind RNA. *Cold Spring Harb Perspect Biol*. 4(1), a003541 (2012).

84. Bernhardt HS. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). *Biol Direct*. 7, 23 (2012).

85. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*. 75(5), 843–854 (1993).

86. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*. 157(1), 77–94 (2014).

87. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*. 19(1), 92–105 (2008).

88. Tomizawa J, Som T. Control of ColE1 plasmid replication: enhancement of binding of RNA I to the primer transcript by the Rom protein. *Cell*. 38(3), 871–878 (1984).

89. Simons RW, Kleckner N. Translational control of IS10 transposition. *Cell*. 34(2), 683–691 (1983).

90. Livny J, Waldor MK. Identification of small RNAs in diverse bacterial species. *Current opinion in microbiology*. 10(2), 96–101 (2007).

91. Altuvia S. Identification of bacterial small non-coding RNAs: experimental approaches. *Current opinion in microbiology*. 10(3), 257–261 (2007).

92. . Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in Salmonella enterica. *Mol Microbiol*. 78, 380–394 (2010).

93. Prasse D, Ehlers C, Backofen R, Schmitz RA. Regulatory RNAs in archaea: first target identification in Methanoarchaea. *Biochem Soc Trans*. 41(1), 344–349 (2013).

94. Omer AD, Ziesche S, Ebhardt H, Dennis PP. In vitro reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. *Proc Natl Acad Sci USA*. 99(8), 5289–5294 (2002).

95. Baker DL, Youssef OA, Chastkofsky MIR, Dy DA, Terns RM, Terns MP. RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP. *Genes & Development*. 19(10), 1238–1248 (2005).

96. Fischer S, Benz J, Späth B, *et al.* Regulatory RNAs in Haloferax volcanii. *Biochem Soc Trans*. 39(1), 159–162 (2011).

97. . Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci USA*. 106(51), 21878 (2009).

98. . Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes Dev*. 16, 1696–1706 (2002).

99. . Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol*. 61, 1013–1022 (2006).

100. . The Escherichia coli OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J*. 17, 6069–6075 (1998).

101. Gottesman S. The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annu Rev Microbiol*. 58(1), 303–328 (2004).

102. Henderson CA, Vincent HA, Casamento A, *et al.* Hfq binding changes the structure of Escherichia coli small noncoding RNAs OxyS and RprA, which are involved in the riboregulation of rpoS. *RNA*. 19(8), 1089–1104 (2013).

103. . Both RNase E and RNase III control the stability of sodB mRNA upon translational inhibition by the small regulatory RNA RyhB. *Nucleic Acids Res*. 33, 1678–1689 (2005).

104. . Hfq is necessary for regulation by the untranslated RNA DsrA. *J Bacteriol*. 183, 1997–2005 (2001).

105. Mohanty BK, Maples VF, Kushner SR. The Sm-like protein Hfq regulates polyadenylation dependent mRNA decay in Escherichia coli. *Mol Microbiol*. 54(4), 905–920 (2004).

106. Møller T, Franch T, Højrup P, *et al.* Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell*. 9(1), 23–30 (2002).

107. . Positive regulation by small RNAs and the role of Hfq. *Proc Natl Acad Sci USA*. 107, 9602–9607 (2010).

108. . RNAs actively cycle on the Sm-like protein Hfq. *Genes & Development*. 24(23), 2621 (2010).

109. Georg J, Hess WR. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*. 75(2), 286–300 (2011).

110. . Hfq binding at RhlB-recognition region of RNase E is crucial for the rapid degradation of target mRNAs mediated by sRNAs in Escherichia coli. *Mol Microbiol*. 79, 419–432 (2011).

111. . Coincident Hfq binding and RNase E cleavage sites on mRNA and small regulatory RNAs. *RNA*. 9, 1308–1314 (2003).

112. Gribskov M, Burgess RR. Overexpression and purification of the sigma subunit of Escherichia coli RNA polymerase. *Gene*. 26(2-3), 109–118 (1983).

113. Campagne S, Marsh ME, Capitani G, Vorholt JA, Allain FH-T. Structural basis for -10 promoter element melting by environmentally induced sigma factors. *Nat Struct Mol Biol*. 21(3), 269–276 (2014).

114. Gross CA, Chan C, Dombroski A, *et al.* The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb Symp Quant Biol*. 63, 141–155 (1999).

115. Muffler A, Fischer D, Hengge-Aronis R. The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for rpoS translation in Escherichia coli. *Genes & Development*. 10(9), 1143–1151 (1996).

116. Gottesman S, Storz G. RNA reflections: converging on Hfq. *RNA*. 21(4), 511–512 (2015).

117. Sledjeski DD, Gupta A, Gottesman S. The small RNA, DsrA, is essential for the low temperature expression of RpoS during exponential growth in Escherichia coli. *EMBO J*. 15(15), 3993–4000 (1996).

118. Updegrove TB, Correia JJ, Chen Y, Terry C, Wartell RM. The stoichiometry of the Escherichia coli Hfq protein bound to RNA. *RNA*. 17(3), 489–500 (2011).

119. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*. 3(12), a003798 (2011).

120. Moon K, Gottesman S. Competition among Hfq-binding small RNAs in Escherichia coli. *Mol Microbiol*. 82(6), 1545–1562 (2011).

121. Guisbert E, Rhodius VA, Ahuja N, Witkin E, Gross CA. Hfq modulates the sigmaE-mediated envelope stress response and the sigma32-mediated cytoplasmic stress response in Escherichia coli. *J Bacteriol*. 189(5), 1963–1973 (2006).

122. Chiang M-K, Lu M-C, Liu L-C, Lin C-T, Lai Y-C. Impact of Hfq on global gene expression and virulence in Klebsiella pneumoniae. *PLoS ONE*. 6(7), e22248 (2011).

123. Sonnleitner E, Schuster M, Sorger-Domenigg T, Greenberg EP, Bläsi U. Hfq-dependent alterations of the transcriptome profile and effects on quorum sensing in Pseudomonas aeruginosa. *Mol Microbiol*. 59(5), 1542–1558 (2006).

124. Meibom KL, Forslund A-L, Kuoppa K, *et al.* Hfq, a novel pleiotropic regulator of virulence-associated genes in Francisella tularensis. *Infect Immun*. 77(5), 1866–1880 (2009).

125. Salvail H, Massé E. Regulating iron storage and metabolism with RNA: an overview of posttranscriptional controls of intracellular iron homeostasis. *Wiley Interdiscip Rev RNA*. 3(1), 26–36 (2011).

126. Prévost K, Salvail H, Desnoyers G, Jacques J-F, Phaneuf E, Massé E. The small RNA RyhB activates the translation of shiA mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. *Mol Microbiol*. 64(5), 1260–1273 (2007).

127. . A small RNA promotes siderophore production through transcriptional and metabolic remodeling. *Proc Natl Acad Sci USA*. 107(34), 15223 (2010).

128. Peterson JW. Bacterial Pathogenesis. In: *Medical Microbiology*. S B (Ed.). University of Texas Medical Branch at Galveston, Galveston, TX (1996).

129. Flemming H-C, Wingender J. The biofilm matrix. *Nat Rev Microbiol*. 8(9), 623–633 (2010).

130. Fazli M, Almblad H, Rybtke ML, Givskov M, Eberl L, Tolker-Nielsen T. Regulation of biofilm formation in Pseudomonas and Burkholderia species. *Environ Microbiol*. 16(7), 1961–1981 (2014).

131. Kulesus RR, Diaz-Perez K, Slechta ES, Eto DS, Mulvey MA. Impact of the RNA chaperone Hfq on the fitness and virulence potential of uropathogenic Escherichia coli. *Infect Immun*. 76(7), 3019–3026 (2008).

132. Mika F, Hengge R. Small Regulatory RNAs in the Control of Motility and Biofilm Formation in E. coli and Salmonella. *Int J Mol Sci*. 14(3), 4560–4579 (2013).

133. Boehm A, Vogel J. The csgD mRNA as a hub for signal integration via multiple small RNAs. *Mol Microbiol*. 84(1), 1–5 (2012).

134. . A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nature Struct Mol Biol*. 20, 36–45 (2013).

135. Leitão JH, Ferreira AS, Ramos CG, Silva IN, Moreira LM. Pathogenicity, virulence factors, and strategies to fight against Burkholderia cepacia complex pathogens and related species. *Appl Microbiol Biotechnol*. 87(1), 31–40 (2010).

136. Rietschel ET, Kirikae T, Schade FU, *et al.* Bacterial endotoxin: molecular relationships of structure to activity and function. *FASEB J*. 8(2), 217–225 (1994).

137. Ansong C, Yoon H, Porwollik S, *et al.* Global systems-level analysis of Hfq and SmpB deletion mutants in Salmonella: implications for virulence and global protein translation. *PLoS ONE*. 4(3), e4809 (2009).

138. Caro-Hernández P, Fernández-Lago L, de Miguel M-J, *et al.* Role of the Omp25/Omp31 family in outer membrane properties and virulence of Brucella ovis. *Infect Immun*. 75(8), 4050–4061 (2007).

139. Cui M, Wang T, Xu J, *et al.* Impact of Hfq on global gene expression and intracellular survival in Brucella melitensis. *PLoS ONE*. 8(8), e71933 (2013).

140. Sittka A, Pfeiffer V, Tedin K, Vogel J. The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. *Mol Microbiol*. 63(1), 193–217 (2006).

141. Delihas N. Annotation and evolutionary relationships of a small regulatory RNA gene micF and its target ompF in Yersinia species. *BMC Microbiol*. 3, 13 (2003).

142. Tsai Y-L, Wang M-C, Hsueh P-R, *et al.* Overexpression of an outer membrane protein associated with decreased susceptibility to carbapenems in Proteus mirabilis. *PLoS ONE*. 10(3), e0120395 (2015).

143. Esterling L, Delihas N. The regulatory RNA gene micF is present in several species of gram-negative bacteria and is phylogenetically conserved. *Mol Microbiol*. 12(4), 639–646 (1994).

144. Nakao H, Watanabe H, Nakayama S, Takeda T. yst gene expression in Yersinia enterocolitica is positively regulated by a chromosomal region that is highly homologous to Escherichia coli host factor 1 gene (hfq). *Mol Microbiol*. 18(5), 859–865 (1995).

145. Sonnleitner E, Sorger-Domenigg T, Madej MJ, *et al.* Detection of small RNAs in Pseudomonas aeruginosa by RNomics and structure-based bioinformatic tools. *Microbiology (Reading, Engl)*. 154(Pt 10), 3175–3187 (2008).

146. Nakano M, Takahashi A, Su Z, Harada N, Mawatari K, Nakaya Y. Hfq regulates the expression of the thermostable direct hemolysin gene in Vibrio parahaemolyticus. *BMC Microbiol*. 8, 155 (2008).

147. Gangaiah D, Labandeira-Rey M, Zhang X, *et al.* Haemophilus ducreyi Hfq contributes to virulence gene regulation as cells enter stationary phase. *MBio*. 5(1), e01081–13 (2014).

148. Bibova I, Skopova K, Masin J, *et al.* The RNA chaperone Hfq is required for virulence of Bordetella pertussis. *Infect Immun*. 81(11), 4081–4090 (2013).

149. . Translational autocontrol of the Escherichia coli hfq RNA chaperone gene. *RNA*. 11, 976–984 (2005).

150. McNealy TL, Forsbach-Birk V, Shi C, Marre R. The Hfq homolog in Legionella pneumophila demonstrates regulation by LetA and RpoS and interacts with the global regulator CsrA. *J Bacteriol*. 187(4), 1527–1532 (2005).

151. Fantappiè L, Metruccio MME, Seib KL, *et al.* The RNA chaperone Hfq is involved in stress response and virulence in Neisseria meningitidis and is a pleiotropic regulator of protein expression. *Infect Immun*. 77(5), 1842–1853 (2009).

152. Panda G, Tanwer P, Ansari S, Khare D, Bhatnagar R. Regulation and RNA-binding properties of Hfq-like RNA chaperones in Bacillus anthracis. *Biochim Biophys Acta*. 1850(9), 1661–1668 (2015).

153. . Effect of RyhB small RNA on global iron use in Escherichia coli. *J Bacteriol*. 187, 6962–6971 (2005).

154. Azam TA, Hiraga S, Ishihama A. Two types of localization of the DNA-binding proteins within the Escherichia coli nucleoid. *Genes Cells*. 5(8), 613–626 (2000).

155. . Cellular electron microscopy imaging reveals the localization of the Hfq protein close to the bacterial membrane. *PLoS ONE*. 4, e8301 (2009).

156. Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol*. 15(2), 108–121 (2014).

157. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci USA*. 74(8), 3171–3175 (1977).

158. . An amazing sequence arrangement at the 5[prime] ends of adenovirus 2 messenger RNA. *Cell*. 12, 1–8 (1977).

159. . Understanding alternative splicing: towards a cellular code. *Nature Rev Mol Cell Biol*. 6, 386–398 (2005).

160. Black DL. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem*. 72(1), 291–336 (2003).

161. . Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Rev Mol Cell Biol*. 8, 209–220 (2007).

162. Burge CB, Padgett RA, Sharp PA. Evolutionary fates and origins of U12-type introns. *Mol Cell*. 2(6), 773–785 (1999).

163. Schneider C, Will CL, Makarova OV, Makarov EM, Lührmann R. Human U4/U6.U5 and U4atac/U6atac.U5 tri-snRNPs exhibit similar protein compositions. *Mol Cell Biol*. 22(10), 3219–3229 (2002).

164. Yeo GW, Van Nostrand EL, Van Nostrand EL, Liang TY. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genet*. 3(5), e85 (2007).

165. . Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med*. 18, 472–482 (2012).

166. . New connections between splicing and human disease. *Trends Genet*. 28, 147–154 (2012).

167. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet*. 17(1), 19–32 (2015).

168. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA*. 4(1), 61–76 (2012).

169. . Regulation of snRNA gene expression by the Drosophila melanogaster small nuclear RNA activating protein complex (DmSNAPc). *Crit Rev Biochem Mol Biol*. 46, 11–26 (2011).

170. . SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III. *Genes Dev*. 12, 2664–2672 (1998).

171. . Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*. 123, 265–276 (2005).

172. . Formation of the 3[prime] end of U1 snRNA requires compatible snRNA promoter elements. *Cell*. 47, 249–258 (1986).

173. . Interactions of U2 gene loci and their nuclear transcripts with Cajal (coiled) bodies: evidence for PreU2 within Cajal bodies. *Mol Biol Cell*. 11, 2987–2998 (2000).

174. Suzuki T, Izumi H, Ohno M. Cajal body surveillance of U snRNA export complex assembly. *J Cell Biol*. 190(4), 603–612 (2010).

175. Raker VA, Plessel G, Lührmann R. The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro. *EMBO J*. 15(9), 2256–2269 (1996).

176. . Methylation of Sm proteins by a complex containing PRMT5 and the putative U snRNP assembly factor pICln. *Curr Biol*. 11, 1990–1994 (2001).

177. . The methylosome, a 20S complex containing JBP1 and pICln, produces dimethylarginine-modified Sm proteins. *Mol Cell Biol*. 21, 8289–8300 (2001).

178. Grimm C, Chari A, Pelz J-P, *et al.* Structural basis of assembly chaperone- mediated snRNP formation. *Mol Cell*. 49(4), 692–703 (2013).

179. . An assembly chaperone collaborates with the SMN complex to generate spliceosomal snRNPs. *Cell*. 135, 497–509 (2008).

180. Kroiss M, Schultz J, Wiesner J, Chari A, Sickmann A, Fischer U. Evolution of an RNP assembly system: a minimal SMN complex facilitates formation of UsnRNPs in Drosophila melanogaster. *Proc Natl Acad Sci USA*. 105(29), 10045–10050 (2008).

181. Zhang R, So BR, Li P, *et al.* Structure of a key intermediate of the SMN complex reveals Gemin2's crucial function in snRNP assembly. *Cell*. 146(3), 384–395 (2011).

182. . Gemin5 delivers snRNA precursors to the SMN complex for snRNP biogenesis. *Mol Cell*. 38, 551–562 (2010).

183. . An essential signaling role for the m3G cap in the transport of U1 snRNP to the nucleus. *Science*. 249, 786–790 (1990).

184. . Interaction between the small-nuclear-RNA cap hypermethylase and the spinal muscular atrophy protein, survival of motor neuron. *EMBO Rep*. 4, 616–622 (2003).

185. . Nuclear speckles: a model for nuclear organelles. *Nature Rev Mol Cell Biol*. 4, 605–612 (2003).

186. Stanek D, Neugebauer KM. The Cajal body: a meeting place for spliceosomal snRNPs in the nuclear maze. *Chromosoma*. 115(5), 343–354 (2006).

187. . A multiprotein complex mediates the ATP-dependent assembly of spliceosomal U snRNPs. *Nature Cell Biol*. 3, 945–949 (2001).

188. Pellizzoni L, Yong J, Dreyfuss G. Essential role for the SMN complex in the specificity of snRNP assembly. *Science*. 298(5599), 1775–1779 (2002).

189. . A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc Natl Acad Sci USA*. 96, 6307–6311 (1999).

190. . Inactivation of the survival motor neuron gene, a candidate gene for human spinal muscular atrophy, leads to massive cell death in early mouse embryos. *Proc Natl Acad Sci USA*. 94, 9920–9925 (1997).

191. . Ribonucleoprotein assembly defects correlate with spinal muscular atrophy severity and preferentially affect a subset of spliceosomal snRNPs. *PLoS ONE*. 2, e921 (2007).

192. . Alternative splicing events are a late feature of pathology in a mouse model of spinal muscular atrophy. *PLoS Genet*. 5, e1000773 (2009).

193. . Drosophila model of spinal muscular atrophy uncouples snRNP biogenesis functions of survival motor neuron from locomotion and viability defects. *Cell Rep*. 1, 624–631 (2012).

194. Schramm L, Hernandez N. Recruitment of RNA polymerase III to its target promoters. *Genes & Development*. 16(20), 2593–2620 (2002).

195. Boelens WC, Palacios I, Mattaj IW. Nuclear retention of RNA as a mechanism for localization. *RNA*. 1(3), 273–283 (1995).

196. Kufel J, Bousquet-Antonelli C, Beggs JD, Tollervey D. Nuclear pre-mRNA decapping and 5' degradation in yeast require the Lsm2-8p complex. *Mol Cell Biol*. 24(21), 9646–9657 (2004).

197. Kufel J, Allmang C, Verdone L, Beggs JD, Tollervey D. Lsm proteins are required for normal processing of pre-tRNAs and their efficient association with La-homologous protein Lhp1p. *Mol Cell Biol*. 22(14), 5248–5256 (2002).

198. Kufel J, Allmang C, Petfalski E, Beggs J, Tollervey D. Lsm Proteins are required for normal processing and stability of ribosomal RNAs. *Journal of Biological Chemistry*. 278(4), 2147–2156 (2002).

199. Boeck R, Lapeyre B, Brown CE, Sachs AB. Capped mRNA degradation intermediates accumulate in the yeast spb8-2 mutant. *Mol Cell Biol*. 18(9), 5062–5072 (1998).

200. Bouveret E, Rigaut G, Shevchenko A, Wilm M, Séraphin B. A Sm-like protein complex that participates in mRNA degradation. *EMBO J*. 19(7), 1661–1671 (2000).

201. Dominski Z, Marzluff WF. Formation of the 3' end of histone mRNA: getting closer to the end. *Gene*. 396(2), 373–390 (2007).

202. Albrecht M, Lengauer T. Novel Sm-like proteins with long C-terminal tails and associated methyltransferases. *FEBS Lett*. 569(1-3), 18–26 (2004).

203. Anantharaman V, Aravind L. Novel conserved domains in proteins with predicted roles in eukaryotic cell-cycle regulation, decapping and RNA stability. *BMC Genomics*. 5(1), 45 (2004).

204. Wilhelm JE, Buszczak M, Sayles S. Efficient protein trafficking requires trailer hitch, a component of a ribonucleoprotein complex localized to the ER in Drosophila. *Dev Cell*. 9(5), 675–685 (2005).

205. Decker CJ, Teixeira D, Parker R. Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body assembly in Saccharomyces cerevisiae. *J Cell Biol*. 179(3), 437–449 (2007).

206. Decker CJ, Parker R. CAR-1 and trailer hitch: driving mRNP granule function at the ER? *J Cell Biol*. 173(2), 159–163 (2006).

207. Boag PR, Nakamura A, Blackwell TK. A conserved RNA-protein complex component involved in physiological germline apoptosis regulation in C. elegans. *Development*. 132(22), 4975–4986 (2005).

208. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA*. 87(12), 4576–4579 (1990).

209. Sugahara J, Kikuta K, Fujishima K, Yachie N, Tomita M, Kanai A. Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. *Mol Biol Evol*. 25(12), 2709–2716 (2008).

210. Marck C, Grosjean H. Identification of BHB splicing motifs in intron-containing tRNAs from 18 archaea: evolutionary implications. *RNA*. 9(12), 1516–1531 (2003).

211. Abelson J, Trotta CR, Li H. tRNA splicing. *Journal of Biological Chemistry*. 273(21), 12685–12688 (1998).

212. Kanai A. Molecular Evolution of Disrupted Transfer RNA Genes and Their Introns in Archaea. Springer Berlin Heidelberg, Berlin, Heidelberg, 181–193 (2013).

213. Soma A, Onodera A, Sugahara J, *et al.* Permuted tRNA genes expressed via a circular RNA intermediate in Cyanidioschyzon merolae. *Science*. 318(5849), 450–453 (2007).

214. Randau L, Münch R, Jahn D, Söll D. Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5′- and 3′-halves. *Nature*. 433(7025), 537 (2005).

215. Plagens A, Tripp V, Daume M, *et al.* In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Res*. 42(8), 5125–5138 (2014).

216. Marchfelder A, Fischer S, Brendel J, *et al.* Small RNAs for defence and regulation in archaea. *Extremophiles*. 16(5), 685–696 (2012).

217. Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Hüttenhofer A. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. *Mol Microbiol*. 55(2), 469–481 (2005).

218. Soppa J, Straub J, Brenneis M, *et al.* Small RNAs of the halophilic archaeon Haloferax volcanii. *Biochem Soc Trans*. 37(Pt 1), 133–136 (2009).

219. Straub J, Brenneis M, Jellen-Ritter A, Heyer R, Soppa J, Marchfelder A. Small RNAs in haloarchaea: identification, differential expression and biological function. *RNA Biol*. 6(3), 281–292 (2009).

220. Bachellerie J-P, Rozhdestvensky T, Bortolin M-L, *et al.* Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc Natl Acad Sci USA*. 99(11), 7536–7541 (2002).

221. Eddy SR. Computational genomics of noncoding RNA genes. *Cell*. 109(2), 137–140 (2002).

222. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci USA*. 99(11), 7542–7547 (2002).

223. Jaschinski K, Babski J, Lehr M, *et al.* Generation and phenotyping of a collection of sRNA gene deletion mutants of the haloarchaeon Haloferax volcanii. *PLoS ONE*. 9(3), e90763 (2014).

224. Törö I, Basquin J, Teo-Dreher H, Suck D. Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile Archaeoglobus fulgidus. *J Mol Biol*. 320(1), 129–142 (2002).

225. Sittka A, Sharma CM, Rolle K, Vogel J. Deep sequencing of Salmonella RNA associated with heterologous Hfq proteins in vivo reveals small RNAs as a major target class and identifies RNA processing phenotypes. *RNA Biol*. 6(3), 266–275 (2009).

226. Krupovic M, Gonnet M, Hania WB, Forterre P, Erauso G. Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. *PLoS ONE*. 8(1), e49044 (2013).

227. Fischer S, Benz J, Späth B, *et al.* The archaeal Lsm protein binds to small RNAs. *Journal of Biological Chemistry*. 285(45), 34429–34438 (2010).

228. Maier L-K, Benz J, Fischer S, *et al.* Deletion of the Sm1 encoding motif in the lsm gene results in distinct changes in the transcriptome and enhanced swarming activity of Haloferax cells. *Biochimie*. 117, 129–137 (2015).

229. Hackl W, Fischer U, Lührmann R. A 69-kD protein that associates reversibly with the Sm core domain of several spliceosomal snRNP species. *J Cell Biol*. 124(3), 261–272 (1994).

230. Lee T, Feig AL. The RNA binding protein Hfq interacts specifically with tRNAs. *RNA*. 14(3), 514–523 (2008).

231. . The archaeal Lsm protein binds to small RNAs. *J Biol Chem*. 285, 34429–34438 (2010).

Chapter 2: Structure and RNA-binding of the Novel Sm-Like Archaeal Protein, *Pyrobaculum Aerophilum* SmAP2

Peter S Randolph[1] and Cameron Mura[1]

[1]University of Virginia, Department of Chemistry, Charlottesville, VA 22904

1 Abstract

In all three domains of life, Sm oligomers interact with RNA either as *chaperones* or as *scaffolding*, facilitating RNA···RNA and RNA···protein interactions. In both bacteria and eukarya, Sm homologs are necessary for RNA metabolism; and in bacteria, they are also instrumental in RNA-mediated post-transcriptional regulation. However, the *in vivo* roles of Sm-like archaeal proteins (SmAPs) remain unknown. The hyperthermophilic crenarchaeote *Pyrobaculum aerophilum* (*Pae*) encodes three SmAPs. The two known *Pae* SmAP structures (SmAP1, SmAP3) illuminated Sm protein evolution and assembly, and implied that these homologs may represent an ancestral form of the complexes that developed into the extant heteromeric Sm assemblies of eukaryotes, such as those at the heart of the spliceosome. We have recently investigated the remaining putative paralog assembly, *Pae* SmAP2. Biophysical characterization via hydrodynamic methods (AUC, ANSEC, SEC-MALS) suggests that *Pae* SmAP2 exists in multiple stable oligomeric states, which are also detectable via chemical cross-linking and mass spectrometry. We have determined the crystal structure of *Pae* SmAP2 in two separate (and rare) space groups. In the crystal, *Pae* SmAP2 forms on octamer, a novel oligomeric feature heretofore unseen in SmAPs. Binding assays with small RNA oligonucleotides show that *Pae* SmAP2 binds to oligo-adenine and oligo-uridine RNA, but not oligo-guanine or oligo-cytosine RNAs.

2 Introduction

Proteins from the Sm family are ubiquitous throughout all domains of life and are considered to be one of the earliest evolving proteins [1]. Sm proteins were first isolated and identified as small antigens involved in systemic lupus erythematosus [2,3]. Initial biochemical work showed that Sm protein were involved in the formation of large ribonucleoproteins (RNP) [4]. Bioinformatics confirmed the presence of Sm homologs in the archaeal branch, with much of the earlier Sm structural work (including the first complete oligomer) done on the Sm-like archaeal proteins (SmAPs) [5-7]. In a separate line of research, the bacterial protein host factor I (HFI) was identified as being necessary for the replication of RNA bacteriophage Qβ, and was re-named Hfq [8]. Hfq was found to bind RNA and exist *in vivo* as a stable hexamer [9,10]. Solving the structure of Hfq unified these separate research paths, showing Hfq shares the same fold as the eukaryal Sm and Sm-like and the archaeal SmAPs [11,12].

Sm and Sm-like proteins share a conserved tertiary structure known as the Sm-fold, but differ in their quaternary structure and functional role between domains. The Sm fold consists of a small five-stranded β-barrel preceded by an α-helix. The Sm fold contains two motifs, Sm 1 and Sm 2, which are separated by a non-conserved linker. The Sm 2 motif contains the β-strands 4 (β4) and 5 (β5), which form the interface between adjacent subunits during oligomer formation, contributing to its overall stability [13]. Loop 4 (L4) contains a variable region that is extended in eukaryal Sm proteins but shortened in bacterial Hfq. The Sm family shows immense plasticity in their oligomer formation; while most have been found as either hexamers (bacterial Hfq) or heptamers (eukaryal Sm and SmAPs) a pentamer and an octamer structure have been solved [14-18]. The bacterial Hfq forms stable homo-hexamers while the canonical Sm core proteins form hetero-heptamers through an intermediate phase as dimers or trimers, only forming their stable, fixed hetero-heptameric rings when bound with their target RNA (assisted by the SMN chaperone complex *in vivo*) [19,20].

Oligomeric variation results in different binding profiles between the bacterial Hfq and the eukaryal Sm proteins. While both Hfq and eukaryal Sm are known to bind uridine-rich single-stranded RNA (ssRNA) around the pore on the face *proximal* to the α-helix (L3 face), the eukaryotic heptamer has a wider pore than the hexameric Hfq, allowing the ssRNA to thread through the pore, rather than run along the face [21-23]. Keeping the faces separate allows Hfq to have additional binding surfaces, binding A-rich sequences on the *distal* face (L4 face), and non-specific RNA binding on the *lateral* rim [21]. Lateral rim binding has also been observed in SmAPs, with the *Pyrococcus abyssi* SmAP co-crystallizing with U_6 RNA [15]. SmAPs are closer in sequence and structure to the eukaryal canonical Sm core and Lsm proteins than the bacterial Hfq. Backbone overlays have shown *Pae* SmAP1 monomer and dimer structures to be almost identical to the SmD1·SmD2 and SmD3·SmB heterodimers [7,24] and all previous SmAP structures have been heptamers, though they form stable homo-oligomers similar to the bacterial Hfq. SmAPs have previously been shown to bind U-rich RNA, but whether they bind through the pore (eukaryotic Sm) or along the face (bacterial Hfq) is unknown [5,15,25].

While Sm and Sm-like proteins bind RNA and are involved in RNA processing they are utilized in different manners in different domains. Eukarya contain large molecular machinery, such as the spliceosome, which are built around interactions between a variety of Sm and Sm-like proteins and RNA sub-units [26]. Eukaryal canonical Sm core and Lsm proteins appear to act as *scaffolding* holding the catalytic RNA components of these complexes in place. Sm proteins were initially discovered as part of the uracil-rich small nuclear ribonucleoproteins (U snRNPs) which form the larger spliceosome complex [19]. U snRNPs cores have two major components, the specific snRNA component (U1 snRNA › U1 snRNP, U2 snRNA › U2 snRNP, …) and the canonical Sm core proteins. Canonical Sm core proteins nucleate snRNP formation, enclosing the U snRNAs around the shared Sm-binding site (Purine-A(U)$_{4\text{-}6}$G-Purine, flanked by two stem-loops), allowing a stable platform for the other components of the complex to build off [27]. While the majority of eukaryal Sm research has focused on their role in splicing, eukaryotic Sm and Lsm proteins underpin many RNA processing pathways including chromosome end maintenance (telomerase), ribosomal RNA processing (snoRNPs), and transfer RNA maturation (RNase P) [28].

Bacterial Hfq does not form large complexes, but instead operates as a free hexamer that functions as a *chaperone* during RNA···RNA and RNA···protein interactions, controlling post-transcription regulation [29]. Hfq has a flexible sequence recognition capacity, helping diverse small non-coding RNAs (sRNAs) bind their target mRNAs, either up or down-regulation translation. Hfq regulates polyadenylation, translation, and degradation of RNA *in vivo* through direct interactions with messenger RNA and/or small regulatory RNA. Hfq is necessary for pathways required for rapid adaptation to the environment including: stress response, nutrient uptake, biofilm formation, and virulence [30]. A common example of Hfq's regulatory role is the sRNA-regulated translation of the RNA polymerase subunit σ^s. A factor in environmental stress response, σ^s requires Hfq to facilitate the annealing of the *trans*-encoded regulatory sRNAs (DsrA, RprA, or ArcZ) to the target mRNA transcript (*rpoS*) [31-33]. In addition to its chaperone role, Hfq has also been shown to play an important role in RNA stability [34,35]. Specifically, Hfq can impede RNase E activity in the absence of sRNA by binding to a site near the RNase E cleavage site; when the sRNA is present Hfq facilitates annealing allowing for degradation by nucleases within the double stranded region (i.e. RNase III) or promote cleavage by RNase upstream of the regulatory region [36-38].

Despite the wealth of information on eukaryotic and bacterial Sm proteins, few studies have addressed SmAPs. To date, only two studies have examined the physiological role of SmAPs.

Co-immunoprecipitation studies on the Sm-like proteins in the Euryarchaote *Archaeoglobus fulgidas* (*Afu*) a role in tRNA intron processing [5,39]. *Afu* SmAPs were shown to interact with each other and associate with two forms of RNase P RNA (possibly precursor and mature) [5]; part of the RNase P RNP complex responsible for cleaving the leader sequence of pre-tRNA [40], possibly signifying a role similar to the eukaryotic Sm and Sm-like proteins. A more recent immunoprecipitation and Δ*smap* mutant study on *Haloferax volcanii*, showed *Hvo* SmAP binds to a variety of RNA, including sRNA within the cell. *Hvo* SmAP pulling down regulatory sRNAs hints at a role closer to *chaperoning* role of bacterial Hfq [41]. These two studies present opposing views of SmAP function, SmAP RNase P interactions are reminiscent of the eukaryal Sm proteins role in RNA processing, while the non-specific sRNA binding of *Hvo* SmAP resembles the *chaperoning* function of bacterial Hfq. Notably, these studies are confined to the Euryachaeota, with no current work examining SmAP function in either Crenarchaea or Thaumarchaea.

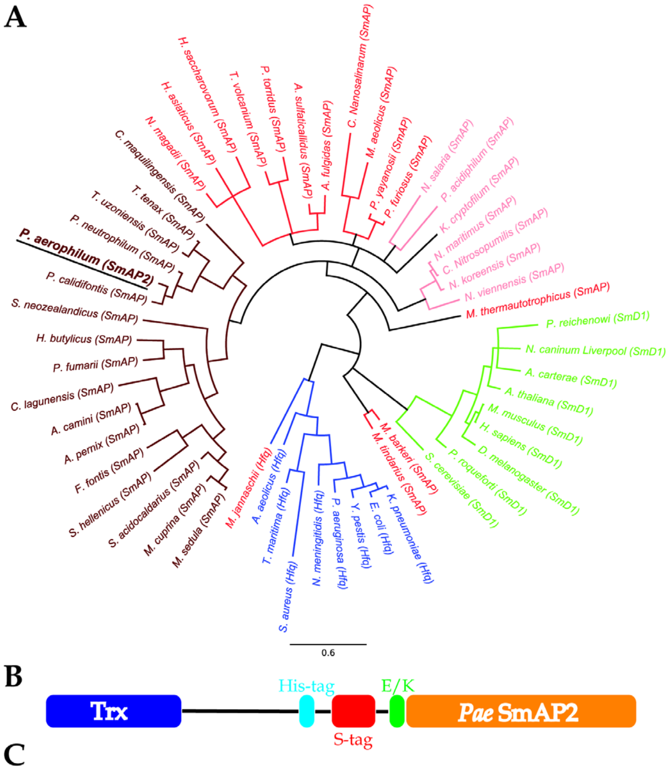
To elucidate the function of SmAPs we decided to examine the three SmAPs encoded by the hyperthermophillic crenarchaea *Pyrobaculum aerophilum*. *P. aerophilum* was chosen because of it is a deep-branching crenarchaea, with possible unique RNA metabolism. *P. aerophilum* encodes a variety of tRNA necessitating splicing (see Chapter 1, 5.4.1). *P. aerophilum* encodes three SmAPs, with unique features differentiating them. *Pae* SmAP1 is electrostatically negative, while *Pae* SmAP2 is highly positive, and *Pae* SmAP3 is augmented with a C-terminal domain that interacts with neighboring *Pae* SmAP3 C-terminal domains, forming a pair of joined SmAP3 heptamers in a head-to-head alignment. The structures of two *Pae* SmAPs (SmAP1 and SmAP3) have previously been solved, illuminated Sm protein assembly [7,18,25]. Here we examine the last Sm homolog of *Pae*, SmAP2. Biophysical characterization via hydrodynamic methods (AUC, ANSEC, SEC-MALS) suggests that *Pae* SmAP2 exists in multiple stable oligomeric states, which are also detectable via chemical cross-linking and mass spectrometry. We have determined the crystal structure of *Pae* SmAP2 in two separate (and rare) space groups. In the crystal, *Pae* SmAP2 forms on octamer, a novel oligomeric feature heretofore unseen in SmAPs.

3 Results

3.1 Cloning and Purification

Pae676 was successfully cloned (*Pae* SmAP2cm) and over-expressed in pET-32a(+) with a N-terminal thioredoxin tag followed by a polyhistidine tag, an S-tag, and an enterokinase cleavage site (Figure 1B, 1C). After initially cloning *Pae* SmAP2cm was successfully purified though a heat cut followed by immobilized metal affinity chromatography (cobalt) utilizing the his-tag. The purified complex was successfully digested with enterokinase and purified by cation exchange chromatography. There were two significant issues with *Pae* SmAP2cm purification: *i)* precipitation of the entire complex during preparation for digestion (dialysis into a low salt solution), lowering the yield, and *ii)* lack of a clean MALDI-TOF spectrum after digestion and cation exchange chromatography. Both of these issues are by-products of the enterokinase digestion: *i)* necessity to dialyze into a low-salt buffer for digestion causes the precipitation, *ii)* and improper digestion results in multiple products in the MALDI-TOF spectrum even though SDS-PAGE shows a single pure band (Figure 2A, 2C).

To remove the necessity for an enterokinase digestion step, *pae676* gene was sub-cloned into a pET-22a vector lacking tags needing cleavage (*Pae* SmAP2pr1). *Pae* SmAP2pr1 was successfully purified with a heat cut step, cation affinity chromatography in denaturing conditions, and gel filtration. SDS-Page and an improved MALDI-TOF (Figure 2D) confirm *Pae* SmAP2pr1 purity.



C

```
        10         20         30         40         50
MSDKIIHLTD DSFDTDVLKA DGAILVDFWA EWCGPCKMIA PILDEIADEY
        60         70         80         90        100
QGKLTVAKLN IDQNPGTAPK YGIRGIPTLL FLKNGEVAAT KVGALSKGQL
       110        120        130        140        150
KEFLDANLAG SGSGHMHHHH HHSSGLVPRG SGMKETAAAK FERQHMDSPD
       160        170        180        190        200
LGTDDDDKAM AKAQQVKLPS PIKVLTKMLN KEIIARLKGG VAVKGVLTAY
       210        220        230        240   244
DGCMNLVLND AAELDKSGEP KTRYGRIVIR GSQVIYVSTS EVTP
```

Figure 1. Sm family phylogenetic tree and *Pae* SmAP2 constructs. (A) Phylogenetic tree of Sm proteins from all families of life. Bacterial Sm-like proteins (Hfq) are shown in blue and root the tree, with eukaryal SmD1 proteins in green, and Sm-like Archaeal proteins (SmAPs) in red. SmAPs are subdivided into crenarchaea (dark red), euryarchaea (red), and thaumarchaea/nanoarchaea (light red). The different families cluster into separate branches, showing while Sm proteins are closely related in structure they differ between families in sequence. (B), (C) *Pae* SmAP2cm sequence and construct schematic. SmAP2 preceded by an N-terminal thioredoxin tag (blue), a polyhistidine tag (purple), an S-tag (red), and an enterokinase cleavage site (green). *Pae* SmAP2pr1 construct contains just the SmAP2 section (orange) beginning with the underlined methionine.

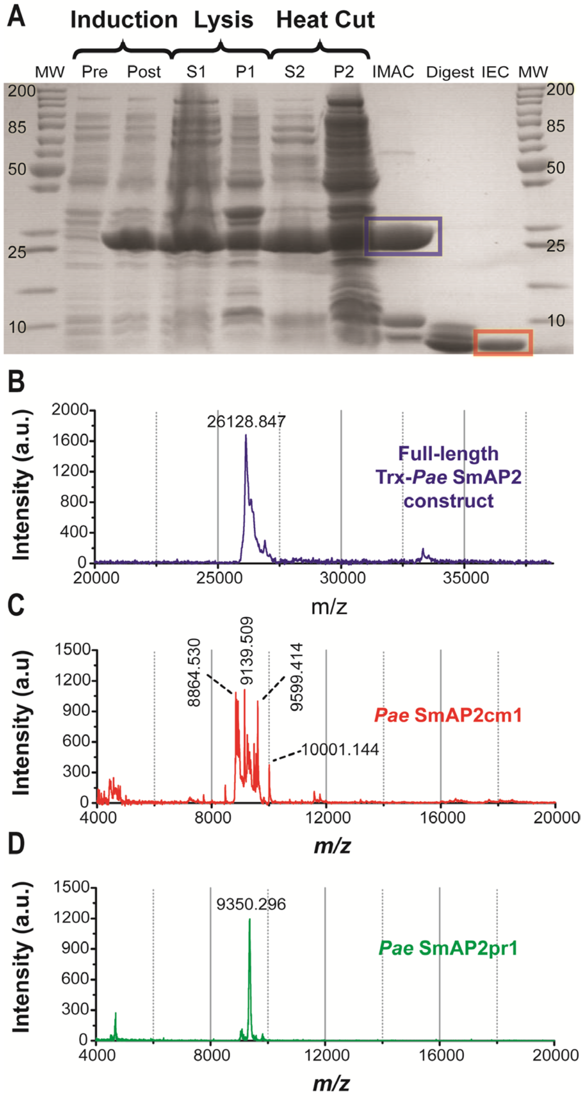


Figure 2. Lysis and purification of SmAP2. (A) SDS-PAGE overview of *Pae* SmAP2cm1 from growth through final purification. From left to right: pre induction (Pre), post induction (Post), post lysing supernatant (S1), post lysing pellet (P1), post heat-cut supernatant (S2), post heat-cut pellet (P2), pooled IMAC peak (IMAC), post enterokinase digestion (Digest), and pooled IEC sample (IEC). Purified *Pae* SmAP2cm1 full construct is highlighted by the blue box; while post digestion, purified *Pae* SmAP2 is highlighted by the red box. (B) MALDI TOF MS spectrum of purified *Pae* SmAP2cm1 construct before digestion shows a single peak at the correct molecular weight. (C) MALDI TOF MS spectrum of purified SmAP2 post digestion of *Pae* SmAP2cm1 construct. While the SDS-PAGE appears clean (1A red box), MALDI TOF spectrum shows a number of peaks around the expected molecular weight. (D) MALDI TOF spectrum of purified *Pae* SmAP2pr1 construct has a single peak at the expected *Pae* SmAP2 molecular weight.

3.2 Bioinformatics

Initial bioinformatics analysis found *Pae* SmAP2 to cluster within the crenarchaeota branch, with other *Pyrobaculum* species (Figure 1A). The bacterial Hfq, eukaryal Sm, and SmAPs clustered together into small clades, except for *Methanococcus jannaschi* SmAP, which had previously been shown to be closer in sequence and structure with the bacterial Hfq than other SmAPs [42].

3.3 Fluorescence Polarization

Binding affinities of *Pae* SmAP2···FAM-r(U) RNA and *Pae* SmAP2···FAM-$(A)_{18}$ RNA were determined by measuring fluorescence polarization (Figure 4A). Varying the amount of *Pae* SmAP2 resulting in a dissociation constant (K_d) of 0.198 µM (N=3) for FAM-$r(U)_6$ and 1.22 µM for FAM-$(A)_{18}$ RNA (N=3) after fitting the data to the Boltzmann equation (Figure 4A). Competition assays revealed that FAM-$(U)_6$ and FAM-$(A)_{18}$ bind non-competitively (data not shown). Additional fluorescence polarization assays demonstrated that *Pae* SmAP2 does not display significant binding to FAM-$r(C)_6$ or FAM-$r(G)_6$. *Pae* SmAP2 binding adenine-rich RNA is length dependent, with significant binding for FAM $r(A)_{18}$ (above), but no appreciable binding for FAM-$r(A)_6$.

3.4 Oligomerization

3.4.1 Hydrodynamic Methods

Pae SmAP2 oligomeric state was examined using hydrodynamic methods: *i)* analytical size exclusion (AnSEC), *i)* size-exclusion chromatography - multi-angle light scattering (SEC-MALS), and *i)* analytical ultracentrifugation (AUC) (Table 1). Depending on the specific

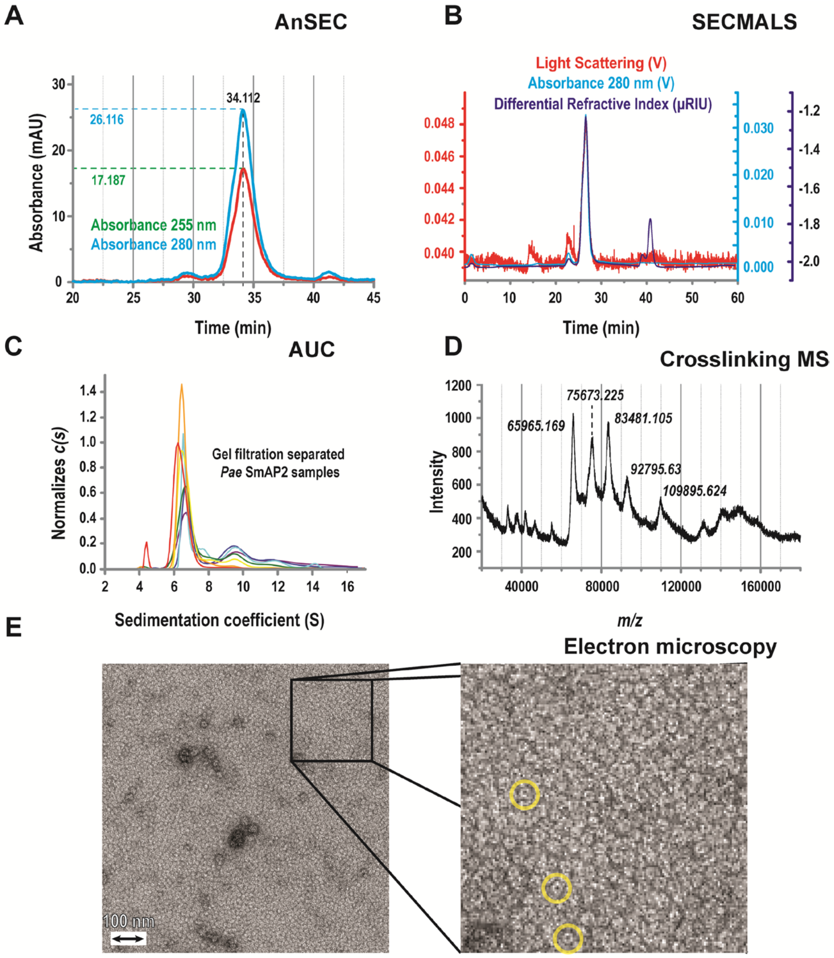


Figure 3. Biophysical characterization of *Pae* SmAP2 oligomeric state. (A) Representative AnSEC chromatogram with *Pae* SmAP2 peak. Molecular weight is generated from elution time compared to known standards. (B) Representative SEC-MALS chromatogram. The major *Pae* SmAP2 peak in homogeneous, while the smaller, early eluting peaks are heterogeneous aggregation. (C) Sedimentation AUC on *Pae* SmAP2 samples separated by size exclusion reveal *Pae* SmAP2 can have multiple stable oligomers. (D) Representative crosslinking followed by MALDI TOF MS spectrum displays multiple resolved *Pae* SmAP2 peaks corresponding to a heptamer (7), octamer (8), and nonamer (9). Additional crosslinking MALDI TOF MS spectra displayed a decamer (10), undecamer (11), and duodecamer (12). (E) Electron microscopy image revealing *Pae* SmAP2 forms stable rings in solution, highlighted in yellow.

purification run, *Pae* SmAP2 eluted at between an octamer and a decamer during AnSEC, based on a generated standard curve of protein calibrants. All *Pae* SmAP2 samples ran from one purification prep eluted at consistent volumes; but the elution volume varied between preps. This higher than expected molecular weight for *Pae* SmAP2 was also observed by SEC-MALS, which does not depend on outside standards (Table 1). SEC-MALS suggested that the bulk *Pae* SmAP2 in solution was isomorphous (i.e. all in the same oligomeric form), with one main peak. A heterogeneous larger molecular weight peak (between 20 and 30 *Pae* SmAP2 subunits) was also seen, possibly a result of aggregation (Figure 3B). AUC returned possible oligomers of 9 ± 1 and 12 ± 1 *Pae* SmAP2 subunits. AUC appeared to have multiple stable oligomers co-existing in solution simultaneously. Additionally, *Pae* SmAP2 was examined by electron microscopy, which showed discrete, stable rings in solution, but the resolution was insufficient to determine the oligomeric state in solution (Figure 3E).

3.4.2 Crosslinking with MALDI TOF MS

Pae SmaP2 was cross-linked (formaldehyde and glutaraldehyde) followed by MALDI-TOF MS. The resulting MALDI TOF MS spectra showed multiple baseline resolved peaks, representing multiple possible oligomeric states. A representative crosslinking MALDI TOF MS spectra in Figure 3D has well resolved peaks for the expected heptamer, but also additional peaks for an octamer, a nonamer, and a smaller peak for a decamer. Crosslinking MALDI TOF MS spectra were collecting with peaks from 7 to 13 *Pae* SmAP2 subunits.

3.4.3 RNA Binding Effect on Solution State

Incubating *Pae* SmAP2 with either U_5 RNA, A_{18} RNA before AnSEC did not significantly affect the elution time (Figure 4 B). *Pae* SmAP2 run without RNA eluted at a calculated molecular

Table 1. Oligomeric states observed by AnSEC, SEC-MALS, AUC, and crosslinking followed by MALDI TOF MS of *Pae* SmAP2

Technique	***Pae* SmAP2** **Oligomeric State**
AnSEC	8, 9, 10
SEC-MALS	8, 9, 10
AUC	9 ± 1, 12 ± 1
Crosslinking MS	7, 8, 9, 10, 11, 12

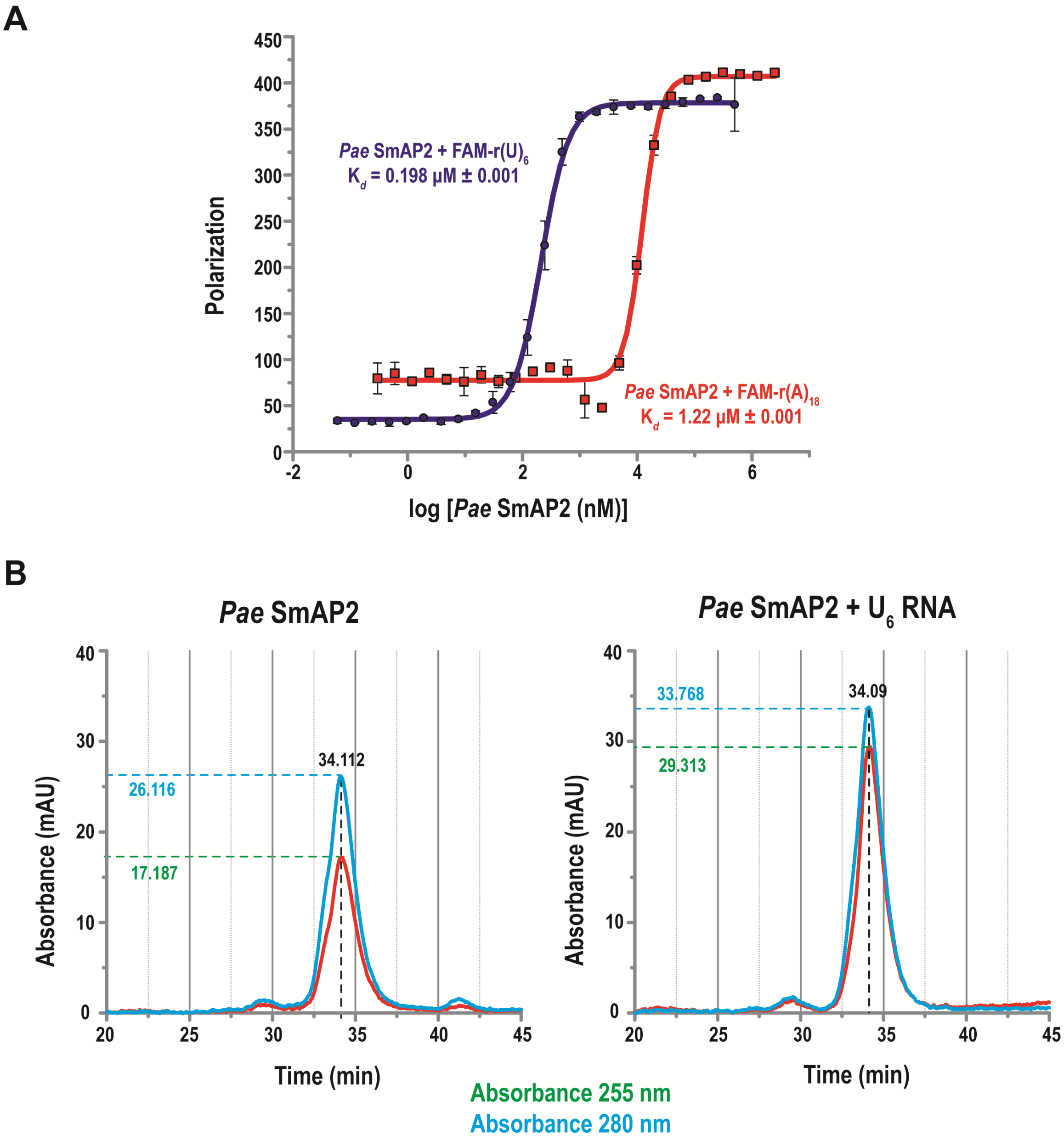


Figure 4. RNA binding of *Pae* SmAP2 and its role in oligomerization. (A) Flourescence anisotropy, fluorescence polarization (FA/FP) binding assays reveal that *Pae* SmAP2 binds both Uridine- and Adenine-rich RNA, with a ten-fold higher affinity for U-rich RNA. (B) AnSEC of *Pae* SmAP2 incubated with RNA does not shift *Pae* SmAP2 into a new oligomeric form.

Table 2. Molecular weight observed for *Pae* SmAP2 samples incubated with RNA. ΔMW is the difference between the sample molecular weight and *Pae* SmAP2 run without binding partner (row 1). (SmAP)*ₙ* is observed oligomeric state, molecular weight divided by the *Pae* SmAP2 monomer weight of 9350 Da. A255/A280 ratio of the absorbance of light at 255 nm and 280 nm reports the amount RNA in a sample. The higher the A255/A280 ratio, the more RNA is present.

Sample	**MW (Da)**	**ΔMW (Da)**	**(SmAP)$_n$**	**A255/A280**
SmAP2	86687.3 ± 141.6	-----	9.27	0.65
SmAP2 + r(A)$_{18}$*	87845.1	1157.8	9.39	0.77
SmAP2 + r(U)$_6$	87457.5 ± 121.3	770.2	9.35	0.87
SmAP2 + r(U)$_6$ + r(A)$_{18}$*	88625.6	1938.3	9.48	0.86

* Sample collecting in duplicate instead of triplicate

weight of ~86.7 kDa (depending on the prep and run). After incubation *Pae* SmAP2 samples showed increased absorption at both 255 nm and 280 nm, with a larger increase at 255 nm (higher A260/A280 ratio) and a small shift in the elution volume, confirming RNA binding (Figure 4B).

3.5 Structural Determination

3.5.1 Crystal Development and Data Collection

Initial crystal trials with the ammonium sulfate grid screen (Hampton Research) resulted in successful crystal formation with at 1.6 M ammonium sulfate, 0.1 M citrate at pH 4.0. These conditions were expanded and produced diffracting, reproducible crystals in a range of 1.33 to 1.86 M ammonium sulfate, with and without 0.1 M citrate at pH 4.0. Crystals formed in two different habits. One habit was cubic (Figure 5A) and the other plate-like (2D growth) (Figure 5B). Both habits were indexed, processed, and scaled using XDS [43]. Space groups were identified by XDS and confirmed through systematic absences as belonging to the $P42_12$ (plate-like) and P23 (cubic shape) spacegroups. P23 and $P42_12$ are extremely rare space groups, comprising only 0.00055% and 0.0035% of the structures in the PDB, respectively. *Pae* SmAP2 crystals in the $P42_12$ space group diffracted to a higher resolution (to 1.85 Å) than P23 (3.10 Å) (Figure 5). Before phasing trials, datasets were run through Xtriage and the UCLA anisotropy server and showed no signs of twinning, or significant anisotropy [44,45]. Calculated Matthews coefficients suggested the $P42_12$ asymmetric unit (ASU) contains between 8 (2.2 % probability) and 19 (3.5 %) *Pae* SmAP2 monomers, with 14 being most probable (13.5%); and P23 ASU contains between 11 (2.4 %) and 24 (2.7%), with joint highest probability for 18 and 19 *Pae* SmAP2 monomers (10.6% probability) [46].

Datasets from both space groups were used in molecular replacement trials with multiple Sm proteins probes covering a variety of oligomers (5-8). Initial molecular replacement phasing

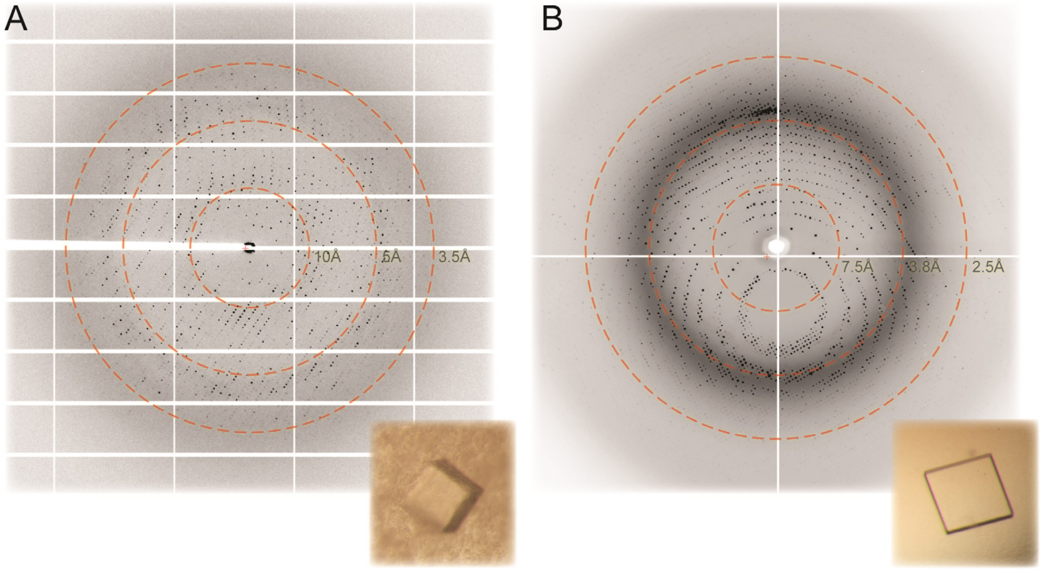
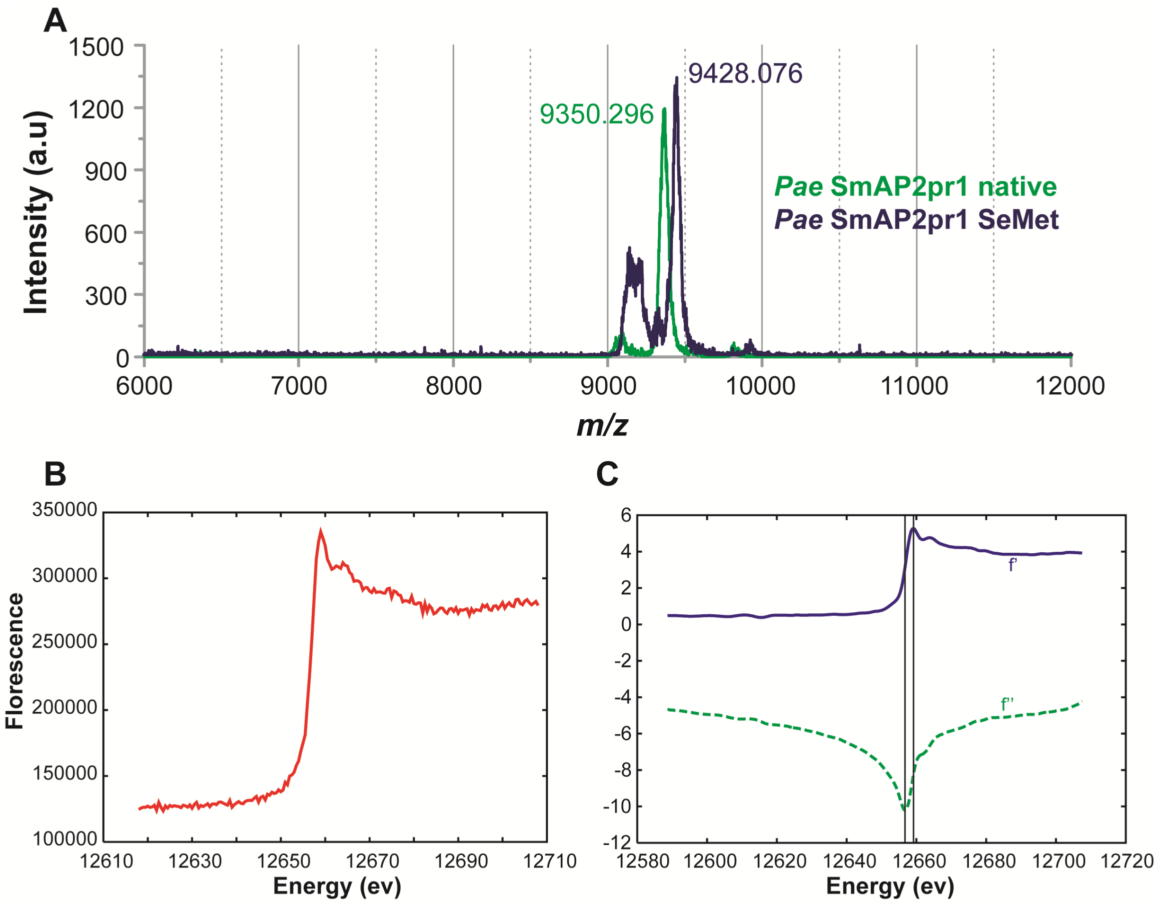


Figure 5. Crystallization and diffraction of *Pae* SmAP2. (A) Diffraction of P23 crystals (inset) developed in 1.6 M ammonium sulfate and 0.1 M citrate pH 4.0 collected at APS 24-IDC beamline shows well-resolved reflections to 3.08 Å. (B) Diffraction of P42$_1$2 crystals (inset) developed in 0.9 M ammonium sulfate and 0.1 M citrate pH 4.0 collected at ALS 5.0.2 shows well-resolved reflections to 1.85 Å.



Wavelengths -	**Peak**	**Edge**	**Low**	**High**
Wavelength -	0.97930 Å	0.97940 Å	0.98271 Å	0.97568 Å
Energy -	12659.18 eV	12656.72 eV	12616.54 eV	12707.48 eV
f' -	-8.20 e-	-10.22 e-	-5.43 e-	-4.21 e-
f" -	5.28 e-	3.12 e-	0.40 e-	3.92 e-

Figure 6. Selenomethionine incorporation and fluorescence scan for *de novo* phasing. (A) MALDI TOF MS confirms selenomethionine incorporation in *Pae* SmAP2. (B) Fluorescence scan to determine the x-ray absorption peak and edge energies for collection.

trials were unsuccessful, so we turned to *de novo* phasing. MALDI TOF MS confirmed incorporation of selenomethionine (Figure 6A) and selenomethionine crystals were successfully developed based on native conditions. Fluorescence scans confirmed stable, detectable selenomethionine in the crystals, and determined peak and inflection wavelengths of 0.97930 and 0.97940 Å, respectively (Figure 6B, 6C). Initially, neither P23 nor $P42_12$ developed crystals with sufficient diffraction resolution to allow *de novo* phasing. Additive screening improved diffraction, until a single $P42_12$ selenomethionine crystal collected at the NE-CAT 24C line, diffracted to 2.7 Å, within the limits of resolution for *de novo* phasing. Datasets were collected at the peak, inflection, high remote energies. Significant anomalous signal (mean anomalous difference > 1) was observed, making them suitable for phasing (see Table 3 for statistics).

3.5.2 Structure Solution

The structure of *Pae* SmAP2 was solved in $P42_12$ by d*e novo* phasing in SHELX (HKL2MAP) [47,48]. Based on the number of methionines in the *Pae* SmAP2 sequence we predicted 2 seleniums per monomer (start methionine residue is usually too labile for use in structure determination) and between 16 and 28 selenomethionines present in the ASU, based off Matthews coefficient. SHELXD (integrating Patterson and direct methods to determine anomalous scattering locations) located 24 seleniums in the ASU and determined that *Pae* SmAP2 crystallizes with 12 subunits per ASU in $P42_12$, arranging as an octamer with two flanking dimers (Figure 10B). The unexpected oligomeric state is most likely reason molecular replacement was unsuccessful. The octameric *Saccharomyces cerevisiae* Lsm3 Loop L4 is quite extended compared to the resulting *Pae* SmAP2 structure, which could have hampered molecular replacement [16]. *Pae* SmAP2 $P42_12$ selenomethionine datasets were isomorphous with previous native datasets, allowing the resolution to be extended to 1.85Å (statistics Table 3 and Table 4). The *Pae* SmAP2 $P42_12$ structure was refined with Phenix (Autobuild, phenix.refine), and coot. *Pae* SmaP2 P23 structure was solved by molecular replacement utilizing the $P42_12$ octamer as the probe and refined using Phenix (Autobuild, phenix.refine) and coot. A large, globular artifact was present in later rounds of refinement in the *Pae* SmAP2 P23 structure, which was resolved by switching to REFMAC for further refinement. Data collection statistics and refinement statistics are in Table 3 and Table 4

3.5.3 *Pae* SmAP2 Monomer

Pae SmAP2 folds as a strongly, bent five stranded β-sheet preceded by a small α-helix, features common to all Sm proteins (Figure 7A). β-strands 2, 3, and 4 have the highest degree of bend, with their two-ends almost overlapping. The linker connecting β-strands 4 and 5 runs along the groove created by the fold, creating a closed system and positioning β-strand 5 to run parallel to β-strand 4 of the adjacent monomer forming the oligomer interface, which is consistent with other Sm proteins. Five hydrogen bonds are formed between the backbones of the β4 residues: Arg-68, Val-70, Arg-72; and β5 residues: Ile-77, Val-79, Thr-81, with additional, but minimal contributions from hydrophobic interactions between Val-70·Val-79 and Ile-71·Tyr78 (Figure 7E). Because the majority of the interactions between β-strands 4 and 5 are between the backbone, this is one of the areas within the minimal Sm fold with the highest sequence variation between domains, which could affect oligomerization.

Structural alignment of the *Pae* SmAP2 monomer backbone with Sm proteins from archaea (*Pae* SmAP1), bacteria (*S. aureus* Hfq) and eukarya (*H. sapiens* SmD3) demonstrate the conservation of the Sm fold between domains (Figure 7D) [18,23,49]. The only area with significant differences is Loop L4, which is extended in the archaeal SmAPs and eukaryal SmD3, compared to the bacterial Hfq. B-factors of *Pae* SmAP2 indicate that Loops L2 and L4 are most likely dynamic compared to the rest of the monomer (Figure 7F). The mobility of the L2 and L4 Loops can also be seen in the variation between monomers within the ASUs. Aligning all 28 monomers (12 from $P42_12$, 16 from P23) results in an overall RMSD of 0.739 Å, with most of the positional variation constrained to Loops L2 and L4 (Figure 7B, 7C).

3.5.4 *Pae* SmAP2 Crystal Structure Oligomerization

Pae SmAP2 octamer is 70.0 Å in diameter (measured from Cα of extended residue Glu-83) with a pore diameter of 17.2 Å across at chokepoint at the extended Lys-30 side chain, (26.0 Å from Lys-30 Cα). This is significantly larger than heptameric SmAPs (*Pae* SmAP1 7.9 Å) or the hexameric Hfq (*Sau* Hfq 5.7 Å), but comparable to the Human U4 snRNP Sm ring (14.8 Å) (Figure 8B). *Pae* SmAP2 is about 30 Å in thickness with the flexible *N*-terminal tail extending another 17 Å (Figure 7A). Interestingly, even though the monomers align with a 0.739 Å RMSD, there is a small, but obvious shift (~6 Å) in the overall octamer when one monomer is aligned between P23 and $P42_12$, which could account for the difference in crystal packing (Figure 8A). The ASU of $P42_12$ forms an octamer flanked by two dimers (Figure 9B), while P23 ASU contains two complete octamers (Figure 9A), the first Sm structures to have complete octamers in their ASU. Both of the *Pae* SmAP2 crystal forms are on the high end of the Matthews coefficients (2.81 for P23 and 3.52 for $P42_12$), containing a large amount of solvent in the crystal (56.2% solvent for P23 and 65.1 % for $P42_12$).

In both space groups, *Pae* SmAP2 octamers form identical higher-ordered 'box' oligomers, which are constructed of 6 octamers, one octamer per face (Figure 10A). Each *Pae* SmAP2 octamer is arranged with its L3 face directed out, and the L4 face, α-helix, and N-terminal tail facing inward. *Pae* SmAP2 boxes packs in two different forms, one with orderly arrays (P23) (Figure 9B, 9C) and the other with a small turn between box layers ($P42_12$) (Figure 9E, 9F). In P23, a second inverted box is formed by faces of the adjacent boxes, with the L3 face on the inside. In $P42_12$, each box is distinct and the faces do not line up, existing as discrete units. The *Pae* SmAP2 box is held together

Table 3. Selenomethionine *Pae* SmAP2 P42$_1$2 dataset statistics for multi-wavelength anomalous dispersion (MAD) *de novo* phasing

Diffraction Statistics	***Pae* SmAP2** **Peak**	***Pae* SmAP2** **Inflection**	***Pae* SmAP2** **High Remote**
Diffraction Source	APS NE-CAT 22IDC	APS NE-CAT 22IDC	APS NE-CAT 22IDC
Wavelength (Å)	0.97930	0.97940	0.97090
a, b, c (Å)	132.5, 132.5, 157.49	132.33, 132.33, 157.34	132.5, 132.5, 157.24
α, β, γ	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution Range (Å)	80.56 – 2.72	93.63 – 2.71	80.52 – 2.71
Completeness (%)	99.8 (98.3)	99.5 (94.7)	99.5 (95.6)
$\langle I/\sigma(I)\rangle$	23.9 (4.1)	23.4 (4.6)	19.7 (2.9)
R_{sym}[†]	8.8 (58.6)	8.9 (51.1)	10.7 (84.2)
R_{meas}[‡]	9.3 (62.3)	9.5 (54.3)	11.3 (89.4)
R_{pim}[¥]	3.1 (20.7)	3.2 (18.2)	3.7 (29.4)
CC(1/2)[§]	99.9 (89.7)	99.9 (90.6)	99.8 (79.4)
Anomalous Correlation[ɸ]	54 (4)	63 (7)	31 (1)
Significant Anomalous[£]	1.49 (0.77)	1.67 (0.83)	1.12 (0.73)

[†] $R_{\text{sym}} = (\sum_{hkl} \alpha \sum_i |I_i(hkl) - \langle I_i(hkl)\rangle|)/(\sum_{hkl} \sum_i I_i(hkl))$, where $I_i(hkl)$ is the intensity of the i^{th} observation of reflection hkl, $\langle\cdot\rangle$ denotes the mean of symmetry-related (or Friedel-related) reflections, and the coefficient $\alpha = 1$; the outer summations run over only unique hkl with multiplicities greater than one.

‡ R_{meas} is defined analogously as R_{sym}, save that the prefactor $\alpha = \sqrt{N_{hkl}/(N_{hkl}-1)}$ is used; N_{hkl} is the number of observations of reflection hkl (index $i = 1 \rightarrow N_{hkl}$).

¥ $R_{\text{p.i.m.}}$, the precision-indicating merging R-factor, is defined as above but with the prefactor $\alpha = \sqrt{1/(N_{hkl}-1)}$.

§ $CC_{1/2}$ is the correlation coefficient between intensities chosen from random halves of the full dataset.

ɸ Percentage of correlation between random half-sets of anomalous intensity difference.

£ Mean anomalous difference in units of its estimated standard deviation ($|F(+) - F(-)|/\sigma$). $F(+)$, $F(-)$ are structure factor estimates obtained from the merged intensity observations in each parity class.

Table 4. Native datasets and structure refinement statistics

Diffraction Statistics	***Pae* SmAP2 P4$2_1$2**	***Pae* SmAP2 P23**
	Native	**Native**
Diffraction Source	ALS 5.0.2	APS NE-CAT 22IDC
Wavelength (Å)	1.100	0.9791
a, b, c (Å)	131.57, 131.57, 157.28	173.44, 173.44, 173.44
α, β, γ	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Resolution Range (Å)	93.15 – 1.85	86.75 – 3.10
Completeness	99.8 (98.6)	99.9 (99.5)
$\langle I/\sigma(I)\rangle$	22.9 (2.4)	18.6 (2.8)
R_{sym}[†]	7.1 (114.1)	10.3 (77.9)
R_{meas}[‡]	7.5 (119.7)	10.9 (82.9)
R_{pim}[¥]	2.3 (35.8)	3.6 (28.1)
CC(1/2)[§]	99.9 (73.0)	99.9 (82.7)
Refinement Statistics	***Pae* SmAP2 P4$2_1$2**	***Pae* SmAP2 P23**
R_{WORK}/R_{FREE}	19.68/22.60	18.60/23.56
R.M.S.		
Bonds (Å)	0.0088	0.011
Angles (˚)	1.157	1.548
Ramachandran (%)		
Favored	97.0	96.2
Allowed	3.0	3.6
Outliers	0.0	0.2
Avg. B-factor	37.70	64.8

† $R_{\mathrm{sym}} = (\sum_{hkl} \alpha \sum_i |I_i(hkl) - \langle I_i(hkl) \rangle|)/(\sum_{hkl} \sum_i I_i(hkl))$, where $I_i(hkl)$ is the intensity of the i^{th} observation of reflection hkl, $\langle \cdot \rangle$ denotes the mean of symmetry-related (or Friedel-related) reflections, and the coefficient $\alpha = 1$; the outer summations run over only unique hkl with multiplicities greater than one.

‡ R_{meas} is defined analogously as R_{sym}, save that the prefactor $\alpha = \sqrt{N_{hkl}/(N_{hkl} - 1)}$ is used; N_{hkl} is the number of observations of reflection hkl (index $i = 1 \rightarrow N_{hkl}$).

¥ $R_{\mathrm{p.i.m.}}$, the precision-indicating merging R-factor, is defined as above but with the prefactor $\alpha = \sqrt{1/(N_{hkl} - 1)}$.

§ $CC_{1/2}$ is the correlation coefficient between intensities chosen from random halves of the full dataset.

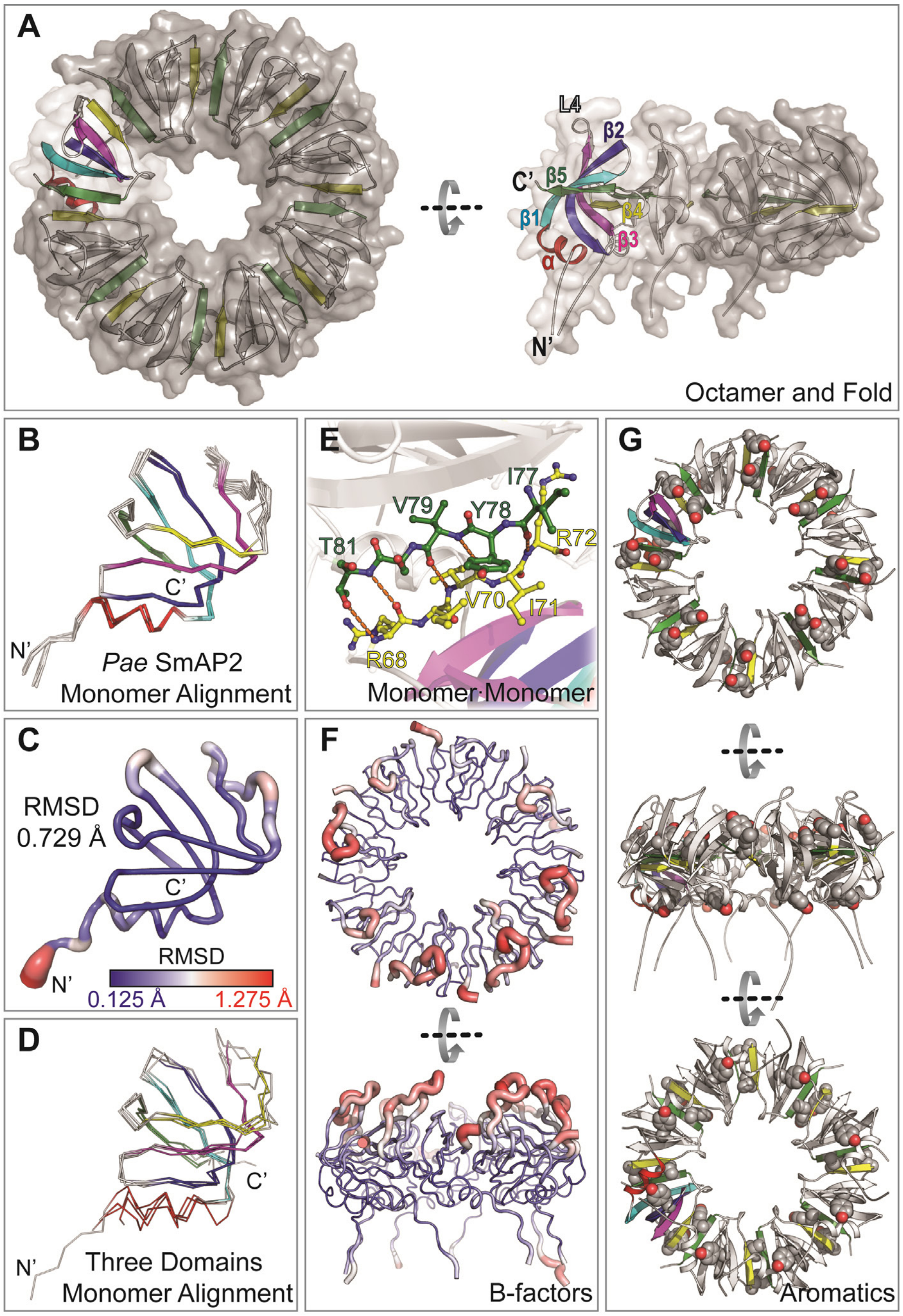

A
L4
β2
C'
β5
β1
β4
β3
α
N'
Octamer and Fold
B
C'
N'
Pae SmAP2
Monomer Alignment
C
RMSD
0.729 Å
C'
N'
RMSD
0.125 Å
1.275 Å
D
C'
N'
Three Domains
Monomer Alignment
E
I77
V79
Y78
T81
R72
V70
I71
R68
Monomer-Monomer
F
B-factors
G
Aromatics

Figure 7. Structure analysis of *Pae* SmAP2. (A) *Pae* SmAP2 adopts the conserved Sm fold and crystallizes as an octamer. (B)(C) Alignment of *Pae* SmAP2 monomers from both the P23 and $P42_12$ structures show minimal dynamics except at the Loops L2 and L4. (D) Alignment of *Pae* SmAP2 with representative bacterial Hfq (*Staphylococcus aureus*), eukaryal (*Homo sapiens* SmD3), and archaea (*Pae* SmAP1) demonstrate the conservation of the Sm fold. (E) β-strand 4 and 5 form the interface between monomers, with backbone interactions driving oligomerization. (F) B factors of *Pae* SmAP2 reveal mobile sections. (G) *Pae* SmAP2 only encodes 3 aromatic residues (Tyr-42, Tyr-66, and Tyr-78) which could bond with RNA, Tyr-78 is buried beneath the L4 loop while Tyr-66 and Tyr-42 are exposed on opposite faces of the lateral rim (Tyr-42 L3 face, Tyr-66, L4 face).

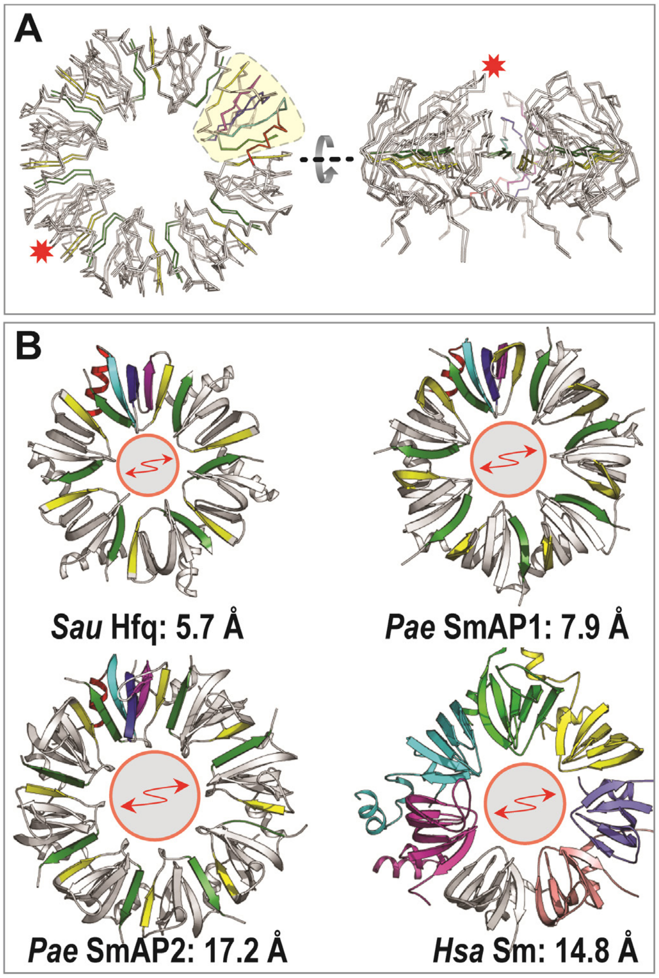


Figure 8. *Pae* SmAP2 P42$_1$2 and P23 octamers alignment and pore diameter. (A) Aligning all monomers of *Pae* SmAP2 results in an RMSD of only 0.729 Å, but if one of the monomers in an octamer is aligned (highlighted in yellow), the octamers do not align, with the backbones shifted by ~6 Å on the opposite monomer. (B) The varying pore diameters of Sm proteins can differentiate between the bacterial and eukaryal binding profiles. The small pore diameter of the bacterial Hfq (top left, *Sau* Hfq) prevents RNA from threading, separating the faces into two distinct binding surfaces. The larger pore diameter of the eukaryal Sm (bottom right, *Hsa* Sm) allows threading, with one RNA domain spanning both faces. *Pae* SmAP1's smaller pore appears to place it in the bacterial profile, while the larger *Pae* SmAP2 pore is reminiscent of the eukaryal pore.

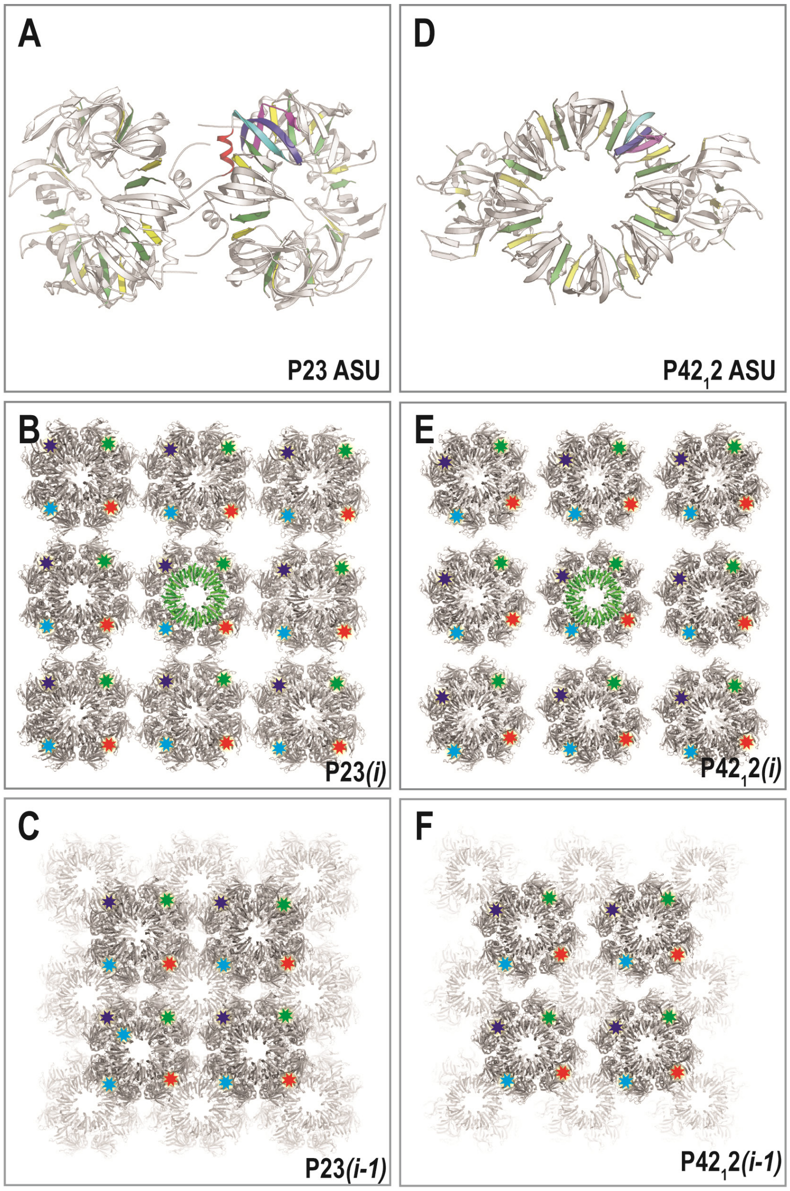


Figure 9. ***Pae* SmAP2 crystal packing.** *Pae* SmAP2 crystallizes in two rare space groups as an octamer. (A) The P23 asymmetric unit (ASU) consists of two full octamers, while (D) P42$_1$2 ASU consists of one octamer and flanked by two dimers. Both space groups form higher ordered six sided boxes but the P23 (B)(C) are arranged in a straight array, with each face in the same plane as the equivalent face (marked by same color stars). (E)(F) P42$_1$2 boxes have a small rotation in the array, with the equivalent faces slightly out of plane.

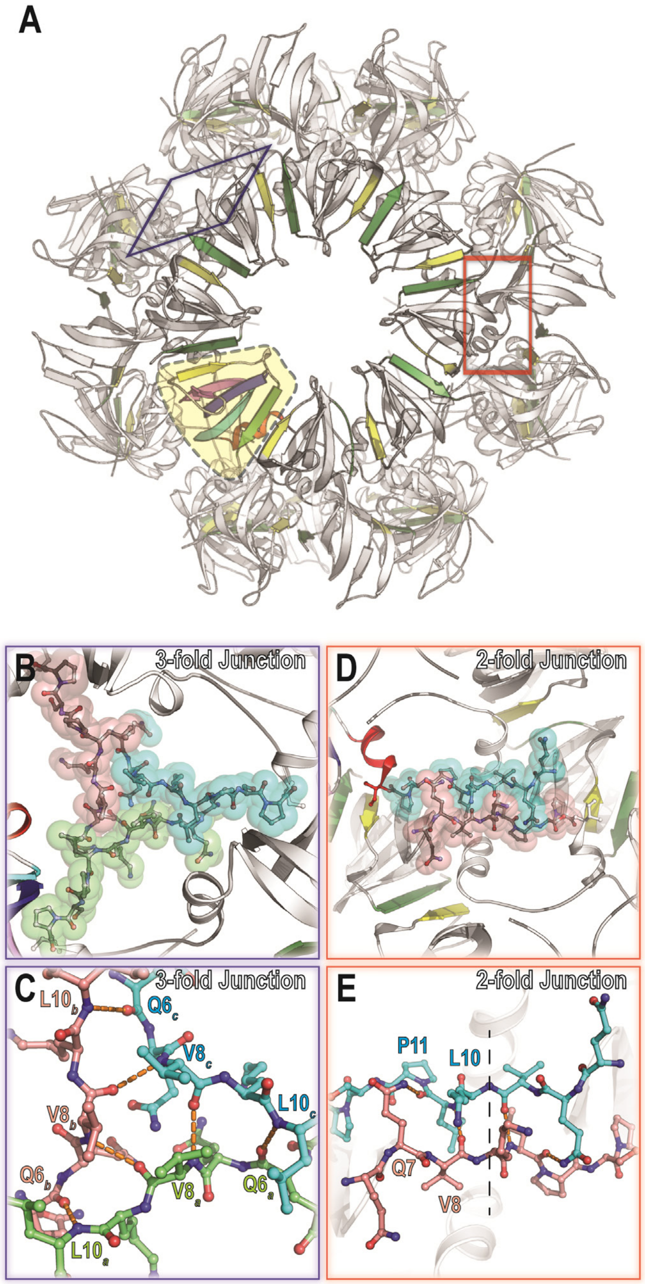


Figure 10. *Pae* SmAP2 crystallizes in higher order boxes stabilized by the N-terminal tail. (A) Both crystal habits ($P42_12$ and P23) form a higher-ordered 'box' oligomer, (either three P23 ASU's or four $P42_12$ ASU's) with an octamer on each face. The 'box' is held together by the *N*-terminal tail, which is generally too disordered in Sm proteins for structure determination, but in the 'box' alternates between (B)(C) forming the contacts for the corner and (D)(E) edge.

through interactions of adjacent N-terminal tails (alternating between the corner and the edge). Sm protein N-terminal tails are usually highly disordered, but in both *Pae* SmAP2 P23 and P42$_1$2 are well-resolved. At the edges of the box, two *Pae* SmAP2 N-terminal tails run parallel to each other, with mirrored paired residues interacting: *i)* backbone amide of Gln-7 hydrogen bonds to the backbone carbonyl oxygen of Pro-11, *ii)* hydrophobic interactions between Val-8 and Leu-10, in addition to their backbones forming a hydrogen bond, and *iii)* Lys-9 hydrogen bonding (backbone and side-chain) with Lys-9 from the adjacent tail (Figure 10A, 10B). The same patch of residues interacts in the corner of the box, except, instead of a two-fold axis, there is a three-fold axis, with the N-terminal tails forming a triangle. Each *Pae* SmAP2 N-terminal tail forms 2 hydrogen bonds to each neighbor (4 bonds per tail), for a total network of 6 hydrogen bonds. If we designate the *Pae* SmaP2 corner N-terminal tails *a*, *b*, and *c*; the bonding network progresses in a loop, *i)* Leu-10$_a$ backbone amide to Gln-6$_b$ backbone carbonyl oxygen; *ii)* Leu-10$_b$·Gln6$_c$, and *iii)* Leu-10$_c$·Gln-6$_a$. This pattern is conserved for Val-8 backbone carbonyl oxygen to the adjacent Val-8 backbone amide (*a*·*b*, *b*·*c*, *c*·*a*). In addition, the 3 corner Val-8s contribute hydrophobic interactions on the outer face of the triangle; and on the inner face, the 3 Gln-7s contribute both hydrophobic interactions and hydrogen bonding between the head groups (Figure 10C, 10D).

3.5.5 Possible *Pae* SmAP2···RNA Binding Sites

Having determined that *Pae* SmAP2 binds both U-rich and A-rich RNA, we examined the crystal structures to determine where and how. Previous Sm structures have demonstrated multiple binding motifs including the eukaryal mode: U-rich binding on the L3 face which threads through the pore; and the multiple bacterial Hfq modes: *i)* U-rich binding on the L3 face in the same pocket as eukaryal binding but not threading through the pore, *ii)* A-rich binding on the L4 face, and *iii)* UA-rich binding on a lateral surface. Within the archaea domain, multiple SmAPs have demonstrated the conserved U-rich binding pocket, and *P. abyssi* SmAP has also bound U-rich

RNA on its lateral rim [5,15,18]. *Pae* SmAP2 does not show any distinct electrostatic pockets for RNA binding, but instead has a large positive pore area that could be ideal for non-specific RNA binding (Figure 11C). The conserved aromatic residue (*S. aureus* Hfq: Tyr-42, *P. aerophilum* SmAP1: His-44, *H. sapiens* SmG: Phe-34) on Loop L3 necessary for π-stacking with the uridine base in U-rich binding in all domains, is substituted in *Pae* SmAP2 with a cysteine (Cys-45), necessitating a different binding motif (Figure 11A). The putative *Pae* SmAP2 binding site resembles *H. sapiens* SmD1 U-rich site which contains a serine (Ser-35) (Figure 11B) instead of an aromatic. In the U4 snRNP core, Ser-35 does not demonstrate U-rich specificity, instead hydrogen bonding with the ribose of the U4 snRNA G-130.

The common Hfq A-rich binding sites are also absent on *Pae* SmAP2. An arginine (Arg-28) replacing the conserved aromatic (*E. coli* Tyr-45, *B. subtilis* Phe-24) from both the Gram-negative ARN-motif (R-binding site) and the Gram-positive AG-motif (A-binding site) on *Pae* SmAP2 [50,51]. There is an adjacent tyrosine (Tyr-78), which could play a role in RNA binding, but the nucleotide would need to be in a different orientation (Figure 12A, 12B, 12C, 12D). Loop L4 is extended compared to bacterial Hfq, occluding the RNA backbone path and conserved A-rich Hfq binding site (Figure 7A).

The lateral binding site observed in Hfq and *P. abyssi* SmAP consists of a conserved aromatic residue near the α-helix (*E. coli* and *Pseudomonas aeruginosa* Hfq: Phe-39, and *P. abyssi* Tyr-34) [15,52]. Mutagenesis has identified a positive patch farther down the rim that is involved in lateral rim-binding, most likely by stabilizing the negatively charged backbone [52]. *P. abyssi* SmAP bound with U_6 RNA and *P. aeruginosa* Hfq bound with UTP in the lateral binding sites allow a closer examination of a possible lateral binding site on *Pae* SmAP2. Aligning all three structure reveals that *Pae* SmAP2 contains the necessary aromatic residue in the conserved location (Tyr-42), but is lacking the positive arginine patch (Figure 13A, 13B). On *P. abyssi* SmAP, U_6

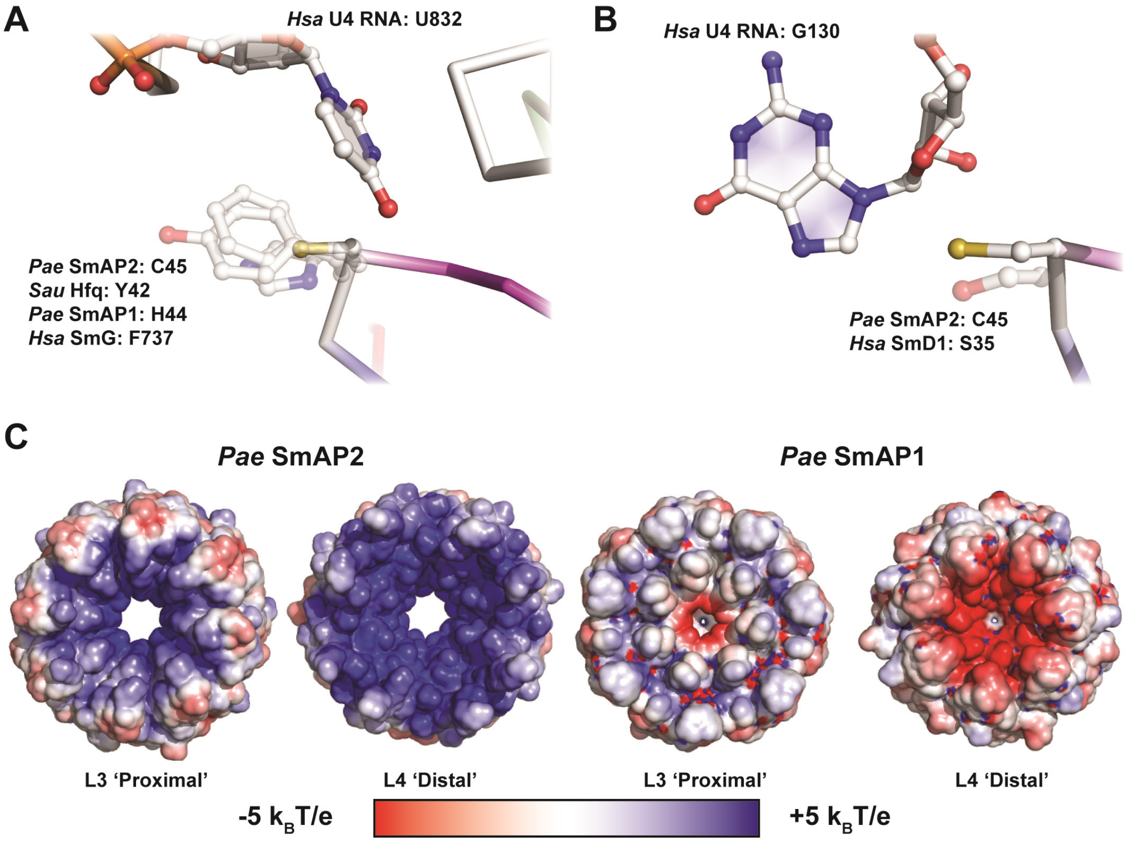


Figure 11. Conserved Sm U-rich binding pocket on *Pae* SmaP2 and electrostatics. (A) *Pae* SmAP2 lacks a highly conserved aromatic residue (bacteria – *Sau* Hfq: Tyr-42; eukaryal – *Hsa* SmG: Met-38; and archaeal – *Pae* SmAP1: His-44), that is responsible for Sm binding to U-rich RNA, necessitating a different binding profile. (B) *Hsa* SmD1 protein (Ser-35) encodes a similar residue in that position. The *Hsa* SmD1 monomer is where the U4 snRNA turns to enter the Sm pore, and Ser-35 hydrogen bonds to the closest nucleotide, G-130. (C) The highly positively

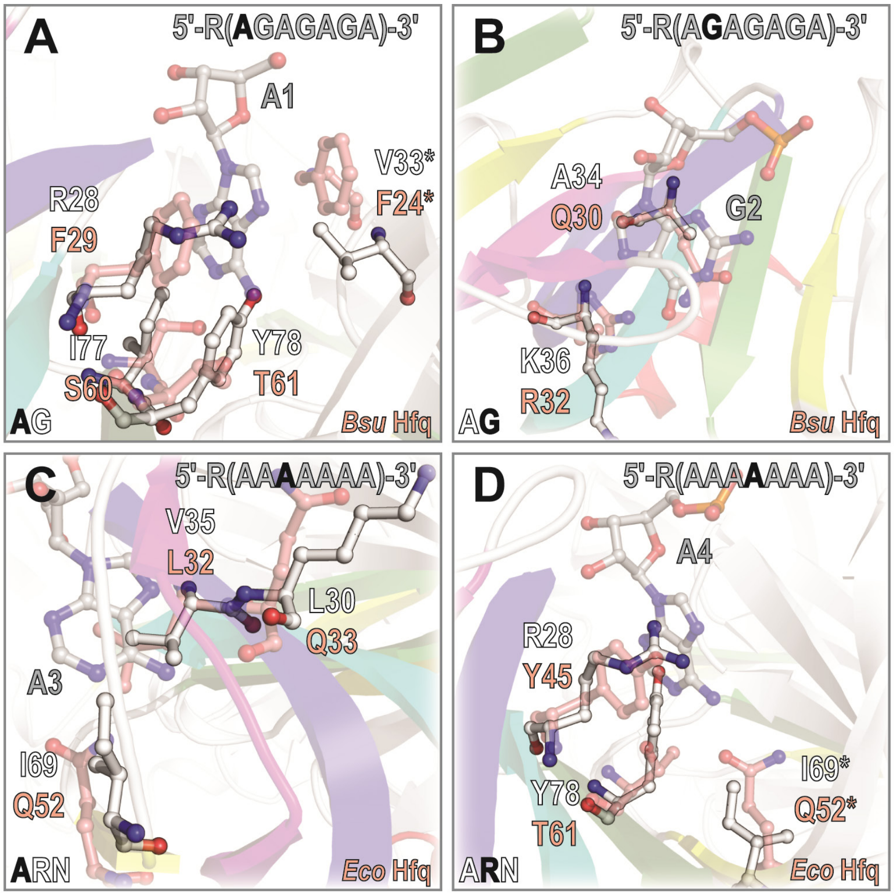


Figure 12. Overlay of *Pae* SmAP2 with A-rich binding sites of bacterial Hfq. (A)(B) Overlays of *Pae* SmAP2 and Gram-positive *Bacillus subtillis* (*Bsu*) Hfq bound with r(AG)$_3$A. Gram-positive bacterial Hfq bind A-rich rna with an AG-repeat. (A) A-binding site of *Bsu* Hfq has two conserved aromatics (F-29 and F-24 of the adjacent monomer) which are substituted for R-28 and V-33 in *Pae* SmAP2. *Pae* SmAP2 Y-78 is adjacent to the site, but the orientation would prevent base stacking. (B) Q-30 hydrophobically interacts with the A-34 in the G-binding site, which in *Pae* SmAP2 is replaced with A-34, which would not extend to the nucleotide. (C)(D) Overlays of *Pae* SmAP2 and Gram-negative *Escherichia coli* (*Eco*) Hfq bound to r(A)$_7$. Gram-negative bacteria bind (C) *Pae* SmAP2 does not contain the charged Q-52 residue in *Eco* Hfq responsible for adenine specificity. (D) The R-site of *Eco* Hfq is the same as the A-site for *Bsu* Hfq, except *Eco* Hfq only has one aromatic residue (Y-45).

RNA continues from Tyr-34 towards the pore, with the next nucleotide stacking with His-10, which is substituted in *Pae* SmAP2 for Thr-18 (Figure 13C). charged surface of *Pae* SmAP2 in contrast to other Sm proteins hints at non-specific binding, which runs contrary to experimental results.

4 Discussion

Early studies on *Pyrobaculum aerophilum* SmAPs have shown that it is promising organism for SmAP functional studies because of the features differentiating the *Pae* SmAP paralogs: *i) Pae* SmAP1 is electrostatically negative, *Pae* SmAP2 is extremely positive, and *ii) Pae* SmAP3 contains an augmented C-terminal domain that forms a tetradecamer. Variation between *Pae* SmAPs suggest they have distinct and either separate functions. The previously determined structures of *Pae* SmaP1 and *Pae* SmAP3 elucidated many areas of Sm function and oligomerization, so here we examine the structure and function of the remaining *Pae* SmAP2.

Overall we found that *Pae* SmAP2 forms higher ordered oligomer than expected in solution, possibly forming multiple stable high-order oligomers. Electron microscopy shows that these oligomers are rings in solution, but we were unable to determine the exact oligomeric state, instead detecting a range between 7 and 12 subunits. The crystal structure revealed that *Pae* SmAP2 crystallizes as an octamer, a heretofore unseen oligomer for SmAPs, opening up the possibility of higher ordered oligomers for Sm proteins. *Pae* SmAP2 was found to bind both U-rich and A-rich RNA non-competitively, but not bind G-rich or C-rich RNA. Binding of RNA did not shift the elution volume of *Pae* SmAP2, so oligomerization was not dependent on RNA binding.

Our initial phylogenetic work suggested that *Pae* SmAP2 would be similar to other SmAPs, but when we were unable to determine the structure of high-resolution crystals by molecular replacement it became apparent *Pae* SmAP2 was unique. Hydrodynamic methods including AnSEC, SEC-MALS, and AUC revealed that *in vitro Pae* SmAP2 is a higher ordered oligomer than expected, between 8 and 13 subunits (Figure 3A, 3B, 3C). Two possibilities existed, *i) Pae* SmAP2 oligomerizes as a small pentameric or hexameric ring, two of which interact similar to *Pae* SmAP3, forming a pair of rings, or *ii) Pae* SmAP2 is a heretofore unseen SmAP oligomer greater than a heptamer. Solving the crystal structure revealed *Pae* SmAP2 crystallizes as an octamer. Even though the *Pae* SmAP2 oligomerizes as an octamer in the crystal, there are unanswered questions about the *in vitro* oligomeric state of *Pae* SmAP2. During analytical ultracentrifugation multiple stable species (not transitioning between states) were identified, consisting of 9 ± 1 and 12 ± 1 *Pae* SmAP2 monomers. *Pae* SmAP2 traveling at 9 ± 1 could be the result of the hydrodynamic radius of the octamer, but that does not explain the higher ordered 12 ± 1. Cross-linking followed by MALDI TOF MS supports the multiple stable oligomers in solution, with spectrum containing multiple baseline resolved peaks between 7 and 12 monomers (Figure 3D). Electron microscopy revealed that *Pae* SmAP2 exists as a stable ring in solution, though whether the majority or just a percentage is unknown (Figure 3E). Another possibility is *Pae* SmAP2 is similar to the eukaryal canonical Sm core proteins, which have multiple oligomeric states, beginning as either dimers or trimers before the canonical Sm core proteins come into contact with their target U snRNA, when they finally oligomerize into a heptamer. To determine if *Pae* SmAP2 only forms its stable single oligomer in the presence of a binding partners or could move between oligomers depending on the current binding partner, we first needed to determine possible binding partners.

To determine possible *Pae* SmAP2 binding partners we started with known binding partners of the Sm family, single stranded RNA, mainly U-rich RNA. Starting with U_6 RNA we found *Pae* SmAP2 binds both U-rich and A-rich RNA, though the A-rich binding is length dependent, with A_{18} exhibiting significant binding, and no binding seen for shorter A-rich strands. *Pae* SmAP2 binds U-rich RNA at 10-fold the affinity as A-rich RNA. Because of the strong positive electrostatics, particularly in the pore region of *Pae* SmAP2, it was possible that *Pae* SmAP2 could bind the negatively charged backbone or RNA non-specifically, explaining the A-rich binding and lesser affinity, but binding assays with C-rich and G-rich RNA saw no significant binding. In addition, competition assays demonstrated that the U-rich and A-rich binding was non-competitive, meaning they bind at different sites. The *Pae* SmAP2 binding profile (A- and U-rich RNA) is redolent of the bacterial Hfq binding profile rather than the eukaryal Sm proteins (U-rich only). To test our theory that *Pae* SmAP2 may only form its complete oligomer when bound to an RNA target, we incubating *Pae* SmAP2 with A-rich RNA, U-rich RNA, or both before AnSEC. While we were able to see binding by an increase in the absorbance at 260 nm, the shift in the peak elution only corresponded to addition of an RNA, with no change in *Pae* SmAP2 oligomer (Figure 3A). The lack of shift in the elution of the *Pae* SmAP2 peak suggests that *i) Pae* SmAP2 does not shift oligomeric state when bound with an RNA target and *ii)* the bound RNA sits flush against *Pae* SmAP2 and does not significantly change the hydrodynamic radius.

While *Pae* SmAP2 binds both A-rich and U-rich RNA, the crystal structure shows that *Pae* SmAP2 does not have the conserved residues in the binding pockets observed in other Sm proteins. RNA binding sites usually contain aromatic residues, and their positions can be clues to locating unknown binding pockets. *Pae* SmAP2 only encodes 3 aromatic residues, all tyrosines, with two, Tyr-42 and Tyr-66, exposed (Figure 7G) on *Pae* SmAP2. Tyr-78 is located in the cleft between Loops L2 and L4, and while not exposed in the crystal structure, could be exposed in solution depending on the mobility of the L2 and L4 Loops. Tyr-42 will be discussed later in relation to the lateral site. Tyr-66 is located on the lateral rim opposite Tyr-42 (near the L4 face) at the beginning of the Loop L4. To date, there has been no confirmed binding at this location on any Sm proteins. The conserved U-rich binding pocket shared by all Sm proteins, around the pore of the L3 face usually consists of an aromatic residue which π-stacks with the uridine base, while hydrogen bonds to the uracil supply specificity (Figure 11A). *Pae* SmAP2 lacks this aromatic residue, instead *Pae* SmAP2 substitutes a cysteine, reminiscent of the serine at this same location in SmD1. The SmD1 binding site is the transition point before the U snRNA threads through the pore, and appears to bind weakly and non-specifically. *Pae* SmAP2 demonstrates specificity, which could be supplied by hydrogen bonding we have yet to identify, and the tight binding being a result of RNA backbone interactions with the highly positive electrostatic surfaces of the *Pae* SmAP2 pore, rather than π-stacking with the nucleotide bases.

Since *Pae* SmAP2 binds A-rich RNA, we examined the common A-rich binding sites from bacterial Hfq. Both Gram-negative and Gram-positive bacteria bind A-rich RNA, though they have different binding motifs, with Gram negative bacteria having an ARN-repeat motif and Gram positive bacteria having a AG-repeat motif. Both A-rich motifs share a conserved location, a binding pocket at the interface between adjacent monomers on the L4 face, which is the R-site in the ARN motif, and the A-site in the AG motif. *Pae* SmAP2 is unlikely to bind RNA in this cleft, because *Pae* SmAP2 lacks the conserved aromatic in Hfq, instead replacing it with an arginine (Arg-28) and a valine (Val-33) on the opposite face of the cleft. In addition, the site is shielded by the extended Loop L4 of *Pae* SmAP2. The A-binding site in *Pae* SmAP2 cannot be completely ruled out as Loop L4 is the most mobile loop of *Pae* SmAP2 (based on B-factor and variations between monomers), and could shift, exposing the site and a nearby tyrosine (Tyr-78) which flanks Arg-28.

An alternative to the common U-rich and A-rich sites seen in most Sm and Hfq proteins, is the newly discovered lateral rim site. The lateral rim site is known to bind UA-rich RNA in Hfq, and has been observed in one SmAP and multiple Hfqs [15,21,52]. Examining the area identified as the lateral site, *Pae* SmAP2 does contain the conserved aromatic residue, Tyr-42 (Figure 13A, 13B). In Hfq, the lateral rim site π-stacks with the conserved aromatic before the backbone contacts an arginine-rich positive patch farther down the rim toward the L4 face. In *Pae* SmAP2 there is no

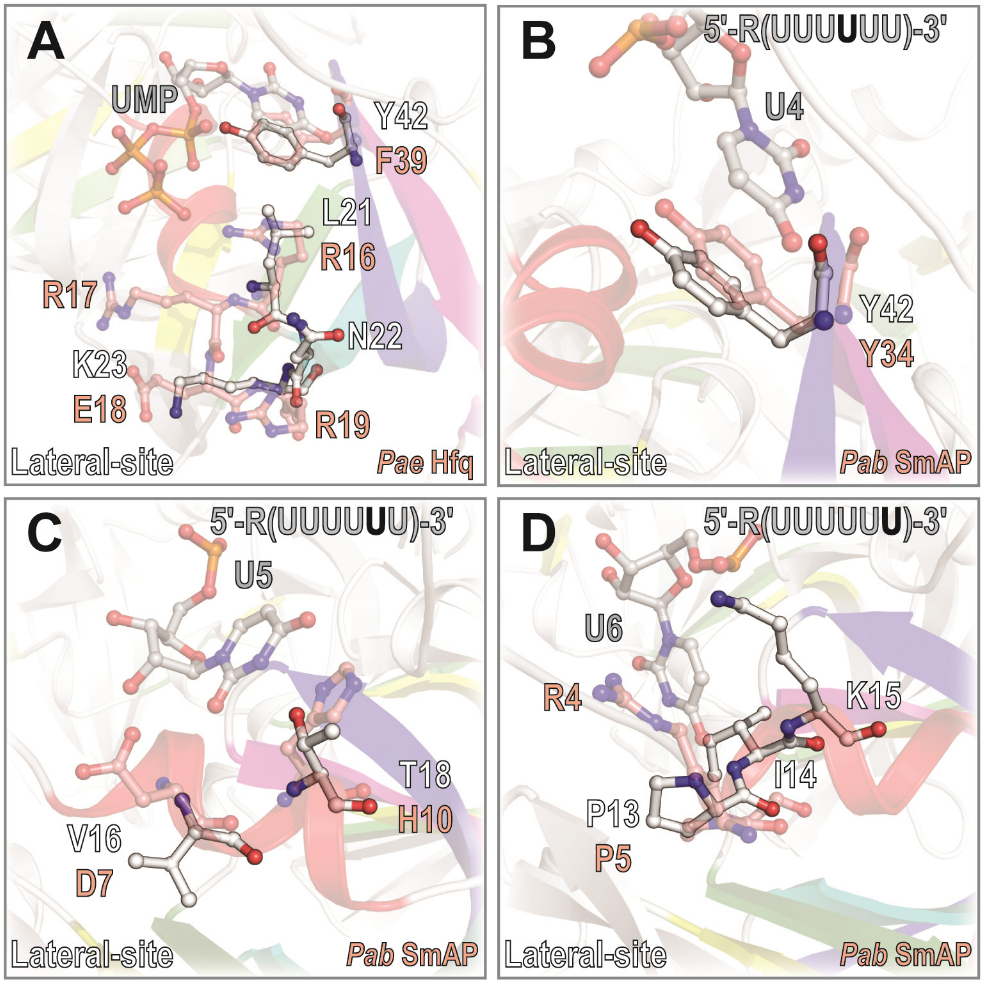


Figure 13. Lateral rim binding site. A lateral binding site has been observed in multiple Hfqs and the archaeal *Pyrococcus abyssi* (*Pab*). (A) Overlay of *Pae* SmAP2 and *Pseudomonas aeruginosa* (*Pae*) Hfq bound to UMP at the lateral site. The conserved aromatic (*Pae* Hfq Phe-39 and *Pae* SmAP2 Tyr-42) overlay, suggesting lateral rim binding in *Pae* SmAP2, though it lacks the positive arginine patch farther down the rim, towards the L4 face. (B)(C)(D) Overlays of *Pae* SmAP2 and *Pab* SmAP lateral bound to r(U)$_6$. (B) π-stacking tyrosine (*Pab* SmAP Tyr-34, *Pae* SmAP2 Tyr-42) is conserved. (C) *Pab* SmAP His-10 stacks with the next uridine nucleotide, but is absent in *Pae* SmAP2. (D) The third uridine hydrogen bonds with a proline that is conserved between *Pab* SmAP (Pro-5) and *Pae* SmAP2 (Pro-13).

positive patch on the outer rim, in fact, the outer rim of *Pae* SmAP2 is one of the few places on the protein that is not highly electrostatically positive. The lack of a positive rim does not rule out RNA stacking on Tyr-42, as the RNA backbone can extend toward the positive pore in *Pae* SmAP2 as seen in the *P. abyssi* SmAP. In the *P. abyssi* SmAP···U_6 structure, the RNA runs both 3' to 5' and 5' to 3' across the lateral binding site. In addition to the uridine stacked with the conserved tyrosine, the neighbor ($n + 1$) nucleotide π-stacks with a His-10 which is absent in *Pae* SmAP2, but a proline which hydrogen bonds with the ($n + 2$) nucleotide is conserved between *P. abyssi* SmAP and *Pae* SmAP2. It is possible that both the A-rich and U-rich RNA binding we observed could be a result of this site, with the RNA strand continuing into the positive pore or running across the face. The sites are adjacent enough that competition might not be observed, but it is more likely that one of A-rich and U-rich RNA bind on an additional, unidentified site. Of interest is the Tyr-66 location on the lateral rim of the L4 face, which could be a new binding site, or a continuation of the lateral-rim binding.

5 Concluding Remarks

The Sm superfamily's central role in RNA processing and regulation, combined with their existence in all three domains, makes them a model system for exploring RNA processing evolution. Sm-mediated interaction between RNAs play vital roles in important pathways such as virulence, quorum sensing, cell death and aging, and mRNA splicing. The largest gap in our knowledge of the Sm superfamily is in the archaeal branch. Many archaeal systems can provide invaluable knowledge about their more complicated analogous eukaryal systems by supplying a simpler model to work with. Initial work on the Sm-like archaeal proteins (SmAPs) were crucial to our understanding of how Sm proteins oligomerize and bind RNA. Unfortunately, since this early work, the study of SmAPs has been limited, and SmAP *in vivo* functions are virtually unknown. Understanding these *in vivo* functions of SmAPs would allow us a better understanding of basic

Sm protein function, provide a window into the evolution of the large eukaryal ribonucleoprotein complexes, and possibly link the evolution of bacterial Hfq and eukaryal Sm proteins. The crenarchaea *Pyrobaculum aerophilum* is a deep-branching, hyperthermophile that encodes multiple SmAP paralogs. The two known *Pae* SmAP structures (SmAP1, SmAP3) illuminated Sm protein evolution and assembly, and implied that these homologs may represent an ancestral form of the complexes that developed into the extant heteromeric Sm assemblies of eukaryotes, such as those at the heart of the spliceosome. Our work on the final *Pae* SmAP, SmAP2, reveals that *Pae* SmAP2 oligomerizes as a unique octamer (unseen in previous SmAPs), and binds both A-rich and U-rich RNA reminiscent of the bacterial Hfq (*chaperone*). The crystal structure revealed that *Pae* SmAP2 lacks the conserved residues seen in the common U-rich and A-rich binding pockets of other Sm proteins, but does contain the aromatic (Tyr-42) necessary for lateral-rim binding. Further research is necessary to determine the specific binding mechanisms of *Pae* SmAP2···RNA binding, the *Pae* SmAP2 solution state, and determine the individual functions of the SmAP paralogs in *Pyrobaculum aerophilum*.

6 Methods

6.1 Phylogenetic Analysis

A phylogenetic tree containing representative members of the eukaryal, bacterial, and archaeal Sm proteins was initially created to determine the clustering of the various Sm proteins. This included SmD1 proteins from eukarya, Hfq's from bacteria, and various SmAPs chosen randomly from crenarchaea, euryarchaea, and thaumarchaea/nanoarchaea. Alignment and tree generation were done in Geneious [53] using MAFFT [54] for alignment and MrBAYES[55] for tree generation. Proteins used included GID's: 549635379, 499164542, 50504551, 504370326, 500145274, 503791707, 503503781, 500271789, 62464807, 499180836, 505404260, 503671430, 499490831, 499219074, 269986972, 500681920, 339756221, 494813972, 407463901, 161527598, 647810924, 495573562, 501137396, 499316925, 500176350, 501319673, 503893548, 503445978, 501267038, 490146347, 494978361, 490327461, 564598001, 499637770, 499179510, 317135015, 332641050, 17864386, 48734707, 19354162, 325115253, 584410573, 641580737, 1323102, 499182470, 499163297, 660682089, 662228596, 218927576, 658112580, 663091812, 640835739, 640835739, 160286013.

6.2 Expression and Purification

The vector pET-32a(+) was cloned to contain the recombinant *Pae* SmAP2 gene locus *pae676* and an N-terminal thioredoxin tag followed by a polyhistidine tag, an S-tag, and an enterokinase cleavage site (SmAP2cm) (Fig 1B). From this vector, the *pae676* gene was sub-cloned using the PIPE method into a pET22b vector with no tags (SmAP2pr1). Constructs were transformed into *Escherichia coli* BL21(DE3) cells and grown in Luria-Bertani (LB) medium, containing 100 μg ml^{-1} ampicillin at 37° C until an OD_{600} of 0.8 was reached. Protein expression was induced by 1 mL of 1.0 mM isopropyl β-D-1-thiogalcatopyranoside (IPTG) and cells were cultured for an additional 3 hrs. Cells were harvested by centrifugation for 5 min at 15,000 g (Sorvall RC 6+) at 4° C and then frozen (-20° C).

Initial lysis and purification steps were identical for both SmAP2cm and SmAP2pr1 constructs. Cells were resuspended in Lysis Buffer (20 mM HEPES pH 7.8, 1.5 M NaCl, 1 mM $MgCl_2$, 0.1% Triton (v/v), 0.4 mM PMSF). 300 µg of lysozyme (Fisher) per liter of culture were added and incubated at 37° C for 30 minutes. Cells were mechanically lysed by passing through a microfluidizer (Microfluidics) three times at 60 psi. Lysate was centrifuged at 35,000 g for 20 min at 4° C then supernatant (S1) and pellet (P1) were separated. S1 was heated at 72° C for 10 min, denaturing the mesophilic *E. coli* proteins which precipitate. The sample was then centrifuged again for 35,000 g for 20 min and supernatant (S2) and pellet (P2) were separated. At this juncture the SmAP2cm and SmAP2pr1 procedures diverged.

SmAP2cm was purified by immobilized metal affinity chromatography (IMAC). S2 was diluted (1:1) with IMAC Wash Buffer (20 mM HEPES pH 7.8, 350 mM NaCl, 10 mM Imidazole pH 7.0) and loaded onto a chelating column (5 ml, HiTrap Chelating Column, GE Healthcare) bound with $CoCl_2$, then eluted with a gradient from 0 to 80% IMAC Elution Buffer (20 mM HEPES pH 7.8, 350 mM NaCl, 520 mM Imidazole pH 7.0). Fractions containing SmAP2cm construct were pooled and dialyzed into Dialysis Buffer 1 (20 mM HEPES pH 7.8, 250 mM NaCl, 15 mM EDTA), followed by a series of dialysis stages (Dialysis Buffer 2: 20 mM HEPES pH 7.8, 150 mM NaCl, 15 mM EDTA; Dialysis Buffer 3: 20 mM HEPES pH 7.8, 100 mM NaCl, 15 mM EDTA; Dialysis Buffer 4: 20 mM HEPES 7.8, 20 mM NaCl, 1 mM CaCl2, 0.1% (w/v) Tween; Enterokinase Cleavage Buffer: 50 mM Tris pH 7.8, 15 mM NaCl, 1 mM CaCl2, 0.1% (w/v) Tween) to lower the salt concentration to an acceptable level for enterokinase digestion while preventing precipitation. Enterokinase (New England Biolabs) was added at a ratio of $1:1x10^6$ (enterokinase:SmAP2cm construct) and incubated at 37°C for 16 hours. The calculated isoelectric point of 9.7 for SmAP2 allows separation using cation exchange from the other digestion products. The solution was dialyzed into IEC Wash Buffer 1 (20 mM HEPES pH 7.8) then loaded onto a cation exchange column (5ml, SP column, GE Healthcare) and eluted using a gradient from 0 to 100% IEC Elution Buffer 1 (20 mM HEPES pH 7.8, 2M NaCl).

SmAP2pr1 is tagless, and is purified by cation exchange chromatography done under denaturing condition. Supernatant (S2) was dialyzed against IEC Wash Buffer 2 (20 mM HEPES pH 6.8, 200 mM NaCl, 8M Urea) for four hours, then loaded onto a cation exchange column (5ml, SP column, GE Healthcare) and eluted using a gradient from 0 to 100% IEC Elution Buffer 2 (20 mM HEPES pH 7.8, 2M NaCl, 8M urea). SmAP2pr1 is then pooled and dialyzed successively against 20 mM HEPES pH 6.8, 1M NaCl then 20 mM HEPES pH 6.8, 1M NaCl for 4 hours each to remove the urea. The sample was then run through gel filtration to ensure pure properly folded *Pae* SmAP2.

Purification of both constructs was confirmed using MALDI-TOF (described below) in conjunction with SDS-PAGE.

6.3 Selenomethionine (SeMet) Incorporation

M9 minimal media (M9 MM) was prepared by combining 220 mL of 5X M9 salts (239 mM Na_2HPO_4, 110 mM KH_2PO_4, 42.5 mM NaCl, 93.5 mM NH_4Cl) and 1.1 ml of $MgSO_4$ and 855.8 ml ddH_2O. The solution was autoclaved and cooled, then 22 ml of sterile 20% (v/v) glycerol, 1.1 ml of thiamine solution (1 mg/ml) and 1.1 ml of ampicillin (100 mg/ml) were added. *Pae* SmAP2pr1 construct was transformed into *E. coli* BL21 (DE3) and grown in LB media as described above. Cells were grown to an OD600 of 0.9 at which point the cultures were centrifuged at 4000 g for 30 min at 277 K. Cell pellets were resuspended in 100 ml M9 MM and re-pelleted at 4000 g for 30 min at 277 K, the supernatant discarded and the pellet resuspended in 50 ml of M9 MM which was then added to the full liter of M9 MM. Free amino acids were then added (50 mg/L of leucine, isoleucine, and valine; 100 mg/L of phenylalanine, lysine, and threonine; 75 mg/L of selenomethionine (Acros Organics)) and the culture was incubated at 37°C for 30-40 minutes to suspend methionine production. A pre-induction sample was taken prior to induction with IPTG. *Pae* SmAP2 production was induced with IPTG and the culture was returned to the incubator shaker at 37°C. After 6 hours the media was centrifuged at 15,000g for 5 min at 4°C, the supernatant discarded and the pellet frozen overnight. Lysis and purification steps are the same as previously mentioned. MALDI-TOF MS confirmed SeMet incorporation.

6.4 Protein Crystallization

Pae SmAP2 was dialyzed into 25 mM Tris at pH 7.8, 250 mM NaCl and concentrated to 50 mg/ml measured by absorbance at 280 nm (Nanodrop 1000, Thermo Scientific), with an expected molecular weight of 9243 Da. Commercial screens (Qiagen JCSG Core Suite 1-3) were set up using a nano-liter volume robot (Mosquito, TTC Inc) in 96-well plates with 100 μl of mother liquor, and hanging drops with a volume of 0.2 μl of 1:1 (protein solution:mother liquor). PEG/Ion and Ammonium Sulfate (A/S) Grid Screens (Hampton Research) was set-up by hand in a 24 well VDX plate with 600 μl mother liquor and a 4 μl hanging drop containing 1:1 (protein solution:mother liquor). Crystallization trials were incubated at 277 and 291 K. Conditions in which crystals developed were expanded upon until by systematically varying precipitant parameters. In addition, 96-well additive screens (Hampton Research) were applied. Reproducible well diffracting crystals were obtained in conditions containing 1.2 to 1.6 M ammonium sulfate either with or without 100 mM citrate at pH 4.0. Additives including cobalt chloride, nickel chloride and guanidine-HCl assisted in crystal formation time and diffraction. SeMet crystals were developed by expanding on native conditions and followed a similar pattern but at a lower ammonium sulfate range (0.8 to 1.2 M) than native. Various cryogenic conditions were explored and an initial solution of 50% saturated ammonium sulfate and 50% glycerol was used but later was replaced with a solution of 65% mother liquor and 35% glycerol. Crystals were screened on a Rigaku MicroMax 007 generator with a Saturn 92 detector, and a Bruker APEX2.

6.5 Data Collection

Single crystals were harvested in nylon cryoloops (Hampton Research) and washed in a 10 µl drop of cryo-protectant for 10-45 sec. The crystals were promptly flash frozen in liquid nitrogen. Diffraction datasets were collected at the ALS and APS synchrotron on either beamline 5.0.2 using a ADSC Quantum 315r detector; SER-CAT 22ID or 22BM beamline, using a MAR 300 and MAR225 CCD detector; or NEC-CAT 22BM and 24C beam lines using a Pilatus detector. Data sets were collected at a variety of oscillation ranges (0.25° to 1.0°) and total crystal rotations (90° to 720°). Multi-anomalous dispersion (MAD) datasets were collected at four energies: *i)* peak, *ii)* inflection, and *iii)* high remote.

6.6 Processing, Structure Determination, And Refinement

Datasets were integrated and scaled in XDS [43]. Crystal quality was examined in phenix.Xtriage, determining possible twinning, as well as verifying indexing, gauging diffraction anisotropy, detecting pseudo-translational symmetry, and calculating the Matthews coefficient [45]. Phaser and MolRep were used for molecular replacement trials [45,56]. Hexameric (1U1S), heptameric (1I8F), and octameric (3BW1) Sm oligomers from bacteria, archaea, and eukarya respectively were used as molecular replacement probes. Two forms of each probe were screened: cleaned of ions and waters (PyMol), and as an alanine chain made using Sculptor [57]. *De novo* multi-anomalous dispersion (MAD) phasing was done using the SHELX suite in HKL2MAP [47,48]. The structures were rebuilt with density modification in Autobuild and further refinement rounds of used a combination of either phenix.refine or Refmac with direct structure manipulation in COOT [58]. Structures were validated using Molprobity [59]. The majority of structural analysis was done in PyMol. Electrostatics were calculated using the ABPS plugin [60]. ProFit was used to align the structures.

6.7 Fluorescence Anisotropy and Fluorescence Polarization

Pae SmAP2 was dialyzed into 25 mM HEPES pH 7.8 and 250 mM NaCl prior to binding studies. Fluorescence anisotropy/polarization measurements were collected with a PHERAstar microplate reader. 5-Carboxyfluorescein-labeled $r(U)_6$ (FAM-$r(U)_6$), FAM-$r(C)_6$, and FAM-$r(A)_{18}$ RNA (Integrated DNA Technologies) were used to probe the binding sites of *Pae* SmAP2. FAM-$r(A)_{18}$ was annealed prior to binding assays by incubating at 85°C for 3 minutes and then placed on ice for 10 minutes [61]. Samples were excited at 490 nm and emission was measured at 522 nm [62]. To determine the apparent equilibrium dissociation constant (K_d) for FAM-RNA, *Pae* SmAP2 was serially diluted in 96-well black polystyrene assay plates (Costar) in the presence of 5 nM FAM-RNA; the final volume in each well was 150 µL. The binding assay samples were incubated in the dark for 45 minutes at RT prior to measurements to ensure equilibrium binding. Data was collected in three independent replicates.

Fluorescence data was fit to a model that assumed that a 1:1 complex formed between *Pae* SmAP2 and FAM-RNA. [61,63]. The model involves fitting the data to a sigmoidal Boltzmann function, which is related to the Hill equation [64,65], and can be rearranged to read

$$y = \frac{(A_1 - A_2)}{1 + e^{\frac{(x - x_0)}{dx}}} + A_2 \qquad \text{(Equation 1)}$$

where x_0 is the inflection point sigmoidal curve, dx is the width of the transition and A_1 and A_2 are the fluorescence polarization intensities of the initial and final states, respectively [63,66,67]. Nonlinear least-squares fits of the equation to the data were performed in OriginPro7.5.

6.8 Chemical Crosslinking

Pae SmAP2 was dialyzed into 25 mM HEPES pH 7.8 and 250 mM NaCl prior to crosslinking with formaldehyde or glutaraldehyde using an 'indirect' method [68]. Experimental setup for this indirect method consisted of a Linbro plate with a microbridge and coverslip upon which 40 μL of the crosslinking agent (25% v/v) and 15 μL of *Pae* SmAP2 (1 mg/mL) were aliquoted, respectively. The chamber was sealed with vacuum grease. The crosslinking agent was acidified with 122 mM HCl before incubation. Samples were incubated at 37 °C for 40 minutes and the reaction was stopped by the addition of 5 μL 1 M Tris pH 7.8. Salts and crosslinking reagents were removed from samples using a C_4 zip tip (Millipore)[69], and the molecular weight of the crosslinked oligomers was assessed by MALDI-TOF MS (protocol described below). Data was collected in three independent replicates for both formaldehyde and glutaraldehyde. RNA containing samples were prepared as previously mentioned (3.7), and incubated with *Pae* SmAP2 for 45 minutes before run.

6.9 MALDI-TOF Mass Spectrometry

MALDI-MS was performed on a Bruker Microflex MALDI. Proteins of approximately 1 mg/mL were diluted 1:4 with 0.01% trifluoroacetic acid (TFA). The protein sample was spotted onto a MSP 96 target ground steel sample plate (Bruker Daltonics) with an equal volume of SA (15 mg/mL sinapinic acid in 50% acetonitrile and 0.05% TFA) and allowed to air dry. The instrument was calibrated by the close external method using a series of low molecular weight (Insulin, Cytochrome C, Ubiquitin I, and Myoglobin) or high molecular weight (Protein A, Trypsinogen, Protein A, and Bovine Albumin) protein calibrants. Spectra were obtained by averaging approximately 50 laser shots with the following settings: positive ion, linear mode, grid voltage 40-75%, m/z range 4,000-20,000 or 20,000-200,000.

6.10 Analytical Size Exclusion (AnSEC)

Pae SmAP2 was dialyzed into 25 mM HEPES pH 8.0 and 250 mM NaCl. The chromatography system consisted of a Superdex 200 10/300 Increase column (spherical composite of crosslinked agarose and dextran matrix) and a Biologic NGC or a Waters Breeze 1525 Binary HPLC with a Waters 2489 UV/Visible detector at room temperature. *Pae* SmAP2 and molecular weight standards were injected and eluted with approximately seven column volumes of 25 mM HEPES pH 8.0 containing 250 mM NaCl at 0.2 mL/min; absorbance was monitored at 260 and 280 nm throughout each run. A standard curve was generated using Gel Filtration Markers Kit for Protein Molecular Weights 29,000 – 700,000 Da (MWGF1000-1KT, Sigma-Aldrich). RNA containing samples were prepared as previously mentioned (3.7), and incubated with *Pae* SmAP2 for 45 minutes before run.

6.11 Size Exclusion Chromatography – Multi-Angle Light Scattering (SEC-MALS)

Pae SmAP2 was dialyzed into 25 mM HEPES pH 8.0 and 250 mM NaCl. SEC-MALS system consisted of a Waters Breeze 1525 Binary HPLC with a Waters 2489 UV/Visible detector, in-line with a Wyatt miniDAWN TRIOS (light scatterer) and tREX (refractometer). Analytical separation was accomplished with a Superdex 200 10/300 Increase column (spherical composite of crosslinked agarose and dextran matrix). Bovine serum albumin (BSA) was used as an isotropic standard and to calibrate the instrument. *Pae* SmAP2 and BSA were injected and eluted with approximately seven column volumes of 25 mM HEPES pH 8.0 containing 250 mM NaCl at 0.2 mL/min; absorbance was monitored on the Waters 2489 UV/Visible detector at 260 and 280 nm throughout each run. Wyatt ASTRA software was used for analysis. RNA containing samples were prepared as previously mentioned (3.7), and incubated with *Pae* SmAP2 for 45 minutes before run.

6.12 Electron Microscopy

Pae SmAP2 samples were analyzed by application to glow-discharged, carbon-coated Cu grids, negative staining using uranyl acetate (2% wt/v), and imaging in a FEI CM12 TEM at an accelerating voltage of 120 keV, equipped with a TVIPS 0124 CCD camera (1k × 1k pixel). Grids were imaged using a Tecnai F20 microscope (FEI) at an accelerating voltage of 120 keV and a nominal magnification of 30,000x. Micrographs were scanned with a Nikon Coolscan 8000 at a raster of 1.25 Å per pixel.

6.13 Analytical ultracentrifugation

Samples were prepared in house and shipped to Peter Schuck and Joy Zhou at NIH, Bethesda, MD, for ultracentrifugation. Samples were run in both sedimentation and velocity ultracentrifugation and processed in SEDPHAT. SV-AUC experiments were conducted using a Beckman ProteomeLab XL-I (Beckman-Coulter) following the standard procedures and boundary structure analysis. Absorbance profiles at 280 nm acquired at a rotor speed of 50,000 rpm were analyzed with a c(s) sedimentation coefficient distribution to determine the overall weighted-average s-value and the s-value of the reaction boundary, respectively, as a function of the composition of the loading mixture.

7 REFERENCES

1. Scofield DG, Lynch M. Evolutionary diversification of the Sm family of RNA-associated proteins. *Mol Biol Evol*. 25, 2255–2267 (2008).

2. Tan EM, Kunkel HG. Characteristics of a soluble nuclear antigen precipitating with sera of patients with systemic lupus erythematosus. *J Immunol*. 96(3), 464–471 (1966).

3. Lerner MR, Steitz JA. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci USA*. 76(11), 5495–5499 (1979).

4. Salgado-Garrido J, Bragado-Nilsson E, Kandels-Lewis S, Séraphin B. Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J*. 18(12), 3451–3462 (1999).

5. Törö I, Thore S, Mayer C, Basquin J, Séraphin B, Suck D. RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex. *EMBO J*. 20(9), 2293–2303 (2001).

6. Collins BM, Harrop SJ, Kornfeld GD, Dawes IW, Curmi PM, Mabbutt BC. Crystal structure of a heptameric Sm-like protein complex from archaea: implications for the structure and evolution of snRNPs. *J Mol Biol*. 309(4), 915–923 (2001).

7. Mura C, Cascio D, Sawaya MR, Eisenberg DS. The crystal structure of a heptameric archaeal Sm protein: Implications for the eukaryotic snRNP core. *Proc Natl Acad Sci USA*. 98(10), 5532–5537 (2001).

8. de Fernandez MTF, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. *Nature*. 219(5154), 588–590 (1968).

9. de Haseth PL, Uhlenbeck OC. Interaction of Escherichia coli host factor protein with oligoriboadenylates. *Biochemistry*. 19, 6138–6146 (1980).

10. Carmichael GG, Weber K, Niveleau A, Wahba AJ. The host factor required for RNA phage Qbeta RNA replication in vitro. Intracellular location, quantitation, and purification by polyadenylate-cellulose chromatography. *Journal of Biological Chemistry*. 250(10), 3607–3612 (1975).

11. Arluison V, Derreumaux P, Allemand F, Folichon M, Hajnsdorf E, Régnier P. Structural Modelling of the Sm-like Protein Hfq from Escherichia coli. *J Mol Biol*. 320(4), 705–712 (2002).

12. Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G. The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol Cell*. 9(1), 11–22 (2002).

13. Hermann H, Fabrizio P, Raker VA, *et al.* snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. *EMBO J*. 14(9), 2076–2088 (1995).

14. Sauter C, Basquin J, Suck D. Sm-like proteins in Eubacteria: the crystal structure of the Hfq protein from Escherichia coli. *Nucleic Acids Res*. 31(14), 4091–4098 (2003).

15. Thore S, Mayer C, Sauter C, Weeks S, Suck D. Crystal structures of the Pyrococcus abyssi Sm core and its complex with RNA. Common features of RNA binding in archaea and eukarya. *Journal of Biological Chemistry*. 278(2), 1239–1247 (2002).

16. Naidoo N, Harrop SJ, Sobti M, *et al.* Crystal structure of Lsm3 octamer from Saccharomyces cerevisiae: implications for Lsm ring organisation and recruitment. *J Mol Biol*. 377(5), 1357–1371 (2008).

17. Das D, Kozbial P, Axelrod HL, *et al.* Crystal structure of a novel Sm-like protein of putative cyanophage origin at 2.60 A resolution. *Proteins*. 75(2), 296–307 (2009).

18. Mura C, Kozhukhovsky A, Gingery M, Phillips M, Eisenberg D. The oligomerization and ligand-binding properties of Sm-like archaeal proteins (SmAPs). *Protein Sci*. 12(4), 832–847 (2003).

19. Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol*. 15(2), 108–121 (2014).

20. Vogel J, Luisi BF. Hfq and its constellation of RNA. *Nat Rev Microbiol*. 9(8), 578–589 (2011).

21. Murina V, Lekontseva N, Nikulin A. Hfq binds ribonucleotides in three different RNA-binding sites. *Acta Crystallogr D Biol Crystallogr*. 69(Pt 8), 1504–1513 (2013).

22. Krummel DAP, Oubridge C, Leung AKW, Li J, Nagai K. Crystal structure of human spliceosomal U1 snRNP at 5.5 A resolution. *Nature*. 458(7237), 475–480 (2009).

23. Leung AKW, Nagai K, Li J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. *Nature*. 473(7348), 536–539 (2011).

24. Kambach C, Walke S, Young R, *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell*. 96(3), 375–387 (1999).

25. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D. Structure and assembly of an augmented Sm-like archaeal protein 14-mer. *Proc Natl Acad Sci USA*. 100(8), 4539–4544 (2003).

26. Thore S. Crystal Structures of the Pyrococcus abyssi Sm Core and Its Complex with RNA. COMMON FEATURES OF RNA BINDING IN ARCHAEA AND EUKARYA. 278, 1239–1247 (2002).

27. Matera AG, Terns RM, Terns MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Rev Mol Cell Biol*. 8, 209–220 (2007).

28. Mura C, Randolph PS, Patterson J, Cozen AE. Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. 10, 636–651 (2013).

29. Valentin-Hansen P, Eriksen M, Udesen C. The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Mol Microbiol*. 51(6), 1525–1533 (2004).

30. Sobrero P, Valverde C. The bacterial protein Hfq: much more than a mere RNA-binding factor. *Crit Rev Microbiol*. 38(4), 276–299 (2012).

31. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. *Proc Natl Acad Sci USA*. 107(21), 9602–9607 (2010).

32. McCullen CA, Benhammou JN, Majdalani N, Gottesman S. Mechanism of positive regulation by DsrA and RprA small noncoding RNAs: pairing increases translation and protects rpoS mRNA from degradation. *J Bacteriol*. 192(21), 5559–5571 (2010).

33. Mikulecky PJ, Kaw MK, Brescia CC, Takach JC, Sledjeski DD, Feig AL. Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nat Struct Mol Biol*. 11(12), 1206–1214 (2004).

34. Panja S, Woodson SA. Hexamer to monomer equilibrium of E. coli Hfq in solution and its impact on RNA annealing. *J Mol Biol*. 417(5), 406–412 (2012).

35. Updegrove TB, Correia JJ, Chen Y, Terry C, Wartell RM. The stoichiometry of the Escherichia coli Hfq protein bound to RNA. *RNA*. 17(3), 489–500 (2011).

36. Afonyushkin T, Vecerek B, Moll I, Bläsi U, Kaberdin VR. Both RNase E and RNase III control the stability of sodB mRNA upon translational inhibition by the small regulatory RNA RyhB. 33, 1678–1689 (2005).

37. Sledjeski DD, Whitman C, Zhang A. Hfq is necessary for regulation by the untranslated RNA DsrA. *J Bacteriol*. 183, 1997–2005 (2001).

38. Bandyra KJ, Said N, Pfeiffer V, Górna MW, Vogel J, Luisi BF. The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. 47, 943–953 (2012).

39. Törö I, Basquin J, Teo-Dreher H, Suck D. Archaeal Sm proteins form heptameric and hexameric complexes: crystal structures of the Sm1 and Sm2 proteins from the hyperthermophile Archaeoglobus fulgidus. *J Mol Biol*. 320(1), 129–142 (2002).

40. Frank DN, Pace NR. RIBONUCLEASE P: Unity and Diversity in a tRNA Processing Ribozyme. 67, 153–180 (1998).

41. Fischer S. The archaeal Lsm protein binds to small RNAs. *J Biol Chem*. 285, 34429–34438 (2010).

42. Nielsen JS, Bøggild A, Andersen CBF, *et al.* An Hfq-like protein in archaea: crystal structure and functional characterization of the Sm protein from Methanococcus jannaschii. *RNA*. 13(12), 2213–2223 (2007).

43. Kabsch W. XDS. 66, 125–132 (2010).

44. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D. Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proc Natl Acad Sci USA*. 103(21), 8060–8065 (2006).

45. Adams PD, Afonine PV, Bunkóczi G, *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. 66, 213–221 (2010).

46. Matthews BW. Solvent content of protein crystals. *J Mol Biol*. 33(2), 491–497 (1968).

47. Pape T, Schneider TR. HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs. 37, 843–844 (2004).

48. Sheldrick GM. A short history of SHELX. 64, 112–122 (2008).

49. Schumacher MA, Pearson RF, Møller T, Valentin-Hansen P, Brennan RG. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. *EMBO J*. 21(13), 3546–3556 (2002).

50. Wang W, Wang L, Wu J, Gong Q, Shi Y. Hfq-bridged ternary complex is important for translation activation of rpoS by DsrA. *Nucleic Acids Res*. 41(11), 5938–5948 (2013).

51. Someya T, Baba S, Fujimoto M, Kawai G, Kumasaka T, Nakamura K. Crystal structure of Hfq from Bacillus subtilis in complex with SELEX-derived RNA aptamer: insight into RNA-binding properties of bacterial Hfq. *Nucleic Acids Res*. 40(4), 1856–1867 (2011).

52. Sauer E, Schmidt S, Weichenrieder O. Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. *Proc Natl Acad Sci USA*. 109(24), 9396–9401 (2012).

53. Kearse M, Moir R, Wilson A, *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. 28, 1647–1649 (2012).

54. Katoh KMK-IKTMK. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. 30, 3059 (2002).

55. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. 17, 754–755.

56. Vagin A, Teplyakov A. MOLREP: an Automated Program for Molecular Replacement. 30, 1022–1025 (1997).

57. Bunkóczi G, Read RJ. Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr D Biol Crystallogr*. 67(Pt 4), 303–312 (2011).

58. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr*. 66(Pt 4), 486–501 (2010).

59. Chen VB, Arendall WB, Headd JJ, *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr*. 66(Pt 1), 12–21 (2009).

60. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA*. 98(18), 10037–10041 (2001).

61. Sun X, Wartell RM. Escherichia coli Hfq binds A18 and DsrA domain II with similar 2:1 Hfq6/RNA stoichiometry using different surface sites. 45, 4875–4887 (2006).

62. Oefner PJ, Huber CG, Umlauft F, Berti GN, Stimpfl E, Bonn GK. High-resolution liquid chromatography of fluorescent dye-labeled nucleic acids. 223, 39–46 (1994).

63. Schumacher MA, Pearson RF, Moller T, Valentin-Hansen P, Brennan RG. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. 21, 3546–3556 (2002).

64. Gesztelyi R, Zsuga J, Kemeny-Beke A, Varga B, Juhasz B, Tosaki A. The Hill equation and the origin of quantitative pharmacology. 66, 427–438.

65. . Nonlinear Regression [Internet]. John Wiley & Sons, Inc., Hoboken, NJ, USA Available from: http://doi.wiley.com/10.1002/0471725315.

66. Hunke C, Antosch M, Müller V, Grüber G. Binding of subunit E into the A-B interface of the A(1)A(O) ATP synthase. 1808, 2111–2118 (2011).

67. Hanes MS, Ratcliff K, Marqusee S, Handel TM. Protein-protein binding affinities by pulse proteolysis: application to TEM-1/BLIP protein complexes. 19, 1996–2000 (2010).

68. Fadouloglou VE, Kokkinidis M, Glykos NM. Determination of protein oligomerization state: two approaches based on glutaraldehyde crosslinking. 373, 404–406 (2008).

69. Warren ME, Brockman AH, Orlando R. On-probe solid-phase extraction/MALDI-MS using ion-pairing interactions for the cleanup of peptides and proteins. 70, 3757–3761 (1998).

Chapter 3: A *Thermotoga* Hfq homolog, from cloning to initial structure solution

Peter S Randolph[a1], Jennifer Patterson[a1], and Cameron Mura[1]

[1]Department of Chemistry, University of Virginia, 409 McCormick Road,
Charlottesville, VA, 22904, USA

[a]PSR and JP are equally contributing authors

1 Abstract

Hfq proteins are the bacterial branch of a ubiquitous RNA-associated 'Sm' protein superfamily. Hfq interacts with myriad RNA species and chaperones RNA···RNA interactions, *e.g.* between small regulatory RNAs and their mRNA targets. Hfq-mediated RNA regulatory networks enable bacteria to rapidly alter their metabolic circuitry in response to environmental fluctuations (temperature, osmotic shock, nutrient gradients, *etc.*), and also underpin bacterial virulence pathways, quorum sensing and biofilm formation. Hfq proteins typically self-assemble into toroidal hexamers that bind RNA, consistent with their homology to the oligomeric Sm proteins. To begin elucidating the roles of Hfq in phylogenetically deep-branching bacteria, including their relationships to archaeal Sm proteins, we have bioinformatically detected a putative Hfq homolog in the genome of the hyperthermophilic bacterium *Thermotoga maritima* (*Tma*). We have cloned and over-expressed *Tma* Hfq, and have purified the recombinant protein to apparent homogeneity. Chemical cross-linking and mass spectrometry, as well as size-exclusion chromatography, reveals *Tma* Hfq hexamers and other oligomeric states *in vitro*. Well-diffracting crystals have been

obtained, with Bragg reflections to better than 2.5 Å; diffraction quality was enhanced by the inclusion of a U_5 RNA in the crystallization conditions. The crystals reproducibly adopt space-group $P2_12_12_1$ (a = 38.7 Å, b = 133.1 Å, c = 205.9 Å), with twelve subunits per asymmetric unit estimated from the Matthews coefficient. Calculation of the self-rotation function indicates that the twelve monomers are related by two-fold and six-fold axes of non-crystallographic symmetry.

2 Introduction

Bacteria inhabit microenvironments that are often spatially heterogeneous or prone to fluctuations in nutrient gradients, temperature, pH, osmolality, oxidative stresses, bacteriophages, local cell density (quorum sensing), *etc*. To endure extrinsic stresses and respond to external stimuli—and, indeed, to leverage changing conditions—bacteria have evolved intricate regulatory circuits that employ small, non-coding RNAs (sRNAs) as generic post-transcriptional modulators of gene expression [1,2]. Unlike the roughly analogous eukaryotic microRNAs, most identified sRNAs are ≈50–250 nucleotides in length, and exert both negative and positive downstream effects (not just RNA silencing); the biogenesis of sRNAs likely stems from 'pervasive transcription' of the bacterial genome [3,4]. In order to rapidly alter cellular metabolic programs and otherwise serve as *in vivo* ribo-regulators, many sRNAs require an abundant (>50 000 copies/cell) RNA-associated protein known as Hfq [5].

Hfq was discovered as an *Escherichia coli* host factor required for bacteriophage Qβ RNA replication, and structural, biochemical and bioinformatic analyses revealed it to be the bacterial branch of the Sm superfamily [6-11]. Homologs of Sm proteins pervade RNA-associated biochemistry in eukaryotes, including pre-mRNA splicing and mRNA decay pathways [12,13]. The relationships between Hfq and other Sm lineages have been recently reviewed [14]. Whereas eukaryotic Sm proteins act as structural scaffolds for snRNPs and other ribonucleoprotein assemblies, Hfq acts as an RNA chaperone—*i.e.*, an RNA-binding protein with flexible sequence recognition capacity, such that it can facilitate base-pairing interactions between a diverse array of regulatory sRNAs and mRNA targets [15]. Because many sRNAs are *trans*-encoded, antisense sRNA···RNA$_{target}$ interactions often feature only partial base-pair complementarity and require Hfq for productive annealing. By independently binding cognate RNAs, Hfq increases their local effective concentration and thus enhances binding affinities. In addition, Hfq can reduce the thermodynamic stability of sRNA hairpins in isolation and modulate the *in vivo* stability (mean lifetimes) of transcripts by tuning mRNA polyadenylation and either enhancing or reducing an RNA's susceptibility to RNase E–based degradation [16-18]. Consistent with the broad range of its ribo-regulatory roles, Hfq has been linked to many pathogenic pathways, including microbial quorum sensing, biofilm formation and virulence factor expression [19-21].

Hfq-based RNA interactions can alter gene expression in seemingly incongruous ways. For instance, Hfq helps repress some protein expression pathways but activate others, depending on the sRNA/mRNA pair triggered by an upstream environmental or physiological cue. Hfq-mediated sRNA···mRNA interactions often occur in the untranslated region (UTR) of a transcript, *e.g.* the 5′ leader region of a target mRNA, and result in stem-loops or other RNA structures. Positive regulation can occur, for instance, if the ribosome-binding site (RBS) in the 5′ UTR of an un-activated mRNA is sequestered in a hairpin; in concert with Hfq, a 'seed region' in the sRNA can hybridize to the mRNA stem-loop, disrupting local secondary structure and unmasking the RBS to enable translation initiation. The DsrA/*rpoS* pair is a classic example, where *rpoS* mRNA encodes the *E. coli* stationary phase/stress-response sigma factor (σ^s) and DsrA is a temperature-sensitive sRNA. Conversely, other sRNAs (*e.g.*, OxyS, induced by oxidative stress) might bind to and occlude an otherwise exposed RBS, thereby hindering ribosome recruitment and repressing protein expression. In addition to sRNA/antisense-based control, some Hfq systems regulate translational activity *via* Hfq-mediated effects on mRNA stability and turnover [18]. Dozens of sRNA-based pathways are now well-documented [22-25], as are *(i)* potential mechanisms by which

Hfq might enable productive sRNA···mRNA interactions [26,27] and *(ii)* methods for identifying and characterizing Hfq···RNA interactions at the molecular and 'omics' scales [28]. Nevertheless, questions persist about the precise (atomic) structural and dynamical basis of Hfq-facilitated RNA annealing and ribo-regulation.

Hfq-chaperoned sRNA···mRNA interactions stem from the RNA– and protein–binding properties of Hfq in its monomeric and hexameric states, as modulated by the equilibria between various oligomeric states [29]. Hfq monomers form homo-hexameric rings in the absence of RNA, unlike the stepwise, RNA-templated assembly of eukaryotic Sm hetero-oligomers [14]; Hfq hexamers also associate into double-ring dodecamers [30,31] and polymeric fibrils [32], at least *in vitro*. All such properties ultimately derive from the 3D structures of individual Hfq subunits. Hfq monomers fold as a highly-bent, five-stranded antiparallel *β*-sheet capped by an *N*-terminal α-helix; like most Sm proteins, these subunits associate into cyclic oligomers (≈ 60-70 Å diameter) that bind RNA. Hfq can bind a great variety of RNAs, and the two faces of the toroidal disc exhibit some sequence specificity: A–rich RNAs, such as poly(A) tails of a target mRNA, preferentially bind the L4 (or 'distal') face, while U–rich RNAs bind the opposite L3 (or 'proximal') face, near the ≈ 10 Å–wide central pore [26,33,34]. This sequence specificity is recapitulated when ribonucleoside triphosphates (ATP, UTP, CTP, *etc.*) are used to crystallographically survey the RNA-binding preferences of the *Pseudomonas aeruginosa* Hfq ring [35]. A third binding surface, the lateral rim binding, has been identified in multiple Hfqs. Lateral rim binding displays a UA-rich specificty, but does not demonstrate the affinity of the other faces, and could represent a transfer point between the L3 and L4 faces. Despite much progress, the generality of our Hfq knowledge is limited by the fact that most studies have examined a narrow range of bacterial lineages—chiefly *E. coli* and other mesophilic *γ*-proteobacteria. Here, we report initial structural studies of an Hfq homolog from the hyperthermophilic, phylogenetically ancient marine bacterium *Thermotoga maritima* (*Tma*).

3 Materials and Methods

3.1 Cloning, Over-expression and Purification

Plasmid DNA containing the *Tma* Hfq open reading frame (TM0526) was provided by the JCSG. Cloning proceeded *via* the ligase-free Polymerase Incomplete Primer Extension (PIPE) [36] approach, with PIPE primers designed for the *Nde*I and *Xho*I restriction sites of the pET-28b(+) expression vector (Table 1); *Pfu* DNA polymerase was used for all PCRs. The recombinant expression construct, denoted His_6–*Tma* Hfq, consists of wild-type *Tma* Hfq linked to an *N*-terminal His_6 tag (Figure 1*a*). Upon completing the requisite pair of PIPE reactions (*i.e.*, insert, vector), PCR-amplified vectors and inserts were mixed to allow DNA hybridization, and co-transformed into chemically competent *E. coli* TOP10 cells (Invitrogen) for *in vivo* ligation and maintenance. The plated TOP10 cells were incubated at 310 K overnight (≈ 16 h) to amplify/propagate this new recombinant expression clone, and the plasmid was then mini-prepped (Qiagen) from overnight cultures for long-term maintenance and downstream applications such as DNA sequencing (Genewiz) or transformation into an expression host. Next, *(i)* the purified recombinant plasmid was chemically transformed into BL21(DE3) *E. coli*, *(ii)* cells were plated on kanamycin-supplemented LB-agar media to select for transformants and *(iii)* 10-mℓ overnight starter cultures were used to inoculate larger-scale (≈ 1–ℓ) production cultures. All microbiological steps utilized standard protocols (Ausubel, 2002), such as growth at 310 K in lysogeny broth (LB), selection with ≈ 50 µg $\text{m}\ell^{-1}$ kanamycin in LB, and aeration of cultures *via* shaking at ≈ 230 rpm. Once cultures reached high confluence ($\text{OD}_{600\text{nm}}$ ≈ 0.8-1), recombinant Hfq was over-expressed by using 1 m*M* isopropyl-1-thio-β-D-galactopyranoside (IPTG) to induce the T7*lac*-based promoter. After ≈ 3-4 h of induction, cells were harvested *via* centrifugation (15000*g*, 5 min, 277 K) and stored at 253 K until further use.

To begin purifying *Tma* Hfq, cells were simultaneously thawed to ambient room temperature (≈ 293 K) and re-suspended in a moderately high-salt buffer, and were lysed by high-pressure mechanical shearing and lysozyme treatment. Specifically, a buffer of 50 m*M* Tris-HCl pH 7.5, 750 m*M* NaCl, 0.4 m*M* phenylmethylsulfonyl fluoride (PMSF) and 0.01 mg mℓ$^{-1}$ hen egg-white lysozyme (Fisher) was added to frozen cell pellets at a ratio of ≈ 40-mℓ buffer per 1-ℓ cell-culture pellet, at room temperature. This mixture was then incubated with gentle shaking at 310 K for 30 min in order to fully thaw and resuspend the cells, and initiate lysis. Cell suspensions were then mechanically disrupted *via* repeated passes through a microfluidizer (Microfluidics), and the resultant lysate was centrifuged (35000*g*, 20 min, 277 K). Initial purification of His$_6$–*Tma* Hfq was achieved by heating this cleared supernatant to ≈ 358 K for 20 min, followed by centrifugation (35000*g*, 30 min, 277 K) to remove the bulk of denatured *E. coli* proteins (Figure 1*b*). Because Hfq tends to promiscuously bind nucleic acids, His$_6$–*Tma* Hfq was freed of co-purifying nucleic acids [37] *via* treatment with guanidinium chloride (GndHCl); specifically, dry GndHCl was added to the buffered His$_6$–*Tma* Hfq sample to a final concentration of 6 *M*. Such treatment reduced any nucleic acid contamination of Hfq, as assessed by A_{260}/A_{280} ratios below ≈ 0.80 (even for samples recalcitrant to nuclease treatment; Patterson & Mura, data not shown). The His$_6$–*Tma* Hfq sample was then loaded onto a 5-mℓ Ni^{2+}-charged iminodiacetic acid–sepharose affinity column (GE Healthcare HiTrap) pre-equilibrated with wash buffer (50 m*M* Tris-HCl pH 7.5, 150 m*M* NaCl, 6 *M* GndHCl, 10 m*M* imidazole), and this was followed by the addition of ≈ 8-10 column volumes of wash buffer. An elution buffer identical to the wash buffer, apart from 600 m*M* imidazole, was then applied as a 0–100% gradient over several column volumes. Elution fractions were analysed by SDS-PAGE, and sufficiently pure fractions (estimated as ≳ 95% in gel lanes) were pooled.

Next, for crystallization and other downstream steps, purified His$_6$–*Tma* Hfq samples were *(i)* exchanged into a buffer free of GndHCl and *(ii)* subjected to limited proteolysis in order to remove the affinity tag (red arrow, Figure 1*a*). Step *(i)* was achieved *via* extensive dialysis, at room temperature, against a buffer composed of 25 m*M* Tris-HCl pH 8.0, 1.0 *M* arginine, 0.2 m*M* PMSF (arginine was found to aid protein solubility). To prepare for step *(ii)*, the resulting protein sample was dialyzed into a digestion buffer (50 m*M* Tris-HCl pH 8.0, 150 m*M* NaCl, 12.5 m*M* EDTA pH 8.0) and samples were then digested with a 1:600 molar ratio of thrombin:His$_6$–*Tma* Hfq. Proteolysis was allowed to proceed overnight (≈ 14 h) at 310 K, and cleavage was monitored *via* matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (Figure 1). Proteolysis was halted by addition of PMSF (to 0.2 m*M*), and the *Tma* Hfq sample—now containing *N*-terminal fragments, thrombin and potentially other contaminants—was applied to a benzamidine affinity column in order to remove thrombin. The eluate from this step was pooled, concentrated so as to reduce volume, and then subjected to a final step of purification *via* size-exclusion chromatography. This step was performed at 277 K using a preparative-grade, HiLoad Superdex 75 column pre-equilibrated with the digestion buffer; this resin is well-suited to a molecular weight range ($3000 < M_r < 70000$) that brackets *Tma* Hfq monomers (10.8 kDa) and potential hexamers (64.8 kDa). For crystallization trials, purified *Tma* Hfq was dialyzed into buffer 'TmaHfqXB', consisting of 25 m*M* Tris pH 8.5 and 100 m*M* NaCl (the p*I* of *Tma* Hfq is ≈ 7.0, based on sequence) [38]. Protein concentrations were quantified by measuring the absorbance at 280 nm and using an extinction coefficient of $\epsilon_{280} = 5960\ M^{-1}\ \mathrm{cm}^{-1}$, as estimated from the amino acid composition of the 95-residue, 10.8-kDa final construct (Figure 1*a*) [39]. All protein concentration steps employed centrifugal ultrafiltration units with 3350-Da molecular weight cut-off membranes (Millipore).

3.2 Chemical Cross-linking and Mass Spectrometry

Cross-linking studies employed the reagent 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC; Pierce); this zero-length cross-linker conjugates carboxylate and primary amine functionalities without being included in the resulting amide bond. Reactions were prepared by mixing EDC and samples of purified *Tma* Hfq that had been dialyzed into a buffer

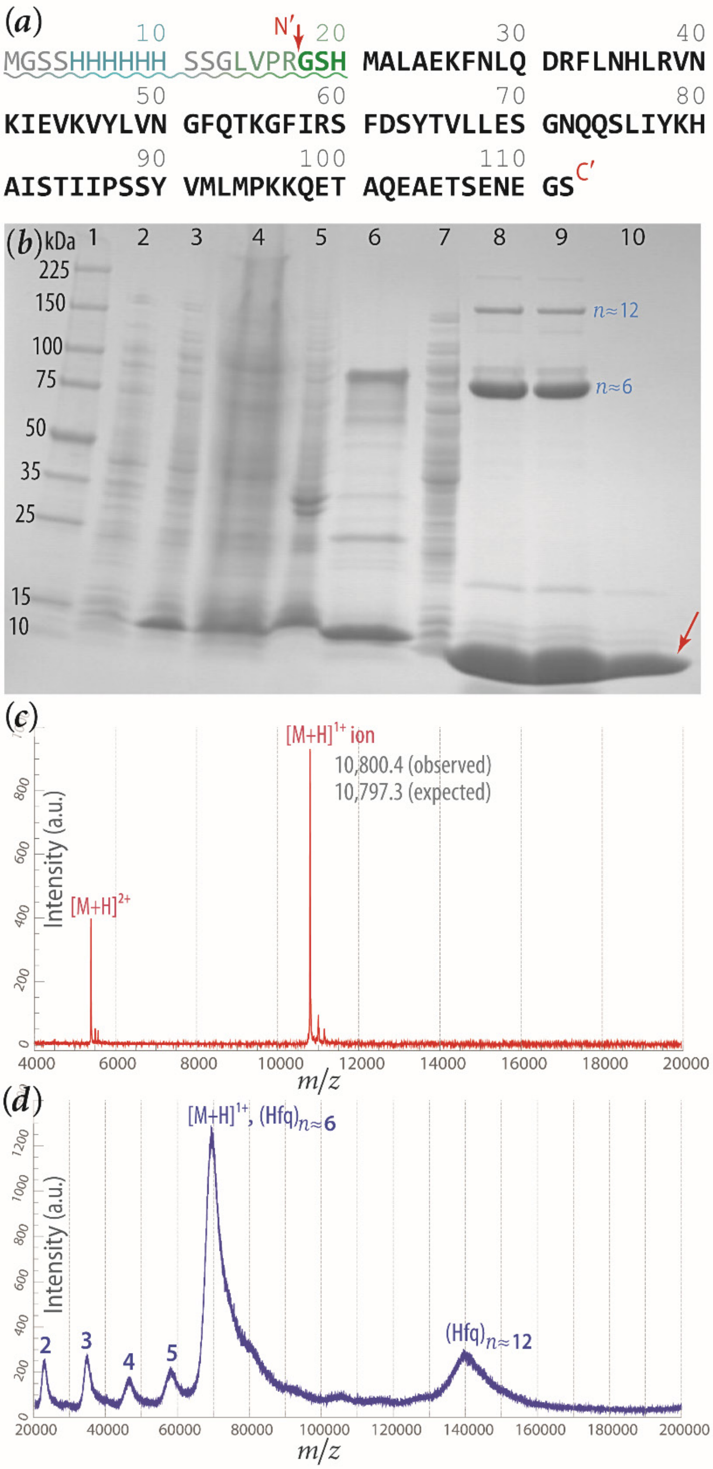

(a)
10
N′
20
30
40
MGSSHHHHHH SSGLVPRGSH MALAEKFNLQ DRFLNHLRVN
50
60
70
80
KIEVKVYLVN GFQTKGFIRS FDSYTVLLES GNQQSLIYKH
90
100
110
AISTIIPSSY VMLMPKKQET AQEAETSENE GS
C′
(b)
kDa
1
2
3
4
5
6
7
8
9
10
225
150
100
75
50
35
25
15
10
n≈12
n≈6
(c)
[M+H]1+ ion
10,800.4 (observed)
10,797.3 (expected)
[M+H]2+
Intensity (a.u.)
0
200
400
600
800
4000
6000
8000
10000
12000
14000
16000
18000
20000
m/z
(d)
[M+H]1+, (Hfq)n≈6
(Hfq)n≈12
2
3
4
5
Intensity (a.u.)
0
200
400
600
800
1000
1200
20000
40000
60000
80000
100000
120000
140000
160000
180000
200000
m/z

Figure 1. Expression and purification of *Tma* Hfq. The recombinant *Tma* Hfq construct: Over-expression, purification and initial characterization of oligomers *via* chemical cross-linking and mass spectrometry. The amino acid sequence of the Hfq^{Tma} construct (*a*) consists of the 92-amino acid wild-type *Tma* Hfq sequence (bold font) with an *N*-terminal, vector-derived $(His)_6$ affinity tag (cyan) and a recognition site for thrombin (green); the red arrow marks the proteolytic cleavage site. The final, purified protein, delimited by *N′* and *C′*, corresponds to wild-type Hfq^{Tma} prepended with a tripeptide (GSH). The over-expression and purification workflow is illustrated in this SDS-PAGE gel (*b*) by steps of pre– and post–induction (lanes 2, 3); the post-lysis supernatant and pellet (lanes 4, 5); soluble and insoluble fractions after a heat-cut (lanes 6, 7); initially pure Hfq^{Tma}, after Ni^{2+}–IMAC and thrombin cleavage (lane 8), and then a benzamidine column (lane 9); and, finally, after size-exclusion chromatography (lane 10 is the monomer peak from the chromatogram). MALDI-TOF mass spectra are shown for purified recombinant Hfq^{Tma} before (*c*) and after (*d*) treatment with the cross-linking reagent EDC; the 1+ and 2+ molecular ions are indicated. The Hfq sample in (*c*) is highly pure, and matches the expected molecular mass of recombinant Hfq^{Tma} monomer (10,797.3 Da). The predominant peak in the cross-linked Hfq^{Tma} mass spectrum (*d*) is a hexamer, while a dodecameric species is also detected; subsidiary peaks correspond to $(Hfq)_n$ molecular masses for the $n \approx 2, 3, 4, 5$ oligomeric states, as labelled. The red arrow in the rightmost lane of the gel (*b*) corresponds to monomeric Hfq^{Tma}; presumptive hexamers and dodecamers, such as detected by cross-linking and mass spectrometry, are marked in blue (lanes 8, 9).

of 25 m*M* HEPES pH 8.0, 200 m*M* NaCl; typical experiments contained 1.0 mg mℓ$^{-1}$ *Tma* Hfq (≈ 93 μ*M* monomer), 67 m*M* EDC and 167 m*M* *N*-hydroxysulfosuccinimide (sulfo-NHS, to enhance coupling efficiency). Reactions were generally incubated at room temperature for 4 h and then quenched *via* the addition of β-mercaptoethanol (βME) to a final concentration of ≈ 18 m*M*. The reaction progress and resulting distribution of cross-linked oligomers were assayed by MALDI-TOF mass spectrometry (Figure 1*d*). Samples were prepared for MALDI by either *(i)* passage through a C4 ZipTip (Millipore) for analyte purification and concentration or *(ii)* dilution of ≈ 1 mg mℓ$^{-1}$ samples, at a ratio of 1:4, with 0.01% trifluoroacetic acid (TFA). Taking the dried-droplet approach, the resulting samples were then mixed ≈ 1:1 with a matrix solution (15 mg mℓ$^{-1}$ sinapinic acid in 50% acetonitrile, 0.05% TFA) on a stainless steel MALDI plate (giving ≈ 2-5–μℓ spots) and crystallized *in situ via* solvent evaporation. MALDI-TOF spectra were acquired on a Bruker microflex instrument operated with routine settings (linear, positive-ion mode; 25 kV accelerating voltage; 40-75% grid voltage); final spectra were obtained by averaging ≈ 50 laser shots, and low–(4-20 kDa) and high (20-100 kDa)–molecular weight calibrants were used for these different *m*/*z* ranges.

3.3 Protein Crystallization

Crystallization trials began with freshly purified *Tma* Hfq samples concentrated to ≈ 10-20 mg mℓ$^{-1}$ in buffer TmaHfqXB (above). JCSG Core Suites I, II, III and IV (Qiagen), as well as PEG/Ion screens (Hampton Research), were used for sparse-matrix screening. Screens were deployed in 96-well plates *via* a nanolitre-scale liquid-handling robot (TTP Labtech's mosquito Crystal). The hanging-drop vapour diffusion format was used, with a 100-μℓ reservoir and 200-nℓ droplet (composed as 1:1 protein:reservoir). All crystallization trials were incubated at 291 K. Visual inspection of the initial screens suggested five crystallization leads. These leads were then refined on a larger-volume scale (24-well VDX plates) by systematically varying typical crystallization parameters [40]—buffer pH, protein concentration and precipitant types/concentrations (*e.g.*, a PEG-200, 400, 1000, … series). In addition to fine grids centred on the initial hits, 96-well additive screens (Hampton Research) were also applied to the leads. Two improved conditions were found, corresponding to 150 m*M* tri-potassium citrate and 30% w/v PEG-3350 (Table 2) with either 1.0 *M* glycine or 0.1 *M* sarcosine as an additive. Final, optimized crystals grew as rhombic prisms within 1 week and developed to maximal dimensions of ≈ 50 μm/edge by 2 weeks (Figure 2*a*); many crystalline specimens began to fracture beyond ≈ 3 weeks. A working cryo-protection procedure for both the glycine- and sarcosine-based conditions was found to be gentle passage of a crystal (in a nylon cryo-loop), over the course of ≈ 10-15 s, through 8 μℓ of mother liquor supplemented with 0.6 μℓ of neat PEG-400. Suitable cryo-protection (lack of ice-rings, diffraction quality, *etc*.) was screened on a Rigaku MicroMax-007 rotating anode X-ray generator equipped with a Saturn-92 charged-coupled device (CCD) detector.

3.4 Diffraction Data Collection, and Processing

Single crystals were harvested and mounted in nylon loops, and instantly transferred to a cryoprotective solution (Table 2; described above) before flash-cooling in a liquid nitrogen stream, for either long-term storage or shipment to synchrotron beamlines. Diffraction datasets were collected on either *(i)* SER-CAT 22-ID or 22-BM beamlines, using a MAR 300 or MAR 225 CCD detector, respectively; or *(ii)* NE-CAT 24-ID-C, using a DECTRIS Pilatus pixel array detector. Complete diffraction datasets were collected for several crystals, and the best apo crystal was found to diffract to at least 2.7 Å. Full datasets were also collected for crystals grown in drops that contained ≈ 3 μ*M* of the oligoribonucleotide r(U)$_5$. Raw diffraction data were indexed, integrated and scaled using the programs XDS and XSCALE [41]. Reduced datasets were examined with Xtriage and other utilities in the PHENIX suite in order to verify indexing, gauge diffraction anisotropy, detect pseudo-translational symmetry, and so on; twinning tests were also performed, though merohedral twinning was not a concern here because of the orthorhombic crystal system and unequal cell edges [42]. Non-crystallographic symmetry (NCS) was evaluated by computing *(i)* the Matthews coefficient; *(ii)* native Patterson maps, using CCP4's FFT module; and *(iii)* the self-rotation function, using GLRF [43-45].

4 Results and Discussion

T. maritima is a hyperthermophilic, Gram-negative species with an unusually compact genome, and is among the most deeply branching and slowly evolving of bacterial lineages [46,47]. That the 'streamlined' *Tma* genome is partly archaeal in origin, and is predicted to encode relatively few sRNAs, makes *Tma* Hfq (and any associated sRNA regulatory networks) salient to comparative analyses of Sm structure and function. We identified a putative *Tma* Hfq homolog *via* iterative PSI-BLAST [48] searches of the most current non-redundant database of protein sequences at NCBI, as well as manual exploration of the TOPSAN (Krishna *et al.*, 2010) and JCSG [49] knowledgebases. The *hfq* gene that was identified (accession numbers in Table 1) had been targeted by JCSG's structural genomics project, but became frozen at the crystallization stage; thus, to avoid potential downstream difficulties, a *Tma* Hfq construct was cloned *de novo*. DNA sequencing confirmed that the recombinant His_6–*Tma* Hfq expression construct (Figure 1*a*) was successfully created *via* PIPE cloning. The protein over-expressed effectively, as indicated by a comparison of SDS-PAGE lanes for pre– and post–induction whole-cell lysates (Figure 1*b*, lanes 2 and 3, respectively). As shown in Figure 1*b*, over-expressed His_6–*Tma* Hfq was purified *via* successive stages of heat-cut → Ni^{2+}–affinity chromatography → proteolysis/clean-up → size-exclusion chromatography.

Hfq tends to bind nucleic acids promiscuously, and to self-assemble into hexamers (and supra-hexameric states) that resist thermal and chemical denaturation. These properties can aid or hinder purification efforts. Initial purification of *Tma* Hfq was achieved by heating cell lysates to 358 K; like other homologs, and consistent with the optimal growth temperature of *Tma* ($T_{\text{opt}} \approx 353$ K), *Tma* Hfq remains soluble at this elevated temperature (Figure 1*b*, lane 6). The oligomers also

Table 1. Macromolecular cloning and expression.

Source organism	*Thermotoga maritima* strain MSB8
DNA source	*Tma* locus TM0526
PIPE forward primer (*insert*)	$^{5'}$GCGCGGCAGCCA↓TATGGCCTTGGCGGAGAAGTTCAACCTTCAG$^{3'}$
PIPE reverse primer (*insert*)	$^{5'}$GTGGTGC↓TCGAGTCAAGATCCCTCGTTTTCAGAGGTCTCAGCC$^{3'}$
PIPE forward primer (*vector*)	$^{5'}$C↓TCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAAC$^{3'}$
PIPE reverse primer (*vector*)	$^{5'}$CA↓TATGGCTGCCGCGCGGCACCAGGCCGCTGCTGTGATGATGATG$^{3'}$
Cloning vector	pET-28b(+)
Expression vector	pET-28b(+)
Expression host	*Escherichia coli* BL21(DE3)
Complete amino acid sequence of the recombinant construct	See Fig 1*a*. Other accession codes and database identifiers: UniProt ID Q9WYZ6; RefSeq WP_010865141; GenPept 4981039; GeneID 897578

The *Nde*I and *Xho*I restriction sites are underlined, and arrows indicate the endonucleolytic cut-sites; vector-derived amino acids in the recombinant protein, including the *N*-terminal His$_6$ affinity tag, are specified in Fig 1*a*. Recombinant wild-type *Tma* Hfq was expressed in native form, without selenomethionine or other forms of labelling.

seem SDS-resistant: Faint bands corresponding to hexamers and other oligomers persist even in denaturing gels (Figure 1*b*, lanes 6, 8, 9), similarly to *E. coli* Hfq [29,50]. As described above, the initial affinity chromatography step required the chaotrope GndHCl to strip away nucleic acids found to co-purify with His_6–*Tma* Hfq. The GndHCl was removed before proteolysis, and downstream experiments, such as CD spectroscopy and size-exclusion chromatography, suggested that the GndHCl-treated samples had resisted denaturation (data not shown). Proteolysis of these samples removed the His_6 affinity tag, yielding recombinant *Tma* Hfq that is nearly identical to the wild-type sequence (Figure 1*a*). Size-exclusion chromatography served both as a final purification step (Figure 1*b*, lane 10) and to verify that previously Gnd-treated *Tma* Hfq could form hexamers (Chapter 4). The purity and identity of *Tma* Hfq samples were confirmed by MALDI-TOF MS (Figure 1*c*), with the expected molecular mass matched to within 0.01%; additionally, the final samples are free of detectable contaminants, including in higher-mass regions beyond 20 kDa (*e.g.*, cross-linked sample in Figure 1*d*).

To prepare for structure determination, and because of potential ambiguity regarding the oligomeric states of Hfq—from monomers, to hexamers, to higher-order species such as stacked double-rings and polymeric fibrils [32,51]—we used chemical cross-linking, followed by MS, to characterize the oligomeric states of *Tma* Hfq *in vitro*. Such an approach has been used to examine the subunit stoichiometry of *E. coli* Hfq in hexameric rings and potentially higher-order (double-ring) assemblies bound to RNA [6,30]. Purified *Tma* Hfq (Figure 1*c*) was found to cross-link as hexamers *in vitro*, consistent with the oligomerization behaviour of known Hfq homologs and with size-exclusion data (See Chapter 4). The $(\text{Hfq})_n$ oligomeric states n = 2, 3, 4, 5, 6 and 12 also appear in the MALD-TOF spectra of cross-linked samples (Figure 1*d*); notably, these low abundancy states also occur in gels of non–cross-linked samples, prior to size-exclusion chromatography (Figure 1*b*, lanes 8, 9).

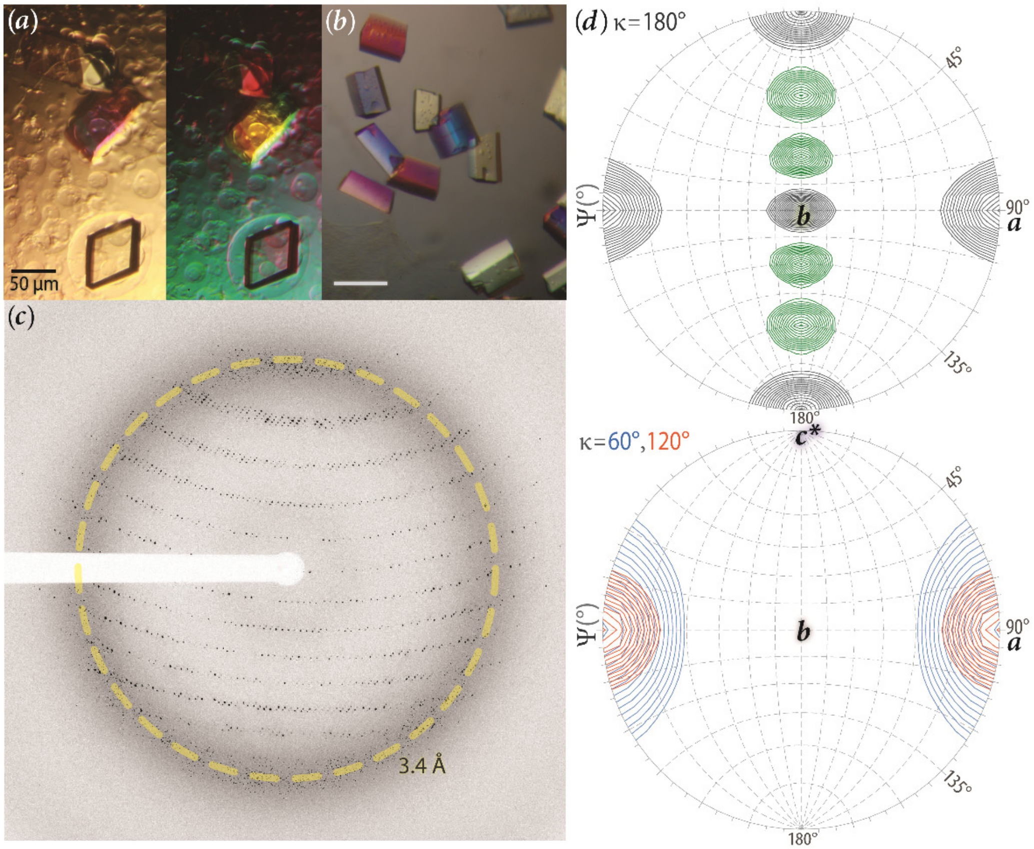


Figure 2. Crystallization and diffraction of *Tma* Hfq. Well-diffracting crystals of *Tma* Hfq were found to exhibit non-crystallographic symmetry. Diffraction-grade crystals (*a*, *b*) were grown by optimizing leads that were discovered by sparse-matrix screening. The scale-bar in (*a*) and (*b*) represents 50 µm. Consistent with a non-cubic space-group, images taken under cross-polarized light (*a*, *b*) exhibit birefringence, as most clearly seen when the analyser is rotated between otherwise identical images in the left/right halves of panel (*a*). A representative diffraction pattern (*c*) exhibits Bragg peaks to beyond 3.0 Å; the yellow circle marks a characteristic ≈ 3.4 Å-ring, arising from diffuse scattering of vitreous ice. The κ = 60°, 120° and 180° sections of the self-rotation function, computed in spherical polar coordinates, are shown as stereographic projections in (*d*); the maps are contoured at values starting at 2.25σ above the mean, with a contour interval of 0.25σ. For clarity, peaks corresponding to NCS are coloured blue (60°), orange (120°) or green (180°).

Screening of crystallization and cryo-protection conditions for purified *Tma* Hfq (Table 2) yielded reproducible, well-diffracting specimens (Figure 2). Many initial crystals exhibited anisotropic diffraction, with a substantial fraction of reflections distorted over large wedges of data collection; anisotropy varied among the crystals within a drop, and screening of diffraction quality ultimately led to usable native datasets (Table 3, Specimen-1). Notably, improved diffraction ($d \approx 2.1$ Å) was observed for crystals grown in drops with low concentrations of U_5 RNA (Table 3, Specimen-2); this RNA was sub-stoichiometric (≈ 3 μ*M*, *vs* m*M*-range [Hfq]), serving as more of an additive than a proper co-crystallizing agent. Crystals grown in the absence or presence of RNA were isomorphous, and belonged to space-group $P2_12_12_1$ with unit cell dimensions of $a \approxeq 39$ Å, $b \approxeq 133$ Å, $c \approxeq 206$ Å. The final, optimized diffraction datasets show insignificant anisotropy and no twinning (consistent with an orthorhombic space-group), and the statistics are of sufficient quality for phasing efforts (Table 3). The Matthews coefficient/solvent content (Table 3) is most compatible with ≈ 9–12 Hfq subunits per asymmetric unit (AU).

Available Hfq crystal structures suggest that 12 subunits would arrange as two hexamers rather than, *e.g.*, one 12 dodecamer. Notably, the $\vec{a}$ cell edge (Table 3) approximates the thickness of a typical Sm disc, implying that two *Tma* Hfq rings must pack laterally in the AU, potentially in an arrangement similar to the side-by-side heptameric rings in the $P2_12_12_1$ form of *M. thermautotrophicum* SmAP1 [51]. To assess rotational symmetry within and between Hfq rings, the self-rotation function was computed. As expected for crystals of 222 point-group symmetry, three peaks for mutually perpendicular 2–fold axes occur on the $\kappa = 180°$ section (Figure 2*d*, black peaks). More interesting, a great circle of six equally spaced 2–folds occurs perpendicular to $\vec{a}$ (Figure 2*d*, green peaks), along with 6– and 3–fold axes on the $\kappa = 60°$ and $120°$ sections, respectively. The spherical coordinates (φ, ψ) of the 3– and 6–fold axes show them to be parallel to one another and to the crystallographic $\vec{a}$ direction, and therefore perpendicular to the aforementioned band of crystallographic (and non-crystallographic) 2–fold axes. The self-rotation

Table 2. Macromolecular crystallization.

Method	Hanging-drop vapour diffusion
Plate type	VDX plates
Temperature (K)	291
Protein concentration	15.3 mg ml^{-1} (≈ 1.4 m*M* in monomer)
Buffer composition of protein solution	25 m*M* Tris, pH 8.5; 100 m*M* NaCl
Buffer composition of reservoir solution	150 m*M* tri-potassium citrate; 30% (w/v) PEG-3350
Buffer composition of additives to drop	1.0 *M* glycine (and, optionally, U$_5$ RNA [see text])
Drop composition (protein + reservoir + additive)	5 µl (2.5 µl + 2.0 µl + 0.5 µl)
Volume of reservoir	600 µl
Cryoprotectant	Mother liquor supplemented with 4.4% v/v PEG-200

Table 3. X-ray diffraction data collection and processing. Values for the highest resoltution shell are given in parentheses

	Tma Hfq [apo]	*Tma* Hfq [r(U)$_5$ additive]
X-ray source (beamline)	APS SER-CAT 22-ID	APS NE-CAT 24-ID-C
Wavelength (Å)	0.97879	0.97920
Temperature (K)	100	100
Detector	MAR CCD 300mm	Pilatus–6MF
Crystal-to-detector distance (mm)	300	300
Rotation range per image (°)	1.0	0.5
Total rotation range (°)	360.0	180
Exposure time per image (s)	1.0	0.5
Space group	$P2_12_12_1$	$P2_12_12_1$
a, *b*, *c* (Å)	39.09, 133.50, 206.18	38.67, 133.13, 205.85
α, *β*, *γ* (°)	90, 90, 90	90, 90, 90
Mosaicity (°)	0.242	0.092
Resolution range (Å)	81.594 – 2.653	81.43 – 2.05
No. of reflections (total)	469 249	447 385
No. of reflections (unique)	32 350	68 036
Completeness (%)	99.8 (98.1)	99.7 (98.7)

Redundancy	14.5 (14.1)	6.6 (6.4)
Signal-to-noise of merged intensities, $\langle I/\sigma(I)\rangle$	21.5 (2.5)	19.1 (2.3)
R_{sym} † (%)	13.9 (150.2)	4.6 (80.6)
R_{meas} ‡ (%)	14.5 (155.9)	5.0 (87.6)
$R_{\text{p.i.m.}}$ ‡ (%)	3.8 (89.2)	2.0 (34.1)
$CC_{1/2}$ § (%)	99.9 (89.2)	99.9 (94.5)
Overall B-value from Wilson plot (Å^2)	51.1	36.39
Matthews coefficient, V_{M} (Å^3 Da^{-1})	2.08 (for 12 subunits/AU)	2.04 (for 12 subunits/AU)
Solvent content (% volume)	40.7	39.8

†$R_{\text{sym}} = (\sum_{hkl} \alpha \sum_i |I_i(hkl) - \langle I_i(hkl)\rangle|)/(\sum_{hkl} \sum_i I_i(hkl))$, where $I_i(hkl)$ is the intensity of the i^{th} observation of reflection hkl, $\langle\cdot\rangle$ denotes the mean of symmetry-related (or Friedel-related) reflections, and the coefficient $\alpha = 1$; the outer summations run over only unique hkl with multiplicities greater than one.

‡ R_{meas} is defined analogously as R_{sym}, save that the prefactor $\alpha = \sqrt{N_{hkl}/(N_{hkl} - 1)}$ is used; N_{hkl} is the number of observations of reflection hkl (index $i = 1 \rightarrow N_{hkl}$). Similarly, the precision-indicating merging R-factor, $R_{\text{p.i.m.}}$, is defined as above but with the prefactor $\alpha = \sqrt{1/(N_{hkl} - 1)}$.

§ $CC_{1/2}$ is the correlation coefficient between intensities chosen from random halves of the full dataset.

peaks are consistent with a *Tma* Hfq hexamer, and also imply additional 2–fold symmetry among the two hexamers in the AU (again, assuming a 2x6 arrangement). The two hexamers are not related by a pure translation, given the absence of significant non-origin peaks in native Patterson maps (data not shown). Along with the cell dimensions and 2x6 rim-to-rim packing argument (above), the lack of native Patterson peaks implies that the non-crystallographic 2–folds detected in the self-rotation function are not *perfectly* parallel to the crystallographic 2_1 screw axes of $P2_12_12_1$. Such could occur if *(i)* the 6–folds of the hexamers in the AU are not perfectly parallel to the crystallographic $\vec{a}$, *i.e.*, the Hfq ring is tilted with respect to that cell face, causing the NCS 2–folds to misalign with the 2_1 axes; and *(ii)* the NCS 2–folds within(/among) the rings are not perfectly parallel to the 2_1 screw axes, *i.e.* the ring is not in perfectly ideal rotational 'register' (with respect to the cell edges) in the $\vec{b}$ x $\vec{c}$ plane.

Tma Hfq crystal structure was solved utilizing a probe of *Bacillus subtillus* Hfq hexamer in Phaser [42]. As expected, *Tma* Hfq crystallizes as a pair of hexamers side-by-side (not stacked), with the hexamers tilted in respect to each other, which is why there was no pure translation peak in the Patterson maps. Initial refinement was hampered by systematic alternating positive and negative difference density (F_o-F_c) (Figure 3) distributed throughout the structure (in protein and solvent channels. After refinement with the alternate density (Chapter 4, Table 4 for refinement statistics), we applied an alternative, more rigorous, anisotropy correction method (UCLA anisotropy correction server [52]) on the initial data and resolved the structure using the same methodology. This corrected the alternating density and improved the overall structure quality (Chapter 4 Table 4). With the improved data-quality we were able to identify bound co-purified nanoRNA U_6 RNA in the proximal (L3) pore, which was not apparent in the initial solution. The *Tma* Hfq monomer is conserved among the two hexamer, with only small variations in the Loop L5 and the N-terminal tail. *Tma* Hfq crystal structure contains a stable N-terminal tail which extends almost parallel to $\vec{a}$, contacting the 'lower' ASU. The N-terminal tail bridges a solvent channel ~12

Å wide, running the length of the ASU. Two parallel ASUs run diagonally from the corner of the cell, stacked head-to-tail, separated by the previously mentioned solvent channel. The ASU (*n-2*) in $\vec{a}$ faces the opposite direction, stacking head-to-head with (*n-1*) ASU. This can be seen in the lattice packing in Figure 4, with minimal crystal contacts between parallel stacked ASUs, and the majority of crystal contacts between side-by-side ASUs which tilt down, creating a spiral or spring. Few contacts between adjacent ASUs can generate disordered crystals, possibly explaining the observed anisotropy. U RNA and other possible binding sites will be discussed in Chapter 4.

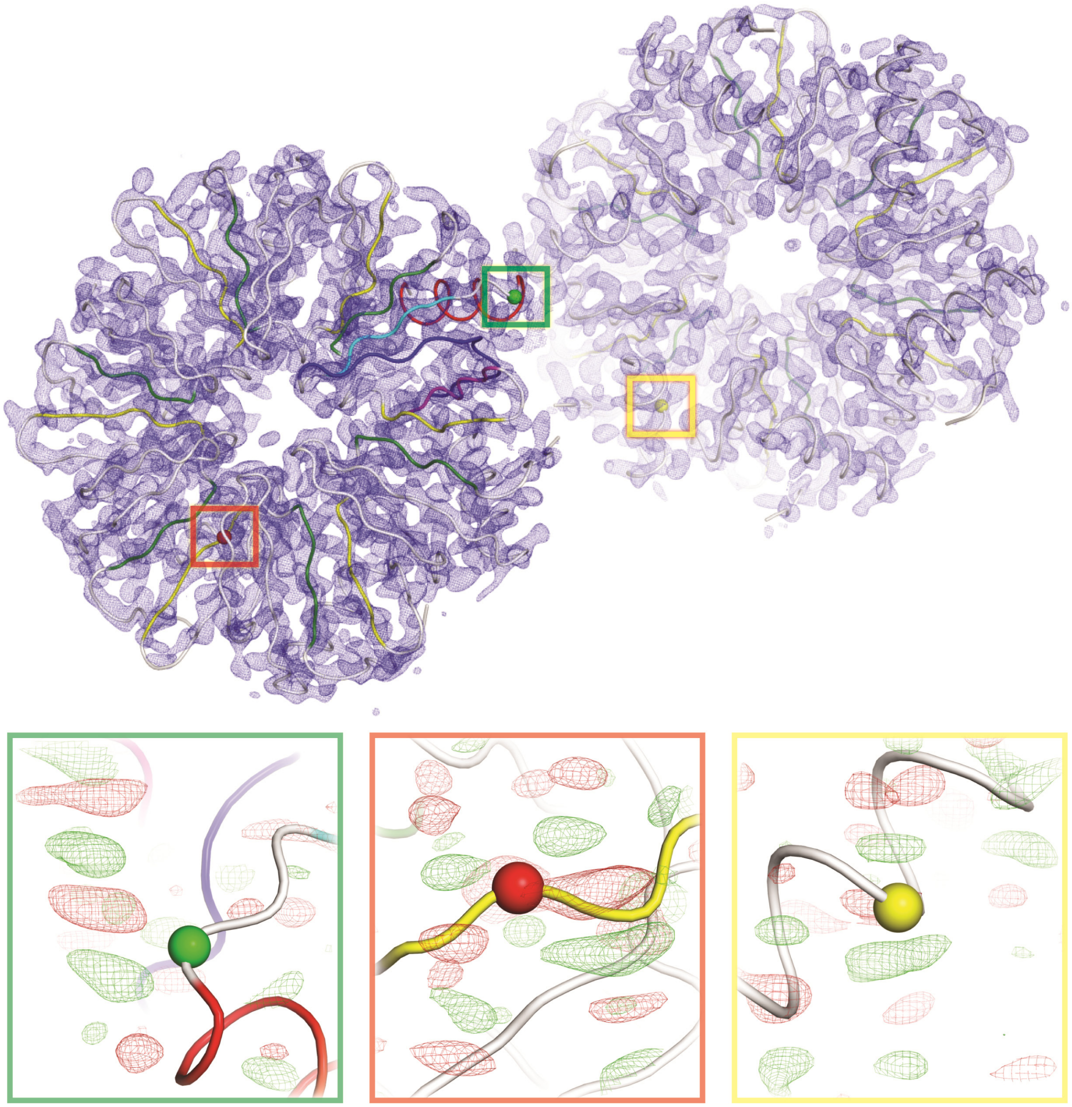


Figure 3. Systematic alternating density, a product of anisotropy. During initial refinement, systematic patches of alternating positive and negative difference density ($F_o - F_c$) was observed throughout the structure. The alternating density was resolved by using the UCLA anisotropy server and resolving the structure.

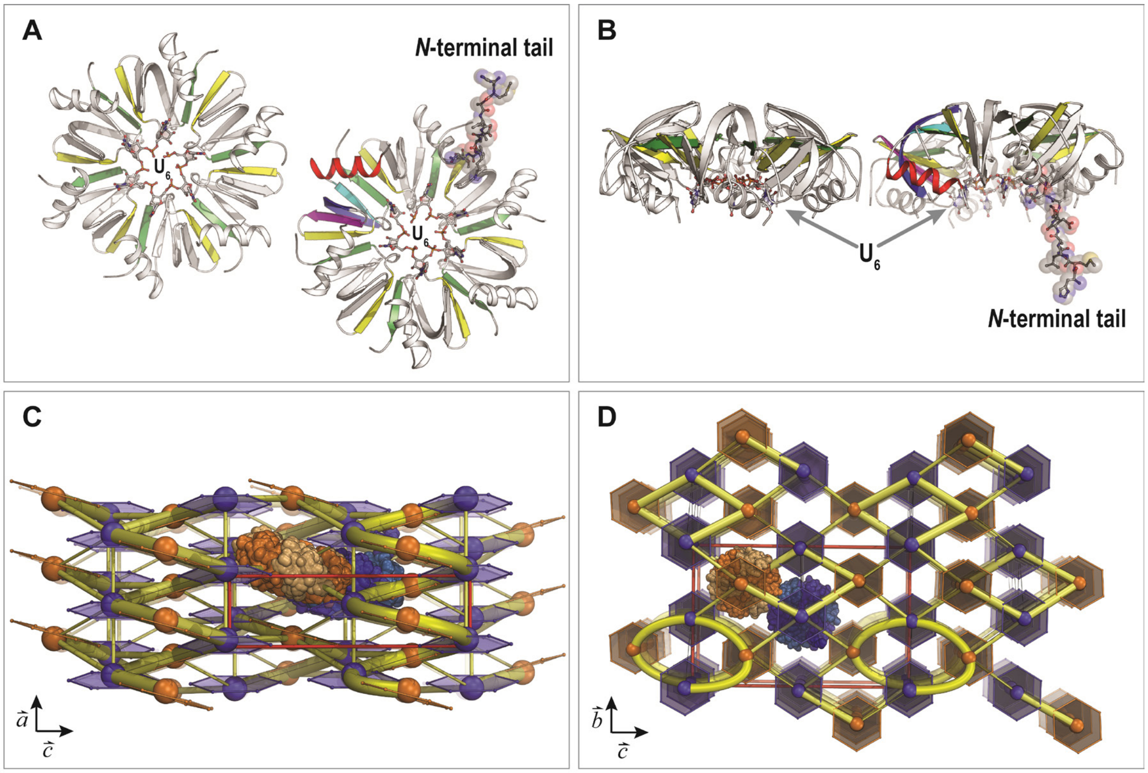


Figure 4. Lattice packing of *Tma* Hfq crystal. (A)(B) *Tma* Hfq crystallizes in the $P2_12_12_1$ space group with two hexamers per ASU. The mobile N-terminal tail contacts the neighboring ASU, allowing its structure to be resolved. (C) Lattice packing viewed down the $\vec{b}$ and (D) $\vec{a}$ axes. *Tma* Hfq packs with few direct crystal contacts between parallel layers in the $\vec{b}$ x $\vec{c}$ plane, instead the majority of the crystal contacts are made between neighbors, which combined with the tilt of the adjacent monomers, spirals down creating multiple ‘springs’ with minimal crystal contacts connecting them.

5 References

1. Altuvia S, Weinstein-Fischer D, Zhang A, Postow L, Storz G. A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. *Cell*. 90, 43–53 (1997).

2. Michaux C, Verneuil N, Hartke A, Giard J-C. Physiological roles of small RNA molecules. *Microbiology (Reading, Engl)*. 160(Pt 6), 1007–1019 (2014).

3. Lybecker M, Bilusic I, Raghavan R. Pervasive transcription: detecting functional RNAs in bacteria. *Transcription*. 5(4), e944039 (2014).

4. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell*. 136(2), 215–233 (2009).

5. Azam TA, Iwata A, Nishimura A, Ueda S, Ishihama A. Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J Bacteriol*. 181, 6361–6370 (1999).

6. Møller T, Franch T, Højrup P, *et al.* Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell*. 9(1), 23–30 (2002).

7. de Fernandez MTF, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qβ-RNA. *Nature*. 219, 588–590 (1968).

8. Arluison V, Derreumaux P, Allemand F, Folichon M, Hajnsdorf E, Régnier P. Structural Modelling of the Sm-like Protein Hfq from Escherichia coli. *J Mol Biol*. 320(4), 705–712 (2002).

9. Zhang A. The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein. *EMBO J*. 17, 6061–6068 (1998).

10. Schumacher MA, Pearson RF, Moller T, Valentin-Hansen P, Brennan RG. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. 21, 3546–3556 (2002).

11. Song T, Mika F, Lindmark B, *et al.* A new Vibrio cholerae sRNA modulates colonization and affects release of outer membrane vesicles. *Mol Microbiol*. 70(1), 100–111 (2008).

12. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. *Int Rev Cell Mol Biol*. 272, 149–189 (2009).

13. Will CL, Lührmann R. Protein functions in pre-mRNA splicing. *Curr Opin Cell Biol*. 9(3), 320–328 (1997).

14. Mura C, Randolph PS, Patterson J, Cozen AE. Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. 10, 636–651 (2013).

15. Moll I, Afonyushkin T, Vytvytska O, Kaberdin VR, Bläsi U. Coincident Hfq binding and RNase E cleavage sites on mRNA and small regulatory RNAs. *RNA*. 9(11), 1308–1314 (2003).

16. Aiba H. Mechanism of RNA silencing by Hfq-binding small RNAs. *Curr Opin Microbiol*. 10, 134–139 (2007).

17. Jain K. A Thermodynamic Perspective of sRNA-mRNA Interactions and the Role of Hfq. In: *ACS Symposium Series*. American Chemical Society, 111–131 (2011).

18. De Lay N, Schu DJ, Gottesman S. Bacterial small RNA-based negative regulation: Hfq and its accomplices. *Journal of Biological Chemistry*. 288(12), 7996–8003 (2013).

19. Bardill JP, Hammer BK. Non-coding sRNAs regulate virulence in the bacterial pathogen Vibrio cholerae. *RNA Biol*. 9(4), 392–401 (2012).

20. Tu KC, Bassler BL. Multiple small RNAs act additively to integrate sensory information and control quorum sensing in Vibrio harveyi. *Genes & Development*. 21(2), 221–233 (2007).

21. . Impact of the RNA chaperone Hfq on the fitness and virulence potential of uropathogenic Escherichia coli. *Infect Immun*. 76(7), 3019–3026 (2008).

22. Storz G, Vogel J, Wassarman KM. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell*. 43(6), 880–891 (2011).

23. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*. 3(12), a003798 (2011).

24. Fröhlich KS, Vogel J. Activation of gene expression by small RNA. *Current opinion in microbiology*. 12(6), 674–682 (2009).

25. Beisel CL, Storz G. Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiol Rev*. 34(5), 866–882 (2010).

26. Sauer E. Structure and RNA-binding properties of the bacterial LSm protein Hfq. *RNA Biol*. 10(4), 610–618 (2013).

27. Wagner EGH. Cycling of RNAs on Hfq. *RNA Biol*. 10(4), 619–626 (2013).

28. Faner MA, Feig AL. Identifying and characterizing Hfq-RNA interactions. *Methods*. 63(2), 144–159 (2013).

29. Panja S, Woodson SA. Hexamer to monomer equilibrium of E. coli Hfq in solution and its impact on RNA annealing. 417, 406–412 (2012).

30. Updegrove TB, Correia JJ, Chen Y, Terry C, Wartell RM. The stoichiometry of the Escherichia coli Hfq protein bound to RNA. *RNA*. 17(3), 489–500 (2011).

31. Wang W, Wang L, Zou Y, *et al.* Cooperation of Escherichia coli Hfq hexamers in DsrA binding. *Genes & Development*. 25(19), 2106–2117 (2011).

32. Arluison V, Mura C, Guzman MR, *et al.* Three-dimensional structures of fibrillar Sm proteins: Hfq and other Sm-like proteins. 356, 86–96 (2006).

33. Mikulecky PJ. Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nature Struct Mol Biol*. 11, 1206–1214 (2004).

34. Link TM, Valentin-Hansen P, Brennan RG. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. *Proc Natl Acad Sci USA*. 106, 19292–19297 (2009).

35. Murina V, Lekontseva N, Nikulin A. Hfq binds ribonucleotides in three different RNA-binding sites. *Acta Crystallogr D Biol Crystallogr*. 69(Pt 8), 1504–1513 (2013).

36. Klock HE, Lesley SA. The Polymerase Incomplete Primer Extension (PIPE) method applied to high-throughput cloning and site-directed mutagenesis. *Methods Mol Biol*. 498, 91–103 (2008).

37. Patterson J, Mura C. Rapid Colorimetric Assays to Qualitatively Distinguish RNA and DNA in Biomolecular Samples. (2013).

38. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 31(13), 3784–3788.

39. Gill SC, Hippel von PH. Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem*. 182(2), 319–326 (1989).

40. McPherson A, Gavira JA. Introduction to protein crystallization. 70, 2–20 (2014).

41. Kabsch W. XDS. *Acta Crystallogr D Biol Crystallogr*. 66(2), 125–132.

42. Adams PD, Afonine PV, Bunkóczi G, *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. 66, 213–221 (2010).

43. Matthews BW. Solvent content of protein crystals. *J Mol Biol*. 33(2), 491–497 (1968).

44. Tong L, Rossmann MG. Rotation function calculations with GLRF program. 276, 594–611 (1997).

45. Winn MD, Ballard CC, Cowtan KD, *et al.* Overview of the CCP4 suite and current developments. 67, 235–242 (2011).

46. Latif H, Lerman JA, Portnoy VA, *et al.* The genome organization of Thermotoga maritima reflects its lifestyle. *PLoS Genet*. 9(4), e1003485 (2013).

47. Nelson KE, Clayton RA, Gill SR, *et al.* Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature*. 399(6734), 323–329 (1999).

48. Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25(17), 3389–3402 (1997).

49. Deacon AM, Godzik A, Lesley SA, Wooley J, Wüthrich K, Wilson IA. The JCSG high-throughput structural biology pipeline. *Acta Crystallogr Sect F Struct Biol Cryst Commun*. 66(Pt 10), 1137–1142 (2010).

50. Sun X, Wartell RM. Escherichia coli Hfq binds A18 and DsrA domain II with similar 2:1 Hfq6/RNA stoichiometry using different surface sites. 45, 4875–4887 (2006).

51. Mura C, Kozhukhovsky A, Gingery M, Phillips M, Eisenberg D. The oligomerization and ligand-binding properties of Sm-like archaeal proteins (SmAPs). *Protein Sci*. 12(4), 832–847 (2003).

52. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D. Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. 103, 8060–8065 (2006).

Chapter 4 RNA binding, structure and self-assembly of *Thermotoga maritima* Hfq: A regulatory mechanism involving two distinct oligomeric states?

Peter S Randolph[a], Jennifer Patterson[a], Shugeng Cao[2], Dan Fox[1], Jon Clardy[2], and Cameron Mura

[1]University of Virginia, Department of Chemistry, Charlottesville, VA 22904 USA

[2]Harvard University, School of Medicine, Department of Biological Chemistry & Molecular Pharmacology, Boston, MA 02115 USA

[a] PSR and JP are equally contributing authors

1 Abstract

The bacterial Sm protein, host factor for Qβ (Hfq), plays a vital role in regulation of bacterial gene expression *via* its function as an RNA chaperone. Post-transcriptional, RNA-based regulatory pathways enable bacteria to rapidly adapt to fluctuations in the bacterial micro-environment. Hfq regulates polyadenylation, translation, and degradation of RNA *in vivo* through direct interactions with messenger RNA and/or small regulatory RNA. In this study, we found that the putative Hfq homolog from the thermophillic *Thermotoga maritima* (*Tma*) heterologously co-purifies with U/C-rich nanoRNAs, binding with a nanomolar affinity. Identified nanoRNA sequences all contain a 5' monophosphate and a 3' hydroxyl and compete with U-rich sequences for the proximal face of Hfq. Data suggests that the position of cytosine within the sequence, rather than the absolute number of

cytosines is the key factor in determining affinity. The crystal structure shows that, even under denaturing condition, a small amount of the heterologous nanoRNA remains bound. *Tma* Hfq forms a hexamer within the crystal, agreeing with previous studies on the functional form of Hfq in *Escherichia coli* (*Eco*) and other bacterial species. However, our studies of *Tma* Hfq suggest that an equilibrium exists between a homo-hexamer and a homo-dodecamer. Both oligomeric states are capable of binding poly-adenine and poly-uracil RNA with low nanomolar binding affinities, with poly-A and poly-U RNA preferentially binding to the dodecamer and hexamer, respectively. This leads to a shift in equilibrium between the states; poly-U shifting the equilibrium toward the hexameric state, and poly-A having no effect.

2 Introduction

Bacteria utilize complicated RNA-based regulatory pathways to allow for rapid adaptation to their changing micro-environment. The level of protein production is not static, but is highly regulated by the rate of translation and the availability of the mRNA transcript, which is dependent on both transcription and degradation. Many small, non-coding RNAs (sRNAs) combined with the *chaperone* Hfq add a level of post-transcriptional regulation to gene expression. Hfq was first identified as a host factor essential for the replication of RNA phage Qβ in *Escherichia coli* (*Eco*) [1]. Hfq has a flexible sequence recognition capacity; helping diverse *trans*-sRNAs bind their target mRNAs, either up or down-regulating translation. This allows bacteria to rapidly respond to fluctuations in environmental conditions including temperature, oxidative stresses, local cell density (quorum sensing), *etc* and pathogenic functions including bacterial virulence, and biofilm formation [2,3]. The central regulatory role of Hfq in numerous cellular pathways results in a pleiotropic phenotype when Hfq is disrupted in *E. coli,* by either mutation or by knockout. These phenotypes include increased sensitivity to low pH, UV light and high temperatures, as well as a decreased growth rate [4]. A common example of Hfq's regulatory role is the sRNA-regulated translation of the RNA polymerase subunit σ^s. A factor in environmental stress response, σ^s requires

Hfq to facilitate the annealing of the *trans*-encoded regulatory sRNAs (DsrA, RprA, or ArcZ) to the target mRNA transcript (*rpoS*) [5-9]. In addition to its chaperone role, Hfq has also been shown to play an important role in RNA stability [10,11]. Specifically, Hfq can impede RNase E activity in the absence of sRNA by binding to a site near the RNase E cleavage site; when the sRNA is present Hfq facilitates annealing allowing for degradation by nucleases within the double stranded region (i.e. RNase III) or promote cleavage by RNase upstream of the regulatory region [12,13]. The precise molecular mechanism underlying RNA stabilization/destabilization and mediation and modulation of RNA···RNA interactions by Hfq, remains a challenging question. There exist a few hypothetical models, but due to gaps in our knowledge a single model has yet to be generally accepted for the molecular basis of Hfq activity [14].

In order to dissect the molecular mechanism of Hfq activity, we must understand the structure and dynamics of Hfq. Hfq belongs to the Sm protein superfamily, which is ubiquitous across all three domains of life and considered to be one of the earliest evolving proteins [15]. The Sm fold consists of a small, highly-bent, five-stranded β-barrel preceded by a small N-terminal α-helix. β-strand 4 (β4) and β-strand 5 (β5) form the interface between adjacent subunits in the protein oligomer, contributing to its overall thermostability [16,17]. Structural data have shown that Hfq self-assembles into stable homo-hexameric toroid structures (~70 Å in diameter, with a ~10 Å pore) in the absence of RNA, unlike the eukaryal canonical Sm core proteins, which assemble in a stepwise fashion around a template RNA [16-18]. The α-helixes lie on one side of the ring referred to as the Loop-3 (L3) or *proximal* face, whereas the opposite side is referred to as the Loop-4 (L4) or *distal* face. Both faces bind RNA and display distinct sequence specificity, the L3 face primarily binding uridine-rich RNA and the L4 face primarily binding adenine-rich RNA. An additional binding site on the *lateral* face (toroid edge), has been confirmed to bind single nucleotides [19]. The homo-hexamer model has been identified as the functional form of Hfq in *E. coli* and other bacterial species using multiple physical techniques (analytical ultracentrifugation, gel filtration chromatography, and electron microscopy) [14,20,21]. Though some studies have suggested the occurrence of higher-order oligomeric states, the potential functional roles of such states remain elusive [18,22,23].

In this study, we identified nanoRNAs heterologously co-purified with *Thermotoga maritime* (*Tma*) and determined their binding affinities. We also solved the *Tma* Hfq crystal structure bound with co-purified r(U)$_6$ RNA, and characterize the self-assembly of *Tma* Hfq. We identified U/C-rich nanoRNAs that co-purify with *Tma* Hfq when heterologously expressed in *E. coli,* by liquid chromatography – tandem mass spectrometry (LC-MS/MS). We demonstrate that these nanoRNAs interact with Hfq with nanomolar affinities that vary based on cytosine substitution into the pentauridine or hexauridine sequences. Denaturing scheme purification removes most of the co-purified nano RNAs, but the crystal structure shows a small amount of r(U)$_6$ remains bound on the *L3* face. Two *Tma* Hfq···r(U)$_6$ structures are presented; a lower resolution structure with no additional RNA added and a higher resolution structure co-crystallized with sub-stoichiometric amounts of r(U)$_5$ RNA. The crystalline *Tma* Hfq exists as a homo-hexamer, but in solution we established that a dodecamer is in equilibrium with the hexameric state through a combination of biophysical techniques including chemical cross-linking, analytical size exclusion chromatography (AnSEC), ‘semi-native’ Western blots, and isothermal titration calorimetry (ITC). We were able to separate these two states with a combination of cross-linking and AnSEC. In order to assess the RNA-binding properties of these two oligomeric states, a series of binding assays were conducted, measuring the ability of *Tma* Hfq to bind both 5-Carboxyfluorescein labeled poly-adenosine (FAM-r(A)$_{18}$) and poly-uracil (FAM-r(U)$_6$) RNAs. In contrast to previous studies with Hfq from other bacterial species, we find that both the hexameric and dodecameric states of *Tma* Hfq are able to interact with U-rich and A-rich RNA. Interestingly, the presence of r(U)$_6$ shifts the equilibrium between the hexamer and dodecamer toward the hexameric state, whereas r(A)$_{18}$ RNA does not alter the equilibrium.

3 Results

3.1 *Tma* Hfq Co-purifies with Nucleic Acid

Throughout protein purification, spectrophotometric readings indicated a 260nm/280nm ratio of approximately 1.80. To assess whether *Tma* Hfq was co-purifying with nucleic acid the sample was separated on an anion exchange column to remove components with a large negative charge. The chromatograph (Figure 6C), contained three 'peaks'. The 260nm/280nm ratios of these three are approximately 0.9-1.7, 1.7-1.9, and 2.0 in order of elution.

Fractions from the first and second elution peak were analyzed by MALDI-MS, which contained a peak corresponding to the +1 and +2 charge state of *Tma* Hfq. The third elution peak did not contain protein within the detection limit of MALDI-MS.

The co-purifying nanoRNAs (CPNs) from peak 1 were isolated by phenol-chloroform extraction and concentrated by ethanol precipitation. CPNs were subjected to a series of colorimetric assays to determine the sugar component [24]. The CPNs did not produce a colored product for the Benedict's assay (free or reducing sugar) or the Dische diphenylamine assay (deoxyribose), (Figure 7A, 7C), whereas a notable blue product was observed for the Bial's orcinol assay (pentose sugar), (Figure 7B).

CPNs were further characterized by ^{1}H and ^{31}P NMR spectroscopy. The ^{1}H spectrum, (Figure 8A), contained peaks corresponding to both the sugar and base protons present in a nucleotide, but a specific base was not apparent. The ^{31}P spectrum (Figure 8B) contained three peaks corresponding to a phosphodiester at ~0ppm, phosphate monoester at ~4 ppm, and phosphonate at ~21 ppm.

3.2 Identification of Hfq-binding nanoRNAs (Shugeng: HPLC and LC-MS/MS)

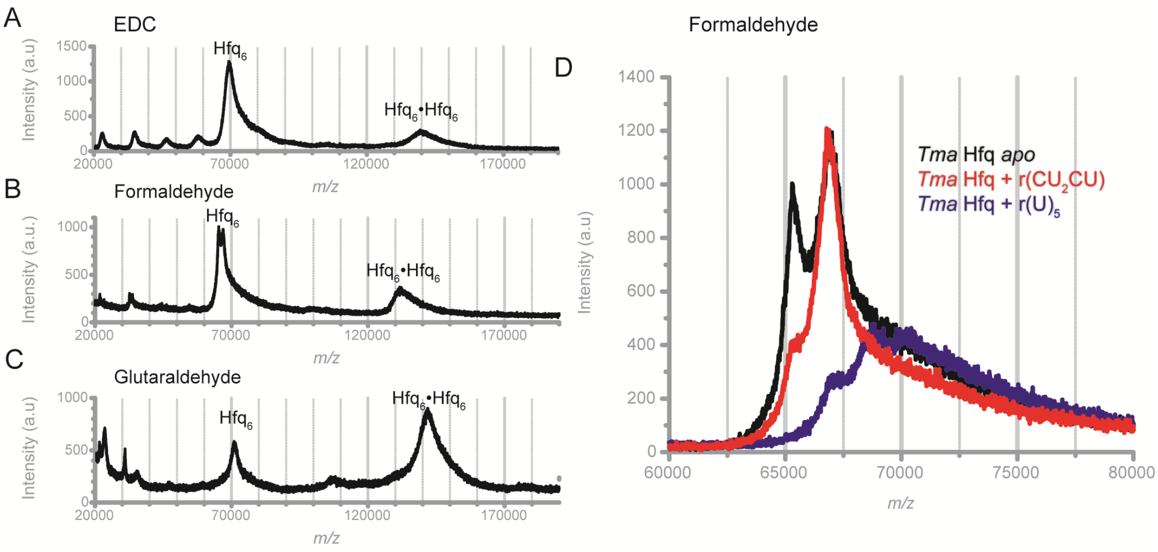


Figure 1. MALDI-TOF MS spectra of crosslinked *Tma* Hfq. Two distinct oligomeric states, hexamer and dodecamer, were observed for all three cross-linkers: A) 1-ethyl-3-[3-dimethylamino-propyl] carbodiimide hydrochloride (EDC), B) formaldehyde, and C) glutaraldehyde; formaldehyde shows a split peak. D) **Black** spectrum is of *Tma* Hfq purified under non-denaturing conditions. The peak around 65 kDa is consistent with the expected molecular weight for a *Tma* Hfq hexamer, whereas the peak around 67 kDa is not consistent with an oligomeric state of *Tma* Hfq. Hfq was spiked with r(CU_2CU) (**red**), indicated a 1:1 stoichiometry between *Tma* Hfq and r(CU_2CU). Hfq spiked with r$(U)_5$ (**blue**), suggests that stoichiometry of Hfq to r$(U)_5$ can be either 1:1, 1:2, or 1:3. The shift in molecular weight upon addition of isolated binding partner suggests that the peak at 67 kDa is due to crosslinking *Tma* Hfq to oligonucleotides that co-purify with the protein.

The different components of the binding partner sample were separated by HPLC; the sequences of six of the components were determined by LC-MS/MS, (Table 1). All six were 5 to 6 nts in length with a 5’ monophosphate and 3’ hydroxyl. Identified nanoRNAs were synthesized and purified by integrated DNA Technologies (IDT) for further studies.

3.3 Binding Partners Interact with the Hfq Hexamer

Hfq was crosslinked with formaldehyde indirectly via vapor diffusion [25]. The MALDI TOF MS spectrum of crosslinked *Tma* Hfq, (Figure 1C, D), contained two peaks not baseline resolved in the 65 kDa to 67 kDa region of the MS spectrum. The peak around 65 kDa is consistent with the expected molecular weight for a *Tma* Hfq hexamer, whereas the peak around 67 kDa is not consistent with an oligomeric state of *Tma* Hfq.

Tma Hfq spiked with fraction 13 (r(CU_2CU)) and 23 (r(U_5)) from the HPLC purification of the isolated binding partners were crosslinked with formaldehyde. The relative intensity of the peak at 67 kDa was increased in the spectrum of *Tma* Hfq with 13, (Figure 1D). The spectrum of *Tma* Hfq with r(U_5) had 3 peaks that are not baseline resolved in the 66kDa to 77kDa region of the spectrum (Figure 1D).

3.4 Hfq Binds nanoRNAs with Nanomolar Affinity

Binding affinity of FAM-r$(U)_6$ and *Tma* Hfq was determined by measuring fluorescence polarization with varying *Tma* Hfq concentrations (Figure 9B), resulting in a dissociation constant (K_d) of 9.64 n*M* (N=4) after fitting the data to the Boltzmann equation. The concentration of the Hfq-binding nanoRNAs required to displace 50% of FAM-r$(U)_6$ (IC50) (Table 1), was determined by measuring fluorescence polarization at varying nanoRNA concentrations, (Figure 9A). Inhibition constant (K_i) for each nanoRNA (Table 1), were calculated using:

$$Ki = \frac{IC50}{1+\frac{[S]}{Kd}} \quad (2)$$

Table 1. ***Tma*** **Hfq co-purified oligonucleotide properties**

Oligo Name	Sequence	Chromatograph Fraction	MW (g/mol)	*N*	IC50 (nM)	K_i[c]
r(U)$_5$	5'Phos-rUrUrUrUrU	f23	1548.9	4	6 ± 1	4.0 ± 0.7
r(U$_3$CU)	5'Phos-rUrUrUrCrU	f17	1547.9	4	13 ± 1	8.8 ± 0.8
r(CU$_2$CU)	5'Phos-rCrUrUrCrU	f13	1546.9	4	24 ± 1	16.0 ± 0.7
r(U$_4$C)	5'Phos-rUrUrUrUrC	f19	1547.9	4	25 ± 1	16.6 ± 0.7
r(CU$_2$CU$_2$)	5'Phos-rCrUrUrCrUrU	f21	1853.1	5[b]	4 ± 1	2.7 ± 0.7
r(U$_4$CU)	5'Phos-rUrUrUrUrCrU	f28	154.0	4	28 ± 1	18.5 ± 0.8

[a] Oligonucledotides identified via LC-MS/MS and their binding affinity probed using a competitive polarization assay utilizing FAM-r(U)$_6$ as labeled ligand

[b] Data points at 10 µM and 20 µM were collected in 3 replicates

[c] The K_i was calculated using the K_d for FAM-r(U)$_6$, 9.64 under experimental conditions (*N*=4)

where [S] is the concentration of FAM-r(U)$_6$ and K$_d$ is the binding constant for FAM-r(U)$_6$ under experimental conditions.

3.5 *Tma* Hfq Assembles into Two Distinct Oligomeric States

Tma Hfq purity was assessed pure by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE, Figure 6A) and MALDI-TOF MS (Figure 6B), which revealed two peaks corresponding to the +1 and +2 charge states of *Tma* Hfq. Crosslinking (EDC, formaldehyde, and glutaraldehyde) followed by MALDI-TOF MS gave mass spectra containing two predominant peaks in the 20,000 to 200,000 mass-to-charge (m/z) ratio range (Figure 1A, B, C), roughly correspond to the expected molecular weight for *Tma* Hfq hexamer (64.8 kDa) and dodecamer (129.6 kDa). Specific molecular weights and oligomeric states measured by MALDI-TOF MS for each cross-linking reagent are summarized in (Table 3).

Additional peaks were observed corresponding to the +2 charge state of the protein hexamer and intermediate oligomeric states; presumably due to incomplete cross-linking of the sample (Figure 1A, B, C). A secondary peak at a molecular weight approximately 1500 Da greater than the Hfq hexamer peak was found in the formaldehyde treated sample (Figure 1B, 1D) and was confirmed to be the previously mentioned nanoRNAs.

Analytical size exclusion (AnSEC) was used to determine molecular weights of the *Tma* Hfq oligomers present in solution. Based on a standard curve, the molecular weight of *Tma* Hfq in solution is approximately 98.4 kDa (Figure 2A, blue), corresponding to a nonamer. To resolve whether the discrepancy between oligomeric states determined by AnSEC and by cross-linking/MALDI-TOF MS is due to the hydrodynamic envelope of the Hfq hexamer, versus a relatively rapid equilibrium between the hexamer and dodecamer, a EDC-crosslinked *Tma* Hfq sample was also run on AnSEC. The crosslinked sample (Figure 2A, red) exhibits two peaks corresponding to molecular weights of approximately 72.8 kDa and 187.9 kDa, which can be

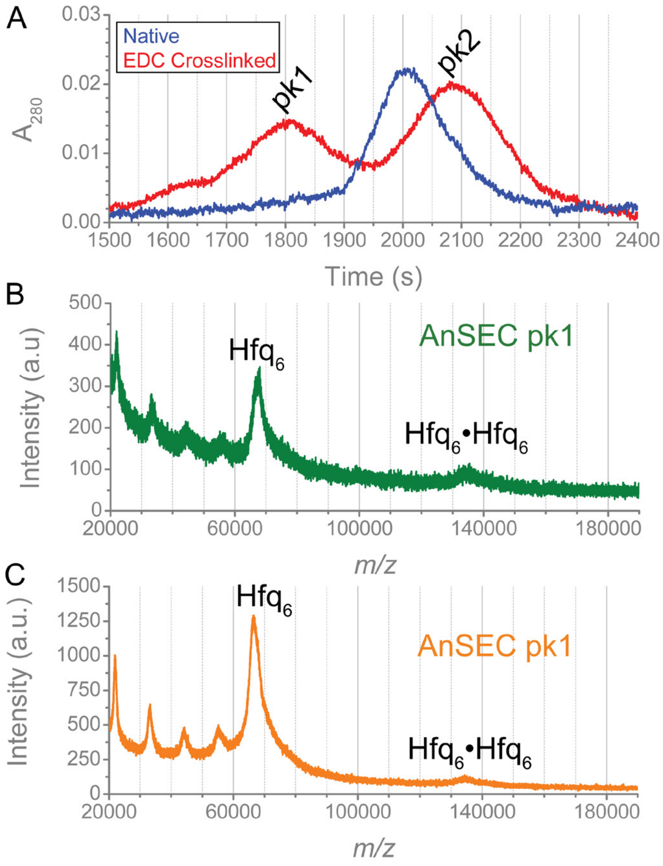


Figure 2. Chromatogram of AnSEC separation of native and crosslinked *Tma* Hfq. A) Chromatogram of the native sample contained a single peak corresponding to a 9-mer (9.11 ± 0.07 subunits) of *Tma* Hfq, whereas the crosslinked sample contained two peaks corresponding an Hfq hexamer (6.75 ± 0.09 subunits) and ~17-mer (17.4 ±0.6 subunits). The presence of a single peak with an intermediate retention time in the native sample suggests equilibrium exists between the two oligomeric states. The predominant peaks observed in the MALDI-TOF spectrum of recombinant *Tma* Hfq from B) peak 1 and C) peak 2 of the AnSEC separation can be contributed to the hexameric form of *Tma* Hfq. A significantly smaller peak corresponding to the *Tma* Hfq dodecamer is observed in both spectra. Additional peaks are observed below 60 kDa that correspond to lower level oligomeric states of Hfq that are presumably due to incomplete cross-linking. No peak was observed around 183.6 kDa, which would correspond to a *Tma* Hfq 17-mer.

attributed to a *Tma* Hfq hexamer (AnSEC pk2) and 17-mer (AnSEC pk1). MALDI-TOF MS spectra for AnSEC pk2 contained a peak corresponding to the Hfq hexamer along with a significantly smaller peak corresponding to the expected molecular weight of an Hfq dodecamer; lower molecular weight peaks were also observed suggesting that cross-linking was not complete (Figure 2C). MALDI-TOF MS spectra for AnSEC pk1 (Figure 2B) contained two peaks corresponding to the Hfq hexamer and dodecamer, but no peak corresponding to a 17-mer was observed.

3.6 Structure Determination and Analysis

Crystals were grown in two reproducible crystal conditions: 150 m*M* tri-potassium citrate and 30% w/v PEG-3350 with either 1.0 *M* glycine or 0.1 *M* sarcosine as an additive. Crystals were isomorphous, belonging to space group $P2_12_12_1$ with cell dimensions *a* = 39 Å, *b* = 133 Å, *c* = 206 Å (±1%). After repeated trials, solvable datasets were collected with a significant amount of anisotropy. These were initially solved with anisotropy correction from Phenix, but systematic alternating density was observed in the structure during refinement which was rectified using the UCLA Anisotropy Server [26], and the structure was refined without issue (See Chapter 3). Crystals were co-crystallized with $r(U)_5$ resulting in significantly improved diffraction from 2.65 to 2.1 Å. Both *Tma* Hfq structure were solved using the same methodology (See Chapter 3 for methodology) (For diffraction and refinement statistics see Table 4). *Tma* Hfq packed as two homo-hexamer rings per asymmetric unit (ASU) with contacts between the lateral faces. The hexameric ring exhibits a diameter of ~70 Å with a ~10 Å pore, consistent with other Hfq structures [16,17,27] (Figure 4A). *Tma* Hfq monomer contains an N-terminal α-helix followed by a β-sheet with the standard β5-β1-β2-β3-β4 topology of the Sm fold (Figure 4A, 4B). One of twelve monomers in the ASU contains a structured *N*-terminal chain extending perpendicular to the toroid plane stretching to a hexamer in the adjacent ASU (Figure 3A). Density on the L3 face was identified as a strand of $r(U)_6$ bound around the pore; crystal averaging creating a seemingly continuous $r(U)_6$ ring (Figure 4B). The uridines are bound in the known Hfq uridine-binding pockets, with a 1-to-1 ratio with

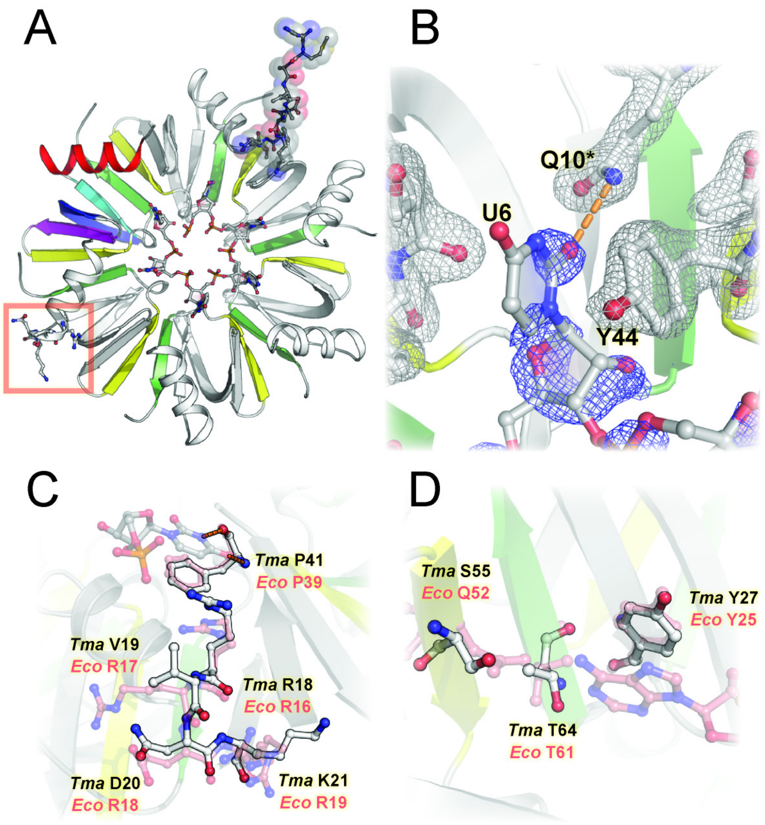


Figure 3. *Tma* Hfq co-crystallizes with a small $r(U)_6$ RNA strand and has an ordered *N*-terminal tail. (A) One of the twelve monomers in the ASU has an ordered *N*-terminal tail which extends to an adjacent ASU. (B) $r(U)_6$ binds in the known U-rich RNA binding pocket on the *L3* face. Tyr-44 base stacks with the uridine base, while glutamine-10 from the adjacent monomers forms a hydrogen bond which is known to be key for specificity. Electron density maps are $2mF_o$-DFc shown at 1.2σ, with the *Tma* Hfq protein map colored gray and the bound $r(U)_6$ map in blue. (C) Lateral rim binding site of *E. coli* Hfq consists of an aromatic residue near the L3 face (Phe-39) and an arginine patch farther down the rim. Phe-39 (*Tma* Hfq Phe-41) and Arg-16 (*Tma* Hfq Arg-18) are conserved in *Tma* Hfq. (D) The conserved A-binding site (R-site in Gram-negative bacteria, A-site in Gram-positive bacteria) is conserved in *Tma* Hfq.

Tma Hfq monomers. Uridine bases are π-stacked with tyrosine 44 (Y44) and hydrogen-bonded to glutamine 10 (Q10) of the previous monomer (Figure 4C). Occupancy of the uridines varies between 0.54 and 0.65, suggesting only some of the *Tma* Hfqs contain bound r(U)$_6$ (and interestingly occupancies are slightly lower for the higher resolution r(U)$_5$-additive structure, ranging from 0.48 – 0.63). The conserved aromatics (Tyr-25 and Phe-39 in *E. coli*) responsible for A-rich RNA binding and the lateral rim binding are conserved in *Tma* Hfq (Tyr-27 and Phe-41) (Figure 4D), though *Tma* Hfq lacks some of the positive arginine patch on the lateral rim.

3.7 Hfq is in Equilibrium Between Hexamer and Dodecamer States

To further elucidate the equilibrium between the hexamer and dodecamer states of Hfq, oligomer dissociation upon dilution was monitored by ITC (Figure 4A, B). As the concentration of *Tma* Hfq increases in the sample cell, the thermal heat released upon injection of *Tma* Hfq decreased until the sample cell reached a concentration of *Tma* Hfq at which the dodecamer no longer dissociated into hexamer (Figure 4A, B); therefore, the forward direction of the hexamer to dodecamer transition is monitored. ITC experiments (N=4) indicated that the association constant (K_a) and the enthalpy of association (ΔH_a) are 0.5 ± 0.1 μ*M* and 36.7 ±0.6 kcal/mol at 25°C, respectively. Gibbs free energy change was calculated to be −7.8 ± 0.1 kcal/mol (N=4) for the association of hexamers to dodecamer, indicating the transition is favorable.

To verify that the oligomeric state transition being monitored by ITC (in the range of ~183 n*M* to 4 μ*M* *Tma* Hfq) is genuinely a dodecamer to hexamer transition, we performed semi-native Western blots, which indicated the two predominant oligomeric states present in the concentration range (20 μ*M* to 78 n*M* *Tma* Hfq) are, as predicted, hexamer and dodecamer (Figure 4C, 4D, 4E). The absence of *Tma* Hfq monomer in this concentration range was not due to a preference of the rabbit anti-*Tma* Hfq pAb for the hexamer. Fluorescence signal intensity of each band was used to

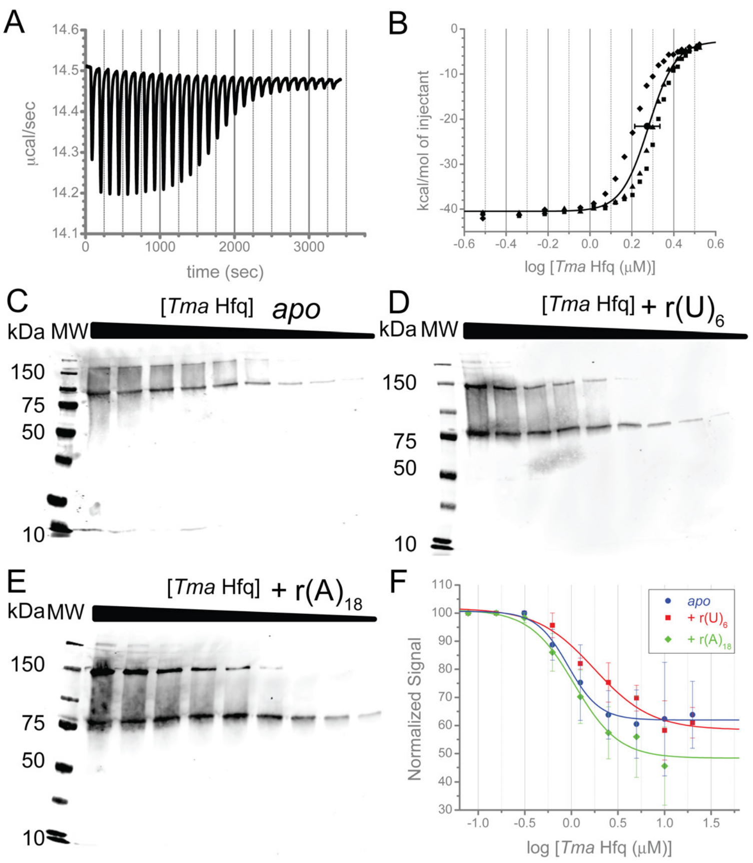


Figure 4. Representative isothermal titration calorimetry (ITC) data of *Tma* Hfq dissociation (dodecamer to hexamer) (A) Representative thermograph of thermal power (µcal/sec) as a function of time (sec); peaks are at constant intervals correspond to the injection of *Tma* Hfq (titrant) into dialysis buffer (sample cell). (B) Area under each peak is integrated then normalized to the moles of injected *Tma* Hfq (kcal/mol of injectant), and plotted against the concentration of *Tma* Hfq. Data was fit to a modified Hill's equation and the K_a and ΔH were determined to be 0.5 ± 0.1µM and 36.7 ± 0.6 kcal/mol (N=4), respectively. Semi-native Western blots of *Tma* Hfq: (C) Semi-native Western blot of (D) *Tma* Hfq *apo*, (E) *Tma* Hfq with r(U)$_6$, and E) *Tma* Hfq with r(A)$_{18}$. F) Flourescence was plotted and fit, determing K$_d$ for each trial.

Table 2. Summary of apparent thermodynamic parameters for dodecamer-hexamer equilibria. Calculated by ITC dilution experiments and semi-native western blot analysis. Dissociation constants (K_d) determined for FAm-r(U)$_6$ and FAM-r(A)$_{18}$ with difference oligomeric states of *Tma* Hfq

Equilibrium	**K_d*** **(µM)**	**ΔG_d** **(kcal/mol)**	**ΔH_d** **(kcal/mol)**	**ΔS_d** **(cal/mol·K)**
$Hfq_6 \rightleftharpoons Hfq_6::Hfq_6$	1.9 ± 0.3	7.81 ± 0.08	−36.9 ± 0.1	−150.1 ± 0.7
$Hfq_6::Hfq_6 \rightleftharpoons Hfq_6$	1.0 ± 0.2	–	–	–
$Hfq_6::Hfq_6 + U_6 \rightleftharpoons$ $Hfq_6 + U_6$	1.8 ± 0.4	–	–	–
$Hfq_6::Hfq_6 + A_{18} \rightleftharpoons$ $Hfq_6 + A_{18}$	1.0 ± 0.2	–	–	–

RNA Affinity	**Sample Description**	**FAM-r(U)$_6$** ***K_d*** *(nM)***	**FAM-r(A)$_{18}$** ***K_d*** *(nM)***
Native	Equilibrium	61 ± 10	236 ± 38
EDC	'static'	116 ± 29	231 ± 30
AnSEC pk1	Dodecamer	79 ± 10	163 ± 24
AnSEC pk2	Hexamer	57 ± 23	333 ± 31

* K_d values calculated using Hfq monomer concentrations

estimate the fraction of hexamer and dodecamer at each concentration, generating binding curves approximating the K_a (Figure 4F) (Table 2).

3.8 Both Oligomeric States have Nanomolar Affinities for Poly-A and Poly-U RNA

RNA-binding properties of the hexamer and dodecamer states of *Tma* Hfq were monitored by fluorescence polarization assays using FAM-r(A)$_{18}$ and FAM-r(U)$_6$ to explore the *L4* and *L3* binding faces respectively. *Tma* Hfq states used in these experiments include: *(i)* native *Tma* Hfq ("native"), *(ii)* EDC-crosslinked *Tma* Hfq EDC ("EDC"), *(iii)* EDC-crosslinked dodecamer ("AnSEC pk1") and *(iv)* EDC-crosslinked hexamer ("AnSEC pk2"). Fluorescence polarization data from four independent experiments were averaged and plotted against Hfq monomer concentration yielding dissociation constants (K_d) for each state of *Tma* Hfq with FAM-r(A)$_{18}$ and FAM-r(U)$_6$ (Table 2). K_d for FAM-r(A)$_{18}$ was approximately equivalent for the native and EDC states of *Tma* Hfq, whereas the K_d for AnSEC pk1 and AnSEC pk2 are significantly different ($P<0.1$, based on unpaired t-test, not assuming the same standard deviation [56]).

Semi-native Western blots of *Tma* Hfq in the presence of stoichiometric equivalents of each U-rich and A-rich RNAs (1:1 Hfq to RNA) revealed that in the presence of r(A)$_{18}$ the hexamer to dodecamer K$_d$ was unaffected, whereas in the presence of r(U)$_6$ the K_d increased by roughly two-fold (from 1.0 ± 0.2 μM to 1.8 ± 0.4 μM) (Figure 4C, D, E, F, Table 2).

4 Discussion

The preference of Hfq to bind of A/U rich RNA sequences is well documented in previous studies [28-31]. Crystal structures have indicated that A-rich and U-rich RNA sequences bind on opposite sides of the Hfq hexamer referred to as the *L4* and *L3* faces, respectively [18,32]. In this study, we identified U/C-rich nanoRNA sequences of 5 to 6 nucleotides, which co-purify with *Tma* Hfq when heterologously expressed in *E. coli*. The crystal structure of *Tma* Hfq has shown that these nanoRNAs do not completely dissociate during denaturing conditions (due to a combination of high stability of the *Tma* Hfq hexamer and high affinity of *Tma* Hfq to its RNA targets) and bind in the known uridine-binding pocket. While *Tma* Hfq crystallizes as a hexamer, in solution there is equilibrium between a hexamer and a higher-ordered dodecamer. The hexamer and dodecamer both bind A-rich and U-rich RNA, with dodecamer formation lowering the K_d of A-rich strands, and U-rich RNA binding shifting the equilibrium to hexamer.

The overwhelming majority of *in vitro* biochemical studies have utilized only a few homologs of Hfq, mainly *E. coli* and other mesophilic γ-proteobacteria,; and data from other Hfq systems would allow comparative analysis that would broaden our understanding of this RNA regulatory system, and better understanding of the mechanism by which Hfq facilitates RNA···RNA interactions [19,33]. To examine the evolutionary conservation and for comparative analysis of Hfq binding, we decided to look into the structure and binding of the deep branching hyperthermophilic *Thermotoga maritima* (*Tma*) Hfq. The thermophilic *Tma* ecosystem (~90ºC) could necessitate tighter binding and greater thermostability of the hexamer. During initial purification (including a heat cut step at 80º C) a high A260nm/A280nm ratio during spectrophotometric measurement of the protein sample indicated *Tma* Hfq was purifying with nucleic acid. The measured A260nm/A280nm ratio throughout purification was approximately 1.80; accepted A260nm/A280nm ratios for "pure" DNA, RNA, and protein are 1.8, 2, and 0.57, respectively [34]. A series of colorimetric chemical reactions determined that *Tma* Hfq was, in all likelihood, co-purifiying with RNA. The predominant phosphodiester peak at 0 ppm in the ^{31}P NMR spectrum (Figure 8A) indicated oligonucleotides and the ^{1}H NMR spectrum (Figure 8B) verified the presence of a nucleotide in the sample, but could not conclusively identify the base. Peak broadness in the ^{1}H NMR spectrum indicated the sample contained multiple components, which was further supported by thin layer chromatography and MALDI-MS (data not shown). Binding partners were separated by HPLC and sequences were determined by LC-MS/MS; six nanoRNAs identified are listed in Table 1. All six oligonucleotides have a 5' monophosphate 3' hydroxyl groups suggesting that this end chemistry is favorable for Hfq binding. Further studies need to be performed to determine the effect of the other end chemistry (i.e. 5' hydroxyl) on the affinity of nanoRNAs for *Tma* Hfq.

The two peaks around 65 kDa and 67 kDa of the MALDI-MS spectrum of formaldehyde crosslinked *Tma* Hfq, black spectrum in Figure 1D, are consistent with Hfq in its hexamer form in the presence and absence of nucleic acid. To verify that the binding partners being isolated by HPLC account for the peak at 67 kDa, the protein sample purified under the denaturing scheme was spiked with the isolated co-purified nanoRNAs separated into HPLC fractions. The relative intensity of the peak at 67 kDa was increased in the spectrum of *Tma* Hfq with r(CU_2CU) (Figure 1D). The single peak generated by addition of r(CU_2CU) indicates a 1:1 stoichiometry between (*Tma* Hfq)$_6$ and r(CU_2CU). Similar results were observed for r(U_3CU), r(U_4C), r(CU_2CU_2), and r(U_4CU) (data not shown). Interestingly, the spectrum of *Tma* Hfq with f23, corresponding to r(U)$_5$ (Figure 1D), had 3 non-baseline resolved peaks in the 66kDa to 77kDa region of the spectrum, suggesting a higher stoichiometry for (*Tma* Hfq)$_6$:r(U)$_5$: 1:1, 1:2, or 1:3. The higher stoichiometry would be possible if multiple U-rich RNA strands could 'share' the *L3* binding pocket, with each having only 1 or 2 nucleotides bound; or the result of additional binding sites on the lateral surface, which have previously been identified in *E. coli* but without a high specificity for U-rich RNA [35,36].

Variance in the IC50 and K$_i$ for the nanoRNAs reported in Table 1 shows a preference for *Tma* Hfq interaction with uracil. Specifically, the data indicates a uracil at the 3' end of the nanoRNA is important for high affinity interactions with *Tma* Hfq, whereas a cytosine in the fifth position of a 5nt or 6nt nanoRNA drastically decreases the affinity of the nanoRNA. These finding are consistent with a previous study that showed that binding to the *L4* face of Hfq from *Salmonella typhimurium* is specific to the 3' end of RNA [31]. Single nucleotide substitutions in a hexa-uridine substrate at either the first or sixth position, indicated that cytosine in the sixth position should have little to no effect on binding affinity. It should be noted that a 13-20 nM binding affinity was observed for *Sty* Hfq with r(U)$_6$, r(U$_5$C), and r(U$_5$A), whereas our data indicates that *Tma* Hfq has a much higher affinity (3-6 nM) for r(U)$_5$ and r(CU$_2$CU)$_2$. This discrepancy is likely due to differences in the *L3* binding site between the homologs, specifically residues *Sty* F42/ *Tma* Y44 and *Sty* Q41/ *Tma* S43. *Eco* Hfq has the same binding site residues as *Sty* Hfq. The higher affinity could be necessitated by the thermophilic environment of *Tma*, where dissociation and cycling would be more rapid. Due to high conservation in the binding site we would predict that the nanoRNAs identified in this study interact with *Eco* Hfq with nanomolar affinities (proximal binding site of *Sty* Hfq and *Eco* Hfq are identical). Interestingly, the two cytosines in r(CU$_2$CU$_2$) do not adversely affect the binding affinity, on the contrary a slightly higher binding affinity is observed compared to r(U)$_5$. This indicates that two consecutive uracils are required at either the 5' or 3' end of the sequence. This data suggests that cytosine isn't discriminated against as previously hypothesized, but likely reduces the number of hydrogen bonds resulting in a lower binding affinity when at the 3' end of the sequence [18].

Purification of *Tma* Hfq under denaturing conditions lowers the A260nm/A280 ratio to 0.83 or lower, indicating a removal of most nucleic acid (pure protein is ~0.67). However, when crystallized without the addition of RNA density corresponding to a small RNA strand was discovered on the *L3* face in the traditional uridine-binding pocket. Each base nestles in a pocket at the monomer interface, with the phosphates in the pore. The uridine base is π-stacked with tyrosine-44 and hydrogen bonded to glutamine-10 of the adjacent monomer, similar to previously models. Sterics do not appear to preclude a cytosine from binding within the uridine pocket (Figure 3C).

Numerous studies have demonstrated that hexamers are a functional form of Hfq [7,10,14,33]. The crystal structure of *Tma* Hfq (Figure 3B) supports the formation of a *Tma* Hfq hexamer, with each monomer exhibiting the Sm fold that is characteristic of Hfq proteins. Some studies have indicated the presence of higher order oligomeric states *in vitro*, though presently any *in vivo* physiological roles of these higher order oligomeric states remains elusive [17-19,22,23,33]. A recent study of the oligomerization of Hfq from *E. coli* found that Hfq adopts multiple oligomeric states at μM concentrations [10]. Furthermore, this study indicated that the midpoints of the monomer-to-hexamer and the hexamer-to-multimer equilibria are 0.8 μM and 4.9 μM, respectively. The midpoints of these two transitions suggest that at intracellular concentrations (1 μM in *E. coli*) the predominant oligomeric states are the monomer and hexamer [37]. During crosslinking of *Tma* Hfq, the presence of a higher order peak corresponding to a dodecamer in addition to a hexamer was seen with all three cross-linking reagents (formaldehyde, glutaraldehyde, EDC) (Table 3). The relative size of the dodecamer peak with respect to the hexamer peak, increases as the cross-linker length increases (EDC<formaldehyde<glutaraldehyde), suggesting while higher order states are present *in vitro*, the hexameric rings are not in as close contact as the individual monomers in the hexamer.

AnSEC elution volumes observed for the native and crosslinked *Tma* Hfq samples suggest that the hexamer and higher order oligomeric state are in relatively rapid equilibrium. Native *Tma* Hfq elutes as one species at a molecular weight corresponding to a nonamer (Figure 2A). Crosslinking with EDC prior to AnSEC yields two peaks corresponding to a hexamer and a higher-ordered oligomer confirmed by MALDI-TOF MS to be a dodecamer (Figure 2B, C). The presence of a hexamer in the crosslinked sample agrees with the results obtained by MALDI-TOF MS analysis of the same sample and suggests that the shape of the protein is not the sole factor causing the higher apparent molecular weight in AnSEC. The orientation or spacing between the hexamers (linked tightly as a double layered ring or as two separate rings connected by a flexible linker) in the dodecamer could correspond to a large change in hydrodynamic radius resulting in the higher apparent molecular weight. Further studies are needed to determine the three-dimensional structure of the putative dodecamer, specifically determining the interface between the hexamers. A K_a of 0.5 ± 0.1 μM determined by ITC of the dodecamer-to-hexamer dissociation is 10-fold lower than what was previously determined for *Eco* Hfq [10]. Semi-native Western blot verified that the two predominant oligomeric states in the concentration range being monitored by ITC (20 μM to 78 nM *Tma* Hfq) are indeed the hexamer and dodecamer (Figure 4C, D, E).

It is common for proteins from thermophilic bacteria to adopt higher-order oligomeric states versus their mesophilic homologs [38,39]. These higher-order states tend to be associated with enhanced thermostability; with formation of the proper oligomeric state critical for function. Fluorescence polarization utilizing FAM-r$(A)_{18}$ (*L4* face) and FAM-r$(U)_6$ (*L3* face) probes revealed that both hexamer and dodecamer states of *Tma* Hfq bound poly-A and poly-U RNA strands with nanomolar affinity. The K_d for FAM-r$(U)_6$ for the native and EDC crosslinked samples were statistically different ($P<0.01$, based on unpaired t-test with Welch's correlation), suggesting the populations of dodecamer and hexamer varied between the two samples. For FAM-r$(U)_6$, the K_d was similar for the native and hexameric states, but differed significantly between the dodecamer and native states. In contrast, the K_d for FAM-r$(A)_{18}$ was approximately equivalent for the native and EDC states of *Tma* Hfq, whereas the K_d for crosslinked dodecamer and crosslinked hexamer are significantly different ($P<0.01$, based on unpaired t-test with Welch's correlation), with FAM-r$(A)_{18}$ exhibiting a 2-fold higher affinity for *Tma* Hfq dodecamer. Cumulatively, these finding suggest that the relative population of the two oligomeric states is altered by the presence of poly-U RNA, shifting the equilibrium toward the hexameric state.

Semi-native Western blots of *Tma* Hfq in the presence of RNA support r$(U)_6$ shifting the equilibrium toward a hexamer, whereas r$(A)_{18}$ does not appear to alter the equilibrium either direction, suggesting that the *L3* faces form the interface between the two hexamers in the dodecameric state (Figure 5). This proposed model is supported both by the RNA-binding properties of the two oligomeric states and by the effect of poly-U RNA on the equilibrium between

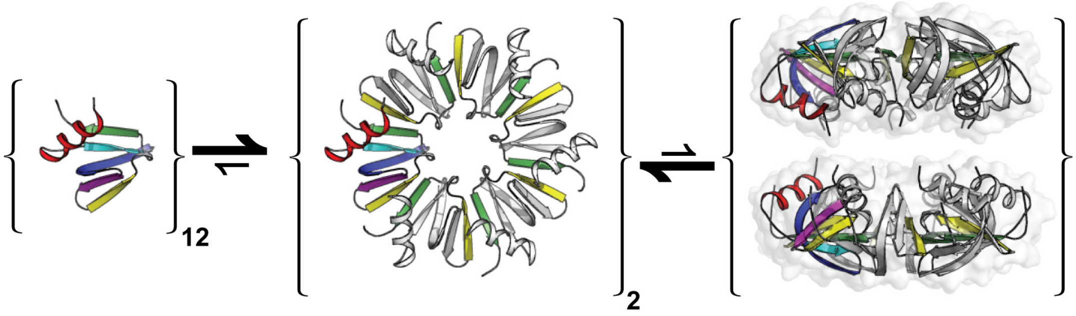


Figure 5. Working model of *Tma* Hfq oligomerization. Working model for the equilibrium that exist between the oligomeric states of Hfq in solution. The dodecamer forms with the proximal faces forming the interface between the two Hfq hexamers.

the two states. Unfortunately, the nanomolar binding affinity of the dodecamer for FAM-r(U)$_6$ is not consistent with an occluded binding site. A possible explanation for this inconsistency is the presence of an additional binding site on the Hfq toroid structure. The lateral surface of the *Eco* Hfq hexamer has been shown to play an important role in Hfq-RNA interactions. This lateral surface of *Eco* Hfq has no apparent nucleic acid sequence specificity and was inefficient for r(U)$_6$ RNA-binding in a *L3* site mutant [19,31,36]. While the *L3* and *L4* binding sites are fairly conserved between *Eco* Hfq and *Tma* Hfq, the lateral binding surface is drastically different. Of the residues shown to be involved in RNA interactions on the later surface of *Eco* Hfq, phenylalanince-39 (F39), arginine-16 (R16), arginine-17 (R17), glutamic acid-18 (E18), and arginine-19 (R19)[19,19], only F39 and R16 are conserved in Tma Hfq. Instead of the arginine-rich rim present in E. coli Hfq, the equivalent surface consists of arginine-18 (R18), valine-19 (V19), asparagine-20 (N20), and lysine-21 (K21). Future studies could concentrate on characterizing the lateral binding site through the use of *Tma* Hfq mutants and strategically designed RNA-probes. An alternative explanation for the nanomolar binding affinity of r(U)$_6$ RNA is that the two hexameric rings are not stacked in perfect alignment, but instead staggered, or loosely tethered, allowing for poly-U RNA to bind to the dodecameric state in a manner that decreases the stability of the oligomer and results in dissociation of the two hexamers. To determine the mode of binding, future experiments could examine the orientation of the two hexamers in the dodecamer, as well as the location of poly-U binding on the dodecamer.

5 Methods

5.1 Cloning and Expression

The coding sequence for *T. maritima (Tma)* Hfq (geneID = 897578) was cloned into the pET-28b(+) vector (Novagen) using polymerase incomplete primer extension (PIPE) [40] between the NdeI and XhoI restriction sites. The vector and insert were transformed into chemically competent *E. coli* TOP10 cells for ligation and amplification. Plasmid was purified from TOP10 cells using the QIAprep Spin Miniprep Kit (QIAGEN); purified plasmid was transformed into competent *E. coli* BL21 (DE3) cells for expression. The transformed cells were grown at 37°C in LB broth and induced with 1 m*M* IPTG at an OD600 between 0.8 and 1.0 for 4 hrs. The culture was centrifuged at 15,000g for 5 min at 4°C and stored at -20°C overnight.

5.2 Denaturing Scheme Protein Purification

Cell pellet was subjected to a freeze/thaw cycle prior to resuspension in lysis buffer (50 m*M* Tris pH 7.5, 750 m*M* NaCl, 0.4 m*M* phenylmethylsulfonyl fluoride (PMSF), 0.01 µg/mL lysozyme, egg white (Fisher)). Cells were incubated at 37°C for 30 min followed by mechanical lysis using a microfluidizer. Lysate was centrifuged at 35,000g for 20 min at 10°C. Supernatant was incubated at 85°C for 20 minutes then centrifuged at 35,000g for 20 min at 10°C. The sample was denatured by addition of guanidinium hydrochloride (gnd-HCl) to a final concentration of 6 *M*. Supernatant was loaded onto a Ni^{2+} charged affinity column (HiTrap^{TM} chelating Hp) using AKTAprime^{TM} plus, followed by 10 column volumes of wash buffer (50 m*M* Tris pH 7.5 containing 150 m*M* NaCl, 6*M* gnd-HCl and 10 m*M* imidazole) and eluted with elution buffer (50 m*M* Tris pH 7.5 containing 150 m*M* NaCl, 6 *M* gnd-HCl and 600 m*M* imidazole) using a 0-100% gradient. The elution fractions were analyzed with SDS-PAGE then refolded by slow removal of denaturant via dialysis into 25 m*M* Tris pH 8.0, 1 *M* arginine, and 0.2 m*M* PMSF overnight. The sample was then dialyzed in 50 m*M* Tris pH 8 containing 150 m*M* NaCl and 12.5 m*M* EDTA for digestion. Protein was digested with thrombin (1:600) overnight at 37°C and cleavage was verified by matrix assisted laser desorption/ionization spectrometry (MALDI TOF MS). After digestion, the sample was run over a benzamidine column to remove thrombin then a preparative gel filtration column to ensure that the protein was folded.

5.3 Oligonucleotide Purification

The protein purification protocol was followed through the heat cut step. The sample was loaded onto a Ni^{2+} charged affinity column (HiTrap^{TM} chelating Hp) using AKTAprime^{TM} plus, followed by 10 column volumes of wash buffer (50 m*M* Tris pH 7.5 containing 150 m*M* NaCl, and 10 m*M* imidazole) and eluted with elution buffer (50 m*M* Tris pH 7.5 containing 150 m*M* NaCl, and 600 m*M* imidazole) using a 0-100% gradient. The elution fractions were analyzed with SDS-PAGE then dialyzed in 50 m*M* Tris pH 8 containing 150 m*M* NaCl and 12.5 m*M* EDTA for digestion; Protein was digested with thrombin (1:600) overnight at 37°C. The sample was run over a benzamidine column to remove thrombin. The digested sample was then diluted with buffer A (25 m*M* Tris pH8.5, 50 m*M* NaCl) and loaded onto a quaternary amine column, followed by 10 column volumes of buffer A and eluted with buffer B (25 m*M* Tris pH8.5, 2 *M* NaCl) using a stepwise gradient. Fractions that eluted at 20% buffer B (440 m*M* NaCl) were combined and concentrated for phenol-chloroform extraction [19]. An equal volume of 100% v/v Ethanol was added to the aqueous phase and incubated at -80˚C for at least 2 hours followed by centrifugation at 9,300 *g* (Eppendorf fixed-angle rotor) for 20 minutes at 4˚C. The supernatant was removed and the pellet was dried at room temperature.

5.4 Colorimetric Sugar Assays [24]

Dried samples were resuspended in distilled water. Benedicts test was performed to test for free reducing sugar; 10 µ*L* of Benedict's Reagent (943 m*M* anhydrous sodium carbonate, 588 m*M* sodium citrate tribasic dehydrate, 68.6 m*M* cupric sulfate pentahydrate) was added to 50 µ*L* of

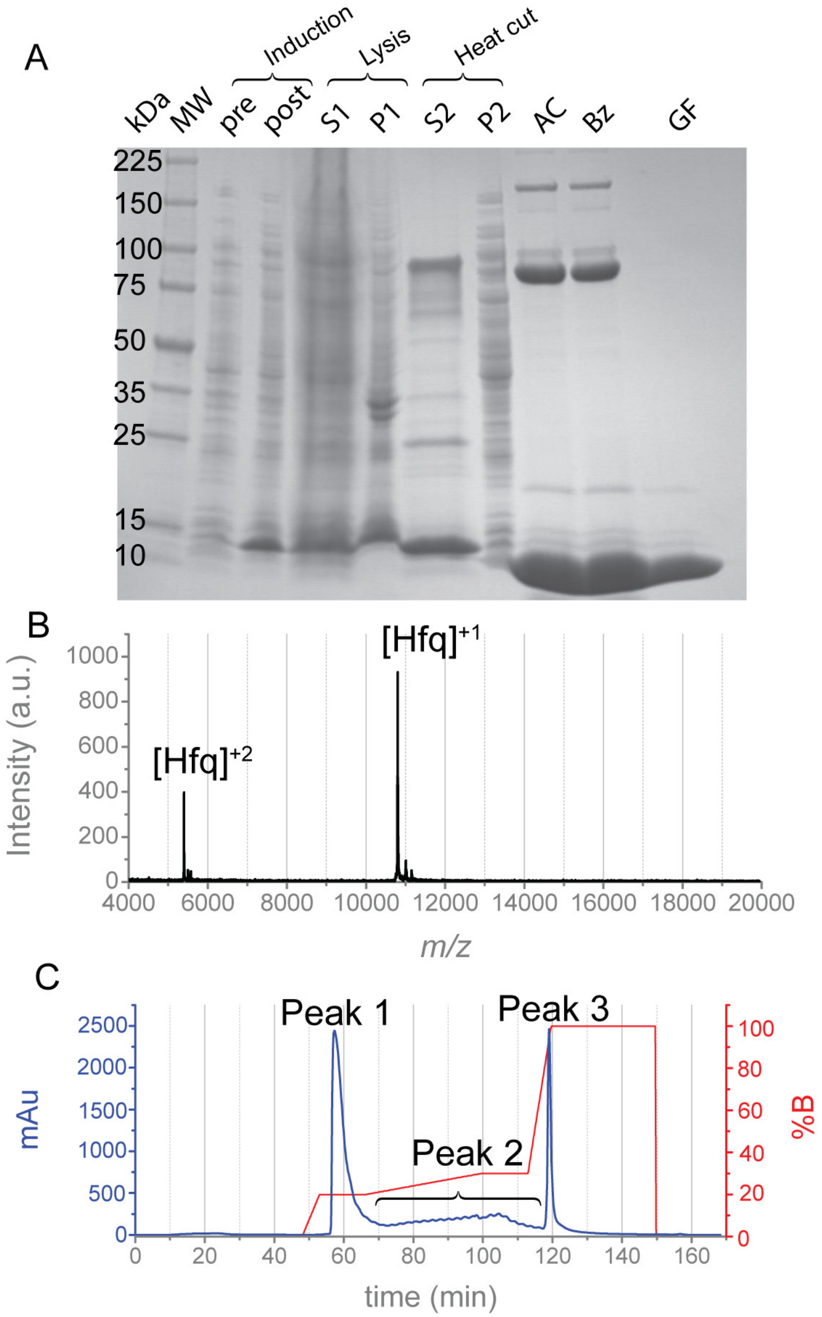

A
Induction
Lysis
Heat cut
kDa
MW
pre
post
S1
P1
S2
P2
AC
Bz
GF
225
150
100
75
50
35
25
15
10
B
Intensity (a.u.)
1000
800
600
400
200
0
[Hfq]+2
[Hfq]+1
4000
6000
8000
10000
12000
14000
16000
18000
20000
m/z
C
Peak 1
Peak 2
Peak 3
mAu
2500
2000
1500
1000
500
0
%B
100
80
60
40
20
0
0
20
40
60
80
100
120
140
160
time (min)

Figure 6. Expression and purification of *Tma* Hfq. Representative SDS-PAGE and MALDI-TOF MS spectrum of recombinant *Tma* Hfq purification and co-purified nanoRNA separation. B) SDS-PAGE samples collected at major steps in the recombinant *Tma* Hfq purification protocol. The samples run on 4-20% w/v TGX gel as follows: Promega Broad Range Protein Molecular Weight Marker (MW), pre-induction (pre), post-induction (post), the soluble (S1) and pellet (P1) fractions from lysis, the soluble (S2) and pellet (P2) fractions from the heat cut step, thrombin-treated sample (AC), benzamine column flow-through (Bz), and gel filtration eluent (GF). A) Two peaks are detected in the MALDI-TOF MS spectrum that can be assigned to the +1 and +2 charge states of the *Tma* Hfq monomer, which has a molecular weight of 10,797.2 Da based on the amino acid sequence. No peaks were observed in the higher molecular weight range (20-100 kDa). C) Chromatogram of *Tma* Hfq being separated from a subset of Hfq-binding RNA via quaternary amine anion-exchange column. The absorbance trace indicates two distinct peaks (Peak 1 and Peak 3) eluting at different salt concentrations, which increases with buffer B (%B). An intermediate "peak" (Peak 2) is observed that is extremely broad and rippled. The two elution peaks, which elute at approximate 20% and 100% buffer B correspond to Hfq bound to oligonucleotides and isolate RNA, respectively.

sample in a PCR tube. Bial's orcinol assay was performed to test for pentose sugar; 50 μ*L* of orcinol reagent (24.2 m*M* orcinol monohydrate, 0.025% w/v ferric chloride hexahydrate, 6 *M* HCl) was added to a PCR tube containing 50 μ*L* of sample. Dische diphenylamine assay was performed to test for deoxyribose; 50 μ*L* of Dische diphenylamine reagent (58.12 m*M* diphenylamine, 0.66% v/v ethanol, 11.4 *M* glacial acetic acid, 17.7 m*M* sulphuric acid) was added to a PCR tube followed by 50 μ*L* of sample. Samples for each assay were sealed and incubated for 20 minutes in boiling water. Controls used throughout these assays include the 0.15 mg/mL ribose (Sigma), 7.5 mg/mL RNA from Baker's yeast (Sigma), 0.45 mg/mL DNA from calf thymus (Sigma), 0.45 mg/mL bovine serum albumin (BSA) (Sigma), and water.

5.5 Chemical Cross-linking.

Tma Hfq was dialyzed into 25 m*M* HEPES pH 8.00 and 200 m*M* NaCl prior to crosslinking with formaldehyde or glutaraldehyde using an 'indirect' method [25]. Experimental setup for this indirect method consisted of a Linbro plate with a microbridge and coverslip upon which 40 μ*L* of the crosslinking agent (25% v/v) and 15 μ*L* of *Tma* Hfq (1 mg/mL) were aliquoted, respectively. The chamber was sealed with vacuum grease. The crosslinking agent was acidified with 122 m*M* HCl before incubation. Samples were incubated at 37 °C for 40 minutes and the reaction was stopped by the addition of 5 μ*L* 1 *M* Tris pH 8.00. Salts and crosslinking reagents were removed from samples using a C_4 zip tip (Millipore), and the molecular weight of the crosslinked oligomers were assessed by MALDI-TOF MS (protocol described below). Data was collected in three independent replicates for both formaldehyde and glutaraldehyde.

Tma Hfq was also crosslinked with 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) after dialysis into 25 m*M* HEPES pH 8.0 and 200 m*M* NaCl. The reaction mixture consisted of 67 m*M* EDC, 167 m*M* N-Hydroxysulfosuccinimide sodium salt (sulfo-NHS), and 1 mg/mL purified recombinant *Tma* Hfq. The reaction was incubated at room temperature (RT) for 4 hours

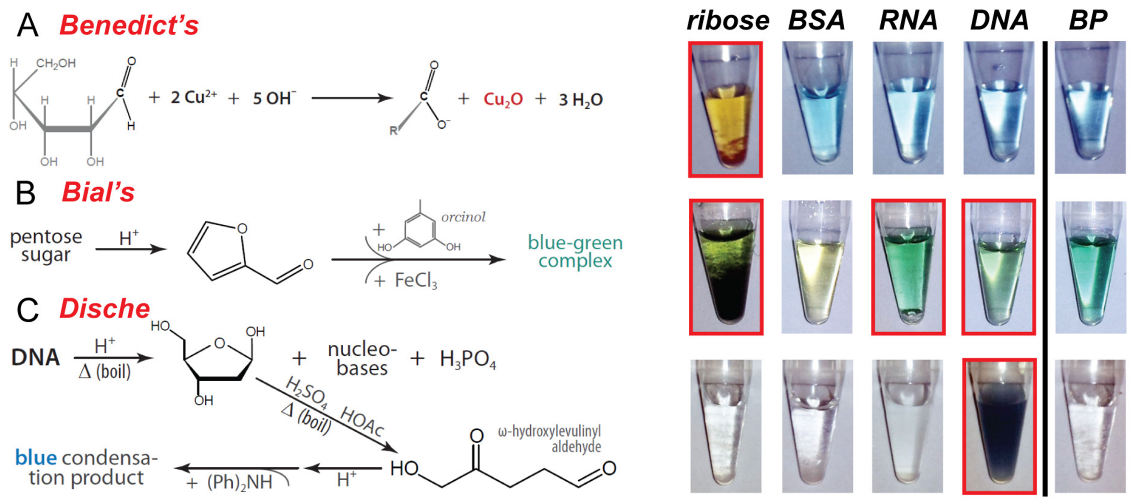


Figure 7. Colorimetric assays of co-purified partners. Colorimetric assays verify co-purified nanoRNAs contains a pentose sugar which is not a free reducing sugar or deoxyribose. A) Benedict's reaction indicated BP does not contain free reducing sugar, B) Bial's reaction indicated the presence of a pentose sugar and C) Dische's diphenylamine reaction indicated that the pentose sugar is not deoxyribose. The reactions utilized in each assay are shown to the left of the results. Controls are listed, with positive controls highlighted in red.

Table 3. Molecular weights and oligomeric states observed by MALDI-TOF for crosslinking *Tma* Hfq

Crosslinking	Molecular Weight (kDa)		Oligomeric State	
Reagent	pk1	pk2	pk1	pk2
Formaldehyde	66 ± 1	133 ± 2	6.15 ± 0.09	12.3 ± 0.2
Gluaraldehyde	71.5 ± 0.3	141 ± 7	6.60 ± 0.3	13.1 ± 0.6
EDC	68.0 ± 0.6	136 ± 2	6.29 ± 0.06	12.6 ± 0.2

then stopped by the addition of β-mercaptoethanol (BME) to a final concentration of approximately 18 m*M*. Cross-linking was assessed by MALDI-TOF MS using the protocol described below. Data was collected in three independent replicates.

5.6 MALDI TOF MS

MALDI-MS was performed on a Bruker Microflex MALDI. Proteins of approximately 1 mg/mL were diluted 1:4 with 0.01% trifluoroacetic acid (TFA). The protein sample was spotted onto a MSP 96 target ground steel sample plate (Bruker Daltonics) with an equal volume of SA (15 mg/mL sinapinic acid in 50% acetonitrile and 0.05% TFA) and allowed to air dry. The instrument was calibrated by the close external method using a series of low molecular weight (Insulin, Cytochrome C, Ubiquitin I, and Myoglobin) or high molecular weight (Protein A, Trypsinogen, Protein A, and Bovine Albumin) protein calibrants. Spectra were obtained by averaging approximately 50 laser shots with the following settings: positive ion, linear mode, grid voltage 40-75%, m/z range 4,000-20,000 or 20,000 -100,000.

5.7 NMR Spectroscopy

Sample was prepared for NMR studies by dissolving approximately 857 µg into 500 µ*L* 90%/10% D_2O/H_2O. The amount of binding partner in the sample was estimated based on absorbance at 260 nm using the molecular weight and extinction coefficient for pure 5'Phos-rUrUrUrCrU, which is 1547.9 g/mol and 46,300 L/(mol·cm), respectively. All NMR spectra for the sample were recorded at 40˚C. ^{1}H spectra were recorded on a Bruker AVANCE spectrometer operating at a proton frequency of 600 MHz with Bruker 5 mm TXI cryoprobe. The ^{1}H data was recorded and processed using Topsin 3.0. ^{1}H one dimensional proton spectra included water suppression using excitation sculpting using gradients. The ^{1}H-^{1}H NOESY spectrum was recorded with a mixing time of 50 ms [22]. The ^{31}P spectrum was recorded on Varian Mercury spectrometer operating at a ^{31}P frequency of 121.3 MHz. The data was processed using Vnmr 6.1c and analyzed using MestReNova suite.

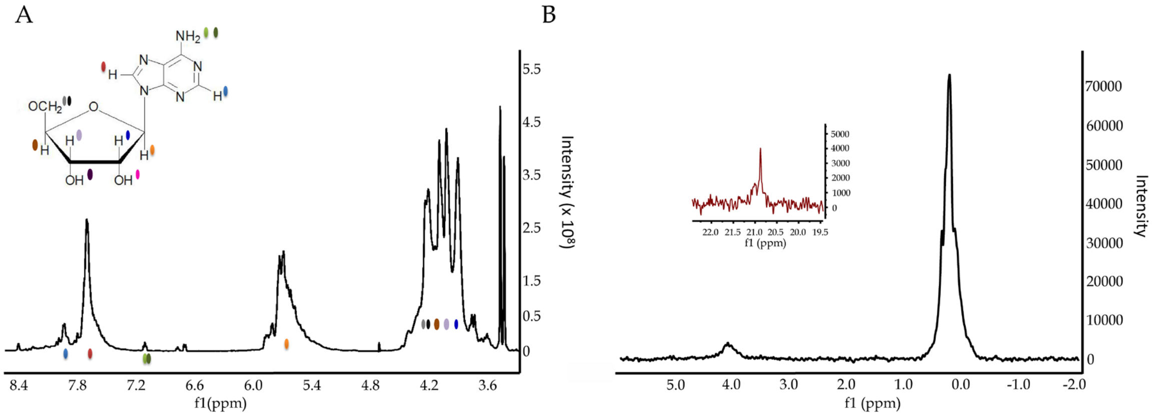


Figure 8. NMR spectra of co-purified partners. A) ^{1}H spectrum and B) ^{31}P spectrum of binding partner in D_2O. The peaks in the ^{1}H spectrum are consistent with a sample containing a nucleotide. ^{31}P spectrum has a peaks corresponding to phosphate diester at ~0 ppm, phosphate monoester at ~4 ppm (not shown), and phosphonate or cyclic phosphate at ~21 ppm (inset).

5.8 Fluorescence Polarization

Tma Hfq was dialyzed into 25 m*M* Tris pH 8.00 and 350 m*M* NaCl prior to binding studies. Fluorescence anisotropy/polarization measurements were collected with a PHERAstar microplate reader. 5-Carboxyfluorescein-labeled r(U)$_6$ (FAM-r(U)$_6$) and FAM-(A)$_{18}$ RNA (Integrated DNA Technologies) were used to probe the binding sites of *Tma* Hfq. FAM-r(A)$_{18}$ was annealed prior to binding assays by incubating at 85°C for 3 minutes and then placed on ice for 10 minutes [41]. Samples were excited at 490 nm and emission was measured at 522 nm [42]. To determine the apparent equilibrium dissociation constant (K_d) for FAM-RNA, *Tma* Hfq was serially diluted in 96-well black polystyrene assay plates (Costar) in the presence of 5 n*M* FAM-RNA; the final volume in each well was 150 μ*L*. The binding assay samples were incubated in the dark for 45 minutes at RT prior to measurements to ensure equilibrium binding. Data was collected in four independent replicates.

5.9 Analysis of Fluorescence Polarization Data.

Fluorescence data was fit to a model that assumed that a 1:1 complex formed between (*Tma* Hfq)$_6$ and FAM-RNA under experimental conditions in which Hfq exists as a hexamer. The model involves fitting the data to a sigmoidal Boltzmann function, which is related to the Hill equation [43,44], and can be rearranged to read

$$y = \frac{(A_1 - A_2)}{1 + e^{\frac{(x - x_0)}{dx}}} + A_2 \qquad \text{(Equation 1)}$$

where x_0 is the inflection point sigmoidal curve, *dx* is the width of the transition and A_1 and A_2 are the fluorescence polarization intensities of the initial and final states, respectively [18,45,46]. Nonlinear least-squares fits of the equation to the data were performed in OriginPro7.5.

5.10 Competitive Binding Assay

Oligonucleotides used in this study include 5'Phos-r(U_5), 5'Phos-r(U_3CU), 5'Phos-r(U_4C), 5'Phos-r(CU_2CU), 5'Phos-r(U_4CU), and 5'Phos-r(CU_2CU_2). Custom oligonucleotides were ordered from integrated DNA technologies and were purified by RNase-free high-pressure liquid chromatography (HPLC) after synthesis. Fluorescence polarization measurements were collected under the settings described above. Samples contained 5 n*M* FAM-r$(U)_6$ and 15 μ*M* (*Tma* Hfq$)_6$ in 25 m*M* Tris pH 8.00 and 350 m*M* NaCl. The concentration of oligonucleotides ranged from 5 μ*M* to 71 a*M*. Samples were incubated in the dark for 45 minutes at room temperature prior to measurements to ensure equilibrium binding.

5.11 Crystallization

Crystallization trials began with freshly purified *Tma* Hfq samples concentrated to ≈ 10-20 mg/ml in 25 m*M* Tris pH 8.5, 100 m*M* NaCl. JCSG Core Suites I, II, III and IV (Qiagen), as well as PEG/Ion screens (Hampton Research), were used for sparse-matrix screening. Nanoliter trials were set in 96-well plates *via* a nanolitre-scale liquid-handling robot (TTP Labtech's mosquito Crystal). Both hanging-drop and sitting-drop vapour diffusion formats were used, with a 100-μ*L* reservoir and 200-n*L* droplet (composed as 1:1 protein:reservoir). All crystallization trials were incubated at 291 K. Leads were then refined on a larger-volume scale (24-well VDX plates) by systematically varying typical crystallization parameters—buffer pH, protein concentration and precipitant types/concentrations [47]. In addition to fine grids centered on the initial hits, 96-well additive screens (Hampton Research) were also applied to the leads. Two improved conditions were found, corresponding to 150 mM tri-potassium citrate and 30% w/v PEG-3350 with either 1.0 *M* glycine or 0.1 *M* sarcosine as an additive. Optimized crystals grew as rhombic prisms within 1 week and developed to maximal dimensions of ~50 μm/edge by 2 weeks. A working cryo-protection procedure for both the glycine- and sarcosine-based conditions was found to be gentle passage of a crystal (in a nylon cryo-loop), over the course of ~10-15 s, through 8 μl of mother liquor supplemented with 0.6 μ*L* of neat PEG-400. Suitable cryo-protection (lack of ice-rings, diffraction

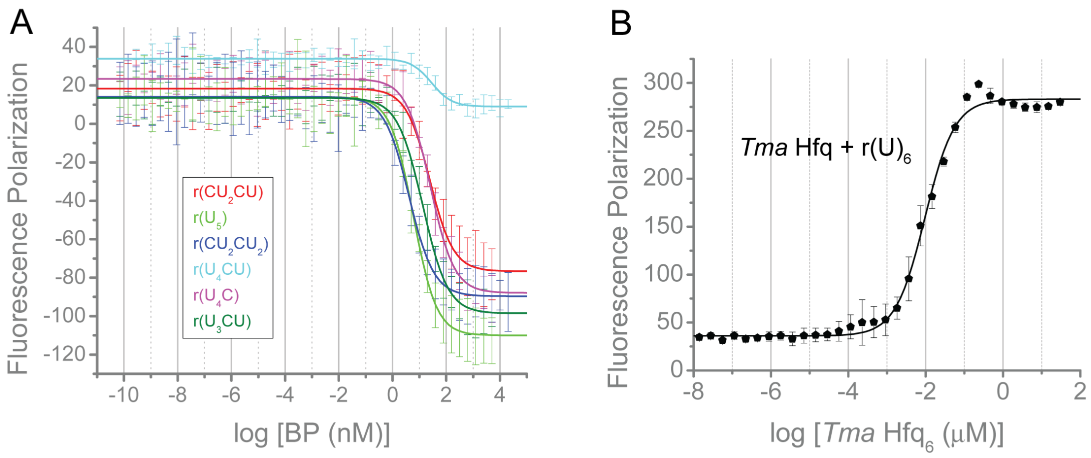


Figure 9. Competition assay curves for *Tma* Hfq with identified nanoRNAs. Competitive assays were performed with FAM-r(U)$_6$ as the target.

quality, *etc.*) was screened on a Rigaku MicroMax-007 rotating anode X-ray generator equipped with a Saturn-92 charged-coupled device (CCD) detector.

5.12 Diffraction Data Collection

Crystals were harvested in nylon loops, incubated in cryoprotectant for 10-15 seconds and then flash-cooled in a bath of liquid nitrogen ($T \approx 77K$). Diffraction datasets were collected on the Southeast Regional Collaborative Access Team (SER-CAT) 22ID and 22BM beamlines, equipped with MAR 300 and MAR 225 CCD detectors, respectively, or on the Northeast Collaborative Access Team (NE-CAT) 24-ID-C, with a DECTRIS Pilatus detector. Diffraction data sets were collected for multiple crystals, with the highest resolution of the crystals grown without $r(U)_5$ at 2.66Å and the diffraction with additive $r(U)_5$ at 2.05Å.

5.13 Diffraction data processing.

Raw diffraction data were indexed, integrated and scaled using the programs XDS and XSCALE [48]. Reduced datasets were examined with Xtriage and other utilities in the PHENIX suite [49] in order to verify indexing, gauge diffraction anisotropy, detect pseudo-translational symmetry, etc; twinning tests were also performed, though merohedral twinning was not a concern here because of the orthorhombic crystal system and unequal cell edges. Non-crystallographic symmetry (NCS) was evaluated by computing *(i)* the Matthews coefficient [50]; *(ii)* native Patterson maps, using CCP4's FFT module [51]; and *(iii)* the self-rotation function, using GLRF [52]. Anisotropy correction was performed by the UCLA-DOE Lab – Diffraction Anisotropy Server [26].

5.14 Structure solution.

In preparation for molecular replacement, the structure of *Bacillus subtilis* Hfq (3HSB) was trimmed to side-chains of conserved residues based on sequence alignment with *Tma* Hfq ($3HSB^{trimmed}$) using Sculptor [53]. Using $3HSB^{trimmed}$ as a probe, *Tma* Hfq was solved by molecular

Table 4. ***Tma*** **Hfq diffraction and refinement statistics**

Diffraction Statistics	***Tma*** **Hfq *apo***	***Tma*** **Hfq *apo* aniso-corrected**	***Tma*** **Hfq r(U_5)**	***Tma*** **Hfq r(U_5) aniso-corrected**
Diffraction Source	APS SER-CAT 22-ID	-	APS NE-CAT 24-ID-C	-
Wavelength (Å)	0.9788	-	0.9792	-
a, b, c (Å)	39.08, 133.50, 206.18	-	38.67, 133.13, 205.58	-
α, β, γ	90.0, 90.0, 90.0	-	90.0, 90.0, 90.0	-
Resolution Range (Å)	81.59 – 2.66	81.59 – 2.66	81.43 – 2.05	81.43 – 2.05
Completeness (%)	99.8 (98.5)	84.7 (29.9)	99.7 (98.7)	71.7 (19.8)
$\langle I/\sigma(I)\rangle$	21.6 (2.6)	25.4 (6.4)	19.1 (2.4)	26.1 (7.2)
R_{sym}†	14.1 (148.7)	11.8 (58.9)	4.6 (76,5)	3.8 (23.0)
R_{meas}‡	14.6 (154.3)	12.3 (61.4)	5.0 (83.2)	4.1 (25.2)
R_{pim}¥	3.8 (40.7)	3.2 (16.9)	2.0 (32.4)	1.6 (10.2)
CC(1/2)§	99.9 (90.2)	99.9 (95.3)	99.9 (95.0)	99.9 (98.1)

Refinement Statistics	***Tma*** **Hfq *apo***	***Tma*** **Hfq apo aniso-corrected**	***Tma*** **Hfq r(U_5)**	***Tma*** **Hfq r(U_5) aniso-corrected**

RWORK/RFREE	19.87/25.82	17.91/24.88	-	19.35/24.11
R.M.S.			-	
Bonds (Å)	0.0114	0.0139	-	0.009
Angles (˚)	1.444	1.556	-	1.155
Ramachandran (%)				
Favored	93.1	93.0	-	97.3
Allowed	6.8	7.0	-	2.7
Outliers	0.1	0.0	-	0.0
Avg. B-factor	55.50	29.60	-	33.60

† $R_{\mathrm{sym}} = (\sum_{hkl} \alpha \sum_i |I_i(hkl) - \langle I_i(hkl) \rangle|)/(\sum_{hkl} \sum_i I_i(hkl))$, where $I_i(hkl)$ is the intensity of the i^{th} observation of reflection hkl, $\langle \cdot \rangle$ denotes the mean of symmetry-related (or Friedel-related) reflections, and the coefficient $\alpha = 1$; the outer summations run over only unique hkl with multiplicities greater than one.

‡ R_{meas} is defined analogously as R_{sym}, save that the prefactor $\alpha = \sqrt{N_{hkl}/(N_{hkl} - 1)}$ is used; N_{hkl} is the number of observations of reflection hkl (index $i = 1 \rightarrow N_{hkl}$).

¥ $R_{\mathrm{p.i.m.}}$, the precision-indicating merging R-factor, is defined as above but with the prefactor $\alpha = \sqrt{1/(N_{hkl} - 1)}$.

§ $CC_{1/2}$ is the correlation coefficient between intensities chosen from random halves of the full dataset.

replacement using Phaser from PHENIX. The model was confirmed and initially refined in AutoBuild. Further refinement was done using REFMAC5 [54]and the phenix.refine module. Model building and structure manipulation steps were performed in the Coot software environment [55]. The higher resolution r(U)$_5$ additive structure was solved using the same probe and procedure.

5.15 Analytical size-exclusion chromatography (AnSEC).

Recombinant *Tma* Hfq was dialyzed into 25 m*M* Tris pH 8.0 and 1 *M* NaCl. The chromatography system consisted of a Superdex 200 10/300GL column (spherical composite of crosslinked agarose and dextran matrix) and a Biologic DuoFlow^{TM} system (BioRad) at 4°C. *Tma* Hfq and molecular weight standards were injected and eluted with approximately seven column volumes of 25 m*M* Tris pH 8.0 containing 1 *M* NaCl at 0.4 mL/min; absorbance was monitored at 280 nm throughout each run. A standard curve was generated using Gel Filtration Markers Kit for Protein Molecular Weights 29,000 – 700,000 Da (MWGF1000-1KT, Sigma-Aldrich).

5.16 Production and purification of anti-*Tma* Hfq pAb.

Recombinant *Tma* Hfq was dialyzed into 10 m*M* Na_2HPO_4, 1.8 m*M* NaH_2PO_4, 2.7 m*M* KCl, and 137 m*M* NaCl at pH 7.40 (PBS) and concentrated to 1 mg/mL using a 3350-Da molecular weight cut off (MWCO) concentrator. pAbs were produced against purified recombinant *Tma* Hfq by Covance (Denver, PA) using their 77-day production protocol in four specific-pathogen-free New Zealand White rabbits. Sera was obtained from a pre-inoculation bleed, three production bleeds, and a terminal bleed from each rabbit, and were stored at −80 °C.

In preparation for purification, serum was thawed at 4 °C with gentle mixing and then clarified by centrifugation at 5,000*g* for 15 minutes. The serum was diluted threefold with binding buffer (20 m*M* phosphate pH 7.00) and then loaded onto a HiTrap^{TM} protein G column using an AKTAprime HPLC. The column was washed with 10 column volumes of binding buffer, and then eluted with 10 column volumes of 0.1 *M* glycine pH 2.70 using a step gradient. Immediately upon elution, the samples were neutralized with 1 *M* Tris pH 9.00 at 0.2 volumes per volume of eluent. Fractions containing eluted pAbs were combined and dialyzed into 20 m*M* HEPES pH 8.0 with 300 m*M* NaCl. Purified pAbs (at ~38.2 μM) were stored at −20 °C in storage buffer consisting of 10 m*M* HEPES pH 7.80, 150 m*M* NaCl, 100 μg/mL BSA, and 25% v/v glycerol.

5.17 Semi-Native Western Blot Analysis.

Recombinant *Tma* Hfq was dialyzed into 25 m*M* Tris pH 8.00, 350 m*M* NaCl and serially diluted from 20 μ*M* to 78 n*M* across nine samples. The diluted samples were incubated at RT for 30 minutes and 4x *semi-native* loading buffer was added to each sample (final 1x working concentrations: 50 m*M* Tris pH 6.8, 0.5% sodium dodecyl sulfate, 10% glycerol, 12.5 m*M* ethylenediaminetetraacetic acid (EDTA), and 0.02% bromophenol blue). Protein samples were separated on a 7.5% w/v TGX gel (Bio-Rad) at RT using 1x Tris-glycine sodium dodecyl sulfate (SDS) running buffer (200V, 28 min), and then transferred to a nitrocellulose membrane using the Trans-Blot Turbo transfer system (Bio-Rad) with a Trans-Blot Turbo Transfer pack (Bio-Rad) at 2.5 A (up to 25 V) for 3 minutes. The Odyssey one-color protein molecular weight marker (Li-Cor), which fluoresces in the 700nm channel of any of the Odyssey Imaging Systems, was run in parallel to the protein samples. The membrane was blocked with 5% w/v dry milk and probed with rabbit anti-*Tma* Hfq polyclonal (pAb) antibodies (Covance; see Appendix 2). Goat anti-rabbit IgG IRDye 800CW was used as the secondary antibody (Li-Cor) for visualization using an Odyssey Li-Cor imaging system. The signal intensity of each band was quantified using the Image Studio software. The signal was then normalized by the total signal intensity of each lane in order to determine the fraction of the sample in each oligomeric state at a specific concentration (i.e. in a particular lane).

5.18 Isothermal Titration Calorimetry (ITC).

Measurements of heat changes associated with the oligomerization of *Tma* Hfq were made on a MicroCal VP^{TM}-ITC system at 25˚C. Recombinant *Tma* Hfq was dialyzed against 25 m*M* Tris pH 8.00, 350 m*M* NaCl, diluted with dialysis buffer to 22.5 μ*M* and degassed at 25 ˚C before loading into the ITC syringe. Dialysis buffer was filtered through a 0.22 μm filter membrane (Millipore) and degassed at 25˚C in a ThermoVac (MicroCal) before loading the sample cell (1.44 m*L* volume). Injections were set for 10 μ*L* of injectant (22.5 μ*M Tma* Hfq), with a 2-minute spacing interval between injections. Raw data was collected as thermal power (μcal/s) over time (min).

5.19 Analysis of ITC data.

Each titration peak in the thermograph (thermal power (μcal/s) as a function of time (min)) was integrated using MicroCal software (MicroCal LLC) to determine the thermal heat per injection (μcal/injection). To generate a binding curve, the thermal heat per injection was converted to the thermal heat per mole of injectant (μcal/mole or injectant) and plotted against the log of the concentration of *Tma* Hfq. The binding curve was then fit to a sigmoidal Boltzmann function, which is related to the Hill equation, and which can be rearranged to read

$$y = \frac{(A_1 - A_2)}{1 + e^{\frac{(x - x_0)}{\partial x}}} + A_2 \qquad \text{(Equation 1)}$$

where x_0 is the inflection point of the sigmoidal curve, *dx* is the width of the transition, and A_1 and A_2 are equal to the enthalpy (in kcal/mol) of the initial and final state, respectively[43,44]. Nonlinear least-squares fits of the equation to the data were performed in OriginPro7.5.

6 References

1. de Fernandez MTF, Hayward WS, August JT. Bacterial proteins required for replication of phage Q ribonucleic acid. Pruification and properties of host factor I, a ribonucleic acid-binding protein. *Journal of Biological Chemistry*. 247(3), 824–831 (1972).

2. Kulesus RR, Diaz-Perez K, Slechta ES, Eto DS, Mulvey MA. Impact of the RNA Chaperone Hfq on the Fitness and Virulence Potential of Uropathogenic Escherichia coli. 76, 3019–3026 (2008).

3. Kint G, De Coster D, Marchal K, Vanderleyden J, De Keersmaecker SC. The small regulatory RNA molecule MicA is involved in Salmonella enterica serovar Typhimurium biofilm formation. 10, 276 (2010).

4. Tsui HC, Leung HC, Winkler ME. Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in Escherichia coli K-12. *Mol Microbiol*. 13, 35–49 (1994).

5. McCullen CA, Benhammou JN, Majdalani N, Gottesman S. Mechanism of positive regulation by DsrA and RprA small noncoding RNAs: pairing increases translation and protects rpoS mRNA from degradation. *J Bacteriol*. 192(21), 5559–5571 (2010).

6. Brescia CC, Mikulecky PJ, Feig AL, Sledjeski DD. Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure. *RNA*. 9, 33–43 (2003).

7. Mikulecky PJ. Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. *Nature Struct Mol Biol*. 11, 1206–1214 (2004).

8. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. *Proc Natl Acad Sci USA*. 107, 9602–9607 (2010).

9. Soper TJ, Woodson SA. The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA. *RNA*. 14(9), 1907–1917 (2008).

10. Panja S, Woodson SA. Hexamer to monomer equilibrium of E. coli Hfq in solution and its impact on RNA annealing. 417, 406–412 (2012).

11. Updegrove TB, Correia JJ, Chen Y, Terry C, Wartell RM. The stoichiometry of the Escherichia coli Hfq protein bound to RNA. *RNA*. 17(3), 489–500 (2011).

12. Afonyushkin T, Vecerek B, Moll I, Bläsi U, Kaberdin VR. Both RNase E and RNase III control the stability of sodB mRNA upon translational inhibition by the small regulatory RNA RyhB. 33, 1678–1689 (2005).

13. Bandyra KJ, Said N, Pfeiffer V, Górna MW, Vogel J, Luisi BF. The seed region of a small RNA drives the controlled destruction of the target mRNA by the endoribonuclease RNase E. 47, 943–953 (2012).

14. Sharma CM, Papenfort K, Pernitzsch SR, Mollenkopf H-J, Hinton JCD, Vogel J. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. 81, 1144–1165 (2011).

15. Scofield DG, Lynch M. Evolutionary diversification of the Sm family of RNA-associated proteins. *Mol Biol Evol*. 25(11), 2255–2267 (2008).

16. Sauter C. Sm-like proteins in Eubacteria: the crystal structure of the Hfq protein from Escherichia coli. 31, 4091–4098 (2003).

17. Nikulin A, Stolboushkina E, Perederina A, *et al.* Structure of Pseudomonas aeruginosa Hfq protein. 61, 141–146 (2005).

18. Schumacher MA, Pearson RF, Moller T, Valentin-Hansen P, Brennan RG. Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. 21, 3546–3556 (2002).

19. Panja S, Schu DJ, Woodson SA. Conserved arginines on the rim of Hfq catalyze base pair formation and exchange. (2013).

20. Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G. The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol Cell*. 9(1), 11–22 (2002).

21. de Fernandez MTF, Hayward WS, August JT. Bacterial Proteins Required for Replication of Phage Qβ Ribonucleic Acid. (1971).

22. Arluison V, Derreumaux P, Allemand F, Folichon M, Hajnsdorf E, Régnier P. Structural Modelling of the Sm-like Protein Hfq from Escherichia coli. *J Mol Biol*. 320(4), 705–712 (2002).

23. Arluison V, Mura C, Guzman MR, *et al.* Three-dimensional structures of fibrillar Sm proteins: Hfq and other Sm-like proteins. 356, 86–96 (2006).

24. Patterson J, Mura C. Rapid Colorimetric Assays to Qualitatively Distinguish RNA and DNA in Biomolecular Samples. (2013).

25. Fadouloglou VE, Kokkinidis M, Glykos NM. Determination of protein oligomerization state: two approaches based on glutaraldehyde crosslinking. 373, 404–406 (2008).

26. Strong M, Sawaya MR, Wang S, Phillips M, Cascio D, Eisenberg D. Toward the structural genomics of complexes: Crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. 103, 8060–8065 (2006).

27. Møller T, Franch T, Højrup P, *et al.* Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol Cell*. 9(1), 23–30 (2002).

28. Vecerek B, Moll I, Bläsi U. Translational autocontrol of the Escherichia coli hfq RNA chaperone gene. 11, 976–984 (2005).

29. Moll I, Leitsch D, Steinhauser T, Blasi U. RNA chaperone activity of the Sm-like Hfq protein. *EMBO Rep*. 4, 284–289 (2003).

30. Geissmann TA, Touati D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*. 23(2), 396–405 (2004).

31. Sauer E, Weichenrieder O. Structural basis for RNA 3'-end recognition by Hfq. *Proc Natl Acad Sci USA*. 108(32), 13065–13070 (2011).

32. Link TM, Valentin-Hansen P, Brennan RG. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. *Proc Natl Acad Sci USA*. 106, 19292–19297 (2009).

33. Updegrove TB, Wartell RM. The influence of Escherichia coli Hfq mutations on RNA binding and sRNA•mRNA duplex formation in rpoS riboregulation. *Biochim Biophys Acta*. 1809(10), 532–540 (2011).

34. Warburg O, Christian W. Isolation and crystallization of the glycolytic enzyme enolase [Internet]. Biochem Zeitschrift Available from: http://scholar.google.com/scholar?q=related:lsg7XotXKxIJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5.

35. Murina V, Lekontseva N, Nikulin A. Hfq binds ribonucleotides in three different RNA-binding sites. *Acta Crystallogr D Biol Crystallogr*. 69(Pt 8), 1504–1513 (2013).

36. Sauer E, Schmidt S, Weichenrieder O. Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. *Proc Natl Acad Sci USA*. 109(24), 9396–9401 (2012).

37. Kajitani M, Kato A, Wada A, Inokuchi Y, Ishihama A. Regulation of the Escherichia coli hfq gene encoding the host factor for phage Q beta. *J Bacteriol*. 176(2), 531–534 (1994).

38. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. 65, 1–43 (2001).

39. Thoma R, Hennig M, Sterner R, Kirschner K. Structure and function of mutationally generated monomers of dimeric phosphoribosylanthranilate isomerase from Thermotoga maritima. 8, 265–276 (2000).

40. Klock HE, Koesema EJ, Knuth MW, Lesley SA. Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. 71, 982–994 (2008).

41. Sun X, Wartell RM. Escherichia coli Hfq binds A18 and DsrA domain II with similar 2:1 Hfq6/RNA stoichiometry using different surface sites. 45, 4875–4887 (2006).

42. Oefner PJ, Huber CG, Umlauft F, Berti GN, Stimpfl E, Bonn GK. High-resolution liquid chromatography of fluorescent dye-labeled nucleic acids. 223, 39–46 (1994).

43. Gesztelyi R, Zsuga J, Kemeny-Beke A, Varga B, Juhasz B, Tosaki A. The Hill equation and the origin of quantitative pharmacology. 66, 427–438.

44. Seber GAF, Wild CJ. Nonlinear Regression [Internet]. John Wiley & Sons, Inc., Hoboken, NJ, USA Available from: http://doi.wiley.com/10.1002/0471725315.

45. Hanes MS, Ratcliff K, Marqusee S, Handel TM. Protein-protein binding affinities by pulse proteolysis: application to TEM-1/BLIP protein complexes. 19, 1996–2000 (2010).

46. Hunke C, Antosch M, Müller V, Grüber G. Binding of subunit E into the A-B interface of the A(1)A(O) ATP synthase. 1808, 2111–2118 (2011).

47. McPherson A, Gavira JA. Introduction to protein crystallization. 70, 2–20 (2014).

48. Kabsch W. XDS. 66, 125–132 (2010).

49. Adams PD, Afonine PV, Bunkóczi G, *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. 66, 213–221 (2010).

50. Matthews BW. Solvent content of protein crystals. *J Mol Biol*. 33(2), 491–497 (1968).

51. Winn MD, Ballard CC, Cowtan KD, *et al.* Overview of the CCP4 suite and current developments. 67, 235–242 (2011).

52. Tong L, Rossmann MG. Rotation function calculations with GLRF program. 276, 594–611 (1997).

53. Bunkóczi G, Read RJ. Improvement of molecular-replacement models with Sculptor. *Acta Crystallogr D Biol Crystallogr*. 67(Pt 4), 303–312 (2011).

54. Murshudov GN, Skubák P, Lebedev AA, *et al.* REFMAC5 for the refinement of macromolecular crystal structures. 67, 355–367 (2011).

55. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. 60, 2126–2132 (2004).