Structural and functional studies of the Hfq family of ancient bacterial RNA-binding proteins

Kimberly Ann Stanek
Greece, NY

Bachelor of Science in Biochemistry, University at Buffalo, 2012
Bachelor of Arts in Chemistry, University at Buffalo, 2012

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Chemistry

University of Virginia
May, 2018

Acknowledgements

The conclusion of my graduate degree would never have been possible without the encouragement and support of all those around me. First and foremost I would like to thank my advisor, Cameron Mura, who introduced me to field of structural biology and afforded me the ability to engage in research through a myriad of techniques, both computational and at the lab bench. I am grateful for his insight and guidance, both scientifically and professionally, during my time as a graduate student. I would also like to thank all of my committee members—Don Hunt, Wladek Minor, Martin Wu, Phil Bourne, and Linda Columbus—for the dedication of their time and their helpful feedback.

I would like to thank members of the Mura lab, both past and present. Specifically, I thank Jennifer Patterson-West, for helping me get started as a new graduate student, mentoring me in the lab, and maintaining a lasting friendship. I thank Peter Randolph for sharing in his crystallographic expertise and being a constant source for helpful discussion and troubleshooting. I thank Charles McAnany for always willing to help me with any computational aspects of my work; I am grateful for our many discussions, scientific and otherwise. I also thank the many undergraduates that I have had the privilege of mentoring or working collaboratively with in my time here—Sebastian Coupe, Berk Ekmekci, Sooraj Anchar, and Jane Nguyen.

I would also like to thank Linda Columbus and members of her lab—Jennifer Martin, Marissa Kieber, Nicole Swope, Steven Keller, Jason Kuhn, and Tracy Kieber—for providing helpful discussion and feedback on many of my presentations, posters, and experiments over the years. I am so grateful for the close relationships between our two labs and the collaborative environment that has been fostered. I also thank Carol Price for her guidance, both scientifically, and personally. I appreciate of all the support and patient listening she has provided throughout the years.

I would also like to thank all of our collaborators, and anyone that has helped me scientifically in my time as a graduate student. Specifically, I thank Mike Saweya and Duilio Cascio for allowing us to

accompany them on many trips to APS, and for endowing me with the essential skills necessary to succeed as a crystallographer. I am so grateful for all of our contact throughout the years. I also thank all of the staff at SER-CAT and NE-CAT, for their technical assistance at all hours of the day. I would also like to thank my undergraduate advisors at Buffalo—Gail Willsky and Michael Garrick—for encouraging me to pursue my doctorate at the University of Virginia.

Finally, I would like to thank all of my friends and family that have stood by me in support throughout the years. I appreciate all of your understanding and patience during this journey. I thank my parents for always cheering me on and being supportive in everything that I do. I also thank Chris DeRosa for his meaningful friendship. Finally, I would like to thank Mike Hague, for continuously proving to be a positive force in my life, and for helping me in more ways than I can count.

Abstract

The bacterial host factor Hfq is an RNA-binding protein that facilitates the interaction of mRNAs with small regulatory RNAs (sRNAs) Hfq self-assembles as hexameric toroid and functions by simultaneously binding U-rich regions of sRNA on one face and A-rich regions of mRNA on the other face. More recently a third site on the lateral rim of Hfq has been implicated in binding RNA, though little is known about the molecular details of RNA-binding at this site. Due to Hfq's broad functional role in binding RNAs, it has been demonstrated to be required for numerous physiological pathways, including stress response, quorum sensing, and biofilm formation. Through structural similarities, Hfq has been identified as the bacterial branch of the Sm superfamily of proteins. While bacteria typically have one Hfq homolog, eukaryotes encode multiple Sm paralogs, which oligomerize as heteroheptamers through a complex chaperoned assembly pathway. The eukaryotic Sm and Sm-like (LSm) proteins are involved in splicing and various other mRNA processing pathways.

Several species of bacteria have been identified as having two putative Hfq paralogs. Such proteins could provide valuable insight into the evolutionary transition from bacterial Hfq to eukaryotic Sm. Currently though, little is known about the physiological role of additional Hfq paralogs and no structural information is available. This work presents a new phylogenetic analysis of Hfq sequences, demonstrating the presence of 2 (or more) Hfq paralogs in bacteria of diverse lineages including the deep-branching extremophilic Aquificae. Here, the structures of two such paralogs, from *Aquifex aeolicus*, have been determined to atomic resolution, offering the first instance of two genuine Hfq proteins from a single species. Atomic-level details of the conserved lateral rim site have also been revealed through co-crystallization of Hfq1 with U_6 RNA. Intriguingly this lateral rim site is not conserved in Hfq2, which was also found to bind RNAs with pH dependence and co-purify with endogenous DNA. These results suggest that Hfq1 and Hfq2 function independent, likely in different cellular pathways.

A single *hfq* open reading frame with two putative linked Sm domains (Hfq_N- Hfq_C) has also been identified in the α-proteobacterium *Novosphingobium aromaticivorans*. Here *N. aromaticivorans* (*Nar*) Hfq has been expressed, purified and biochemically characterized. The oligomeric state of *Nar* Hfq in solution is a trimer, likely assembling as a ring of alternating Hfq_N and Hfq_C subunits. This represents the first known instance of "pseudo-heteromeric" Hfq, which poses intriguing consequences for differences in RNA-binding affinities. Initial efforts to crystallize *Nar* Hfq were hindered by the presence of an ~40 residue proline-rich N-terminal tail which appears to increase the conformational heterogeneity of the *Nar* Hfq crystals. Ongoing efforts are currently aimed at crystallization of ΔN-*Nar* Hfq construct.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Chapter 1: The bacterial RNA chaperone Hfq

Kimberly Stanek

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

1 Overview

The central dogma of molecular biology describes the fundamental processes by which genetic information flows: deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA) and RNA is translated into protein [1]. In this simplified model, DNA acts as the carrier of hereditary information, proteins as the functional unit of the living cell, and RNA as an intermediary between the two. RNA has been shown to be an extremely diverse molecule, with functions ranging from being a primary information carrier itself [2] to an enzymatic catalyst [3]. Due to its myriad functional roles and the ability of RNA to self-replicate, it has long been suspected that RNA-based organisms existed prior to modern life [4,5]. In what is known as the "RNA World hypothesis", these RNA-based lifeforms replicated in the absence of catalytic proteins, and were perhaps the first living entities [6,7].

Even after the incorporation of DNA and protein as functional biomolecules, RNA has continued to be an indispensable molecule in the cell. In addition to its dominant role in protein translation, both as the mRNA template and as the core catalytic component of the ribosome [8], RNA has more recently been shown to have a multitude of other regulatory roles. Indeed, an entire field of small, non-coding RNA biology has developed within the past twenty years.

2 Small regulatory RNAs in Bacteria

Small RNA, or sRNA, refers to a class of regulatory RNAs that are generally ~20-300 nucleotides in length, and found in bacteria [9,10]. These sRNAs modulate post-transcriptional regulation through a variety of mechanisms, enabling bacteria to efficiently and precisely tune protein expression, which in turn allows for rapid response to environmental signals. Examples of these crucial pathways include (i) response to environmental stresses, such as changes in temperature or pH, (ii) quorum sensing, which is essentially a method of cell-cell communication, (iii) regulation of metabolic pathways in response to environmental metabolite concentrations, and (iv) expression of virulence factors during pathogen-host interactions. Each of these cellular pathways relies upon an sRNA-based regulatory circuit [9].

2.1 Mechanisms of sRNA action

Most known sRNAs function by base pairing to one (or more) mRNA sequences and affecting translation of the mRNA target. It should be noted that other forms of sRNA are also present in bacteria; for example the *Eshcerichia coli* 6S RNA regulates transcription by mimicking a DNA promoter open complex and binding directly to RNA polymerase [11]. For the sake of simplicity, this work will focus on the former type. There are two major types of sRNAs which bind mRNA. The first, known as *cis*-encoded or antisense sRNAs, are transcribed from the DNA strand opposite the mRNA template, and thus have perfect base-pairing complementarity to their target. The second are *trans*-encoded sRNAs, which are transcribed from a different genomic region, and thus generally have only partial sequence complementarity.

Antisense sRNAs have been best characterized for their role in the downregulation of plasmid-encoded toxin proteins [12,13]. This effect can occur through base pairing of the sRNA to the 5' end, 3' end, or entire sequence of the target mRNA encoding the toxin, and subsequent inhibition of translation. sRNAs acting in *trans* are typically expressed under specific growth conditions and have been shown to function through a variety of mechanisms. These sRNAs have been shown not only to downregulate [14], but also upregulate [15,16] their mRNA targets, with the mode of action depending on the specific mRNA-sRNA pairing. *Trans*-acting sRNAs typically feature limited base-pairing complementarity, and they generally bind to a seed region that is typically 6-8 nucleotides in length at the 5'-end of the mRNA [17]. This type of mechanism also allows for one sRNA to target multiple mRNAs, which can vary greatly in sequence and structure. The sRNA RyhB for example, which is expressed under iron-limiting conditions, has a plethora of mRNA targets, some of which are downregulated and others of which are upregulated [18,19]. Thus the precise function and cellular effect of a given sRNA is highly context-dependent.

Conversely, a single mRNA can be targeted by multiple sRNAs. One example of this is the *rpoS* mRNA, which encodes the sigma S (σ^S) factor, a crucial transcriptional regulator during the stationary growth phase. The OxyS sRNA, expressed under conditions of oxidative stress, downregulates expression of *rpoS* mRNA by binding near the ribosomal binding site (RBS) and thus preventing translation of σ^s [20]. Several other sRNAs—DsrA (induced at low temperature), RprA (induced during cell surface stress), and ArcZ (repressed during anaerobic conditions)—all upregulate σ^S expression by inducing a conformational change upon binding *rpoS*. Specifically, this conformational change opens an inhibitory hairpin present on the 5'-leader region of *rpo*S [15,21,22].

Binding of an sRNA can also affect the cellular lifetime of mRNAs, as mRNA that is not being actively translated will be susceptible to enzymatic degradation. RNase E is one such ubiquitous enzyme that mediates much of a bacterium's mRNA decay process. Through inhibition or enhancement of ribosomal binding, sRNAs can increase or decrease the degradation of their mRNA target (respectively) as a secondary effect [18,23]. In some cases, sRNAs may specifically recruit RNase E as their 5'-monophosphate groups are the preferred regulation elements for RNase E; this also positions the enzyme near the imminent cleavage site of the mRNA [24]. This mechanistic feature also explains why, in most known instances, the sRNA is degraded along with the mRNA target [25].

2.2 Protein chaperones of sRNAs

In many cases the annealing of sRNA-mRNA duplexes additionally require a protein chaperone for productive pairing. This requirement is tied to several factors, including the relative abundance of each RNA species, the degree of base-pairing complementarity, and the presence of inhibitory structural elements that would otherwise preclude annealing [26]. The most well-studied of these RNA chaperones, a protein known as Hfq, is required for the successful pairing of a large subset of *trans*-acting sRNAs with their target mRNAs [27–29]. Hfq, which is described in detail below, binds sRNA and mRNA simultaneously, increasing the local concentration of both species. The protein may also restructure, or remodel the RNAs, affecting stability, or may play a role in the recruitment of RNase E.

Initially, it was expected that only *trans*-acting sRNAs, with limited base-pairing specificity, would require a protein chaperone. However, in more recent years the FinO family of proteins has emerged as a chaperone of *cis*-encoding sRNAs [30,31]. FinO was first identified as a plasmid-encoded protein [32], but additional, chromosomally-encoded homologs have

been characterized since [33]. One FinO protein, known as ProQ, has been shown to bind to *trans*-acting sRNAs as well *cis*-acting, including some overlap with Hfq targets [34]. Conversely, Hfq has recently been shown to function as a chaperone for the *cis*-encoded RNA-IN/RNA-OUT pair, involved in the Tn10/IS10 transposition system [35]. Needless to say, the complex interplay between multiple sRNAs, mRNAs, and protein chaperones allows for finely-tuned bacterial regulation. The mechanistic and structural foundations of these processes remains largely enigmatic.

3 Hfq as a central hub in sRNA-based regulatory networks

This thesis focuses on the bacterial RNA chaperone Hfq. Hfq was initially identified as a host factor required for the replication of bacteriophage Qβ in *E. coli* [36]; later studies, during the 1990s, showed that knocking out the *hfq* gene resulted in pleiotropic phenotypic effects (decreased growth rate, increased ultraviolet sensitivity, etc.), indicating that this protein might have a much more extensive regulatory role than was originally expected [37]. The regulatory function of Hfq became evident when it was shown to be required for translation of the broadly regulatory σ^s factor from the *rpoS* mRNA [38]. Since the mid-90s, an increasing number of new sRNAs have been identified through genome-wide analyses (both experimentally and computationally), and a large subset of these sRNAs have been further shown to interact with Hfq in many bacterial species [39–41]. Hfq has now been linked to such various pathways as stress response [42–44], quorum sensing [45], biofilm formation [46], and expression of virulence factors [44,47,48], and the Hfq “master regulator” protein is considered a central hub for post-transcriptional regulation in bacteria.

3.1 The RNA-binding landscape of Hfq

The Hfq protein is rather small (typically 80-100 amino acids), consisting of an N-terminal α-helix followed by five highly-curved antiparallel β-strands [49,50] (Fig 1A). However, the protein self-assembles as a stable, toroidal hexamer (Fig 1B), greatly increasing the potential surface area for binding RNA. One face of the Hfq ring, termed *proximal* (with respect to the α-helix), binds a stretch of up to six uridines in a cyclic fashion around the ring pore [49,51] (Fig 2A). In these six equivalent proximal-binding pockets, the uracil bases intercalate between two highly conserved Phe42 side chains, one from each of two adjacent monomers. This proximal pore of Hfq preferably binds the U-rich 3'-ends of sRNAs; a highly conserved His57 residue is further selective for the hydroxyl group present on the 3'-end of these sRNAs as a result of rho-independent transcription termination [51,52].

The opposite face of the Hfq ring, known as the *distal* face, binds adenosine-rich sequences via a mechanism that varies between gram-negative and gram-positive bacteria. In gram-negative bacteria, the distal face of Hfq recognizes a tripartite $(AAN)_n$ motif (Fig 2B), where A is adenosine and N is non-specific [53,54]. The first adenine in this motif inserts in a pocket formed between *β*2 and *β*4 strands of one Hfq monomer, while the second adenine π-stacks with between conserved Tyr25 residues found on the *β*2 strands of adjacent subunits. The third nucleotide base is flipped away from the Hfq surface and does not make contact with the protein, acting as a bridge between adjacent distal site pockets. In this way, gram-negative Hfq can accommodate a sequences of up to 18 nucleotides. In contrast, gram-positive Hfq recognizes a bipartite $(AN)_n$ motif (Fig 2C), with a single adenine-binding pocket similar to the second A site of gram-negative Hfq, and a non-specific linker (thus accommodating sequences of up to 12 nucleotides) [55]. *In vivo*, the distal face is thought to bind A-rich regions of mRNAs.

Such Hfq-binding regions have been identified in the 5'-untranslated regions (5'-UTRs) and 3'-poly(A) tails of mRNAs [56,57].

3.2 The stoichiometry of Hfq and RNA: mechanistic implications

The proximal and distal faces of Hfq can bind RNA independently and simultaneously [58–60]. This alludes to a simple mechanism whereby Hfq binds an sRNA on its proximal face and an mRNA on its distal face, and—functioning as a chaperone—allows the two species to productively base-pair (Fig 2D). Realistically, the structural and dynamical features of this process are much more complex; the detailed mechanism of Hfq action likely varies as well, depending on the sRNA-mRNA pairing involved. Remarkably, there is still debate over the prevalence and relevance of the ternary sRNA:Hfq:mRNA complex *in vivo* as other stoichiometries, including 2:1 Hfq_6:RNA have also been found [59–62]. One consideration however, is that these stoichiometries, observed *in vitro* may not occur under physiologically relevant concentrations of Hfq and RNA. Studies in *E. coli* show that Hfq is a highly expressed protein, with ~10,000 copies of Hfq_6 per cell during exponential growth phase [63]. Concentrations of sRNAs will vary greatly depending on environmental conditions, for example OxyS is found at ~4,500 copies per cell during oxygen stress [64,65]. Due to the large pool of Hfq binding partners (in terms of variety and sample concentration), the concentrations of Hfq are most likely limiting *in vivo.* Accordingly, a mechanism whereby RNAs cycle on and off the surface of Hfq is thought to occur [66].

Higher-order oligomers of Hfq, due to stacking of the hexameric rings, have also been observed via a number of methods [67–69]. Hfq dodecamers have been crystallized in all possible orientations: proximal-to-proximal stacking is seen with Cyanobacterial *Synechosystis* Hfq [70], a distal-to-distal interface is observed with *Staphylococcus aureus* Hfq [49], and

proximal-to-distal stacking is found in both the *apo* (unbound) and RNA-bound forms of *Listeria monocytogenes* Hfq [71]. While the role of such oligomers *in vivo* is still unclear, it is worth noting that these differently-oriented dodecamers could conceivably alter RNA-binding affinities (and therefore physiological functionalities) by occluding certain RNA-binding surfaces. More recently, a structure of *E. coli* Hfq with A_{18} bound on the distal face revealed that additional base-stacking interactions between the N-site bases of the 'AAN' motif of neighboring A_{18} molecules resulted in a supramolecular $(Hfq_6)_2$:$(A_{18})_2$ complex [72]. This structure corroborates (and may explain) previous reports of Hfq dodecamers formed in the presence of RNA [61,69].

In addition to the relatively straightforward binding of unstructured regions of single-stranded RNA, Hfq has been shown to remodel secondary structural elements, for both mRNAs and sRNAs [35,73,74]. These structural changes can also affect the thermodynamic stability of the RNA (e.g. melting temperature), suggesting a rather direct role for Hfq in RNA regulation. For example, OxyS and RprA, which were both found to undergo structural rearrangements upon Hfq binding, were also made more susceptible to (OxyS) or protected from (RprA) degradation by RNase E as a result of Hfq being present [74]. This suggests that Hfq can function by binding a single RNA and affecting its stability. sRNAs may also act to sequester Hfq, limiting the availability of the protein for binding to and favorably restructuring mRNA targets [75].

3.3 Evidence for additional RNA-binding surface on Hfq

Recent studies have suggested that Hfq must have additional surfaces for binding RNA, beyond the proximal and distal regions described above. This idea is based on the finding that sRNAs were able to bind Hfq even when accessibility to the proximal site was blocked [51]. Furthermore, the presence of additional internal U-rich regions are sometimes required for the

efficient binding of sRNAs to Hfq [76]. Mutational analysis indicated that several residues on the outer rim, including Phe39 and Arg16, were important for the binding of sRNA [61,77]. A stretch of arginines, found on the N-terminal α-helix, has been defined as the arginine-rich patch, or *lateral* rim of Hfq, and includes Arg16, Arg17, and Arg19 in *E. coli* (Fig 3). This lateral rim has been revealed to be important for facilitating annealing between sRNA and mRNA and providing a platform for RNAs to cycle on and off of Hfq [77,78] . Mutation of the arginine-rich patch showed that, while Hfq was still able to bind RNAs at the proximal and distal sites, release of the resulting dsRNA duplex was impaired [78].

The exact mechanism by which Hfq binds RNA at the lateral rim is still unclear. Some findings demonstrate a structural preference for UA-rich sequences [79] while others suggest that the site does not exhibit nucleotide specificity, but rather that the arginines interact non-specifically with RNA through favorable electrostatic interactions with the phosphate backbone [77]. A recent structure of *E. coli* Hfq in complex with a RydC sRNA (Fig 3) suggests the importance of a pocket formed by Asn14, Arg16, Arg17, and Phe39 that binds two nucleotides of uridine [80]. This structure elucidates the previous finding that Phe39 is also important for binding sRNA at the lateral rim. Interestingly, Arg19 does not actively engage with the RNA in this co-crystal structure. Conceivably, the lateral rim could consist of a pre-formed binding pocket in addition to the arginine-rich patch, which could serve to guide RNA towards that binding pocket. Whether such a binding pocket would (or should) exhibit any nucleotide specificity is unclear.

3.4 Class I and Class II sRNAs bind Hfq via distinct mechanisms

With the definition of the lateral rim as a third RNA-binding surface on Hfq, a new functional mechanism has been proposed. In this model, mRNA binds to the distal face of Hfq,

while sRNA interacts with the proximal and lateral surfaces. However, as more sRNA-mRNA pairings were studied in the context of proximal, distal, and lateral Hfq-binding, this configuration did not always appear to hold, and a further categorization of Hfq-binding sRNAs as 'Class I' or 'Class II' in *E. coli* was proposed [81,82]. Typically, class I sRNAs act as emergency responders (in response to environmental cues, metabolite levels, etc.), whereas class II sRNAs are more stable in the cell and generally act as silencers. Class I sRNAs, including DsrA and ArcZ, have internal UA-rich sequences in addition to their U-rich 3'-ends, and bind the proximal and lateral sites of Hfq. The mRNA targets of class I sRNAs (in this case *rpoS*) have the classic 'AAN' motif that binds to the distal face of Hfq.

Class II sRNAs, on the other hand, have internal 'AAN' motifs and interact instead with the proximal and distal faces of Hfq. Conversely, the mRNA targets of class II sRNAs contain UA-rich sequences that bind to the lateral face. Class II sRNAs include ChiX, which targets *chiP* mRNA, encoding a chitoporin sugar transporter. ChiX acts as a silencer of *chiP* by pairing to the transcript and, intriguingly, an antisense ChiX RNA known as *chB* has been shown to further bind and destabilize ChiX. In this system ChiX is therefore both an effector and target of post-transcriptional regulation [83]. In general, class II sRNAs appear to have a higher stability, and may be involved in multiple rounds of pairing with mRNAs. However, as has become apparent with many aspects of Hfq-sRNA biology, strict distinctions between these class I and class II sRNAs are not always straightforward, and certain sRNAs, including OxyS, appear to have a more complex behavior that falls somewhere in between. Likewise, some mRNAs have been found to be targeted by both class I and class II sRNAs [82].

3.5 Beyond the Hfq core: the role of the C-terminal tail

In addition to the ~60 amino acid structural core of Hfq, the protein has a C-terminal tail extension, which varies in length from just a few residues [84], to ~100 in members of the γ-proteobacterial *Moraxellaceae* [85,86]. This region is sometimes referred to as the C-terminal 'domain' of Hfq. However, the absence of clear electron density for the C-terminal tail in crystal structures, coupled with analysis of the overall structural envelope of *E. coli* Hfq through small angle x-ray scattering (SAXS), indicate that this region is unstructured [87]. Thus "C-terminal region" (CTR) is the terminology that will be used here. The general functional importance of the CTR is still debated, and studies of *E. coli* Hfq (with a CTR of ~40 residues) have produced conflicting results on whether or not the tail is required for efficient binding of RNAs [88–90]. Subsequent studies have revealed that likely only a subset of sRNAs, such as those with longer sequences, require the CTR [87].

While the sequences of the CTRs are not well-conserved among Hfq homologs, they tend to be enriched for acidic residues, particularly glutamate (although other variations are found, such as the glycine-rich tails of the *Moraxallaceae*); thus the tails are thought to have a function unrelated to direct binding of negatively-charged nucleic acid. Recently, the *E. coli* CTR has been shown to self-associate with the lateral rim of Hfq, as well as compete for RNA-binding at the lateral site [91,92]. These findings suggest that the CTR could serve as an auto-regulator of Hfq, by *(i)* increasing the rate at which sRNAs and mRNAs are productively paired, *(ii)* facilitating the release of the sRNA-mRNA duplex after annealing, and *(iii)* preventing the non-specific binding of nucleic acid.

4 Hfq as seen through a structural lens

To fully understand the mechanism by which Hfq functions, molecular details of the protein and its complexes with RNAs are necessary. Low resolution methods, such as atomic force microscopy (AFM), circular dichroism (CD), and small-angle X-ray scattering (SAXS) can provide information on the overall shape and stoichiometry of Hfq:RNA assemblies. For atomic-level resolution detail, the prominent methodologies have been nuclear magnetic resonance spectroscopy (NMR) [93], X-ray crystallography [94], and, more recently, cryo-electron microscopy (cryo-EM) [95]. Of these, approaches, the Hfq protein (with a molecular weight of ~60 kDa for the hexamer) is best suited to structural characterization via X-ray crystallography; thus far, all of the available high-resolution Hfq structures have been determined using this method.

Since the early 2000s, a multitude of Hfq structures, representing homologs from 11 different bacterial species and one archaeon, have been determined, along with several co-crystal structures in complex with short strands (~5-20 nucleotides in length) of RNA [96]. This expanding database of Hfq and Hfq-RNA structures offers a wealth of insight into Hfq function, including details about Hfq oligomerization, the precise definition of the proximal, distal, and lateral RNA-binding pockets, and the substrate specificity at each of these sites. *E. coli* Hfq was recently co-crystallized with the sRNA RydC, revealing how simultaneous binding at the proximal and lateral sites is achieved [80]. In general, sRNAs are difficult to crystallize, as they are not as structurally homogeneous as proteins; however the compact pseudoknot structure of RydC, coupled with its role in forming crystal contacts, lent itself to successful crystallization.

In cases where RNAs are longer and less structured, useful information can still be gained by co-crystallizing Hfq with short oligonucleotides thought to be representative of Hfq-binding sequences, such as those obtained through systematic evolution of ligands by exponential enrichment (SELEX). For example *Someya et al.* determined the structure of Hfq from gram-positive *Bacillus subtilis* bound to SELEX-derived $(AG)_3A$ RNA [97]. There are currently no high-resolution Hfq-RNA complexes determined via cryo-EM, though low-resolution data on the Hfq ring have been obtained [98,99]. Cryo-EM has been highly successful with the eukaryotic homologs of Hfq, know as Sm proteins (discussed below), which require RNA in order to form stable oligomers; thus, this approach may also be possible for larger Hfq complexes (for example, an sRNA:Hfq:mRNA ternary complex).

4.1 Hfq belongs to the Sm protein superfamily

Hfq was initially identified in 1968 [36], but it wasn't until the early 2000s, through structural characterization, that the protein was revealed to be the bacterial member of the Sm superfamily of proteins [84,98,100]. The Sm proteins were discovered in eukaryotes in the 1970s by Joan Steitz [101], where they act as scaffolds in mRNA splicing and in other RNA processing pathways [102,103]. Sm archaeal proteins (SmAPs) were established some 20 years later, where their precise physiological function is still unclear [104]. The homology between Hfq, SmAP, and Sm initially came as a surprise. Sm proteins were not expected to be found in bacteria or archaea, as both lack the required spliceosomal machinery. Nevertheless, as revealed by the first crystallographic structure of an Hfq [49], the 3D fold is extremely well-conserved between bacteria and eukarya (1.2 Å rmsd for *S. aureus* Hfq and human SmD3).

There are, however, some key structural differences between the bacterial Hfq, the Sm archaeal SmAPs, and the eukaryotic Sm and Sm-like (LSm) proteins (Fig 4). Most bacterial species have one copy of Hfq, which spontaneously self-assembles as a hexamer. Archaeal species have between one and three paralogs, which have been shown to self-associate into stable hexamers [105], heptamers [106], and octamers (Randolph & Mura, personal communication). Eukaryotes have many (>20) paralogs of Sm and LSm proteins, that are believed to have originated through multiple gene duplication events [107]. Through a complex pathway involving several protein chaperones, the Sm and LSm proteins are assembled as heteroheptamers [108]. The function of each Sm ring is in turn determined by the monomeric subunits which comprise it. For example, in *Saccharomyces cerevisiae*, a ring comprised of LSm subunits 2-8 localizes to the nucleus where it associates with the U6 small nuclear RNA, forming the U6 small nuclear ribonucleoprotein (U6 snRNP) spliceosomal complex [109]. In contrast, the LSm1-7 ring, which differs by just one subunit, localizes to the cytoplasm, where it instead plays a role in mRNA degradation [110].

The shift in oligomeric state, from hexamer to heptamer, corresponds to a major difference in the mechanism of RNA-binding as well. While the amino acids that define the proximal site are largely conserved between Hfq and Sm, the pore of the Sm heptamer is large enough that RNA may thread through it, as is seen in the structures of U snRNPs [111,112]. The distal and lateral modes of binding RNA appear to be absent from the eukaryotic Sm proteins. It is still unclear whether the archaeal SmAPs are more “Hfq-like” or “Sm-like” in terms of cellular function and RNA-binding [104]. However, crystal structures of *Pyrococcus abyssi* SmAP in complex with oligonucleotides show that the proximal as well as lateral modes of RNA-binding are conserved between SmAP and Hfq [106,113]. While SmAPs are capable of

forming heptamers and octamers (as well as higher order species such as 14-mers), it is unclear whether single-stranded RNA is able to thread through the pore of the ring [114].

4.2 The Sm fold is that of a small β-barrel

Presumably, a close examination of the Sm fold would give some insight into the origins of the broad structural and functional plasticity described above. The Sm fold belongs to a large class of evolutionarily ancient and functionally diverse proteins known as small β-barrels (SBBs). The SBB domain is, as the name implies, rather small (≤100 residues), and consists of five to six β-strands arranged as two closely-packed, roughly orthogonal β-sheets that exhibit two-fold rotational pseudo-symmetry (Fig 5). Although the Sm domain consists of five β-strands, the elongated β2 strand features a highly conserved glycine residue that enables severe bending of the strand, essentially segmenting it. This results in one β-sheet comprised of β2C, β3 and β4, and the other sheet consisting of strands β5, β1 and β2N. The SBB is an extremely modular structural unit, and additional structural motifs, such as the N-terminal α-helix, or L4 loop insertion of the Sm proteins, are also commonly seen (Fig 5A), as well as tandem and mixed domains. Furthermore, SBBs exhibit a strong propensity to form higher-order oligomers, for example, the toroidal Sm proteins and pentameric OB-fold proteins.

Considered across all SBB proteins, this domain is unique in that it exhibits very limited sequence restraints. Analysis of a broad range of SBB sequences reveals the conservation of only ~7 hydrophobic residues that form the SBB core (Youkharibache *et. al.,* in revision). Consequently, SBBs are found in a plethora of physiological roles with drastically different functions. These include: *(i)* sRNA-based regulation by Hfq [27], *(ii)* the spliceosomal scaffolding and RNA processing functionalities of Sm rings [102,103], *(iii)* binding of single and double-stranded DNA by OB-fold proteins [115], *(iv)* SH3 domains that recognize poly-proline

motifs in signal transduction and epigenetic regulation [116], *(v)* tandem OB ad SH3 folds present in the ribosomal protein L2 [117], and *(vi)* membrane transport by the Sm-containing MscS channel [118].

The SBB is thought to be a very ancient fold, as it occurs in such essential cellular components as the ribosome [119]. One open question is whether the fold has undergone convergent or divergent evolution (or both). SBBs with very different sequences and even topologies can nevertheless provide extremely similar 3D structural platforms for nucleic acid interactions. For example, the shiga-like toxin from *E. coli* adopts an OB-fold and has extremely high structural similarity with Hfq (0.625 Å RMSD, Fig 5B,C). However the strand order between OB and Hfq is permuted, such that the N-terminal α-helix of Hfq sits instead between strands β3 and β4 for OB. One phylogenetic study of SBB sequences did find that protein function was in fact more closely related to protein sequence than to structural similarity [120]. Intriguingly, a recent structure of ProQ (Fig 5D) revealed an N-terminal domain that is similar to the SBB class of Tudor domains [121], thus linking the two bacterial RNA chaperones, Hfq and ProQ, as members of the SBB structural family.

5 Bridging knowledge gaps in Hfq function

Much of our knowledge of Hfq derives from the study of enterobacterial homologs, such as *E. coli*, although putative *hfq* genes have been identified in at least 50% of bacterial genomes [84]. As an identifiable Hfq does not appear to be ubiquitous in all bacterial species, this could indicate that the gene either had a common ancestral origin and was subsequently lost in certain bacterial clades (e.g phyla Chlamydia and Actinomycetes), or that it was acquired through horizontal gene transfer. However, due to the extremely divergent sequences of Hfq

homologs (and SBBs in general), it is also possible that there are additional genes that have yet to be identified through sequence homology. Such was the case with cyanobacteria, which initially were thought to be missing an *hfq* gene [84]. Putative Hfq gene sequences were later identified in two cyanobacterial species, *Synechocystis* and *Anabaena*, and were then confirmed structurally as Hfq proteins through X-ray crystallography [70].

As the sequences of Hfq homologs vary greatly, we might expect the mechanistic action of the protein to differ as well between species. *Synechocystis* and *Anabaena* Hfq for example, lack the conserved 'YKHAI' motif associated with U-rich RNA binding at the proximal site, and bind RNA with only weak affinity *in vitro* [70]. Furthermore, they did not complement an *E. coli Δhfq* knockout *in vivo*. The requirement of Hfq for post-transcriptional regulation in gram-positive bacteria is still debated as well. For example, *Staphylococcus aureus* Hfq, is expressed at low cellular levels, and while it co-purifies with an sRNA known as RNAIII, it does not appear to be required for the annealing of RNAIII with its mRNA targets [122]. Furthermore, knockout of the *S. aureus hfq* gene has no apparent phenotypic effect [123]. Several explanations have been proposed for these observations: *(i)* the requirement for Hfq may be relieved in bacteria with a low-GC content and/or a more compact genome (e.g. *S, aureus*, with a GC content of ~33%), or *(ii)* there is another RNA chaperone that fulfills this function [26]. Still, the question remains: what is Hfq doing in these species if not acting as a chaperone?

Intriguingly, several species of bacteria have been found with at least two putative *hfq* genes. Recent studies have verified the presence of multiple Hfq paralogs in *Burkholderia cenocepacia* [124,125] and *Bacillus anthracis* [126,127]. Studies in *B. cenocepacia* show that the bacterium has two authentic Hfq paralogs, one of which is more highly expressed during the log growth phase, while the other is preferentially expressed during stationary phase. Both proteins were also shown to affect *B. cenocepacia* virulence. *Bacillus anthracis* has three copies

of the *hfq* gene, two of which are expressed chromosomally and the third which is found on a plasmid. Only two of the *B. anthracis* Hfqs were able to partially restore function in an *E. coli* Δ*hfq* knockout, and the role of the third Hfq is still unclear. None of the paralogs were shown to associate with one another, and it is likely that they have distinct functional roles in the cell.

6 Objectives of this work

This thesis seeks to identify and characterize Hfq orthologs and paralogs from a diverse range of bacterial species, in order to *(i)* fill existing knowledge gaps and discrepancies in Hfq function, and *(ii)* to better understand how the evolutionary transition from the relatively simple, homomeric bacterial Hfq chaperone to the intricate, heteromeric eukaryotic Sm scaffold might have occurred. The first objective of this work has been to perform a new bioinformatic survey of the Hfq family of proteins, with the intent of identifying new Hfq homologs. New Hfq orthologs have been identified in the phylum Actinobacteria, which were previously thought to lack an Hfq; we find that these genes were likely acquired through lateral gene transfer. Multiple Hfq paralogs (two or more) have also been identified in the phylum Aquificales and the γ-proteobacterial order Aeromonadales. Lateral gene transfer and copy number variation were also observed with different strains of gram-positive pathogen *Streptococcus pneumoniae*, hinting at a potential role for the *hfq* gene in virulence pathways.

The second objective of this work was to structurally and functionally characterize the two Hfq paralogs found in thermophilic bacterium *Aquifex aeolicus*, a member of the deep-branching Aquificales clade. The determination of the structures of both paralogs (Hfq1 and Hfq2) represents the first structural characterization of two Hfqs from the same species. Hfq1 was co-crystallized with U_6 RNA in the lateral-binding pocket, demonstrating that this mode

of binding is deeply conserved among bacterial species. At 1.5 Å resolution, this structure also provides an unprecedented level of detail about the sequence specificity and mechanism of RNA recognition at this binding site. While the structures of Hfq1 and Hfq2 are highly similar (to within a 0.93 Å RMSD between the hexameric rings of Hfq1 of Hfq2), Hfq2 demonstrates altered RNA-binding affinities, including dependence on pH and the ability to co-purify with endogenous (cellular) DNA.

The final objective of this work has involved characterization of Hfq from the soil-dwelling γ-proteobacterial species *Novosphingobium aromaticivorans.* This homolog is unique in that in codes for two tandem Sm domains separated by a short linker, with a domain organization that we denote Hfq_N-Hfq_C. *N. aromaticivorans* Hfq appears as a trimer in solution, suggesting a pseudo-hexameric ring of alternating Hfq_N and Hfq_C domains. Initial efforts to obtain high-resolution X-ray diffraction data were hindered by crystalline disorder, presumably due to an ~40 residue proline-rich N-terminal tail. Proline-rich regions are a common protein structural motif, and often play a role in signaling pathways [128]. They are also the preferred binding partners of SH3 domains, which is intriguing, as both the Sm and SH3 folds are instances of small β-barrels. Future efforts will be aimed at crystallizing a ΔN-Hfq construct in order to determine the crystal structure of this unique Hfq ortholog.

References

1. Crick F. Central dogma of molecular biology. Nature. 1970;227: 561–563.
2. Drake JW, Holland JJ. Mutation rates among RNA viruses. Proc Natl Acad Sci U S A. 1999;96: 13910–13913.
3. Cech TR. Ribozymes, the first 20 years. Biochem Soc Trans. 2002;30: 1162–1166.
4. Orgel LE. Evolution of the genetic apparatus. J Mol Biol. 1968;38: 381–393.
5. Crick FH. The origin of the genetic code. J Mol Biol. 1968;38: 367–379.
6. Atkins JF, Gesteland RF, Cech TR, editors. RNA Worlds: From Life's Origins to Diversity in Gene Regulation. 1 edition. Cold Spring Harbor Laboratory Press; 2010.
7. Robertson MP, Joyce GF. The origins of the RNA world. Cold Spring Harb Perspect Biol. 2012;4. doi:10.1101/cshperspect.a003608
8. Steitz TA, Moore PB. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem Sci. 2003;28: 411–418.
9. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003798
10. Wagner EGH, Romby P. Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. Adv Genet. 2015;90: 133–208.
11. Wassarman KM. 6S RNA: a small RNA regulator of transcription. Curr Opin Microbiol. 2007;10: 164–168.
12. Gerdes K, Wagner EGH. RNA antitoxins. Curr Opin Microbiol. 2007;10: 117–124.
13. Fozo EM, Hemm MR, Storz G. Small toxic proteins and the antisense RNAs that repress them. Microbiol Mol Biol Rev. 2008;72: 579–89, Table of Contents.
14. De Lay N, Schu DJ, Gottesman S. Bacterial small RNA-based negative regulation: Hfq and its accomplices. J Biol Chem. 2013;288: 7996–8003.
15. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. Proc Natl Acad Sci U S A. 2010;107: 9602–9607.
16. Fröhlich KS, Vogel J. Activation of gene expression by small RNA. Curr Opin Microbiol. 2009;12: 674–682.
17. Sharma CM, Darfeuille F, Plantinga TH, Vogel J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. Genes Dev. 2007;21: 2804–2817.
18. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron

metabolism in Escherichia coli. Proc Natl Acad Sci U S A. 2002;99: 4620–4625.

19. 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. 2007;64: 1260–1273.

20. Zhang A, Altuvia S, Tiwari A, Argaman L, Hengge-Aronis R, Storz G. The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein. EMBO J. 1998;17: 6061–6068.

21. 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. 1996;15: 3993–4000.

22. Majdalani N, Hernandez D, Gottesman S. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. Mol Microbiol. 2002;46: 813–826.

23. Morita T, Mochizuki Y, Aiba H. Translational repression is sufficient for gene silencing by bacterial small noncoding RNAs in the absence of mRNA destruction. Proc Natl Acad Sci U S A. 2006;103: 4858–4863.

24. Prévost K, Desnoyers G, Jacques J-F, Lavoie F, Massé E. Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage. Genes Dev. 2011;25: 385–396.

25. Massé E, Escorcia FE, Gottesman S. Coupled degradation of a small regulatory RNA and its mRNA targets in Escherichia coli. Genes Dev. 2003;17: 2374–2383.

26. Jousselin A, Metzinger L, Felden B. On the facultative requirement of the bacterial RNA chaperone, Hfq. Trends Microbiol. 2009;17: 399–405.

27. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nat Rev Microbiol. 2011;9: 578–589.

28. Sauer E. Structure and RNA-binding properties of the bacterial LSm protein Hfq. RNA Biol. 2013;10: 610–618.

29. Weichenrieder O. RNA binding by Hfq and ring-forming (L)Sm proteins: a trade-off between optimal sequence readout and RNA backbone conformation. RNA Biol. 2014;11: 537–549.

30. Attaiech L, Glover JNM, Charpentier X. RNA Chaperones Step Out of Hfq's Shadow. Trends Microbiol. Elsevier; 2017;25: 247–249.

31. Olejniczak M, Storz G. ProQ/FinO-domain proteins: another ubiquitous family of RNA matchmakers? Mol Microbiol. 2017;104: 905–915.

32. Meynell E, Meynell GG, Datta N. Phylogenetic relationships of drug-resistance factors and other transmissible bacterial plasmids. Bacteriol Rev. 1968;32: 55–83.

33. Attaiech L, Boughammoura A, Brochier-Armanet C, Allatif O, Peillard-Fiorente F, Edwards RA, et al. Silencing of natural transformation by an RNA chaperone and a multitarget small

RNA. Proc Natl Acad Sci U S A. 2016;113: 8813–8818.

34. Smirnov A, Förstner KU, Holmqvist E, Otto A, Günster R, Becher D, et al. Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. Proc Natl Acad Sci U S A. National Academy of Sciences; 2016;113: 11591–11596.

35. Ross JA, Ellis MJ, Hossain S, Haniford DB. Hfq restructures RNA-IN and RNA-OUT and facilitates antisense pairing in the Tn10/IS10 system. RNA. 2013;19: 670–684.

36. Franze de Fernandez MT, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. Nature. 1968;219: 588–590.

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

38. 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 Dev. 1996;10: 1143–1151.

39. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev. 2001;15: 1637–1651.

40. Sittka A, Lucchini S, Papenfort K, Sharma CM, Rolle K, Binnewies TT, et al. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. PLoS Genet. 2008;4: e1000163.

41. Sharma CM, Vogel J. Experimental approaches for the discovery and characterization of regulatory small RNA. Curr Opin Microbiol. 2009;12: 536–546.

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

43. Sledjeski DD, Whitman C, Zhang A. Hfq is necessary for regulation by the untranslated RNA DsrA. J Bacteriol. 2001;183: 1997–2005.

44. Fantappiè L, Metruccio MME, Seib KL, Oriente F, Cartocci E, Ferlicca F, 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. 2009;77: 1842–1853.

45. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. Cell. 2004;118: 69–82.

46. Mika F, Hengge R. Small Regulatory RNAs in the Control of Motility and Biofilm Formation in E. coli and Salmonella. Int J Mol Sci. 2013;14: 4560–4579.

47. Sittka A, Pfeiffer V, Tedin K, Vogel J. The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. Mol Microbiol. 2007;63: 193–217.

48. Chao Y, Vogel J. The role of Hfq in bacterial pathogens. Curr Opin Microbiol. 2010;13:

24–33.

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. John Wiley & Sons, Ltd; 2002;21: 3546–3556.

50. Brennan RG, Link TM. Hfq structure, function and ligand binding. Curr Opin Microbiol. 2007;10: 125–133.

51. Sauer E, Weichenrieder O. Structural basis for RNA 3'-end recognition by Hfq. Proc Natl Acad Sci U S A. 2011;108: 13065–13070.

52. Wilson KS, von Hippel PH. Transcription termination at intrinsic terminators: the role of the RNA hairpin. Proc Natl Acad Sci U S A. 1995;92: 8793–8797.

53. Link TM, Valentin-Hansen P, Brennan RG. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. Proc Natl Acad Sci U S A. 2009;106: 19292–19297.

54. Robinson KE, Orans J, Kovach AR, Link TM, Brennan RG. Mapping Hfq-RNA interaction surfaces using tryptophan fluorescence quenching. Nucleic Acids Res. 2014;42: 2736–2749.

55. Horstmann N, Orans J, Valentin-Hansen P, Shelburne SA 3rd, Brennan RG. Structural mechanism of Staphylococcus aureus Hfq binding to an RNA A-tract. Nucleic Acids Res. 2012;40: 11023–11035.

56. Folichon M, Arluison V, Pellegrini O, Huntzinger E, Régnier P, Hajnsdorf E. The poly(A) binding protein Hfq protects RNA from RNase E and exoribonucleolytic degradation. Nucleic Acids Res. 2003;31: 7302–7310.

57. Soper TJ, Woodson SA. The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA. RNA. Cold Spring Harbor Laboratory Press; 2008;14: 1907–1917.

58. 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. 2004;11: 1206–1214.

59. 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. 2013;41: 5938–5948.

60. Updegrove TB, Correia JJ, Chen Y, Terry C, Wartell RM. The stoichiometry of the Escherichia coli Hfq protein bound to RNA. RNA. 2011;17: 489–500.

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. Biochemistry. American Chemical Society; 2006;45: 4875–4887.

62. Wang W, Wang L, Zou Y, Zhang J, Gong Q, Wu J, et al. Cooperation of Escherichia coli Hfq hexamers in DsrA binding. Genes Dev. 2011;25: 2106–2117.

63. 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. 1994;176: 531–534.

64. 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. 1997;90: 43–53.

65. Papenfort K, Said N, Welsink T, Lucchini S, Hinton JCD, Vogel J. Specific and pleiotropic patterns of mRNA regulation by ArcZ, a conserved, Hfq-dependent small RNA. Mol Microbiol. 2009;74: 139–158.

66. Fender A, Elf J, Hampel K, Zimmermann B, Wagner EGH. RNAs actively cycle on the Sm-like protein Hfq. Genes Dev. 2010;24: 2621–2626.

67. Arluison V, Mura C, Guzmán MR, Liquier J, Pellegrini O, Gingery M, et al. Three-dimensional structures of fibrillar Sm proteins: Hfq and other Sm-like proteins. J Mol Biol. 2006;356: 86–96.

68. Obregon KA, Hoch CT, Sukhodolets MV. Sm-like protein Hfq: Composition of the native complex, modifications, and interactions. Biochim Biophys Acta. 2015;1854: 950–966.

69. Stanek KA, Patterson-West J, Randolph PS, Mura C. Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: conservation of the lateral RNA-binding mode. Acta Crystallogr D Struct Biol. 2017;73: 294–315.

70. Bøggild A, Overgaard M, Valentin-Hansen P, Brodersen DE. Cyanobacteria contain a structural homologue of the Hfq protein with altered RNA-binding properties. FEBS J. Blackwell Publishing Ltd; 2009;276: 3904–3915.

71. 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. 2014;20: 1548–1559.

72. Schulz EC, Seiler M, Zuliani C, Voigt F, Rybin V, Pogenberg V, et al. Intermolecular base stacking mediates RNA-RNA interaction in a crystal structure of the RNA chaperone Hfq. Sci Rep. 2017;7: 9903.

73. Soper TJ, Doxzen K, Woodson SA. Major role for mRNA binding and restructuring in sRNA recruitment by Hfq. RNA. 2011;17: 1544–1550.

74. Henderson CA, Vincent HA, Casamento A, Stone CM, Phillips JO, Cary PD, 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. 2013;19: 1089–1104.

75. 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. 2011;1809: 532–540.

76. Ishikawa H, Otaka H, Maki K, Morita T, Aiba H. The functional Hfq-binding module of bacterial sRNAs consists of a double or single hairpin preceded by a U-rich sequence and followed by a 3’ poly(U) tail. RNA. 2012;18: 1062–1074.

77. Panja S, Schu DJ, Woodson SA. Conserved arginines on the rim of Hfq catalyze base pair

formation and exchange. Nucleic Acids Res. Oxford University Press; 2013;41: 7536–7546.

78. Zheng A, Panja S, Woodson SA. Arginine Patch Predicts the RNA Annealing Activity of Hfq from Gram-Negative and Gram-Positive Bacteria. J Mol Biol. 2016;428: 2259–2264.

79. 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 U S A. 2012;109: 9396–9401.

80. Dimastrogiovanni D, Fröhlich KS, Bandyra KJ, Bruce HA, Hohensee S, Vogel J, et al. Recognition of the small regulatory RNA RydC by the bacterial Hfq protein. eLife Sciences. eLife Sciences Publications Limited; 2014;3: e05375.

81. Zhang A, Schu DJ, Tjaden BC, Storz G, Gottesman S. Mutations in interaction surfaces differentially impact E. coli Hfq association with small RNAs and their mRNA targets. J Mol Biol. 2013;425: 3678–3697.

82. Schu DJ, Zhang A, Gottesman S, Storz G. Alternative Hfq-sRNA interaction modes dictate alternative mRNA recognition. EMBO J. 2015;34: 2557–2573.

83. Figueroa-Bossi N, Valentini M, Malleret L, Fiorini F, Bossi L. Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. Genes Dev. 2009;23: 2004–2015.

84. Sun X, Zhulin I, Wartell RM. Predicted structure and phyletic distribution of the RNA-binding protein Hfq. Nucleic Acids Res. 2002;30: 3662–3671.

85. Attia AS, Sedillo JL, Wang W, Liu W, Brautigam CA, Winkler W, et al. Moraxella catarrhalis expresses an unusual Hfq protein. Infect Immun. 2008;76: 2520–2530.

86. Schilling D, Gerischer U. The Acinetobacter baylyi Hfq gene encodes a large protein with an unusual C terminus. J Bacteriol. 2009;191: 5553–5562.

87. Beich-Frandsen M, Vecerek B, Konarev PV, Sjöblom B, Kloiber K, Hämmerle H, et al. Structural insights into the dynamics and function of the C-terminus of the E. coli RNA chaperone Hfq. Nucleic Acids Res. 2011;39: 4900–4915.

88. Sonnleitner E, Napetschnig J, Afonyushkin T, Ecker K, Vecerek B, Moll I, et al. Functional effects of variants of the RNA chaperone Hfq. Biochem Biophys Res Commun. 2004;323: 1017–1023.

89. Vecerek B, Rajkowitsch L, Sonnleitner E, Schroeder R, Bläsi U. The C-terminal domain of Escherichia coli Hfq is required for regulation. Nucleic Acids Res. 2008;36: 133–143.

90. Olsen AS, Møller-Jensen J, Brennan RG, Valentin-Hansen P. C-terminally truncated derivatives of Escherichia coli Hfq are proficient in riboregulation. J Mol Biol. 2010;404: 173–182.

91. Santiago-Frangos A, Kavita K, Schu DJ, Gottesman S, Woodson SA. C-terminal domain of the RNA chaperone Hfq drives sRNA competition and release of target RNA. Proc Natl

Acad Sci U S A. 2016;113: E6089–E6096.

92. Santiago-Frangos A, Jeliazkov JR, Gray JJ, Woodson SA. Acidic C-terminal domains autoregulate the RNA chaperone Hfq. eLife Sciences. eLife Sciences Publications Limited; 2017;6: e27049.

93. Cavanagh J, Fairbrother WJ, Arthur G. Palmer I, Skelton NJ, Rance M. Protein NMR Spectroscopy: Principles and Practice. Elsevier; 2010.

94. Sherwood D, Cooper J. Crystals, X-rays and Proteins: Comprehensive Protein Crystallography. Oxford University Press; 2011.

95. Cheng Y, Grigorieff N, Penczek PA, Walz T. A primer to single-particle cryo-electron microscopy. Cell. 2015;161: 438–449.

96. Stanek KA, Mura C. Producing Hfq/Sm Proteins and sRNAs for Structural and Biophysical Studies of Ribonucleoprotein Assembly. In: Arluison V, Valverde C, editors. Bacterial Regulatory RNA: Methods and Protocols. New York, NY: Springer New York; 2018. pp. 273–299.

97. 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. 2012;40: 1856–1867.

98. Møller T, Franch T, Højrup P, Keene DR, Bächinger HP, Brennan RG, et al. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Mol Cell. 2002;9: 23–30.

99. Arluison V, 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. 2007;16: 1830–1841.

100. 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. 2002;320: 705–712.

101. Lerner MR, Steitz JA. Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. Proc Natl Acad Sci U S A. 1979;76: 5495–5499.

102. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003707

103. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. Int Rev Cell Mol Biol. 2009;272: 149–189.

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

105. 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. 2002;320: 129–142.

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

107. 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. Public Library of Science; 2009;5: e1000315.

108. Fischer U, Englbrecht C, Chari A. Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdiscip Rev RNA. 2011;2: 718–731.

109. Achsel T, Brahms H, Kastner B, Bachi A, Wilm M, Lührmann R. A doughnut-shaped heteromer of human Sm-like proteins binds to the 3'-end of U6 snRNA, thereby facilitating U4/U6 duplex formation in vitro. EMBO J. 1999;18: 5789–5802.

110. Bouveret E, Rigaut G, Shevchenko A, Wilm M, Séraphin B. A Sm-like protein complex that participates in mRNA degradation. EMBO J. 2000;19: 1661–1671.

111. Leung AKW, Nagai K, Li J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;473: 536.

112. Urlaub H, Raker VA, Kostka S, Lührmann R. Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. EMBO J. 2001;20: 187–196.

113. 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. J Biol Chem. 2003;278: 1239–1247.

114. 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 U S A. National Academy of Sciences; 2001;98: 5532–5537.

115. Murzin AG. OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. 1993;12: 861–867.

116. McCarty JH. The Nck SH2/SH3 adaptor protein: a regulator of multiple intracellular signal transduction events. Bioessays. 1998;20: 913–921.

117. Nakagawa A, Nakashima T, Taniguchi M, Hosaka H, Kimura M, Tanaka I. The three-dimensional structure of the RNA-binding domain of ribosomal protein L2; a protein at the peptidyl transferase center of the ribosome. EMBO J. 1999;18: 1459–1467.

118. Bass RB, Strop P, Barclay M, Rees DC. Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. Science. 2002;298: 1582–1587.

119. Klein DJ, Moore PB, Steitz TA. The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. J Mol Biol. 2004;340: 141–177.

120. Theobald DL, Wuttke DS. Divergent evolution within protein superfolds inferred from profile-based phylogenetics. J Mol Biol. 2005;354: 722–737.

121. Gonzalez GM, Hardwick SW, Maslen SL, Skehel JM, Holmqvist E, Vogel J, et al. Structure of theEscherichia coliProQ RNA-binding protein. RNA. 2017;23: 696–711.

122. Huntzinger E, Boisset S, Saveanu C, Benito Y, Geissmann T, Namane A, et al. Staphylococcus aureus RNAIII and the endoribonuclease III coordinately regulate spa gene expression. EMBO J. 2005;24: 824–835.

123. Bohn C, Rigoulay C, Bouloc P. No detectable effect of RNA-binding protein Hfq absence in Staphylococcus aureus. BMC Microbiol. 2007;7: 10.

124. Ramos CG, Sousa SA, 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. 2011;193: 1515–1526.

125. Ramos CG, da Costa PJP, Döring G, Leitão JH. The novel cis-encoded small RNA h2cR is a negative regulator of hfq2 in Burkholderia cenocepacia. PLoS One. 2012;7: e47896.

126. Vrentas C, Ghirlando R, Keefer A, Hu Z, Tomczak A, Gittis AG, et al. Hfqs in Bacillus anthracis: Role of protein sequence variation in the structure and function of proteins in the Hfq family. Protein Sci. 2015;24: 1808–1819.

127. 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. 2015;1850: 1661–1668.

128. Zarrinpar A, Bhattacharyya RP, Lim WA. The structure and function of proline recognition domains. Sci STKE. 2003;2003: RE8.

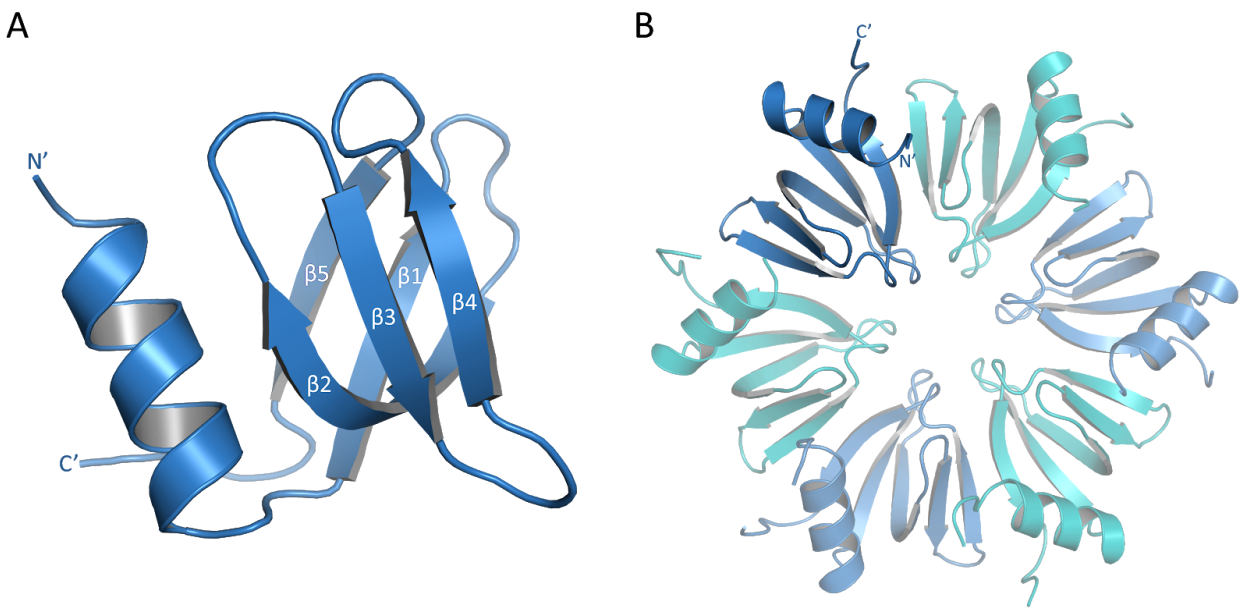


Figure 1 Structure of the bacterial RNA chaperone Hfq. (A) The Hfq fold consists of an N-terminal α-helix followed by five antiparallel β-strands (labelled here, along with the N' and C'-termini). Note that the C-terminal tail (~40 residues for *E. coli*) is disordered and typically cannot be modelled in Hfq crystal structures. (B) Hfq spontaneously self-oligomerizes into a stable toroidal hexamer with an outer diameter of ~65 Å and pore diameter of ~10 Å. Oligomerization is facilitated through backbone hydrogen-bonding interactions between strands β4 and β5 of adjacent monomers. *E. coli* Hfq (PDB 1HK9) is shown here. Hfq monomers are colored alternatingly as blue or teal for clarity. The view shown here in onto the proximal face.

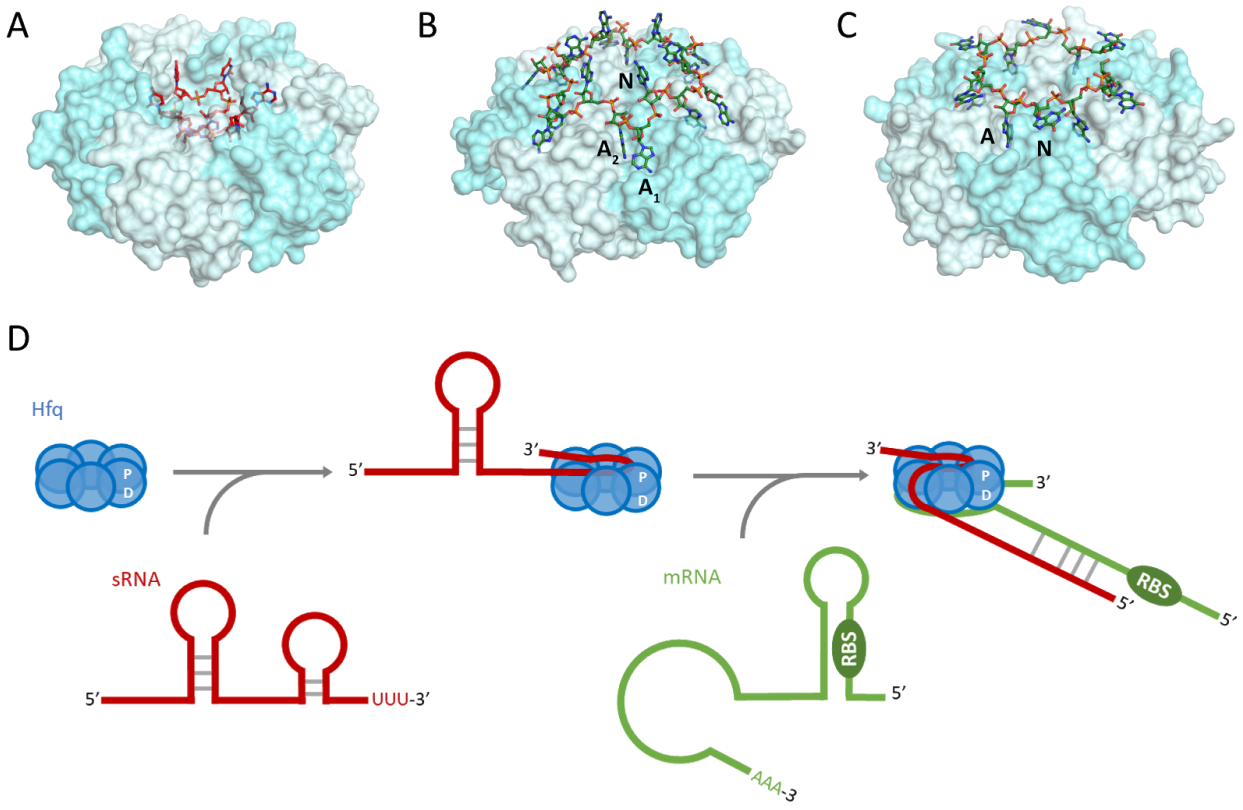


Figure 2 The RNA-binding landscape of Hfq. (A) Proximal-facing view of *S. aureus* Hfq crystallized with AU_5G RNA (PDB 1KQ2). The proximal site consists of six equivalent binding pockets with uridine specificity; the uracil bases stack between a conserved Phe/Tyr residue from adjacent subunits (Phe42 in *E. coli* Hfq numbering). (B) The distal mode of RNA-binding in gram-negative bacteria. Distal-facing view of *E. coli* Hfq with A_9 RNA (PDB 3GIB) is shown with the same representation as in (A) except the RNA is colored green. The gram-negative 'AAN' motif that binds Hfq is labelled. (C) The distal mode of RNA-binding in gram-positive bacteria. Distal-facing view of *B. subtilis* Hfq with $(AG)_3A$ RNA (PDB 3AHU) is shown with the same representation as in (B). The gram-positive 'AN' motif that binds Hfq is labelled. (D) One mechanism of Hfq action. The Hfq hexamer (blue) simultaneously binds sRNA (red) on the proximal (P) face and mRNA (green) on the distal (D) face, resulting in a ternary

sRNA:Hfq:mRNA complex. In this example, the sRNA is restructured upon binding Hfq. Hfq brings both RNAs into close proximity, allowing them to base-pair; here, this results in remodelling of the mRNA, and concomitant release of an occluded ribosomal binding site (RBS). This process, in turn, leads to an increase in translation.

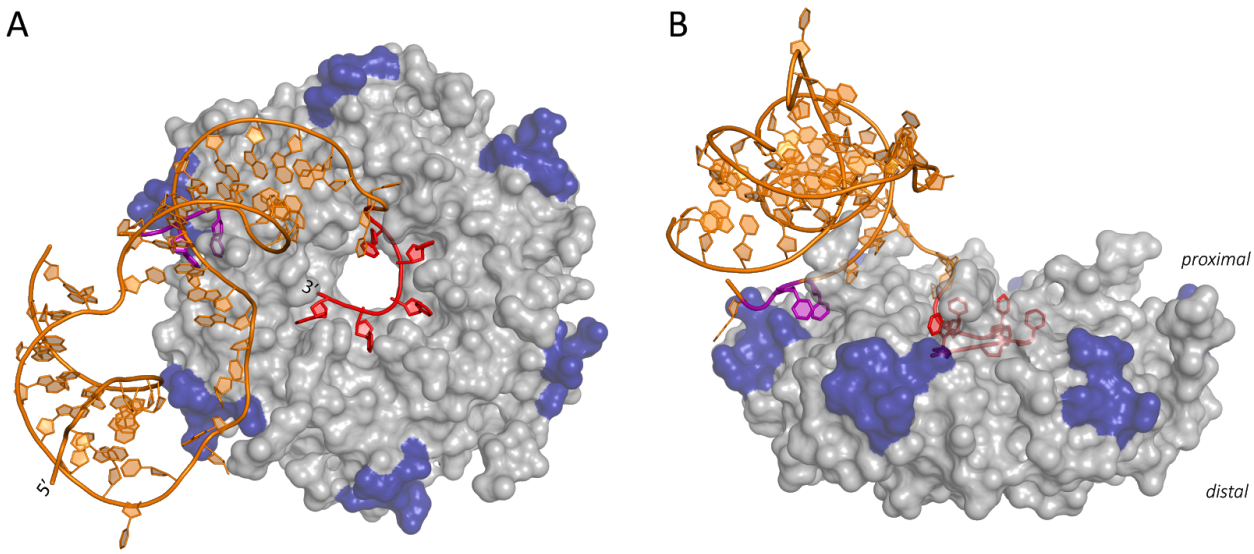


Figure 3 RNA-binding at the lateral surface of Hfq. The structure of *E. coli* Hfq (grey) in complex with RydC sRNA (orange) is shown (PDB 4V2S) from (A) proximal and (B) lateral orientations. The 3'-end of RydC (highlighted in red) binds to the proximal pore region, while an internal U-rich sequence makes further contacts with the lateral surface. Two nucleotides of uridine (highlighted in purple) bind in a deep cleft near the proximal side of the rim. The arginine-rich patch of Hfq is colored blue; note that only a subset of these arginines are involved in binding of the two uridines. The arginine-rich patch likely associates RNA non-specifically, increasing subsequent binding to the lateral pocket.

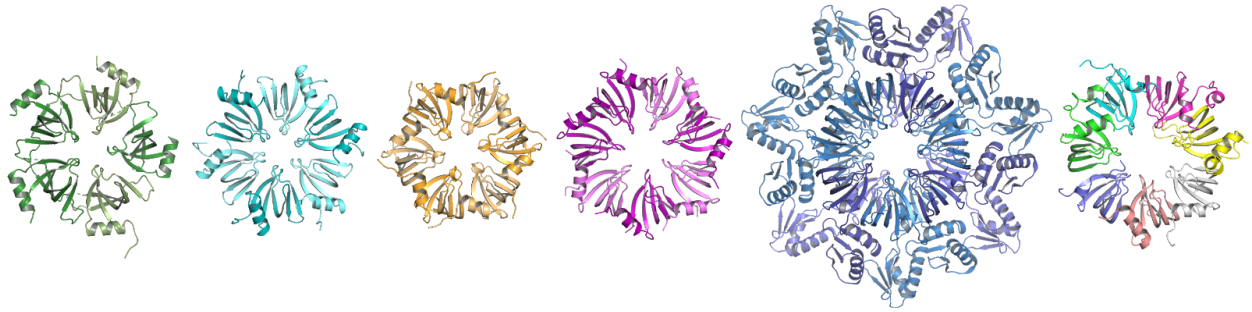


Figure 4 Varying oligomeric plasticity in the Sm protein superfamily. From left to right: An Sm-like pentamer was identified in an uncultured marine organism (PDB 3BY7). Bacterial Hfq from *E. coli* (PDB 1HK9) self-assembles as a stable homohexamer. The euryarchaeotal *A. fulgidus* encodes two SmAPs; SmAP1 oligomerizes as a heptamer and SmAP2, shown here, forms a hexamer (PDB 1LJO); The chrenarchaeotal *P. aerophilum* encodes three SmAPs; SmAP1 (PDB 1LNX) forms a heptamer while SmAP3 (PDB 1M5Q) assembles as a 14-mer due to the stacking of two heptameric rings. More recently, SmAP2 was found to oligomerize as an octamer. Eukaryotic Sm rings such as the one found in the human U4 snRNP (PDB 4WZJ) exist as heteroheptamers and are formed through a chaperoned assembly pathway. In the U4 snRNP, these paralogs are SmB/B', SmD1, SmD2, SmD3, SmE, SmF, and SmG. All of the structures shown here are on a common length scale.

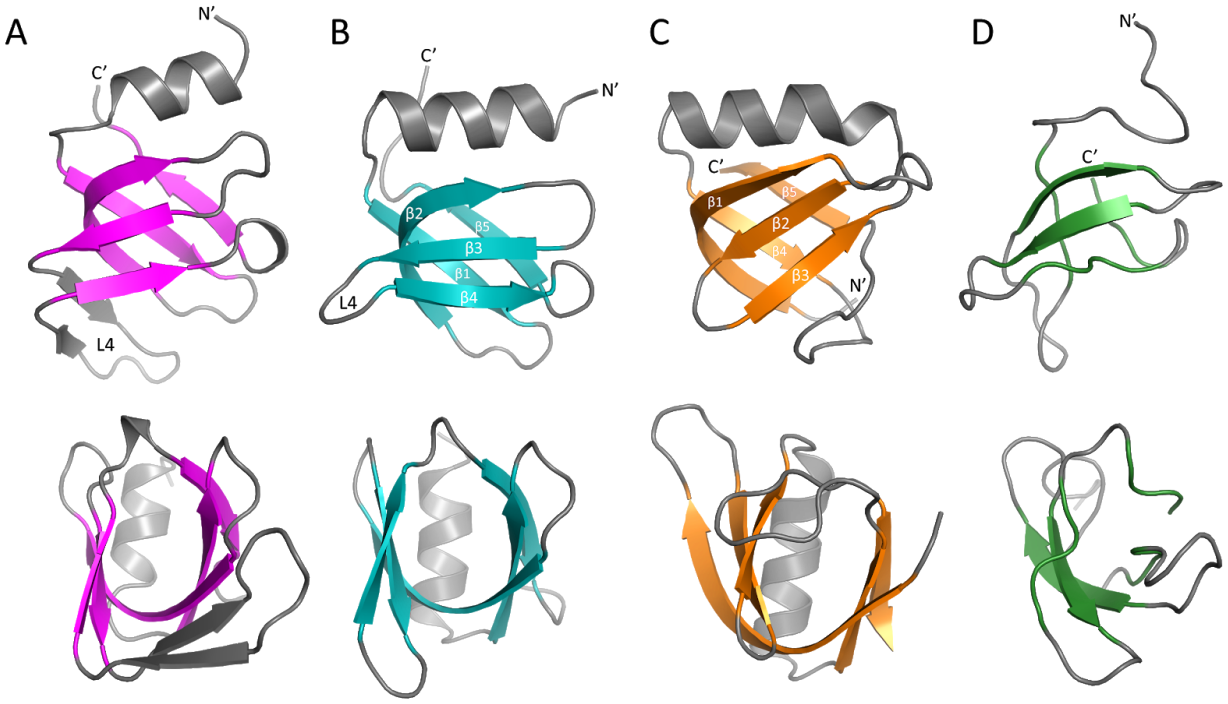


Figure 5 The topology of small β-barrels. The β-rich core of the SBB domain is highlighted in color and extra loops and motifs are shown in grey. (A) Human SmD3 (PDB 1D3B) exhibits SH3-like topology and features an extension of the L4 loop (grey β-strands). (B) *E. coli* Hfq (PDB 1HK9) is a bacterial Sm protein, featuring the same topology as SmD3 (other than the shortened L4 loop), however the protein also exhibits high structural similarity with the OB-fold. (C) The shiga-like toxin from *E. coli* is an OB-fold with 0.625 Å RMSD to Hfq. Note however, the strand permutation between OB and Hfq; for example, the OB-fold α-helix falls between strands β3 and β4. (D) The C-terminal domain of the bacterial RNA chaperone ProQ is a more flexible Tudor-like SBB.

Chapter 2: Phylogenetic analysis of the Hfq family of RNA-binding proteins

Kimberly Stanek and Cameron Mura

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

Abstract

The bacterial branch of the Sm superfamily of RNA-binding proteins, known as Hfq, functions as a central hub in post-transcriptional regulation of gene expression by chaperoning the actions of small regulatory RNAs (sRNAs) with their mRNA targets. Bacterial genomes typically encode one Hfq homolog that oligomerizes as a hexamer, while eukaryotes have multiple (>7) Sm paralogs that form heteroheptamers. There are many unresolved questions surrounding evolutionary transition from Hfq to Sm. As a first step, some insight could be gained with a more complete understanding of Hfq phylogeny. Here, we have surveyed available bacterial genomic sequences with the intent of identifying new Hfq homologs and constructing a modern, more complete phylogeny of the bacterial subset of Sm proteins. We have detected Hfq homologs in bacterial phyla previously thought to be missing this protein, including the Actinobacteria; these homologs were likely acquired through horizontal gene transfer. In addition, we have identified clades of bacteria, including the Aquificae, that contain two or more Hfq paralogs. Together, these results demonstrate that Hfq likely has a more intricate evolutionary history than previously suspected, that includes multiple gene duplication and

transfer events, and points to its functional importance in such rapidly evolving pathways as pathogenesis.

1. Introduction

The bacterial protein Hfq, originally identified as a host factor required for the replication of bacteriophage Qβ [1], has since been shown to function generally as an RNA chaperone and major hub of post-transcriptional regulation [2,3]. This small (~80 amino acids) protein adopts the Sm fold, which consists of an N-terminal α-helix followed by five highly-curved antiparallel β-strands. Hfq spontaneously self-assembles as a hexameric toroid, which greatly increases its available surface for binding RNA [4–6]. Hfq functions by binding small, non-coding regulatory RNA (sRNA) on the so-called 'proximal' face (with respect to the α-helical region) and mRNA on the 'distal' face of the hexamer, and facilitating their annealing. The Hfq-dependent sRNAs are then able to upregulate [7], or downregulate [8] expression of their mRNA targets. As a result, Hfq has been linked to numerous physiological pathways, including stress response [9,10], quorum sensing [11], iron metabolism [12], and expression of virulence factors [13].

Hfq and other Sm fold proteins comprise an ancient superfamily found in all three domains of life and are believed to share a single common origin in bacteria. However, the exact phylogenetic relationships between Sm proteins from the three domains of life are unresolved as sequences are quite divergent, and this homology is largely inferred from structural similarity [14]. As is commonly found to be the pattern for homologous proteins across bacteria and eukarya, a series of extensive gene duplications of the Sm proteins occurred during the early stages of eukaryote evolution, resulting in eukaryotic species typically having more than 20 Sm and Sm-like (LSm) paralogs [15]. Eukaryotic Sm proteins assemble as

heteroheptamers through a complicated, chaperoned biogenesis pathway [16], and they function in splicing (in snRNP cores) or in other RNA processing pathways [17,18]. The Sm-like archaeal proteins (SmAPs) are closer in sequence similarity to their eukaryotic counterparts, yet in other ways they are more similar to Hfq. For example, archaeal species typically have between one and three paralogs which, like Hfq, spontaneously oligomerize as homomeric assemblies. However the oligomeric state of SmAPs has been shown to vary from a hexamer to a 14-mer comprised of two heptameric rings [19,20]. A hexameric Hfq-like protein has also been characterized in the archaeal species *Methanocaldococcus jannaschii [21]*. Notably, the physiological functions of SmAPs are still largely unknown [22].

To gain a better understanding of the evolutionary transition from bacterial Hfq to eukaryotic Sm, a complete phylogenetic history of Hfq is necessary. An initial survey of Hfq sequences in 2002 revealed that ~50% of sequenced bacterial genomes included a putative *hfq* gene (out of roughly 140 genomic sequences available [6]. In addition to pervasiveness within the more recently-branching proteobacterial lineages (which includes *E. coli*), Hfq sequences were also found in two early-branching phyla: Thermotogae and Aquificae. However, Hfq sequences were absent in several other more basal phyla, including Chlamydia, Spirochaetes, Actinomycetes and Cyanobacteria. This pattern could indicate that either *(i)* Hfq is a very ancient protein that was retained only in the clade leading to proteobacteria, or *(ii)* that Hfq emerged later, in proteobacteria, and was subsequently obtained in more basal lineages through horizontal gene transfer (HGT). The authors of the 2002 report found no evidence for HGT and concluded that the first scenario (i.e. gene loss) was more likely. Intriguingly, this early work also identified members of the *Bacillus cereus* group (including *Bacillus anthracis*) and *Burkholderia* as having two Hfq sequences; these paralogs were proposed to have been obtained through recent gene duplication events.

Within the past 15 years, as sequencing efforts have expanded, and the sensitivity of sequence-based homology detection algorithms have improved, the number of annotated Hfq sequences has increased. With these new developments and additional datasets it has become apparent that Hfq sequences vary immensely, and there are likely additional homologs in bacterial clades previously thought to be lacking Hfq. For example in recent years, highly divergent Hfqs have been annotated in several species of cyanobacteria, and crystallographic studies have verified the presence of genuine Hfq proteins in the cyanobacterial species *Anabaena* and *Synechocystis*, which share ~25% seq ID with *E. coli* Hfq [23]. More recently, biochemical and biophysical studies in *Bacillus anthracis* showed that this gram-positive bacterium actually has three Hfq paralogs, the third of which is located on a plasmid [24,25]. In addition to demonstrating that bacteria can in fact harbor more than two Hfq paralogs, the presence of a plasmid-encoded Hfq also suggests the occurrence of HGT.

As more bacteria are found to contain multiple copies of Hfq, a systematic classification of these paralogs becomes important for clarity and consistency. However, we have found that there is currently some inconsistency in naming, even within the same species. For instance, the two research groups that recently published on the *B. anthracis* Hfqs used different naming schemes, such that the copies referred to as ‘Hfq1’ and ‘Hfq2’ in one study were reversed in the other study [24,25]. By constructing a phylogenetic tree of Hfq sequences, we can infer the evolutionary relationship of different Hfq homologs, which in turn, should provide functional insight and as well as guide future classification schemes. In the case of *B. anthracis*, we will use the the notation adopted by *Panda et al.* in this text, as ‘Hfq1’ was found to share a common ancestor with other Bacillus Hfqs whereas ‘Hfq2’ and ‘Hfq3’ are more divergent.

In this work, we have surveyed available bacterial genomes with the intent of identifying new putative Hfq sequences, and constructing a more complete phylogenetic history of Hfq. We

report the presence of two (or three) Hfq paralogs in members of phylum Aquificae, as well as two Hfq paralogs in the γ-proteobacterial Aeromonodales. We have also identified putative Hfq sequences in members of Actinobacteria, which were previously thought to lack Hfq. These Hfq homologs appear to have been acquired fairly recently through HGT. Finally, we have identified several strains of *Streptococcus pneumoniae*—which also were thought to lack Hfq—and which we find in fact have between one and five Hfq copies. These sequences appear to have been acquired through a combination of HGT and gene duplication events.

2. Methods

2.1 Database searches

A reference set of Hfq sequences representing "canonical Hfq" we selected several representative sequences (including Hfqs verified through structural or biochemical characterization) from a diverse number of phyla: Thermotogae, Firmicutes (note that only sequences from class Bacilli were used, as Firmicutes sequences are in general very divergent and produced low bootstrap values when using more diverse lineages), Spirochaetes, Acidobacteria, α-proteobacteria, β-proteobacteria, and γ-proteobacteria. Position-specific iterative (PSI)-BLAST searches of the NCBI non-redundant protein database were performed using the amino acid sequences of our reference set of sequences. A full list of Hfq sequences used in this study can be found in Table S1. δ-proteobacterial and cyanobacterial Hfq sequences were omitted from our analyses as their sequences are too divergent (with sequence identities below 30%). In our searches we identified 2 Hfq paralogs in 11 members of Aquificales, 3 Hfq paralogs in 5 members of Desulfurobacteriales, and 2 Hfq paralogs in 31

members of Aeromonodales. We also found Hfq sequences in 23 Actinobacterial genomes (including 11 strains of *Mycobacterium abscessus*) and in 39 different strains of *Streptococcus pneumoniae* (strains had anywhere from one to five paralogs).

2.2 Sequence alignments and phylogenetic tree construction

For phylogenetic analysis of the inter-relationships between multiple Hfq paralogs (Fig 1), we selected Hfq sequences from five Aquificales members, three Aeromonadales, and three from the Bacillus cereus group (Table S1) and aligned them to our reference set of Hfq sequences. Similarly, to construct an Actinobacterial tree (Fig 3), we selected Hfq sequences from six representative species (Table S1) and aligned them to our reference set of Hfq sequences. Sequences were manually trimmed to the Sm core domain (~60 amino acids) and aligned using *MUSCLE* [26]. This approach was employed as the the N- and C-termini of the Hfq sequences vary greatly in terms of sequence identity as well as length, and do not provide useful additional information. Maximum likelihood phylogenetic trees were constructed with *RAxML* using a Gamma model of rate heterogeneity [27]. Non-parametric bootstrap support values were calculated from 1000 bootstrapping replicates.

3. Results and discussion

To detect, potential novel Hfq homologs, we have performed PSI-BLAST searches of the NCBI database of non-redundant protein sequences using as our search sequences a set of representative Hfq sequences from a diverse number of phyla in which species are known to have only one copy of Hfq and where HGT is not thought to have occured (Table S1). Here we take these sequences to represent “canonical Hfq”. In these searches, we identified two to three

Hfq paralogs in members of the Aquificae, as well as two paralogs in the γ-proteobacterial order Aeromonadales. We found that for many of the Aquificae, one of the paralogs was annotated as iron-sulfur cluster assembly protein HesB. However, this is likely the result of a propagated function mis-annotation [28], as manual inspection of these sequences reveals telltale characteristics of Hfq, such as a highly conserved 'YKHAI' motif (present near the proximal RNA-binding site). Furthermore, these putative HesB/Hfq sequences share no significant sequence similarity with verified HesB proteins. Ultimately, structural and functional characterization will be required to verify that these putative sequences encode authentic Hfq proteins.

We also detected Hfq sequences in 22 Actinobacterial genomes, and in 49 different strains of *Streptococcus pneumoniae*—both of which were previously thought to lack an Hfq. An estimated phylogeny would likely help to ascertain the origins of these new Hfq sequences we have identified. From our reconstructed phylogenies (Fig 1,3,4) we were able to recapitulate the separate clades for each phylum of bacteria. However, we were unable to resolve any evolutionary relationships between the clades, as our bootstrapping support values were too low (typically <30). This is consistent with the observations of *Sun et al.* [6]. Nevertheless, our phylogenetic trees provide some insight into the origins of paralogous Hfq sequences, as well as evidence for HGT; these major findings are reported below.

3.1 Bacterial species with multiple Hfq paralogs

3.1.1 Aquificae

Species from the phylum Aquificae comprise a diverse set of extremophiles that may represent one of the most early-branching bacterial lineages. However, the exact phylogenetic

placement of the Aquificae is still under debate [29]. Two conflicting theories suggest a most recent ancestor with either the Thermotogae or ε-proteobacteria [30,31]. Rampant gene transfer also appears to have occurred with the Aquificae, which could explain the apparent late-branching in some studies [29]. However, early-branching of the Aquificae based on 16S rRNA sequences has also been attributed as a spurious feature of the high G+C content generally seen for thermophilic bacteria [32]. Initially, members of Aquificae were identified as having a single *hfq* gene, but the sequences were too diverse to be linked to any shared origin with another bacterial phyla [6]; *Aquifex aeolicus* Hfq as an example, shares 47.5% sequence identity with *E. coli* Hfq.

In our present survey and phylogenetic analysis we have identified additional Hfq sequences in Aquificae: members of order Aquificales (which includes *A. aeolicus*) have two putative Hfq paralogs while members of order Desulfurobacteriales possess three Hfq paralogs. Here, we refer to the previously identified paralog as 'Hfq1' and the new paralogs as 'Hfq2' and 'Hfq3' (for the Desulfurobacteriales). We note that the putative Hfq2 and Hfq3 paralogs are more divergent in sequence from canonical Hfq than is Hfq1. The Desulfurobacteriales paralogs share limited sequence similarity to the Aquificales paralogs and were omitted from our phylogenetic analysis (as their inclusion resulted in low bootstrap support values). Hfq sequences from these two orders did not group together in our initial phylogenetic analysis (not shown), and it is likely that the additional Hfq copies were acquired after the Aquificales and Desulfurobacteriales diverged.

While it is uncertain whether Aquificales Hfq2 arose through gene duplication or HGT, we found that the clade containing Aquificales Hfq2 sequences and Bacillus Hfq2 and Hfq3 sequences, with 73% bootstrap support, shares higher sequence identity within members of the clade than with sequences outside the clade (Fig 1). Furthermore, the ~10-15 residue, acidic

C-terminal extensions found in Aquificales Hfq1 or Bacillus Hfq1 are absent in all members of this Hfq2/3 clade. Other canonical Hfqs have been characterized with disordered C-terminal extensions of up to ~100 amino acid residues [33,34]. In *E. coli* Hfq, the ~40 residue, acidic C-terminal tail has been shown to be important for efficient release of the sRNA-mRNA duplex from Hfq after annealing [35,36].

The Desulfurobacteriales Hfq2 and Hfq3 sequences are unique in that they both have N-terminal extensions of ~100 residues (Fig 2). Extended N-termini are much less common than C-terminal extensions with Hfq homologs, and the potential functional role of the N-terminal region of Hfq in general has not been as well-studied as the C-terminus. While the *in vivo* function is unclear, we have recently found that the N-terminus can help mediate the formation of dodecamers, as well as non-specifically participate in RNA-binding at the lateral surface of Hfq [37]. PSI-BLAST searches of the N-termini alone do not yield any hits of high sequence similarity, although Desulfurobacteriales Hfq2 does show limited homology with the GntR family of DNA-binding transcriptional regulators. Intriguingly, both Hfq2 and Hfq3 N-termini exhibit several lysine-rich 'KKXK' repeats, which are physicochemically similar to the 'RRER' motif identified as an additional region for binding RNA on the lateral rim of canonical Hfq [38,39]. While Desulfurobacteriales Hfq1 homologs have this conserved 'RRER' motif, the Hfq2 and Hfq3 paralogs do not. Interestingly, the highly conserved 'YKHAI' motif from the β4-β5 strand region of canonical Hfq is 'YKHSI' in Desulfurobacteriales Hfq1 sequences, but is strictly conserved in the Hfq2 and Hfq3 paralogs.

3.2.1 Aeromonadales

We have also identified two putative Hfq paralogs in members of the γ-proteobacterial order Aeromonadales (Fig 1). To our knowledge, this is the first example of a γ-proteobacterial

species with multiple Hfqs. In our phylogeny, Aeromonadales Hfq1 groups with other γ-proteobacterial Hfqs (*Oceanomas smirnovii* as an example shares 69.3% sequence identity with *E. coli* Hfq). It is not clear how Hfq2 originated from these lineages, although it was likely not a recent event, as no common ancestor for the Aeromonadales Hfq2 paralogs was found in our analysis. While Aeromonodales Hfq1 has an ~10 residue, acidic C-terminal tail that is similar to canonical Hfq, Hfq2 has a longer tail (~35 residues) that is more enriched for basic residues. This could suggest a role in binding nucleic acid, as was originally expected for canonical Hfq. Hfq2 also has a signature, ~10 residue N-terminal tail enriched for threonine and serine residues, and an altered 'YKHAI' motif that is instead 'YKAKI'. The potential physiological role of such amino acid substitutions is as of yet unclear.

3.2 Bacterial species that likely acquired Hfq through horizontal gene transfer

3.2.1 Actinobacteria

We have identified putative Hfq-coding open reading frames in a small number of Actinobacterial species, including several species known to be drug-resistant pathogens, such as the *Mycobacterium* and *Streptomyces* genuses. A comparative phylogenetic analysis against known Hfq sequences strongly indicates that these newly identified homologs were likely acquired through HGT (Fig 3). Such a mechanism would also explain why the *hfq* gene is not ubiquitous amongst all Actinobacteria. Multiple HGT events appear to have occured, with the donor sequences originating from diverse lineages of bacteria, (including Bacilli, β-proteobacteria, and γ-proteobacteria), and some events occurring relatively recently. For

example, *M. tuberculosis* Hfq, which despite intense focus on this species, evaded detection until now, shares 82.5% sequence identity with *E. coli* Hfq, the only differences in sequence being found in the disordered C-terminal tails.

Corynebacterium striatum, a recently emerging drug-resistant pathogen [40], appears to have obtained Hfq through HGT from *Staphylococcus aureus* (Fig 3). *S. aureus* Hfq is somewhat unique, in that it is rather acidic for an RNA-binding Hfq protein (with an isolelectric point of 4.69), and its sequence is extremely divergent from other Hfq homologs (25.9% sequence identity to *E. coli* Hfq). The *in vivo* functional importance of *S. aureus* Hfq is still unclear. Initial studies found it was expressed at low cellular levels, and that knockout of this *hfq* gene had no apparent phenotypic effect [41,42]. However, later efforts showed that expression levels may be strain dependent; for example multi-resistant *S. aureus* (MRSA) strains expressed Hfq at higher levels, and pathogenicity was decreased in Δ*hfq* mutants of these strains [43].

We found nine different Hfq protein sequences in various strains of *Mycobacterium abscessus*, and it appears that these train variants have gained the gene through multiple HGT events. In our phylogenies, most of these sequences share a common ancestor with *S. aureus* Hfq. One such strain (*M. abscessus subsp. abscessus str. 1000*) features two paralogs, with 73.8% pairwise sequence identity, indicating a possible additional gene duplication event (Fig 3). Several other *M. abscessus* strains grouped with members of order Bacillales, while another populated a clade with the β-proteobacteria. Two Hfq sequences were also found in a member of the Propionibacteriales, *Mumia flava*. Intriguingly, these *M. flava* Hfq sequences are nearly identical to the two Hfq paralogs found in *Burkholderia cenocepacia* (100% sequence identity for the Hfq1 proteins and 92.7% for the Hfq2 paralogs).

3.2.2 *Streptococcus pneumoniae*

The pathogenic bacterium *Streptococcus pneumoniae*, which is a member of the order Lactobacillales, was previously thought to lack an Hfq [6,13]. As other orders of Bacilli do possess one (or more) Hfqs, gene loss may have occurred prior to the Lactobacillales radiation. As part of this present study, we have found 49 different strains of *S. pneumoniae* with anywhere from one to five copies of Hfq. Most of these sequences share a common ancestor with one of the three Hfq paralogs found in members of the *Bacillus cereus* group (Fig 4). Due to the close phylogenetic relationship between the Bacillales and Lactobacillales, we cannot distinguish whether Hfq was completely lost before the branching of the Lactobacillales and later reacquired, or whether an Hfq gene was retained variably in different members of the Lactobacillales.

S. pneumoniae type strain: N was the strain with the highest copy number we found, with a total of five Hfq-like sequences. Hfq1 (to use the numbering as annotated in GenBank) is most similar to *B. anthracis* Hfq1, while Hfq2 and Hfq5 group with *B. anthracis* Hfq2, and Hfq3 and Hfq4 with *B. anthracis* Hfq3. Mostly likely, three Hfqs were acquired from the *B. cereus* group and then two gene duplication events led to the total of five Hfqs within this strain. Note that all five copies of Hfq are chromosomally encoded. Another strain, *S pneumoniae 2842STDY5753546*, has four Hfq paralogs, one of which more clearly originated through HGT (Fig 4). In this strain, Hfq1 shares a clade with *Bacillus subtilis*, while Hfq2 appears to have been acquired from γ-proteobacteria, sharing the highest sequence identity with *Haemophilus influenzae*. The Hfq3 and Hfq4 sequences of this strain share a common ancestor with the highly divergent *S. aureus* Hfq. As Hfq3 and Hfq4 share ~85% sequence identity to *S. aureus* Hfq, this is likely not due to the phenomenon of long branch attraction.

4. Conclusions

Here, we have shown that the bacterial Hfq protein is far more widespread than previously thought, having been acquired through HGT in the phylum Actinobacteria and in several strains of *S. pneumoniae* (both of which were thought to lack Hfq). Gene duplication of Hfq also appears to be a fairly common phenomenon and we now know of species with two (or more) Hfqs in Aquificae, Bacilli, β-proteobacteria, and γ-proteobacteria. Our results also suggest that multiple gene duplication events likely led to these paralogous Hfqs within the different phyla. Typically, gene duplication of ancient protein families is observed extensively only after the eukaryotic branching point [14]. Thus it is intriguing, from the perspective of molecular evolution, to find such duplication events in even such potentially early-branching bacterial species as the Aquificae.

Structural and functional characterization of the Hfq sequences proposed herein is required to understand their potential physiological roles. The paralogous Hfqs we have found in the Aquificales and Aermonadales likely serve distinct roles *in vivo*, as has been suggested for the Hfq paralogs in *B. anthracis* and *B. cenocepacia* [24,25,44]. It is also of note that (sometimes multiple) Hfq sequences have been retained or seemingly reacquired in several pathogenic species, such as *S. pneumoniae*. The importance of Hfq to the full pathogenic capability of several bacterial species has been clearly demonstrated and is reviewed in [13]. In some species such as *S. aureus*, it has also been shown that Hfq expression levels are tied to the pathogenicity of the host strain [42,43]. Useful insight could be gained by comparing the pathogenicity of strains (ex. *S. pneumoniae*) that have acquired Hfq versus those that lack it.

This would be particularly interesting in those strains which have multiple Hfq sequences, as it as yet unknown whether all of the encoded sequences would be actively expressed.

Acknowledgements

We thank M. Wu and M. Hague (UVA Biology) for helpful discussion. This work was funded by NSF career award 1350957 and Jeffress Memorial Trust award J-971.

References

1. Franze de Fernandez MT, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. Nature. 1968;219: 588–590.

2. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nat Rev Microbiol. 2011;9: 578–589.

3. Sauer E. Structure and RNA-binding properties of the bacterial LSm protein Hfq. RNA Biol. 2013;10: 610–618.

4. 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. John Wiley & Sons, Ltd; 2002;21: 3546–3556.

5. 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. 2002;320: 705–712.

6. Sun X, Zhulin I, Wartell RM. Predicted structure and phyletic distribution of the RNA-binding protein Hfq. Nucleic Acids Res. 2002;30: 3662–3671.

7. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. Proc Natl Acad Sci U S A. 2010;107: 9602–9607.

8. De Lay N, Schu DJ, Gottesman S. Bacterial small RNA-based negative regulation: Hfq and its accomplices. J Biol Chem. 2013;288: 7996–8003.

9. 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 Dev. 1996;10: 1143–1151.

10. Gottesman S, McCullen CA, Guillier M, Vanderpool CK, Majdalani N, Benhammou J, et al. Small RNA regulators and the bacterial response to stress. Cold Spring Harb Symp Quant Biol. 2006;71: 1–11.

11. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. Cell. 2004;118: 69–82.

12. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. Proc Natl Acad Sci U S A. 2002;99: 4620–4625.

13. Chao Y, Vogel J. The role of Hfq in bacterial pathogens. Curr Opin Microbiol. 2010;13: 24–33.

14. Aravind L, Iyer LM, Koonin EV. Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr Opin Struct Biol. 2006;16: 409–419.

15. 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. Public Library of Science; 2009;5: e1000315.

16. Fischer U, Englbrecht C, Chari A. Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdiscip Rev RNA. 2011;2: 718–731.

17. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003707

18. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. Int Rev Cell Mol Biol. 2009;272: 149–189.

19. 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. 2002;320: 129–142.

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

21. Nielsen JS, Bøggild A, Andersen CBF, Nielsen G, Boysen A, Brodersen DE, et al. An Hfq-like protein in archaea: crystal structure and functional characterization of the Sm protein from Methanococcus jannaschii. RNA. 2007;13: 2213–2223.

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

23. Bøggild A, Overgaard M, Valentin-Hansen P, Brodersen DE. Cyanobacteria contain a structural homologue of the Hfq protein with altered RNA-binding properties. FEBS J. Blackwell Publishing Ltd; 2009;276: 3904–3915.

24. Vrentas C, Ghirlando R, Keefer A, Hu Z, Tomczak A, Gittis AG, et al. Hfqs in Bacillus anthracis: Role of protein sequence variation in the structure and function of proteins in the Hfq family. Protein Sci. 2015;24: 1808–1819.

25. 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. 2015;1850: 1661–1668.

26. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

27. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30: 1312–1313.

28. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5: e1000605.

29. Eveleigh RJM, Meehan CJ, Archibald JM, Beiko RG. Being Aquifex aeolicus: Untangling a

hyperthermophile's checkered past. Genome Biol Evol. 2013;5: 2478–2497.

30. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature. 2009;462: 1056–1060.

31. Boussau B, Guéguen L, Gouy M. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol. 2008;8: 272.

32. Griffiths E, Gupta RS. Signature sequences in diverse proteins provide evidence for the late divergence of the Order Aquificales. Int Microbiol. 2004;7: 41–52.

33. Attia AS, Sedillo JL, Wang W, Liu W, Brautigam CA, Winkler W, et al. Moraxella catarrhalis expresses an unusual Hfq protein. Infect Immun. 2008;76: 2520–2530.

34. Schilling D, Gerischer U. The Acinetobacter baylyi Hfq gene encodes a large protein with an unusual C terminus. J Bacteriol. 2009;191: 5553–5562.

35. Santiago-Frangos A, Kavita K, Schu DJ, Gottesman S, Woodson SA. C-terminal domain of the RNA chaperone Hfq drives sRNA competition and release of target RNA. Proc Natl Acad Sci U S A. 2016;113: E6089–E6096.

36. Santiago-Frangos A, Jeliazkov JR, Gray JJ, Woodson SA. Acidic C-terminal domains autoregulate the RNA chaperone Hfq. eLife Sciences. eLife Sciences Publications Limited; 2017;6: e27049.

37. Stanek KA, Patterson-West J, Randolph PS, Mura C. Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: conservation of the lateral RNA-binding mode. Acta Crystallogr D Struct Biol. 2017;73: 294–315.

38. Panja S, Schu DJ, Woodson SA. Conserved arginines on the rim of Hfq catalyze base pair formation and exchange. Nucleic Acids Res. Oxford University Press; 2013;41: 7536–7546.

39. 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 U S A. 2012;109: 9396–9401.

40. Alibi S, Ferjani A, Boukadida J, Cano ME, Fernández-Martínez M, Martínez-Martínez L, et al. Occurrence of Corynebacterium striatum as an emerging antibiotic-resistant nosocomial pathogen in a Tunisian hospital. Sci Rep. Nature Publishing Group; 2017;7: 9704.

41. Huntzinger E, Boisset S, Saveanu C, Benito Y, Geissmann T, Namane A, et al. Staphylococcus aureus RNAIII and the endoribonuclease III coordinately regulate spa gene expression. EMBO J. 2005;24: 824–835.

42. Bohn C, Rigoulay C, Bouloc P. No detectable effect of RNA-binding protein Hfq absence in Staphylococcus aureus. BMC Microbiol. 2007;7: 10.

43. Liu Y, Wu N, Dong J, Gao Y, Zhang X, Mu C, et al. Hfq is a global regulator that controls the pathogenicity of Staphylococcus aureus. PLoS One. 2010;5.

doi:10.1371/journal.pone.0013069

44. Ramos CG, Sousa SA, 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. 2011;193: 1515–1526.

Figure 1. Phylogenetic analysis of bacterial species with multiple Hfq paralogs.

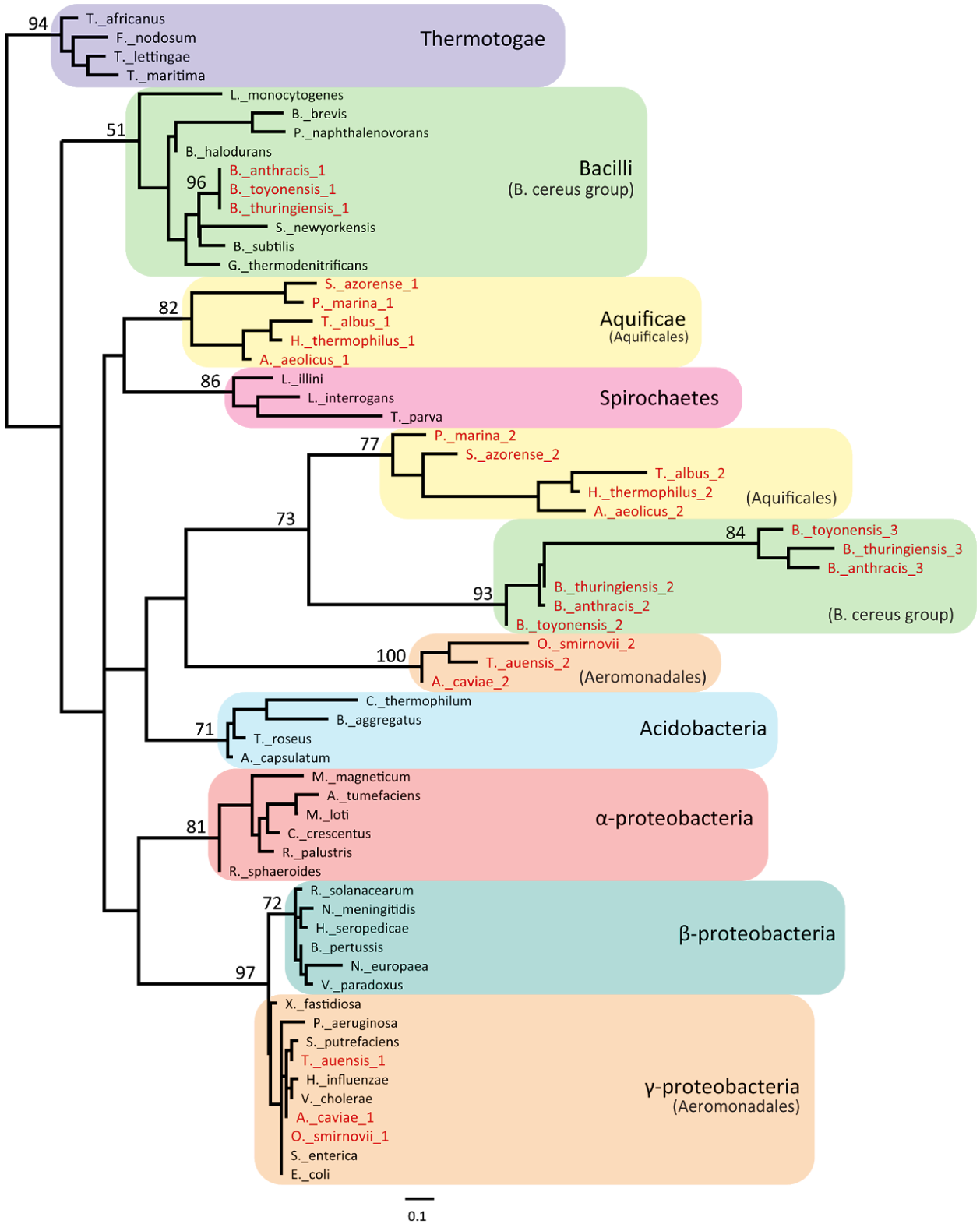


Figure 1. Phylogenetic analysis of bacterial species with multiple Hfq paralogs. Corresponding phyla to which representative Hfq sequences belong are indicated by the colored boxes unless otherwise stated: Thermotogae (navy), Bacilli (green), Aquificae (gold), Spirochaetes (magenta), Acidobacteria (blue), α-proteobacteria (red), β-proteobacteria (teal), and γ-proteobacteria (orange). Species with multiple Hfq paralogs are labelled in red and the taxonomic rank to which these species belong is provided in parentheses. The scale bar represents the number of mutations per site. Bootstrap support values, calculated from 1000 bootstrapping replicates, are shown at basal nodes.

Figure 2. Desulfurobacteriales Hfq2 and Hfq3 feature elongated N-terminal regions

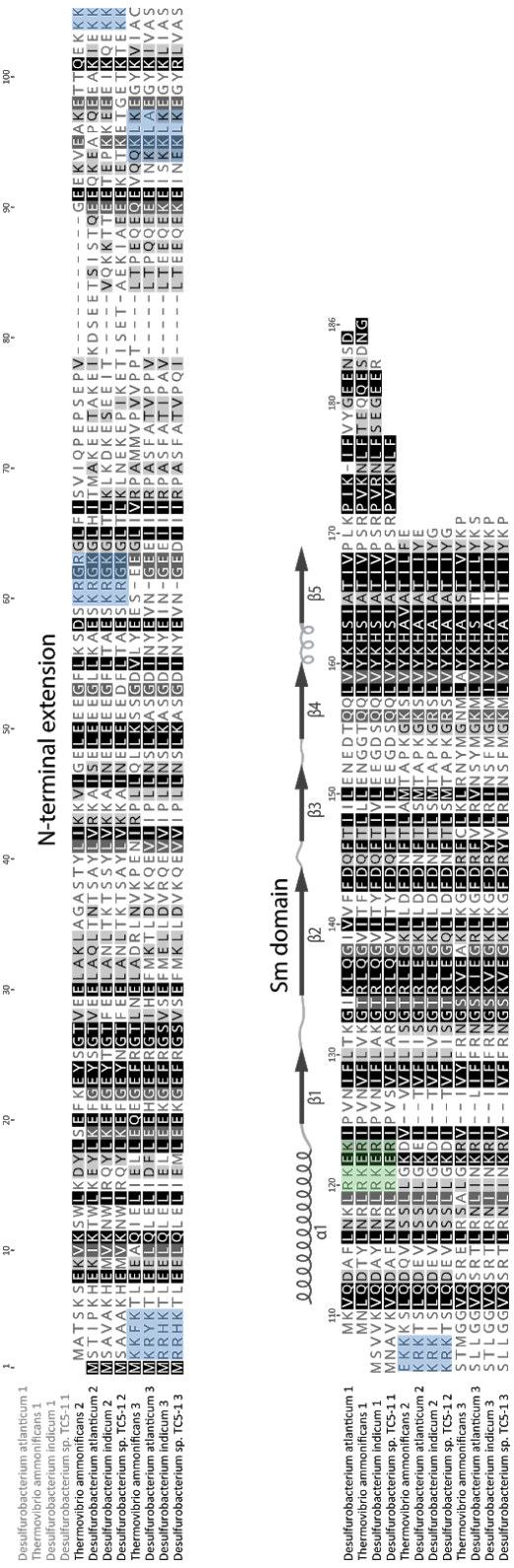


Figure 2. Desulfurobacteriales Hfq2 and Hfq3 feature elongated N-terminal regions. Hfq1, Hfq2 and Hfq3 sequences from *Desulfurobacterium atlanticum*, *Desulfurobacterium indicum*, *Desulfurobacterium sp. TC5-1*, and *Thermovibrio ammonificans* were aligned using MUSCLE. The blue boxes highlight 'KKNK' motifs present on the ~40 residue N-terminal tails of Hfq2 and Hfq3. The green box highlights the 'RRER' signature motif found at the lateral rim of Hfq. Secondary structural elements of the Sm domain are shown above as cartoon. Residues are colored according similarity to the consensus: black boxes with white text, 100% similar; dark grey box with with white text, 80-100% similar; light grey box with black text, 60-80% similar; white box with grey text, <60% similar. Note that this 'RRER' motif is conserved in Hfq1 homologs only.

Figure 3. Horizontal gene transfer among Actinobacterial Hfqs.

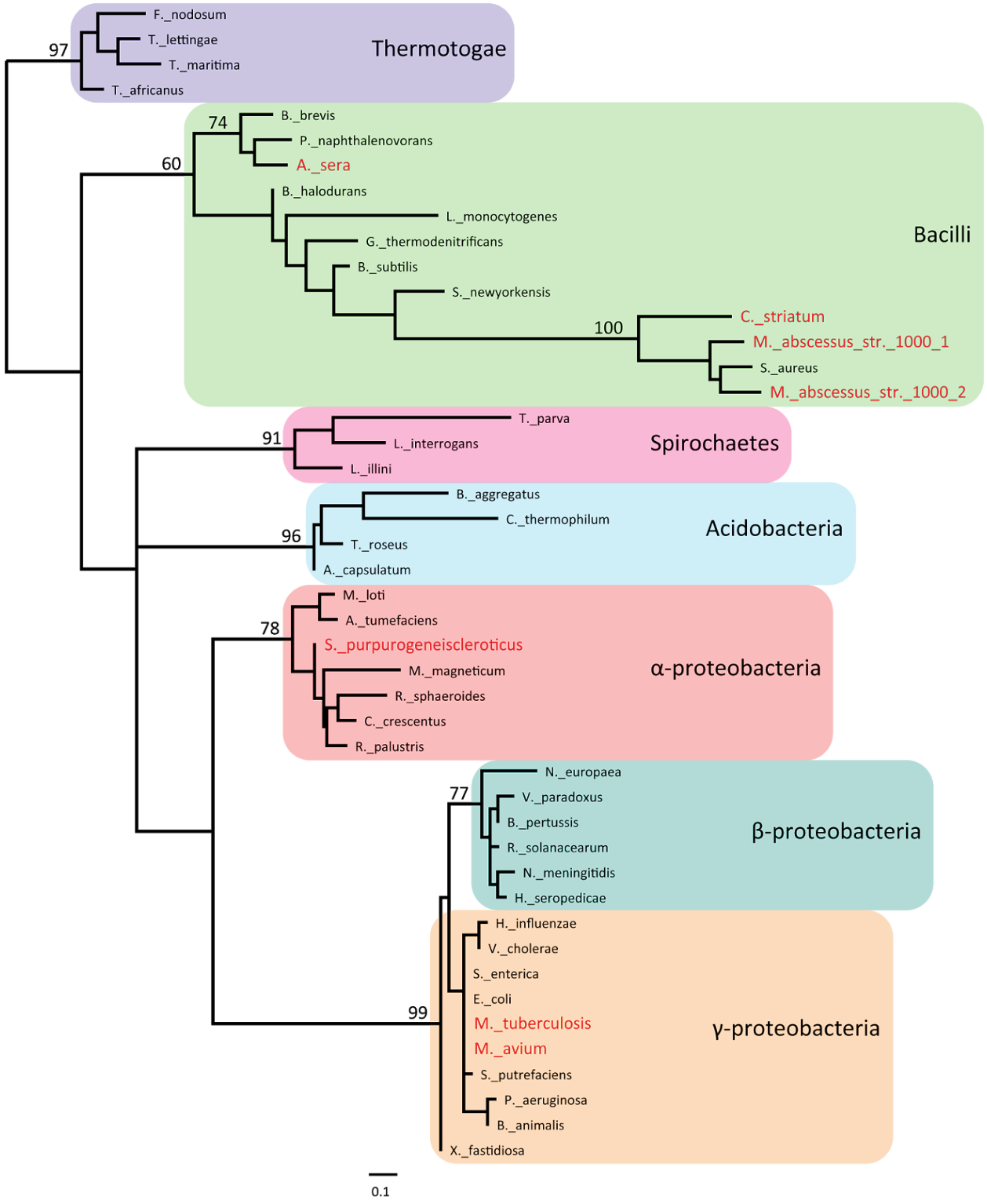


Figure 3. Horizontal gene transfer among Actinobacterial Hfqs. Corresponding phyla to which representative Hfq sequences belong are indicated by the colored boxes unless otherwise stated: Thermotogae (navy), Bacilli (green), Spirochaetes (magenta), Acidobacteria (blue), α-proteobacteria (red), β-proteobacteria (teal), and γ-proteobacteria (orange). Hfq sequences from the following species, representing the phylum actinobacteria, are labelled in red: *Actinocatenispora sera*, *Corynebacterium striatum*, *Mycobacterium abscessus subsp. Abscessus str. 1000* (Hfq1 and Hfq2 are shown), *Streptococcus purpurogenisclerotis*, *Mycobacterium tuberculosis*, and *Mycobacterium avium*. The scale bar represents the number of mutations per site. Bootstrap support values, calculated from 1000 bootstrapping replicates, are shown at basal nodes. Note that Actinobacterial Hfq sequences are polyphyletic, and group with Bacilli, α-proteobacteria, or γ-proteobacteria, indicating multiple instances of HGT.

Figure 4. Phylogeny of Hfq homologs from two different *S. pneumoniae* strains.

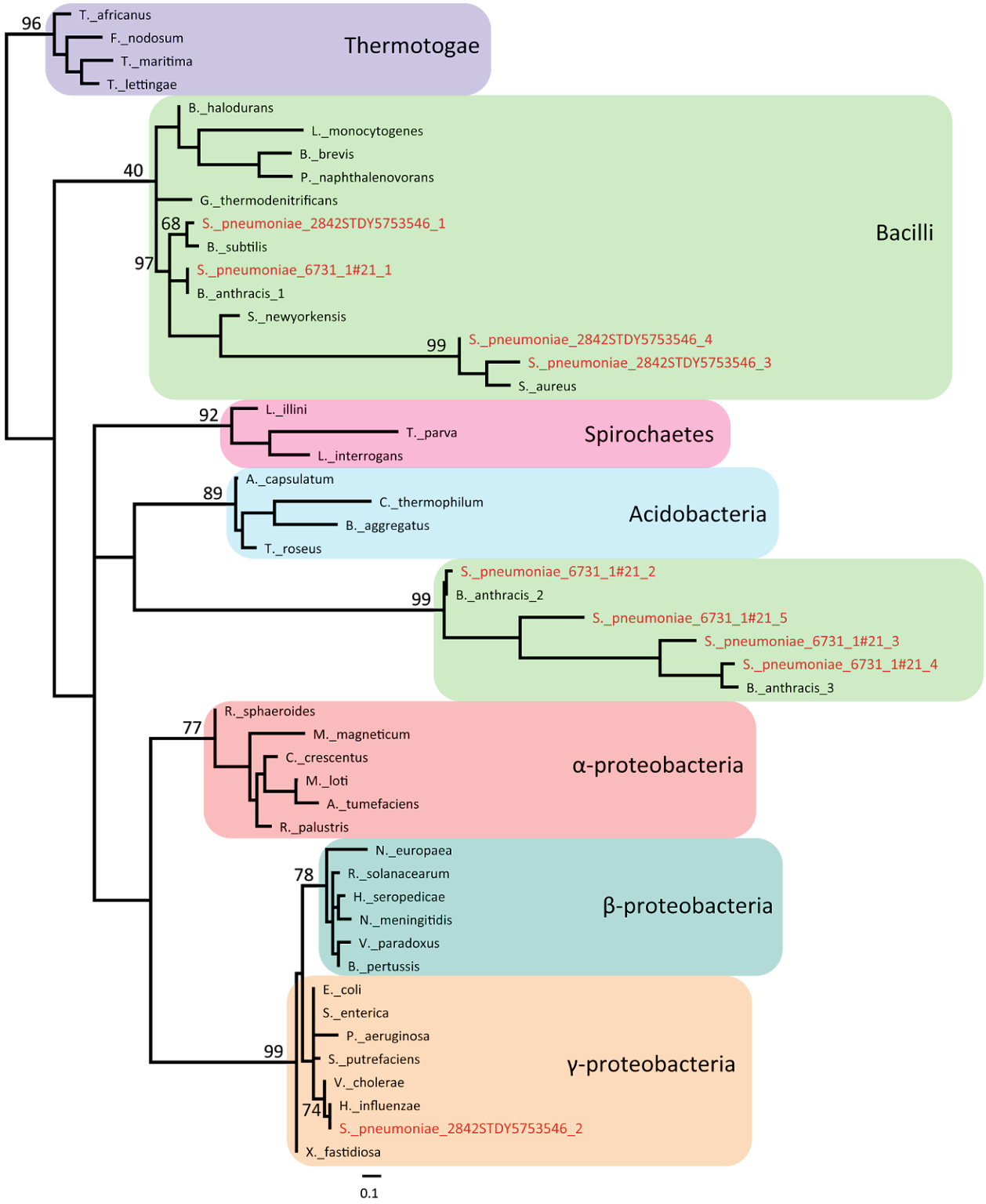


Figure 4. Phylogeny of Hfq homologs from two different *S. pneumoniae* strains. Corresponding phyla to which representative Hfq sequences belong are indicated by the colored boxes unless otherwise stated: Thermotogae (navy), Bacilli (green), Aquificae (gold), Spirochaetes (magenta), Acidobacteria (blue), α-proteobacteria (red), β-proteobacteria (teal), and γ-proteobacteria (orange). *Streptococcus pneumoniae* sequences are labelled in red. For this phylogenetic tree, two strains of *S. pneumoniae* were chosen, type strain: N (6731_#21), which has five Hfq paralogs, and strain 2842STDY57553546, which has four Hfq paralogs. The scale bar represents the number of mutations per site. Bootstrap support values, calculated from 1000 bootstrapping replicates, are shown at basal nodes. A combination of gene duplication and HGT is likely responsible for the great diversity of *S. pneumoniae* Hfq sequences.

Chapter 3: Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: conservation of the lateral RNA-binding mode*

Kimberly Stanek, Jennifer Patterson-West, Peter S. Randolph and Cameron Mura

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

* This chapter is a reprint of the following published article:

Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: conservation of the lateral RNA-binding mode. K. Stanek, J. Patterson-West, P.S. Randolph, & C. Mura. *Acta Cryst.* (2017), *D73*, 294–315

Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: Conservation of the lateral RNA-binding mode

Kimberly A Stanek[a], Jennifer Patterson-West[a], Peter S Randolph[a] and Cameron Mura[a]*

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

Correspondence email: cmura@muralab.org

Synopsis The structure of an Hfq homolog from the deep-branching thermophilic bacterium *Aquifex aeolicus*, determined to 1.5-Å resolution both in *apo* form and bound to a uridine-rich RNA, reveals a conserved, pre-organized RNA-binding pocket on the lateral rim of the Hfq hexamer.

Abstract The host factor Hfq, as the bacterial branch of the Sm family, is an RNA-binding protein involved in post-transcriptional regulation of mRNA expression and turnover. Hfq facilitates pairing between small regulatory RNAs (sRNA) and their corresponding mRNA targets by binding both RNAs and bringing them into close proximity. Hfq homologs self-assemble into homo-hexameric rings, with at least two distinct surfaces that bind RNA. Recently, another binding site—dubbed the 'lateral rim'—has been implicated in sRNA•mRNA annealing; the RNA-binding properties of this site appear to be rather subtle, and its degree of evolutionary conservation is unknown. An Hfq homolog has been identified in the phylogenetically deep-branching thermophile *Aquifex aeolicus* (*Aae*), but little is known about the structures and functions of Hfq from basal bacterial lineages such as the Aquificae. Thus, we have cloned, overexpressed, purified, crystallized, and biochemically characterized *Aae* Hfq. We have determined the structures of *Aae* Hfq in space-groups *P*1 and *P*6, both to 1.5 Å resolution, and we have discovered nanomolar-scale binding affinities for uridine- and adenosine-rich RNAs. Co-crystallization with U_6 RNA reveals that the outer rim of the *Aae* Hfq hexamer features a well-defined binding pocket that is selective for uracil. This *Aae* Hfq structure, combined with biochemical and biophysical characterization of the homolog, reveals deep evolutionary conservation of the lateral RNA-binding mode, and lays a foundation for further studies of Hfq-associated RNA biology in ancient bacterial phyla.

Keywords: Hfq; Sm protein; RNA; *Aquifex aeolicus*; hexamer; evolution

1. Introduction

The bacterial protein Hfq, initially identified as an *E. coli* host factor required for the replication of RNA bacteriophage Qβ (Franze de Fernandez *et al.*, 1968, Franze de Fernandez *et al.*, 1972), is now known to play a central role in the post-transcriptional regulation of gene expression and mRNA metabolism (Vogel & Luisi, 2011, Sauer, 2013, Updegrove *et al.*, 2016). Hfq has been linked to many RNA-regulated cellular pathways, including stress response (Sledjeski *et al.*, 2001, Zhang *et al.*, 2002, Fantappie *et al.*, 2009), quorum sensing (Lenz *et al.*, 2004), and biofilm formation (Mandin & Gottesman, 2010, Mika & Hengge, 2013). The diverse cellular functions of Hfq stem from its fairly generic role in binding small, non-coding RNAs (sRNA) and facilitating base-pairing interactions between these regulatory sRNAs and target mRNAs. A given sRNA might either upregulate (Soper *et al.*, 2010) or downregulate (Ikeda *et al.*, 2011) one or more target mRNAs via distinct mechanisms. For example, the sRNA RhyB downregulates several Fur-responsive genes under iron-limiting conditions (Masse & Gottesman, 2002), whereas the DsrA, RprA and ArcZ sRNAs stimulate translation of *rpoS* mRNA, encoding the stationary-phase σ^s factor (Soper *et al.*, 2010). In general, Hfq is required for cognate sRNA•mRNA pairings to be productive, and abolishing Hfq function typically yields pleiotropic phenotypes, including diminished viability (Fantappie *et al.*, 2009, Vogel & Luisi, 2011).

Hfq is the bacterial branch of the Sm superfamily of RNA-associated proteins (Mura *et al.*, 2013). Eukaryotic Sm and Sm-like (LSm) proteins act in intron splicing and other mRNA-related processing pathways (Will & Luhrmann, 2011, Tharun, 2009, Tycowski *et al.*, 2006), while the cellular functions of Sm homologs in the archaea remain unclear. Though the biological functions and amino acid sequences of Sm proteins vary greatly, the overall Sm fold is conserved across all three domains of life: five antiparallel β-strands form a highly bent β-sheet, often preceded by an *N*-terminal α-helix (Fig 1; (Kambach *et al.*, 1999)). Sm proteins typically form cyclic oligomers via hydrogen bonding between the β4···β5′ (edge) strands of monomers in a head–tail manner, yielding a toroidal assembly of six (Hfq) or seven (other Sm) subunits (Mura *et al.*, 2013); Hfq and other Sm rings can further associate into head–head and head–tail stacked rings, as well as polymeric assemblies (Arluison *et al.*, 2006). The oligomerization mechanism also varies across the Sm superfamily: Sm-like archaeal proteins (SmAPs) and Hfq homologs spontaneously self-assemble into stable homo-heptameric or homo-hexameric rings (respectively) that resist chemical and thermal denaturation, whereas eukaryotic Sm hetero-heptamers form via a chaperoned biogenesis pathway. This intricate assembly pathway (Fischer *et al.*, 2011) involves staged interactions with single–stranded RNA (e.g. small nuclear RNAs of the spliceosomal snRNPs), such that RNA threads through the central pore of the Sm ring (Leung *et al.*, 2011). In contrast, Hfq hexamers expose two distinct RNA-binding surfaces (Mikulecky *et al.*, 2004), termed the 'proximal' and 'distal' (with respect to the α-helix) faces of the ring. These two surfaces can bind RNA independently and simultaneously (Wang *et al.*, 2013), with different RNA sequence specificities along each face.

The proximal face of Hfq preferentially binds uridine-rich single-stranded RNA (ssRNA) in a manner that is well-conserved amongst Gram-positive (Schumacher *et al.*, 2002, Kovach *et al.*, 2014) and Gram-negative bacteria (Weichenrieder, 2014). The binding region, located near the pore, consists of six equivalent ribonucleotide-binding pockets, and can thus accommodate a six-nucleotide segment of ssRNA. Each uracil base π-stacks with a conserved aromatic side-chain (Phe or Tyr) from the L3 loops of adjacent monomers (e.g., F42 in *E. coli*, corresponding to F40 in *A. aeolicus*; Fig 1), and nucleobase specificity is achieved via hydrogen bonding between Q8 and the exocyclic O2 of each uracil. (Unless otherwise noted, residue numbers refer to the *E. coli* Hfq sequence; for clarity, only the *Aae* numbering is shown in Fig 1.) A key physiological function of the proximal face of Hfq is thought to be the selective binding of the U–rich 3′-termini of sRNAs, resulting from *ρ*–independent transcription termination (Wilson & von Hippel, 1995). Hfq's recognition of these 3′ ends is facilitated by the well-conserved H57 of the L5 loop ('3_{10} helix' in Fig 1), which is well-positioned to interact with the unconstrained, terminal 3′-hydroxyl group (Sauer & Weichenrieder, 2011, Schulz & Barabas, 2014). This mode of recognition may also explain the ability of Hfq to bind specifically to sRNAs over DNA, or other RNAs.

In contrast to the uracil-binding proximal region, the distal face of Hfq preferentially binds adenine-rich RNA, with the mode of binding varying between Gram-negative and Gram-positive species. Hfq homologs from Gram-negative bacteria specifically recognize RNAs with a tri-nucleotide motif, denoted $(A–R–N)_n$, where A=adenine, R=purine, N=any nucleotide; this recognition element was recently refined to be a more restrictive $(A–A–N)_n$ motif (Robinson *et al.*, 2014). A-A-N–containing RNAs bind to a large surface region on the distal face, which can accommodate up to 18 nucleotides of a ssRNA (Link *et al.*, 2009), and such RNAs are recognized in a tripartite manner: *(i)* the first *A*-site is formed by residues between the β2 and β4 strands of one monomer (E33 ensures adenine specificity); *(ii)* the second *A site* lies between the β2 strands of adjacent subunits, and includes a conserved Y25 (Fig 1) that engages in π-stacking interactions; and *(iii)* a nonspecific nucleotide (*N*)–binding site bridges to the next A–A pocket. In contrast to this recognition mechanism, the distal face of Gram-positive Hfq recognizes a bipartite adenine–linker $(AL)_n$ motif. This structural motif features an *A*-site that is similar to the first *A*-site of Gram-negative bacteria; in addition, a nonspecific nucleotide-binding pocket acts as a linker (*L*) site, allowing 12 nucleotides to bind in a circular fashion atop this face of the hexamer (Horstmann *et al.*, 2012, Someya *et al.*, 2012). The ability of the distal face to specifically bind A-rich regions, such as the long, polyadenylated 3′-tails of mRNAs (Folichon *et al.*, 2003), leads to several links between Hfq and mRNA degradation/turnover pathways (Mohanty *et al.*, 2004, Bandyra & Luisi, 2013, Regnier & Hajnsdorf, 2013). The general capacity of Hfq to independently bind RNAs at the proximal and distal sites brings these distinct RNA species into close proximity, as part of an sRNA•Hfq•mRNA ternary complex. Indeed, a chief cellular role of Hfq is the

productive annealing of RNA strands in this manner, for whatever downstream physiological purpose (be it stimulatory or inhibitory).

Independent binding of RNAs at the proximal/distal sites elucidates only part of what is known about Hfq's RNA-related activities. For instance, Hfq has been shown to protect internal regions of sRNA (Balbontín *et al.*, 2010, Ishikawa *et al.*, 2012, Updegrove & Wartell, 2011, Zhang *et al.*, 2002) and to reduce the thermodynamic stability (ΔG°_{fold}) of some RNA hairpins (Robinson *et al.*, 2014), but current mechanistic models for Hfq activity do not account for all of these properties. In addition, recent studies have identified a new RNA-binding site on the Hfq ring, beyond the proximal and distal sites (Sauer, 2013). This third site, located on the outer rim of the Hfq toroid and presaged in RNA-binding studies a decade ago (Sun & Wartell, 2006), is variously termed the 'lateral', 'rim' or 'lateral rim' site (the terms are used synonymously herein). Mutational analyses reveal that an arginine-rich patch near the *N*-terminal α-helix, containing the segment $R^{16}R^{17}E^{18}R^{19}$ in *E. coli*, facilitates rapid annealing of Hfq-bound mRNAs and sRNAs (Panja *et al.*, 2013). These arginine residues, along with conserved aromatic (F/Y39; 'φ' in Fig 1) and basic (K47) residues, look to be vital for the binding of full-length sRNAs to Hfq (Sauer *et al.*, 2012). Further understanding of the precise mechanism of RNA binding to the lateral rim site (and any base specificity at that site) has been hindered by a lack of structural information on Hfq_{rim}···RNA interactions. A recent crystal structure of *E. coli* Hfq complexed with the full-length riboregulatory sRNA RydC (a regulator of biofilms and some mRNAs) revealed a potential binding pocket formed by N13, R16, R17 and F39, and capable of accommodating two nucleotides of uridine (Dimastrogiovanni *et al.*, 2014); however, the exact positioning and geometry of the nucleotides were not discernible at the resolution (3.5 Å) of that model.

Our current mechanistic knowledge of Hfq···RNA interactions is based primarily on homologs from proteobacterial species, particularly the γ-proteobacteria *E. coli* and *Pseudomonas aeruginosa*; structural information about nucleotide-binding at the lateral site is available only from these two species. We do not know if the rim RNA-binding mode is conserved in homologs from other bacterial species, or perhaps even more broadly (in archaeal and eukaryotic lineages). Hfq orthologs from phylogenetically deep-branching bacteria, such as *Aquifex aeolicus* (*Aae*), may help clarify the degree of conservation of Hfq's various RNA-binding surfaces, including the lateral rim. *Aae* Hfq has been shown, via immunoprecipitation/deep-sequencing studies, to partially restore the phenotype of a *Salmonella enterica* Hfq knock-out strain, *Δhfq* (Sittka *et al.*, 2009), but nothing else is known about the RNA-binding properties of *Aae* Hfq. Precisely positioning *Aae* within the bacterial phylogeny is difficult given, for instance, that many *Aae* genes are similar to those in ε-proteobacteria (Eveleigh *et al.*, 2013). Nevertheless, 16S rRNA and genomic sequencing data firmly place *Aae*, along with other members of the Aquificales order, among the deepest branches in the bacterial tree—near the bacterial/archaeal divergence. Sequence similarity to proteobacterial genes has been attributed to extensive lateral gene transfer (Oshima *et al.*, 2012, Boto, 2010); importantly, extensive lateral transfer does not

seem to be an issue with Hfq homologs (Sun *et al.*, 2002), and Sm proteins likely have a single, well-defined origin (Veretnik *et al.*, 2009).

Here, we report the crystal structure and RNA-binding properties of an *A. aeolicus* Hfq ortholog. *Aae* Hfq crystallized in multiple space-groups, with both hexameric and dodecameric assemblies in the lattices. These oligomeric states were further examined in solution, via chemical crosslinking assays, analytical size-exclusion chromatography, and light-scattering experiments. We found that *Aae* Hfq binds uridine– and adenosine–rich RNAs with nanomolar affinities *in vitro*, and that the inclusion of Mg^{2+} enhances binding affinities by factors of ≈2× (A-rich) or ≈10× (U-rich). Co-crystallization of *Aae* Hfq with U_6 RNA reveals well-defined electron density (to 1.5 Å) for at least two ribonucleotides in a rim site, suggesting that this auxiliary RNA-binding site is conserved even amongst evolutionarily ancient bacteria. Finally, comparative structural analysis reveals that *(i)* the spatial pattern of Hfq··· RNA interatomic contacts, which effectively defines the rim site, is preserved between *Aae* and *E. coli*, and *(ii)* the residues comprising the *Aae* Hfq rim site are pre-organized for U–rich RNA binding.

2. Materials and Methods

2.1. Cloning, expression and purification of *Aae* Hfq

The *Aae hfq* gene was cloned via the polymerase incomplete primer extension (PIPE) methodology (Klock & Lesley, 2009), using an *A. aeolicus* genomic sample as a PCR template. The T7-based expression plasmid pET-28b(+) was used, yielding a recombinant protein construct bearing an *N*-terminal His6×-tag and a thrombin-cleavable linker preceding the Hfq (Supp Fig S1a, Supp Table S1); in all, the affinity tag and linker extend the 80-aa native sequence by 20 residues, giving the full-length sequence in Supp Fig S1a. Plasmid amplification, and *in vivo* ligation of the vector and insert, were achieved via transformation of the PIPE products into chemically competent TOP10 *E. coli* cells. Recombinant *Aae* Hfq was produced by transforming the plasmid into the BL21(DE3) *E. coli* expression strain, followed by outgrowth in Luria-Bertani media at 310 K. Finally, expression of *Aae* Hfq from the T7*lac*-based promoter was induced by the addition of 1 mM isopropyl-β-D-thiogalactoside (IPTG) when the optical density, measured at 600 nm (OD_{600}) reached ≈ 0.8–1.0. Cell cultures were then incubated at 310 K, with shaking (≈230rpm), for an additional four hours, pelleted at 15,000*g* for 5 minutes at 277 K, and then stored at 253 K overnight.

Cell pellets were re-suspended in a solubilisation and lysis buffer (50 mM Tris pH 7.5, 750 mM NaCl, 0.4 mM PMSF, and 0.01 mg/ml chicken egg white lysozyme (Fisher)) and incubated at 310 K for 30 min. Cells were then mechanically lysed using a microfluidizer. To clarify cell debris, the lysate was pelleted via centrifugation at 35,000*g* for 20 min at 277 K. The supernatant from this step was then incubated at 348 K for 20 min, followed by centrifugation at 35,000*g* for 20 min; this heat-cut step was performed because most Hfq homologs examined thus far have been thermostable, and

because *A. aeolicus* is a hyperthermophile (optimum growth temperature, $T_{opt} \approx 360$ K (Huber & Eder, 2006)). To reduce contamination by any spurious *E. coli* nucleic acids, which have been known to co-purify with other Hfqs, the clarified supernatant from the heating step was treated with high concentrations (≈ 6 M) of guanidinium hydrochloride (GndCl). To remove any particulate matter, Gnd-treated samples were then immediately clarified by 0.2-µm syringe filtration.

Recombinant *Aae* Hfq was then purified via immobilized metal affinity chromatography (IMAC), using a Ni^{2+}–charged iminodiacetic acid–sepharose column with an NGC (BioRad) medium-pressure liquid chromatography system. After loading the clarified supernatant from the heat-cut and GndCl treatment steps, the column was treated with four column volumes of wash buffer (50 mM Tris pH 8.5, 150 mM NaCl, 6 M GndCl, 10 mM imidazole). Next, *Aae* Hfq was eluted by applying a linear gradient, from 0–100% over 10 column volumes, of elution buffer (identical to the wash buffer, but with 600 mM imidazole). Protein-containing fractions, as assessed by the absorbance at 280 nm and chromatogram elution profiles, were then combined and, in order to remove GndCl, dialyzed against a buffer of 25 mM Tris pH 8.0, 1 M arginine. Next, to prepare for removal of the His6×-tag, the protein was then dialyzed into 50 mM Tris pH 8.0, 500 mM NaCl and 12.5 mM EDTA. The *Aae* Hfq sample was subjected to proteolysis with thrombin, at a 1:600 Hfq:thrombin ratio (by mass), by incubating at 315 K overnight (≈16 h), followed by application to a benzamidine affinity column to remove the thrombin. To improve sample homogeneity, *Aae* Hfq was further purified over a preparative-grade gel-filtration column containing *Superdex™ 200 Increase* resin; *Aae* Hfq eluted as a single, well-defined peak. Chromatographic steps were conducted at room temperature; lengthier incubation steps, such as dialysis, were carried out at 310 or 315 K throughout the purification, as *Aae* Hfq samples were found to be relatively insoluble over a few hours at room temperature (≈ 295 K).

Aae Hfq sample purity was generally assayed via SDS-PAGE gels or matrix assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). Samples were prepared for MALDI by diluting 1:4 (v/v) with 0.01% v/v trifluoroacetic acid (TFA) and then spotting on a steel MALDI plate in a 1:1 v/v ratio with a matrix solution (15 mg/ml sinapinic acid in 50% acetonitrile, 0.05% TFA); this mixture crystallized *in situ* via solvent evaporation. Mass spectra were acquired on a Bruker MicroFlex instrument operating in linear, positive-ion mode (25 kV accelerating voltage; 50-80% grid voltage), and final spectra were the result of averaging at least 50 laser shots. Two sets of molecular weight calibrants were used for low (4–20 kDa) and high (20–100 kDa) m/z ranges. Purification progress and sample MALDI spectra are illustrated in Supp Fig S1b and Fig 2, respectively.

2.2. Crosslinking assays

Purified *Aae* Hfq was chemically crosslinked, using formaldehyde, in a so-called 'indirect' (vapor diffusion–based) method (Fadouloglou *et al.*, 2008). First, *Aae* Hfq samples at 0.6 mg/ml were dialyzed into a buffer consisting of 25 mM HEPES pH 8.0 and 500 mM NaCl. Reaction solutions were

prepared in 24-well Linbro plates using micro-bridges (Hampton Research). Immediately before use, 5 N HCl was added to 25% w/v formaldehyde in a 1:40 v/v ratio. Next, 40 µl of this acidified formaldehyde solution was added to the micro-bridge, and 15 µl of the 0.6 mg/ml *Aae* Hfq was added to a silanized coverslip. Greased wells were then sealed by flipping over the coverslips and the reaction was incubated at 310 K for 40 min. Reactions were quenched by the addition of a primary amine; specifically, 5 µl of 1 M Tris pH 8.0 was mixed into the 15 µl protein droplet. Crosslinked samples were then desalted on a C4 resin (using *ZipTip*® pipette tips) in preparation for analysis via MALDI-TOF MS, as described above.

2.3. Analytical size-exclusion chromatography and multi-angle static light scattering

Analytical size-exclusion chromatography (AnSEC) was performed with a pre-packed *Superdex 200 Increase* 10/300 GL column and a Bio-Rad NGC™ medium-pressure liquid chromatography system. Prior to AnSEC, all protein samples were dialyzed into a running buffer consisting of 50 mM Tris pH 8.0 and 200 mM NaCl. In separate experiments, *Aae* Hfq samples (250 µM protein) were mixed 1:1 v/v with RNA sequences (at 50 µM), denoted 'U_6' (5′-monophosphate–r(U)$_6$–3′-OH) or 'A_{18}' (5′-monophosphate–r(A)$_{18}$–3′-OH), and equilibrated by incubation at 310 K for 1 h prior to loading onto the AnSEC column. Elution volumes were measured by simultaneously monitoring the absorbance at 260 nm (RNA) and at 280 nm (protein). A standard curve was generated using the Sigma gel-filtration markers kit, with calibrants in the 12–200 kDa molecular weight range: cytochrome c (12.4 kDa), carbonic anhydrase (29 kDa), bovine serum albumin (66 kDa), alcohol dehydrogenase (150 kDa) and β-amylase (200 kDa); blue dextran was used to calculate the void volume, V_0.

To determine absolute molecular masses (i.e., without reference standards and implicit assumptions about spheroidal shapes), and in order to assess potential polydispersity of *Aae* Hfq in solution, multi-angle static light scattering (MALS) was used in tandem with size-exclusion chromatographic (SEC) separation. A flow-cell–equipped light scattering (LS) detector was used downstream of the SEC, in-line with an absorbance detector (UV) and a differential refractive index (RI) detector. In our SEC–UV/RI/LS system, *(i)* the SEC step serves to fractionate a potentially heterogeneous sample (giving the usual chromatogram, recorded at either 280 or 260 nm on a Waters UV/vis detector), *(ii)* the differential refractometer (RI) estimates the solute concentration via changes in solution refractive index (i.e., dn/dc), and *(iii)* the LS detector measures the excess scattered light. This workflow was executed on a Waters HPLC system equipped with the Wyatt instrumentation noted below, and utilized the same column (*Superdex 200*) and solution buffer conditions as described immediately above. LS measurements were taken at three detection angles, using a Wyatt miniDAWN TREOS (λ = 658 nm), and the differential refractive index was recorded from a Wyatt Optilab T-rEX. This enables the molecular mass of the solute in each fraction to be determined because the amount of light scattered (from the LS data) scales with the weight-averaged molecular masses (desired quantity) and solute

concentrations (from the RI data); if multiple species exist in a given (heterogeneous) fraction, the polydispersity can be quantified as the ratio of the weight-averaged (M_w) and number-averaged molar masses (M_n). Data were processed and analysed using Wyatt's ASTRA software package, applying the Zimm formalism to extract the weight-averaged molecular masses (Folta-Stogniew, 2009).

2.4. Fluorescence polarization–based binding assays

RNA–binding affinities were determined via fluorescence anisotropy/polarization experiments (FA/FP; (Pagano *et al.*, 2011)), using fluorescein-labelled oligoribonucleotides. In particular, the RNA probes 5′-FAM–r(U)$_6$–3′-OH (FAM–U$_6$) and 5′-FAM–r(A)$_{18}$–3′-OH (FAM–A$_{18}$) were used, with 6-carboxyfluorescein amidite (FAM) modification of the 5′ ends; the FAM label features absorption and emission wavelengths, λ_{max}, of 485 nm (excitation) and 520 nm (detection), respectively. FAM-labelled RNAs at 5 nM were added to a serially-diluted concentration series of purified *Aae* Hfq (in 50 mM Tris pH 8.0, 500 mM NaCl), and allowed to equilibrate for 45 min at room temperature. The highest [Hfq] was 30 μM (in terms of monomer), and a total of 18 serial dilutions were performed to produce datasets such as Fig 4. For binding assays that were supplemented with Mg^{2+}, a 1 M $MgCl_2$ stock solution was used and the final [Mg^{2+}] in the binding reaction was 10 mM.

The fluorescence polarization, P, is measured as $P = \frac{I_{\parallel}-I_{\perp}}{I_{\parallel}+I_{\perp}}$, where $I_{\parallel}$ and $I_{\perp}$ are the emitted light intensities in directions parallel and perpendicular to the excitation plane, respectively. FP data were recorded on a PheraSTAR spectrofluorimeter equipped with a plate reader (BMG Labtech), and values from three independent trials were averaged. The effective polarization, in units of millipolarization (*mP*), was plotted against log[(Hfq)$_6$]. Binding data were fit, via nonlinear least-squares regression, to a logistic functional form of the classic sigmoidal curve for saturable binding. Specifically, the four-parameter equation

$$y(x) = A_2 + (A_1 - A_2)\left[\frac{1}{1+e^{(x-x_0)/dx}}\right] \qquad (1)$$

was used, where the independent variable x is the log[(Hfq)$_6$] concentration at a given data point and the fit parameters are: *(i)* A_1, the polarization at the start of the titration (unbound; lower plateau of the binding isotherm); *(ii)* A_2, the final polarization at the end of the titration (saturated binding; upper plateau); *(iii)* x_0, the apparent equilibrium dissociation constant ($K_{D,\,app}$) for the binding reaction in terms of log[(Hfq)$_6$]; and *(iv)* a parameter, dx, giving the characteristic scale/width over which the slope of the sigmoid changes. In this formulation, dx is essentially the classic Hill coefficient, measuring the steepness of the binding curve; the greater the magnitude of dx, the narrower the transition region. In addition to fitting the binding data with the four-parameter logistic model (Eq 1), a simpler, three-parameter model was also applied, with the following functional form:

$$y(x) = A_2 + (A_1 - A_2)\left[\frac{(P_t+\mathcal{L}_t+K_D)-\sqrt{(P_t+\mathcal{L}_t+K_D)^2-4P_t\mathcal{L}_t}}{2\mathcal{L}_t}\right], \qquad (2)$$

where the terms A_1 and A_2 are as above (Eq 1), K_D is the dissociation constant (x_0, above), and the variables $\mathcal{L}_t$ and P_t are the total concentrations of ligand (the FAM-labelled RNA) and receptor (here, taken as an Hfq hexamer), respectively. Though assuming 1:1 stoichiometry between *Aae* $(Hfq)_6$ and RNA, and not capturing potential cooperativity between possibly multiple ligand-binding sites, this second model does account for the effects of receptor depletion on the fitted K_D values. This, in turn, is an important consideration in fitting data points with abscissas near (within ≈10× of) the true K_D, as the assumption that the concentration of ligand•receptor complex, $[\mathcal{L}{\bullet}P]$, is far lower than the total concentrations of each species ($[\mathcal{L}]_t$, $[P]_t$) is violated if $[P]_t \approx K_D$. That is, free $[P] \cong [P]_t$ no longer holds near the K_D. Despite the advantage of accounting for receptor depletion, note that this treatment implicitly takes the Hill coefficient (the 'slope factor' for the transition region) to be one, rather than letting it vary (as in Eq 1); indeed, the only three degrees of freedom with which to describe the binding curve are the upper and lower asymptotes, and the midpoint of the transition (i.e., K_D, or 'x_0' of Eq 1). Assuming a Hill coefficient of unity and a simple (1:1 stoichiometry) $\mathcal{L} + P \rightleftarrows \mathcal{L}{\bullet}P$ equilibrium, one can show that neglecting to account for receptor-depletion phenomena gives an apparent (fitted) dissociation constant, $K_{D,\,app}$, that exceeds by ½$[P]_t$ the 'true' K_D computed via the Eq 2 treatment. For these reasons, both models—Equations 1 and 2—were considered in fitting the data. All calculations described in this section were performed with in-house code written in the R programming language, using the RSTUDIO integrated development environment.

2.5. X-ray crystallography

2.5.1. Crystallization

Prior to crystallization trials, purified *Aae* Hfq was dialyzed into a buffer of 50 mM Tris pH 8.0 and 500 mM NaCl, and concentrated to 4.0 mg/ml. Protein samples were typically stored at 310 K, to retain solubility, and used within two weeks of purification. All crystallization trials were performed using the vapour diffusion method in sitting-drop format. Sparse-matrix screening (Jancarik & Kim, 1991) yielded initial leads (visible crystals) under several conditions, and these were then optimized by adjusting the concentration of protein and precipitating agent, as well as the pH of the mother liquor. Diffraction-grade crystals (Supp Figs S1c, d) were reproducibly obtained with 0.1 M sodium cacodylate pH 5.5, 5% w/v PEG-8000, and 40% v/v 2-methyl-2,4-pentanediol (MPD) as the crystallization buffer. In our final condition, 6-µl sitting drops (3 µl well + 3 µl of 4 mg/ml *Aae* Hfq) were equilibrated, at 291 K, against 600-µl wells containing the crystallization buffer. Initial micro-crystals developed over several days. Optimization of the above condition via additive screens (Hampton Research) led to the discovery of several compounds that, in a 1:4 v/v additive:crystallization buffer ratio, slowed nucleation and increased crystal size. The optimized crystals grew to average dimensions of 50×50×10 µm/edge within 2 weeks and adopted cubic or hexagonal plate morphologies. Three particularly useful additives, used in subsequent crystallization trials, were: *(i)* 0.1 M hexammineco-

balt(III) chloride, $[Co(NH_3)_6]Cl_3$, *(ii)* 1.0 M GndCl and *(iii)* 5% w/v of the non-ionic detergent *n*-octyl-β-D-glucoside. The final *apo*-form *Aae* Hfq crystals were obtained with additive *(i)*; details are provided in Supp Table S2. *Aae* Hfq also was co-crystallized with a U-rich RNA (U_6), under the above crystallization conditions and supplemented with additive *(ii)* instead of additive *(i)*; these crystals were obtained by first incubating the purified protein with 500 μM 5′-monophosphate–r$(U)_6$–3′-OH (hereafter denoted 'U_6'), in a 1:1 ratio, at 310 K for 1 h prior to setting-up the crystallization drop.

2.5.2. Diffraction data collection and processing

The crystallization conditions described above adequately protected *Aae* Hfq crystals against ice formation upon flash-cooling (presumably because of the MPD), making it unnecessary to transfer crystals to an artificial mother liquor/cryo-protectant. Crystals were harvested using nylon loops and flash-cooled with liquid nitrogen. Diffraction data were collected at the Advanced Photon Source (APS) beamlines 24-ID-E and 24-ID-C for the *apo* and U_6-bound crystal forms, respectively. Initial data-processing steps—indexing/integrating, scaling and merging reflections—were performed in XDS (Kabsch, 2010). Space-group assignments and unit cell determinations utilized POINTLESS from the CCP4 suite (Winn *et al.*, 2011). Cell dimensions for the *apo* form (*P*1) are a = 63.46 Å, b = 66.06 Å, c = 66.10 Å, α = 66.05°, β = 83.94°, γ = 77.17°, and the U_6–co-crystals (*P*6) have $a = b$ = 66.19 Å, c = 34.21 Å.

2.5.3. Structure solution, refinement and validation

Initial phases for the diffraction data-sets for both crystal forms were obtained via molecular replacement (MR). Specifically, the PHASER (McCoy *et al.*, 2007) software was used, with the *P. aeruginosa* (*Pae*) hexamer structure (PDB 1U1S) as a search model for phasing of both crystal forms (*Aae* and *Pae* Hfq share high sequence similarity; see Fig 1). Note that initial phases for the *P*1 and *P*6 *Aae* crystal forms were obtained independently of one another, *i.e.* via parallel MR efforts. For the *P*1 (*apo*) form, with 12 monomers/unit cell (indicative of two hexamers), the calculated Matthew's coefficient (V_M) is 2.06 $Å^3$/Da, corresponding to a solvent content of 40.21% by volume. For the *P*6 (U_6-bound) form, only one monomer/ASU is feasible, with a V_M = 2.28 $Å^3$/Da and a 46.08% solvent content. These and related characteristics of the diffraction data are summarized in Table 1.

After obtaining initial MR solutions in PHASER, the correct *Aae* Hfq amino acid sequence was built and side-chains completed in a largely automated manner, using the PHENIX suite's AUTOBUILD functionality (Adams *et al.*, 2010). Individual solvent molecules, including H_2O, MPD, and Gnd, were added in a semi-automated manner (i.e., with visual inspection and manual adjustment) after the initial stages of refinement. Refinement of atomic positions, occupancies and atomic displacement parameters (ADPs)—either as isotropic '*B*-factors' or as full anisotropic ADPs—proceeded over several rounds in PHENIX. Some early refinement steps included simulated annealing optimization of coordinates, via molecular dynamics in torsion-angle space, as well as refinement of translation-

libration-screw (TLS) parameters to account for anisotropic disorder of each subunit chain (one TLS group defined per monomeric Hfq subunit). These steps yielded R_{work}/R_{free} values of 0.194/0.212 and 0.212/0.223 for the *P*1 and *P*6 datasets, respectively. The diffraction limits of the *P*1 and *P*6 forms—1.49 Å and 1.50 Å, respectively—occupy an intermediate zone, between the atomic-resolution ($d \lesssim 1.4$ Å) and medium resolution ($d \gtrsim 1.7$ Å) limits whereupon clearer decisions can be made as to the treatment of *B*-factors (Merritt, 2012). For instance, a relatively simple model (fewer parameters/atom), featuring individual isotropic *B*-factors and one TLS group per chain, might be most justifiable at ≈1.6 Å, depending on the quality of the diffraction data, whereas a more complex *B* model with a greater number of parameters—e.g., full anisotropic ADP tensors, U^{ij}, one per atom—is likely to be statistically valid (and, indeed, advised) at resolutions better than ≈1.3 Å.

For both the *P*1 and *P*6 forms of *Aae* Hfq, a final *B*-factor model was chosen based on analyses of the data/parameter ratio (i.e., number of reflections/atom), Hamilton's generalized residual (Hamilton, 1965) and related criteria, as implemented in the *bselect* routine of the PDB_REDO code (Joosten *et al.*, 2012). The *P*1 and *P*6 data-sets contained 16.5 and 17.5 reflections per atom, respectively, making the anisotropic refinement problem nearly two-fold overdetermined; PDB_REDO's unsupervised decision algorithm identified the fully anisotropic, individual *B*-factor model as being optimal. The structural models resulting from various ADP refinement strategies were assessed using the protein anisotropic refinement validation and analysis tool (PARVATI; (Zucker *et al.*, 2010)). In the final refinement stages for both *Aae* Hfq crystal forms, *P*1 (*Z*=12 monomers/cell) and *P*6 (*Z*=6 monomers/cell), full anisotropic *B*-factor tensors were refined individually for virtually every atom. (A small fraction of atoms in both the *P*1 and *P*6 models were treated isotropically, i.e. by refining individual B_{iso} values; most of these atoms, selected based on per-atom statistical tests in PDB_REDO, were either water or heteroatoms [e.g., Gnd in *P*1, PEG in *P*6].) At no point in the refinement were NCS restraints or constraints imposed for the 12 subunits in the *P*1 cell. All refinement steps involving visual inspection and manual adjustment of the model were done in COOT (Emsley *et al.*, 2010).

After the correct protein sequence had been built and refined against the *P*6 dataset, at least two complete nucleotides of U_6 RNA—including three phosphate groups—were clearly visible in σ_A-weighted difference electron-density maps ($mF_o - DF_c$). Ribonucleotides were built into electron density using the RCRANE utility (Keating & Pyle, 2010), after an initial round of refinement of coordinates, occupancies and individual *B*-factors in PHENIX. Validation of the final structural models included *(i)* inspection of the Ramachandran plot, via PROCHECK (Laskowski *et al.*, 1993); *(ii)* assessment of nonbonded interactions and geometric packing quality, via ERRAT (Colovos & Yeates, 1993); *(iii)* analysis of sequence/structure compatibility, via the profile–based method of VERIFY3D (Eisenberg *et al.*, 1997); and, finally, *(iv)* detailed stereochemical/quality checks with the MOLPROBI-

TY software (Chen *et al.*, 2010). Final structure determination and model refinement statistics are provided in Table 2.

2.6. Sequence and structure analyses

Sequences of verified Hfq homologs, drawn from diverse bacterial phyla, were selected for alignment and analysis against *Aae* Hfq. Here, we take 'verified' to mean that the putative Hfq homolog, from the published literature, has been identified via functional analysis or structural similarity (e.g., shown to adopt the Sm fold). Multiple sequence alignments were computed via two progressive alignment codes: *(i)* the multiple alignment using fast Fourier transform method (MAFFT; (Katoh & Standley, 2013)) and *(ii)* a sequence comparison approach using log-expectation scores for the profile function (MUSCLE; (Edgar, 2004)). The GENEIOUS bioinformatics platform (Kearse *et al.*, 2012) was used for some data/project-management steps and tree visualization purposes. Multiple sequence alignments (Fig 1) were processed using ESPRIPT (Gouet *et al.*, 1999), run as a command-line tool; the resulting PostScript source was then modified to obtain final figures. Iterative PSI-BLAST (Camacho *et al.*, 2009) searches against sequences in the PDB were used to identify homologous proteins as trial MR search models. *Pae* Hfq, with 46% pairwise identity to *Aae* Hfq (across 97% query coverage), exhibited the greatest sequence similarity (≈63%, at the level of BLOSUM62) and was therefore chosen as the initial MR search model.

Structural alignments were performed using a least-squares fitting algorithm (McLachlan, 1982) implemented in the program PROFIT (Martin & Porter, 2009). Multiple structural alignment of the 12 monomeric subunits in the *apo* form of *Aae* Hfq was used to create a mean reference structure, and each monomer was then aligned to that averaged reference. To assess 3D structural similarity between each of the $n(n–1)/2$ distinct pairs of monomers, a pairwise distance matrix was constructed by computing main-chain RMSDs between subunits i and j, giving matrix element (i, j). Agglomerative hierarchical clustering was performed on this distance matrix, using either the complete–linkage criterion or Ward's variance minimization algorithm with a Euclidean distance metric (Jain *et al.*, 1999); in-house code was written for these steps in both the R (within RSTUDIO) and Python languages.

Residues were assigned to secondary structural elements by a consensus approach, via visual inspection in PYMOL as well as the automated assignment tools DSSP and STRIDE; the precise borders can differ between these codes by a residue or two. Normal mode analyses of the *P*1 and *P*6 structures—taken as coarse-grained (C_α-only) representations and treated as anisotropic network models (ANM)—were performed with the PRODY/NMWIZ (Bakan *et al.*, 2011) plugin to VMD (Humphrey *et al.*, 1996). The ANM's Hessian matrix was built using default parameters for the force constant ($\gamma = 1$) and pairwise interaction cut-off distance (15 Å). Of the $3N–6$ nontrivial modes, displacements along the softest ≈ 20 vibrational modes, which correspond to low-frequency/high-amplitude collective motions, were visually inspected in VMD. Other structural analyses (e.g., Fig 6a) entailed com-

puting the principal axes of the moment of inertia tensor and the best-fit plane to 3D structures (in the sense of linear least-squares); the latter task utilized a previously-described singular value decomposition code (Mura *et al.*, 2010), and all other structural analysis tasks employed in-house code written in Python or as Unix shell scripts. Nucleic acid stereochemical parameters and conformational properties, e.g. values of glycosidic torsion angles and sugar pucker phase angles of the U_6 RNA, were analysed and calculated with the program DSSR (Lu *et al.*, 2015). Surface area properties, such as solvent-accessible surface areas (SASA) and buried surface areas (BSA, or ΔSASA), were calculated as averages from five approaches: *(i)* Shrake & Rupley's 'surface-dot' counting method (Shrake & Rupley, 1973), as implemented in AREAIMOL; *(ii)* the classic Lee & Richards 'rolling-ball' method (Lee & Richards, 1971), available in NACCESS; *(iii)* the 'reduced surface' analytical approach of MSMS (Sanner *et al.*, 1996); and more approximate (point-counting) methods from the structural analysis routines available in *(iv)* PYMOL and *(v)* PYCOGENT (Cieślik *et al.*, 2011).

All molecular graphics illustrations in Figs 5–8 and Supp Figs S3–S6 were created in PyMOL, with the exception of Fig S4e, f (created in VMD, rendered with Tachyon). LIGPLOT+ (Laskowski & Swindells, 2011) was used in creating schematic diagrams of interatomic contacts, as in Fig 8. Many of our scientific software tools were used as SBGRID–supported applications (Morin *et al.*, 2013).

3. Results

The organism *A. aeolicus* belongs to the taxonomic order *Aquificales*, in the phylum *Aquificae*, within what may be the most phylogenetically ancient and deeply branching lineage of the *Bacteria*. Thus, this species offers a potentially informative context in which to examine the evolution of sRNA-based regulatory systems, such as those built upon Hfq. The *Aae* genome contains an open reading frame with detectable sequence similarity to characterized Hfq homologs (e.g., from *E. coli* and other proteobacteria), and an RNomics/deep-sequencing study has shown that this putative Hfq homolog, upon heterologous expression in the γ-proteobacterium *Salmonella enterica*, can immunoprecipitate host sRNAs (Sittka *et al.*, 2009). Sequence analysis confirms that this putative Hfq can be identified via database searches (Fig 1), and that this homolog exhibits enhanced residue conservation at sequence positions that correspond to the three RNA-binding sites on the surface of Hfq—*proximal*, *distal* and *lateral rim*, denoted in the consensus line in Fig 1. As the first step in our crystallographic studies, we cloned, expressed and purified recombinant *Aae* Hfq; in these initial experiments, *Aae* Hfq generally resembled hitherto characterized Hfq homologs in terms of biochemical properties (e.g., resistance to chemical and thermal denaturation, hexamer formation).

3.1. Cloning, expression, purification and initial biochemical examination of *Aae* Hfq

Recombinant, wild-type *Aae* Hfq was successfully cloned, over-expressed and purified from *E. coli*, as confirmed by various biochemical and biophysical data, including SDS-PAGE gels (Supp Fig S1)

and MALDI-TOF mass spectra of the native protein (Fig 2a). The His6×-tagged *Aae* Hfq is 100 amino acids (AA) long, with a molecular weight of 11,365.0 Da and a predicted isoelectric point of 9.69; the working *Aae* Hfq construct, obtained via proteolytic removal of the tag (Supp Fig S1a), is 83 AA (9,482.9 Da, pI = 9.45). The expected mass, computed from the AA sequence, is in close agreement with that experimentally characterized by MALDI–TOF, indicating successful (complete) removal of the affinity tag (Fig 2a) at position G^{-2} (residue numbering is such that the wild-type methionine is M^{1}, as indicated in Supp Fig S1a).

Initial *Aae* Hfq purification efforts were hindered by nucleic acid contaminants. Specifically, purified protein samples exhibited A_{260}/A_{280} absorbance ratios of ≈ 1.65, indicative of co-purifying nucleic acids (De Mey *et al.*, 2006, Patterson & Mura, 2013); this problem is perhaps unsurprising, given the known affinity of Hfq for nucleic acids, combined with *Aae* Hfq's particularly high pI. By applying systematic colorimetric assays (Patterson & Mura, 2013) to *Aae* Hfq samples with high A_{260}/A_{280} ratios (Supp Fig S2a), we found that the co-purifying nucleic acids likely comprise a heterogeneous pool of RNAs, with lengths between ≈ 100-200 nucleotides (Supp Fig S2b). Early experiments using anion-exchange chromatography revealed that nucleic acid–bound Hfq would elute at three distinct ionic strengths (in a linear salt gradient), and each peak appeared to contain a population of nucleic acids that varied in length, both within one peak and between the three peaks (data not shown). To obtain well-defined, well-behaved *apo Aae* Hfq samples—for downstream RNA-binding assays, crystallization trials, etc.—relatively high concentrations (≈ 6 M) of guanidinium were added to cell lysates, the aim being to dissociate spurious Hfq-associated nucleic acids. Inclusion of Gnd in the purification workflow (see Methods) yielded samples with improved A_{260}/A_{280} ratios (≈ 0.8), suggesting that nucleic acid contamination had been at least partly alleviated (pure protein has A_{260}/A_{280} ≈ 0.7, and *E. coli* Hfq samples with an A_{250}/A_{274} ≈ 0.8 have been reported to have trace nucleic acid contamination (Updegrove *et al.*, 2010)). Notably, the Gnd denaturant did not appear to unfold or disrupt *Aae* Hfq's oligomerization properties, based on various observations; for instance, a discrete band corresponding to the hexameric assembly persisted on SDS-PAGE gels of Gnd-treated samples (Supp Fig S1b).

As an initial assessment of its self-assembly properties and oligomeric states in solution, purified *Aae* Hfq was examined by analytical size-exclusion chromatography (Fig 3a,b, black traces). The protein elutes as a single, well-shaped peak, with no apparent splitting, broadening, shouldering, tailing, etc. However, the location of this peak is unexpected: the peak's elution volume gives a molecular weight (MW) of ≈ 37 kDa, rather than the ≈ 57 kDa expected for an *Aae* Hfq hexamer. This apparent MW, obtained using a standard curve as described in the Methods section, could indicate a tetrameric assembly, for which the MW is calculated to be 37.9 kDa. Shape-dependent deviations from ideal migration properties would be expected to give an $(Hfq)_6$ species that migrates *faster*, not slower, than anticipated based purely on MW, given the larger effective hydrodynamic radius of a toroidal hex-

amer (versus the roughly globular standards used to calibrate our column elution volumes). However, favourable protein···resin interactions would tend to retard the migration of an *Aae* Hfq oligomer, leading to a smaller apparent MW species. Given the highly basic pI, and resultant charge on *Aae* Hfq at near-neutral pHs, we suspect that the low MW estimate from AnSEC stems from protein···resin interactions, electrostatic or otherwise; spurious *Aae* Hfq retention was also seen in experiments with other, unrelated chromatographic resins. Note that nonspecific protein adsorption to SEC resins was first documented long ago (Belew *et al.*, 1978) and has been reviewed (Arakawa *et al.*, 2010).

The aberrant AnSEC elution behaviour prompted us to assay the *Aae* oligomeric state by alternative means. SEC coupled with multi-angle light scattering (MALS) showed that the *Aae* Hfq eluting at this peak position corresponds to a hexamer, with a weight-averaged molecular weight, M_w, of 58.75 kDa (Fig 3c). A plot of the molar mass distribution (Fig 3c, green circles) exhibits uniform values across this *Aae* Hfq peak (Fig 3c, inset), indicating that this region of the eluted sample is monodisperse. *Aae* Hfq monomers were found to be susceptible to chemical crosslinking with formaldehyde, as analysed by MALDI–TOF MS (Fig 2). The main peak in the mass spectrum of this sample (Fig 2b) corresponds to a hexamer (57,498.0 Da from MS, versus 56,897.4 Da from the sequence); a second peak, near ≈ 115 kDa, corresponds to within 1.5% of the MW of a dodecameric assembly. Some Sm and Hfq orthologues have been found to assemble into stacked double-rings and other higher-order species, based on analytical ultracentrifugation and light-scattering data (Mura, Kozhukhovsky*, et al.*, 2003, Mura, Phillips*, et al.*, 2003, Dimastrogiovanni *et al.*, 2014), electron microscopy (Arluison *et al.*, 2006, Mura, Kozhukhovsky*, et al.*, 2003), gel-shift assays and other approaches; however, an integrated experimental analysis, using multiple independent methodologies on the same Hfq system, strongly suggests that the *E. coli* $(Hfq)_6$•RNA binding stoichiometry is predominantly 1:1 (Updegrove *et al.*, 2011).

3.2. Characterization of RNA-binding by *Aae* Hfq in solution

To evaluate putative RNA interactions with *Aae* Hfq, solution-state binding interactions between *Aae* Hfq and either U_6 or A_{18} (unlabelled) RNAs were examined via analytical size-exclusion chromatography. RNAs that are U-rich (e.g., U_6) or A-rich (e.g., harbouring an $(AAN)_n$ motif) are known to bind at the proximal and distal faces, respectively, of Hfq homologs from Gram-negative species. We found that U_6 RNA binds *Aae* Hfq in solution, based on comparisons of the following elution profiles (Fig 3a): *(i)* Hfq-only (black trace, detected via absorbance at 280 nm), *(ii)* U_6-only (grey, monitored at 260 nm) and *(iii)* an Hfq+U_6 mixture (red, 260 nm). In sample (*iii*), the Hfq+U_6 mixture, note the absence of a U_6 RNA peak near 19.5 mL (Fig 3a, grey), and a concomitant peak shift to a position centred at the Hfq-only trace, indicating saturated binding of the RNA. Properties of the elution profiles for samples *(i)* and *(iii)*—specifically, no shift in the peak position and no alteration of the bilat-

eral symmetry of the peak (no tailing, shouldering, etc.)—suggest that the addition of U_6 does not alter the distribution of apparent oligomeric states of *Aae* Hfq.

In contrast to the U_6 behaviour, adding A_{18} RNA to an *Aae* Hfq sample does appear to shift the Hfq oligomeric state to a higher-order species (Fig 3b, blue trace, major peak) that coexists with the usual hexamer (blue trace, minor peak). This newly-appearing, A_{18}-induced species is hydrodynamically larger than $(Hfq)_6$, as it elutes far earlier than does Hfq in the Hfq-only sample (black trace); the higher-order entity appears to correspond to an *Aae* Hfq dodecamer. This was further verified based on the M_w determined via SEC-MALS experiments done in parallel, which agrees to within 0.5% with the ideal M_w of an $\{(Hfq)_6\}_2 \bullet A_{18}$ complex (Supp Fig S3). Also, note that the Hfq+A_{18} trace is devoid of a peak at the A_{18}-only position (i.e., no peak in the blue trace, near the ≈ 18.5 mL peak location of the grey trace), indicating that binding has saturated with respect to A_{18}.

To further quantify the interactions of Hfq with U-rich and A-rich RNAs, the binding affinities for *Aae* Hfq with 5′-FAM–labelled RNA oligoribonucleotides were determined via fluorescence polarization (FP) assays (Fig 4, Supp Fig S4). FAM-U_6 and FAM-A_{18} probes were taken as proxies for U-rich and A-rich ssRNAs, enabling us to assay the strength of *Aae* Hfq···RNA interactions with these prototypical A/U-rich RNAs (for brevity, we refer to these RNAs as simply 'U_6' and 'A_{18}' if the FAM is obvious from context). Both U_6 and A_{18} were found to bind *Aae* Hfq with similarly high affinities: using a full nonlinear (logistic function) treatment of the sigmoidal binding isotherm (Eq 1 in §2.4), the nanomolar–scale apparent dissociation constants ($K_{D,app}$) are 21.3 nM for U_6 and 17.4 nM for A_{18} (Fig 4, thin, lighter-colour traces). The sigmoidal shape of these binding curves indicates positive cooperativity, and Hill coefficients were calculated to be 1.3 and 2.2 for U_6 and A_{18}, respectively. The inclusion of 10 mM Mg^{2+} in the binding reaction enhanced the U_6–binding affinity by an order of magnitude, yielding a $K_{D,app}$ of 2.1 nM (Fig 4; red, thicker trace) with a Hill coefficient of 1.7; the A_{18}–binding affinity also increased in the presence of Mg^{2+}, though by only two-fold, to a $K_{D,app}$ of 9.5 nM (blue, thicker trace) with a Hill coefficient of 2.4.

Because the apparent K_D values for U_6 and A_{18} binding were found to be in the low nanomolar range, depletion of the Hfq receptor must be accounted for near the lower Hfq concentration range sampled in our binding assays (≈ nM-range, Fig 4). Receptor-depletion phenomena can lead to spuriously high values of $K_{D,app}$ as computed from nonlinear regression against FP data, as detailed in the Methods (§2.4). Thus, to assess the impact of receptor depletion, we also performed a nonlinear least-squares fit of a three-parameter form of the classic binding isotherm (§2.4) against the FP binding data. This model (Eq 2 in §2.4) yielded the results shown in Supp Fig S4, with K_D's that were indeed ≈ 20-40% lower in magnitude than those calculated by fitting with the full sigmoidal/logistical model (i.e., Eq 1, §2.4). Note, however, that this three-parameter model assumes a Hill coefficient fixed at unity, and does not account for the aforementioned positive cooperativity that we detect in *Aae*

Hfq···RNA binding (see the discussion of the *dx* parameter in §2.4). Also, note that the $U_6^{Mg^{2+}}$ and $A_{18}^{Mg^{2+}}$ Hfq-binding reactions, which had the lowest K_D values (2.1 and 9.5 nM, respectively) of the four systems shown in Fig 4 and Supp Fig S4, were also the two systems that featured the greatest discrepancy in the $K_{D,app}$ computed via Eq 1 (includes cooperativity, neglects depletion) versus Eq 2 (neglects cooperativity, accounts for depletion); this is a reassuring finding, in terms of a depletion model for our *Aae* Hfq•RNA system, as the discrepancies that arise from receptor depletion become greater at lower K_D values. Finally, we note that no significant binding was detected between *Aae* Hfq and either FAM-A_6 or FAM-C_6 (data not shown).

3.3. Crystal structures of *Aae* Hfq monomers and oligomers, and their lattice packing

Crystals of *Aae* Hfq were readily obtained in multiple forms, including hexagonal plates and small, birefringent parallelepiped habits (Supp Fig S1c). At least three distinct morphologies could be identified, which we denote *(i)* a '*P*1 form' (*apo* Hfq, without RNA), *(ii)* a '*P*6 form' (with RNA, see §3.5 below) and *(iii)* a third form that likely belongs to space-group $P3_1$ or $P6_2$. Forms *(i)* and *(ii)* were well-diffracting (Supp Fig S1d), leading to the *P*1 and *P*6 structures reported here; the third form yielded diffraction data with potential pathologies, including translational pseudosymmetry or tetartohedral twinning, and its structure is the subject of future work (Stanek & Mura, unpublished data). Initial *Aae* Hfq crystals were obtained with a crystallization reagent comprised of 0.1 M sodium cacodylate, 5% w/v PEG 8000 and 40% v/v MPD; inclusion of $[Co(NH_3)_6]Cl_3$ additive, at ≈10 mM in the final crystallization drop, improved specimen size and quality. These *apo Aae* Hfq crystals formed in space-group *P*1, with cell dimensions a = 63.46 Å, b = 66.06 Å, c = 66.10 Å, α = 60.05°, β = 83.94°, γ = 77.17°. These dimensions are most consistent with $Z \approx$ 10–12 monomers/cell, and a resolution-dependent probabilistic estimator for the Matthews coefficient (Kantardjieff & Rupp, 2003) gives a 12-mer as the second highest peak; also, the $a \approx b \approx c$ geometry is consistent with a model wherein two Hfq hexameric rings, which generally measure ≈ 65 Å in diameter, stack atop one another in the cell.

The *P*1 *Aae* Hfq structure was refined to 1.49 Å resolution, with initial phases obtained by molecular replacement with a *Pae* Hfq hexamer search model (PDB 1U1S; (Nikulin *et al.*, 2005)). The *Pae* homolog was used because sequence analysis (Fig 1) showed it to have the greatest sequence identity (>40%) to *Aae* Hfq. A promising molecular replacement solution was readily identified, and side-chains for the *Aae* Hfq sequence were initially built in an automated manner using PHENIX. As detailed in the Methods section, the number of reflections per atom, as well as other diffraction data quality statistics, prompted us to refine atomic displacement parameters (ADPs) via treatment of the full, anisotropic *B*-factor tensor for essentially all non-hydrogen atoms (most of the isotropically-treated exceptions were atoms of solvent molecules or small-molecule components of the crystallization buffer). Anisotropic treatment of individual ADPs began at a relatively late stage in the overall

refinement workflow, and doing so noticeably improved the R_{work}/R_{free} residuals, from 13.6%/17.2% to 12.8%/15.6% before and after anisotropic treatment, respectively (Table 2). The final, refined *P*1 model was subjected to extensive validation and quality assessment, in terms of both the 3D structure itself (i.e., atomic coordinates) as well as the patterns of *B*-factors (i.e., anisotropic ADPs), as described in the Methods section.

In addition to >400 solvent (H_2O) molecules, the final *P*1 model also includes four PEG fragments, eight Gnd molecules, seven Cl^- ions, and 25 MPD molecules (Table 2). Six each of the Gnd cations and chloride anions bind between the two Hfq rings, in identical positions with respect to the nearest protein subunit (i.e., in a 6-fold–symmetric arrangement; Fig 5); the other Gnd and Cl^- species occur at unremarkable locations. The PEG fragments bind in a concave region on the exposed face of the *DE* ring—i.e., on *Aae* Hfq's distal surface (not shown in Fig 5, for clarity). Notably, this moderately apolar pocket corresponds to the second *A site* in the $(A–A–N)_n$ recognition motif described above (§1). The cleft is formed between adjacent subunits—at the interfaces of chains I/J, J/K, K/L and L/G—and is well-defined in *Aae* Hfq, with one of its walls formed by the phenolic ring of Y23 (homologous to *E. coli* Y25, crucial in A-rich RNA-binding). The PEG fragments bind with similar poses in each of the four sites. Of the 25 MPD molecules, 24 occupy 6-fold–symmetric positions near the proximal face of *Aae* Hfq (the remaining MPD is near the distal face of the *DE* ring). These 24 MPDs bind in a 2×{6+6′} arrangement. Here, the '2' denotes that a set of 12 MPDs binds identically to each of the two Hfq hexamers (i.e., *PE* and *DE* rings in Fig 5), and the prime in '6+6′' indicates two distinct subsets of MPDs: one binds at Hfq's proximal RNA site (below, and Fig 7), while the other MPD is disposed near the α-helix on the proximal site, not far from the lateral rim.

The overall 3D structure of the *Aae* Hfq monomer (Fig 5) is that of the Sm fold, as anticipated based on sequence similarity and the efficacy of MR in phasing the diffraction data. In particular, an *N*-terminal α-helix is followed by five highly-curved β-strands arranged as an antiparallel β-sheet. The secondary structural elements (SSEs), shown schematically in Fig 1, are labelled in the 3D structure of Fig 6b. Precise SSE boundaries in *Aae* Hfq, computed with STRIDE, are residues # 5–16 (α1), 19–24 (β1), 29–38 (β2), 41–46 (β3), 49–54 (β4) and 58–63 (β5); the same ranges are obtained with DSSP, save that DSSP's criteria make F37 (not D38) the end of the most curved strand (β2). Most of *Aae* Hfq's β-strands are delimited by loops that adopt various β-turn geometries (including types I, II′, IV, VIII), with the exception of a short 3_{10} helix (residues 55–57) between β4→β5. These loops contain many of the RNA-contacting residues of Hfq (see below) and, as labelled in Figs 1, 5, 6 and 9, we denote these linker regions as L1→L5. Noncovalent interactions between Hfq monomers include van der Waals contacts and hydrogen bonds between the backbones of strand β4 of one subunit and β5* of the adjacent subunit, effectively extending the β-sheet across the entire toroid; these enthalpically favourable interatomic contacts likely facilitate self-assembly of the hexamer. (Unless otherwise stated, asterisks denote an adjacent Hfq subunit, be it related by crystallographic symmetry or otherwise.)

Residues 1→68 of the native *Aae* Hfq sequence could be readily built into electron density maps for each monomer in the ASU, thus providing a structure of Hfq's *N*-terminal region as well as the entire Sm domain; note that the *N*-terminal tail, illustrated for the *apo*/*P*1 structure in Fig 5 (bottom-right) and Fig 6b, is unresolved in many previous Hfq structures. Most of the *Aae* Hfq *C*-terminal residues 70→80 were not discernible in electron density, and are presumably disordered.

3.4. The *apo* form of *Aae* Hfq

While neither NCS averaging, nor any NCS constraints or restraints, were applied at any point in the phasing and refinement of *Aae* Hfq in the *apo* form, the 12 monomers in the *P*1 cell are virtually indistinguishable from one another (Fig 6a,b, Supp Fig S5), at least at the level of protein backbone structure (there are side-chain variations). The mean pairwise main-chain RMSD, for all monomer pairs in the *P*1 cell, lies below 0.3 Å; this low value is also evident in the magnitude of the ordinate scale of the structural clustering dendrogram in Supp Fig S5c. To systematically compare structures, a matrix of RMSDs was constructed from all pairwise subunit alignments. Agglomerative hierarchical clustering on this distance matrix (Supp Fig S5c) reveals that the subunits partition into two low-level (root-level) clusters so as to recapitulate the natural (structural) ordering found in the crystal: that is, chains A→F cluster together (as the *proximal-exposed*, or *PE*, ring in Fig 5), and likewise chains G→L form a second group (the *distal-exposed*, or *DE*, ring). This finding is illustrated in Fig 6c, which conveys the degree of 3D structural similarity as a circular graph wherein the width of an edge between two chains is inversely scaled by their RMSD.

At the *Aae* Hfq monomer level, the greatest structural variation occurs among the *N*-termini and the L4 loop region between β3→β4; apart from the termini, loop L4 (Fig 6b) is the most variable region in most known protein structures from the Sm superfamily. The conformational heterogeneity in the termini and loops of *Aae* Hfq stems, at least partly, from differing patterns of interatomic contacts for different subunits, at the levels of monomers, hexamers and dodecamers in the overall *P*1 lattice. The patterns of conformational heterogeneity are clear when the dodecameric structure is visualized as a cartoon, with the diameter of the backbone tube scaled by the magnitude of per-atom B_{eq} values (this derived quantity, computed from the trace of the full anisotropic ADP tensor, is taken as an estimate of the true B_{iso} values that would result from refinement of an isotropic model); such renditions are shown in Supp Figs S6a and S6b for the *P*1 and *P*6 structures, respectively. Analogously, Supp Figs S6c and S6d provide thermal ellipsoid representations of the patterns of variation in anisotropic ADPs across the *P*1 dodecamer and *P*6 monomer. In both sets of depictions, Figs S6a/b and S6c/d, colours are graded by the magnitude of per-atom B_{eq} values, from low (blue) to medium (white) to high (red). To initially assess the relative contributions of static disorder (e.g., variation in rotameric states across subunits) and dynamic disorder (e.g., harmonic breathing modes and other collective/global motions) in variable regions such as loop L4 and the termini, a normal mode analysis was

performed on a coarse-grained representation of the *Aae* Hfq structures, using an anisotropic network model of residue interactions (see Methods). Illustrative results for the dodecamer and monomer are shown in Supp Figs S6e and S6f, respectively. The pattern of normal mode displacements for both the dodecamer and monomer do not implicate loop L4 in any especially high-amplitude, low-frequency modes (Supp Fig S6f), suggesting that L4's increased ADPs (elevated B_{eq} values) stem more from static disorder rather than any particular dynamical process involving this loop region (though anharmonic dynamics remains possible). The dodecamer calculation does reveal a significant harmonic mode corresponding to anti-symmetric rotation of the two Hfq rings with respect to one another (PE ↺, DE ↻; Supp Figs S6e). This result is consistent with our observation that the only large-scale (dodecamer-scale) structural difference between the two rings is a slight rotation of one relative to the other (Fig 5, left)—versus, for instance, a rigid-body tilt (Fig 6a, Supp Figs S5a,b).

At the Hfq ring and supra-ring levels, the refined *P*1 structure reveals an *Aae* Hfq dodecamer consisting of two hexameric rings stacked in a head-to-tail orientation (Fig 5). Propagated across the lattice, this arrangement gives cylindrical tubes with a defined polarity. The tubes run along the crystallographic $\vec{a}$ axis, and their lateral packing yields near–six-fold symmetry along this direction; a slight translational shift of the dodecamers in adjacent unit cells, in the plane perpendicular to $\vec{a}$, causes the rings to be slightly offset with respect to the lattice tubes (the tubes are not perfectly cylindrical, insofar as the 6-fold axis of an individual Hfq ring is not coaxial with the principal axis of its parent tube). In the dodecamer, the distal face of one Hfq ring is exposed (termed the *DE ring*), while the other ring features a proximal-exposed face (the *PE ring*, Fig 5, right). The *N*-termini of the *DE* hexamer contact the L2-loop/β2-strand region of the *PE* ring, as illustrated in Fig 5 (the L2 loops mark the beginning of strand β2; see the label in Fig 6a). As is apparent in the axial view of Fig 5 (left), one ring is slightly rotated relative to the other. Geometric analysis of this rotation (denoted 'Δ' in Fig 6a), as well as other rigid-body transformations relating the two rings (Supp Fig S5a, b), shows that the 6-fold symmetry axes of the rings in the dodecamer are not perfectly parallel—a slight tilt occurs between the rings ('δ' in Fig 6a). This tilt appears to stem largely from structural differences in the *N*-terminal regions (Supp Fig S5). Consistent with these observations, the set of six *N*-terminal regions of the *DE* ring (which mediate ring···ring interactions within a dodecamer) exhibit slightly higher B_{eq} values and greater conformational variability than do the six *N*-termini of the *PE* ring (which mediate dodecamer···dodecamer contacts between unit cells), as can be seen in Supp Fig S6a.

Noncovalent molecular interactions between the proximal···distal faces mediate the association of Hfq rings into a dodecamer, and a slightly altered (translationally shifted) version of these same energetically favourable interactions stiches together the dodecamers into a set of crystal lattice contacts in the *P*1 form of *Aae* Hfq. Notably, a *proximal→distal* stacking geometry is the chief mode of ring association in the *Aae P*6 lattice too. *Aae* Hfq dodecamers clearly occur in the *P*1 lattice, with a substantial amount of buried surface area (BSA) defining the ring···ring interface (Fig 5). Specifically,

3663 ± 244 Å^2 of SASA is occluded between the *PE* and *DE* hexamers in the *PE*•*DE* complex. Note that this quantity is being reported as a total $BSA = ASA_{PE} + ASA_{DE} - ASA_{PE\bullet DE}$, where ASA_i is the ASA of species *i*, rather than as the per-subunit value (which would be given by half of the above expression, were we to assume a perfectly 2-fold symmetric interface); also, note that this *mean* ± *standard deviation* is reported from the results of five different surface area calculation approaches, as described in the Methods.

3.5. Crystal structure of *Aae* Hfq bound to U_6 RNA

Upon co-crystallization with U_6 RNA, a second, distinct *Aae* Hfq crystal form was discovered. These crystals could be indexed in *P*6, with unit cell dimensions of $a = b = 66.19$ Å, $c = 34.21$ Å. In this form, the cell geometry, solvent content and molecular mass of *Aae* Hfq are only compatible with a single Hfq monomer/ASU; based on known Hfq structures, the crystallographic 6-fold was presumed to generate intact hexamers, such as shown in Fig 7a. Specifically, co-crystallization of *Aae* Hfq with this model uridine-rich RNA was achieved by incubating purified Hfq samples with 500 µM U_6 RNA prior to crystallization trials. The complex crystallized in 0.1 M sodium cacodylate, 5% w/v PEG 8000 and 40% v/v MPD, and the denaturant compound Gnd was found to be an effective additive (Supp Table S2). The crystal structure of the *Aae* Hfq•U_6 RNA complex was refined to 1.50 Å resolution (Fig 7); we emphasize that the initial solution of this structure was achieved independently of the *apo P*1 form, via molecular replacement, using *P. aeruginosa* Hfq as a search model.

Those residues that are crucial in forming the proximal (U-rich) RNA–binding pocket in *E. coli* and other Hfq orthologues—i.e., *Eco* residues Q8, F42, K56, H57—are conserved in the *Aae* Hfq sequence (Fig 1). This observation led us to anticipate that any bound U_6 would be localized to the proximal pore region. Instead, a molecule of MPD, which served as a precipitant and cryo-protectant in our crystallization experiments (Table S2), was found to occupy the proximal site of the Hfq hexamer, with the MPD hydroxyl groups hydrogen-bonded to the side-chains of *Aae*'s H56 and *Q6 residues (Fig 7c). In addition, the bound MPD makes van der Waals contact with other conserved residues that line the proximal site, specifically *L39 and F40. During refinement of this structure, two nucleotides of the U_6 RNA molecule, including the flanking 5′ and 3′ phosphates (the latter coming from the third U), were readily discernible in $mF_o - DF_c$ difference electron density maps (Fig S7). Notably, processing and reduction of the diffraction data (collected from *P*6-form crystals) in *P*1 yielded similar electron density for the RNA at each lateral binding pocket in the hexamer (Fig S7). Rather than being bound at the proximal site, the uridine residues of U_6 occupied a cleft formed between the *N*-terminal α-helix and strand β2, in a position located roughly near the outer ('lateral') rim of the *Aae* Hfq toroid (Fig 7a,b).

3.6. RNA binding at the outer rim of the *Aae* Hfq hexamer: Structural details

The *Aae* Hfq•U_6 structure reveals a lateral RNA-binding pocket that accommodates two nucleotides of uridine. The *N*-terminal α-helix primarily contacts the phosphodiester and ribose groups, and the β2 strand interacts mostly with the uracil bases (Figs 7a, 7b, 8a). As a consequence of this RNA-binding geometry, both nucleotides that were fully built into electron density (U1, U2) are held in a bridging, *anti*-conformation ($\chi = -165.2°$ for U1, $-116.8°$ for U2), with the ribose moieties extending outward from the pocket (Fig 7b). Interestingly, while the U1 ribose is in the 3′-*endo* conformation typically seen in canonical (*A*-form) RNA structures, with a pseudo-rotation phase angle (*P*) of 17.5º for this North sugar pucker, the U2 ribose adopts a less typical 2′-*endo* conformation (*P* = 163.2º).

Protein···RNA interactions are mediated by both side-chain and backbone atoms of *Aae* Hfq. The full set of interactions is shown in 3D in Fig 7a, b, and schematically in Fig 8a. Two side-chains in *Aae*'s *N*-terminal α-helix, N11 and R14, contact the phosphodiester groups (denoted 'Ⓟ' for brevity), and another cationic residue (K15) is 3.6 Å from the Ⓟ linking the two uridines. Backbone and side-chains atoms from strand β2 hydrogen-bond with the bases, ensuring uridine specificity (Figs 7b, 8, 9). In particular, both the carbonyl oxygen and amide nitrogen of F37 interact with N3 and O4 of U2, respectively, while the hydroxyl side-chain of S36 contacts the exocyclic O4 of the U1 nucleobase. S36 also helps position a pivotal H_2O that directly hydrogen-bonds to both the N3 atom of U1 and the S36 hydroxyl (Fig 8a); this well-ordered (ice-like) water molecule engages in a network of hydrogen-bonds, in a distorted tetrahedral geometry (additional structural waters also contact the uracil and Ⓟ moieties, as shown in Fig 8). Other interactions at the lateral site include a series of three π-stacking interactions (Fig 8a): between the phenyl ring of F37···U2, the U1···U2 bases, and the phenolic ring of *Y3···U1. RNA-binding at the lateral site is composite in nature, involving not just residues of strand β2 and helix α1 of one Hfq subunit, but also the *N*-terminal tail of an adjacent subunit in the ring. The irregularly structured *N*-terminal tail of one Hfq monomer extends into the neighbouring lateral site, where the *N*-terminal sequence $H^0M^1P^2Y^3K^4$ nearly 'covers' that rim site and supplies additional contacts with RNA. For instance, *Y3 engages in the π-stacking mentioned above, as well as a hydrogen-bond between its amide nitrogen and the O2 of U1 (an interaction that does not select between uracil and cytidine). Also in this region, the backbone carbonyl oxygen of *M1 hydrogen-bonds to the ribose O2′ of U1, thus contributing to discrimination between RNA and DNA. Finally, we note that two contacts in this region may be spurious: *(i)* the *H0···Ⓟ interaction, where residue *H0 is from the recombinant construct (not wild-type *Aae* Hfq; see numbering in Supp Fig S1); and *(ii)* the R29′···Ⓟ interaction, which is a crystal lattice contact (the prime symbol on R29′ indicates an adjacent unit cell).

Comparison of the *Aae* Hfq•U_6 structure with the independently-refined *apo Aae* Hfq structure suggests that the lateral RNA-binding site is essentially pre-structured for RNA complexation (Fig 9). In terms of comparative structural analysis, note that the *apo*/*P*1 and RNA-bound/*P*6 structures *(i)* are at equally high resolutions (1.49, 1.50 Å respectively; Table 1), *(ii)* were refined in similar man-

ners (e.g., using anisotropic ADPs), albeit independently of one another, and *(iii)* are of comparable quality in terms of R_{work}/R_{free}, stereochemical descriptors, etc. (Table 2). Residues N11, R14, S36 and F37—which are phylogenetically conserved to varying degrees (Fig 1)—largely define the structural and chemical topography of the lateral site (Fig 7a). As shown in Fig 9, these crucial residues adopt nearly identical rotameric states in the *apo* and U_6–bound forms of *Aae* Hfq. The two principal RNA-related structural differences, in going from the *apo* to the U_6–bound forms, are: *(i)* a shift in the residue E7 rotamer (Fig 9, red label), positioning this side-chain away from the pocket and thus enabling the U2 base to be accommodated, and *(ii)* the precise path of the *N*-terminal tail (i.e., the ≈5 residues preceding helix α1), which varies with respect to the lateral site. In the dodecameric *apo* structure, six of the *N*-termini mediate *ring···ring* contacts (Fig 5, *DE* ring) while the other half (from the *PE* ring) mediate lattice contacts, giving rise to one source of structural heterogeneity in this region. In terms of intrinsic conformational flexibility, normal mode calculations (Supp Fig S6 and Methods) indicate that the *N*-terminal regions in the hexamer are highly flexible when free in solution, but rigidified (as much as any other part of the Sm fold) when sandwiched between the Hfq rings.

4. Discussion

The *apo* form of *Aae* Hfq crystals, refined to 1.49 Å in space-group *P*1, reveals a dodecamer comprised of two hexamers in a head-to-tail orientation. The individual subunits of *Aae* Hfq are similar in structure, with a mean pairwise RMSD less than ≈ 0.3 Å for all monomer backbone atoms. The largest differences among the 13 independently-refined Hfq monomer structures (12 in *P*1, one in *P*6) occur in the *N*-terminal and L4 loop regions; notably, these are the two regions that mediate much of the interface between rings (*distal···proximal* face contacts in Fig 5), as well as the intermolecular contacts between dodecamers across the lattice. The patterns of structural differences are also captured in the symmetric matrix of pairwise RMSDs between chains: hierarchical clustering on this distance matrix results in the monomers that comprise the *PE* (chains A–F) and *DE* (chains G–L) hexameric rings partitioning into two distinct groups (Fig 6c, Supp Fig S5c).

Sm proteins, including Hfq, exhibit a strong propensity to self-assemble into cyclic and higher-order oligomers. These assemblies often crystallize as either *(i)* cylindrical tubes with a defined polarity, via head→tail association of rings (*Aae* Hfq and *Mth* SmAP1 are two examples), or *(ii)* head↔head stacks of cyclic oligomers, often with dihedral point-group symmetry (*Pae* SmAP1 is an example (Mura, Kozhukhovsky, *et al.*, 2003)). An examination of the lattice packing of all known Hfq structures (data not shown) reveals at least one example of each possible ring-stacking mode for a dodecameric assembly: *(i)* a *proximal•proximal* interface, as seen in the extensive interface between hexamers of an Hfq orthologue from the cyanobacterium *Synechocystis* sp. PCC6803 (PDB ID 3HFO; (Bøggild *et al.*, 2009)); *(ii)* a *distal•distal* interface, observed in *S. aureus* Hfq (PDB ID 1KQ2; (Schumacher *et al.*, 2002)) and in *P. aeruginosa* Hfq, with a more modest interface and relative trans-

lational shift of one ring (PDB ID 4MMK; (Murina *et al.*, 2014)); and *(iii)* the *head→tail* packing of two rings in a *L. monocytogenes* (*Lmo*) Hfq structure, in *apo* and RNA-bound forms (PDB ID 4NL2; (Kovach *et al.*, 2014)). The *Aae head→tail* interface (Fig 5) buries more ASA than that between the *Lmo* Hfq rings, but otherwise the stacking in these two Hfq structures resemble one another even in fine geometric details (e.g., the top/bottom, *PE*/*DE*, rings are similarly rotated with respect to one another). Also, the *S. aureus distal•distal* dodecamer buries 2666 $Å^2$ of surface area, which is considerably less than the ≈ 3700 $Å^2$ of ΔSASA determined here for *Aae* Hfq's *distal•proximal* stacking mode.

As a point of reference, note that the above ΔSASA quantities represent less buried surface area than in the *ring•ring* interfaces found in the structures of various Sm and SmAP homologs. (Recall that Hfq rings are hexameric while SmAPs are generally heptameric, meaning that a systematic difference in ΔSASA trends will occur simply by virtue of subunit stoichiometry.) The *ring•ring* interfaces in *P. aerophilum* and *M. thermautotrophicum* 14-mers occlude 7550 and 3000 $Å^2$, respectively. Unlike *P. aerophilum* SmAP3, where the burial of >21,000 $Å^2$ along an intricate interface between stacked rings suggests *bona fide* higher-order oligomers (Mura, Phillips*, et al.*, 2003), the extent of the *Aae* Hfq *distal•proximal* interface does not as clearly indicate whether or not dodecamers exist. The free energy of association between *Aae* Hfq's *PE•DE* rings, ΔG°_{bind}, can be estimated via the linear relationship $\Delta G^{\circ}_{bind} = \gamma \cdot BSA$ (the slope, γ, is often taken as ≈ 20-30 $cal \cdot mol^{-1} \cdot Å^{-2}$ (Janin *et al.*, 2008)); however, *Aae* Hfq's *PE•DE* interface is not primarily apolar in character, so this approach may severely overestimate the ΔG°_{bind}. Also, in terms of the existence and potential relevance of double-rings and higher-order species, recall that *Aae* Hfq can form dodecamers *in vitro*, at least when bound to an A–rich RNA and assayed by AnSEC (Fig 3b, blue arrow). Nevertheless, despite all these observations, *(i)* whether or not Hfq dodecamers actually occur *in vivo*, beyond crystalline and *in vitro* milieus (such as in AnSEC experiments) remains unclear, and *(ii)* even if such dodecamers do exist, the potential physiological activities and functional roles of higher-order oligomeric states of Hfq remains murky.

Intriguingly, our solution-state AnSEC data are consistent with the binding of A_{18}, presumably at the distal face of $(Hfq)_6$, causing a shift in the distribution of *Aae* Hfq oligomeric states from hexamers (only) to a more dodecameric population (Fig 3). This effect may be attributed to the longer A_{18} strand simultaneously binding to two Hfq rings, giving a 'bridged' ternary complex. There also appears to be some length-dependence to the interaction of A-rich RNAs with Hfq, as we found that A_6 did not exhibit high-affinity binding to *Aae* Hfq; this dependence may stem from mechanistic differences in the early (initiation) stages of the kinetic mechanism for Hfq···RNA binding. *Aae* Hfq demonstrates a nanomolar affinity for A_{18} and U_6 RNA that is selective (C_6 does not bind) and that is consistent with the properties of Hfq homologs characterized from other bacteria, both Gram-negative (e.g., proteobacteria such as *E. coli*) and Gram-positive. For instance, the magnesium–dependence of

the *Aae* Hfq•U_6 interaction (Fig 4), with 10-fold stronger binding in the presence of Mg^{2+}, mirrors the Mg^{2+}-dependency of U-rich–binding by Hfq homologs from the pathogenic, Gram-positive bacterium *Listeria monocytogenes* (*Lmo*) and the Gram-negative *E. coli* (*Eco*; (Kovach *et al.*, 2014)). For both *Lmo* and *Eco* Hfq, the inclusion of 10 mM magnesium increased the U_6-binding affinity by >100-fold; the effect was similar, but less pronounced, for U_{16} (≈ 3-4–fold increase). Thus, the Mg^{2+}-dependency of the *Aae* Hfq•U_6 RNA interaction is intermediate between these two extremes.

At present, only two other known Hfq structures contain a nucleic acid bound to the lateral site. These structures are: *(i) Pae* Hfq co-crystallized with the nucleotide uridine–5′–triphosphate (UTP; PDB ID 4JTX; (Murina *et al.*, 2013)) and *(ii) Eco* Hfq bound to a full-length sRNA known as RydC (PDB ID 4V2S; (Dimastrogiovanni *et al.*, 2014)). Comparison of the lateral RNA-binding sites of the *Aae*, *Pae* and *Eco* Hfq structures reveals a highly conserved pocket formed by N13, R16, R17, S38 and F39 (*Eco* numbering; see also Fig 1). In *Aae* Hfq, K15 appears to be homologous to *Eco* R16, insofar as this side-chain is well-positioned to engage in electrostatic and hydrogen-bond interactions with the sugar-phosphate backbone of a bound RNA (Figs 7b, 8, 9). This structural feature can be seen both in *Eco* Hfq (R17 with the phosphate of a neighbouring nucleotide) and in *Pae* Hfq (K17 with the 5′ phosphate tail of UTP). Notably, uridine is the only nucleotide that has been found to bind at the lateral site in all three of these Hfq structures—*Eco*, *Pae*, and now *Aae*.

At a resolution of 1.5 Å, the *Aae* Hfq•U_6 structure offers new insights into the apparent specificity of the lateral pocket for uridine nucleosides. We see that interactions with the backbone of strand β2 provide discrimination between uracil and cytosine bases in the cognate RNA. One uracil base π-stacks with a key phenylalanine residue, while the second uracil stacks atop the preceding nucleobase. The second nucleotide adopts a C2′-*endo* conformation, leading to accommodation of the base in this binding cleft on Hfq's surface. In this configuration, the *N*-terminal region may then provide further enthalpically favorable interactions that stabilize the complex. The *Aae* Hfq lateral site includes two of the three arginine residues of the 'arginine patch', known to be important for annealing of sRNAs and mRNAs (Panja *et al.*, 2013). We propose that the third arginine of this motif acts primarily electrostatically—without directionality, and non-specifically as regards RNA sequence—in order to enhance the diffusional association of an RNA by 'guiding' it towards the lateral pocket. In addition, the physicochemical basis for the phylogenetic conservation of the lateral site may be that it simply provides additional surface area for Hfq···sRNA interactions, perhaps supplying an extended platform for the 'cycling' of RNAs across the surface of the Hfq ring (Wagner, 2013); similarly, the rim site may serve as an additional 'anchor' site for the association of moderate-length, U-rich RNAs that bind with low intrinsic affinity for the proximal site, but which can reach the lateral/rim site. We propose that the lateral site, which is structurally well-defined on the outer rim of the *Aae* Hfq hexamer, is a biologically relevant region that functions in binding $(U)_n$ segments of RNA containing at least two

consecutive uridine nucleotides; moreover, we propose that this RNA-binding region is conserved in even the most ancient bacterial lineages.

The structural features of Hfq···RNA interactions in homologs from evolutionarily ancient bacteria share some similarity with the properties of Sm-like archaeal proteins (SmAPs), such as a SmAP from the hyperthermophile *Pyrococcus abyssi* (*Pab*) that was co-crystallized with U_7 RNA (Thore *et al.*, 2003). Interestingly, the oligoribonucleotide in that crystal structure was found in two sites: the canonical U-rich–binding site near the lumen of the ring (analogous to Hfq's proximal site), as well as a 'secondary' pocket on the same (proximal) face. This secondary site of *Pab* SmAP is distant from the U–binding site, lying between the *N*-terminal α-helix and strand β2 of the Sm fold. Note that the 'lateral site' of Hfq had not yet been discovered as an RNA interaction region at the time of the *Pab* structure determination. The secondary RNA-binding site in *Pab* SmAP also contains a phenylalanine residue that is conserved among Hfq homologs and that is required for π-stacking with the nucleobase. However, the asparagine residue found at the lateral site of all characterized Hfq homologs is instead a histidine in *Pab* SmAP; this residue's imidazole side-chain provides an additional stacking platform for an adjacent ribonucleotide in the *Pab* complex, in an interaction that is unseen with known Hfq homologs. The α-helix of *Pab* SmAP does not extend as far as that of Hfq, and the arginine-rich patch that occurs at this rim area in Hfq homologs is but a single lysine residue in *Pab* SmAP. Nevertheless, the presence of this partially conserved lateral pocket in *Pab* SmAP does suggest an ancient, common origin for this mode of protein···RNA recognition by Hfq and other members of the Sm superfamily. Somewhat similarly, a uridine-binding site was crystallographically identified in *Pyrobaculum aerophilum* SmAP1, in a region on the 'L3 face' (analogous to Hfq's proximal face) that lies distal to the canonical U–rich RNA-binding site at the inner surface of the pore; this L3-face region was described as a 'secondary' binding site because of relatively weak electron density for the phosphoribose (Mura, Kozhukhovsky*, et al.*, 2003). We can now see that the secondary U-rich–binding sites in at least two archaeal Sm proteins, from *Pab* and *P. aerophilum*, occupy a region that is roughly analogous to the lateral rim of Hfq.

The historical lack of structural data on RNA-binding at the Hfq lateral site may be because uridine-rich RNAs—such as might localize to the lateral rim—are also capable of binding to the higher-affinity proximal site. A single binding event is consistent with the idealized shape of our *Aae* Hfq•U_6 binding curves (Fig 4), which bear no hint of multiple transitions or non–two-state binding. This could indicate that U_6 binding at the proximal and distal sites differs by at least an order of magnitude (beyond the detection range of our assay). In terms of the structure of the *Aae* Hfq•U_6 complex reported here, we suspect that two facets of our crystallization efforts serendipitously shifted the RNA–binding propensity towards the lateral site. First, MPD was present at high concentrations in our crystallization condition (many Hfq homologs reported in the literature were crystallized with PEGs, not MPD). MPD is a commonly-used precipitating agent and cryoprotectant, and inspection of electron

density maps reveals it to be associated, at high occupancy, with all 12 subunits of the *apo* form of *Aae* Hfq; specifically, 24 of the 25 MPDs found in the *P*1 electron density are bound in one of two locations (Fig 5), and one of these locations corresponds to what would be a proximal RNA-binding site. Moreover, an MPD molecule was also bound in the *P*6 (U_6-bound) crystal forms, in clear density at the proximal site (Fig 7); notably, this proximal-site MPD almost perfectly superimposes in 3D with the 12 MPDs at this site in the 12 subunits of the *apo*/*P*1 structure. In terms of structural and chemical properties, the hydroxyl groups of MPD closely mimic the ribose and uracil moieties of uridine, as shown in Supp Fig S8. Residues H56 and Q8 have been identified as two key residues in the proximal site that contact the ribose 2′-OH and uracil's exocyclic O2 atom upon binding of U_6 at the proximal site (Schumacher *et al.*, 2002). However, in our *Aae* Hfq structure these two residues instead contact MPD (*Q6 and H56 in Fig 7c). The lateral RNA-binding site, however, does not include many contacts to ribose (versus the phosphate and nucleobase groups), and thus MPD would not be expected to compete as strongly against RNA-binding at that site. The hypothesis that MPD interferes with RNA-binding by localizing at the proximal site (e.g., Fig 7c) is borne out by RNA-binding competition assays, which reveal that exceedingly high concentrations of MPD—such as in our crystallization conditions—can successfully inhibit *Aae* Hfq•U_6 binding (Supp Fig S9). The second unique feature of *Aae* Hfq that may increase the affinity for U-rich RNA at the lateral site is the flexible *N*-terminal tail, which folds over the lateral site when nucleic acid is bound, further stabilizing the associated U_6 RNA. In our work, the *N*-terminus includes three plasmid-derived residues that remain after cleavage of a His6× tag used in protein purification ($G^{-2}S^{-1}H^{0}$; Supp Fig S1a). The additional histidine contacts the phosphate of nucleotide U2 (Figs 7b, 8). In addition, the native sequence includes a tyrosine residue that provides further aromatic stacking interactions with base U1 (residue *Y3 in Figs 7b, 8). This tyrosine residue is not conserved among other Hfq homologs, many of which contain a glutamate at this position (Fig 1).

The crystallographic and biochemical work reported here reveals that the putative Hfq homolog encoded in the *A. aeolicus* genome is an authentic Hfq, as it *(i)* adopts the Sm fold, *(ii)* self-assembles into hexameric rings that can associate into higher-order double rings in the lattice (as do many known Hfqs), and *(iii)* binds A/U-rich RNAs with high affinity (and selectivity). Perhaps most exciting, these structural and functional properties are recapitulated by an Hfq homolog from the Aquificae phylum, which may be the most basal, deeply-branching lineage in the *Bacterial* domain of life (Bocchetta *et al.*, 2000, Burggraf *et al.*, 1992). To date, all Hfq structures have been limited to three phyla: *(i)* most Hfq structures are from the *Proteobacteria*, *(ii)* a few are from the (mostly Gram-positive) *Firmicutes* and, finally, *(iii)* two known homologs are of *Cyanobacterial* origin. Because of its basal phylogenetic position, the *Aae* Hfq structures reported here—the first Hfq structures from outside these three bacterial lineages—suggest that members of the Sm/Hfq superfamily of RNA–

associated proteins, along with at least some of their RNA–binding properties, likely existed in the last common ancestor of the *Bacteria*.

Figure Legends

Figure 1 Multiple sequence alignment of *Aae* Hfq and some representative homologues. Sequence analysis of several Hfq homologues, characterized from various phyla, reveals conservation of key amino acids comprising Hfq's three distinct RNA-binding regions (*distal*, *proximal*, *lateral*). The *Aae* Hfq sequence is numbered at the top, and secondary structural elements are drawn based on the *Aae* crystal structures reported herein; helices are schematized as spirals, strands as arrows, and numbered loop labels are shown (a short 3_{10} helix forms loop L5, coloured brown). Strictly identical amino acids are in bold blue text on a yellow background, while sites with highly similar residues are highlighted with a grey background; these blocks of partially conserved residues are also lightly boxed. In the consensus sequence shown at the bottom, uppercase letters indicate strict identity and lowercase letters correspond to physicochemically equivalent residues that meet a similarity threshold (≥85% sites in a given column). Residues known to contact RNA at the proximal, distal or lateral sites are marked with red, blue or green square symbols, respectively. Note the high level of conservation of residues involved in all three RNA-binding sites. In addition to *Aae* Hfq (of the phylum *Aquificae*), the twelve aligned sequences include *(i)* three Hfq homologs from the mostly Gram–positive Firmicutes (*Sau*, *Lmo*, *Bsu*), *(ii)* a homologue from the ancient phylum *Thermotogae* and *(iii)* several characterized Hfq orthologues from the α–, β– and γ–proteobacteria. The relationships between these species are indicated in the dendrogram (left), obtained during the progressive alignment calculation and coloured so as to highlight phylum-level differences. The genus/species and sequence accession codes [GenBank] follow: *Aae*, *A. aeolicus* [AAC06479.1]; *Sau*, *Staphylococcus aureus* [ADC37472.1]; *Tma*, *T. maritima* [AGL49448.1]; *Lmo*, *Listeria monocytogenes* [CBY70202.1]; *Bsu*, *Bacillus subtilis* [BAM57957.1]; *Rsp*, *Rhodobacter sphaeroides* [A3PJP5.1]; *Atu*, *Agrobacterium tumefaciens* [EHH08904.1]; *Nme*, *Neisseria meningitidis* [P64344.1]; *Hse*, *Herbaspirillum seropedicae* [ADJ64436.1]; *Pae*, *P. aeruginosa* [B3EWP0.1]; *Eco*, *E. coli* [BAE78173.1]; and *Vch*, *Vibrio cholerae* [A5F3L7.1].

Figure 2 *Aae* Hfq monomers and oligomers, as assayed by crosslinking and mass spectrometry. MALDI-TOF spectra are shown for (*a*) native, untreated (non-crosslinked) *Aae* Hfq monomers, with an expected MW of 9482.9 Da based on the recombinant protein sequence (Supp Fig S1), as well as (*b*) a chemically crosslinked *Aae* Hfq sample. As detailed in the Methods (§2.2), crosslinking assays employed a gentle ('indirect') method, using formaldehyde as a crosslinking agent. The main peaks in the crosslinked sample correspond to hexamers and dodecamers, with expected MWs of 56,897.4 and 113,794.8 Da, respectively. The singly-charged molecular ion peaks, $[M+H]^{1+}$, are accompanied by schematics (blue and orange balls) that indicate the anticipated architecture of the oligomeric states, alongside the peak's MW, as determined from the mass spectrum (crosslinked species are better char-

acterised by a MW range, rather than a single value, because of variability in the number of crosslinker molecules that react).

Figure 3 The solution-state distribution of *Aae* Hfq oligomers shifts in the presence of short RNAs. Elution profiles are shown for analytical size-exclusion chromatography of *Aae* Hfq samples incubated with either (*a*) U_6 or (*b*) A_{18} RNAs (specifically, 250 μM *Aae* Hfq was incubated in a 1:1 v/v ratio with 50 μM of either U_6 or A_{18} RNA). The elution of *Aae* Hfq was detected via the absorbance at 280 nm (A_{280}), and RNA and Hfq•RNA complexes were monitored at A_{260}. While putative Hfq···U_6 interactions do not appear to shift the oligomeric state, as indicated by the close alignment of the black (Hfq alone) and red (Hfq•U_6) peaks in (*a*), Hfq interactions with A_{18} do shift the oligomeric species towards a higher-order state (blue arrow in *b*, denoting apparent dodecamers). This shift could correspond to the simultaneous binding of A_{18} to two Hfq hexamers, potentially via two modes: *(i)* as an $(Hfq)_6$•A_{18}•$(Hfq)_6$ 'bridged' complex, or *(ii)* as A_{18} bound to one of the two distal faces that would be exposed on an independently-stable $(Hfq_6)_2$ double-ring dodecamer. These two models cannot be distinguished via AnSEC. *(c)* To verify the molecular weight of the *Aae* Hfq elution peak, the protein was analysed via SEC fractionation followed by multi-angle static light scattering and refractive index measurements. The SEC elution profile (black trace) is taken as the absorbance at 280 nm. Light-scattering and refractive index data can be used to compute molar masses, and the open circles shown here (semi-transparent green) are the molar mass distribution data (i.e., masses [in kDa] as a function of elution volume). The weight-averaged molecular weight, M_w, of the Hfq sample is computed for the entire peak from this distribution, and the scale is given by the vertical axis on the right-hand side (green numbers; note that this scale applies to the main plot, not the inset). The apparent M_w that was computed, 58.75 kDa, corresponds to a hexameric assembly of *Aae* Hfq.

Figure 4 High-affinity binding of *Aae* Hfq to A– and U–rich RNAs, with variable Mg^{2+} dependencies. Binding was quantified via fluorescence polarization assays using 5 nM FAM-U_6 (red) or FAM-A_{18} (blue) and varying concentrations of Hfq, either in the absence (thin lines) or presence (thick lines) of 10 mM $MgCl_2$. For each binding reaction, data from three replicates (standard errors given by vertical bars) were fit using a four-parameter logistic function to model the sigmoidal binding isotherm; nonlinear fits were also performed with an alternative model, accounting for receptor-depletion but neglecting cooperativity (§2.4 and Supp Fig S4). The computed binding constants are given (inset) in terms of the $(Hfq)_6$ concentration, as the stoichiometry of all characterized Hfq•RNA complexes, as well as the structural results reported herein, suggest that a hexamer is the active/functional unit. The addition of Mg^{2+} increases the binding affinity for both FAM-U_6 and FAM-A_{18}, albeit with a greater influence for the U-rich (proximal site–binding) RNA. Significant binding was not detected for a shorter A-rich (FAM-A_6) or C-rich (FAM-C_6) ssRNA.

Figure 5 Crystal structure of *Aae* Hfq in the *apo* form, with head–to–tail stacking of hexameric rings. The *apo* form of *Aae* Hfq crystallized in *P*1 as a dodecameric assembly of hexamers, stacked in a *proximal*-to-*distal* orientation in the lattice. Ribbon diagrams of the final, refined structure are shown here, from perpendicular viewpoints. The proximal-exposed (*PE*) hexamer is coloured blue and cyan, and subunits in the distal-exposed (*DE*) hexamer are coloured alternatingly yellow and orange. Co-crystallizing molecules of MPD (grey carbons) and GndCl (green carbons) are shown as ball-and-stick representations, and Cl^- ions are rendered as yellow spheres scaled to the van der Waals radius. Note that many of the Gnd cations and Cl^- anions are coplanar, where they form a 'salty' layer at the ring interface (this is most clearly seen in the transverse view). Contacts between hexamers are mediated by the *N*-termini of the *DE* hexamer (top) and the loop L2/strand β2 regions of the *PE* hexamer (bottom); the approximate location of one of the lateral rim RNA-binding sites is labelled on the *DE* ring.

Figure 6 Structural variation across the *Aae* Hfq monomer (*P*6) and dodecamer (*P*1) crystal forms. At a gross structural level, the two Hfq rings in the head-tail dodecamer of the *P*1 crystal form (Fig 5, axial view) appear to be related by a rigid-body rotation. The two rings—the proximal-exposed (*PE*) and distal-exposed (*DE*) hexamers—were brought, via pure rigid-body translation, to a common origin, indicated by the blue sphere in (*a*). Best-fit planes to each ring were then computed, as described in the Methods section (§2.6) and shown here as semi-transparent hexagonal plates of either orange (*DE* ring) or cyan (*PE* ring) colour. For clarity, the *DE* ring (orange/yellow in Fig 5) is omitted in panel (*a*), and a couple of the L2 loops are labelled (in the *PE* ring) simply as a structural landmark. The three principal axes of the moment of inertia tensor are shown in either orange (*DE* ring) or blue (*PE* ring); large differences in the orientation of these principal axes are marked by green and red 'Δ' symbols, while a 'δ' symbol (blue) denotes smaller-scale differences. The rotation between the rings is clear from the relative disposition (Δ) of two of the principal axes. Furthermore, a small—but discernable—difference (δ) in the directions of the normal axes indicates a slight tilt between the rings; this direction would correspond to the 6-fold axis in a perfectly symmetric double-hexamer. A multiple structural alignment of the 12 subunits in the *P*1 cell (*b*) reveals little structural variation of the Sm core (shown as C_α backbone traces), while there are many examples of side-chain variability (as noted in the panel). The defining secondary structural elements of the Sm fold (L1 loop, β1 strand, etc.), as well as the termini, are labelled. The two regions of *Aae* Hfq that most extensively engage in interactions between rings (*hexamer···hexamer* contacts in Fig 5), and in forming crystal contacts, are the L4 loops and the irregularly-structured ≈5 residues at the *N*-terminus (preceding α1). These also are the two most variable regions in Hfq, both in terms of sequence length (and composition) as well as 3D structure, as seen in (*b*). The side-chain variability shown in (*b*) takes two forms: *(i)* alternate conformers that could be built for a single residue, such as the Q52 example highlighted to the left, and *(ii)* rotameric variation for a single residue across the 12 subunits, such as the groups of

three residues shown as sticks near the top of (*b*). In many instances of the latter case, the 12 residue states clustered into two groups, corresponding to the *DE* or *PE* hexamer. In the diagram of panel (*c*), the Hfq subunits in *P*1, labelled by chain ID, are evenly spaced about a circle; arcs are drawn between the most structurally similar pairs of subunits, with the line thickness inversely scaled by the RMSD for the given pair. For clarity, not all $\approx n^2$ edges are shown here, but rather only at the levels of subunit pairs and triples (i.e., the deepest and second-deepest levels of leaf-nodes in the full dendrogram of Supp Fig S5c). This result, from hierarchical clustering on backbone RMSDs, shows that pairs of monomers within a given hexamer are structurally more similar to each other than are pairs between hexamers (chains A→F comprise the *PE* ring, and G→L are the *DE* ring).

Figure 7 Crystal structure of *Aae* Hfq with U–rich RNA bound at the lateral rim. The asymmetric unit of the *P*6 form contains a single Hfq subunit, shown as a tan-coloured ribbon diagram (*a*), in addition to 36 H_2O molecules (red spheres), a molecule of PEG (lime-coloured carbons), a molecule of MPD (gray carbons) and one molecule of U_6 RNA (green carbons). Non–protein atoms are represented as balls-and-sticks, using CPK colours (except as noted above for carbons). Expansion of the ASU to the full *P*6 cell gives an intact Hfq hexamer, shown onto the proximal face in (*a*). The meshes delimit the $2mF_o - DF_c$ electron density map, contoured at 1.5σ and shown only in the regions of RNA (dark blue) or MPD (light blue). The fragment of U_6 that could be unambiguously built into electron density contained two complete uridines and the 5′ phosphate moiety of the next residue; the path of this RNA strand is denoted by orange ① and ③ symbols for the ribonucleotides, from 5′→3′. Unexpectedly, U_6 nucleotides were found on the outer rim of *Aae* Hfq, in a position analogous to the lateral site of other Hfqs (*b*), while a molecule of MPD occupied the U-rich–binding pore as shown in (*c*). This magnified view (*b*) of the lateral site (same colour scheme as *a*) shows the RNA-contacting residues (labelled) in greater detail; asterisks distinguish residues from the *N*-termini of a neighbouring subunit, as also indicated in (*a*). Electron density maps such as this one were readily interpretable as RNA (see also Supp Fig S7). The magenta dashed lines (hydrogen bonds) and semi-transparent green cylinders (π-stacking interactions) indicate enthalpically favourable Hfq···RNA contacts. Most such contacts are mediated by both backbone and side-chain atoms of *Aae* Hfq, as well as the nucleobase and phosphodiester groups of the RNA; the ribose rings project outward from the cleft, and interact with Hfq more sparsely. (*c)* MPD binds at the pore and mimics the Hfq···uridine contacts found at the proximal RNA-binding site in some Hfq homologs. Contacts denoted by magenta dashed lines identically match the contacts to a uridine nucleotide in other Hfq structures containing U-rich RNA (see also Supp Fig S8). The green line indicates a van der Waals contact between L39 and MPD, and the green cylinder denotes another apolar interaction between *Aae* Hfq···MPD; this latter contact would presumably be replaced by a π-stacking interaction between F40 and a U base, were a U-rich RNA (rather than MPD) bound at the proximal site.

Figure 8 Conserved pattern of interatomic contacts at the lateral RNA-binding site of Hfq hexamers. In this schematic diagram of the interatomic contacts between the lateral site of *Aae* Hfq, U_6 RNA and nearby H_2O molecules (*a*), protein atoms are shown as ball-stick representations (CPK colouring, light grey carbons) and covalent bonds in the nucleotides are drawn as thicker, orange-coloured lines. For clarity, only a subset of H_2O molecules is drawn (green, labelled 'W#'). Here, asterisks denote another Hfq chain in the same unit cell and the prime symbol denotes a neighbouring cell. Hydrogen bonds are magenta for protein···RNA interactions, while those to H_2O are green. Stacking interactions between aromatic entities φ_1 and φ_2 are indicated by green circles from $\varphi_1 \cdots \varphi_2$. Two nucleotides of uridine (labelled) appear in an open, bridging conformation with the α-helix and β2 strand of an Hfq monomer (grey flanking regions). The phosphate groups are hydrogen bonded to N11 and R14 of the *N*-terminal α-helix, while the nucleobase hydrogen bonds with the backbone atoms of strand β2 (specifically, S36 and F37), thus imparting specificity for uridine. Note that additional π-stacking interactions are present between the side-chain of F37 and RNA base U2, as well as within the RNA (between U2···U1; not shown, for clarity). The lateral pocket of *Eco* Hfq is shown in (*b*), complexed with the sRNA RydC (same colouring scheme and conventions as in *a*). The U46 and U47 bases adopt conformations similar to those seen in (*a*), with the phosphate groups contacting residues of the α-helix. F39 π-stacks with U47, analogous to the interaction seen in *Aae* Hfq. Note that the adjacent G45 and A48 bases are flipped away from the pocket, and are shown here to offer context in the overall sequence of the sRNA. While not strictly conserved in terms of precise amino acid sequence, the *N*-terminal regions of the *Aae* and *Eco* Hfq homologues do provide similar backbone interactions with U1 and U46, respectively. Note also the directionality of the RNA backbone, which follows the same 5′→3′ path along the lateral site on the surface of the *Aae* and *Eco* Hfq rings (see also Figs 7a, b).

Figure 9 The lateral site of *Aae* Hfq is pre-structured for RNA-binding. The 3D structure of the single, unique monomer from the Hfq-U_6 co-crystal structure (teal backbone) was superimposed with the twelve subunits of the *apo* Hfq structure (grey). Residues that contact RNA, to within ≈ 3.6 Å in the *P*6 Hfq•U_6 structure, are shown as sticks for both the *P*6 and *P*1 structures. Apart from residue E7, which sterically occludes the binding pocket and thus likely adopts a different conformation upon RNA binding, note that the side-chains in the *apo* structure adopt rotameric states quite similar to those in the 3D structure of U_6-bound *Aae* Hfq. This finding suggests pre-organization of *Aae* Hfq's RNA-binding site.

Table 1 X-ray diffraction data collection and processing statistics

Values for the highest-resolution shell are given in parentheses.

	Aae Hfq, *apo* form ('*P*1')	*Aae* Hfq•U_6 RNA ('*P*6')
Diffraction source	APS NE-CAT 24-ID-E	APS NE-CAT 24-ID-C
Wavelength (Å)	0.9792	0.9195
Temperature (K)	100	100
Detector	ADSC Q315 CCD	Dectris Pilatus 6MF
Crystal-detector distance (mm)	200	300
Rotation range per image (°)	1.0	1.0
Total rotation range (°)	400.0	300.0
Exposure time per image (s)	1.0	1.0
Space-group	*P*1	*P*6
a, *b*, *c* (Å)	63.46, 66.06, 66.10	66.19, 66.19, 34.21
α, β, γ (°)	60.05, 83.94, 77.17	
Mosaicity (°)	0.143	0.107
Resolution range (Å)	57.27–1.49 (1.53–1.49)	34.21–1.50 (1.55–1.50)
Total No. of reflections	299 450	46 203
No. of unique reflections	138 120	13 177
Completeness (%)	93.7 (83.7)	94.9 (93.4)
Redundancy	2.2 (2.1)	3.5 (3.5)
$\langle I/\sigma(I)\rangle$	14.0 (3.4)	12.3 (3.6)
R_{sym}[†]	0.039 (0.258)	0.056 (0.292)
R_{meas}[‡]	0.052 (0.349)	0.065 (0.345)
$R_{p.i.m.}$[‡]	0.035 (0.234)	0.032 (0.179)
$CC_{½}$[§]	0.998 (0.886)	0.998 (0.942)
Overall *B*-value from Wilson plot (Å^2)	12.62	15.87
Matthews coefficient, V_M (Å^3 Da^{-1})	2.06 (for 12 subunits/AU)	2.28 (for 1 subunit/AU)
Solvent content (% volume)	40.21	46.08

[†] $R_{sym} = (\sum_{hkl} \alpha \sum_i |I_i(hkl) - \langle I_i(hkl)\rangle|)/(\sum_{hkl}\sum_i I_i(hkl))$, 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, $R_{p.i.m.}$, 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.

Table 2 Structure determination and model refinement

Values for the highest-resolution shell are given in parentheses.

	Aae Hfq, *apo* form ('*P*1')	*Aae* Hfq•U_6 RNA ('*P*6')
Resolution range (Å)	46.35–1.49 (1.51–1.49)	34.21–1.50 (1.56–1.50)
Completeness (%)	93.9	94.9
No. of reflections, working set	138104 (12739)	13171 (1308)
No. of reflections, test set	10625 (983)	662 (70)
Final R_{cryst}	0.1323 (0.1531)	0.1443 (0.1499)
Final R_{free}	0.1696 (0.2108)	0.1719 (0.1933)
No. of non-H atoms		
Macromolecules	7670 Hfq	598 Hfq, 43 RNA
Ligands	200 MPD, 32 Gnd, 7 Cl^-, 28 PEG	8 MPD, 7 PEG
Solvent	413 H_2O	36 H_2O
Total	8350	692
No. residues of protein, solvent or ligand molecules included in the final, refined structure	*Aae* Hfq 848 (over 12 subunits) H_2O 413 MPD 25 Cl^- 7 Gnd 8 PEG‡ 4	*Aae* Hfq 71 (over 1 subunit) U_6 RNA ≈ 2–3† H_2O 36 MPD 1 PEG‡ 1
R.m.s. deviations		
Bonds (Å)	0.005	0.005
Angles (°)	0.75	0.76
Average *B*-factors (Å^2)		
Protein	19.32	22.18
Ligand	25.89	30.44
Ramachandran plot		
Most favoured (%)	98	97
Allowed (%)	1.7	2.9
Outliers (%)	0	0
Rotamer outliers (%)	0.34	1.5
PDB ID	5SZD	5SZE

†This value is given as a range because two complete U nucleotides, plus a fragment of a third residue, could be built into electron density maps. ‡Fragments of polyethylene glycol could be built in both structures, generally of two to three repeat units (i.e., $(O–C–C)_2–O$, neglecting hydrogens).

Acknowledgements We thank H. Huber (Regensburg) for providing a sample of *A. aeolicus* genomic material; J. Bushweller (UVa) for access to a fluorescence plate reader; J. Shannon (UVa) for assistance with MALDI-TOF instrumentation; D. Cascio and M. Sawaya (UCLA) for crystallographic advice; and L. Columbus (UVa) and C. McAnany (UVa) for helpful discussions. Beamlines NE-CAT 24-ID-C/E at Argonne National Lab's *Advanced Photon Source* are DOE facilities (DE-AC02-06CH11357), with NIH funding for general operations (GM103403) and for the Pilatus detector (RR029205). This work was funded by the Jeffress Memorial Trust (J-971), NIH training grant T32GM080186 (P.S.R.) and NSF CAREER award MCB–1350957.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 213-221.

Arakawa, T., Ejima, D., Li, T. & Philo, J. S. (2010). *J Pharm Sci* **99**, 1674-1692.

Arluison, V., Mura, C., Guzmán, M. R., Liquier, J., Pellegrini, O., Gingery, M., Régnier, P. & Marco, S. (2006). *J Mol Biol* **356**, 86-96.

Bakan, A., Meireles, L. M. & Bahar, I. (2011). *Bioinformatics* **27**, 1575-1577.

Balbontín, R., Fiorini, F., Figueroa-Bossi, N., Casadesús, J. & Bossi, L. (2010). *Mol Microbiol* **78**, 380-394.

Bandyra, K. J. & Luisi, B. F. (2013). *RNA Biol* **10**, 627-635.

Belew, M., Porath, J., Fohlman, J. & Janson, J.-C. (1978). *Journal of Chromatography A* **147**, 205-212.

Bocchetta, M., Gribaldo, S., Sanangelantoni, A. & Cammarano, P. (2000). *Journal of Molecular Evolution* **50**, 366-380.

Bøggild, A., Overgaard, M., Valentin-Hansen, P. & Brodersen, D. E. (2009). *Febs J* **276**, 3904-3915.

Boto, L. (2010). *Proceedings. Biological sciences / The Royal Society* **277**, 819-827.

Burggraf, S., Olsen, G. J., Stetter, K. O. & Woese, C. R. (1992). *Systematic and Applied Microbiology* **15**, 352-356.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. (2009). *BMC Bioinformatics* **10**, 421.

Chen, V. B., Arendall, W. B., 3rd, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Crystallographica Section D, Biological crystallography* **66**, 12-21.

Cieślik, M., Derewenda, Z. S. & Mura, C. (2011). *Journal of Applied Crystallography* **44**, 424-428.

Colovos, C. & Yeates, T. O. (1993). *Prot Sci* **2**, 1511-1519.

De Mey, M., Lequeux, G., Maertens, J., De Maeseneire, S., Soetaert, W. & Vandamme, E. (2006). *Analytical Biochemistry* **353**, 198-203.

Dimastrogiovanni, D., Frohlich, K. S., Bandyra, K. J., Bruce, H. A., Hohensee, S., Vogel, J. & Luisi, B. F. (2014). *Elife* **3**.

Edgar, R. C. (2004). *Nucleic Acids Res* **32**, 1792-1797.

Eisenberg, D., Lüthy, R. & Bowie, J. U. (1997). *Methods in enzymology* **277**, 396-404.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Crystallographica Section D* **66**, 486-501.

Eveleigh, R. J., Meehan, C. J., Archibald, J. M. & Beiko, R. G. (2013). *Genome Biol Evol* **5**, 2478-2497.

Fadouloglou, V. E., Kokkinidis, M. & Glykos, N. M. (2008). *Anal Biochem* **373**, 404-406.

Fantappie, L., Metruccio, M. M., Seib, K. L., Oriente, F., Cartocci, E., Ferlicca, F., Giuliani, M. M., Scarlato, V. & Delany, I. (2009). *Infect Immun* **77**, 1842-1853.

Fischer, U., Englbrecht, C. & Chari, A. (2011). *Wiley Interdiscip Rev RNA* **2**, 718-731.

Folichon, M., Arluison, V., Pellegrini, O., Huntzinger, E., Regnier, P. & Hajnsdorf, E. (2003). *Nucleic Acids Res* **31**, 7302-7310.

Folta-Stogniew, E. J. (2009). *Macromolecular Interactions: Light Scattering*, *Encyclopedia of Life Sciences (ELS)*. Chichester: John Wiley & Sons, Ltd.

Franze de Fernandez, M. T., Eoyang, L. & August, J. T. (1968). *Nature* **219**, 588-590.

Franze de Fernandez, M. T., Hayward, W. S. & August, J. T. (1972). *J Biol Chem* **247**, 824-831.

Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. (1999). *Bioinformatics* **15**, 305-308.

Hamilton, W. (1965). *Acta Crystallographica* **18**, 502-510.

Horstmann, N., Orans, J., Valentin-Hansen, P., Shelburne, S. A., 3rd & Brennan, R. G. (2012). *Nucleic Acids Res* **40**, 11023-11035.

Huber, R. & Eder, W. (2006). *Aquificales*, *Proteobacteria: Delta and Epsilon Subclasses. Deeply Rooting Bacteria*, Third ed., edited by M. Dworkin, S. Falkow, E. Rosenberg, K. H. Schleifer & E. Stackebrandt, pp. 925-938: Springer New York.

Humphrey, W., Dalke, A. & Schulten, K. (1996). *J Mol Graph* **14**, 33-38, 27-38.

Ikeda, Y., Yagi, M., Morita, T. & Aiba, H. (2011). *Mol Microbiol* **79**, 419-432.

Ishikawa, H., Otaka, H., Maki, K., Morita, T. & Aiba, H. (2012). *RNA* **18**, 1062-1074.

Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). *ACM Computing Surveys* **31**, 264-323.

Jancarik, J. & Kim, S.-H. (1991). *Journal of Applied Crystallography* **24**, 409-411.

Janin, J., Bahadur, R. P. & Chakrabarti, P. (2008). *Q Rev Biophys* **41**, 133-180.

Joosten, R. P., Joosten, K., Murshudov, G. N. & Perrakis, A. (2012). *Acta Crystallogr D Biol Crystallogr* **68**, 484-496.

Kabsch, W. (2010). *Acta Crystallographica Section D* **66**, 125-132.

Kambach, C., Walke, S., Young, R., Avis, J. M., de la Fortelle, E., Raker, V. A., Luhrmann, R., Li, J. & Nagai, K. (1999). *Cell* **96**, 375-387.

Kantardjieff, K. A. & Rupp, B. (2003). *Prot Sci* **12**, 1865-1871.

Katoh, K. & Standley, D. M. (2013). *Molecular Biology and Evolution* **30**, 772-780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. (2012). *Bioinformatics* **28**, 1647-1649.

Keating, K. S. & Pyle, A. M. (2010). *Proc Natl Acad Sci USA* **107**, 8177-8182.

Klock, H. E. & Lesley, S. A. (2009). *Methods in Molecular Biology* **498**, 91-103.

Kovach, A. R., Hoff, K. E., Canty, J. T., Orans, J. & Brennan, R. G. (2014). *RNA* **20**, 1548-1559.

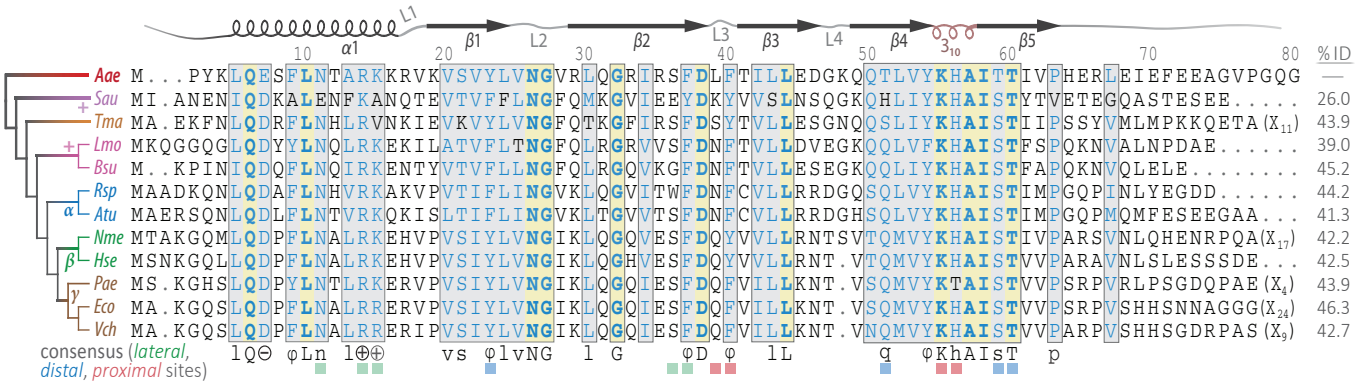
Laskowski, R. A., Macarthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *Journal of Applied Crystallography* **26**, 283-291.

Laskowski, R. A. & Swindells, M. B. (2011). *J Chem Inf Model* **51**, 2778-2786.

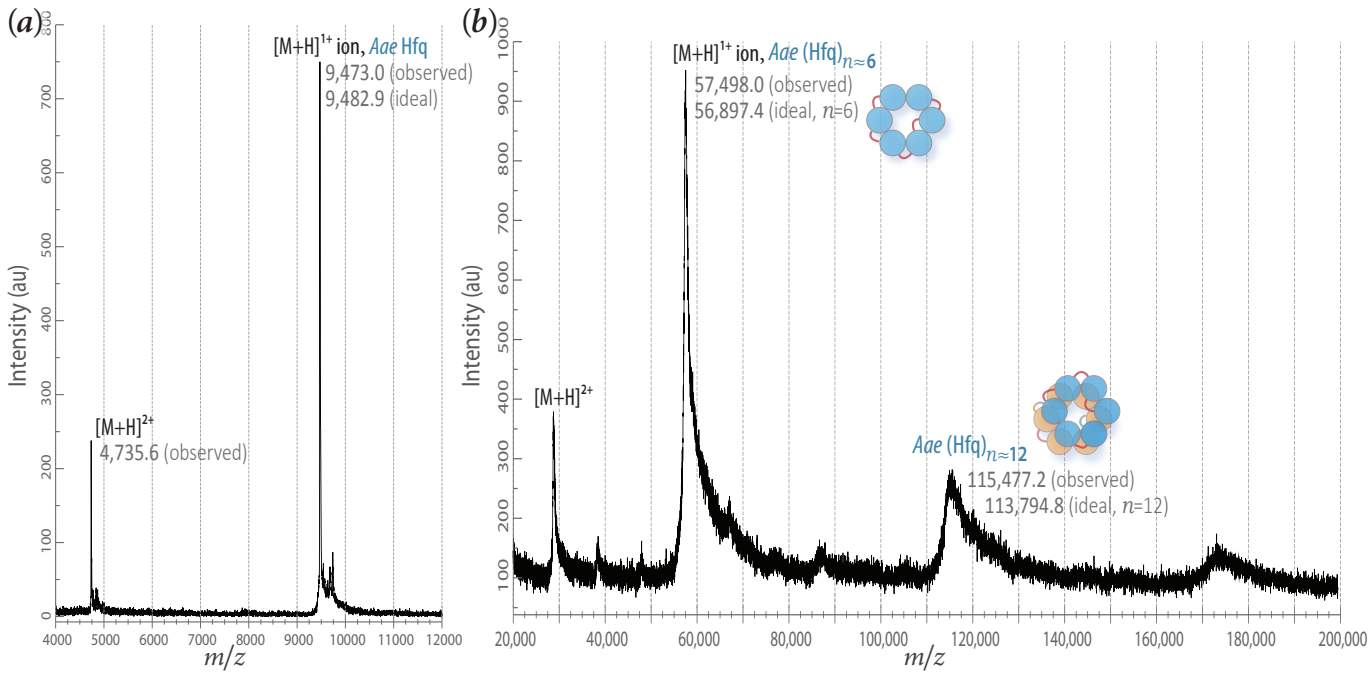
Lee, B. & Richards, F. M. (1971). *J Mol Biol* **55**, 379-400.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S. & Bassler, B. L. (2004). *Cell* **118**, 69-82.
Leung, A. K., Nagai, K. & Li, J. (2011). *Nature* **473**, 536-539.
Link, T. M., Valentin-Hansen, P. & Brennan, R. G. (2009). *Proc Natl Acad Sci USA* **106**, 19292-19297.
Lu, X.-J., Bussemaker, H. J. & Olson, W. K. (2015). *Nucleic Acids Res*.
Mandin, P. & Gottesman, S. (2010). *The EMBO Journal* **29**, 3094-3107.
Martin, A. C. R. & Porter, C. T. (2009). *ProFit,* http://www.bioinf.org.uk/software/profit.
Masse, E. & Gottesman, S. (2002). *Proc Natl Acad Sci USA* **99**, 4620-4625.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Journal of Applied Crystallography* **40**, 658-674.
McLachlan, A. (1982). *Acta Crystallographica Section A* **38**, 871-873.
Merritt, E. A. (2012). *Acta Crystallogr D Biol Crystallogr* **68**, 468-477.
Mika, F. & Hengge, R. (2013). *Int J Mol Sci* **14**, 4560-4579.
Mikulecky, P. J., Kaw, M. K., Brescia, C. C., Takach, J. C., Sledjeski, D. D. & Feig, A. L. (2004). *Nat Struct Mol Biol* **11**, 1206-1214.
Mohanty, B. K., Maples, V. F. & Kushner, S. R. (2004). *Mol Microbiol* **54**, 905-920.
Morin, A., Eisenbraun, B., Key, J., Sanschagrin, P. C., Timony, M. A., Ottaviano, M. & Sliz, P. (2013). *Elife* **2**, e01456.
Mura, C., Kozhukhovsky, A., Gingery, M., Phillips, M. & Eisenberg, D. (2003). *Prot Sci* **12**, 832-847.
Mura, C., McCrimmon, C. M., Vertrees, J. & Sawaya, M. R. (2010). *PLOS Comput Biol* **6**.
Mura, C., Phillips, M., Kozhukhovsky, A. & Eisenberg, D. (2003). *Proc Natl Acad Sci (USA)* **100**, 4539-4544.
Mura, C., Randolph, P. S., Patterson, J. & Cozen, A. E. (2013). *RNA Biology* **10**, 636-651.
Murina, V., Lekontseva, N. & Nikulin, A. (2013). *Acta Cryst. D* **69**, 1504-1513.
Murina, V. N., Melnik, B. S., Filimonov, V. V., Uhlein, M., Weiss, M. S., Müller, U. & Nikulin, A. D. (2014). *Biochemistry (Moscow)* **79**, 469-477.
Nikulin, A., Stolboushkina, E., Perederina, A., Vassilieva, I., Blaesi, U., Moll, I., Kachalova, G., Yokoyama, S., Vassylyev, D., Garber, M. & Nikonov, S. (2005). *Acta Crystallographica Section D* **61**, 141-146.
Oshima, K., Chiba, Y., Igarashi, Y., Arai, H. & Ishii, M. (2012). *International Journal of Evolutionary Biology* **2012**, 9.
Pagano, J. M., Clingman, C. C. & Ryder, S. P. (2011). *RNA* **17**, 14-20.
Panja, S., Schu, D. J. & Woodson, S. A. (2013). *Nucleic Acids Res* **41**, 7536-7546.
Patterson, J. & Mura, C. (2013). *J Vis Exp (JoVE)*, e50225.
Regnier, P. & Hajnsdorf, E. (2013). *RNA Biol* **10**, 602-609.
Robinson, K. E., Orans, J., Kovach, A. R., Link, T. M. & Brennan, R. G. (2014). *Nucleic Acids Res* **42**, 2736-2749.
Sanner, M. F., Olson, A. J. & Spehner, J. C. (1996). *Biopolymers* **38**, 305-320.
Sauer, E. (2013). *RNA Biology* **10**, 610-618.
Sauer, E., Schmidt, S. & Weichenrieder, O. (2012). *Proc Natl Acad Sci USA* **109**, 9396-9401.
Sauer, E. & Weichenrieder, O. (2011). *Proc Natl Acad Sci USA* **108**, 13065-13070.
Schulz, E. C. & Barabas, O. (2014). *Acta Cryst. F* **70**, 1492-1497.
Schumacher, M. A., Pearson, R. F., Moller, T., Valentin-Hansen, P. & Brennan, R. G. (2002). *EMBO J* **21**, 3546-3556.
Shrake, A. & Rupley, J. A. (1973). *J Mol Biol* **79**, 351-371.
Sittka, A., Sharma, C. M., Rolle, K. & Vogel, J. (2009). *RNA Biology* **6**, 266-275.
Sledjeski, D. D., Whitman, C. & Zhang, A. (2001). *J Bacteriol* **183**, 1997-2005.
Someya, T., Baba, S., Fujimoto, M., Kawai, G., Kumasaka, T. & Nakamura, K. (2012). *Nucleic Acids Res* **40**, 1856-1867.
Soper, T., Mandin, P., Majdalani, N., Gottesman, S. & Woodson, S. A. (2010). *Proc Natl Acad Sci USA* **107**, 9602-9607.
Sun, X. & Wartell, R. M. (2006). *Biochemistry* **45**, 4875-4887.
Sun, X., Zhulin, I. & Wartell, R. M. (2002). *Nucleic Acids Res* **30**, 3662-3671.
Tharun, S. (2009). *Int Rev Cell Mol Biol* **272**, 149-189.
Thore, S., Mayer, C., Sauter, C., Weeks, S. & Suck, D. (2003). *The Journal of Biological Chemistry* **278**, 1239-1247.
Tycowski, K. T., Kolev, N. G., Conrad, N. K., Fok, V. & Steitz, J. A. (2006). *12 The Ever-Growing World of Small Nuclear Ribonucleoproteins*.
Updegrove, T. B., Correia, J. J., Chen, Y., Terry, C. & Wartell, R. M. (2011). *RNA* **17**, 489-500.
Updegrove, T. B., Correia, J. J., Galletto, R., Bujalowski, W. & Wartell, R. M. (2010). *Biochim Biophys Acta* **1799**, 588-596.
Updegrove, T. B. & Wartell, R. M. (2011). *Biochim Biophys Acta* **1809**, 532-540.
Updegrove, T. B., Zhang, A. & Storz, G. (2016). *Curr Opin Microbiol* **30**, 133-138.
Veretnik, S., Wills, C., Youkharibache, P., Valas, R. E. & Bourne, P. E. (2009). *PLOS Comput Biol* **5**, e1000315.
Vogel, J. & Luisi, B. F. (2011). *Nat Rev Microbiol* **9**, 578-589.
Wagner, E. G. (2013). *RNA Biology* **10**, 619-626.
Wang, W., Wang, L., Wu, J., Gong, Q. & Shi, Y. (2013). *Nucleic Acids Res* **41**, 5938-5948.
Weichenrieder, O. (2014). *RNA Biology* **11**, 537-549.
Will, C. L. & Luhrmann, R. (2011). *Cold Spring Harb Perspect Biol* **3**.
Wilson, K. S. & von Hippel, P. H. (1995). *Proc Natl Acad Sci USA* **92**, 8793-8797.
Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Crystallographica Section D* **67**, 235-242.
Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C. & Storz, G. (2002). *Mol Cell* **9**, 11-22.
Zucker, F., Champ, P. C. & Merritt, E. A. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 889-900.

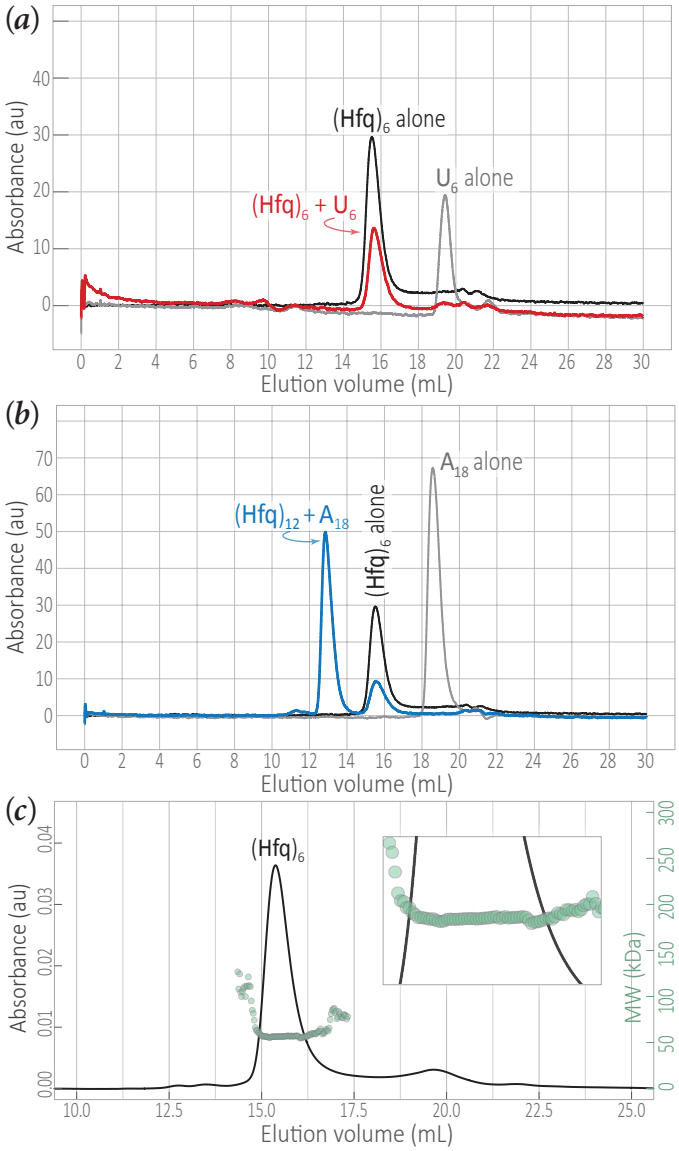
Stanek et al., Figure 1



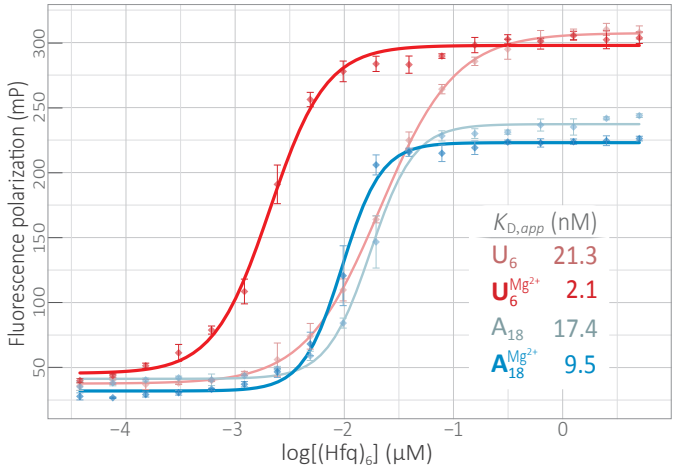
Stanek et al., Figure 2



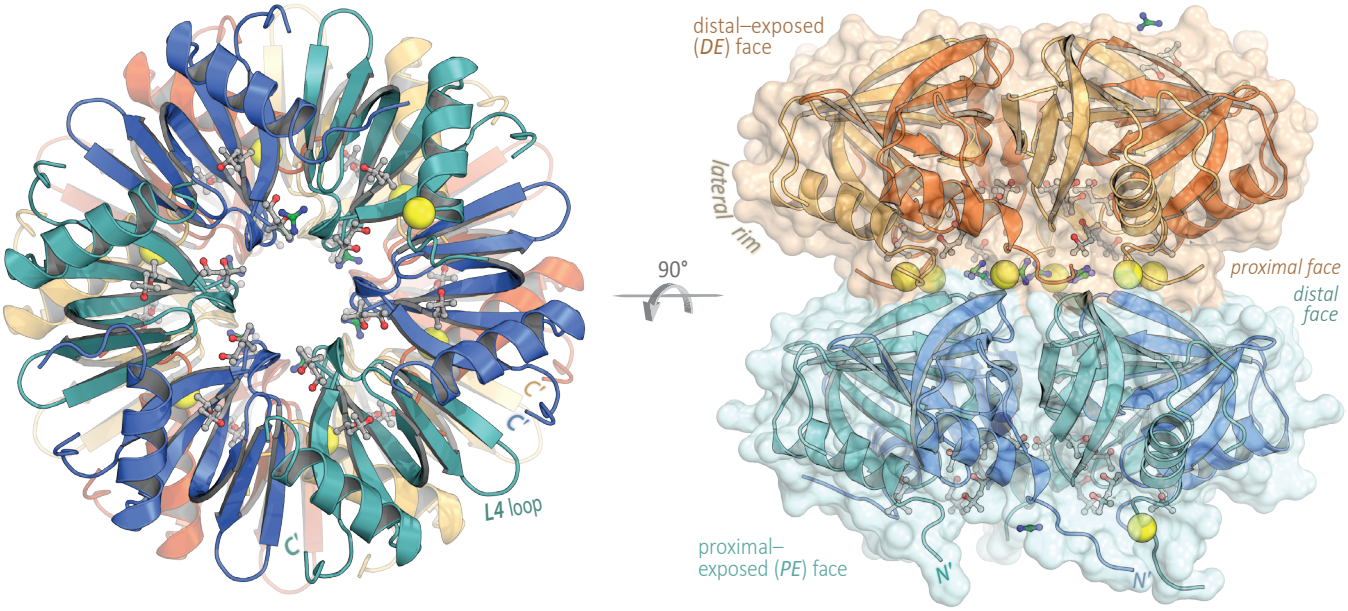
Stanek et al., Figure 3



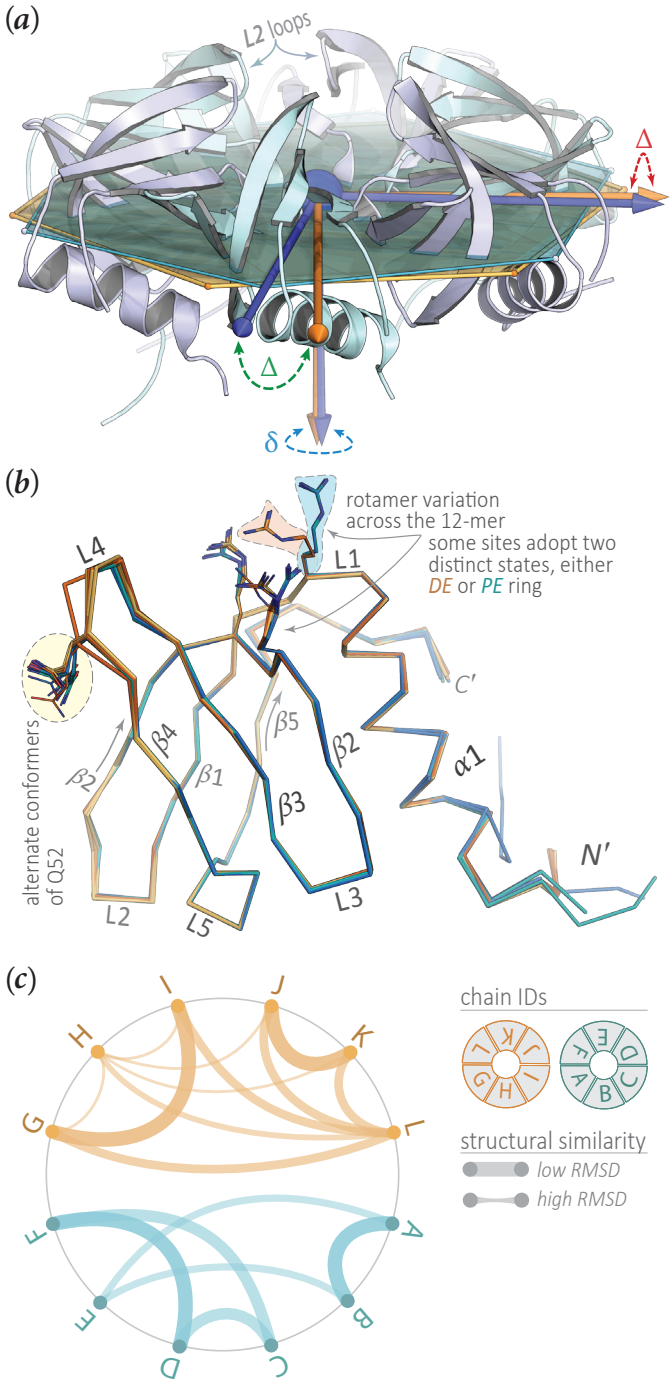
Stanek et al., Figure 4



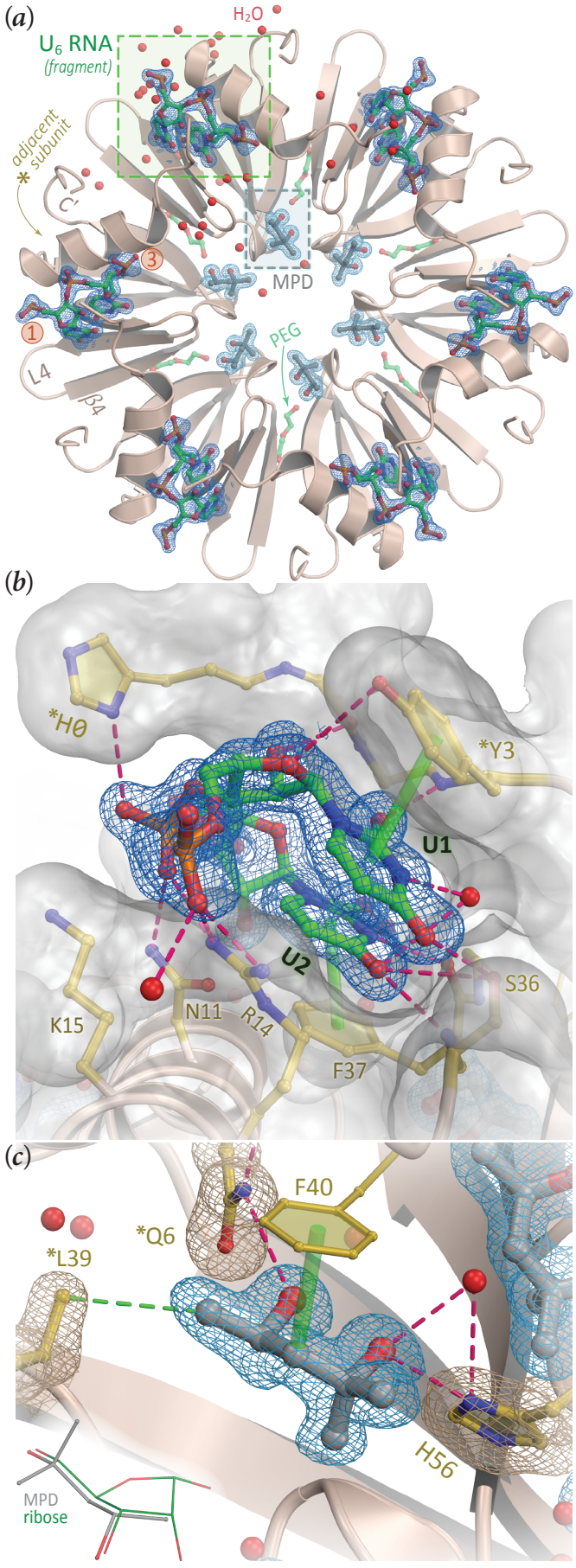
Stanek et al., Figure 5



Stanek et al., Figure 6

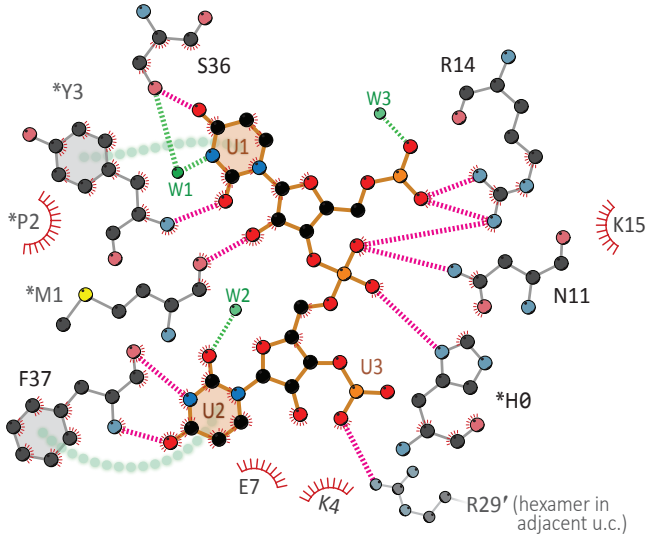


Stanek et al., Figure 7

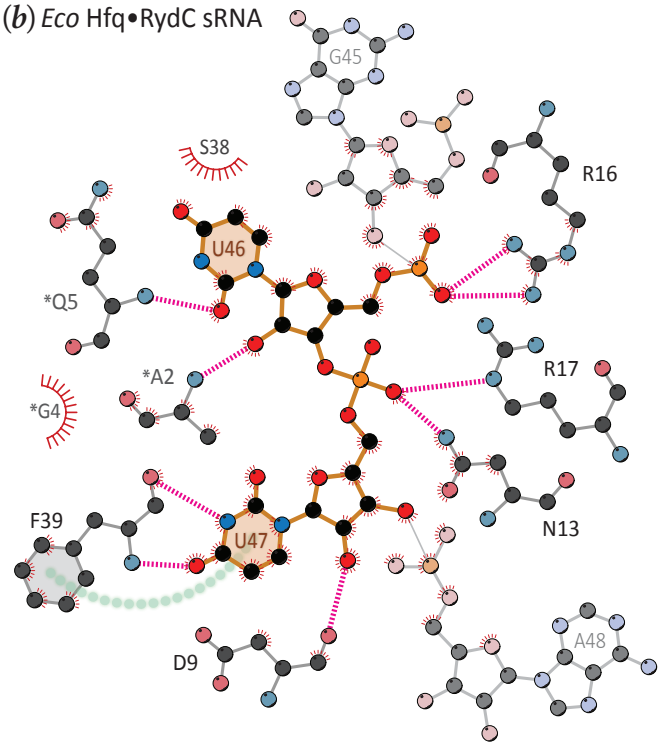


Stanek et al., Figure 8

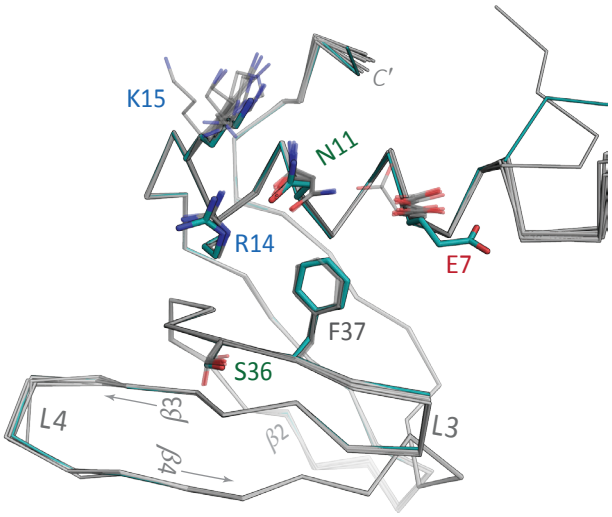
(*a*) *Aae* Hfq•U_6 RNA



(*b*) *Eco* Hfq•RydC sRNA



Stanek et al., Figure 9



Chapter 4: Structure of the second Hfq homolog from Aquifex aeolicus

Kimberly Stanek and Cameron Mura

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

Abstract

The bacterial host factor Hfq is an RNA-binding protein that facilitates the interaction of mRNAs with small regulatory RNAs (sRNAs), acting as a hub in transcriptional regulatory networks. Hfq acts in numerous physiological pathways, including stress response, quorum sensing, and expression of virulence factors. Several species of bacteria encode a second Hfq protein, denoted Hfq2, but little is known about the function of these paralogs. Furthermore, an Hfq2 structure also has yet to be reported. We have identified a second Hfq paralog in the genome of the deep-branching thermophile *Aquifex aeolicus* (*Aae*), and here we report the first structure of *Aae* Hfq2, to 2.0 Å resolution. Several known properties of Hfq, such as the overall Sm fold and propensity to form hexamers and dodecamers, are conserved in Hfq2. We have also co-crystallized *Aae* Hfq2 with uridine tri-phosphosphate, revealing that the proximal, U-rich RNA-binding site of Hfq1 is also conserved in Hfq2. Intriguingly, the arginine-rich lateral patch of Hfq1 is absent in Hfq2. *Aae* Hfq2 also exhibits higher-affinity binding to RNA at low pH and co-purifies with RNA as well as DNA when expressed in *E. coli*. Hfq1 and Hfq2 do not appear to associate with one another, and we suspect that Hfq2 serves a distinct functional role *in vivo*.

1. Introduction

The bacterial protein Hfq, discovered as a host factor required for the replication of the RNA bacteriophage Qβ [1], has since been found to act as a hub of post-transcriptional regulation. Hfq has roles in numerous pathways, including quorum sensing [2], stress response [3–5], and the expression of virulence factors [6,7]. This functional versatility is achieved via the ability of Hfq to bind to a large variety of mRNAs and small regulatory RNAs (sRNAs) [8–10]. The resulting physiological effect is that these sRNAs can upregulate [11] or downregulate [12,13] their mRNA target(s) through various mechanisms, and in many cases Hfq is required for these pairings to be successful [14].

Based on structural similarity, Hfq was identified as the bacterial branch of the Sm superfamily of proteins [15–17]. This ancient set of proteins is present in all domains of life and is ubiquitous in RNA processing [18]. In eukaryotes, the Sm and Sm-like (LSm) proteins act as scaffolds in snRNP assembly and mRNA splicing, as well as other other RNA-processing pathways [19,20]; the role of Sm homologs in archaea is still unclear. While the sequences and functions of Sm proteins vary greatly, the 3D structural fold of the family is highly conserved. The Sm fold can be thought of as a small β-barrel, consisting of an N-terminal α-helix followed by five highly-bent antiparallel β-strands [21]. The Sm domain is quite small (~60 residues for Hfq), however Sm proteins can oligomerize to form toroidal rings via backbone hydrogen-bonding interactions between the β4 and β5 strands of adjacent monomers; this provides a greatly enhanced surface for potentially binding RNA.

A long-standing question in the field of Sm biology is how these proteins have evolved such drastically different functionalities over time. Part of the answer likely lies in the oligomeric

propensities of the proteins. While bacterial genomes typically have just one copy of Hfq, which spontaneously oligomerizes as a hexamer, eukaryotes typically have numerous (> 7) Sm paralogs. These eukaryotic Sm proteins form heteroheptameric rings that encircle a U snRNA through a chaperoned assembly pathway [22]. As the pore of the Hfq hexamer is narrower, ssRNA is sterically unable to thread through the ring, and instead binds in a manner that encircles the pore [23]. Additional surface patches on the Hfq ring are also available for alternative RNA interactions; these do not seem to occur with the the eukaryotic Sm proteins [24]. The Sm-like archaeal proteins (SmAPs) provide some insight as an evolutionary intermediate; while they are more similar in sequence to their eukaryotic counterparts, they also exhibit behavior that is more Hfq-like [25]. For instance, SmAPs spontaneously self-oligomerize as homomeric rings, although their oligomeric state can vary from a hexamer to a 14-mer [26,27]. SmAP rings additionally bind RNA along their outer rim, in a manner similar to the recently characterized arginine-rich lateral binding site of Hfq [25,28,29].

When the first structures of bacterial Hfq became known, it was observed that several species of bacteria have two or more putative *hfq* genes [16]. Bacterial species encoding more than one Hfq may offer insight into the evolutionary transition from a single homohexameric Hfq to multiple paralogs in a heteroheptameric Sm. As the sequences of Hfq proteins vary greatly, beyond the limits of what can be reliably detected by sequence similarity alone, we predict that new Hfq paralogs will continue to be identified in bacterial species. Until very recently however, none of these putative ‘Hfq2’ or ‘Hfq3’ homologs had been verified as genuine Hfqs. These results from several groups indicate the presence of multiple Hfq paralogs in *Burkholderia cenocepacia* [30,31] and *Bacillus anthracis* [32,33]. However, we still have little information on the role of the additional Hfq copies *in vivo* and to date no atomic-resolution structures are

available for multiple Hfq paralogs from a single species. Note that Vrentas *et. al.* did construct homology models for *B. anthracis* Hfq1, Hfq2, and Hfq3 [32].

In the case of *B. cenocepacia*, the two *hfq* genes were found to be differentially expressed: Hfq1 was most highly expressed during the late log growth, while Hfq2 was maximally expressed during stationary phase [30]. Both proteins were shown to affect the virulence of *B. cenocepacia*. In addition, Hfq2 was shown to bind DNA and was negatively regulated by the sRNA h2cR [31]. *B. anthracis* has been demonstrated to have three *hfq* genes, two of which are genomically encoded and a third present on a plasmid [32,33]. Note different nomenclature was used in these two cited studies of *B. anthracis* Hfqs, such that the copies referred to as 'Hfq1' and 'Hfq2' in one study were reversed in the other study. Here we will use the naming scheme adopted by *Panda et al*., where Hfq2 and Hfq3 (plasmid-encoded) refer to the two more divergent paralogs. Biochemical characterization revealed that *B. anthracis* Hfq1 and Hfq3 form hexamers *in vitro*, and they can partially complement a Δ*hfq E. coli* strain. Hfq2, however, was unable to form a stable hexamer or restore phenotype in the Δ*hfq* strain and its role *in vivo* is still unclear; Vrentas *et. al.* suggest it may function to inhibit Hfq1.

Here we have verified the presence of a second Hfq, dubbed 'Hfq2', in a representative species from the Aquificae: *Aquifex aeolicus* (*Aae*). This early-branching extremophile additionally provides broad evolutionary insight into the Hfq family and Sm superfamily. We recently characterized the first *Aae* Hfq homolog, which we will now refer to as 'Hfq1'. We have now recombinantly expressed, purified, and crystallized the second *Aae* Hfq protein, and here we report the first crystal structure of a genuine 'Hfq2' homolog. We have also characterized the nucleic acid-binding properties of *Aae* Hfq2 through a combination of fluorescence polarization assays and co-purification with nucleic acids *in vivo*. This work led us to find that *Aae* Hfq2 binds to RNA and DNA with different affinities than Hfq1. We also examined the potential

association of *Aae* Hfq1 and Hfq2 and found no evidence for an interaction, indicating that Hfq2 act autonomously and likely has a distinct cellular function.

2. Materials and Methods

2.1 Cloning, expression and purification of *Aae* Hfq2

Aae hfq2 was first cloned from the full-length *Aae* genome into the pCR-Blunt vector using the Zero Blunt PCR Cloning Kit (Invitrogen). The *hfq2* gene insert was then introduced into the T7-promoter based pET-28b(+) *E. coli* expression plasmid through double-digestion of the *hfq2* insert and empty pET28b(+) vectors, with restriction enzymes NdeI and XhoI, followed by ligation of both products at room temperature at 18 °C for 4 hours with T4 DNA ligase. The *hfq2*-pET28b(+) plasmid was then amplified by transformation of chemically competent TOP10 *E. coli* cells and purified using a miniprep kit (Qiagen). The final recombinant construct consisted of an *N*-terminal His-6x tag followed by a thrombin-cleavable linker.

To recombinantly express *Aae* Hfq2, chemically competent BL21(DE3) *E. coli* cells were transformed with the *hfq2*-pET28b(+) plasmid and plated on LB-agar supplemented with 0.05 mg/ml kanamycin, followed by outgrowth in Lysogeny Broth (LB) media overnight with shaking (225 rpm) at 37 °C. Expression of *Aae* Hfq2 was induced by adding isopropyl-β-D-thiogalactoside (IPTG) to 1 mM when the cell cultures reached an optical density (OD_{600}) of 0.8-1.0. Cells were incubated for an additional 3-4 hours, followed by pelleting at 15,000*g* for 5 min at 5 °C and stored at -20 °C overnight. Cell pellets were resuspended in a partial lysis buffer (50 mM Tris pH 8, 750 mM NaCl, 0.4 mM PMSF) and chemically lysed with the addition of 0.01 mg/ml chicken egg white lysozyme (Fisher) followed by incubation at 37 °C

for 20 min. To ensure complete lysis, cells were then mechanically lysed using a microfluidizer. The cell lysate was then clarified by centrifugation at 35,000*g* for 20 min at 18 °C. Since most known Hfq proteins are thermostable, a heat-cut step was performed by incubating the lysate for 20 min at 75 °C and then pelleting at 35,000*g* for 20 min at 18 °C. The lysate was then filtered through a 0.2 µm syringe filter.

The His-tagged *Aae* Hfq2 construct was isolated via immobilized metal affinity chromatography (IMAC) using a pre-packed iminodiacetic acid Sepharose column (GE Lifesciences) charged with Ni^{2+}; this step was performed on an NGC medium-pressure liquid chromatography system (Bio-Rad). The lysate was loaded onto the column, which was then washed with four column volumes of wash buffer (50 mM Tris pH 8.5, 150 mM NaCl, 10 mM imidazole). The protein was eluted by applying a linear gradient of elution buffer (50 mM Tris pH 8.5, 150 mM NaCl, 600 mM), from 0 to 100% over 10 column volumes. The flow-through was fractionated and monitored via absorbance at 280 nm. Fractions containing protein, as assessed by A_{280}, were pooled and dialyzed into a buffer consisting of 50 mM Tris pH 8.0, 150 mM NaCl, and 12.5 mM EDTA. To cleave the 6x-His tag, the Hfq2 protein was incubated at RT overnight with a 1:600 mass ratio Hfq:thrombin. A benzamidine column was then used to remove the thrombin. Finally, to improve sample homogeneity, the Hfq2 protein was run through a preparative-grade HiPrep 16/600 Sephacryl S-300 HR gel-filtration column. *Aae* Hfq2 sample purity was assessed via SDS-PAGE and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF spectra), as described previously [29].

2.2 X-ray crystallography

2.2.1 Crystallization

Prior to crystallization attempts, *Aae* Hfq2 samples were dialyzed into 50 mM HEPES pH 8.0 and 200 mM NaCl, and then concentrated to 7.5 mg/ml. Initial sparse-matrix crystallization trials were performed under vapor diffusion/sitting drop format using the JCSG Core grid screens. Trays were set using a Mosquito liquid-handling robot and 96 well-format Intelliplates (Art Robbins) and equilibrated at 18 °C. Several initial hits were identified through visual inspection under a microscope and further screened by systematically varying pH and concentration of components of the crystallization condition. Grid screening and further crystallization efforts were conducted through hanging-drop vapor diffusion with 24 well-format plates. 6 µl drops (3 µl protein + 3 µl crystallization buffer) were equilibrated with 600 µl wells of crystallization buffer at 18 °C. Reproducible, birefringent and well-diffracting crystals were obtained using 0.1 M citric acid pH 4.0, 1 M LiCl, and 30% v/v PEG 400. The *Aae* Hfq2 crystals exhibited a parallelepiped morphology, and developed to a maximal size of 200 µm x 100 µm x 50 µm within 2 days.

2.2.2 Co-crystallization with single nucleotides

A co-crystallization screen was made by preparing 100 mM stocks of the following 15 nucleotides in ddH_2O: dAMP, dGMP, dCMP, dATP, dGTP, dCTP, dTTP, AMP, GMP, CMP, UMP, ATP, GTP, CTP, and UTP (stored at -20 °C until use). For these experiments, *Aae* Hfq2 was crystallized as described in §2.2.1, except that drops consisted of 3 µl protein + 2.4 µl crystallization buffer + 0.6 µl of 100 mM nucleotide, giving a 10 µM final concentration of

nucleotide in each case. Crystals developed within 2 days for half of the nucleotides tested, and within one week for the majority of the nucleotides tested. These co-crystals exhibited various morphologies (as well as multiple morphologies within the same well), including parallelepipeds, hexagonal plates, and most commonly, rounded ovals.

2.2.3 Diffraction data collection and processing

Crystals were harvested with nylon loops and flash cooled in liquid nitrogen. Diffraction data were collected at the Advanced Photon source (APS), at beamlines 24-ID-C and 22-ID for the *apo* and UTP-bound Hfq2 structures, respectively. Initial processing of the diffraction data (indexing, scaling, and merging) was performed in XDS [34]. *Aae* Hfq2 crystallized in spacegroup $P2_1$ with unit cell dimensions of a = 60.47, b = 101.06, c = 65.66 Å and β = 107.8°. Isomorphous crystals were obtained when *Aae* Hfq2 was co-crystallized with UTP. The *apo* and UTP-bound crystal forms diffracted to 1.98 Å and 1.85 Å, respectively. The initial quality of all diffraction datasets was analyzed using Xtriage within the PHENIX suite of programs [35]. Both the *apo* and UTP-bound datasets exhibited severe anisotropy and were submitted to the UCLA anisotropy server [36] for further analysis. Ultimately, ellipsoidal truncation was required only for the Hfq2-UTP dataset, to 2.5 Å in the c* direction. The data collection statistics for both datasets (including the ellipsoidally-truncated Hfq2-UTP dataset), are reported in Table 1.

2.2.4 Structure solution, refinement and validation

Initial phases for the *apo* Hfq2 diffraction data were estimated through molecular replacement. The *S. aureus* Hfq hexamer (PDB 1KQ1), with a sequence identity of 38% to *Aae* Hfq2, was used as a search model. The unit cell consisted of twelve monomers (two hexameric rings), with a Matthew's coefficient (V_M) of 1.91 $Å^3$ Da^{-1}. Polypeptide chains with the amino acid

sequence of *Aae* Hfq2 were built into the initial molecular replacement solution using AutoBuild within PHENIX. At this stage, the R_{work}/R_{free} split for the *apo Aae* Hfq2 structure was 0.240/0.271. Refinement proceeded through several cycles of automated refinement of atomic positions, occupancies and atomic displacement parameters (ADPs) as isotropic *B* factors in PHENIX and manual refinement, including placement of waters and ligands, in Coot [37]. Simulated annealing was also performed early in refinement. After an initial round of refinement of the Hfq2-UTP co-crystal dataset in PHENIX, twelve molecules of UTP were identified in the weighted mF_o-DF_c difference density map (one for each monomer) and manually placed into the model. The R_{work}/R_{free} split of the Hfq2-UTP structure before UTP density was built in was 0.225/0.274. The stereochemistry and contacts in the final structural models were validated through inspection in MolProbity [38]. Final refinement statistics for the *apo Aae* Hfq2 structure and the Hfq2-UTP structure are reported in Table 2.

2.2.5 Sequence and structural analysis

Hfq sequences were aligned using MUSCLE [39], within the Geneious bioinformatics platform [40]. The structure of *Aae* Hfq1, used for structural comparisons to Hfq2, is PDB 5SZD (i.e., with no RNA or ligands bound). The average pairwise RMSD of the Hfq2 monomer was calculated by averaging the RMSDs from structural alignment of each pairwise combination of the 12 monomers of *Aae* Hfq2. Pairwise structural alignments were performed through combinatorial extension, using the cealign algorithm within PyMOL. Average pairwise RMSD between *Aae* Hfq1 and Hfq2 was calculated in a similar manner. The buried surface area (or difference in solvent accessible surface area, ΔSASA) associated with dodecamerization was calculated using the point-counting method available in PyMOL. Electrostatic potentials, contoured at ± 5 k_BT/e for visualization purposes, were calculated using the Adaptive

Poisson-Boltzmann Solver (APBS) plugin within PyMOL. All molecular structural illustrations were created in PyMOL.

2.3 Fluorescence polarization-based assays

Fluorescence polarization assays and calculations of RNA-binding affinities were performed as previously described [29]. Briefly, polarization data was obtained using the following 5'-fluorescein–3'-OH-labelled oligoribonucleotides: FAM–U_6, FAM–C_6, FAM–A_6, and FAM–A_{18}. In these assays, 5 nM FAM-RNAs were added to a serial dilution of *Aae* Hfq2 at pH 5.5 (buffered in 25 mM MES and 250 mM NaCl) or pH 8.0 (buffered in 50 mM HEPES and 200 mM NaCl) and equilibrated for 45 minutes at room temperature. For binding assays in the presence of Mg^{2+}, solutions were supplemented with 10 mM $MgCl_2$. Polarization values were plotted against the log[$(Hfq)_6$], and the binding data were fit to the following four-parameter sigmoidal curve,

$$y(x) = A_2 + (A_1 - A_2)\left\{\frac{1}{1+exp[(x-x_0)/dx)}\right\}$$

where x is the log[$(Hfq)_6$], A_1 and A_2 are the polarization values at lower and upper bounds of the binding curve, respectively, x_0 is the apparent equilibrium dissociation constant ($K_{d,app}$), and the shape/slope parameter dx captures the Hill coefficient.

2.4 Nucleic acid co-purification

To co-purify *Aae* Hfq2 with endogenous *E. coli* RNAs, the recombinant construct was purified as in §2.1 except that the following buffer components were prepared at a lower pH of 6.0 (versus 8.0): lysis buffer (25 mM MES pH 6.0, 750 mM NaCl), wash buffer (25 mM MES pH 6.0, 150 NaCl, 10 mM imidazole) and elution buffer (25 mM MES pH 6.0, 150 mM NaCl, 600

mM imidazole). During the IMAC step, the absorbance at 260 nm was continuously monitored in addition to A_{280}. Samples from the fractionated eluate were run on both 1.5% agarose and 15-20% SDS-polyacrylamide gels to determine the size distribution of the nucleic acids. A series of colorimetric assays [41] were performed in order to distinguish between DNA, RNA, and contaminating sugars in the co-purifying nucleic acid samples. Controls consisted of 1 mg/ml ribose, 3 mg/ml DNA, 1 mg/ml RNA, and 3 mg/ml BSA. In these assays, 10 µl of Benedict's reagent, 50 µl of Orcinol reagent, or 10 µl Diphenylamine reagent were added to 50 µl samples to test for the presence of free-reducing sugars, pentose sugars, or DNA, respectively. Samples were placed in a boiling water bath for 20 minutes until any color-changes were observed.

3. Results and discussion

3.1 *Aae* Hfq2, the second Hfq paralog from *A. aeolicus*

Multiple Hfq paralogs have been identified and characterized in the gram-positive *B. anthracis* and the gram-negative β-proteobacterium *B. cenocepacia* [30,32,33]. Through sequence similarity searches, we found that members of class Aquificales also possess two annotated Hfq paralogs (Stanek & Mura, *unpublished*). We note that several of these putative Aquificales Hfq sequences are doubly annotated as iron-sulfur cluster assembly protein HesB, however this is likely an artefact of homology-based transfer of functional annotation [42], as these Aquificales Hfq2 sequences share no significant sequence similarity with HesB proteins that have been functionally characterized. Furthermore visual inspection of the Aquificales Hfq2 sequences reveals they have the conserved 'YKHAI' motif that is characteristic of an Hfq.

The gene for *Aae* Hfq2 encodes a 71-aa protein with a theoretical isoelectric point (pI) of 6.25. This is notably more acidic than *Aae* Hfq1, which has a pI of 9.45, though this pI is not unusual for an Hfq (*S. aureus* Hfq, as an extreme example, features a pI of 4.69). Through phylogenetic analysis of Hfq sequences (*Stanek et al., unpublished*), we found that *Aae* Hfq2 may share a common origin with *B. cereus* group Hfq2 and Hfq3. Notably, *Aae* Hfq2 and *B. anthracis* Hfq2 and Hfq3 share higher sequence identities with each other than with *Aae* Hfq1 or *B. anthracis* Hfq1 (Fig 1A). *Aae* Hfq2 and *B. anthracis* Hfq2 and Hfq3 also lack the acidic C-terminal tails found in other Hfq homologs [43–45]. These similarities could suggest that these additional paralogs have analogous functions in *A. aeolicus* and *B. anthracis*. Intriguingly though, examination of the nearest gene neighbors of *Aae hfq1* and *hfq2* reveals that *hfq2* is found upstream of the GTPase *hflx* (Fig 1B). This gene neighbor is conserved for most gram-negative *hfq* sequences, including *E. coli*.

Here, a recombinant *Aae hfq2* gene was cloned and expressed in *E. coli*. The final Hfq2 protein product, with an expected MW of 8,435 Da, was successfully purified, as assessed by MALDI-TOF spectra (Fig S1). While we previously found *Aae* Hfq1 co-purified with nucleic acids, this was not the case with Hfq2, as demonstrated by an absorbance ratio at 260 and 280 nm (A_{260}/A_{280}) of ~0.8 for Hfq2 samples. Hfq2 ran as an apparent pentamer via analytical size exclusion chromatography (Fig S2) with a molecular weight of 42.51 kDa. Such aberrant elution times have been observed with other Hfqs, including *Aae* Hfq1, due to interactions of the protein with the separation media [11,46]. Therefore, size exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) was used to verify the oligomeric state (Fig S3). SEC-MALS data showed that Hfq2 did indeed oligomerize as a hexamer; the protein eluted as a single, monodisperse peak with a weight-averaged molecular weight (M_W) of 48.56 kDa.

Ultimately, *Aae* Hfq2 was crystallized so that properties of this paralog, including its fold, oligomeric state, and entire 3D structure, could be elucidated at atomic resolution.

3.2 The overall structure of *Aae* Hfq2

We have determined the crystal structure of *Aae* Hfq2 to 2.0 Å resolution (Fig 2). The protein adopts the characteristic Sm fold: an N-terminal α-helix followed by five highly-bent β-strands arranged in an antiparallel sheet (Youkharibache et al., *in revision*). Notably, *Aae* Hfq2 oligomerizes as a hexamer, via hydrogen bonding of β4 and β5 strands of adjacent monomers. This behavior matches that of known Hfq and Sm oligomeric rings. *Aae* Hfq2 was crystallized in spacegroup *P*21, with twelve unique (not symmetry-related) monomers in the asymmetric unit, arranged as a dodecamer of two stacked hexamers (Fig 2A). Non-crystallographic symmetry (NCS) averaging was not used at any point during refinement, and the average pairwise RMSD between each of the twelve monomers in the ASU is 0.246 Å. The first ~10 amino acids of the N-terminus were not resolved in any of the chains and are likely disordered. The full C-terminal region of each subunit could be resolved, which was unsurprising, as *Aae* Hfq2 is missing the unstructured C-terminal extensions characteristic of other Hfq homologs. Instead, the C-terminus of *Aae* Hfq2 instead consists of just two glycine residues beyond the β5 strand.

The overall structure of *Aae* Hfq2 is notably similar to *Aae* Hfq1, with an RMSD of 0.93 Å between Hfq1 and Hfq2 hexamers . The largest differences between the structures occur at the L4 loop, N-terminus (including the positioning of the N-terminal α-helix) and C-terminus (Fig 3). The C-termini of *Aae* Hfq1 wrap around the lateral rim to extend away from the proximal face, whereas the shortened *Aae* Hfq2 C-termini are positioned in the opposite direction, towards the distal face. Notably, the L4 loop and termini of Hfq homologs are generally the least conserved regions of the protein, in terms of amino acid sequence.

Both *Aae* Hfq1 and Hfq2 exhibit the ability to crystallize as dodecamers, though *Aae* Hfq1 hexamers pack in a *proximal*-to-*distal* orientation, while *Aae* Hfq2 is arranged in a *distal*-to-*distal* orientation. This amounts to a buried surface area (ΔSASA) between hexamers of 3600 $Å^2$ for Hfq1 and 3000 $Å^2$ for Hfq2. *S. aureus* Hfq also crystallized in this *distal*-to-*distal* orientation, with a comparable ΔSASA of 2600 $Å^2$. This *distal*-to-*distal* interface in *Aae* Hfq2 is facilitated primarily through a network of side-chain and backbone interactions of Arg35 and Arg38 from each of the twelve monomers. These residues, located on the L2 loop and β2 strand, are in an analogous position to the distal-face residues that mediate the *proximal*-to-*distal* interface of Hfq1 dodecamers. Notably, the two Hfq2 hexamers that compose the dodecamer are not perfectly parallel, and there is a slight tilt of one of the hexameric rings. A similar feature, though not as severe, characterized the ring-ring packing geometry of *Aae* Hfq1.

3.3 Potential RNA-binding surface of Hfq2

We would expect to find regions of positive charge on the surface of *Aae* Hfq2, were this protein able to bind nucleic acid. When comparing the electrostatic potential of *Aae* Hfq1 and Hfq2, we see that the surface of *Aae* Hfq1 has a relatively positive charge distribution over the entire surface of the protein (Fig 4A), whereas the positive surface charge on Hfq2 is localized to the central pore of protein (Fig 4B). Unlike Hfq1, the surface of the lateral rim of Hfq2 lacks positive charge. The lateral rim of Hfq in *E. coli* has also been referred to as the 'arginine patch' and includes an 'RRER' motif, along the N-terminal α-helix, which has been shown to be important for sRNA-mRNA annealing [28,47,48].

Upon detailed examination of the structure, we find that the arginine residues of *Aae* Hfq2 are localized to the pore region, instead of the lateral rim (Fig 4C). The sequence of *Aae* Hfq1, which we previously co-crystallized with U_6 RNA at the lateral surface, includes six

arginine residues per monomer [29]. Five of these are localized to the lateral surface of the protein, with four of them further positioned within the arginine patch. Of the four arginine residues present in the sequence of *Aae* Hfq2, two are found on the distal surface and mediate dodecamer formation (as discussed above) and two are found on the proximal face. The lateral rim 'RRER' motif of *E. coli* Hfq ('RKKR' for *Aae* Hfq1) becomes 'KQQG' in *Aae* Hfq2.

Through structural alignment with *E. coli* Hfq, we have identified the analogous RNA-binding residues in Hfq1 and Hfq2. The proximal pocket of *Aae* Hfq2 is well-conserved, with the exception of R45, which is a glutamine in *E. coli* or leucine in *Aae* Hfq1 (Fig 5A). The residues that contribute to the distal site of *Aae* Hfq2 are relatively well-conserved, though several changes do make this face of the ring more polar and electropositive in Hfq2 than in Hfq1, including V28→N34, T51→H57, and N26→R32 substitutions. In the A site of *E. coli* Hfq [49], the adenosine is positioned between two glutamine residues which contact the base and provide specificity. These residues correspond to V37 and H57 in Hfq2. While the His residue could provide further π-stacking with an RNA base, some affinity may be lost due to the valine substitution.

Analysis of the available structures of RNA bound to the lateral surface of Hfq [46] suggest that the first two residues of the 'RRER' motif, along with Asn11 of the α-helix, and Ser36 and Phe37 of strand β2, form a pocket that binds specifically to two nucleotides of uridine (U_2). The rest of the arginine patch likely binds non-specifically to RNA (e.g. contacts the sugar-phosphate backbone), nucleating further interactions. Comparison of the residues involved in this 'U_2 pocket' with the analogous residues in *Aae* Hfq2 shows that while most of the residues are not strictly conserved, the substitutions do tend to preserve physicochemical properties (ex. Ser→Asp, Phe→His) (Fig 5C). We propose that RNA might bind to the lateral surface of Hfq2, though probably at a much lower affinity than for RNA-Hfq1 interactions. On

purely structural and enthalpic grounds, the lack of an arginine-rich patch would be expected to diminish binding of RNA at the lateral rim.

3.4 The proximal mode of binding is conserved in *Aae* Hfq2

To map the potential nucleic acid-binding sites on the surface of *Aae* Hfq2, we co-crystallized the protein with a suite of di- and triphosphate ribo- and deoxyribonucleotides. Several diffraction datasets were collected and phased via molecular replacement with the *apo* Hfq2 structure. Weighted mF_O-dF_C difference density maps were inspected after an initial round of refinement in order to identify the potential presence of nucleotides. While in most cases nucleotides were not found, in one case UTP could be identified as having co-crystallized with Hfq2. This Hfq2-UTP co-crystal was isomorphous with the *apo* Hfq2 form (in the same space group and with unit cell dimensions that agreed to within 0.5%), with the ASU consisting of two hexamers stacked in a *distal-to-distal* manner.

In this new *Aae* Hfq2 structure, twelve molecules of UTP were found in the proximal pore region of each Hfq2 subunit (Fig 6A). Only one to two phosphate groups were discernible in the difference electron density maps near each UTP molecule. This weak density is likely due to the disordered nature of the triphosphate tails, as was also observed by others with *P. aeruginosa* Hfq co-crystals [24]. Close examination of the UTP-binding site (Fig 6B) shows that it is analogous to the *proximal* site that has been previously described for several other Hfq homologs [23,50]. The uracil bases of each UTP molecule participate in alternating π-stacking interactions with Y46 residues of adjacent Hfq subunits. Further contacts occur between the O2' hydroxyl group and the side chains of Q12 and K61. The ribose O2' and O3' hydroxyls contact an imidazole nitrogen of H62 from the neighboring Hfq subunit, thereby imparting specificity for ribose versus deoxyribose.

In other Hfq homologs from gram-negative bacteria, a glutamine residue (Q41 in *E. coli*) often precedes the aromatic residue that stacks with the uracil bases (F42 in *E. coli*). In these Hfq-RNA co-crystal structures, this glutamine plays a role in uridine selectivity by hydrogen-bonding with O4 of the uracil. Intriguingly, this residue is L39 in Hfq1 and R45 in Hfq2. In the *apo* Hfq2 structure, this R45 residue extends into the pore of the protein, occluding the proximal pockets. Upon binding of UTP, the arginine sidechain presumably re-orients to project away from the proximal face. In our structure, the R45 side chains were usually not discernible in the electron density and are likely disordered and not participating in UTP-binding. However, in the hexamer, this array of six arginines may play a role in guiding the diffusional association of RNA via electrostatic steering [51], to the proximal pore of Hfq2.

3.5 *Aae* Hfq2-RNA binding depends on magnesium and pH

To quantify the affinity of Hfq2-RNA association in solution we utilized fluorescence polarization assays. This technique allows us to determine the apparent dissociation constants ($K_{d,app}$) for Hfq2 binding to short RNAs. Here, we used FAM-labelled U_6 and A_{18} RNAs as representative of the U-rich and A-rich binding motifs found to bind other gram-negative Hfq homologs, as well as FAM-A_6 and C_6 as representative of sequences not found to bind Hfq with high affinity. Under our initial conditions, *Aae* Hfq2 bound with only relatively low affinity to the RNAs tested (Table 3), with binding constants of 0.645 μM and 1.094 μM for U_6 and A_{18}, respectively. As magnesium has been shown to improve binding affinities for Hfq, and is well-known to do so for RNA-binding proteins more generally [48,50], a second suite of fluorescence polarization assays were conducted with the addition of 10 mM $MgCl_2$. Intriguingly, inclusion of Mg^{2+} increased the affinity of Hfq2 for U_6 by several orders of magnitude (to a K_d of 4.4 nM), but it completely abolished the binding of A_{18}.

As *Aae* Hfq2 is, overall, less positively charged than *Aae* Hfq1, and several of the potential RNA-binding pockets contain histidine residues, we performed additional RNA-binding assays at a solution pH of 6.0. Note that at this pH, ~50% of the histidine-imidazole side-chains will be positively charged. Under these conditions, the affinities of A-rich and C-rich RNAs for Hfq2 were greatly enhanced, with binding constants of 5.0, 5.6 and 2.9 nM for A_{18}, A_6 and C_6, respectively. While the U_6 affinity became undetectably low at this pH, the addition of 10 mM Mg^{2+} restored binding, with a K_d of 1.5 nM. Similarly, at low pH, inclusion of Mg^{2+} diminished the affinity for A_{18}, from 5.0 nM in the absence of Mg^2 to 36.7 nM.

3.6 At low pH, *Aae* Hfq2 co-purifies with nucleic acid

Because Hfq2 was found to bind RNAs with higher affinities at lower pH, we sought to co-purify recombinant Hfq2 with endogenous *E. coli* RNAs by altering our Hfq2 purification scheme. Specifically, by using a 2-(*N*-morpholino)ethanesulfonic acid (MES)-based buffer (pH 6.0) for cell lysis (and all subsequent steps), we were able to successfully co-purify the protein with nucleic acid, as assayed by an increased A_{260}/A_{280} ratio of our samples to ~1.5. To distinguish this nucleic acid as DNA or RNA (or both), we subjected the samples to a series of colorimetric assays [41]. While Hfq1 co-purified with RNA only, surprisingly, we found that Hfq2 co-purified with DNA (note that these assays cannot distinguish between a sample consisting of DNA and RNA, or DNA only). While previous studies have found that *E. coli* Hfq can bind to and alters the large-scale structure of DNA [52–54], to our knowledge this is the first example of another Hfq homolog binding DNA. We suggest that Hfq-DNA binding may be more widespread The *in vivo* implications for the interplay of Hfq and DNA are as yet unclear, though *E. coli* Hfq has been demonstrated to associate with the bacterial nucleoid [55].

3.7 Potential interactions of *Aae* Hfq1 and Hfq2

As heteroligomeric Sm rings are found in eukaryotes, this suggests that *Aae* Hfq1 and Hfq2 might interact as well, potentially forming heteromeric species. Several different experimental methods were utilized to test this interaction, from simple mixing (and subsequent analysis via SDS-PAGE and SEC), to immunoblotting (Fig S4) and co-precipitation, using 6x-His tagged and tagless constructs. We found no evidence for Hfq1-Hfq2 interaction in any of these different lines of experimentation, though we do note that our chromatographic data yielded inconclusive results, as both tagless Hfq constructs were found to interact with the nickel resins. In agreement with our results, preliminary work found no evidence for association between Hfq paralogs in the case of *B. anthracis* either [32]. More likely, *Aae* Hfq1 and Hfq2 serve two distinct roles in *A. aeolicus*. This hypothesis is consistent with our finding that Hfq2 has a nucleic acid binding-profile that differs from Hfq1, including dependence on pH and the ability to bind DNA. Perhaps Hfq1 and Hfq2 bind to different subsets of RNAs, that are involved in different regulatory pathways, for example acidic stress response for Hfq2 [56].

Our failure to detect an Hfq1-Hfq2 association is potentially interesting. One explanation could be that once the $(Hfq1)_6$ and $(Hfq2)_6$ rings have formed they are both thermodynamically and kinetically stable and do not dissociate, precluding any interchange of subunits. This is further supported by the observation that samples of His-tagged and tagless Hfq1 (or Hfq2) did not appear to self-associate via far-western dot blot (Fig S4). While Hfq oligomerizes simply by backbone hydrogen-bonding interactions between strands β4 and β5, large changes in the side chains at these positions could potentially affect this zipper-like interaction. A comparison of the β4 and β5 strands of Hfq1 and Hfq2 reveals several differences in the outermost residues of the strand-strand interface. Both Hfq1 and Hfq2 have a conserved glutamine residue at the start of

β4; in Hfq1, this residue sidechain (Q49) was found in alternate conformations interacting with neighboring residue T51 or with the backbone of β5. In Hfq2 this residue (Q55) forms an additional hydrogen bond with the β5 strand. In Hfq1, Intriguingly, there is also a proline residue in β5 of Hfq1 (P63) and β4 of Hfq2 (P56). These opposing prolines likely restrict the available conformations of the two strands and may thereby prevent them from interacting.

4. Conclusions

The work presented here, in conjunction with our previous report of *Aae* Hfq1, represents the first case of two structurally verified Hfq paralogs within the species. Together these results demonstrate that a bacterial species can in fact encode multiple proteins which adopt the characteristic Sm fold of Hfq. This has intriguing evolutionary implications for the many duplication events that occurred with the eukaryotic Sm proteins. Typically these gene duplications are only observed after the branching of eukaryota [57], so it is unusual to see multiple paralogs, especially in a potentially early-branching bacterium such as *A. aeolicus*. These finding also support the gene duplication and drift model of Sm superfamily evolution, whereby early Sm and LSm paralogs diverged greatly in sequence and functionality and were later recruited into the heteroheptameric spliceosomal core [58].

The structure of *Aae* Hfq2 reveals that its fold is highly similar to Hfq1 (at the levels of the monomeric subunit, the assemblies of the hexameric rings, and the formation of higher-order dodecamers, as observed in the crystal structures and in solution). Moreover, through co-crystallization with UTP, we have demonstrated that the proximal mode of RNA-binding is conserved in an Hfq2 paralog. However, we also find that small changes in sequence have lead to large differences in the overall RNA-binding behaviors of Hfq1 and Hfq2, particularly for A-rich RNAs, which would be expected to bind at the distal face. *Aae* Hfq2 binds

to A-rich and C-rich RNAs only at the more acidic pH 6.0, and while Mg^{2+} improves affininities for U-rich RNA, it abolishes binding of A-rich sequences. We also note that the lateral RNA-binding site, which is conserved in *Aae* Hfq1, is absent in Hfq2. Instead the majority of Hfq2 arginine residues are localized to the pore of the hexameric ring.

The physiological role of *Aae* Hfq2 is still unclear, however given our observations it is most likely that Hfq2 performs some unique, distinct function in the cell. We were unable to detect any interaction between Hfq1 and Hfq2 and moreover, we found that Hfq2, but not Hfq1, co-purifies with endogenous DNA, suggesting a separate functional role for Hfq2 only. These results are in agreement with findings for *B. anthracis* and *B. cenocepacia* Hfqs [31–33], as neither set of paralogs was found to interact. *B. cenocepacia* Hfq1 and Hfq2 were maximally expressed during different growth phases, suggesting the two proteins could function similarly, but bind a different population of RNAs. *Aae* Hfq1 and Hfq2 could also be expressed differentially, and perhaps participate in different pathways. For example, Hfq2 could be involved in low-pH stress-response pathways. The *Aae* Hfq2 structure reported here provides the molecular-level details as to how such different functionalities of two Hfq paralogs from a single bacterial species is achieved.

Acknowledgements

We thank H. Huber (Regensburg) for providing a sample of *A. aeolicus* genomic material, J. Bushweller (UVa) for access to the fluorescent plate reader, J. Shannon (UVa) for assistance with MALDI-TOF, D. Cascio and M. Sawaya (UCLA) for crystallographic advice, and L. Columbus, C. McAnany, J. Patterson, and P. Randolph (UVa) for helpful discussions. Beamlines SER-CAT 22-ID and NE-CAT 24-ID-C/E at Argonne National Laboratory's Advanced Photon Source are DOE facilities (DE-AC02-06CH11357).

Funding information

Funding for this research was provided by: National Science Foundation, Division of Molecular and Cellular Biosciences (award No. 1350957); Thomas F. and Kate Miller Jeffress Memorial Trust (award No. J-971).

References

1. Franze de Fernandez MT, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. Nature. 1968;219: 588–590.

2. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. Cell. 2004;118: 69–82.

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

4. Guisbert E, Rhodius VA, Ahuja N, Witkin E, Gross CA. Hfq Modulates the σE-Mediated Envelope Stress Response and the σ32-Mediated Cytoplasmic Stress Response in Escherichia coli. J Bacteriol. 2007;189: 1963–1973.

5. Gottesman S, McCullen CA, Guillier M, Vanderpool CK, Majdalani N, Benhammou J, et al. Small RNA regulators and the bacterial response to stress. Cold Spring Harb Symp Quant Biol. 2006;71: 1–11.

6. Sittka A, Pfeiffer V, Tedin K, Vogel J. The RNA chaperone Hfq is essential for the virulence of Salmonella typhimurium. Mol Microbiol. 2007;63: 193–217.

7. Chao Y, Vogel J. The role of Hfq in bacterial pathogens. Curr Opin Microbiol. 2010;13: 24–33.

8. Sauer E. Structure and RNA-binding properties of the bacterial LSm protein Hfq. RNA Biol. 2013;10: 610–618.

9. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nat Rev Microbiol. 2011;9: 578–589.

10. Updegrove TB, Zhang A, Storz G. Hfq: the flexible RNA matchmaker. Curr Opin Microbiol. 2016;30: 133–138.

11. Soper T, Mandin P, Majdalani N, Gottesman S, Woodson SA. Positive regulation by small RNAs and the role of Hfq. Proc Natl Acad Sci U S A. 2010;107: 9602–9607.

12. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. Proc Natl Acad Sci U S A. 2002;99: 4620–4625.

13. De Lay N, Schu DJ, Gottesman S. Bacterial small RNA-based negative regulation: Hfq and its accomplices. J Biol Chem. 2013;288: 7996–8003.

14. Gottesman S, Storz G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003798

15. Møller T, Franch T, Højrup P, Keene DR, Bächinger HP, Brennan RG, et al. Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. Mol Cell. 2002;9: 23–30.

16. Sun X, Zhulin I, Wartell RM. Predicted structure and phyletic distribution of the RNA-binding

protein Hfq. Nucleic Acids Res. 2002;30: 3662–3671.

17. 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. 2002;320: 705–712.

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

19. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003707

20. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. Int Rev Cell Mol Biol. 2009;272: 149–189.

21. Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA, et al. Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. Cell. 1999;96: 375–387.

22. Fischer U, Englbrecht C, Chari A. Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdiscip Rev RNA. 2011;2: 718–731.

23. 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. John Wiley & Sons, Ltd; 2002;21: 3546–3556.

24. Murina V, Lekontseva N, Nikulin A. Hfq binds ribonucleotides in three different RNA-binding sites. Acta Crystallogr D Biol Crystallogr. 2013;69: 1504–1513.

25. 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. J Biol Chem. 2003;278: 1239–1247.

26. 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. 2002;320: 129–142.

27. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D. Structure and assembly of an augmented Sm-like archaeal protein 14-mer. Proc Natl Acad Sci U S A. National Academy of Sciences; 2003;100: 4539–4544.

28. Panja S, Schu DJ, Woodson SA. Conserved arginines on the rim of Hfq catalyze base pair formation and exchange. Nucleic Acids Res. Oxford University Press; 2013;41: 7536–7546.

29. Stanek KA, Patterson-West J, Randolph PS, Mura C. Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: conservation of the lateral RNA-binding mode. Acta Crystallogr D Struct Biol. 2017;73: 294–315.

30. Ramos CG, Sousa SA, 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. 2011;193: 1515–1526.

31. Ramos CG, da Costa PJP, Döring G, Leitão JH. The novel cis-encoded small RNA h2cR is a negative regulator of hfq2 in Burkholderia cenocepacia. PLoS One. 2012;7: e47896.

32. Vrentas C, Ghirlando R, Keefer A, Hu Z, Tomczak A, Gittis AG, et al. Hfqs in Bacillus anthracis: Role of protein sequence variation in the structure and function of proteins in the Hfq family. Protein Sci. 2015;24: 1808–1819.

33. 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. 2015;1850: 1661–1668.

34. Kabsch W. XDS. Acta Crystallogr D Biol Crystallogr. 2010;66: 125–132.

35. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66: 213–221.

36. 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 U S A. 2006;103: 8060–8065.

37. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66: 486–501.

38. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66: 12–21.

39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

40. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28: 1647–1649.

41. Patterson J, Mura C. Rapid colorimetric assays to qualitatively distinguish RNA and DNA in biomolecular samples. J Vis Exp. 2013; e50225.

42. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5: e1000605.

43. Vecerek B, Rajkowitsch L, Sonnleitner E, Schroeder R, Bläsi U. The C-terminal domain of Escherichia coli Hfq is required for regulation. Nucleic Acids Res. 2008;36: 133–143.

44. Santiago-Frangos A, Jeliazkov JR, Gray JJ, Woodson SA. Acidic C-terminal domains autoregulate the RNA chaperone Hfq. eLife Sciences. eLife Sciences Publications Limited; 2017;6: e27049.

45. Beich-Frandsen M, Vecerek B, Konarev PV, Sjöblom B, Kloiber K, Hämmerle H, et al. Structural insights into the dynamics and function of the C-terminus of the E. coli RNA

chaperone Hfq. Nucleic Acids Res. 2011;39: 4900–4915.

46. Stanek KA, Mura C. Producing Hfq/Sm Proteins and sRNAs for Structural and Biophysical Studies of Ribonucleoprotein Assembly. In: Arluison V, Valverde C, editors. Bacterial Regulatory RNA: Methods and Protocols. New York, NY: Springer New York; 2018. pp. 273–299.

47. Zheng A, Panja S, Woodson SA. Arginine Patch Predicts the RNA Annealing Activity of Hfq from Gram-Negative and Gram-Positive Bacteria. J Mol Biol. 2016;428: 2259–2264.

48. 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 U S A. 2012;109: 9396–9401.

49. Link TM, Valentin-Hansen P, Brennan RG. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. Proc Natl Acad Sci U S A. 2009;106: 19292–19297.

50. 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. 2014;20: 1548–1559.

51. Wade RC, Gabdoulline RR, Lüdemann SK, Lounnas V. Electrostatic steering and ionic tethering in enzyme–ligand binding: Insights from simulations. Proc Natl Acad Sci U S A. National Academy of Sciences; 1998;95: 5942–5949.

52. Takada A, Wachi M, Kaidow A, Takamura M, Nagai K. DNA binding properties of the hfq gene product of Escherichia coli. Biochem Biophys Res Commun. 1997;236: 576–579.

53. Updegrove TB, Correia JJ, Galletto R, Bujalowski W, Wartell RM. E. coli DNA associated with isolated Hfq interacts with Hfq's distal surface and C-terminal domain. Biochim Biophys Acta. 2010;1799: 588–596.

54. Jiang K, Zhang C, Guttula D, Liu F, van Kan JA, Lavelle C, et al. Effects of Hfq on the conformation and compaction of DNA. Nucleic Acids Res. 2015;43: 4332–4341.

55. Azam TA, Ishihama A. Twelve Species of the Nucleoid-associated Protein from Escherichia coli : SEQUENCE RECOGNITION SPECIFICITY AND DNA BINDING AFFINITY. J Biol Chem. 1999;274: 33105–33113.

56. Yang G, Wang L, Wang Y, Li P, Zhu J, Qiu S, et al. hfq regulates acid tolerance and virulence by responding to acid stress in Shigella flexneri. Res Microbiol. 2015;166: 476–485.

57. Aravind L, Iyer LM, Koonin EV. Comparative genomics and structural biology of the molecular innovations of eukaryotes. Curr Opin Struct Biol. 2006;16: 409–419.

58. 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. Public Library of Science; 2009;5: e1000315.

Tables

Table 1 X-ray data collection statistics

	Aae Hfq2	*Aae* Hfq2-UTP
Beamline	APS 22-ID	APS 24-ID-C
Wavelength (Å)	1.0000	0.979
Temperature (K)	100	100
Detector	MARMOSAIC 300	Dectris PILATUS 6MF
Crystal-to-detector distance (mm)	225	300
Rotation range per image (°)	0.5	0.5
Total rotation range (°)	180	90
Exposure time per image (s)	0.5	0.5
Space group	$P2_1$	$P2_1$
a, b, c (Å)	60.47, 101.06, 65.66	60.36, 101.19, 67.06
α, β, γ (°)	90, 107.8, 90	90, 106.2, 90
Mosaicity (°)	0.166	0.102
Resolution range (Å)	62.56-1.98 (2.05-1.98)	64.43-1.85 (1.91-1.85)
Total no. of reflections	184,994 (13,455)	313,307 (30,709)
No. of unique reflections	50,903 (3,861)	65,501 (6,388)
Completeness (%)	97.0 (74.0)	99.0 (96.8)
Multiplicity	3.6 (3.5)	4.8 (4.8)
I/σ(I)	11.0 (3.5)	11.5 (2.0)
R_{merge}	0.067 (0.332)	0.055 (0.555)
R_{meas}	0.078 (0.391)	0.062 (0.619)
$R_{p.i.m.}$	0.041 (0.204)	0.027 (0.275)
$CC_{1/2}$	0.997 (0.956)	0.998 (0.933)
Overall *B* value from Wilson plot (Å^2)	23.83	20.77
Matthews coefficient V_M (Å^3 Da^{-1})	1.91	1.97

Table 2 Structure determination and refinement

	Aae Hfq2	*Aae* Hfq2-UTP
Resolution range (Å)	57.59-2.00 (2.07-2.00)	57.98-1.85 (1.97-1.85)
Completeness (%)	99.16 (96.51)	73.81 (19.96)
No. of reflections, working set	50,693 (4,890)	48,652 (1,316)
No. of reflections, test set	2,535 (244)	2,459 (68)
Final R_{work}	0.210 (0.243)	0.223 (0.237)
Final R_{free}	0.252 (0.315)	0.273 (0.250)
No. of non-H atoms		
Macromolecules	5,987	5,888
Ligands	0	189
Solvent	51	68
Total	6,038	6,145
No. of protein residues	747	745
R.M.S.D.		
Bonds (Å)	0.007	0.008
Angles (°)	0.97	1.09
Average *B* factors (Å^2)		
Protein	33.29	24.81
Ligand		36.64
Solvent	27.94	21.14
Ramachandran Plot		
Most favored (%)	96.54	96.95
Allowed (%)	3.32	2.91
Outliers (%)	0.14	0.14
Rotamer outliers (%)	1.21	3.45

Table 3 Hfq2 affinities for binding short RNAs quantified via fluorescence polarization

K_D (nM)	U_6		A_{18}		A_6	C_6
	$-Mg^{2+}$	$+Mg^{2+}$	$-Mg^{2+}$	$+Mg^{2+}$	$-Mg^{2+}$	$-Mg^{2+}$
pH 8	645	4.4	1094	—	—	—
pH 6	—	1.5	5.0	36.7	5.6	2.9

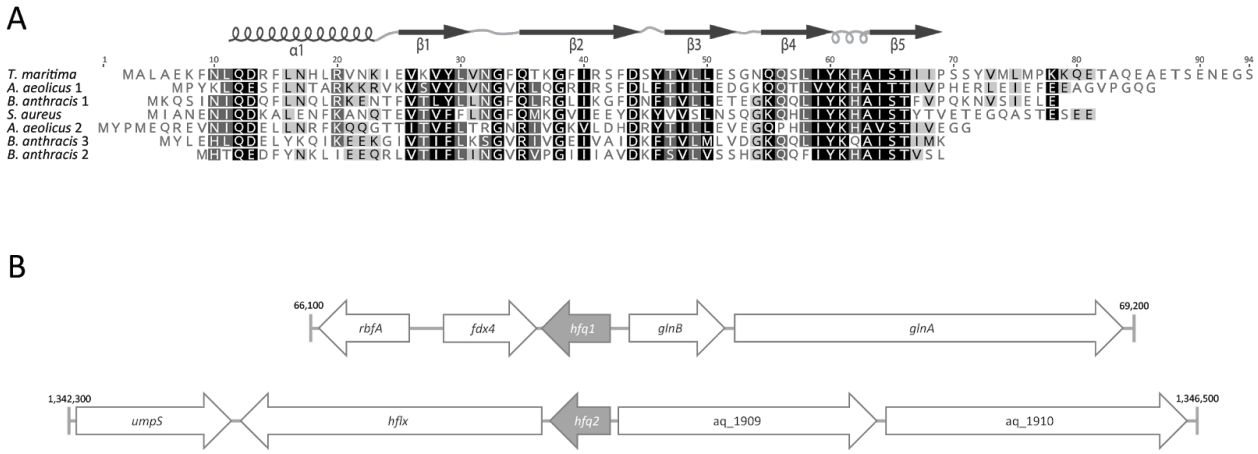


Figure 1. Comparison of *Aae* Hfq1 and Hfq2. (A) Sequence alignment of *Aae* Hfq1 (NP_213072.1) and Hfq2 (YP_008920737.1) with Hfqs from deep-branching *Thermotoga maritima* (Q9WYZ6.1), gram-positive *Staphylococcus aureus* (OYP81766.1), and *Bacillus anthracis* Hfq1 (AAT32953.1), Hfq2 (AAT30766.1), and plasmid-encoded Hfq3 (AAT35495.1). Sequences were aligned in MUSCLE and the Sm domain secondary structural elements are shown above as a cartoon. Residues are colored according similarity to the consensus: black boxes with white text, 100% similar; dark grey box with with white text, 80-100% similar; light grey box with black text, 60-80% similar; white box with grey text, <60% similar. (B) Gene neighbors of *Aae* Hfq1 and Hfq2 are shown as white arrows and genomic position are given in base pairs. Genes encode for the following protein products: *rbfA*, ribosome binding factor A; *fdx4*, ferredoxin; *glnB*, nitrogen regulatory PII protein; *glnA*, glutamine synthetase; *umpS*, uridine-5'-monophosphate synthetase; *hflx*, GTP-binding protein. Note that aq_1909 and aq_1910 are hypothetical proteins with no BLAST hits.

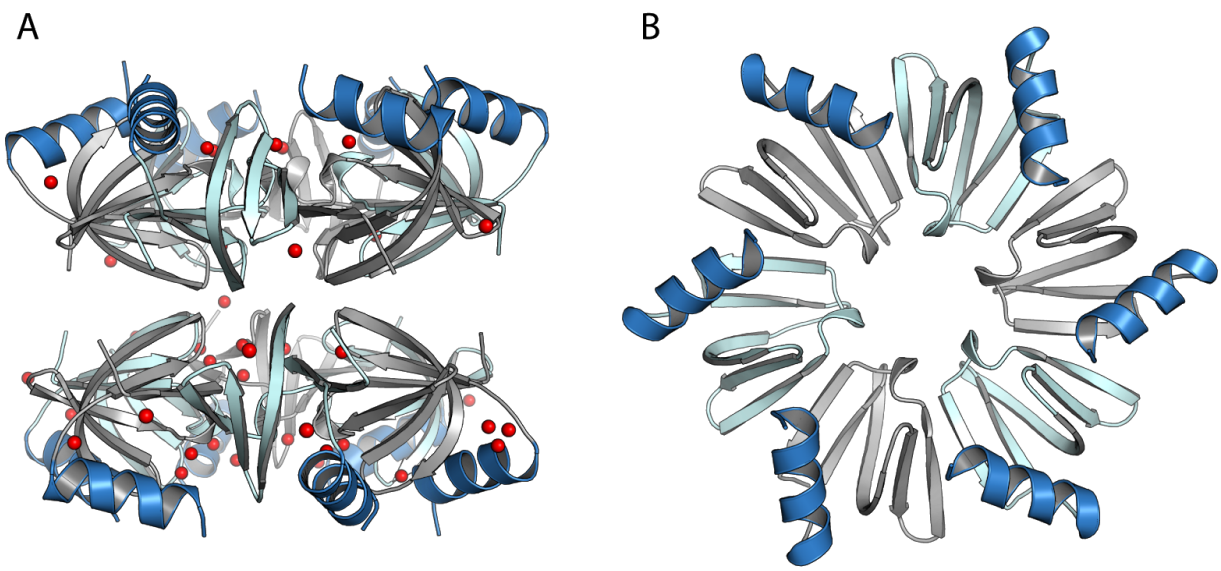


Figure 2. *Aae* Hfq2 crystallized as a dodecamer of two hexamers in a *distal-to-distal* orientation. (A) A cartoon representation of the crystallographic ASU. Hfq2 monomers are alternately colored in grey and cyan, with the N-terminal α-helical regions colored blue. Waters are shown as red spheres. Note the that the planes of the two hexamers do not lie perfectly parallel to one another. (B) A single Hfq2 hexamer, with the same coloring and shown rotated 90° relative to (A), exhibits the conserved Sm fold: five highly-bent antiparallel β-sheets preceded by an N-terminal α-helix. Note that the view in (B) is onto the proximal face.

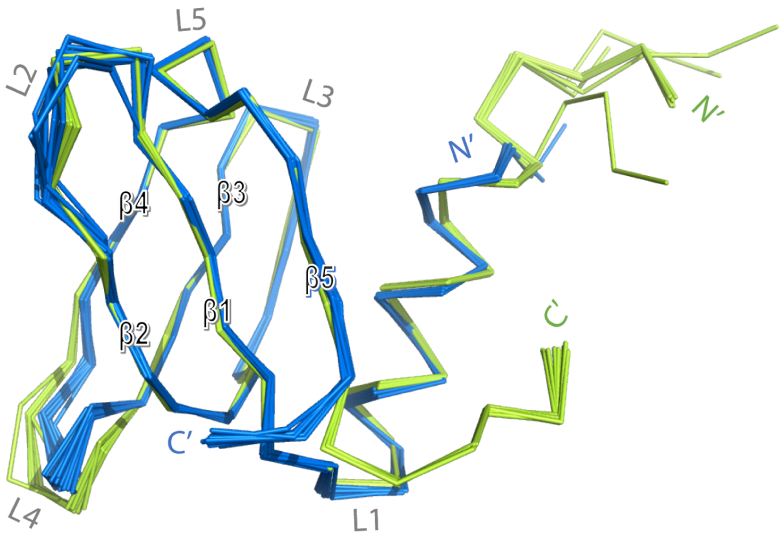


Figure 3. Multiple structural alignment of the twelve unique Hfq1 (green) and Hfq2 (blue) monomeric subunits from the crystallographic ASUs, shown as Cα-ribbon traces. The termini, loops, and β-strands are labelled and numbered. The Hfq1 and Hfq2 monomeric folds are highly conserved, with the greatest variances in the L2 and L4 loop regions, the termini, and the angling of the N-terminal region of the α-helices.

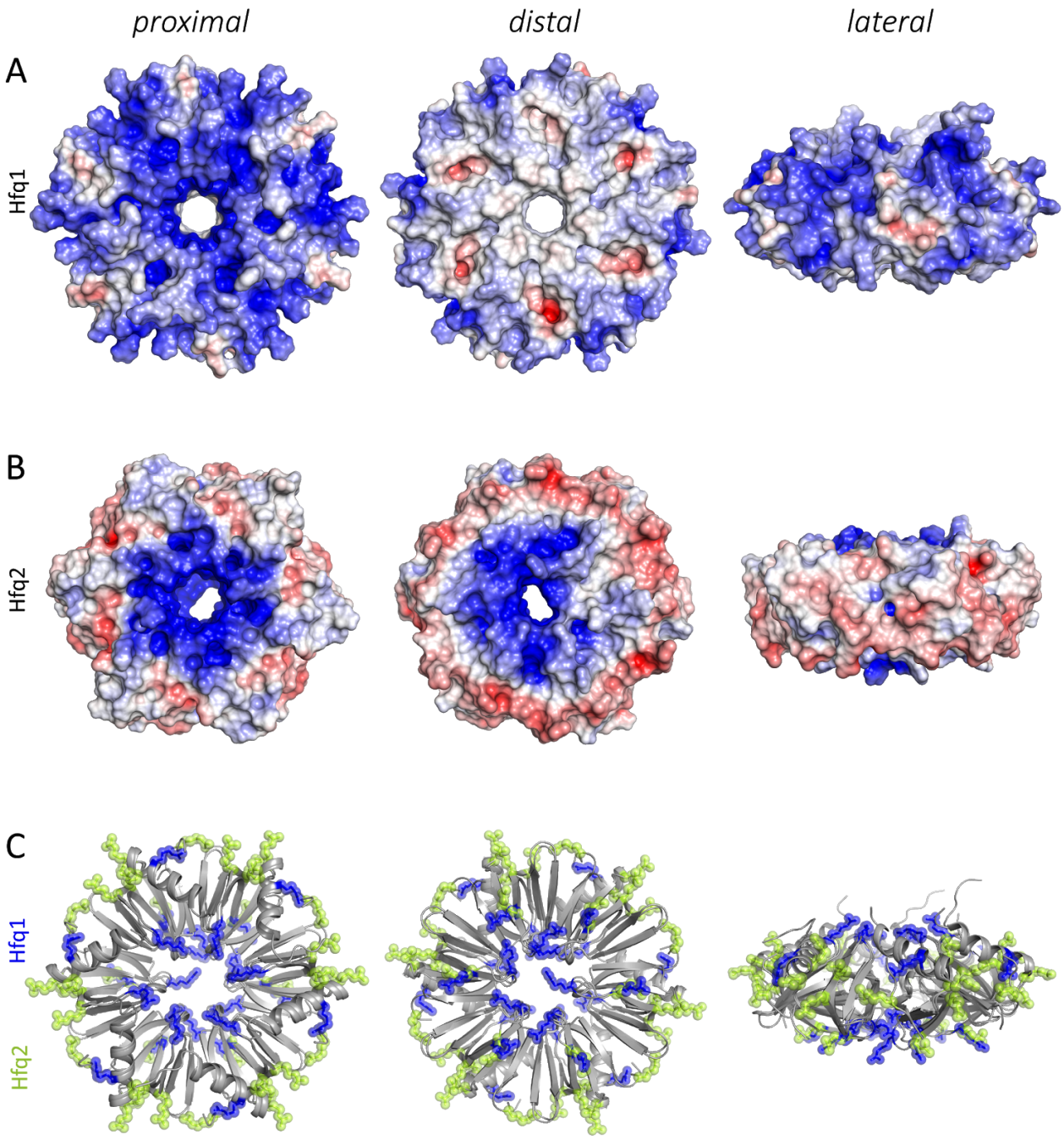
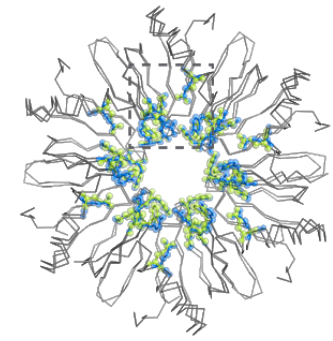
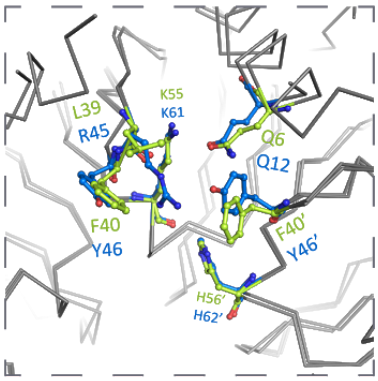


Figure 4. The surface distribution of charge varies between Hfq1 and Hfq2. (A) The Hfq1 hexamer is shown as a surface representation and colored according to the distribution of electrostatic potential, graded from ± 5 k_BT/e (blue is positive and red is negative). This potential was calculated using the APBS tools plug-in in PyMOL. The proximal and lateral surfaces of

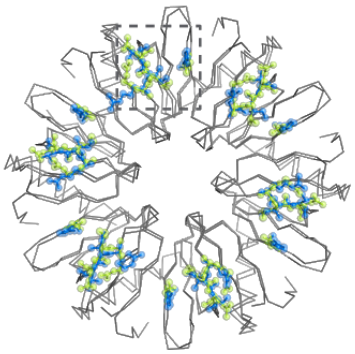
Hfq1 are highly basic, while the distal surface exhibits some apolar patches. (B) The same representation as (A), but for Hfq2. Note that the most electropositive region of Hfq2 is far more constricted to the pore of the hexamer. (C) Structural alignment of the Hfq1 and Hfq2 hexamers, shown as grey cartoon representations. Arginine residues are shown as spheres and colored green (Hfq1) or blue (Hfq2). Note that the Hfq1 arginines are concentrated around the lateral rim of the hexamer, whereas the Hfq2 arginines are found mainly near the pore.

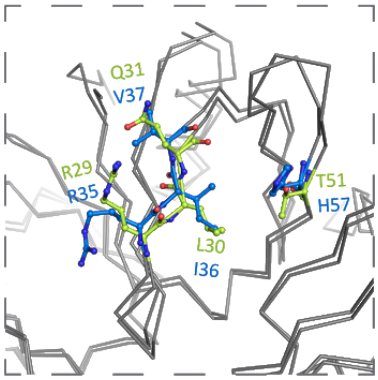
A *proximal site*



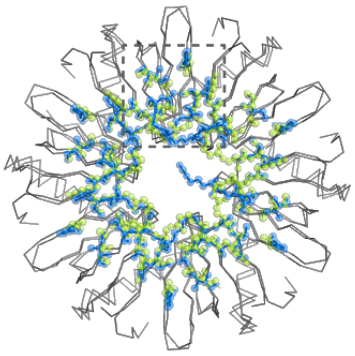


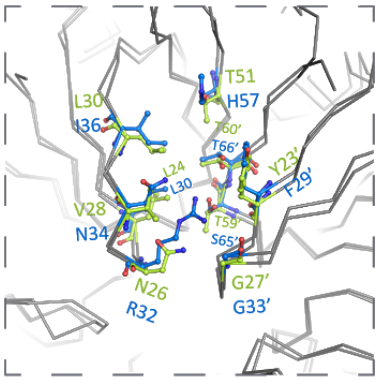
B *distal 'A' site*



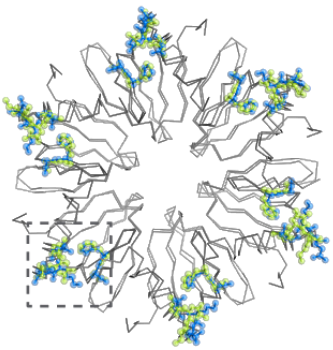


C *distal 'R' site*





D *lateral site*



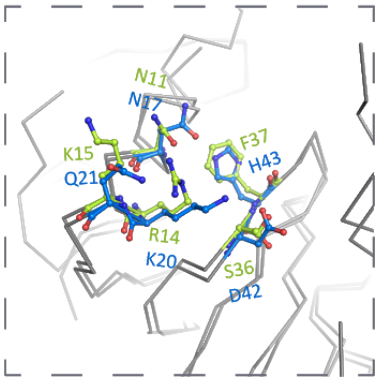


Figure 5. Structural comparison of the proximal, distal, and lateral RNA-binding sites of *Aae* Hfq1 and Hfq2. Hfq1 and Hfq2 hexamers were structurally aligned and are shown as grey ribbons. Residues identified through comparison with *E. coli* Hfq as being involved in binding RNA at the (A) proximal, (B,C) distal (the A and R sites correspond to the 'ARN' tripartite motif) and (D) lateral site are labelled and shown as stick representations. Hfq1 residues are colored green and Hfq2 residues blue, with residues from an adjacent subunit denoted with an apostrophe. The left panels show where these residues are located across the surface of the hexamer, while the right panels give a magnified view of a single site. Overall, the residues identified as potentially being involved in RNA-binding are more highly conserved in Hfq1 than in Hfq2.

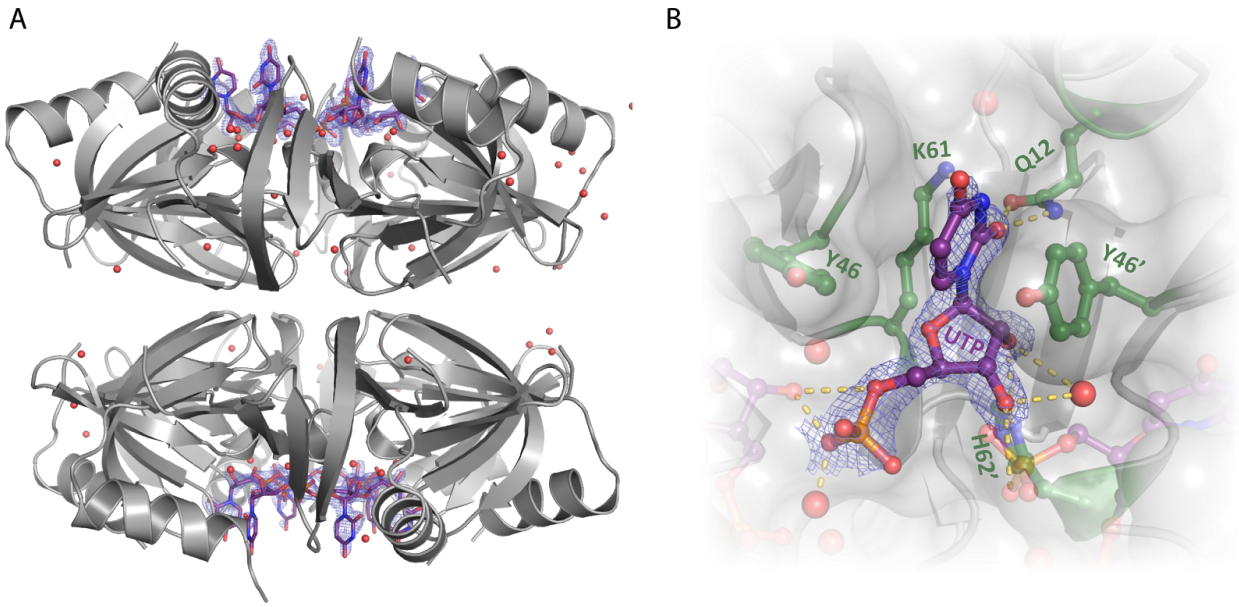


Figure 6. Hfq2 co-crystallized with UTP in the proximal binding pocket. (A) The asymmetric unit of the $P2_1$ form contains two Hfq2 hexamers, shown as grey-colored cartoons. Molecules of UTP, shown in purple ball-and stick representation, were found at each of the 12 proximal sites within the ASU. The meshes delimit the $2mF_o - DF_c$ electron-density map, contoured at 1.5σ. (B) This view of a single proximal site, with same color scheme as (A), shows the RNA-contacting residues (shown in green ball-and-stick representation) in greater detail; an apostrophe denotes residues from the neighboring subunit. The yellow dashed lines (hydrogen bonds) indicate enthalpically favorable Hfq-RNA contacts.

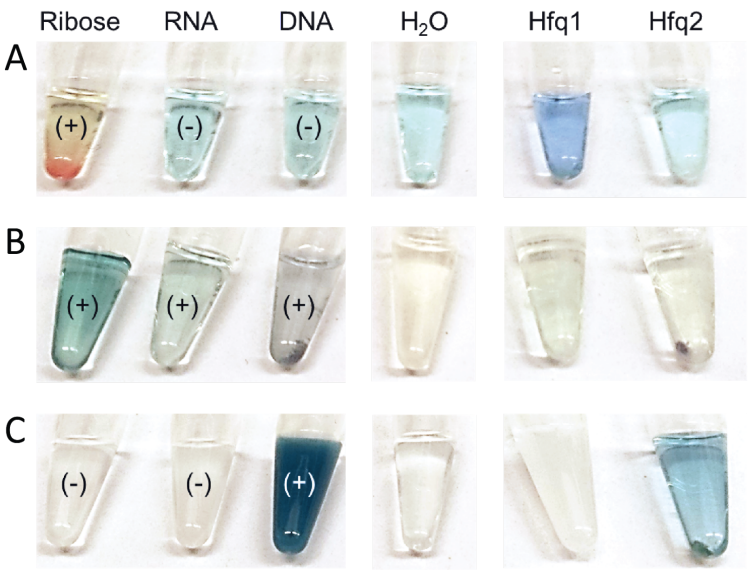


Figure 7. Nucleic acids that co-purify with Hfq1 and Hfq2 were isolated through purification of recombinantly-expressed protein at pH 8.0 (Hfq1) or pH 6.0 (Hfq2). Colorimetric assays were used to identify the binding partners as DNA and RNA. (A) The Benedict's assay tests for free reducing sugars, with a positive resulting producing a red end-product, (B) the Bial's assay detects the presence of any pentose sugar, resulting in a blue-green end-product and (C) the Dische assay tests specifically for deoxyribose, yielding a blue product. Ribose, and RNA and DNA standards are shown as positive (+) or negative (-) controls of each assay. Hfq2 can be seen to have co-purified with DNA, whereas Hfq1 co-purifies with RNA only.

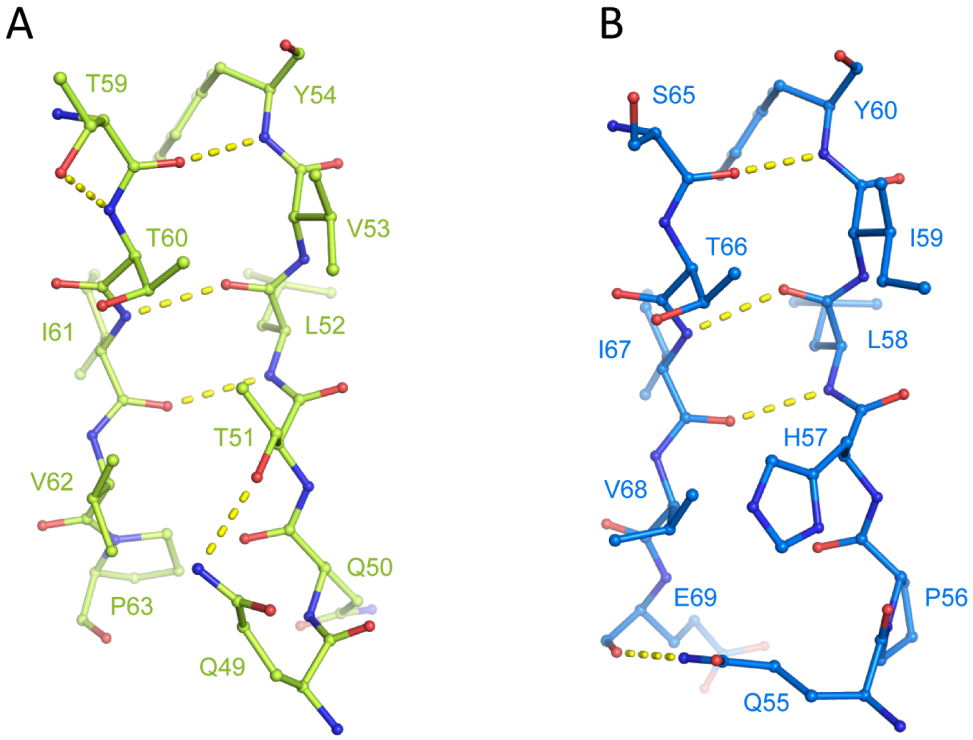


Figure 8. Comparison of the detailed structures of the β4-β5 interfaces of Hfq1 and Hfq2. The β4 and β5 strands of two adjacent monomers within the hexameric ring are shown as stick representation for (a) Hfq1 and (b) Hfq2. Polar contacts, within a cutoff distance of 3.6 Å are shown as dashed yellow lines. Note the alternate positioning of the prolines, present on the β5 strand of Hfq1 versus β4 of Hfq2.

Chapter 5: Crystallization and initial structural characterization of a tandem-domain Hfq homolog from *Novosphingobium aromaticivorans*

Kimberly Stanek, Sebastian Coupe, and Cameron Mura

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

Abstract

The Sm superfamily of RNA-binding proteins, found in all three domains of life, display immense oligomeric and functional plasticity. The bacterial Sm protein, known as Hfq, chaperones the interactions of small regulatory RNA (sRNA) and mRNA, and is involved in numerous cellular processes, including stress response, quorum sensing, sugar metabolism, and virulence. The biologically-functional Hfq oligomer is a hexameric toroid, with two distinct faces that can simultaneously bind RNA. Based on sequences analyses, the α-proteobacterium *Novosphingobium aromaticivorans* has tandemly-repeated Sm domains (Hfq_N-Hfq_C) within its putative *hfq* open reading frame, implying differences in the quaternary structure and potentially in RNA-binding properties. Here *Nar* Hfq has been recombinantly cloned, expressed, purified and biophysically characterized. RNA-binding assays indicate behavior consistent with other (more conventional) bacterial Hfq homologs, and show high affinity binding for A/U-rich ssRNA. The major oligomeric state of *Nar* Hfq *in vitro* is a trimer, suggesting a ring of alternating Hfq_N and Hfq_C domains. Ultimately, we wish to obtain high-resolution structural information on *Nar* Hfq. However, initial crystallization efforts have been hindered by the presence of an N-terminal

proline-rich tail. Ongoing and future work is aimed at crystallizing a ΔN-terminal construct, wherein 38 residues are truncated at the N-terminus.

1. Introduction

The Sm superfamily comprises an ancient family of proteins found in all domains of life, with roles ranging from post-transcriptional regulation to RNA splicing and other processing pathways [1]. The Sm fold consists of an N-terminal α-helix followed by five highly-curved antiparallel β-strands which form a small β-barrel (or SBB, *Youkharibache et al., in revision*). While the Sm fold is only ~60-100 residues in length, the protein is typically found as an oligomeric ring in biochemical, biophysical, and structural characterization, and is thought to act as an oligomeric ring *in vivo*. Ring assembly is facilitated via interactions between β4 and β5 strands of adjacent monomers. Sm rings display immense oligomeric plasticity, ranging from pentamers to 14-mers, and provide an enhanced platform for binding single-stranded RNA [2–5].

The bacterial Sm protein was initially identified as a host factor required for the replication of bacteriophage Qβ, and is referred to as 'Hfq' for this reason [6]. Hfq has been shown to function generally in post-transcriptional regulation by chaperoning the interactions of small regulatory RNAs (sRNAs) and their mRNA targets [7–9], and it has been linked to numerous pathways including quorum sensing [10], stress response [11,12], iron metabolism [13], and expression of virulence factors [14]. Typically, species of bacteria encode one *hfq* gene, although several species have been identified as having two or three copies [15–17]. These Hfq paralogs are also detailed further in Chapter 2 and 4 of this work. Hfq oligomerizes as a homohexameric ring with two distinct faces for binding RNA, termed 'proximal' and 'distal'

(with respect to the α-helix) [2,18]. In general, U-rich regions of sRNA bind to the proximal face while an A-rich region of a target mRNA simultaneously binds to the distal face. More recently, an additional arginine-rich patch (with an 'RRER' signature motif in *E. coli*) has been shown to additionally bind U-rich RNA and facilitate annealing of the sRNA-mRNA duplex [19–22].

The eukaryotic Sm proteins are most widely known for their scaffolding role in the assembly of snRNPS and the higher-order spliceosomal complex, and have also been found to function in other mRNA regulatory pathways [23,24]. Eukaryotes have numerous (>20) Sm and like-Sm (LSm) paralogs that oligomerize as heteroheptameric rings through a complex, chaperoned biogenesis pathway [25]. The subunit composition of each Sm ring determines the functional role, and perhaps cellular localization, of the complex. For example, a ring composed of LSm paralogs 2-8 localizes to the nucleus where it forms the U6 spliceosomal core, whereas a ring comprised of LSm1-7 will localize to the cytoplasm, functioning instead in mRNA degradation, as part of P-bodies [26–28]. The structural mode of RNA-binding to Sm rings also differs from that of Hfq. Due to increased pore size of the heptameric Sm oligomers, the RNA is able to thread through the pore, and interacts only with the proximal face in a manner similar to Hfq [5,29,30].

A major open question in Sm biology is how the evolutionary transition occurred from homohexameric Hfq to heteroheptameric Sm complexes. Some information can be gleaned by considering the Sm-like archaeal proteins (SmAPs) [1]. Archaeons encode anywhere from one to three SmAP paralogs; while these copies don't appear to associate, different SmAPs are capable of spontaneously self-oligomerizing as hexamers [3], heptamers [3,4,31], and even octamers (*Randolph et al.*, unpublished). SmAPs have also been co-crystallized with RNA in binding pockets resembling the proximal and lateral sites of Hfq [32,33]. However, little is known

about the physiological roles of SmAPs (and archaeal RNA processing in general), and archaea lack the sophisticated spliceosomal machinery found in eukaryotes.

Further evolutionary insight, regarding expansion and genomic organization of the Sm family of proteins, can be ascertained from certain bacterial lineages. As previously mentioned, several species of Hfq encode multiple paralogs, indicative of early gene duplication events. *In vitro* biochemical characterization of these copies shows that they likely do not interact, and instead have distinct physiological functions. For example, *Burkholderia cenocepacia* Hfq1 and Hfq2 are maximally expressed during different phases of growth, and could therefore interact differentially with growth-phase specific RNAs [34]. We have also demonstrated the *Aquifex aeolicus* Hfq1 and Hfq2 bind RNA with different affinities *in vitro*, and that Hfq2 co-purifies with DNA as well as RNA when recombinantly expressed and purified in *E. coli* (Chapter 4).

Species from the α-proteobacterial order Sphingomonadales have been identified as having an extremely long Hfq open reading frame, including the possibility of two tandemly-repeated Hfq domains [35]. If these Hfq homologs were to assemble similarly to other characterized Hfq proteins, then they could represent the first “pseudo-heteromeric” Hfqs, making them intriguing candidates of study from an Sm evolutionary perspective. Here, we have successfully cloned, over-expressed, purified, and characterized the tandem domain Hfq from *Novosphingobium aromaticivorans*, a soil-dwelling member of the Sphingomonadales. Notably this bacterium has more recently been linked to cases of primary biliary cirrhosis [36]. We have discovered that *N. aromaticivorans* (*Nar*) Hfq forms a trimer in solution, and binds U-rich and A-rich sequences with high affinity. We have also crystallized *Nar* Hfq; our initial crystals diffracted X-rays poorly, likely due to an N-terminal proline-rich tail (which may exhibit a great degree of conformational heterogeneity). Ongoing efforts are aimed at crystallizing a ΔN-terminal *Nar* Hfq construct , in the hopes of obtaining better diffracting crystals.

2. Materials and Methods

2.1 Cloning, expression and purification of recombinant *Nar* Hfq

The genes for full-length and ΔN-*Nar* Hfq were first cloned from the *N. aromaticivorans* DSM 12444 genome into the pCR-BLUNT vector using the Zero BLUNT PCR Cloning Kit (Invitrogen). The inserts were then introduced into the T7-promoter based pET28b(+) plasmid through double restriction enzyme digestion of insert and empty pET28b(+) vector, followed by ligation of both products using T4 DNA ligase. The pET28b(+) plasmid introduces an *N*-terminal His-6x tag followed by a thrombin-cleavable linker. Plasmids were amplified by transformation of chemically competent TOP10 *E. coli* cells and purified using a miniprep kit (Qiagen).

To over-express recombinant, full-length *Nar* Hfq, chemically competent BL21(DE3) *E. coli* cells were transformed with the *Nar* Hfq-pET28b(+) plasmid and plated on LB-agar supplemented with 0.05 mg/ml kanamycin, followed by outgrowth in Lysogeny broth (LB) media overnight with shaking (225 rpm) at 37 °C. *Nar* Hfq expression was induced by adding 1 mM isopropyl-β-D-thiogalactoside (IPTG) when the cell cultures reached an optical density (OD_{600}) of 0.8-1.0. Cells were incubated for an additional 3-4 hours and then pelleted at 15,000*g* for 5 min at 5 °C and stored at -20 °C overnight. Cell pellets were resuspended in a partial lysis buffer (50 mM Tris pH 7.5, 750 mM NaCl) and a cOmplete protease inhibitor cocktail tablet (Roche) was added to prevent degradation. Cells were chemically lysed with the addition of 0.01 mg/ml chicken egg white lysozyme (Fisher) followed by incubation at 37 °C for 20 min. To ensure complete lysis, cells were then mechanically lysed using a microfluidizer. The cell lysate was then clarified by centrifugation at 35,000*g* for 20 min at RT. A heat-cut step was performed by

incubating the lysate for 20 min at 80 °C and then clarifying the solution via pelleting at 35,000*g* for 20 min at RT. The lysate was then further clarified by passage through a 0.2 µm syringe filter.

The His-tagged *Nar* Hfq construct was isolated via immobilized metal affinity chromatography (IMAC) using a 5 mL pre-packed iminodiacetic acid Sepharose column (GE Lifesciences) charged with Ni^{2+}. The chromatographic steps were conducted at room temperature (RT) on an NGC medium-pressure liquid chromatography system (Bio-Rad). The lysate was loaded onto the column, which was then washed with four column volumes of wash buffer (50 mM Tris pH 7.5, 150 mM NaCl, 10 mM imidazole). The protein was eluted by applying a linear gradient of elution buffer (50 mM Tris pH 7.5, 150 mM NaCl, 600 mM), from 0 to 100%, over 10 column volumes. The flow-through was fractionated and monitored via absorbance at 280 nm. Fractions containing protein were pooled and dialyzed into a buffer consisting of 50 mM Tris pH 7.5, 150 mM NaCl, and 12.5 mM EDTA. To cleave the 6x-His tag, *Nar* Hfq was incubated at RT overnight with a 1:600 w/w ratio of Hfq:thrombin. A benzamidine affinity column was then used to remove the thrombin. Finally, to ensure sample homogeneity, the *Nar* Hfq protein was applied to a preparative-grade HiPrep 16/600 Sephacryl S-300 HR gel-filtration column, run at RT on the NGC system mentioned earlier. *Nar* Hfq sample purity was assessed via SDS-PAGE and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF spectra), as described previously for *Aquifex aeolicus* Hfq1 [22].

2.2 Analytical size-exclusion chromatography and multi-angle light scattering to assay oligomeric states

Prior to chromatographic analysis, samples of *Nar* Hfq were dialyzed into a buffer consisting of 25 mM Tris pH 7.5, 150 mM NaCl and 12.5 mM EDTA. Analytical size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) was performed with a pre-packed Superdex 200 Increase 10/300 GL column on a Waters HPLC system. A Waters UV-Vis detector measured absorbance at 280 nm and light scattering measurements were taken from three detection angles using a Wyatt MiniDAWN TREOS (λ = 658 nm). Refractive index was recorded using a Wyatt Optilab T-rEX. Data processing was carried out using the *ASTRA* software package (Wyatt) in order to obtain the weight-averaged molecular mass (M_w) for select peaks.

2.3 Fluorescence polarization-based assays

Fluorescence polarization assays and calculations of RNA-binding affinities were carried out as previously described for *Aquifex aeolicus* Hfq1 [22]. Briefly, polarization data were obtained using the following four 5'-fluorescein–3'-OH-labelled oligoribonucleotides: FAM–U_6, FAM–C_6, FAM–A_6, and FAM–A_{18}. The FAM-RNAs, at a concentration of 5 nM, were added to a serial dilution of *Nar* Hfq in 25 mM Tris pH 7.5, 150 mM NaCl, and 12.5 mM EDTA, and equilibrated at RT for 45 minutes. Polarization values were plotted against the log[(Hfq)].

For monophasic binding, as in the case of U_6, binding data were fit to a four-parameter sigmoidal curve:

$$y(x) = A_2 + (A_1 - A_2)\left\{\frac{1}{1+exp[(x-x_0)/dx)}\right\}$$

where x is the log[(Hfq)], A_1 and A_2 are the polarization values at the lower and upper bounds of the binding curve, respectively, x_0 is the apparent equilibrium dissociation constant (K_d), and dx is a representation of the Hill coefficient.

For suspected biphasic binding, as observed for A_{18}, binding data were fitted to a five-parameter equation:

$$y(x) = A_2 + (A_1 - A_2)(frac)\left\{\frac{1}{1+exp[(x-x_0)/dx)}\right\}$$

where the additional *frac* term is a free parameter that corresponds to the fraction of binding in the first phase (or binding event) of the overall binding isotherm.

2.4 Crystallization and initial X-ray diffraction studies

2.4.1 Crystallization

Prior to crystallization, purified recombinant *Nar* Hfq was dialyzed into 25 mM Tris pH 7.5, 150 mM NaCl, and 12.5 mM EDTA, and then concentrated to 8.0 mg/ml. Initial sparse-matrix crystallization screening was performed under vapor diffusion in sitting drop format using the JCSG Core grid screens and 96-well format IntelliPlates (Art Robbins). Several initial hits were identified and further screened by systematically varying pH and concentration of the crystallization components. Grid screening and further crystallization trials were conducted via hanging-drop vapor diffusion in 24-well Linbro trays. 6 µl drops (3 µl protein + 3 µl crystallization buffer) were equilibrated against 600-µl wells of crystallization buffer at 18 °C. Reproducible crystals were obtained using 0.1 M CAPS pH 10.8 and 20% w/v PEG 2000. Crystals grew within one week as clusters of thin plates.

Microseeding was performed by setting crystal trays as described above, except with 60-80% of the precipitant concentrations. After equilibrating the tray for 90 minutes, drops were then streak-seeded from a slurry of crushed crystals (at a 1:1, 1:10, 1:500, or 1:1,000 dilution with *Nar* Hfq buffer) using a whisker microtool. Microseeding resulted in thicker, more separate plates with slightly improved diffraction (to ~3.5 Å).

2.4.2 Diffraction data collection and processing

Prior to X-ray diffraction experiments, crystals were transferred to a mother liquor consisting of 0.1 M CAPS pH 10.8, 20% w/v PEG 2000, and 30% v/v MPD in order to cryoprotect. Crystals were harvested with nylon loops and flash cooled in liquid nitrogen. Diffraction data was collected at the Advanced Photon source (APS) at beamline 22-ID. Initial processing of the diffraction data (indexing, scaling, and merging) was performed in XDS and HKL [37,38]. The initial quality of all diffraction datasets was analyzed using Xtriage within the PHENIX suite of programs [39], and anisotropy was measured via UCLA's anisotropy server [40]. Our partial molecular replacement solution was computed in phaser within the PHENIX suite. *E. coli* Hfq, with a sequence identity of 55% to Hfq_N, and 44% sequence identity to Hfq_C, was used as a search model. *Nar* Hfq crystallized in spacegroup *C*2 with unit cell dimensions of a = 111.51 Å, b = 64.22 Å, c = 82.08 Å and β = 95.4°. The AutoBuild functionality within PHENIX was used in an attempt to build the correct *Nar* Hfq sequence into the molecular replacement model.

3. Results and discussion

3.1 A tandem-domain Hfq from *N. aromaticivorans*

The soil-dwelling α-proteobacterium *N. aromaticivorans* possesses an open reading frame that encodes a putative 193-amino acid Hfq protein. The sequence of *N. aromaticivorans* (*Nar*) Hfq includes two tandem (~60 residue) Hfq domains, which we will refer to as N-terminal Hfq (Hfq_N) and C-terminal Hfq (Hfq_C), separated by a 20-residue linker of no notable sequence characteristics. *Nar* Hfq also has a 15 residue acidic C-terminal tail, reminiscent of other single-domain Hfq proteins [41–43], and interestingly, an ~40-residue proline-rich N-terminal tail. Proline-rich polypeptides are well-known to play an important role in cellular signaling pathways, and are also the recognition motifs of SH3 domains [44]. intriguingly, the Sm and SH3 domains are both members of superfold class of small β-barrels. Alignment of Hfq_N and Hfq_C to the Sm domains of other Hfq proteins shows that Hfq_C is more divergent in sequence than is Hfq_N (Fig 1). While the proximal and lateral RNA-binding residues appear to be conserved for Hfq_N and Hfq_C, the lateral binding site is not as well conserved, for example the highly conserved Asn45 (Asn13 in *E. coli* Hfq numbering) is instead a serine (Ser126) for Hfq_C. The arginine-rich 'RRER' motif ('RKNK' in Hfq_N) is also absent in Hfq_C, which instead has the rather different sequence of 'RDSG'.

In order to structurally and functionally characterize *Nar* Hfq, we have cloned, expressed, and purified the recombinant protein in *E. coli*, as verified through analysis via SDS-PAGE and MALDI-TOF mass spectrometry (Fig 2). The 6xHis-tagged construct is 213 amino acids in length with a molecular weight of 23.3 kDa and a predicted isoelectric point (pI) of 9.24. After

proteolytic cleavage of the His-tag (as verified through MALDI-TOF) the final construct is 196 amino acids in length, with a molecular weight of 21.4 kDa and pI of 9.02. In our initial purification efforts we found that *Nar* Hfq co-purified with a small amount of nucleic acid, as assessed by the ratio of absorbance at 260 and 280 nm (A_{260}/A_{280}). Specifically, the A_{260}/A_{280} was ~1.0 for *Nar* Hfq samples. An A_{260}/A_{280} ratio of 0.7 would be expected for pure protein and 1.2 for pure RNA [45]. Various attempts to remove the nucleic acid contaminant using nucleases (RNases and DNases) or denaturants (guanidinium) either proved unsuccessful at removing the A_{260}-absorbing species or else reduced protein solubility, so such efforts were abandoned. Hfq proteins are notorious for co-purifying with nucleic acids that can be difficult to remove; nevertheless, in many cases, Hfq has been successfully crystallized even in the presence of contaminating nucleic acid [46].

3.2 *Nar* Hfq oligomerizes as a trimer

If we assume that *Nar* Hfq does oligomerize similarly to other characterized Hfqs, as a pseudo-hexameric ring comprised of six Sm domains contributed by three *Nar* Hfq polypeptides, one of two likely oligomerization modes can be proposed: *(i) Nar* Hfq forms a trimeric ring of alternating Hfq_N and Hfq_C domains, or *(ii) Nar* Hfq assembles as a hexamer of two stacked rings, one ring composed of Hfq_N domains and the other of Hfq_C domains (Fig 2A). Note that further, supra-ring oligomerization via the stacking of two rings, would also be possible in scenario *(i)*. Also note that scenario *(ii)* allows for greater variability at the single-ring level, as two heptameric rings could also potentially form. As an initial measure of *Nar* Hfq oligomeric state in solution we used size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS). SEC-MALS enables us to determine the absolute molecular mass, even in the

presence of aberrant elution profiles, as have been previously seen for other Hfq homologs due to interactions with chromatographic resins, and aspherical shape of the toroidal disc [22].

Nar Hfq eluted as a uniform, monodisperse peak, with a calculated weight-averaged molecular weight (M_W) of 61.43 kDa (Fig 2B). With a predicted MW of an *Nar* Hfq trimer at 64.2 kDa, this calculated M_W indicates that *Nar* Hfq forms a trimer in solution to within 5% agreement. The presence of such a trimer eliminates the possibility of model *(ii)*, as that stoichiometry would require a total of six subunits to form a full ring. Thus, we suspect that *Nar* Hfq likely forms a trimeric ring in solution, consisting of alternating Hfq_N and Hfq_C domains. Note that this mode of oligomerization could have significant implications for RNA-binding. For example, the proximal site of this pseudo-hexameric ring no longer consists of six equivalent binding pockets but instead, of two alternating pockets with potentially different affinities (or even nucleotide specificities). Similar behavior has been observed, in terms of structure-function relationships, with the heteroheptameric Sm proteins. The U1 snRNP core, for example, recognizes an 'AAUUUG' sequence motif of the U1 RNA, with each Sm subunit binding specifically and sequentially to one nucleotide [30]. Fundamentally, this affords a structural basis/mechanism for subfunctionalization.

3.3 *Nar* Hfq binds U-rich and A-rich RNA

To examine whether *Nar* Hfq was able to bind RNA in a manner similar to other Hfq proteins, we performed fluorescence polarization experiments with 5'-FAM-labelled U_6 and A_{18} (Fig 4). These single-stranded RNA oligonucleotides were used as binding partners represent the proximal (U_6) and distal (A_{18}) motifs for RNA-binding to known Hfqs. Note that we also tested binding of *Nar* Hfq to FAM-A_6 and FAM-C_6, but no significant binding was seen for either of these oligonucleotides (data not shown). FAM-U_6 bound *Nar* Hfq with an apparent dissociation

constant ($K_{d,app}$) of 247 nM, which is consistent with the U_6 binding affinities observed previously for both gram-positive and gram-negative bacteria [22,47–49]. While *Nar* Hfq also associated with A_{18} RNA, the binding observed was biphasic, with a high-affinity event ($K_{d,app}$ of 13.2 nM), and a low-affinity $K_{d,app}$ of 4.48 μM. This biphasic curve most likely stems from either *(i)* differences in binding affinities between Hfq_N and Hfq_C or *(ii)* the subsequent formation of higher-order oligomers, as was found to occur for *E. coli* Hfq presence of A_{18} RNA [18,50,51]. A recent crystal structure elucidated this biochemical behavior, showing that two A_{18} RNAs bound to the distal face of Hfq could further base pair, resulting in an $(Hfq_6\text{-}A_{18})_2$ supramolecular assembly [52]. Previously, we have observed similar behavior with *A. aeolicus* Hfq1 (Chapter 3) and Hfq2 (Chapter 4), though we do note that both proteins exhibited a monophasic binding curve with FAM-A_{18} [22].

3.4 Initial crystallization efforts

Ultimately, atomic-resolution X-ray diffraction data is required to understand the structure and detailed geometry of the *Nar* Hfq oligomer as well as fully understand its RNA-binding properties. *Nar* Hfq was crystallized under several conditions, as determined via sparse-matrix screening. After further (and finer) grid screening, reproducible crystals, in the form of thin clusters of plates (Fig 5A), were produced in the following optimized condition: 0.1M CAPS pH 10.8 and 20% w/v PEG 2000. Initial diffraction data collected from these crystals was severely anisotropic and limited to low-resolution (reflections extended. to ~3.9 Å in a* and b*, and to ~8.5 Å in c*) so further optimization using an additive screen was next attempted. 5% w/v n-dodecyl β-D-maltoside was initially identified as a potential additive, and resultant crystals had a new spherulite habit (Fig 5B). Unfortunately however, these crystals did not diffract to any

noticeable degree. Finally microseeding of the initial crystals was tried; the resultant crystals exhibited thicker more separated plates with slightly improved diffraction (to ~3.5 Å).

We were able to obtain a dataset of high enough quality (3.77 Å resolution, Fig 6A) to get a partial solution through molecular replacement, using *E. coli* Hfq, with a sequence identity of 55% to Hfq_N, and 44% sequence identity to Hfq_C, as our search model. In this form, *Nar* Hfq crystallized in spacegroup *C*2 with unit cell dimensions of a = 111.51 Å, b = 64.22 Å, c = 82.08 Å and β = 95.4°. Examination of our electron density maps provided explanation for the severe anisotropy that we had observed; *Nar* Hfq appeared to crystallize in layers of parallel rings separated by ~40 Å (Fig 6B). Between these layers—and forming the crystal contacts between—them were the proline-rich, N-terminal tails, which in this case had adopted an extended conformation. Proline-rich polypeptides have the propensity to form a unique helix, known as a polyproline type II (PPII) helix [53]. Unfortunately this initial X-ray dataset was of insufficient quality to build a complete model. However, it did provide valuable insight to guide future protein engineering and crystallization efforts.

4. Conclusions and future directions

Here, we have discovered that *N. aromaticivorans* has a tandem-domain Hfq homolog which binds RNA in a manner akin to known Hfqs. We have shown that *Nar* Hfq binds U-rich and A-rich RNAs with nanomolar affinities, and furthermore exhibits biphasic binding to A_{18} RNA. We have also found that *Nar* Hfq oligomerizes as a trimer in solution, indicating a ring assembly of alternating Hfq_N and Hfq_C domains. Such an assembly would be the first example of a (pseudo) heterohexameric Hfq ring. While we were able to crystallize *Nar* Hfq, our initial crystals diffracted poorly and exhibited severe anisotropy. From a partial molecular replacement solution, we can attribute this poor X-ray data quality to the loose packing of *Nar* Hfq rings in the

direction between plates, due to the extended, proline-rich N-terminal tails. This prompted us to design an N-terminal truncation mutant for further crystallographic study.

A new construct was designed which had the first 34 residues truncated (ΔN-*Nar* Hfq). This construct, which also includes an N-terminal 6x-His tag, is 180 residues in length with a molecular weight of 20.1 kDa and an isoelectric point pI of 8.72. The final construct after His-tag cleavage is 163 residues in length, with a molecular weight of 18.2 kDa and pI of 8.06. ΔN-*Nar* Hfq was successfully expressed and purified recombinantly in *E. coli*, as verified through SDS-PAGE. Notably, samples of ΔN-*Nar* Hfq had much lower A_{260}/A_{280} ratios of ~0.8, indicating that the protein was not co-purifying with nucleic acid. ΔN-*Nar* Hfq also tended to precipitate out of solution; a series of storage buffers with varying pH, salt concentration, and temperature were tested, but ultimately pelleting the precipitated protein was most effective. The remaining soluble fraction (at a concentration of ~5 mg/mL) appears stable over an extended period of 2-3 weeks (as assessed by the lack of further precipitation and single peak on SDS-PAGE gels).

To verify that ΔN-*Nar* Hfq behaves similarly to full-length *Nar* Hfq we will repeat the SEC-MALS and FP experiments to confirm that the oligomeric state and RNA-binding affinities (respectively) are preserved in this construct. We will then attempt to crystallize ΔN-*Nar* Hfq through sparse matrix screening, followed by further grid screening as necessary. Ultimately, our goal is to produce higher-quality (in terms of resolution and anisotropy) diffracting crystals to determine the structure of *Nar* Hfq at atomic resolution.

Acknowledgements

We thank J. Bushweller (UVa) for access to the fluorescent plate reader, J. Shannon (UVa) for assistance with MALDI-TOF, D. Cascio and M. Sawaya (UCLA) for crystallographic advice, and L. Columbus, C. McAnany, and P. Randolph (UVa) for helpful discussions. Beamlines SER-CAT 22-ID and NE-CAT

24-ID-C/E at Argonne National Laboratory's Advanced Photon Source are DOE facilities (DE-AC02-06CH11357).

Funding information

Funding for this research was provided by: National Science Foundation, Division of Molecular and Cellular Biosciences (award No. 1350957); Thomas F. and Kate Miller Jeffress Memorial Trust (award No. J-971).

References

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

2. 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. John Wiley & Sons, Ltd; 2002;21: 3546–3556.

3. 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. 2002;320: 129–142.

4. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D. Structure and assembly of an augmented Sm-like archaeal protein 14-mer. Proc Natl Acad Sci U S A. National Academy of Sciences; 2003;100: 4539–4544.

5. Urlaub H, Raker VA, Kostka S, Lührmann R. Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. EMBO J. 2001;20: 187–196.

6. Franze de Fernandez MT, Eoyang L, August JT. Factor fraction required for the synthesis of bacteriophage Qbeta-RNA. Nature. 1968;219: 588–590.

7. Sauer E. Structure and RNA-binding properties of the bacterial LSm protein Hfq. RNA Biol. 2013;10: 610–618.

8. Updegrove TB, Zhang A, Storz G. Hfq: the flexible RNA matchmaker. Curr Opin Microbiol. 2016;30: 133–138.

9. Vogel J, Luisi BF. Hfq and its constellation of RNA. Nat Rev Microbiol. 2011;9: 578–589.

10. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. Cell. 2004;118: 69–82.

11. 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 Dev. 1996;10: 1143–1151.

12. Gottesman S, McCullen CA, Guillier M, Vanderpool CK, Majdalani N, Benhammou J, et al. Small RNA regulators and the bacterial response to stress. Cold Spring Harb Symp Quant Biol. 2006;71: 1–11.

13. Massé E, Gottesman S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. Proc Natl Acad Sci U S A. 2002;99: 4620–4625.

14. Chao Y, Vogel J. The role of Hfq in bacterial pathogens. Curr Opin Microbiol. 2010;13:

24–33.

15. Vrentas C, Ghirlando R, Keefer A, Hu Z, Tomczak A, Gittis AG, et al. Hfqs in Bacillus anthracis: Role of protein sequence variation in the structure and function of proteins in the Hfq family. Protein Sci. 2015;24: 1808–1819.

16. 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. 2015;1850: 1661–1668.

17. Ramos CG, da Costa PJP, Döring G, Leitão JH. The novel cis-encoded small RNA h2cR is a negative regulator of hfq2 in Burkholderia cenocepacia. PLoS One. 2012;7: e47896.

18. 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. 2004;11: 1206–1214.

19. Panja S, Schu DJ, Woodson SA. Conserved arginines on the rim of Hfq catalyze base pair formation and exchange. Nucleic Acids Res. Oxford University Press; 2013;41: 7536–7546.

20. 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 U S A. 2012;109: 9396–9401.

21. Dimastrogiovanni D, Fröhlich KS, Bandyra KJ, Bruce HA, Hohensee S, Vogel J, et al. Recognition of the small regulatory RNA RydC by the bacterial Hfq protein. eLife Sciences. eLife Sciences Publications Limited; 2014;3: e05375.

22. Stanek KA, Patterson-West J, Randolph PS, Mura C. Crystal structure and RNA-binding properties of an Hfq homolog from the deep-branching Aquificae: conservation of the lateral RNA-binding mode. Acta Crystallogr D Struct Biol. 2017;73: 294–315.

23. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol. 2011;3. doi:10.1101/cshperspect.a003707

24. Tharun S. Roles of eukaryotic Lsm proteins in the regulation of mRNA function. Int Rev Cell Mol Biol. 2009;272: 149–189.

25. Fischer U, Englbrecht C, Chari A. Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdiscip Rev RNA. 2011;2: 718–731.

26. Bouveret E, Rigaut G, Shevchenko A, Wilm M, Séraphin B. A Sm-like protein complex that participates in mRNA degradation. EMBO J. 2000;19: 1661–1671.

27. Achsel T, Brahms H, Kastner B, Bachi A, Wilm M, Lührmann R. A doughnut-shaped heteromer of human Sm-like proteins binds to the 3'-end of U6 snRNA, thereby facilitating U4/U6 duplex formation in vitro. EMBO J. 1999;18: 5789–5802.

28. Novotny I, Podolská K, Blazíková M, Valásek LS, Svoboda P, Stanek D. Nuclear LSm8 affects number of cytoplasmic processing bodies via controlling cellular distribution of

Like-Sm proteins. Mol Biol Cell. 2012;23: 3776–3785.

29. Leung AKW, Nagai K, Li J. Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2011;473: 536.

30. 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. 2015;4. doi:10.7554/eLife.04986

31. 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 U S A. National Academy of Sciences; 2001;98: 5532–5537.

32. 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. J Biol Chem. 2003;278: 1239–1247.

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

34. Ramos CG, Sousa SA, 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. 2011;193: 1515–1526.

35. Sun X, Zhulin I, Wartell RM. Predicted structure and phyletic distribution of the RNA-binding protein Hfq. Nucleic Acids Res. 2002;30: 3662–3671.

36. Kaplan MM. Novosphingobium aromaticivorans: a potential initiator of primary biliary cirrhosis. Am J Gastroenterol. 2004;99: 2147–2149.

37. Kabsch W. XDS. Acta Crystallogr D Biol Crystallogr. 2010;66: 125–132.

38. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes. Acta Crystallogr D Biol Crystallogr. 2006;62: 859–866.

39. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66: 213–221.

40. 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 U S A. 2006;103: 8060–8065.

41. Arluison V, Folichon M, Marco S, Derreumaux P, Pellegrini O, Seguin J, et al. The C-terminal domain of Escherichia coli Hfq increases the stability of the hexamer. Eur J Biochem. 2004;271: 1258–1265.

42. Beich-Frandsen M, Vecerek B, Konarev PV, Sjöblom B, Kloiber K, Hämmerle H, et al.

Structural insights into the dynamics and function of the C-terminus of the E. coli RNA chaperone Hfq. Nucleic Acids Res. 2011;39: 4900–4915.

43. Santiago-Frangos A, Jeliazkov JR, Gray JJ, Woodson SA. Acidic C-terminal domains autoregulate the RNA chaperone Hfq. eLife Sciences. eLife Sciences Publications Limited; 2017;6: e27049.

44. Zarrinpar A, Bhattacharyya RP, Lim WA. The structure and function of proline recognition domains. Sci STKE. 2003;2003: RE8.

45. De Mey M, Lequeux G, Maertens J, De Maeseneire S, Soetaert W, Vandamme E. Comparison of DNA and RNA quantification methods suitable for parameter estimation in metabolic modeling of microorganisms. Anal Biochem. 2006;353: 198–203.

46. Stanek KA, Mura C. Producing Hfq/Sm Proteins and sRNAs for Structural and Biophysical Studies of Ribonucleoprotein Assembly. In: Arluison V, Valverde C, editors. Bacterial Regulatory RNA: Methods and Protocols. New York, NY: Springer New York; 2018. pp. 273–299.

47. Sauer E, Weichenrieder O. Structural basis for RNA 3'-end recognition by Hfq. Proc Natl Acad Sci U S A. 2011;108: 13065–13070.

48. Horstmann N, Orans J, Valentin-Hansen P, Shelburne SA 3rd, Brennan RG. Structural mechanism of Staphylococcus aureus Hfq binding to an RNA A-tract. Nucleic Acids Res. 2012;40: 11023–11035.

49. 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. 2014;20: 1548–1559.

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. Biochemistry. American Chemical Society; 2006;45: 4875–4887.

51. Soper TJ, Woodson SA. The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA. RNA. Cold Spring Harbor Laboratory Press; 2008;14: 1907–1917.

52. Schulz EC, Seiler M, Zuliani C, Voigt F, Rybin V, Pogenberg V, et al. Intermolecular base stacking mediates RNA-RNA interaction in a crystal structure of the RNA chaperone Hfq. Sci Rep. 2017;7: 9903.

53. Williamson MP. The structure and function of proline-rich regions in proteins. Biochem J. 1994;297 (Pt 2): 249–260.

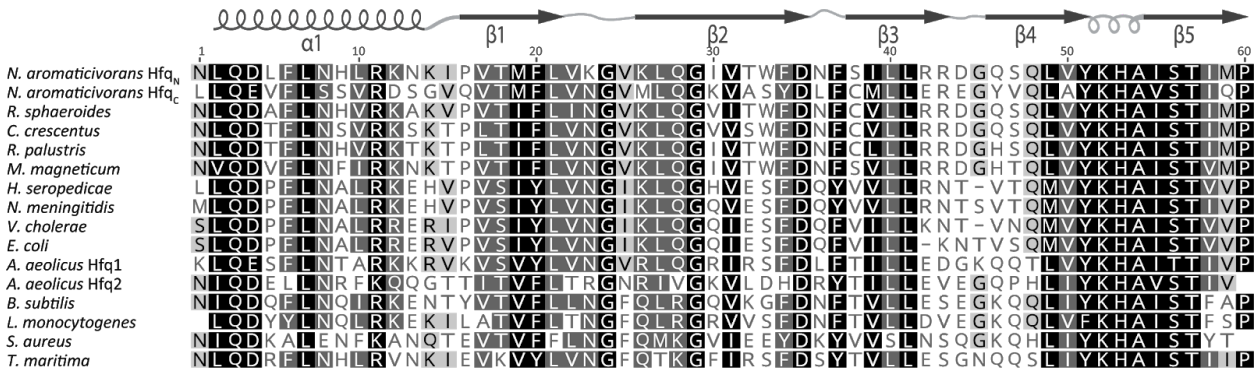


Figure 1. Sequence alignment of *Nar* Hfq_N and Hfq_C with representative Hfq homologs. Sequences were trimmed to the ~60 residue core domain and then aligned using *MUSCLE*. This alignment shows that Hfq_C is more divergent that Hfq_N, and lacks the arginine-rich 'RRER' motif (residues 11-14 here). The secondary structure of the Hfq domain is shown above in cartoon representation. Residues are colored according similarity to the consensus: black boxes with white text, 100% similar; dark grey box with with white text, 80-100% similar; light grey box with black text, 60-80% similar; white box with grey text, <60% similar. The sequences used are as follows: *Rhodobacter sphaeroides* (A3PJP5), *Caulobacter crescentus* (Q9A7H8), *Rhodopseudomonas palustris* (B3QIM2), *Magnetospirillum magneticum* (Q2W4P9), *Herbaspirillum seropedicae* (ADJ64436), *Neisseria meningitidis* (P64344), *Vibrio cholerae* (A5F3L7), *Escherichia coli* (P0A6X3), *Aquifex aeolicus* Hfq1 (WP_010880010) and Hfq2 (WP_0204015108), *Bacillus subtilis* (KFK79581.1), *Listeria monocytogenes* (CBY70202), *Staphylococcus aureus* (ADC37472), and *Thermotoga maritima* (Q9WYZ6).

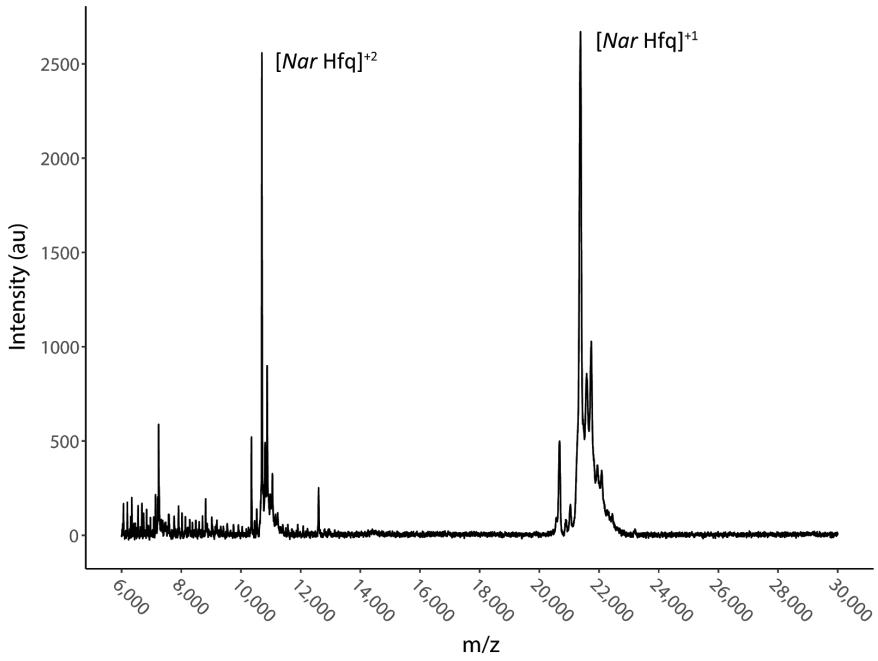


Figure 2. Purification of full-length *Nar* Hfq. MALDI-TOF spectrum of recombinantly expressed and purified *Nar* Hfq after 6x-His tag cleavage is shown. The final recombinant construct has an expected molecular weight of 21,414 Da for the monomer.

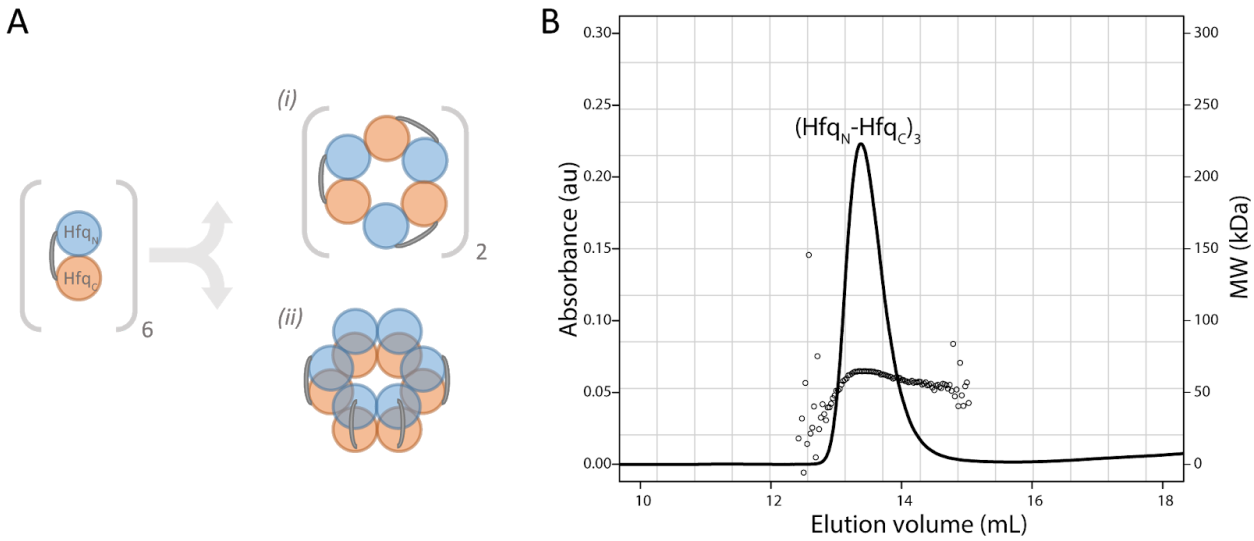


Figure 3. The oligomeric state of *Nar* Hfq. (A) Two proposed models for *Nar* Hfq oligomerization: *(i)* a single trimeric ring is formed from alternating Hfq_N (blue) and Hfq_C (orange) domains, or *(ii)* a hexameric assembly of two stacked rings, the top ring consisting of Hfq_N, and the bottom of Hfq_C. (B) The oligomeric state of *Nar* Hfq in solution was assayed via size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS). Absorbance was recorded at 280 nm to monitor protein elution (black trace), and the molar mass distribution data (open circles) was calculated from measurements of light scattering. The weight-averaged molecular weight (M_W) of the peak was calculated to be 61.43 kDa, indicating a trimeric assembly. This eliminates the possibility of model *(ii)*, and points to model *(i)* as the correct oligomeric assembly.

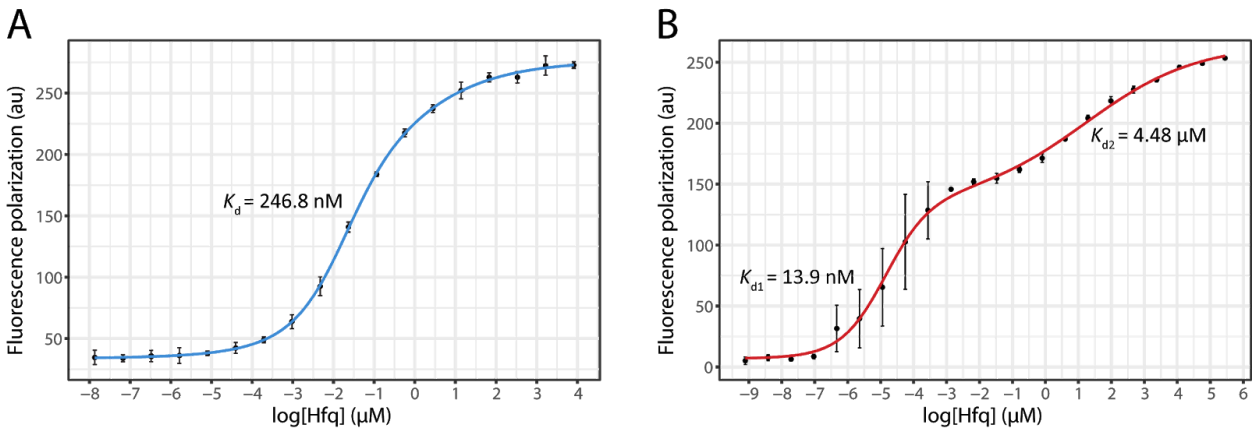


Figure 4. *Nar* Hfq binds to U_6 and A_{18} RNAs with high affinity. Binding affinities for 5 nM FAM-labelled (A) U_6 or (B) A_{18} single-stranded RNA oligonucleotides were measured via fluorescence polarization assays. Data from three replicates were plotted (standard errors are shown as vertical lines) and fitted to (A) a four-parameter sigmoidal curve, representative of monophasic binding, or (B) a five-parameter curve with two inflection points, representative of biphasic binding. The apparent dissociation constants ($K_{d,app}$) are shown in terms of the logarithm of the concentration of *Nar* Hfq monomer.

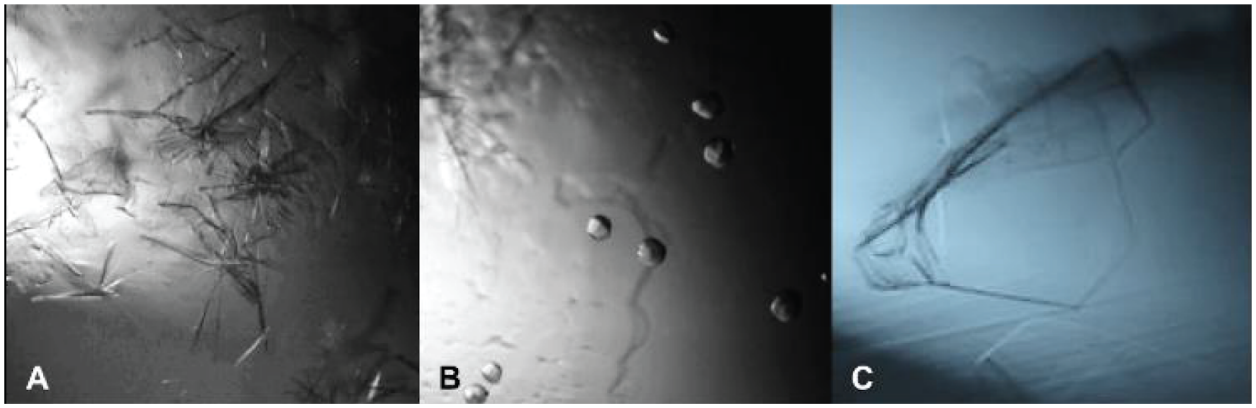


Figure 5. Preliminary crystallization of *Nar* Hfq. (A) Reproducible crystals in the form of clusters of plates formed in 0.1 M CAPS pH 10.8 20% w/v PEG 2000. These crystals were highly diffracted to poor resolution and the resulting diffraction data was highly anisotropic. (B) Use of n-dodecyl-*β*-D-maltoside as an additive resulted in crystals exhibiting a new spherulite habit. However, these crystals did not diffract. (C) Microseeding of the crystals in (A) resulted in thicker, more separated plates, with slightly improved diffraction quality.

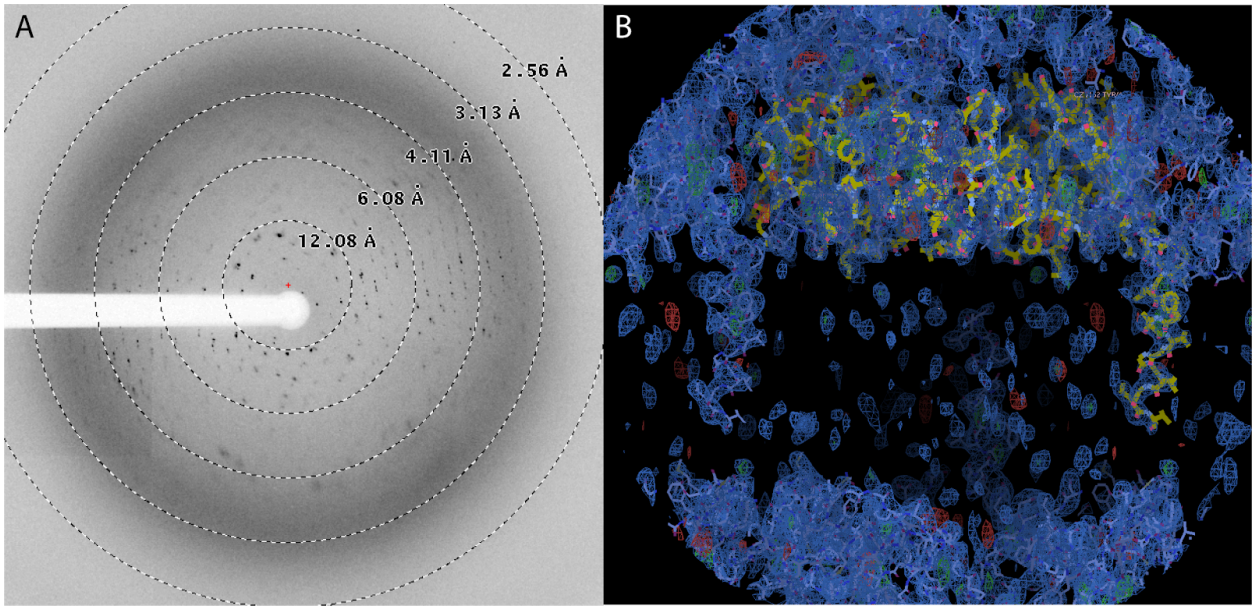


Figure 6. Preliminary crystallographic studies of *Nar* Hfq. (A) X-ray diffraction by *Nar* Hfq crystals. Shown here is a single frame from an *Nar* Hfq X-ray dataset collected on SER-CAT beamline 22-ID. *Nar* Hfq crystallized in 0.1 M CAPS pH 10.8 and 20% w/v PEG 2000. Crystals diffracted anisotropically, to ~3.7 Å in a* and b*, and to ~8.5 Å in c*. (B) A partial molecular replacement solution of *Nar* Hfq is shown, as viewed in Coot. The partial model, shown in yellow, included fragments of nonsense sequence built into extra density between the Hfq rings, which was attributed to the N-terminal regions of *Nar* Hfq. The $2mF_O$-DF_C electron density map, contoured at 1.5 σ, is shown in blue.

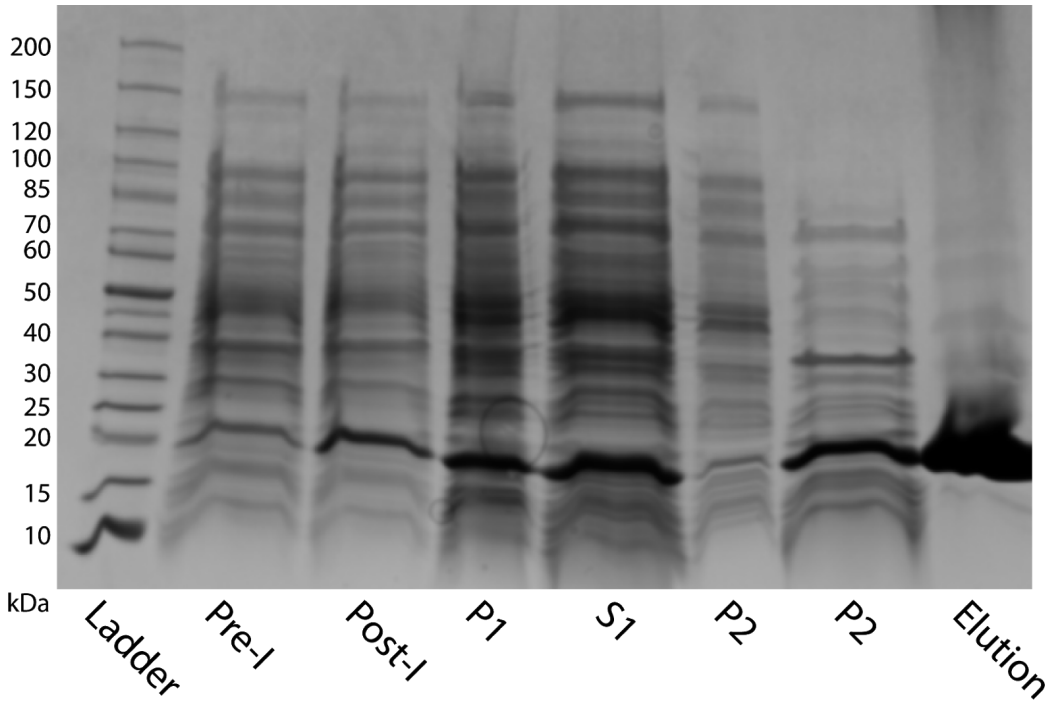


Figure 7. Recombinant expression and purification of ΔN-*Nar* Hfq as assessed via SDS-PAGE gel. Gel lanes are as follows: ladder (EZ-Run *Rec* protein ladder, Fisher), Pre-I (pre-induction), Post-I (post-induction of ΔN-*Nar* Hfq expression with IPTG), P1 (pellet) and S1 (supernatant) after centrifugation of cells following a four hour induction, P2 (pellet) and S2 (supernatant) from a second pelleting step after a 20 minute incubation of cell lysate at 75 °C, and Elution from IMAC purification of the 6xHis-tagged ΔN-*Nar* Hfq. The recombinant ΔN-*Nar* Hfq construct has an expected MW of 20.1 kDa prior to cleavage of the His-tag.