3662–3671 *Nucleic Acids Research, 2002, Vol. 30 No. 17*

Predicted structure and phyletic distribution of the RNA-binding protein Hfq

Xueguang Sun, Igor Zhulin and Roger M. Wartell*

School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received June 15, 2002; Revised and Accepted July 7, 2002

ABSTRACT

Hfq, a bacterial RNA-binding protein, was recently shown to contain the Sm1 motif, a characteristic of Sm and LSm proteins that function in RNA processing events in archaea and eukaryotes. In this report, comparative structural modeling was used to predict a three-dimensional structure of the Hfq core sequence. The predicted structure aligns with most major features of the *Methanobacterium thermoautotrophicum* LSm protein structure. Conserved residues in Hfq are positioned at the same structural locations responsible for subunit assembly and RNA interaction in Sm proteins. A highly conserved portion of Hfq assumes a structural fold similar to the Sm2 motif of Sm proteins. The evolution of the Hfq protein was explored by conducting a BLAST search of microbial genomes followed by phylogenetic analysis. Approximately half of the 140 complete or nearly complete genomes examined contain at least one gene coding for Hfq. The presence or absence of Hfq closely followed major bacterial clades. It is absent from high-level clades and present in the ancient Thermotogales-Aquificales clade and all proteobacteria except for those that have undergone major reduction in genome size. Residues at three positions in Hfq form signatures for the beta/gamma proteobacteria, alpha proteobacteria and low GC Gram-positive bacteria groups.

INTRODUCTION

Hfq, also called HF-I, is a 12 kDa heat-stable protein, encoded by the *hfq* gene at 95 min on the *Escherichia coli* chromosome map (1). Originally discovered as a host factor required for bacteriophage Qβ RNA replication (2), it was later shown to be associated with ribosomes (3) and, to a lesser extent, with the nucleoid (4,5). Hfq is a global regulator of *E.coli* metabolism, and disruption of the *hfq* gene can cause a pleiotropic phenotype (6). The broad impact of the protein appears to stem from its role in regulating the stability and/or translation of mRNAs from a number of regulatory genes. One of these mRNAs is the *rpoS* mRNA that encodes the stationary phase sigma factor σ^s of RNA polymerase (7,8). Mutational studies suggest that Hfq is involved in the processes that affect the secondary structure near the 5′ end of *rpoS* mRNA alleviating an inhibition of ribosome access to a translation start region (9). A similar behavior is inferred from studies that indicate Hfq helps open an inhibitory stem–loop structure at the 3′ end of Qβ plus-strand RNA to mediate access of Qβ replicase (2).

Hfq has been shown to affect the *in vivo* stability of mRNAs expressed from the *ompA*, *mutS*, *miaA* and *hfq* genes (10,11), and to stimulate elongation of the poly(A) tail of the *rpsO* mRNA (12). The mechanism of the effect of Hfq on mRNA stability appears to involve its influence on the interaction of non-coding regulatory RNAs with specific mRNAs. It has been shown to affect the binding of regulatory RNAs DsrA, OxyS, RprA and Spot 42 with their target mRNAs (9,13–16). While functional roles for Hfq have been demonstrated, and general models for its mechanism of action proposed (9), the absence of structural information on Hfq and Hfq–RNA complexes hinders an understanding of its molecular mechanism(s) of action.

Recent amino acid sequence analysis of Hfq has shown that the N-terminal portion of Hfq is highly conserved among a number of bacteria and shares a strong similarity with the Sm1 motif of Sm and Sm-like (LSm) proteins found in eukaryotes and archea (15,16). These results suggest that Hfq is an ancestral Sm protein. Sm proteins are essential components of the small nuclear ribonucleoproteins (snRNPs) that form spliceosomes (17,18). Sequence comparisons of Sm proteins from a range of species showed that the Sm motif is comprised of two conserved regions, Sm1 and Sm2, separated by a region varying in length and sequence (19). Biochemical and crystallographic studies (20,21) have demonstrated that the Sm motif dictates a common folding domain that enables Sm proteins to assemble onto a uridine-rich region of snRNAs and form a ring-like heteroheptamer. Formation of this core structure is essential for the stability and function of the snRNPs (22).

Searches of eukaryotic genome databases have shown that a large number of proteins contain the Sm sequence motif (23,24). Some of these proteins are similar to the originally characterized splicesomal Sm proteins, and others are referred to as LSm proteins. Analysis of archaeal genomes also revealed the presence of ORFs that encode LSm proteins (24). Biochemical and crystal studies of three archaeal LSm proteins revealed that they exhibit properties similar to their counterparts in eukaryotes. They bind to RNA with oligo(U) sequences, and assemble a heptameric ring around the RNA (25,26). A comparison of the monomer subunits in the crystal

*To whom correspondence should be addressed. Tel: +1 404 894 3735; Fax: +1 404 894 0519; Email: roger.wartell@biology.gatech.edu

structures of eukaryotic Sm proteins that form dimers with the monomer subunits of archaeal LSm proteins that form homodimers and heptamers (27,28) show strong similarities. Each subunit has a short alpha helix followed by five interwoven beta strands separated by short loops.

The presence of the Sm1 motif sequence in *E.coli* Hfq, and the ability of Hfq to form a hexameric ring and bind RNA, support the notion that it is evolutionarily related to the Sm family of proteins. However, the absence of the Sm2 motif in Hfq makes the structural relationship of Hfq with the known structures of Sm proteins uncertain. In this report secondary structure prediction, amino acid solvation properties and three-dimensional (3D) threading algorithms were used to predict a 3D structure for the N-terminal domain of Hfq. The predicted structure fits very well with the major features of the C_α backbone of *Methanobacterium thermoautotrophicum* LSm protein. The Sm1 motif sequence in Hfq is structurally aligned with its counterpart in the archaeal LSm protein, and a highly conserved portion of the Hfq sequence assumes a structural fold similar to that of the Sm2 motif of archaeal and eukaryotic Sm proteins. Highly conserved residues of Hfq are also located in the same structural region that is responsible for subunit assembly in the Sm proteins.

The strong structural similarity of Hfq and the LSm protein supports the hypothesis that Hfq is an ancestral Sm protein and contributes confidence in its predicted 3D structure. During revision of this manuscript a paper describing the crystal structure of the *Staphylococcus aureus* Hfq hexameric protein and a complex of this protein with RNA was published by Schumacher *et al.* (29). Our 3D model of the *E.coli* Hfq monomer is in excellent agreement with the monomer of this structure. Comparison of the *S.aureus* Hfq structure with predicted features of Hfq based on our structure/sequence analysis is presented below.

The presence of the Hfq protein in bacteria was explored by BLAST searches against bacterial genomes available in the NCBI databases. Approximately half of the bacterial genomes examined contain an Hfq protein based on strong amino acid sequence similarity, protein sequence length and amino acid conservation pattern. Phyletic distribution of Hfq indicates that it is an ancient protein. We obtained no evidence that Hfq might be a subject of lateral gene transfer and conclude that gene loss played a major role in its evolution. The bacterial species in which Hfq was absent were highly correlated with specific taxonomic or lifestyle trends.

MATERIALS AND METHODS

Database searches

Non-redundant database (NCBI) searches were performed by Position-Specific-Iterative (PSI)-BLAST program (30), using the amino acid sequence of *E.coli* Hfq (GI 16131994) as the primary query sequence. The inclusion threshold (E value) employed was 0.01. A multiple sequence alignment was constructed by the CLUSTAL W program (31). Additional BLAST searches were carried out against the Microbial Genomes database at NCBI (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/genom_table_cgi), using amino acid sequences of Hfq or the consensus Sm motif as the query.

Multiple sequence alignment and secondary structure prediction

A multiple sequence alignment of Hfq proteins was constructed by the CLUSTAL W program (31) using the output of the PSI-BLAST searches. The amino acid conservation pattern was determined by calculating consensus using the Perl script by Nigel Brown and Jianmei Lai (available at http://www.bork.embl-heidelberg.de/Alignment/consensus.html). The secondary structure was predicted using the consensus method JPRED2 (32), which utilizes multiple sequence alignments, along with PSI-BLAST and HMM profiles.

Fold recognition and 3D modeling

The 3D-PSSM (Position Specific Scoring Matrix) program (33) was employed to search for proteins with structural similarity to Hfq. This server provides 3D structural information about the backbone of a query protein by scoring the relationship between the residues of a query sequence with the residues of a homologous protein of known structure. The query protein is scanned against a library composed of proteins with known crystal structures and scored for compatibility using several scoring components. These include amino acid sequence profiles built from relatively close homologs, more general profiles containing more remote homologs, matching of secondary structure elements, and matching the propensities of residues to occupy varying levels of solvent accessibility. Known protein structures within the database with significant homology to the query sequence are used to produce closest fit alignments between query sequence and target structures that maximize position specific scores. The top 20 structural alignments to the query sequence are produced, each illustrating regions of similarity and differences.

The SWISS-MODEL comparative protein modeling server (34) was employed to generate a 3D model of the *E.coli* Hfq protein based on the structural alignment of its sequence with the highest scoring template structure determined by 3D-PSSM. In the initial step of modeling, the Hfq query sequence was modified in order to accommodate the four and six residue segments absent from Hfq when compared with the best template structure, the *M.thermoautotrophicum* LSm_α protein (28) (see Fig. 3A and B). Residues were inserted into the Hfq sequence at the minus-labeled segments shown in Figure 3C and D to produce a query sequence that would match the length of the *M.thermoautotrophicum* template sequence. This modified Hfq query sequence was submitted to the SWISS-MODEL server, which produced a predicted structure. The query sequence was then changed to the correct Hfq sequence by replacing the inserted residues with gaps and the Hfq sequence and template sequence resubmitted to the server in 'Optimize Mode' after aligning the gaps to the template sequence as indicated by the 3D-PSSM model.

RESULTS AND DISCUSSION

Sequence similarity of Hfq and Sm proteins

A search of the non-redundant database (NCBI) using PSI-BLAST was carried out with the *E.coli* Hfq amino acid sequence as the query sequence. Twenty-five similar (statistically significant) sequences were detected from a range of

bacterial species. Multiple alignment of these sequences is shown in Figure 1. Hfq proteins are highly conserved in their N-terminal halves of the molecules. This conserved domain corresponds to residues 7–64 in *E.coli* Hfq. In contrast, the C-termini of Hfq proteins vary greatly among the different species. In some instances it is totally absent, e.g. the 57 amino acid Hfq protein of *Bacillus anthracis*. This result implies that the C-terminal region might not play a significant role in the major function(s) of Hfq and attention was focused on the N-terminal region. This hypothesis is supported by a recent study showing that the Hfq homolog of *Pseudomonas aeruginosa*, consisting of 82 amino acids from the N-terminal end, can functionally replace *E.coli* Hfq for phage Qβ replication and for *rpoS* expression (35).

Two conserved motifs are observed in Hfq. The first motif, Sm1, is a counterpart of the Sm1 motif found in archaeal and eukaryotic Sm and LSm proteins (19). From Figure 1, it can be seen that the Sm1 sequence is well aligned to residues 20–52 of the *E.coli* Hfq sequence. As noted previously (16), the Sm2 motif of archaeal and eukaryotic proteins does not appear to have a counterpart in Hfq. However, Hfq does have an additional conserved region, YKHA, following the Sm1 motif. The relationship of the YKHA motif of Hfq with the Sm2 motif of Sm proteins was explored by generating a structural model of Hfq.

Comparison of predicted structure of Hfq with known structure of the Sm protein

The secondary structure of the consensus Hfq sequence was predicted by JPRED2 (32). Figure 1 shows that Hfq is a β-sheet-rich structure with an α helix at the N-terminus. All the predicted secondary structure elements fall in the region that is conserved in the multiple alignment. In contrast, no secondary structure elements were predicted in the C-terminus of Hfq. Crystal structures of several Sm proteins show a common fold for the Sm motif. The fold contains an N-terminal helix, followed by five segments of β strands (20,26–28). Strands β1, β2 and β3 are part of the Sm1 motif, whereas the Sm2 motif corresponds to β4 and β5 strands (shown in Fig. 1). The topology of the secondary structure elements in an Sm protein is schematically shown in Figure 2A and B. Strands β2, β3 and β4 are strongly bent to allow the formation of the hydrophobic core. The structural plasticity needed for such a high degree of curvature in the β1 strand is provided by several strictly conserved glycines that occur near the pivot points (Gly18, Gly23, Gly53 and Gly59 in Sm consensus; Fig. 1). The segment linking the β4 and β5 lies at the top of the U-shaped trough to close the protein into a β-barrel-like structure.

Hfq has the same predicted secondary structure elements in the Sm1 motif region as does the Sm protein: an α helix followed by β strands. The critical residues that are required for the β1 strand curvature (Gly29 and Gly34 in Hfq) are identical in all Hfq homologs. Interestingly, another long β strand (in some predictions, it was two separate β strands) was predicted in Hfq from Ser51 to Pro65. Although this region could not be aligned to the Sm2 motif in the Sm protein sequence, the length of this region—referred to as Xm2 in Figure 1—closely matches the length of the Sm2 motif. In addition, the Sm2 and Xm2 motifs both have highly conserved residues flanked by hydrophobic residues. There are two highly conserved residues in the middle of the Sm2 motif, Arg–Gly, and four highly conserved residues in the Xm2 motif, Tyr–Lys–His–Ala.

The 3D structure of Hfq was predicted using comparative modeling as described in the Materials and Methods. 3D-PSSM was first employed to thread the Hfq sequence as a query against known protein structures in a fold library. The template structure that produced the highest score for the Hfq sequence was the archaeal LSm_α protein from *M.thermoautotrophicum* (28). The structure of the LSm_α protein is shown in Figure 2A and B, while Figure 2C and D shows the best-fit model structure of Hfq with the locations of deletions and insertions from the template structure that produced the highest score. The Hfq model is well aligned to the LSm_α protein. The major difference is in the β3 and β4 region. Four and six amino acids of the template structure are absent from Hfq in the β3 and β4 strands respectively. Visually, the Hfq structure suggests that if these 10 residues are simultaneously deleted, the isolated loop 4 may be able to connect the remaining fragments of β3 and β4. The missing parts of Hfq, when compared with the LSm_α protein, fall just within the highly variable region between the Sm1 and Sm2 motifs, which includes loop 4 as well as parts of the β3 and β4 strands. This suggests that the amino acid sequences which constitute a minimum Sm fold can be shortened, and may be composed of adjacent Sm1 and Sm2 motifs with no variable linker.

The SWISS-MODEL program was then employed to generate a 3D model of the Hfq protein using the archaeal LSm_α protein as a template, and the information inferred from the 3D-PSSM highest scoring alignment shown in Figure 2C and D. The optimized structural model of Hfq is shown in Figure 3 where it is compared with the structure of the LSm_α protein. The features that are constant in both structures are illustrated in blue. A red ribbon designates Hfq and a green ribbon illustrates the LSm_α protein in the regions where there are differences in their structural features. The β3 and β4 strands of Hfq are shortened relative to the LSm_α protein and connected by loop 4. Loop 4 changes its orientation from up and to the right for the LSm_α protein to a downward direction for Hfq (Fig. 3A and B). The aqua-colored ribbon shows the location of the β4 strand residues SQMVY, and the β5 strand residues AISTVV. Figure 4 shows the β4–loop 5–β5 region in greater detail. The residues in the β5 strand of Hfq, STVVP, appear to occupy similar spatial locations as the corresponding residues of the LSm_α protein, VLISP. Based on sequence alignment, amino acid characteristics and secondary structure prediction, we anticipated that the highly conserved His–Ala residues of Hfq would occupy the same 3D positions as the highly conserved Arg–Gly of an Sm protein. However, a comparison of the structural models in Figure 4 indicates that the His–Ala pair in Hfq is shifted in their relative location two residues downstream when compared with the Arg–Gly residues in the Sm structure.

The predicted structure of Hfq was compared with the Sm protein structure with regard to segments that may be involved in subunit interaction in the formation of multimers. Several studies indicate that Hfq forms a hexamer (2,15,16,36,37), while Sm proteins form a homo- or hetero-heptamer depending on the number of distinct subunits available *in vivo*. Archea species form a heptamer composed of seven identical

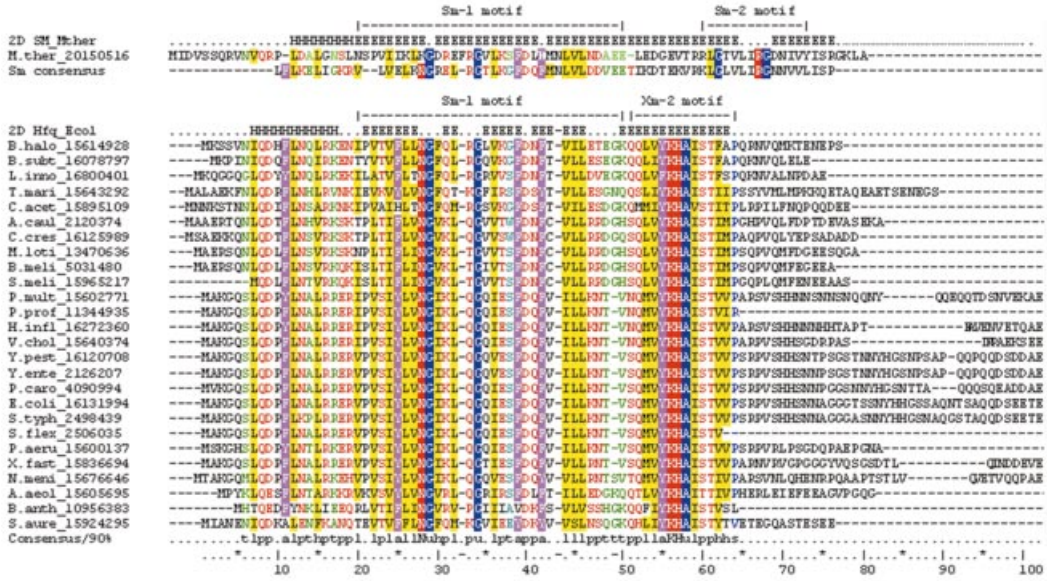

Figure 1. Multiple alignment of Hfq proteins from 26 bacterial genomes compared with the LSm protein from *M.thermoautotrophicum* and a consensus sequence for Sm proteins (shown above the alignment). Known secondary structure for the LSm protein and predicted structure for Hfq proteins are shown above the corresponding sequences: H, α helix; E, β strand. The Sm1 and Sm2 motifs of the Sm protein and the Sm1 and Xm2 motifs of Hfq proteins are shown. The 90% consensus shown below the alignment was derived using the following amino acid groupings. Positively charged residues (RKH) are shown as white letters on a red background; polar residues (p, KRHEDQNST) are shown as red letters; turn-like residues (t, ACDEGKNQRST) are green letters; bulky hydrophobic residues (h, ACLIVMHYFW) and the aliphatic subset of these type residues (l, LIVM) have a yellow background; aromatic residues (a, FHWY) are white letters with a purple background; small residues (s, ACDGNPSTV) are blue letters; tiny (u, AGS) are white letters with a blue background. Sequences are denoted by the species abbreviation followed by GI number. Species abbreviations: M.ther, *M.thermoautotrophicum*; B.halo, *Bacillus halodurans*; B.subt, *Bacillus subtilis*; L.inno, *Listeria innocua*; T.mari, *Thermotoga maritima*; C.acet, *Clostridium acetobutylicum*; A.caul, *Azorhizobium caulinodans*; C.cres, *Caulobacter crescentus*; M.loti, *Mesorhizobium loti*; B.meli, *Brucella melitensis biovar Abortus*; S.meli, *Sinorhizobium meliloti*; P.mult, *Pasteurella multocida*; P.prof, *Photobacterium profundum*; H.infl, *Haemophilus influenzae*; V.chol, *Vibrio cholerae*; Y.pest, *Yersinia pestis*; Y.ente, *Yersinia enterocolitica*; P.caro, *Pectobacterium carotovorum*; E.coli, *E.coli*; S.typh, *Salmonella typhimurium*; S.flex, *Shigella flexneri*; P.aeru, *P.aeruginosa*; X.fast, *Xylella fastidiosa*; N.meni, *Neisseria meningitidis*; A.aeol, *Aquifex aeolicus*; B.anth, *Bacillus anthracis*; S.aure, *S.aureus.*

subunits (26,27). Eukaryotes utilize different polypeptide chains to assemble a hetero-heptamer Sm complex (16,19,20). In both cases, adjacent monomers in the heptamer interact via pairing of the β4 and β5′ strands (the ′ indicates the adjacent subunit). Only the last five residues of the β4 strand in one Sm protein are involved in pairing with the β5′ strand of the adjacent Sm protein.

In the structural model of Hfq the first six residues of the β4 strand are absent when compared with the Sm protein. However, the five remaining residues that form the β4 strand and the beginning of loop 5 are located in similar positions to the residues of the LSm_{α} protein that participate in quaternary interactions (Fig. 4). The residues spanning the β4–β5 strand region of the human Sm protein that are involved in pairing adjacent subunits are LVLLRGSVIVV (20), while in the archaeal LSm they are TVLIRGQNIVY (28). In both cases, a group of hydrophobic residues flank a positively charged Arg that is engaged in several hydrogen bonds with main chain and side chain atoms of the adjacent subunit's β5′ strand. In the predicted Hfq structure, the residues spanning the β4 strand–loop–β5 strand region are SQMVYKHAISTVV. One again has hydrophobic residues flanking positively charged residues, in this case lysine and histidine. Although, as mentioned above, the His–Ala residues of Hfq are not in the same structural position as the Arg–Gly residues of the Sm protein, the similar nature of the residues in this region suggest that the predicted Hfq structure also supports multimer formation through β4–β5′ strand pairing. It is worth noting that the sequence in this part of the Hfq structure, VYKHAIST, is almost completely conserved among Hfq proteins (Fig. 1).

The recently determined *S.aureus* Hfq structure (29) shows that β4–β5′ strand interface is indeed a key part of intersubunit interactions. In this structure, H bonds occur between the highly conserved Tyr56 in β4 and Tyr63 in β5′. In the *E.coli* Hfq sequence valine occurs at position 63. This is the more dominant amino acid at this location in Hfq proteins (Fig. 1) and suggests that Tyr56 H-bonds with a different residue in β5′ in *E.coli* Hfq or this H bond is not essential for this interface. The *S.aureus* Hfq structure also shows that contacts between α helix residues and loop L3 residues of the adjacent subunit and between side chains in β strands contribute to the dimer interface.

The 3D model of Hfq also provides an opportunity to consider its potential sites of interaction with RNA. The RNA determinants important for Sm core assembly appear to be complex. One prerequisite for an RNA to be bound by an Sm protein heptamer is an 'Sm site element', a 7–10 nt single-stranded segment that has the consensus sequence $PuAU_{3-6}GPu$ usually flanked by stem–loop structures (38,39). *In vitro* analysis with an RNA oligonucleotide consisting of a minimal Sm site element revealed that the 5′

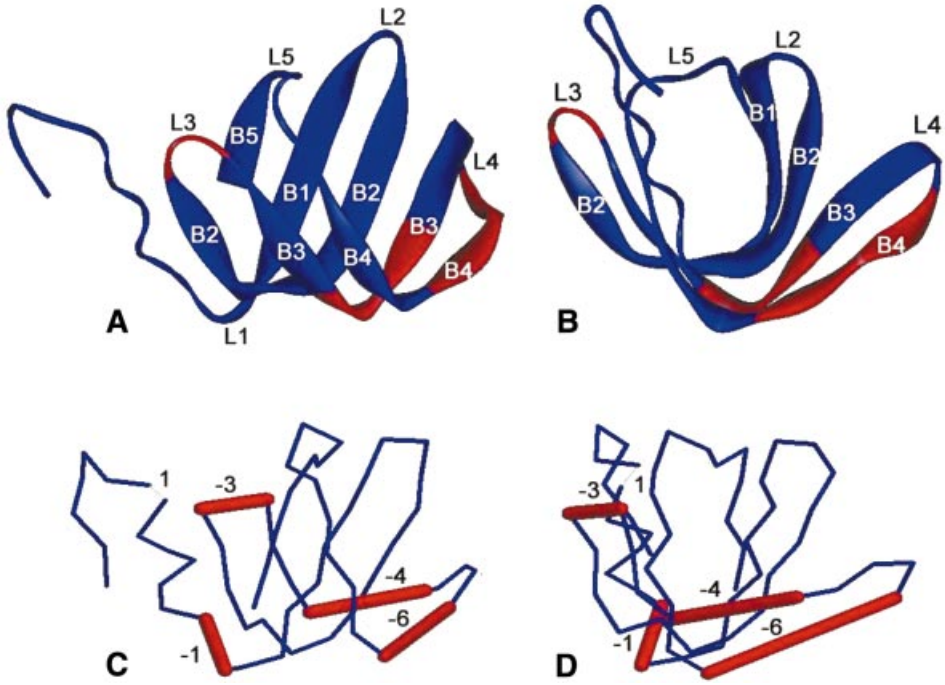

Figure 2. (**A** and **B**) Ribbon representations of two views of the crystal structure of an archaeal Sm protein (PDB accession number 1i81) rotated by 90E. Images are produced by RasMol program (http://www.bernstein-plus-sons.com/software/RasMol_2.7.1). (**C** and **D**) 3D line representations of the Hfq structure predicted by the 3D-PSSM web server using the above archaeal Sm protein as template. The views shown in (C) and (D) are the same as in (A) and (B) respectively. The locations of Hfq residues that are inserted or deleted when compared with the template are represented by thin and thick bars respectively, and accompanied by numbers indicating the number of residues involved. Labels B1–B5 correspond to the β strands β1–β5; labels L1–L5 correspond to the loops.

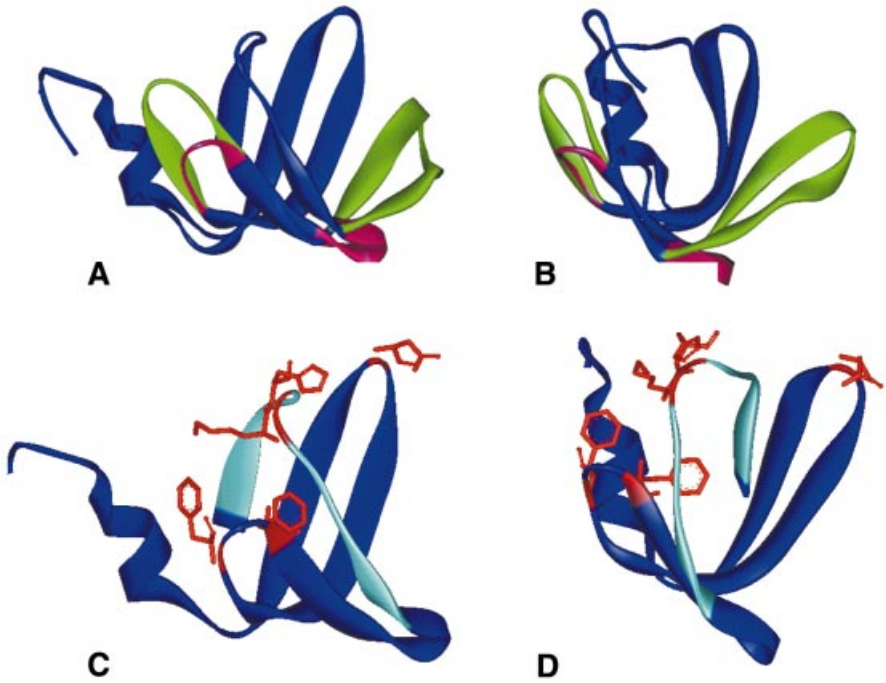

Figure 3. The 3D structure of Hfq generated by SWISS-MODEL program using the same archaeal LSm_{α} protein determined to be the best template by 3D PSSM. (**A** and **B**) Front and side views of the predicted Hfq structure as well as the template Sm structure. The backbone features that are constant in both structures are illustrated in blue. Differing structural elements are shown by using a red ribbon for Hfq and a green ribbon to illustrate the LSm_{α} protein backbone. The aqua ribbon illustrates the β4 strand residues SQMVY and β5 strand residues AISTVV. (**C** and **D**) Front and side views of the predicted Hfq model with several potential RNA-interacting residues shown in stick model representation: Lys31 in loop 2, Phe39 and Phe42 in loop 3, Lys56 and His57 in loop 5.

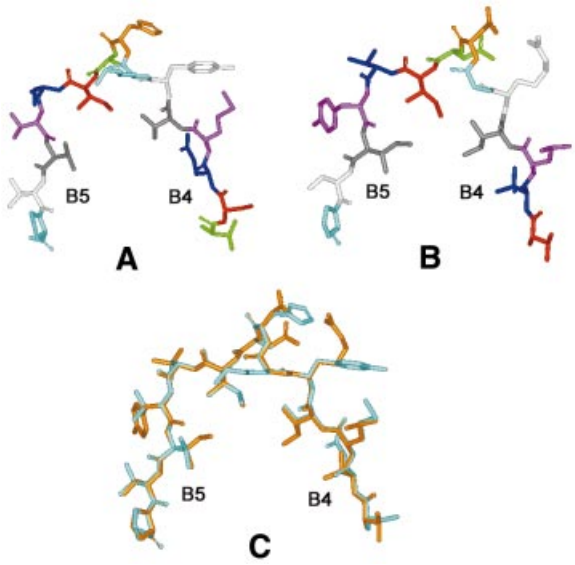

Figure 4. Molecular representation of the β4 and β5 strands in Hfq model (**A**) and LSm$_\alpha$ protein (**B**). Overlapping representation of both is shown in (**C**).

adenosine of the element plays a critical role in the heptamer's association, while the uridine bases and the 2′ hydroxyl groups collectively provide a binding determinant (39,40).

In human snRNP core, several Sm proteins were shown to interact with the uridine stretch of the Sm site element by UV cross-linking experiments. The most efficient cross-links were observed for the G and B/B′ proteins, which are linked to the first and third uridines of the Sm site element respectively (41). The residues (His37 for B/B′, Phe37 for G) involved in contacting the RNA are located at equivalent regions in both proteins, namely in loop L3 of the Sm1 motif. In contrast, crystal structure of the archaeal SmAP protein suggests that residues in other loops (Arg29 in L2, Asp57 in L4 and Glu71 in L5) are more likely to interact with the RNA Sm site element (27). All four of these loops jut into the inner ring or pore of the doughnut-shaped heptamer (Fig. 2). The corresponding regions of Hfq, which by analogy would be expected to be oriented toward the inner ring of the hexamer, also have conserved residues (for the *E.coli* sequence: Lys31 in loop 2, Phe39 and Phe42 in loop 3, Lys56 and His57 in loop 5). The location of some of these residues is illustrated in Figure 3. The Phe39 and Phe42 in loop 3 and Lys56–His57 in loop 5 are almost 100% conserved among different bacterial species examined, implying they have critical roles in structure and function.

Hfq has been shown to be an essential participant in facilitating the interaction of some small riboregulator RNAs, such as DsrA (9) and Spot42 (16), with their target mRNAs. It was proposed that the role of Hfq might be analogous to Rop, in which two phenylalanines intercalate into base pairs and facilitate the pairing of two RNA molecules (42). If Hfq functions in this way, the highly conserved Phe42 in loop 3 and its nearby Phe39 are candidates for this role (Fig. 3C and D).

Several of the above predictions are verified in the recently published crystal structure of the *S.aureus* Hfq–RNA complex (29). In this structure, the backbone of the oligoribonucleotide 5′-AUUUUG-3′ was found to form a circular conformation as it bound to an electropositive patch around one face of the pore of the hexameric Hfq. Residues in the Sm1 and Sm2 motifs of adjacent subunits are utilized to build six nucleotide-binding pockets. There are no intramolecular base stacking interactions within the RNA as the bases are splayed out, fitting into the individual binding pockets. Each base is sandwiched between two Tyr42 side chains from adjacent subunits. The presence of Phe42 instead of Tyr42, which occurs for most Hfq sequences, appears to be able to serve the same function for the nucleotide-binding pockets. The highly conserved Lys–His motif located in loop 5 and facing the pore also contacts the RNA. Lys57 (shifted by one amino acid in *S.aureus* due to an extra residue relative to the *E.coli* sequence) H-bonds with uracil and His58 makes contacts with the phosphate oxygens of one nucleotide as well as the ribose O2′ hydroxyl of the adjacent nucleotide.

Phyletic distribution of the Hfq protein

In order to determine the pervasiveness of the Hfq gene and related Sm proteins in bacteria and explore its evolution, a BLAST search was conducted against bacterial genomes available at the NCBI database. Fifty-eight completed and 82 unfinished bacterial genomes were examined. Approximately half of the bacterial genomes contain at least one gene that codes for an Hfq protein based on strong amino acid sequence similarity to the *E.coli* Hfq sequence, sequence length and amino acid conservation pattern. The presence and absence of the Hfq protein in particular species closely follows recently redefined major bacterial clades (43,44), as shown in Table 1.

Hfq is missing from three high-level bacterial clades: Chlamydia-Spirochaetes, Actinomycetes-Deinococcus-Cyanobacteria and Green sulfur bacteria-Cytophagales. However, it is present in the most deeply branched Thermotogales-Aquificales. The Hfq protein is present in all alpha, beta, gamma and delta proteobacteria except for those that have experienced a massive genome reduction due to their parasitic lifestyle, e.g. *Buchnera* sp. (45), *Rickettsia prowazekii* (46) and *Brucella melitensis* (47). This type of phyletic distribution suggests two possible scenarios for the evolution of the Hfq protein. First, Hfq might be an ancient protein, which was lost early in evolution by major clades and retained only by the lineage leading to proteobacteria. Second, the Hfq protein evolved late in evolution during the separation of the proteobacterial clade and was transferred laterally to several species outside the proteobacteria.

A phylogenetic tree built from the multiple alignment of the Hfq protein sequences is shown in Figure 5. The tree has a topology expected from Table 1. Three major clades of alpha proteobacteria, beta/gamma proteobacteria and low GC Gram-positive bacteria are well defined and supported by bootstrap analysis. A comparison of Hfq sequences from bacteria in the three groups illustrated in Figure 5 is given in Figure 6. The results reveal three positions, at the borders of loop 2 and loop 3, and in loop 4 of the predicted Hfq structure that have residues characteristic to each group. For the low GC Gram-positive bacteria, the dominant residues corresponding to the

Table 1. Presence or absence of Hfq sequence from BLAST search of bacterial genomes

Phyla	Species with Hfq[a]	Species without Hfq
Thermotogales-Aquificales	*Aquifex aeolicus*	
	Thermotoga maritima	
Chlamydia-Spirochetes		*Chlamydia muridarum*
		Chlamydia trachomatis
		Chlamydophlia pneumoniae
		Borrelia burgdorferi
		Treponema pallidum
Green sulfur-Cytophagales		*Cytophaga hutchinsonii*
Actinomycetes-Deinococcales		*Deinococcus radiodurans*
Cyanobacteria		*Mycobacterium leprae*
		Thermobifida fusca
		Streptomyces coelicolor
		Nostoc sp.
		Nostoc punctiforme
		Synechocystis sp.
		Prochlorococcus marinus
		Synechococcus sp
Low GC Gram-positive	*Bacillus halodurans*	*Mycoplasma genitalium*
	Bacillus subtilus	*Mycoplasma pneumoniae*
	Clostridium acetobutylicum	*Mycoplasma pulmonis*
	Clostridium perfringens	*Ureaplasma urealyticum*
	Listeria innocua	*Lactococcus lactis*
	Listeria monocytogenes	*Streptococcus pneumoniae*
	Staphylococcus aureus	*Streptococcus pyogenes*
	Bacillus anthracis(2)	*Enterococcus faecium*
∈-proteobacteria		*Helicobacter pylori*
		Campylobacter jejuni
α-proteobacteria	*Agrobacterium tumefaciens*	*Brucella melitensis*
	Caulobacter crescentus	*Rickettsia conorii*
	Mesorhizobium loti	*Rickettsia prowazekii*
	Sinorhizobium meliloti	
	Magnetococcus sp.	
	Rhodopseudomonas palustris	
	Rhodobacter sphaeroides	
	Magnetospirillum magnetotacticum(2)	
	Novosphingobium aromaticivorans(2)	
β/γ-proteobacteria	*Neisseria meningitidis*	*Buchnera* sp.
	Ralstonia solanocearum	
	Burkholderia fungorum(2)	
	Burkholderia mallei(2)	
	Burkholderia pseudomallei(2)	
	Haemophilus influenza	
	Pasteurella multocida	
	Xylella fastidiosa	
	Nitrosomonas europea	
	Pseudomonas aeruginosa	
	Pseudomonas fluorescens	
	Vibrio cholerae	
	Yersinia pestis	
	Salmonella enterica	
	Escherichia coli	
	Shewanella putrefaciens	

[a]Species with two copies of Hfq sequence per genome are followed by (2).

E.coli sequence positions 30, 43 and 50 are phenylalanine, tyrosine and lysine respectively. The alpha proteobacteria are dominated at these positions by valine, cysteine, and histidine or glutamine, while the beta/gamma proteobacteria universally have isoleucine, valine and valine at these locations. The positions are highlighted in Figure 6. These residues may provide some specificity in the interactions of Hfq with RNA or the interactions governing subunit oligomerization.

Since the above analysis is based on the assumption that the model structure for the *E.coli* Hfq is appropriate for other Hfq proteins, it is worth noting that our predicted structure is in excellent agreement with the recently published crystal structure of the *S.aureus* Hfq (29). We also note that four independent algorithms which utilize a single amino acid sequence as a query predict secondary structures for *S.aureus* Hfq and other relatively distant Hfq sequences (e.g. *Geobacter sulfureducens* Hfq) that closely fit the LSm and consensus Hfq structures (data not shown).

Another outcome from the microbial genome search worth noting was that 6 of the 140 eubacterial genomes examined

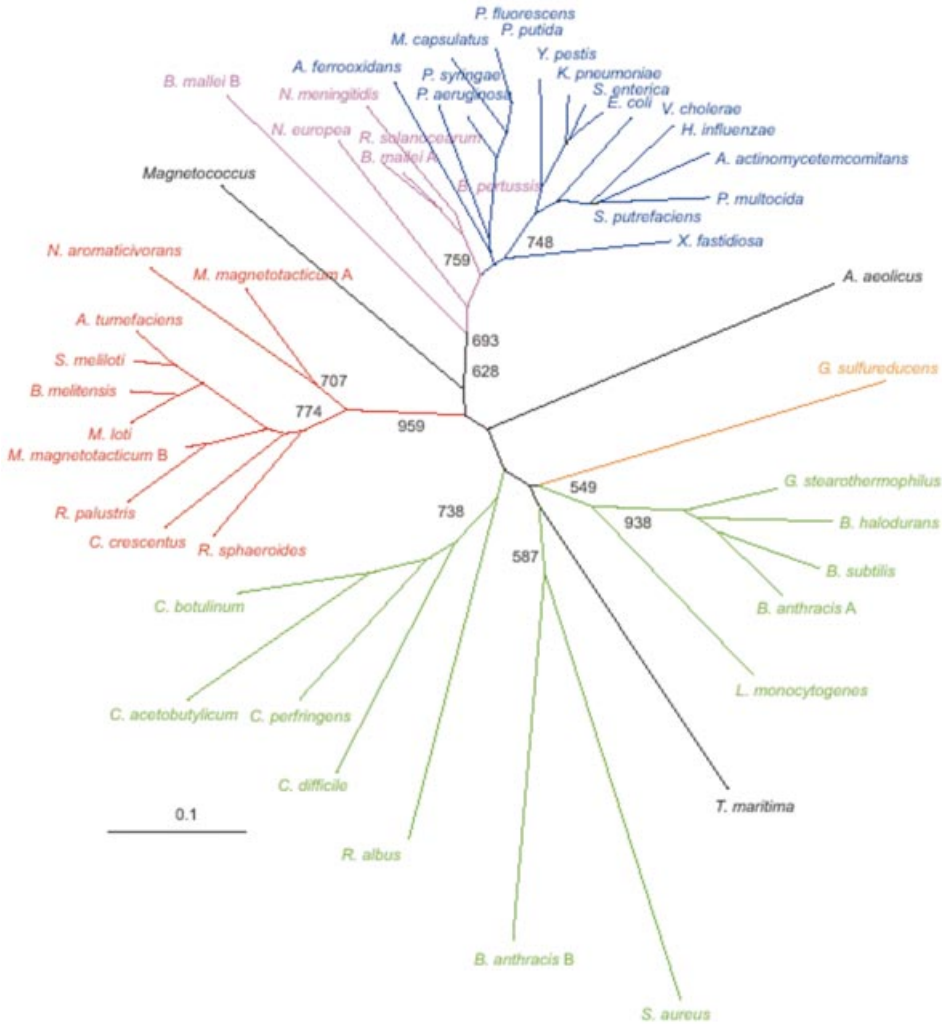

Figure 5. Unrooted neighbor-joining tree inferred by analysis of Hfq protein sequences. Sequences were aligned using CLUSTAL program and all positions with gaps were excluded from the analysis. Bootstrap values of >600 are displayed at deep nodes only. Color code: green, low GC Gram-positive bacteria; red, alpha proteobacteria; purple, beta proteobacteria; blue, gamma proteobacteria; orange, delta proteobacteria. Aquifecales-Thermatogales and unclassified *Magnetococcus* are shown in black.

contained two distinct copies of an Hfq protein coding sequence (Table 1). Duplicated *hfq* genes are always found within the same clade on the phylogenetic tree as original copies (Fig. 5) indicating the likelihood of paralogous relationships over lateral gene transfer. Two Hfq sequences are found within a single 193-residue protein of the bacterium *Novosphingobium aromaticivorans*, which reinforces the notion that Hfq is a subject of relatively frequent gene duplication events. Twenty-three residues separate the two distinct 59 residue Hfq motifs (E values of 3×10^{-14} and 2×10^{-10}) in the *N.aromaticivorans* protein. This protein may code for a heterodimeric version of an Hfq structural unit similar to the heterodimers observed for the eukaryotic Sm proteins.

Our phylogenetic analysis produced no evidence for lateral transfer of Hfq. This is consistent with the proposal that the bacterial (Hfq) and archaeal/eukaryotic (Sm and LSm) versions of this important RNA-binding protein shared a common ancestor prior to the separation of bacteria and archaea–eukarya. Gene loss appears to be a major driving force in the evolution of Hfq.

ACKNOWLEDGEMENTS

This study was supported by an Emory-Georgia Tech Biomedical Center Research Grant (to R.M.W.) and by start-up funds from Georgia Institute of Technology (to I.Z.). We acknowledge the following sequencing centers and their funding agencies for the availability of preliminary data on unfinished microbial genomes: the Institute for Genomic Research (US Department of Energy, US Department of Agriculture, National Institutes of Health), the Joint Genome Institute (US Department of Energy), the Sanger Centre (Beowulf Genomics), University of Oklahoma (National Institutes of Health, National Science Foundation), and Genome Therapeutics.

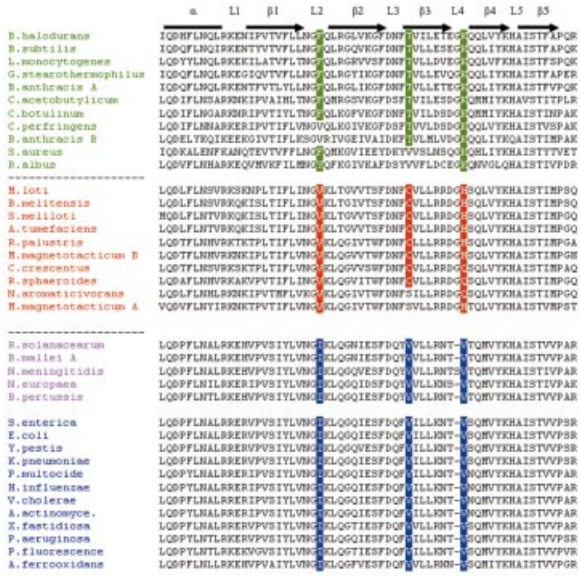

Figure 6. Conserved amino acid residues specific to Hfq proteins from major bacterial groups defined by phylogenetic analysis. Multiple alignment of Hfq sequences is subdivided according to bacterial groups inferred from the tree shown in Figure 5. Positions where amino acid conservation are group specific are shown.

REFERENCES

1. Kajitani,M. and Ishihama,A. (1991) Identification and sequence determination of the host factor gene for bacteriophage Q beta. *Nucleic Acids Res.*, **19**, 1063–1066.
2. Franze de Fernandez,M.T., Eoyang,L. and August,J.T. (1968) Factor fraction required for the synthesis of bacteriophage Q beta-RNA. *Nature*, **219**, 588–590.
3. Carmichael,G.G., Weber,K., Niveleau,A. and Wahba,A.J. (1975) The host factor required for RNA phage Qbeta RNA replication *in vitro*. Intracellular location, quantitation, and purification by polyadenylate-cellulose chromatography. *J. Biol. Chem.*, **250**, 3607–3612.
4. Kajitani,M., Kato,A., Wada,A., Inokuchi,Y. and Ishihama,A. (1994) Regulation of the *Escherichia coli* hfq gene encoding the host factor for phage Q beta. *J. Bacteriol.*, **176**, 531–534.
5. Azam,T.A, Hiraga,S. and Ishihama,A. (2000) Two types of localization of the DNA-binding proteins within the *Escherichia coli* nucleoid. *Genes Cells*, **5**, 613–626.
6. Tsui,H.C., Leung,H.C. and Winkler,M.E. (1994) Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in *Escherichia coli* K-12. *Mol. Microbiol.*, **13**, 35–49.
7. Muffler,A., Fischer,D. and Hengge-Aronis,R. (1996) The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for rpoS translation in *Escherichia coli*. *Genes Dev.*, **10**, 1143–1151.
8. Brown,L. and Elliott,T. (1997) Mutations that increase expression of the rpoS gene and decrease its dependence on hfq function in *Salmonella typhimurium*. *J. Bacteriol.*, **179**, 656–662.
9. Sledjeski,D., Whitman,C. and Zhang,A. (2001) Hfq is necessary for regulation by the untranslated RNA DsrA. *J. Bacteriol.*, **183**, 1997–2005.
10. Tsui,H.C., Feng,G. and Winkler,M.E. (1997) Negative regulation of mutS and mutH repair gene expression by the Hfq and RpoS global regulators of *Escherichia coli* K-12. *J. Bacteriol.*, **179**, 7476–7487.
11. Vytvytska,O., Moll,I., Kaberdin,V.R., von Gabain,A. and Blasi,U. (2000) Hfq (HF1) stimulates ompA mRNA decay by interfering with ribosome binding. *Genes Dev.*, **14**, 1109–1118.
12. Hajnsdorf,E. and Regnier,P. (2000) Host factor Hfq of *Escherichia coli* stimulates elongation of poly(A) tails by poly(A) polymerase I. *Proc. Natl Acad. Sci. USA*, **97**, 1501–1505.
13. Altuvia,S., Zhang,A., Argaman,L., Tiwari,A. and Storz,G. (1998) The *Escherichia coli* OxyS regulatory RNA represses fhlA translation by blocking ribosome binding. *EMBO J.*, **17**, 6069–6075.
14. Majdalani,N., Chen,S., Murrow,K.J., St-John,K. and Gottesman,S. (2001) Regulation of RpoS by a novel small RNA: the characterization of RprA. *Mol. Microbiol.*, **39**, 1382–1394.
15. Zhang,A., Wassarman,K.M., Ortega,J., Steven,A.C. and Storz,G. (2002) The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs. *Mol. Cell*, **9**, 11–22.
16. Möller,T., Franch,T., Højrup,P., Keene,D.R., Bächinger,H.P., Brennan,R.G. and Valentin-Hansen,P. (2002) Hfq: a bacterial Sm-like protein that mediates RNA-RNA interaction. *Mol. Cell*, **9**, 23–30.
17. Lerner,M.R. and Steitz,J.A. (1979) Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc. Natl Acad. Sci. USA*, **76**, 5495–5499.
18. Luhrmann,R., Kastner,B. and Bach,M. (1990) Structure of spliceosomal snRNPs and their role in pre-mRNA splicing. *Biochim. Biophys. Acta*, **1087**, 265–292.
19. Hermann,H., Fabrizio,P., Raker,V.A., Foulaki,K., Horning,H., Brahms,H. and Luhrmann,R. (1995) snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. *EMBO J.*, **14**, 2076–2088.
20. Kambach,C., Walke,S., Young,R., Avis,J.M., de la Fortelle,E., Raker,V.A., Luhrmann,R., Li,J. and Nagai,K. (1999) Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell*, **96**, 375–387.
21. Walke,S., Bragado-Nilsson,E., Seraphin,B. and Nagai,K. (2001) Stoichiometry of the Sm proteins in yeast spliceosomal snRNPs supports the heptamer ring model of the core domain. *J. Mol. Biol.*, **308**, 49–58.
22. Fischer,U., Sumpter,V., Sekine,M., Satoh,T. and Luhrmann,R. (1993) Nucleo-cytoplasmic transport of U snRNPs: definition of a nuclear location signal in the Sm core domain that binds a transport receptor independently of the m3G cap. *EMBO J.*, **12**, 573–583.
23. Seraphin,B. (1995) Sm and Sm-like proteins belong to a large family: identification of proteins of the U6 as well as the U1, U2, U4 and U5 snRNPs. *EMBO J.*, **14**, 2089–2098.
24. Salgado-Garrido,J., Bragado-Nilsson,E., Kandels-Lewis,S. and Seraphin,B. (1999) Sm and Sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J.*, **18**, 3451–3462.
25. Achsel,T., Stark,H. and Luhrmann,R. (2001) The Sm domain is an ancient RNA-binding motif with oligo(U) specificity. *Proc. Natl Acad. Sci. USA*, **98**, 3685–3689.
26. Töro,I., Thore,S., Mayer,C., Basquin,J., Seraphin,B. and Such,D. (2001) RNA binding in an Sm core domain: X-ray structure and functional analysis of an archaeal Sm protein complex. *EMBO J.*, **20**, 2293–2303.
27. Mura,C., Cascio,D., Sawaya,M.R. and Eisenberg,D.S. (2001) The crystal structure of a heptameric archaeal Sm protein: Implications for the eukaryotic snRNP core. *Proc. Natl Acad. Sci. USA*, **98**, 5532–5537.
28. Collins,B.M., Harrop,S.J., Kornfeld,G.D., Dawes,I.W., Curmi,P.M. and Mabbutt,B.C. (2001) Crystal structure of a heptameric Sm-like protein complex from archaea: implications for the structure and evolution of snRNPs. *J. Mol. Biol.*, **309**, 915–923.
29. Schumacher,M.A., Pearson,R.F., Moller,T., Valentin-Hansen,P. and Brennan,R.G. (2002) Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. *EMBO J.*, **21**, 3546–3556.
30. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
31. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
32. Cuff,J.A., Clamp,M.E., Siddiqui,A.S., Finlay,M. and Barton,G.J. (1998) JPred: a consensus secondary structure prediction server. *Bioinformatics*, **14**, 892–893.
33. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
34. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.

35. Sonnleitner,E., Moll,I. and Blasi,U. (2002) Functional replacement of the *Escherichia coli* hfq gene by the homologue of *Pseudomonas aeruginosa*. *Microbiology*, **148**, 883–891.
36. Franze de Fernandez,M.T., Hayward,W.S. and August,J.T. (1972) Bacterial proteins required for replication of phage Q ribonucleic acid. Purification and properties of host factor I, a ribonucleic acid-binding protein. *J. Biol. Chem.*, **247**, 824–831.
37. Kamen,R., Kondo,M., Romer,W. and Weissmann,C. (1972) Reconstitution of Q replicase lacking subunit with protein-synthesis-interference factor i. *Eur. J. Biochem.*, **31**, 44–51.
38. Branlant,C., Krol,A., Ebel,J.P., Lazar,E., Haendler,B. and Jacob,M. (1982) U2 RNA shares a structural domain with U1, U4, and U5 RNAs. *EMBO J.*, **1**, 1259–1265.
39. Raker,V.A., Hartmuth,K., Kastner,B. and Luhrmann,R. (1999) Spliceosomal U snRNP core assembly: Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner. *Mol. Cell. Biol.*, **19**, 6554–6565.
40. Hartmuth,K., Raker,V.A., Huber,J., Branlant,C. and Luhrmann,R. (1998) An unusual chemical reactivity of Sm site adenosines strongly correlates with proper assembly of core U snRNP particles. *J. Mol. Biol.*, **285**, 133–147.
41. Urlaub,H., Raker,V.A., Kostka,S. and Luhrmann,R. (2001) Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. *EMBO J.*, **20**, 187–196.
42. Predki,P.F., Nayak,M., Gottlieb,M.B. and Regan,L. (1995) Dissecting RNA-protein interactions: RNA-RNA recognition by Rop. *Cell*, **80**, 41–50.
43. Wolf,Y.I., Rogozin,I.B., Grishin,N.V., Tatusoy,R.L. and Koonin,E.V. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, **1**, 8–22.
44. Brochier,C., Bapteste,E., Moreira,D. and Philippe,H. (2002) Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.*, **18**, 1–5.
45. Gil,R., Sabater-Munoz,B., Latorre,A., Silva,F.J. and Moya,A. (2002) Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. *Proc. Natl Acad. Sci. USA*, **99**, 4454–4458.
46. Andersson,S.G., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C., Podowski,R.M., Naslund,A.K., Eriksson,A.S., Winkler,H.H. and Kurland,C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**, 133–140.
47. DelVecchio,V.G., Kapatral,V., Redkar,R.J., Patra,G., Mujer,C., Los,T., Ivanova,N., Anderson,I., Bhattacharyya,A., Lykidis,A., Reznik,G., Jablonski,L., Larsen,N., D'Souza,M., Bernal,A., Mazur,M., Goltsman,E., Selkov,E., Elzer,P.H., Hagius,S., O'Callaghan,D., Letesson,J.J., Haselkorn,R., Kyrpides,N. and Overbeek,R. (2002) The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. *Proc. Natl Acad. Sci. USA*, **99**, 1–3.