Prediction of Hfq in actinobacteria

Nagamani Bora[1], Alan C Ward[2,3] and Wonyong Kim[2*]

Affiliations:

[1] School of BioSciences, Nottingham University, LE12 5RD, UK

[2] Department of Microbiology, College of Medicine, Chung-Ang University, Seoul 156-756, South Korea

[3] School of Biology, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

Email: kimwy@cau.ac.kr

Abstract

Hfq is the bacterial orthologue of the eukaryotic (L)Sm family of proteins found across all domains of life and potentially an ancient protein, but it has not been found in all phyletic lines. A careful search successfully identified a distant hfq orthologue in the cyanobacteria leaving the actinobacteria as the major phylum with no known hfq orthologue. A search for hfq in actinobacteria, using domain enhanced searching (DELTA-BLAST) with cyanobacterial hfq, identified a conserved actinobacterial specific protein as remotely homologous. Structural homology modelling using profile matching to fold libraries and *ab initio* 3D structure determination supports this prediction and suggests module shuffling in the evolution of the actinobacterial hfq. Our results provide the basis to explore this prediction, and exploit it, across diverse taxa with potentially important post-transcriptional regulatory effects in virulence, antibiotic production and interactions in human microbiomes. However, the role of hfq in gram positive bacteria has remained elusive and experimental verification will be challenging.

Introduction

Hfq is the bacterial orthologue of the eukaryotic (L)Sm family of proteins, responsible for multiple RNA processing activities[1-3] such as RNA splicing, nuclear RNA processing and mRNA decay. The simpler hfq hexameric homopolymers provide models to understand the more complex eukaryotic pentameric heteropolymers[2]. In prokaryotes hfq[4,5] is involved in post-transcriptional regulation[6] acting as an RNA chaperone in facilitating interactions between small non-coding RNAs and mRNA in processes like stress responses, quorum sensing and virulence[7].

The (L)Sm family of proteins, found across all domains of life, is potentially an ancient protein, but it has not been found in all phyletic lines[4]. A careful search successfully identified a distant hfq orthologue in the cyanobacteria[8,9] leaving its sister clade, the actinobacteria[10,11] , with major pathogens, human commensals and industrially important species, as the major phyletic line with no hfq orthologue.

Here we show that new blast analyses of cyanobacterial hfq against actinobacterial protein sequences predict a conserved actinobacterial specific protein as distantly homologous. Secondary and tertiary structure prediction supports this with structural homology and suggests module shuffling in its evolutionary divergence. An hfq orthologue has significance for regulatory mechanisms in the corynebacteria, mycobacteria, streptomycetes and other actinobcteria.

Our results provide the basis to explore this prediction, and exploit it, across diverse taxa, in: amino acid production; nitrogen fixation; virulence; antibiotic production; and interactions in human microbiomes. Increasingly sophisticated protein structure prediction tools[12] should uncover the roles of more small, taxon specific proteins of unknown function[13].

Small non-coding RNAs

Small non-coding RNAs (ncRNA) exert post-transcriptional regulation enabling rapid response to changing environmental conditions. *Cis*-antisense RNAs, transcribed from the strand opposite the target gene are complementary, and specific, to their target mRNA. *Trans*-encoded small RNAs, from intergenic regions, are weakly complementary to multiple mRNA targets, requiring hfq to facilitate interactions.

These regulatory RNAs are found across prokaryote diversity[6]. This regulatory mechanism is expected to be important in the actinobacteria. *Streptomyces* exhibit a wide range of complex regulatory processes to cope with rapidly changing environmental conditions, and diverse ncRNAs have been detected[14]. They are likely to be important tools in genetic engineering and synthetic biology[15,16]. Many *cis*- and *trans*- ncRNAs have been found in the work horse of the fine chemical industry, *Corynebacterium glutamicum*[17,18] and their genome-wide expression demonstrated in *Mycobacterium tuberculosis*[19,20].

Hfq-binding small non-coding RNAs

Hfq recognises structurally diverse ncRNAs and facilitates their interaction with their partially complementary mRNA targets[1]. Hfq is the only bacterial orthologue of the (L)Sm protein superfamily. Sm proteins show a conserved Sm fold with an N-terminal α-helix followed by a twisted 5-stranded antiparallel β-pleated-sheet, with conserved motifs, Sm1 and Sm2. In hfq Sm1 is partly conserved with a different Sm2 motif, but structural studies reveal the distinct Sm fold.

A single ncRNA targeting the expression of several genes and operons can be tuned to synchronize a coordinated response to stressful environmental conditions. Such a multi-target activity requires specificity and coordinated regulation which is facilitated by hfq, which is the only factor in bacteria described as facilitating the interaction between ncRNAs and their target mRNAs[5].

Searching for missing Hfq orthologues

The view that blast searching might not find all hfq orthologues was supported by the failure to find hfq candidates in two major clades, the actinobacteria and cyanobacteria[4]. However, a careful search using sequence length, pattern and motifs successfully identified an ORF in the *Anabaena* PCC 7120 genome as an Hfq orthologue[8]. Blast then identified many putative homologues in unicellular and filamentous cyanobacteria and *Prochlorococcus*. The weak sequence homology was bolstered by determining the 3D structure[9] of the *Synechocystis* sp. PCC 6803 (ssr3341 - 3HFO) and *Anabaena* PCC 7120 (asl2047 – 3HFN) hfq and their structural homology to the *Escherichia coli* hfq.

Results

BLAST

The actinobacteria are a sister clade to the cyanobacteria in whole organism phylogenies[10,11]. Blastp of the cyanobacterial hfq sequences against the actinobacteria found few hits with low E-values and only 2 are short (*Mycobacterium vaccae* Mvac_26751 E-value 2.1 and *Mycobacterium vanbaalenii* Mvan_1794 E-value 7.8). Blastp of the *Escherichia coli* hfq against cyanobacteria also finds few hits with E-values above the threshold, 2 are short (*Fischerella muscicola* WP_016860369.1 E-value 0.090; *Synechococcus* sp. CB0205 WP_010317051 E-value 4.1) both now annotated as hfq.

DELTA-BLAST[21] detects remote protein homologs using domain enhanced searching and found 67 hfq matches to the *E. coli* hfq sequence in the cyanobacteria, only the last 3 are not short proteins. The top hit was the *Fischerella muscicola* hfq with an E-value of 2e-08.

DELTA-BLAST of the *Synechocystis* sp. PCC 6803 hfq against the actinobacteria matched 93 short proteins in the conserved hypothetical protein pfam family DUF3107 and aligns the DUF3107 conserved domain with the hfq and Sm-like domains in *Synechocystis* sp. PCC 6803 hfq. The

top hit is to *Mycobacterium vanbaaleni* Mvan_1794 (E-value 2e-05). DELTA-BLAST of *E. coli* hfq against the actinobacteria found no hits.

DELTA-BLAST of the *Synechocystis* sp. PCC 6803 hfq against the whole nr database found 2346 hits above the default threshold. The DELTA-BLAST distance tree of the results, downloaded in Newick format and displayed in Dendroscope[22] is shown in Figure 1.

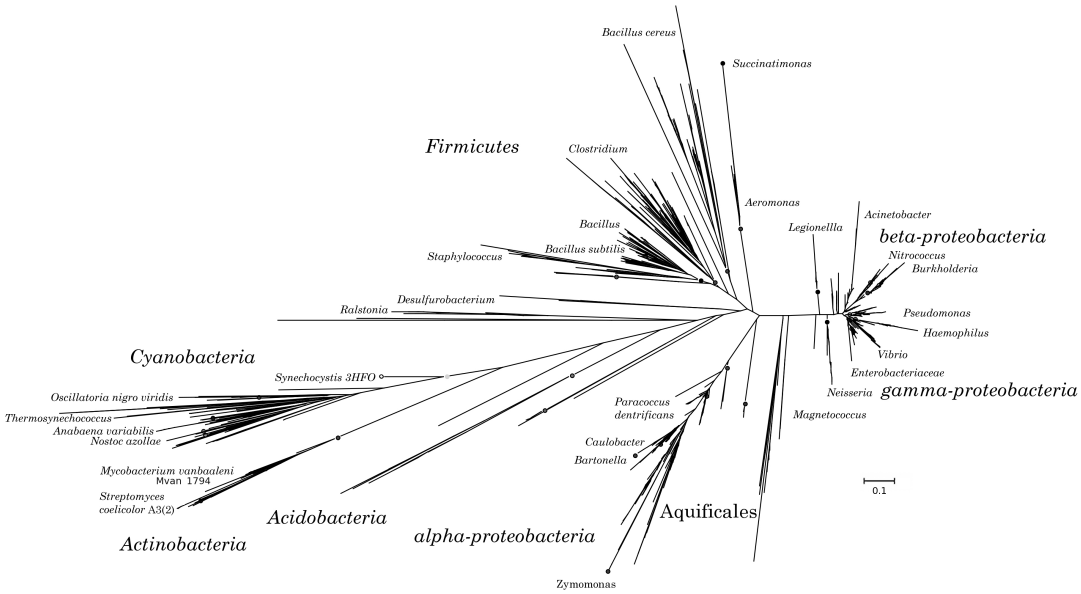

Figure 1. Phylogenetic tree of protein distances of significant matches in DELTA-BLAST to *Synechocystis* hfq PCC 6803 (distances for each hit are the distance from *Synechocystis* hfq)[21].

A full, multiple sequence alignment[23] and maximum likelihood tree[24] displays a more distant relationship of the actinobacterial protein sequences (Supplementary data S1).

Structural Homology

Predicting protein tertiary structure from sequence data is still challenging[12,25,26] but *ab initio* prediction is most successful for small proteins. The Phyre2 web server[12] generates a sequence profile of the submitted protein sequence, predicts the secondary structure (Figure 2A) and matches the profile to the profiles in a fold library.

The secondary structure prediction for Mvan_1794 does not show an N-terminal α-helix followed by 5-strands of β-pleated-sheet as found in hfq. However the top hit for *Mycobacterium vanbaaleni* Mvan_1794 in the Phyre2 fold library was to an Sm-like fold (Sm2) in hfq (3HFO-A) from *Synechocystis* PCC 6803 (Figure 2B). *Mycobacterium vaccae* MVAC_26751 and *Streptomyces coelicolor* A3(2) SCO5169, DUF3107 proteins with significant DELTA-BLAST matches to *Synechocystis* hfq, and *Corynebacterium glutamicum* ATCC 13032 Cgl0772 and *Bifidobacterium bifidum* ATCC 29521 BIFBIF00292, DUF3107 proteins not matched by DELTA-BLAST of the *Synechocystis* sp. PCC 6803 hfq (Supplemental data S2), returned the same hit in Phyre2.

The secondary structure of the DUF3107 sequences is 2 strands of β-pleated sheet followed by an α-helix and 4 β-pleated sheets (Figure 2A). In the 3D structure of 3HFO-A β2 is twisted, so that β2i forms an anti-parallel sheet with β1, while β2ii aligns antiparallel to β3 in the β3/β4 sheet. Threading amino acids 19-83 of *Mycobacterium vanbaaleni* Mvan_1794, which includes the α-helix and succeeding β-pleated sheets, onto the template structure of *Synechocystis* 3HFO-A matches the α-helix and β-pleated sheet β3/β4, β5 of 3HFO-A (Figure 2C). The secondary structure of the DUF3107 sequences, including *Mycobacterium vanbaaleni*, and one to one threading (Figure 2) indicate that protein modules (α-helix, β1/β2i, β2ii/β3/β4, β5 and C-terminal tail) are shuffled relative to other eubacterial hfq, as β1/β2i, α1, β2ii/β3/β4, β5 and C-terminal (Figure 2D). If protein modules are identified based upon homology modelling, and the amino acids corresponding to β1/β2i in *Mycobacterium vanbaaleni* Mvan_1794 are re-ordered to follow α1, the re-ordered primary sequence threads onto the *Synechocystis* hfq sequence (Figure 2D).

The tree from the DELTA-BLAST domain-based distances (Figure 1) shows the actinobacterial sequences as a sister clade to the cyanobacterial hfq. Multiple sequence alignment and the maximum likelihood

phylogenetic tree (Supplementary data S1) show the actinobacterial sequences as significantly more distant.

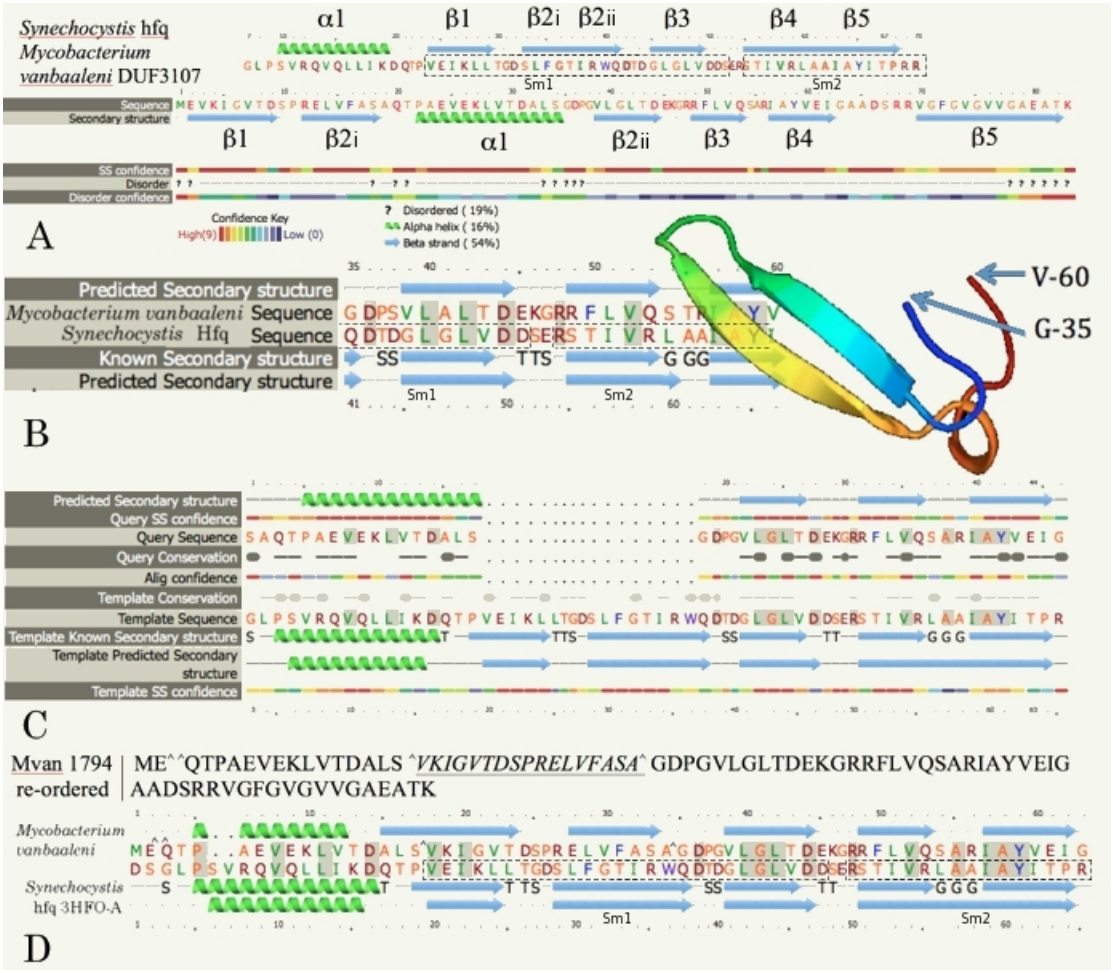

Figure 2. (A) Comparison of predicted secondary structure of *Synechocystis* sp. PCC 6803 Hfq. and *Mycobacterium vanbaalenii* DUF 3107 protein (Mvan_1794) from Phyre2[12]. (B) Homology modelling of the Sm-like fold in *Mycobacterium vanbaaleni* DUF3107 protein Mvan_1794 in Phyre2 to template *Synechocystis* 3HFO-C (C) One to one threading of amino acids 19-83 of *Mycobacterium vanbaaleni* Mvan_1794 against *Synechocystis* sp. PCC 6803 hfq (Phyre2 Expert mode) (D) One to one threading of the re-ordered *Mycobacterium vanbaaleni* Mvan_1794 sequence against PDB 3HFO-A, hfq from *Synechocystis* sp. PCC 6803 (secondary structure: top is actual; bottom is predicted). Mvan_1794 re-ordering: Sequence cut ^^ ^ *SEQ INSERT*^. ------ Sm1 and Sm2 in *Synechocystis* sp. PCC 6803 hfq.

Linear alignment of the sequences cannot align the shuffled homologous regions, the sequence in the order β1/β2i, α1 in DUF3107 does not align with the sequence for α1, β1/β2 in 3HFO-A. The same actinobacterial primary sequences, re-ordered to match the re-arranged Mvan_1794 template (Figure 2D), align to generate a maximum likelihood tree topology (Figure 3) similar to DELTA-BLAST (Figure 1).

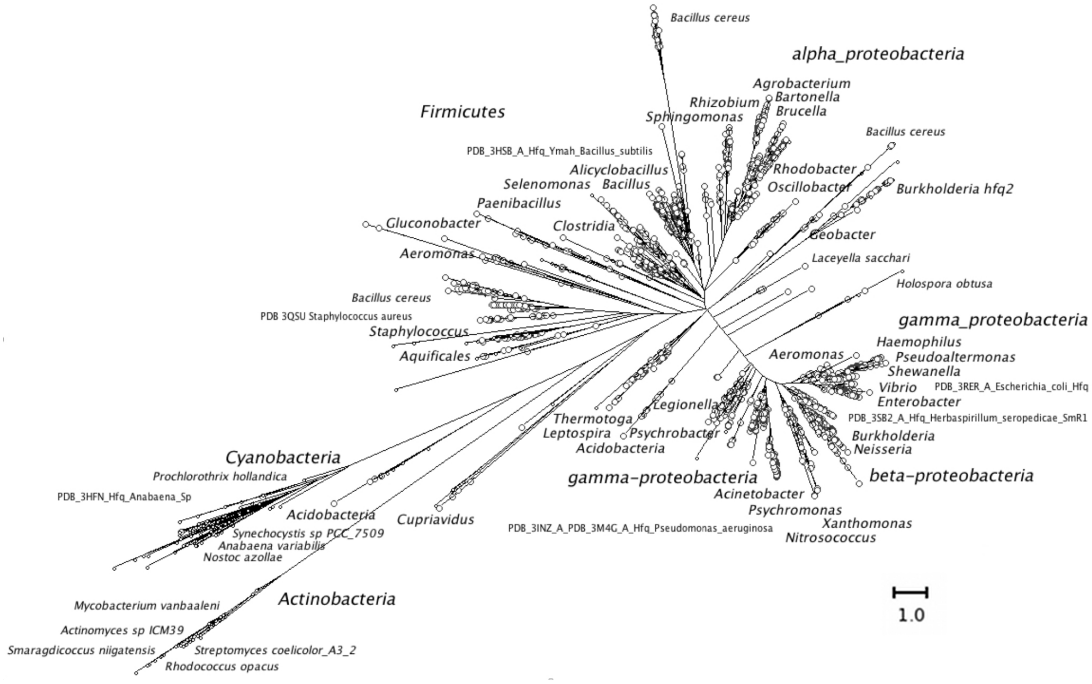

Figure 3. Maximum likelihood phylogenetic tree based upon multiple sequence alignment of the protein sequences retrieved using DELTA-BLAST of the *Synechocystis* sp. PCC 6803 Hfq. Aligned with the re-ordered actinobacterial sequences. ◦ sequences ○ sequences annotated as hfq, host-factor 1 or Sm-like.

Ab initio modelling by Phyre2[12] of the *Mycobacterium vanbaaleni* Mvan_1794 and *Corynebacterium glutamicum* ATCC 13032 Cgl0772 protein sequences only recovers fragments of the complete structure (Figure 2B). However, *ab initio* modelling by QUARK[26] recovers complete 3D structure predictions, with modules that correspond to the α1, β1/β2i, β2ii/β3/β4 and β5 found in *Synechocystis* sp. PCC 6803 hfq (Figure 4 and Figure S4, supplemental data).

The best full model (Mvan_model3.pdb) aligns with the α1 and β2ii/β3/β4 modules in the *Synechocystis* sp. PCC 6803 hfq (Figure 4) with Tm-align[27].

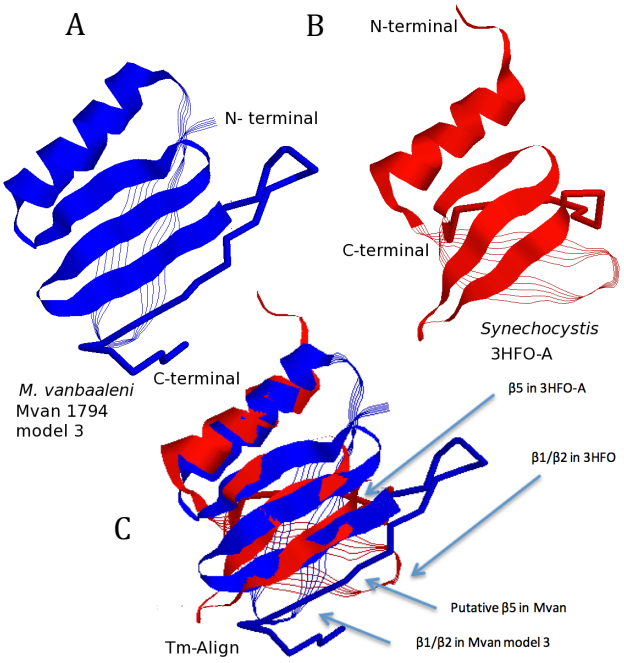

Figure 4. Comparison of the tertiary structures for (A) *Mycobacterium vanbaaleni* Mvan_1794 DUF3107 (Mvan_model3.pdb) (B) *Synechocystis* sp. PCC 6803 hfq (C) Aligned structures (Tm-align[27]). Aligned structures (α1 and β2ii/β3/β4) are displayed as ribbon. The β1/β2i sheet as strands and β5 in 3HFO-A, and the putative β5 and extended C-terminal in Mvan_1794, as backbone.

In Figure 4 the β1/β2i sheet of Mvan_1794 does not thread onto 3HFO-A and β5 forms antiparallel β-pleated sheet with β4 while in 3HFO-A β5 forms antiparallel β-pleated sheet with β1. The re-ordered sequence (2D) does thread the β1/β2i sheet of Mvan_1794 onto the corresponding structure in 3HFO-A (Figure S4, supplementary data).

If the predicted structure for *Mycobacterium vanbaaleni* Mvan_1794 is oriented (Figure 5A) to match the *Synechocystis* sp. PCC 6803 hfq (Figure 5D) structure, it can be assembled to match the full hexamer for 3HFO (Figure 5C and E), similar to the structure for *Salmonella enterica* hfq in[1]. In the 3HFO hexamer β5 forms antiparallel β-pleated sheet with both β1,

in its own subunit, and β4, in the adjacent subunit, forming the bonds holding the hexamer together.

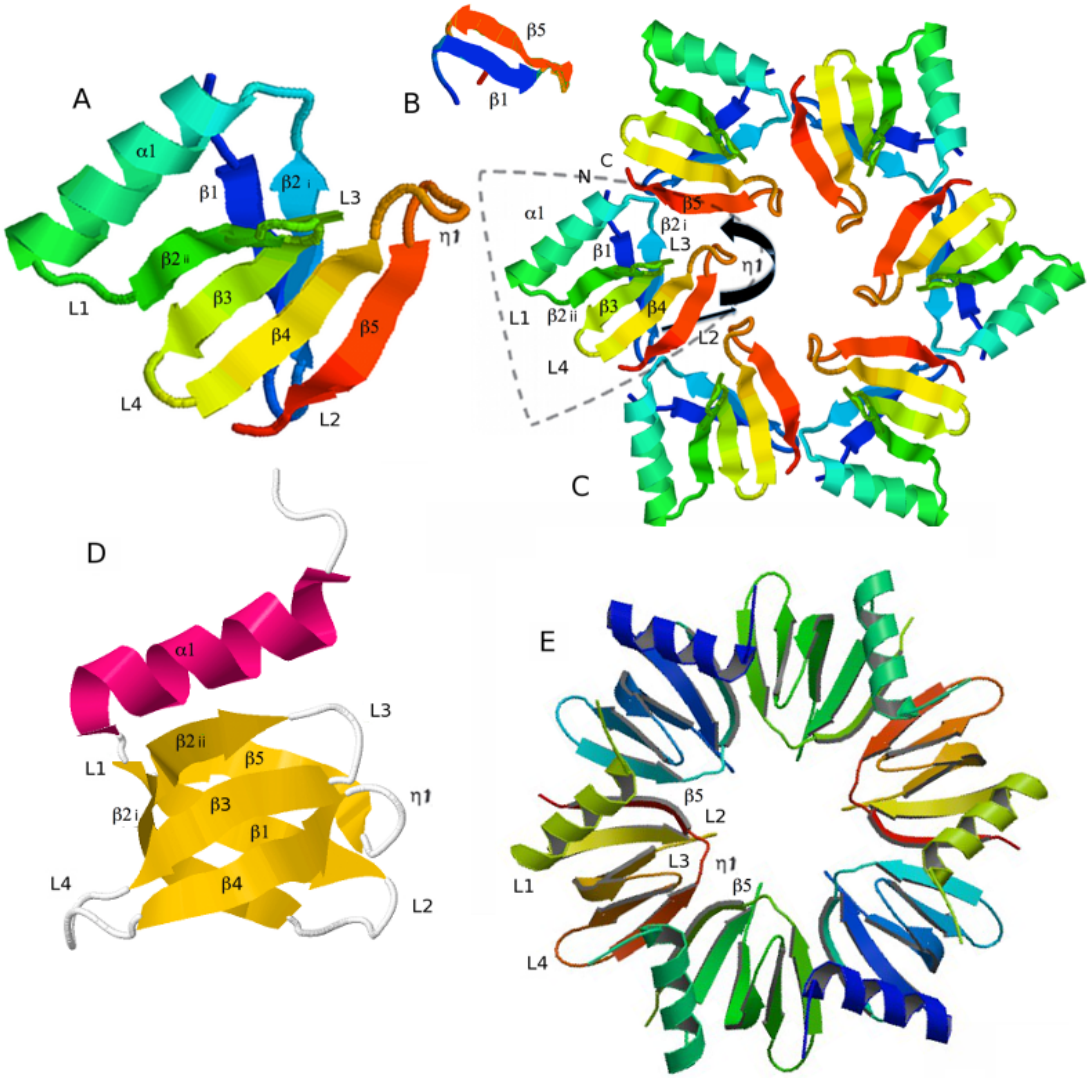

Figure 5. Comparison of *Mycobacterium vanbaaleni* Mvan_1794 proposed hexamer with the *Synechocystis* sp. PCC 6803 hexameric hfq.
(A) predicted structure of the monomer protein of Mvan_1794 (B) - structure prediction of Mvan_1794 β1/β5 with loop η1 as linker (C) – proposed hexamer sof Mvan_1794 protein subunits. Curved arrow - proposed re-alignment of β5, straight arrow - shift of β1/β2 to interact with re-aligned β5 as modelled in (B). (D) - structure of *Synechocystis* sp. PCC 6803 hfq subunit (E) - *Synechocystis* sp. PCC 6803 hexamer.

In the model structure for *Mycobacterium vanbaaleni* Mvan_1794 the structure prediction is based upon a single polypeptide chain, and the

strand proposed as the structural homologue of β5 forms anti-parallel β-pleated sheet with β4 (Figures 5A and 4). Modelling the polypeptide for *Synechocystis* sp. PCC 6803 hfq (3HFO-A) sometimes recovers the same structure in models (data not shown). In the full crystallographic structure for 3HFO β5 forms antiparallel β-pleated sheet with β1, but forms an antiparallel sheet with β4 in the adjacent subunit (Figure 4) forming the inter-subunit bonds. The β5 strand in Mvan_1794 occupies the same position as β5 strands in 3HFO, but shifted one subunit, and aligns to β4 in the same chain (Figure 5C and E). In each Mvan_1794 chain β5 could re-orientate to form β-pleated sheet with β1 (Figure 5B) and align with β4, as it currently does, but in the adjacent subunit. The sequences from β1 and β5 from the putative *Mycobacterium vanbaaleni* hfq subunit, joined with a linker polypeptide from loop η1, are modelled by Quark as β-pleated sheet (Figure 5B). This would pull β1/β2i over to move loop L2 into an even more similar position to L2 in 3HFO.

RNA binding

The potential RNA binding residues determined from the primary sequence by BindN[29] are shown in Figure S3 and compared with the RNA binding sites on *Salmonella enterica*[1] and *Synechocystis* sp. PCC 6803 hfq[9] (Supplementary data S3).

SPOT-seq[30], template-based RNA-binding detection, was more accurate than other sequence or structure-based methods tested in[31] and predicts hfq as an RNA-binding protein (z-scores of 18.5/18.3 for *Bacillus subtilis* and *Escherichia coli* hfq). The *Synechocystis* sp. PCC 6803 hfq and *Mycobacterium vanbaaleni* Mvan_1794, are not predicted as RNA-binding proteins but align with 3/5 of the same templates, but with lower z-scores (~11 and 5 respectively).

Conclusion

Bacterial whole genomes are full of small, hypothetical and often taxon specific genes, originally seen as potential artefacts, but now frequently,

frustatingly, enigmatic. Of 29 signature proteins uniquely defining the actinobacteria[32] only 2 were not annotated as hypothetical genes of unknown function. Successful searches for the function of such taxon specific hypothetical proteins e.g. Yeats *et al.* (2003)[33], even with new tools and exponentially increasing databases, is slow.

One actinobacterial signature protein, represented by ML0814 (NP_301620) from *Mycobacterium leprae* in[32], is a protein of unknown function containing the DUF3107 domain, present across the actinobacteria. BLAST or DELTA-BLAST of members of the DUF3107 family does not detect significant homology to non-actinobacterial proteins. A subset (93) of the DUF3107 family of proteins (>475) are identified as homologous to hfq by DELTA-BLAST of *Synechocystis* sp. PCC 6803 Hfq (Figure 1) against actinobacterial proteins.

The putative hfq orthologue in the actinobacteria has diverged in both sequence homology and module order, making identification[32,33] as an Sm-like RNA-binding protein problematic. The determination of structural homology[9] has been important in demonstrating the orthology and function of the cyanobacterial hfq. The phylogenetic relationship of the actinobacteria and cyanobacteria in the clade described as the "Terrabacter" by Battistuzzi *et al.*[34] suggests the phylogeny guided search strategy used in this paper. The extent of the divergence of the putative hfq orthologue in actinobacteria, with a reorganised fold (Figure 5A and B), may indicate that more orfans may be divergent examples of functional proteins known in other taxa. But new tools and the detection of structural homology, with advances in protein structure prediction, may be needed for their identification. DELTA-BLAST looks like a useful tool for the putative identification of such partial homology. Module shuffling is common in higher eukaryotes, through mRNA isoforms with alternative exons, but is rarely documented in prokaryotes[35].

The detection of a putative actinobacterial hfq orthologue opens new avenues for research in pathogens like mycobacteria[19], the industrially

important corynebacteria[36] and streptomycetes[37]. However, although hfq is a key regulatory molecule in the proteobacteria[5], its role in gram positive bacteria is less clear. An *hfq* deletion mutant of *Bacillus subtilis* showed no phenotypic changes over a wide range of growth conditions, except in survival in stationary phase in rich media[38]. It is highly conserved in all sequenced *B. subtilis* strains[38] so survival may be a powerful selective force! Transcriptomic analysis detected changes in the levels of over 100 transcripts[38] many linked to sporulation (although the strain used in the study was non-sporulating so these changes did not explain the stationary phase phenotype). Even a specialised role in stationary phase post-transcriptional regulation may be important in actinobacteria, for example in survival of latent pathogenic mycobacteria or secondary metabolism in actinomycetes. The *B. subtilis* hfq complemented the lack of hfq in *Salmonella* for only one hfq-dependent regulatory activity[38]. And the hfq from *Synechocystis* sp. PCC 6803 showed altered RNA binding[9] and very limited ability to bind to *Escherichia coli* Hfq target RNAs *in vitro*. So the elusive role of DUF3107 proteins[33] may be difficult to pin down experimentally.

Methods

Blast analysis

Blast and DELTA-BLAST analysis of the *Synechocystis* sp. PCC 6803 (3HFO_A) and *Anabaena* sp. PCC 7120 (3HFN_A) hfq sequences was performed at NCBI using blastp and DELTA-BLAST, with default parameters, except for increased maximum sequence hits returned. Data accessed February 2014.

Multiple sequence alignment

Sequences were imported into SeaView[23] and aligned using Muscle[39]. Maximum likelihood trees were generated in FastTree2[24] with the -gamma option and visualised in Dendroscope[22]. Trees were imported into SeaView and sequences ordered to follow tree order, then manually re-aligned in SeaView following iterative tree generation and sequence

ordering. All gaps inserted into variable C terminal tails were removed and matching tails manually aligned. Manual re-alignment was monitored with the cat20 gamma likelihood from FastTree2 using the -gamma option.

Structural homology

Protein sequences were submitted to Phyre2[12] and Quark[26] for *ab initio* protein prediction. Models were aligned with reference structures using one to one threading in expert mode in Phyre2 and using Tm-align[27]. Structures were displayed in Rasmol 2.7.5.2 (http://rasmol.org/) and model quality assessed in PROQ2[28].

RNA binding

Potential RNA binding sites were identified with BindN[29] and SPOT-seq[30].

Acknowledgements

The authors would like to acknowledge the powerful tools and data made available for *in silico* analysis, including DELTA-BLAST, PHYRE2[12] and Quark[26] and useful comments from the authors of PHYRE2[12] and Quark[26].

References

1. Sauer, E. Structure and RNA-binding properties of the bacterial LSm protein Hfq. *RNA Biol* **10**, 582-90 (2013).

2. Wilusz, C.J. & Wilusz, J. Eukaryotic L(Sm) proteins: lessons from bacteria. *Nat Struct Mol Biol* **12**, 1031-6 (2005).

3. Wilusz, C.J. & Wilusz, J. L(Sm) proteins and Hfq: Life at the 3′ end. *RNA Biol* **10**, 564-73 (2013).

4. Sun, X., Zhulin, I. & Wartell, R.M. Predicted structure and phyletic distribution of the RNA-binding protein hfq. *Nucl Acid Res* **30**, 3662-3671 (2002).

5. Sobrero, P. & Valverde, C. The bacterial protein Hfq: much more than a mere RNA-binding factor. *Crit Rev Microbiol* **38**, 276 -99 (2012).

6. Waters, L.S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136**, 615–

628 (2009)

7. Gripenland, J., Netterling, S., Loh, E., Tiensuu, T., Toledo-Arana, A. & Johansson, J. RNAs: regulators of bacterial virulence. *Nature Rev Microbiol* **8**, 857-866 (2010)

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

9. Bøggild, A., Overgaard, M., Valentin-Hansen, P. & Brodersen, D.E. Cyanobacteria contain a structural homologue of the Hfq protein with altered RNA-binding properties. *FEBS Journal* **276**, 3904–3915 (2009).

10. Ward, A.C. & Bora, N. The Actinobacteria. *In* Practical Handbook of Microbiology, Second Edition (Edited by Goldman E Green LH) Chapter 27 pp. 375-444, CRC Press (2008).

11. Lang, J.M., Darling, A.E. & Eisen, J.A. Phylogeny of bacterial and archaeal genomes using conserved Genes: Supertrees and Supermatrices. *PLoS ONE* **8**, e62510 (2013).

12. Kelley, L.A. & Sternberg, M.J.E. Protein structure prediction on the web: a case study using the Phyre server. *Nature Protocols* **4**, 363 – 371 (2009).

13. Tautz, D & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nature Rev Genet* **12**, 692-702 (2011)

14. Moody, M.J., Young, R.A., Jones, S.E. & Elliot, M.A. Comparative analysis of non-coding RNAs in the antibiotic-producing *Streptomyces* bacteria. *BMC Genomics* **14**, 558 (2013).

15. Medema, M.H., Alam, M.T., Breitling, R. & Takano, E. The future of industrial antibiotic production: From random mutagenesis to synthetic biology. *Bioeng Bugs* **2**, 230–233 (2011).

16. Uguru, G.C., Mondhe, M., Goh, S., Hesketh, A., Bibb, M.J., Good, L. & Stach, J.E.M. Synthetic RNA Silencing of Actinorhodin Biosynthesis in

Streptomyces coelicolor A3(2). *PLoS ONE* **8**, e67509 (2013).

17. Zemanová, M., Kaderábkovà, P., Pátek, M., Knoppová, M., Šilar, R. & Nešvera, J. Chromosomally encoded small antisense RNA in *Corynebacterium glutamicum*. *FEMS Microbiol Lett* **279**, 195–201 (2008).

18. Mentz, A, Neshat, A, Pfeifer-Sancar, K, Puḧler, A, Ruc̈kert, C & Kalinowski, J. Comprehensive discovery and characterization of small RNAs in *Corynebacterium glutamicum* ATCC 13032. *BMC Genomics* **14**, 714 (2013).

19. Miotto, P., Forti, F., Ambrosi, A., Pellin, D., Veiga, D.F., Balazsi, G., Gennaro, M.L., Serio, C.D., Ghisotti, D., Cirillo, D.M. Genome-Wide Discovery of Small RNAs in *Mycobacterium tuberculosis*. *PLoS ONE* **7**, e51950 (2012).

20. Li S, Ng PK, Qin H, Lau JK, Lau JP, Tsui SK, Chan T & Lau TC Identification of small RNAs in *Mycobacterium smegmatis* using heterologous Hfq. *RNA* **19**, 74–84 (2013)

21. Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J. & Madden, T.I. Domain enhanced lookup time accelerated BLAST. *Biology Direct* **7**, 12 (2012).

22. Huson, D., Richter, D., Rausch, C., Dezulian, T., Franz, M. & Rupp, R. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**, 460 (2007).

23. Gouy, M., Guindon, S. & Gascuel, O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol Biol Evol* **27**, 221-224 (2010).

24. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).

25. Kryshtafovych, A., Fidelis, K. & Moult, J. CASP9 results compared to those of previous casp experiments. *Proteins*: Structure, Function, and Bioinformatics **79** (S10), 196-207 (2011).

26. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715-1735 (2012).

27. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on TM-score. *Nucl Acid Res* **33**, 2302-2309 (2005).

28. Ray, A., Lindahl, E. & Wallner, B. Improved model quality assessment using ProQ2. *BMC Bioinformatics* **13**, 224 (2012).

29. Wang, L. & Brown, S.J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucl Acid Res* **34**, W243-W248 (2006).

30. Zhao, H., Yang, Y. & Zhouz, Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biology* **8**, 988-996 (2011).

31. Zhao, H., Yang, Y. & Zhouz, Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol BioSyst* **9**, 2417 (2013).

32. Gao, B., Paramanathan, R. & Gupta, R.S. Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Antonie van Leeuwenhoek* **90**, 69-91 (2006).

33. Yeats, C., Bentley, S. & Bateman, A. New knowledge from old: *In silico* discovery of novel protein domains in *Streptomyces coelicolor*. *BMC Microbiol* **3**, 3 (2003).

34. Battistuzzi, F.U., Feijao, A. & Hedges, S.B. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of the land. *BMC Evol Biol* **4**, 44 (2004).

35. de Chateau, M. & Bjorck, L. Identification of interdomain sequences promoting the intronless evolution of a bacterial protein family. *Proc Natl Acad Sci* USA 93, 8490-8495 (1996)

36. Schröder, J. & Tauch, A. Transcriptional regulation of gene expression

in *Corynebacterium glutamicum*: the role of global, master and local regulators in the modular and hierarchical gene regulatory network. *FEMS Microbiol Rev* **34**, 685-737 (2010).

37. Zakeri, B. & Lu, T.K. Synthetic Biology of Antimicrobial Discovery. *ACS Synth Biol* **2**, 358−372 (2013).

38. Rochat, T, Delumeau, O, Figueroa-Bossi, N, Noirot, P, Bossi, L, Dervyn, E, Bouloc, P. Tracking the elusive function of Bacillus subtilis hfq. *PloS ONE* 10(4), e0124977 (2015)

39. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* **32**, 1792-97 (2004).

Supplementary data

S1 Phylogenetic tree of multiple sequence alignment of sequences retrieved from DELTA-BLAST of the *Synechocystis* sp. PCC 6803 hfq NP_441518.1 sequence against the Genbank nr database.

S2 Phylogenetic tree of DUF3107 sequences ○ genera matched to *Synechocystis* sp. PCC 6803 hfq (3HFO-A) by DELTA-BLAST (Figure S2) ● predicted structure in paper for *Corynebacterium glutamicum*,

S3 Two RNA binding sites on (A) Salmonella enterica hfq, (B) Synechocystis sp. 6803 hfq and (C) Mycobacterium vanbaaleni DUF3107 (amino acids 1-63) omitting putative RNA binding residues on the C-terminal tail.

S4 Alignment of the ab initio predicted structure, QUARK24, of Mycobacterium vanbaaleni protein Mvan_1794 (DUF3107) against the crystal structure of Synechocystis sp. PCC 6803 hfq (3HFO-A) by Tm-Align25 and one-to-one threading12. (A) Predicted backbone for Mvan_1794 with β1/β2i, α1, β2ii/β3/β4 segments labelled (B) corresponding structure for 3HFO-A (C) Tm-alignment of M_van_model3.pdb against 3HFO-A showing alignment of α1, β2ii, β3, β4 (compare with one-to-one threading in Figure 2D) (D) view of C along the α-helix and edge of the β2ii, β3, β4 sheet (E/F) the β1, β2i sheets, shown in grey in A and B, for Mycobacterium vanbaaleni (E) and Synechocystis (F) are in the opposite orientation, show limited sequence similarity and don't align with Tm-Align. However, re-ordering the Mycobacterium vanbaaleni Mvan_1794 sequence segments from β1, β2i, α1, β2ii, β3, β4, β5 to α1, β1, β2, β3, β4, β5 (Figure 2D) allows one to one threading, with the β1/β2i sheet aligned, as shown in E/F

Supplementary data S1

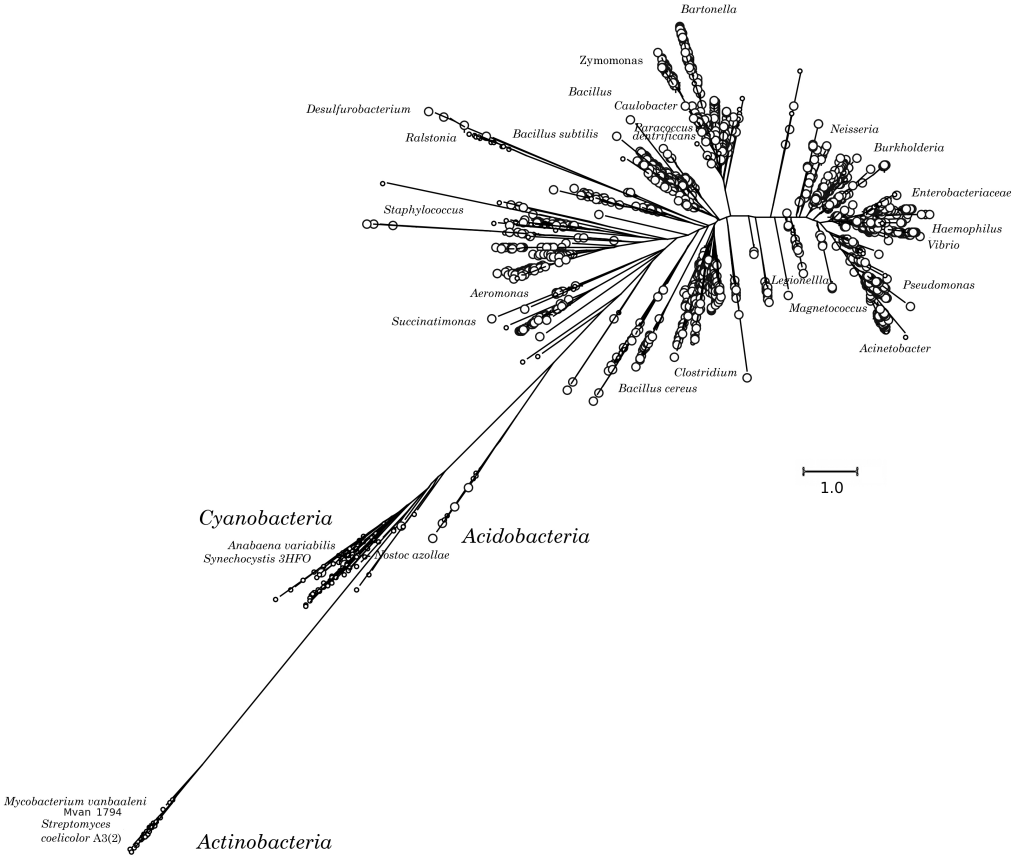

Figure S1. Phylogenetic tree for multiple sequence alignment of protein sequences retrieved by DELTA-BLAST of the *Synechocystis* sp. PCC 6803 Hfq. ○ sequences ○ sequences annotated as hfq, HF1 or Sm-like

Supplementary data S2

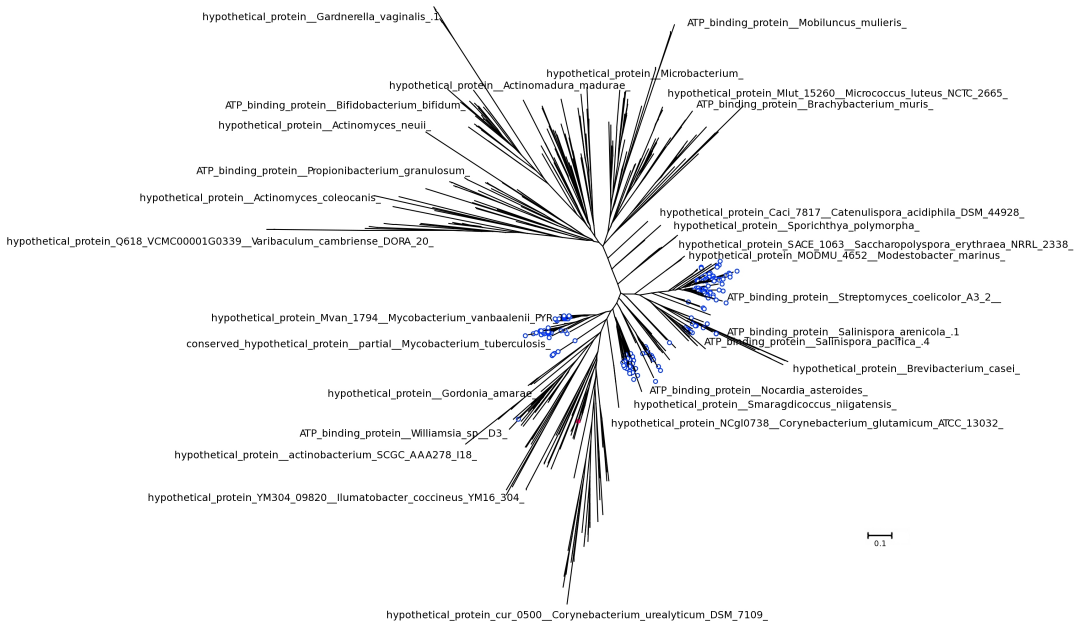

Figure S2. Phylogenetic tree of DUF3107 sequences ○ taxa matched to *Synechocystis* sp. PCC 6803 hfq (3HFO-A) by DELTA-BLAST ✹ predicted structure for *Corynebacterium glutamicum*.

Supplementary data S3

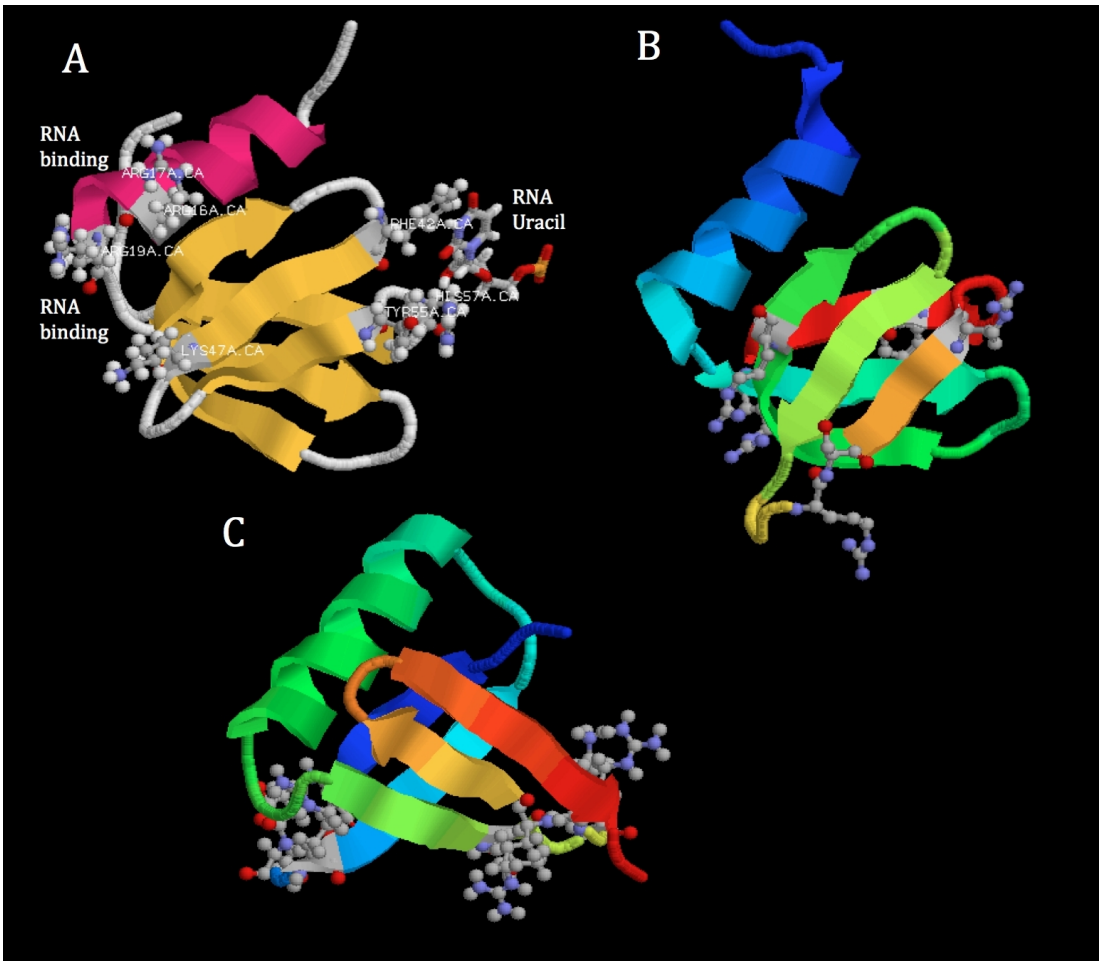

Figure S3. Two RNA binding sites on (A) *Salmonella enterica* hfq, (B) *Synechocystis* sp. 6803 hfq and (C) *Mycobacterium vanbaaleni* DUF3107 (amino acids 1-63) omitting putative RNA binding residues on the C-terminal tail.

Supplementary data S4

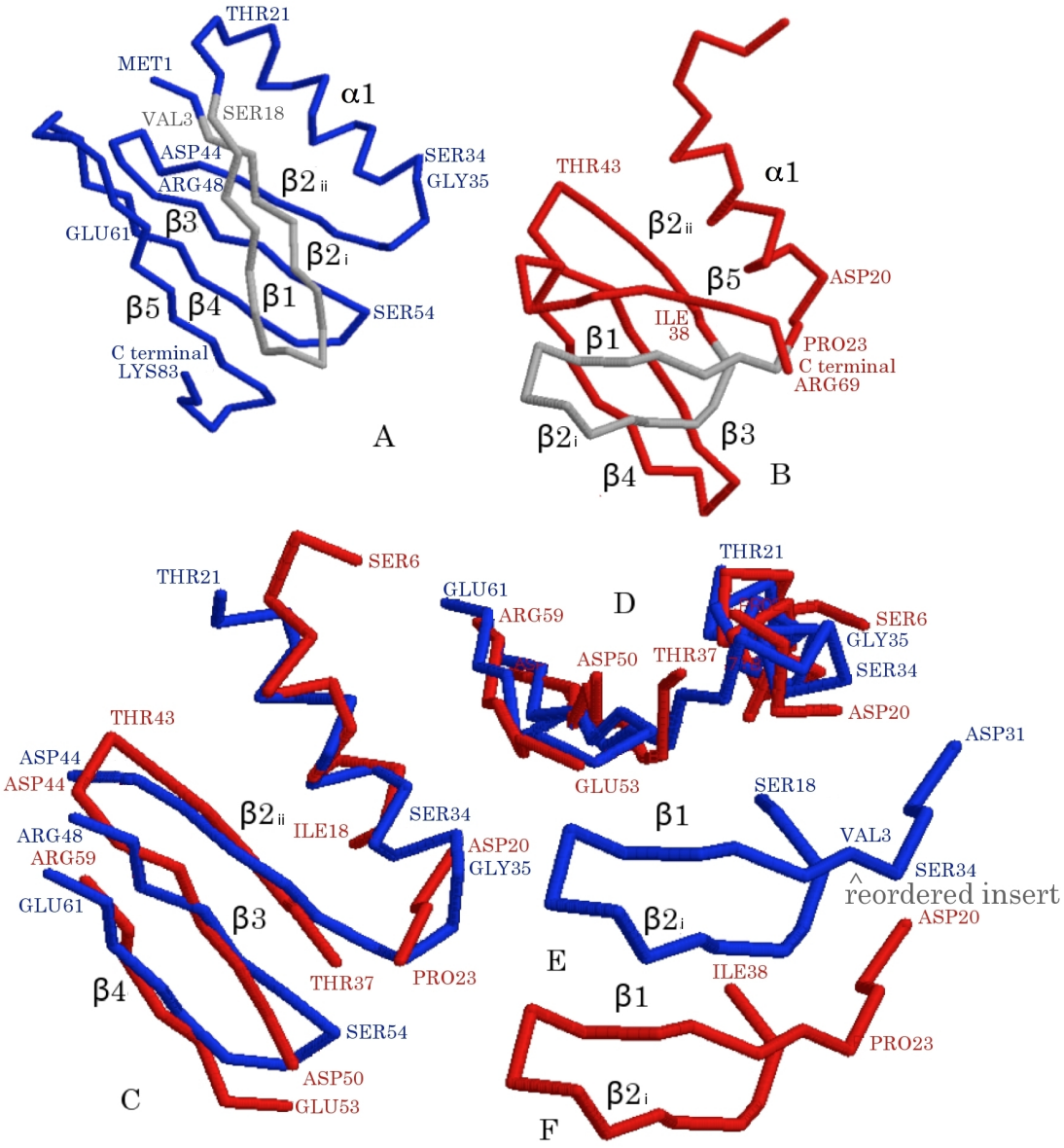

Figure S4. Alignment of the *ab initio* predicted structure, QUARK[26], of *Mycobacterium vanbaaleni* protein Mvan_1794 (DUF3107 - blue) against the crystal structure of *Synechocystis* sp. PCC 6803 hfq (3HFO-A - red) by Tm-Align[27] and one-to-one threading[12].

(A) Predicted backbone for Mvan_1794 with β1/β2i, α1, β2ii/β3/β4 segments labelled (B) corresponding structure for 3HFO-A (C) Tm-alignment of M_van_model3.pdb against 3HFO-A showing alignment of α1, β2ii, β3, β4 (compare with one-to-one threading in Figure 2D) (D) view of C along the α-helix and edge of the β2ii, β3, β4 sheet (E/F) the β1, β2i sheets, shown in grey in A and

B, for *Mycobacterium vanbaaleni* (E) and *Synechocystis* (F) are in the opposite orientation, show limited sequence similarity and don't align with Tm-Align. However, re-ordering the *Mycobacterium vanbaaleni* Mvan_1794 sequence segments from β1, β2i, α1, β2ii, β3, β4, β5 to α1, β1, β2, β3, β4, β5 (Figure 2D) allows one to one threading, with the β1/β2i sheet aligned, as shown in E/F