ABSTACT

ELLIS, JOSEPH. The RNA World: A Look at Ribonuclease P RNA, Small Nucleolar RNA, 6S RNA, and the Small Ribosomal Subunit. (Under the Direction of James W. Brown).

The term RNA world was first coined in 1986 by W. Gilbert. It was largely based on the observation that RNA, not protein, was responsible for the most critical roles in Bacteria, Archaea and Eukaryotes. Although the RNA world means different things to different researchers Gerald Joyce and Leslie Orgel were able to surmise three common characteristic of all RNA World hypotheses: 1) Genetic continuity was dependent on the replication of RNA; 2) Base pairing was predicated on the Watson - Crick Model; 3) Genetically encoded proteins were not catalytic. An example of an ancient catalytic RNA observed in modern cells is ribonuclease P (RNase P). RNase P is responsible for the maturation of pre-tRNA by cleaving the 5' leader to form the mature tRNA and is widely believed to be a relic from the RNA world. Other functionally important RNAs are: small nucleolar RNAs (snoRNAs) which generally catalyze sequence specific 2'-O- ribose methylation and pseudouridylation of ribosomal RNAs, 6S RNAs responsible for the modulation of RNA polymerase, and the small ribosomal subunit which plays a significant role in the synthesis of proteins and peptides. Recent research advances have shown all of these RNAs are important in medical and biotechnology applications. Here we describe our research efforts with these functionally important and essential RNAs: ribonuclease P RNA, small nucleolar RNA, 6S RNA, and the small ribosomal subunit. The cellular processes, database generation, bioinformatics approaches, and the application of RNA in biotechnology are detailed in this work.

THE RNA WORLD: A LOOK AT RIBONUCLEASE P RNA, SMALL NUCLEOLAR RNA, 6S RNA, AND THE SMALL RIBOSOMAL SUBUNIT

by
Joseph Ellis

A dissertation submitted to the Graduate Faculty of
the North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

MICROBIOLOGY

Raleigh, North Carolina

2006

APPROVED BY:

_____	_____
Amy Grunden, Ph.D .	Stu Maxwell, Ph.D.
_____	_____
Robert Kelly, Ph.D. Minor Chair	James Brown, Ph.D. Chair of Advisory Committee

DEDICATION

This work is dedicated to my wife Heather Ellis and our three wonderful children Lauren, Zachary, and Annabelle. Thank you for always listening, for always caring, and for all the smiles we shared along the way.

BIOGRAPHY

I graduated from North Carolina State University with a B.S. in microbiology in 1997. I received a temporary position at Eli Lilly and company a month later and shortly after a full time position as an associate biochemist until 2001. I returned to North Carolina State University in 2001 on an educational leave of absence from Eli Lilly to earn my PhD in Microbiology.

ACKNOWLEDGEMENTS

I would like to thank my family Heather, Lauren, Zachary, and Annabelle for their love and support. I would to thank my mother and father Ron and Pat Ellis for all they have done for me, not only during my graduate career, but throughout my life. I would not be who I am today without them. I would also like to thank my advisor Jim Brown for allowing me to pursue all my interests and for all his advice, encouragement, and unfaltering belief, in words and in actions, in my research ability. Thank you Jim! I would also like to thank Amy Grunden for believing in my research and programming abilities and for sticking with me until we could complete our research efforts, even when difficult or confounding. I would also like to thank the entire microbiology department faculty, too numerous to name individually, for all of their support, encouragement, and kindness. I cherish my time and my research while at North Carolina State University and will miss all of you greatly. Most of all I would like to thank my lord and savior Jesus Christ for guiding and strengthening me through all my work and trials. I drew great strength from the verse in Psalms 46:10 “Be still and know that I am God…”.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1: LITERATURE REVIEW: RIBONUCLEASE P IN BACTERIA AND ARCHAEA

INTRODUCTION

Ribonuclease P (RNase P) is an ancient ribonuclease responsible for the processing of precursor tRNA (pre-tRNA) by removing the 5' leader sequence but also participates in the maturation of other RNAs such as 2S, 4.5S, tmRNA (10S), some polycistronic mRNA, and some viral RNAs. [1-6] Unlike most enzymes, RNase P is composed of both RNA and protein. The RNase P RNA is found in all three domains of life (Bacteria, Archaea, and Eukaryotes) including mitochondria and plastids. Furthermore, it is the RNA, not the protein(s), that catalyze the site specific cleavage reaction and is therefore, by definition, a ribozyme.

BACTERIAL RIBONUCLEASE P

Bacterial RNase P RNA Biophysical Diversity

Two distinct biophysical RNase P RNA structures predominate in the bacterial domain, Type A and Type B. Both Type A and Type B RNase P RNAs are catalytically active *in vivo* without protein, however their secondary structure is significantly different. Type A RNase P RNAs, such as those encoded by *E. coli*, are dependent on the protein for stabilizing the tertiary structure of the molecule compared to the Type B RNase P RNA, such as those encoded by *B. subtilis,* which is not dependent on the protein subunit for stabilizing its tertiary structure.[7]

Type A RNase P RNAs are the most common (and ancestral) form of the RNA; type B RNase P RNAs are found only in some Gram-positive Bacteria. Gram-positive Bacteria can be subdivided into two phylogenetic groups: low-G+C (firmicutes) and high-G+C (actinobacteria). High-G+C Gram-positive Bacteria contain the ancestral Type A RNase P RNA, however the main branch of the low-G+C Bacteria RNase P RNAs share several unique structural elements not found in other RNase P RNAs (P5.1, P10.1, P15.1, P15.2), and lack several that are otherwise highly conserved (P6, P13, P14, P16, P17, P18). (Figure 1) The RNase P RNAs found in low-G+C Gram-positive Bacteria, with these unique RNA elements, underwent a dramatic and abrupt biophysical change to their structural and sequence morphology (Figure 2).[8] Haas et al. also noted a difference in tRNA biogenesis between low G+C Gram positive Bacteria and all other Bacteria. In low G+C Gram-positive Bacteria, unlike all other Bacteria, the 3' terminal CCA sequence of tRNA, an important substrate recognition element for the RNase P Loop 15 or L15, is generally not encoded in the gene sequences, but are predominantly added postranscriptionally.

Despite these biophysical differences it has been shown by Walker and Engel that the two types of RNase P RNA are interchangeable in vivo.[9] This finding suggests the structural idiosyncrasies of Type A and B enzymes are not crucial for RNase P function *in vivo*.

The Functional Role of Bacterial RNase P Protein Subunit

In Bacteria, the RNase P ribonucleoprotein complex is comprised of a single RNA and a small protein subunit to form the complete holoenzyme. However, it is the

RNA not the protein that catalyzes the cleavage reaction of the pre-tRNA substrate forming the mature tRNA. The holoenzyme complex of RNA and protein form a dimer in solution in both *E. coli* and *B. subtilis*.[10] Buck et al demonstrated that although both could form dimers, in *E. coli* the holoenzyme forms a heterogeneous mixture of dimers and monomers and shift nearly completely to monomers in the presence of mature tRNA.[7] However in *B. subtilis* holoenzyme exist almost exclusively as dimers, and do not shift to monomers in the presence of mature tRNA.[7] Interestingly, Buck et al argue that it is the RNA attributes, not the attributes of the protein, responsible for dimer formation in both *E. coli* and *B. subtilis*. [11] Because the RNA plays the crucial role in dimerization, it is likely that the additional RNA elements (i.e. P5.1, P10.1, P15.1, P15.2) in Type B RNase P RNA are associated with stable dimer formation in *Bacillus* and the absence of these RNA elements in Type A RNase P RNA contribute to an unstable dimer complex that disassociates in the presence of substrate.

Unfortunately the biochemical significance of stable dimer formation *in vitro* and its biological relevance *in vivo* is not understood.

ARCHAEAL RIBONUCLEASE P

Archaeal RNase P

Archaeal RNase P holoenzymes are typically comprised of a single RNA and four protein subunits. These proteins are homologues to four proteins conserved in eukaryotic nuclear RNase P enzymes. At sufficiently high salt concentrations (4 M ammonium acetate and 300 mM MgCl2) the RNAs of some Archaea, those most

resembling Bacterial RNase P RNAs in both sequence and structure, are catalytically active in the absence of proteins *in vitro*. [12]

One commonality among all archaeal RNase P holenzymes is the presence of the eukaryote RNase P protein subunit ortholog Pop4. However, there does not appear to be any sequence homology between the bacterial RNase P protein C5 and any of the four archaeal RNase P subunits. Because there is no clear sequence homology between bacterial and archaeal or eukaryotic RNase P proteins it is likely the RNase P RNA evolved first and the proteins evolved independently following the split of the three domains of life.

Interestingly no RNase P RNA or associated proteins were annotated with the release of the genomic sequences for *Pyrobacterium* and *Nanobacterium*. Further research by Li *et al* using novel computational methodologies to search for RNase P RNA were also unable to identify an RNase P RNA in either of these two archaeal genera. [13] *Nanoarchaeum* is a deep branching obligate symbiont with *Ignicoccus*. It has an amazingly small genome (490,885bp) and encodes some of its tRNAs in halves, each half contains the appropriate sequence and structures to be joined by splicing in *trans*. [14-18] The tRNAs in *Nanoarchaeum* have an upstream Box A RNA polymerase binding site very close to the ORF of the tRNA itself. Although not demonstrated experimentally, it is likely due to the close proximity of Box A to the ORF of the tRNA (~13nt), that *Nanoarchaeum* does not transcribe a pre-tRNA with a 5' leader. (Figure 3) The lack of a 5' leader on the tRNA means the RNase P holoenzyme is not required for post transcriptional modification (i.e. removal of the 5' leader sequence). This may also be the case in *Pyrobacterium* as well. This raises interesting questions regarding this

evolutionary leap. How did the genome cope with the loss of the normally essential function of RNase P and what gains (if any) are made by the loss of this multifunctional holoenzyme?

Archaeal RNase P RNA Biophysical Diversity

As in Bacteria, most archaeal RNase P RNAs are Type A RNAs, resembling bacterial RNase P RNAs in both sequence and structure. However, despite the structural similarity between bacterial RNase P RNA and the archaeal RNA it was widely thought previously that, like eukaryotic RNase P RNA, the archaeal RNase P RNA was absolutely dependent on their associated proteins for catalytic activity. Pannucci et al described for the first time that some archaeal RNase P RNA are catalytically active in the absence of protein, like bacterial RNase P RNAs, but that they required extreme ionic conditions (4 M ammonium acetate, 300 mM MgCl2, 50 mM Tris-Cl). [12] As in bacteria, the protein subunits in the archaeal enzymes seem not to contribute directly to catalysis, but at least predominantly toward stabilization of the superstructure of the RNA subunit.

One of the subsets of archaeal RNase P RNAs which did not display catalytic activity even under high ionic conditions was Type M RNase P RNAs. To date only five archaeal species (*Archaeoglobus fulgidus, Methanocaldococcus jannaschii, Methanococcus marapaludis, Methanothermococcus thermolithotrophicus, and Methanococcus vannielii)* have been described with a Type M RNA. [19] Type M RNAs are essentially similar to Type A RNAs but lack two essential substrate recognition elements. One such element found in all other bacterial and archaeal RNase P RNAs is

P8. P8 is located in the region of the RNA that forms a highly conserved cruciform consisting of P7, P8, P9, and P10. [20, 21] (Figure 4) P8 recognizes substrate at the T loop of the pre-tRNA. Additionally, in Bacteria P8 stabilizes, via tertiary interactions, P18, P4, and P14. Another structural element, L15, is also missing in Type M RNAs. L15 is distal of P15, and like P8 is also essential for substrate recognition. L15 recognizes and binds the 3'-NCCA tail of the pre-tRNA which is necessary for efficient substrate cleavage. [22, 23] Type M RNase P RNAs, then, have specifically lost all of the regions known to be directly involved in substrate recognition outside of the active site.

The absence of essential RNA elements involved in substrate recognition and the absence of novel structural elements that could replace the functional roles P8 and L15 raises the question, "How does RNase P compensate for the loss of these critical secondary structures in its RNA?" Two scenarios could answer this question: 1) One or more of the four known protein homologues associated with other archaeal RNase P RNAs are able to specifically recognize the tRNA substrate, or 2) an additional protein or proteins has/have taken over the functional role of substrate recognition typically associated with P8 and L15.

PYROCOCCS HORIKOSHII OT3 AS AN ARCHAEAL MODEL

Determining the structure of human RNase P proteins is of interest to many researchers. Unfortunately, human RNase P proteins and more generally eukaryotic RNase P proteins have been difficult to crystallize for structural analysis. However, because many Archaea are hyperthermophiles, share strong sequence homology with

eukaryotic RNase P proteins, and by the nature of their extreme environment produce highly stable proteins, they make excellent model systems for structural analysis. One such archaeon is *Pyrococcus horikoshii OT3.* All four RNase P proteins from *P. horikoshii OT3* have had their structure determined by X-ray crystalography. Here we will briefly examine each of the four protein structures.

Ph1771p

Ph1771p is a homologue of human RNase P protein Rpp29 and is composed of four α-helices and six antiparallel ß-sheets and one ß-strand near the C-terminus that protrudes away from the globular portion of the protein forming a ß-barrel structure. [24] (Figure 5) The protruding ß-strand (ß7) forms a ß-sheet with ß4 which Kouzuma et al suggests is involved in protein-protein interaction with other RNase P subunits. Two possible RNA binding sites were identified in Ph1771p. The first is a loop region connecting strands ß2 and ß3 which are composed of hydrophilic residues exposed to solvent and the other is composed of α-helices 1-4 and ß-strand ß6 forming a cluster of positively charged amino acids. [24] Kouzuma et al also noted Ph177p shared structural similarity to other RNA binding proteins such as *Staphylococcus aureus* translational regulator Hfq and *Haloarcula marismortui* ribosomal protein L21E. This suggests a common ancestor among these RNA binding proteins that underwent divergent evolution to bind and perform specific functions with their associated RNAs respectively.

Ph1877p

Ph1877p is a homologue to human RNase P protein Rpp30 and is composed of ten α-helices and seven β-strands forming a TIM barrel. [25, 26] (Figure 6) Takagi et al utilized site directed mutagenesis to determine the functional role of 12 amino acids of Ph1877p: Lys42, Arg68, Arg87, Arg90, Asp98, Arg107, His114, Lys123, Lys158, Arg176, Asp180, and Lys196. They found the amino acids that affected RNase P's ability to cleave substrate were Arg90, Arg107, Lys123, Arg176, and Lys196 resulting in reduced activity of the holoenzyme by 32-48% versus wild type activity. [25] Yeast two-hybrid analysis showed that Ph1877p interacts with Ph1481p in the RNase P holoenzyme, which is consistent with the interactions determined in other archaeal RNase P protein-protein interactions. [27, 28] Utilizing the Ph1481p-Ph1877p relationship Kawano et al were able to co-crystallize Ph1481p and Ph1877p. When the crystal structure data was coupled with the site directed mutation results obtained by Takagi et al, the role of some of the important amino acids could be identified. Ph1877p's Arg107 and Arg176 in the crystal structure appear to be directly involved in binding Ph1481 via hydrogen bonding. The disruption of this important interaction would likely affect the heterodimers formation between Ph1887p and Ph1481p resulting in the reduced activity reported by Takagi et al. [25, 26]

Ph1601p

Ph1601p is a homologue to human RNase P protein Rpp21 and is composed of an N-terminal domain comprised of two α-helices, while the central domain and C-terminal domain together form a zinc ribbon domain, giving the protein an L-shape. [29] (Figure

7) Several clusters of positively charged amino acids along one face of the L-arms, suggest an RNA binding role, whereas four Cys residues bind a zinc molecule stabilizing the N-terminus and C-terminus domains. [29] When Kakuta et al performed site-directed mutagenesis it was observed that Lys69, Arg86, and Arg105 play important functional roles in RNase P activity.

Ph1481p

Ph1481p is a homologue to human RNase P protein Pop5 and is composed of five antiparallel ß-sheets and five α-helices forming a α/β globular protein. [26] (Figure 8) When this structure is compared to RNase P protein structure from *Bacillus subtilis*, *Staphylococcus aureus*, and *Thermatoga maritima,* no meaningful sequence homology or structural similarity is discernable. [25, 30-32] The interpretations of the structural similarity between Ph1481p and eubacterial RNase P proteins proposed by Wilson et al differ from Kawano et al. Wilson et al report the structure of Ph1481p is similar to the bacterial RNase P protein subunits . [33] They find the primary structural difference between eubacterial RNase P protein subunit and that of Ph1481p is the orientation of helix α4 of Ph1491p and its counterpart in bacterial RNase P proteins α1. Wilson et al go on to note that the structural similarity is remarkable because of the primary sequence and secondary topologies differ significantly from the eubacterial RNase P protein, suggesting a different evolutionary origin. However, Kawano et al argue that the structure of Ph1481 is much more similar to the generic ribonucleoprotein (RNP) domains found in a number of RNA binding proteins suggesting a role in the binding of

the RNase P RNA molecule itself. There is probably no meaningful similarity between the RNP domain fold and the unique fold of bacterial RNase P proteins.

Ph1877p and Ph1481p form a heterotetrameric structure, with a homodimer of Ph1481p in the center of two monomers of Ph1877p. Kawano et al found the α1 and α2 loop are involved in the dimerization of Ph1481p and the heterodimerization of Ph1877p. [26] Their data also strongly supports that the dimerization of Ph1481p is required for RNase P activity and the RNase P particle contains a heterotetramer composed of Ph1481p and Ph1877p. [26] Whether this heterotetramer is associated with one or two molecules of the RNA subunit has not been determeind.

Is L7Ae a fifth RNase P protein in Pyrococcus horikoshii OT?

L7Ae protein was first described as a protein associated with the small ribosomal subunit. [34] However, the crystallization of the complete large ribosomal subunit (50S) with the associated proteins reveled L7Ae was actually associated with the large subunit not the small subunit as previously thought. [35] More recently L7Ae was found to have another RNA binding role in Archaea. Kuhn et al demonstrated, via protein binding studies, that L7Ae binds to small nucleolar RNAs (sRNA) with the C/D motif and forms specific complexes with some H/ACA sRNAs as well. [36, 37] Further studies on L7Ae's association with sRNAs revealed a specific structural motif used as a binding site by L7Ae called a kink-turn or k-turn motif that is found in both ribosomal RNA and sRNA. [37] This motif is composed of a kink in the phosphodiester backbone that causes a sharp turn in the RNA helix. [38]

More recently, interest was generated with the finding that the L7Ae protein in *Haloarcula marismortui* shared a strong sequence similarity with human RNase P protein subunit Rpp38. [39] Furthermore, reconstitution experiments with *Pyrococcus horikoshii's* four known RNase P proteins and RNA revealed a lower than expected optimal reaction temperature of 55°C, much lower than the 70°C reported for the wild type RNase P holoenzyme. [40] These findings led Fukuhara et al to examine the role L7Ae may play in RNase P holoenzyme in *Pyrococcus horikoshii OT3.*

While Fukuhara et al are quick to point out they currently have no direct evidence that Ph1496p is an actual protein component in the *P. horikoshii* RNase P *in vivo*, they do have substantial circumstantial evidence. Most notably, when L7Ae (Ph1496p) is added to the four know RNase P proteins (Ph1481p, Ph1601p, Ph1877p, and Ph1771p) and RNase P RNA, optimal enzymatic activity is restored to the wild type temperature of 70°C.

Based on L7Ae's ability to bind RNA, Fukuhara et al examined its ability to bind RNase P RNA. Utilizing mobility shift assays they showed that Ph1496p binds specifically to RNase P RNA in the stem-loop structure composed of residues 229-276 and the terminal stem-loop (116-201 or P12) under excess amount of L7Ae protein. [39] Interestingly, this is different from other RNA structural motifs (i.e. k-turn) described in the literature as the typical binding sites for L7Ae.

Is Alba an Additional Archaeal RNase P Protein Subunit?

The protein "alba" has been characterized in *Sulfolobus* as a dimeric, highly basic protein that binds cooperatively and at high density to DNA, inducing negative

supercoiling. [41-43] While the dense coating of *Sulfolobus*'s DNA by Alba, and resulting negative supercoiling protects it from nuclease digestion, no significant compaction of genomic DNA has been observed. [41, 44] DNA binding affinity of Alba in Archaea is regulated by *Sir2* which deacetylates lysine 11 in *Archaeoglobus fulgidus* or lysine 16 in *Sulfolobus solfataricus*, modulating gene expression. [41] While the role of Alba as a DNA binding protein is clear, new insights into its structure and sequence analysis suggested a potential dual role in Crenarchaea: one in DNA binding and the other in RNA binding.

The crystal structure of Alba shows a fold of the N-terminus similar to the DNA binding domain of DNase I followed by an initiation factor IF3 domain at the C-terminus. [45, 46] The IF3 RNA binding motif is part of the YhbY group of RNA binding peptides. For this reason Alba has been hypothesized to have originally been an RNA binding protein that has retained this function but subsequently evolved DNA binding properties in the crenarchaeal lineage. [45]

Using sequence analysis methodologies such as PSI-Blast L Aravind *et al* was able to suggest homology between Alba proteins and RNase P subunit Rpp25 with ~30% similarity between the two proteins. [43] This data in conjunction with the structural RNA motif of Alba is suggestive that it could also have a role in the RNase P holoenzyme. Further evidence suggesting the presence of yet-unidentified protein constituent(s) of RNase P in *Methanothermobacter thermoautotrophicus* is the buoyant density which is 1.42 g/mL in cesium sulfate. This density corresponds to an RNA:protein ratio of 0.96. Since the RNA is 98 kDa, there must be ~93 kDa of protein in the RNA:protein complex. However, only 70 total kDa of protein is accounted for by the

4 known proteins; either the ratio of one or more of the known protein subunits to the RNA is not 1:1 (e.g. two Mth11 proteins required per single RNase P RNA to form a functional holoenzyme) or there is 23 kDa of protein yet to be identified in *M. thermoautotrophicus*.

In chapter 2 we will discuss that despite Alba's homology to the known human RNase P subunit Rpp25, an RNA binding motif, and an unaccounted portion of protein within the RNase P holoenzyme of *M. thermoautotrophicus* there is no molecular or biochemical evidence supporting the hypothesis that Alba has a secondary function as an RNase P subunit in *M. thermoautotrophicus*.

REFERNECES

1. Bothwell, A.L., R.L. Garber, and S. Altman, *Nucleotide sequence and in vitro processing of a precursor molecule to Escherichia coli 4.5 S RNA.* J Biol Chem, 1976. **251**(23): p. 7709-16.
2. Hori, Y., T. Tanaka, and Y. Kikuchi, *In vitro cleavage of Drosophila 2S rRNA by M1 RNA.* Nucleic Acids Symp Ser, 2000(44): p. 93-4.
3. Ushida, C., D. Izawa, and A. Muto, *RNase P RNA of Mycoplasma capricolum.* Mol Biol Rep, 1995. **22**(2-3): p. 125-9.
4. Bothwell, A.L., B.C. Stark, and S. Altman, *Ribonuclease P substrate specificity: cleavage of a bacteriophage phi80-induced RNA.* Proc Natl Acad Sci U S A, 1976. **73**(6): p. 1912-6.
5. Yang, Y.H., et al., *Engineered external guide sequences are highly effective in inducing RNase P for inhibition of gene expression and replication of human cytomegalovirus.* Nucleic Acids Res, 2006. **34**(2): p. 575-83.
6. Gimple, O. and A. Schon, *In vitro and in vivo processing of cyanelle tmRNA by RNase P.* Biol Chem, 2001. **382**(10): p. 1421-9.
7. Buck, A.H., et al., *Structural perspective on the activation of RNAse P RNA by protein.* Nat Struct Mol Biol, 2005. **12**(11): p. 958-64.
8. Haas, E.S., et al., *Structure and evolution of ribonuclease P RNA in Gram-positive bacteria.* Nucleic Acids Res, 1996. **24**(23): p. 4775-82.
9. Walker, S.C. and D.R. Engelke, *Ribonuclease p: the evolution of an ancient RNA enzyme.* Crit Rev Biochem Mol Biol, 2006. **41**(2): p. 77-102.
10. Buck, A.H., et al., *Protein activation of a ribozyme: the role of bacterial RNase P protein.* Embo J, 2005. **24**(19): p. 3360-8.
11. Haas, E.S. and J.W. Brown, *Evolutionary variation in bacterial RNase P RNAs.* Nucleic Acids Res, 1998. **26**(18): p. 4093-9.
12. Pannucci, J.A., et al., *RNase P RNAs from some Archaea are catalytically active.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 7803-8.
13. Li, Y. and S. Altman, *In search of RNase P RNA from microbial genomes.* Rna, 2004. **10**(10): p. 1533-40.
14. Randau, L., M. Pearson, and D. Soll, *The complete set of tRNA species in Nanoarchaeum equitans.* FEBS Lett, 2005. **579**(13): p. 2945-7.
15. Waters, E., et al., *The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12984-8.
16. Randau, L., et al., *The heteromeric Nanoarchaeum equitans splicing endonuclease cleaves noncanonical bulge-helix-bulge motifs of joined tRNA halves.* Proc Natl Acad Sci U S A, 2005. **102**(50): p. 17934-9.
17. Di Giulio, M., *Nanoarchaeum equitans is a living fossil.* J Theor Biol, 2006.
18. Randau, L., et al., *Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves.* Nature, 2005. **433**(7025): p. 537-41.
19. Brown, J.W., *The Ribonuclease P Database.* Nucleic Acids Res, 1999. **27**(1): p. 314.
20. Harris, J.K., et al., *New insight into RNase P RNA structure from comparative analysis of the archaeal RNA.* Rna, 2001. **7**(2): p. 220-32.

21. Brown, J.W. and E.S. Haas, *Ribonuclease P structure and function in Archaea.* Mol Biol Rep, 1995. **22**(2-3): p. 131-4.
22. Svard, S.G., U. Kagardt, and L.A. Kirsebom, *Phylogenetic comparative mutational analysis of the base-pairing between RNase P RNA and its substrate.* Rna, 1996. **2**(5): p. 463-72.
23. Svard, S.G. and L.A. Kirsebom, *Several regions of a tRNA precursor determine the Escherichia coli RNase P cleavage site.* J Mol Biol, 1992. **227**(4): p. 1019-31.
24. Numata, T., et al., *Crystal structure of archaeal ribonuclease P protein Ph1771p from Pyrococcus horikoshii OT3: an archaeal homolog of eukaryotic ribonuclease P protein Rpp29.* Rna, 2004. **10**(9): p. 1423-32.
25. Takagi, H., et al., *Crystal structure of the ribonuclease P protein Ph1877p from hyperthermophilic archaeon Pyrococcus horikoshii OT3.* Biochem Biophys Res Commun, 2004. **319**(3): p. 787-94.
26. Kawano, S., et al., *Crystal structure of protein Ph1481p in complex with protein Ph1877p of archaeal RNase P from Pyrococcus horikoshii OT3: implication of dimer formation of the holoenzyme.* J Mol Biol, 2006. **357**(2): p. 583-91.
27. Kifusa, M., et al., *Protein-protein interactions in the subunits of ribonuclease P in the hyperthermophilic archaeon Pyrococcus horikoshii OT3.* Biosci Biotechnol Biochem, 2005. **69**(6): p. 1209-12.
28. Hall, T.A. and J.W. Brown, *Interactions between RNase P protein subunits in archaea.* Archaea, 2004. **1**(4): p. 247-54.
29. Kakuta, Y., et al., *Crystal structure of a ribonuclease P protein Ph1601p from Pyrococcus horikoshii OT3: an archaeal homologue of human nuclear ribonuclease P protein Rpp21.* Biochemistry, 2005. **44**(36): p. 12086-93.
30. Stams, T., et al., *Ribonuclease P protein structure: evolutionary origins in the translational apparatus.* Science, 1998. **280**(5364): p. 752-5.
31. Spitzfaden, C., et al., *The structure of ribonuclease P protein from Staphylococcus aureus reveals a unique binding site for single-stranded RNA.* J Mol Biol, 2000. **295**(1): p. 105-15.
32. Kazantsev, A.V., et al., *High-resolution structure of RNase P protein from Thermotoga maritima.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7497-502.
33. Wilson, R.C., et al., *Structure of Pfu Pop5, an archaeal RNase P protein.* Proc Natl Acad Sci U S A, 2006. **103**(4): p. 873-8.
34. Wittmann-Liebold, B., *The Ribosome: Structure, Function, and Evolution,*, in *American Society for Microbiology, Washington, DC,*. 1990. p. 598-616.
35. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.
36. Kuhn, J.F., E.J. Tran, and E.S. Maxwell, *Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein.* Nucleic Acids Res, 2002. **30**(4): p. 931-41.
37. Rozhdestvensky, T.S., et al., *Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea.* Nucleic Acids Res, 2003. **31**(3): p. 869-77.
38. Klein, D.J., et al., *The kink-turn: a new RNA secondary structure motif.* Embo J, 2001. **20**(15): p. 4214-21.

39. Fukuhara, H., et al., *A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3.* Biochem Biophys Res Commun, 2006. **343**(3): p. 956-64.
40. Kouzuma, Y., et al., *Reconstitution of archaeal ribonuclease P from RNA and four protein components.* Biochem Biophys Res Commun, 2003. **306**(3): p. 666-73.
41. Bell, S.D., et al., *The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation.* Science, 2002. **296**(5565): p. 148-51.
42. Wardleworth, B.N., et al., *Structure of Alba: an archaeal chromatin protein modulated by acetylation.* Embo J, 2002. **21**(17): p. 4654-62.
43. Aravind, L., L.M. Iyer, and V. Anantharaman, *The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism.* Genome Biol, 2003. **4**(10): p. R64.
44. Sandman, K. and J.N. Reeve, *Archaeal chromatin proteins: different structures but common function?* Curr Opin Microbiol, 2005. **8**(6): p. 656-61.
45. Chou, C.C., et al., *Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 Angstroms.* J Bacteriol, 2003. **185**(14): p. 4066-73.
46. Wang, G., et al., *Crystal structure of a DNA binding protein from the hyperthermophilic euryarchaeon Methanococcus jannaschii.* Protein Sci, 2003. **12**(12): p. 2815-22.

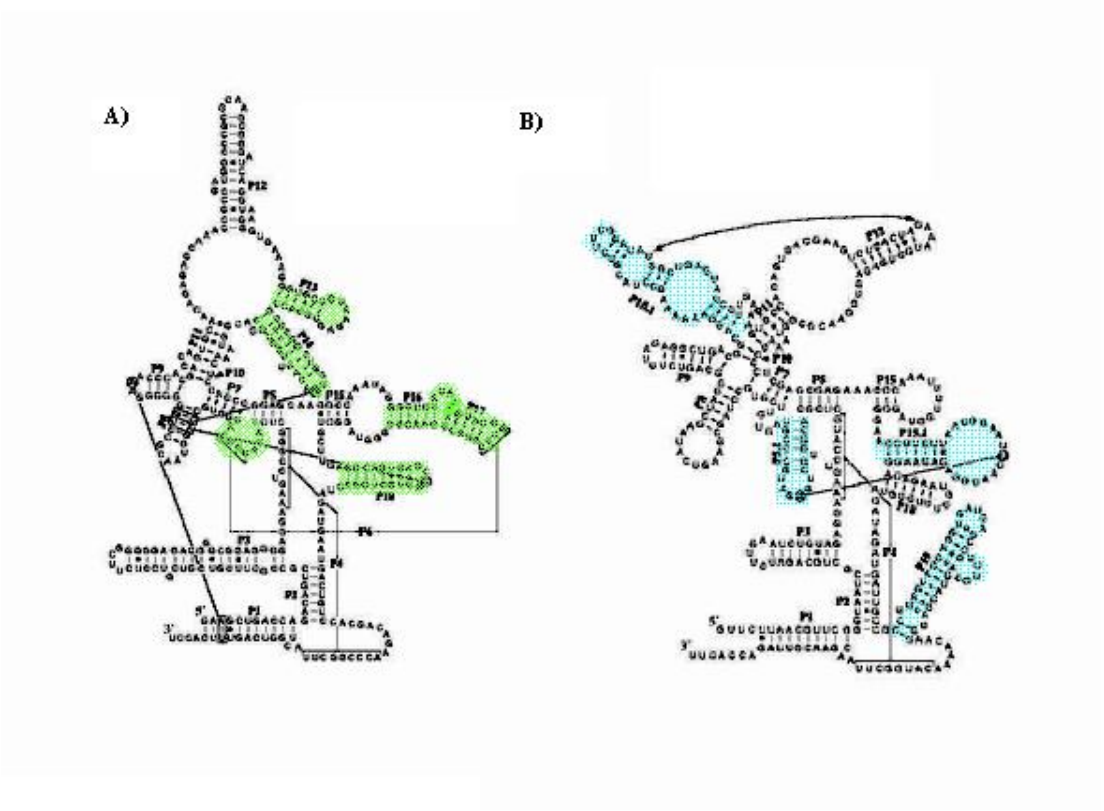


Figure 1: Type A and Type B RNase P RNA secondary structure

A typical Type A RNase P RNA from *E. coli* as seen in (A) compared to a Type B RNase P RNA as seen in (B). Type B RNAs are typically characterized as having all the secondary elements such as: P5.1, P10.1, P15.1, but lacking the RNase P RNA elements commonly associated with Type A RNAs such as: P6, P13, P14, P16, P17, P18

Figure 1 Reference:

Brown, J.W., *The Ribonuclease P Database.* Nucleic Acids Res, 1999. **27**(1): p.314.

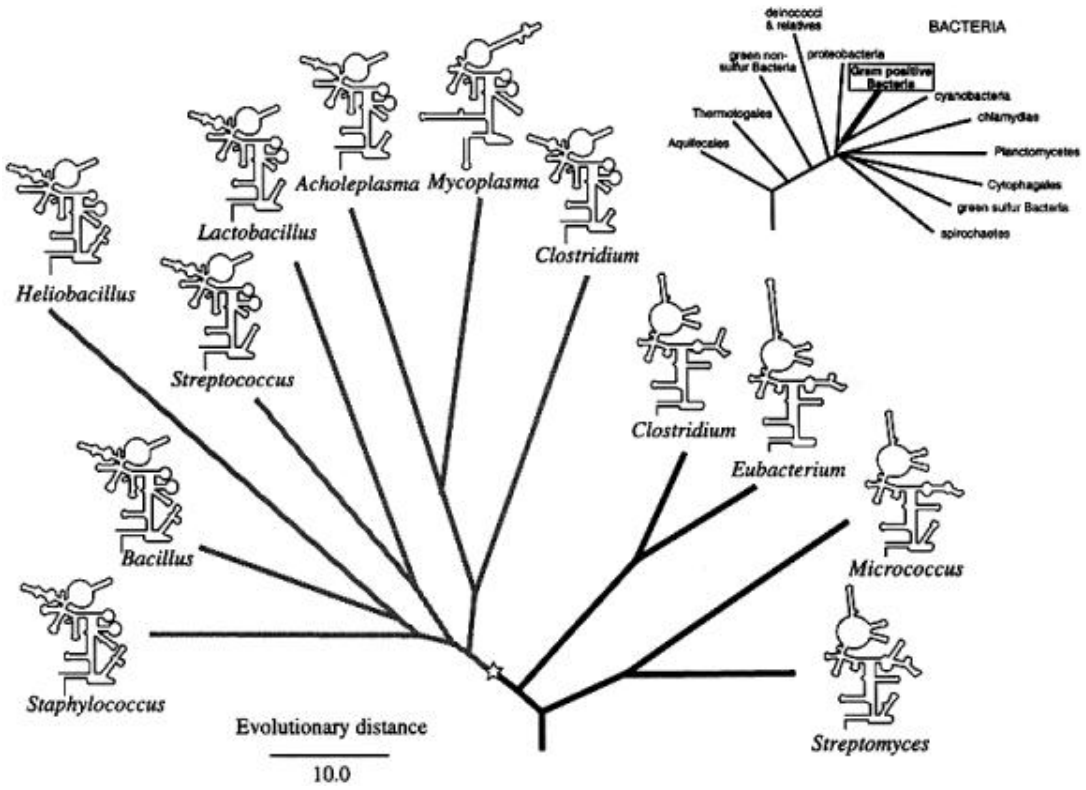


Figure 2: Phylogenetic tree of Gram positive bacteria with RNase P RNA Structure

An RNase P RNA phylogenetic tree of Gram positive bacteria is shown with their associated RNase P RNA secondary structures as reported by Haas et al. The star indicates the position in the tree separating species with type A RNase P RNA (*Clostridium, Eubacterium, Micrococcus, Streptomyces*) and species with type B RNase P RNA structures (*Clostridium, Mycoplasma, Acholeplasma, Lactobacillus, Streptococcus, Heliobacillus, Bacillus, Staphylococcus*). Note the distinct biophysical RNA divergence in *Clostridium* species, with some containing type A RNase P RNAs and other *Clostridium* species containing type B RNase P RNAs.

Figure 2 Reference:

Haas, E.S., et al., *Structure and evolution of ribonuclease P RNA in Gram-positive bacteria.* Nucleic Acids Res, 1996. **24**(23): p. 4775-82.

Figure 3: The tRNA of Nanoarchaeum

The tRNAs in *Nanoarchaeum* have an upstream Box A RNA polymerase binding site very close to ORF of the tRNA itself. While not shown experimentally it is likely, due to the close proximity of Box A to the ORF of the tRNA (~13nt), that *Nanoarchaeum* does not transcribe a pre-tRNA with a 5' end.

Figure 3 Reference:

Randau, L., et al., *Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves.* Nature, 2005. **433**(7025): p. 537-41.

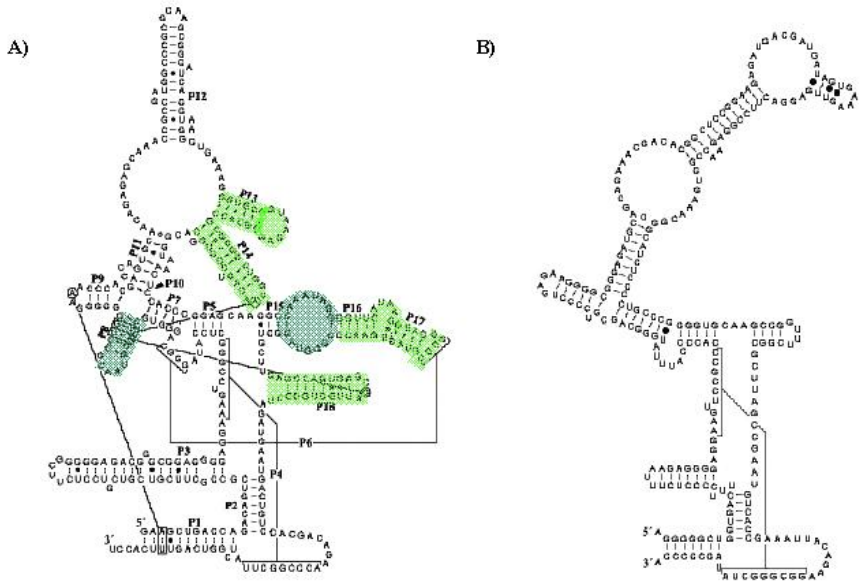


Figure 4: Type A and Type M RNase P RNA secondary structure

Secondary structure of Type A RNase P RNAs (a) and Type M (b). Type M RNAs are essentially the same as Type A RNAs but lack two essential substrate recognition elements. One such element found in all other bacterial and archaeal RNase P RNAs is P8 (blue). P8 recognizes tRNA substrate at the T loop of the pre-tRNA. Another structural element, L15 (blue), is also missing in Type M RNAs. L15 is distal of P15 and like P8 is also essential for substrate recognition. L15 recognizes and binds the 3'-NCCA tail of the pre-tRNA which is necessary for efficient substrate cleavage.

Figure 4 Reference:

Brown, J.W., *The Ribonuclease P Database.* Nucleic Acids Res, 1999. **27**(1): p.314.

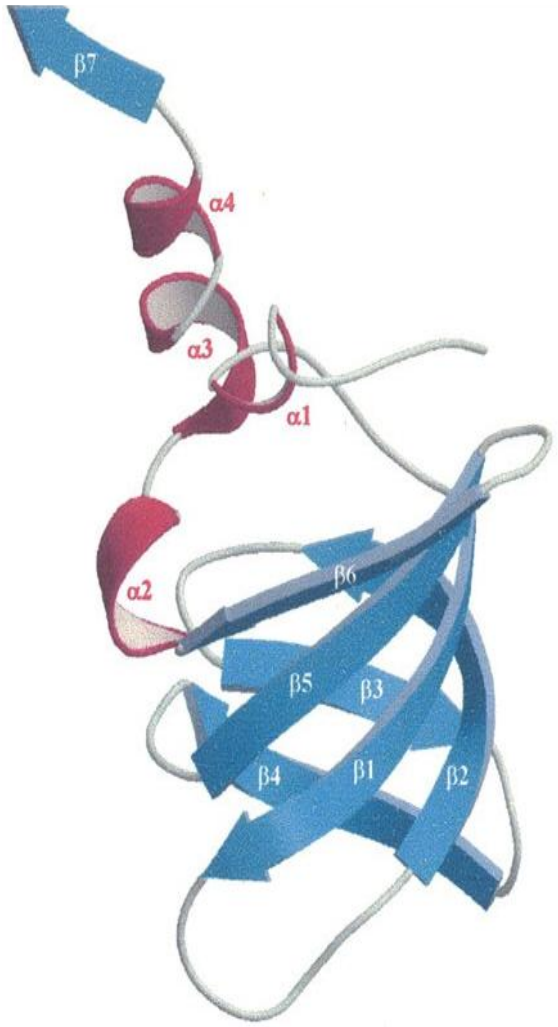


Figure 5: Ribbon structure of RNase P protein Ph1771p

Ribbon representation of the RNase P protein Ph1771p from *P. horikoshii*. Ph1771p is a homologue of human RNase P protein Rpp29 and is composed of four α-helices and six antiparallel ß-sheets and one ß-strand near the C-terminus that protrudes away from the globular portion of the protein forming a ß-barrel structure

Figure 5 Reference:

Numata, T., et al., *Crystal structure of archaeal ribonuclease P protein Ph1771p from Pyrococcus horikoshii OT3: an archaeal homolog of eukaryotic ribonuclease P protein Rpp29*. Rna, 2004. **10**(9): p. 1423-32.

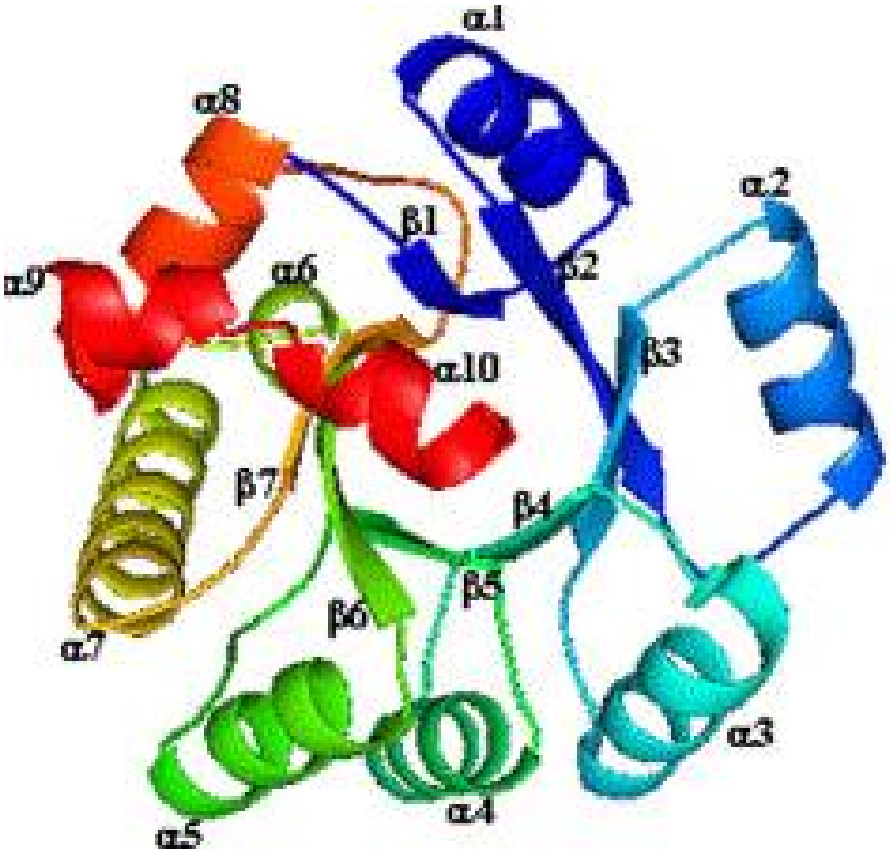


Figure 6: Ribbon structure of RNase P protein Ph1877p

Ribbon representation of the RNase P protein Ph1877p from *P. horikoshii*. Ph1877p is a homologue of human RNase P protein Rpp30 and is composed of ten α-helices and seven β-strands forming a TIM barrel.

`

Figure 6 Reference:

Kawano, S., et al., *Crystal structure of protein Ph1481p in complex with protein Ph1877p of archaeal RNase P from Pyrococcus horikoshii OT3: implication of dimer formation of the holoenzyme.* J Mol Biol, 2006. **357**(2): p. 583-91.

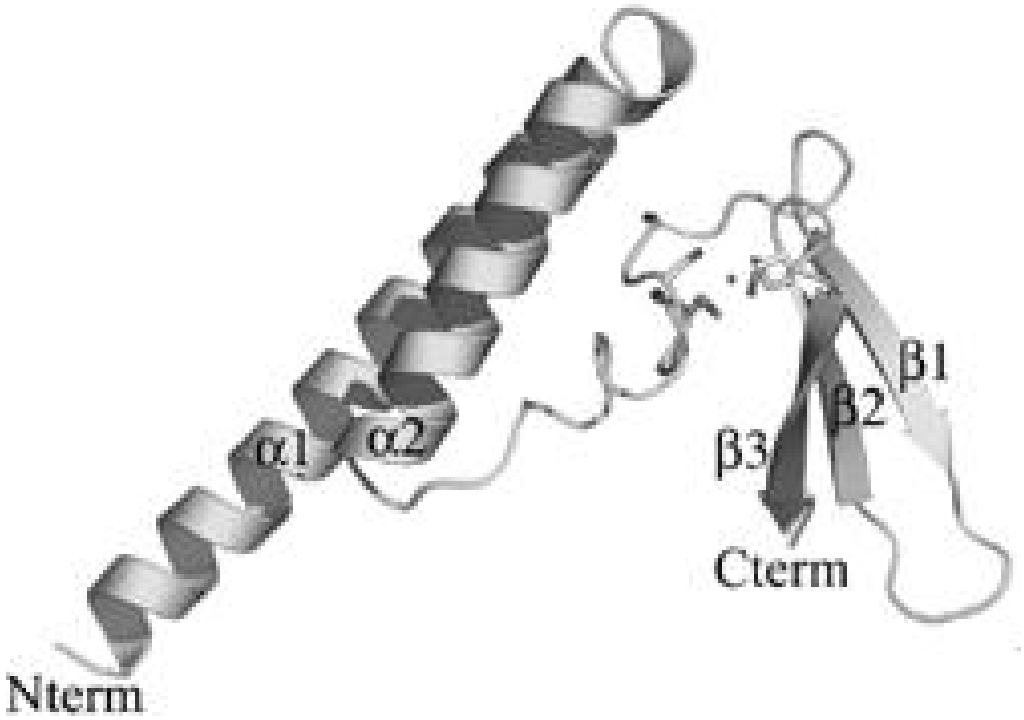


Figure 7: Ribbon structure of RNase P protein Ph1601p

Ribbon representation of the RNase P protein Ph1601p from *P. horikoshii*. Ph1601p is a homologue to human RNase P protein Rpp21 and is composed of an N-terminal domain comprised of two α-helices, while the central domain and C-terminal domain together form a zinc ribbon domain, giving the protein an L-shape.

Figure 7 Reference:

Kakuta, Y., et al., *Crystal structure of a ribonuclease P protein Ph1601p from Pyrococcus horikoshii OT3: an archaeal homologue of human nuclear ribonuclease P protein Rpp21*. Biochemistry, 2005. **44**(36): p. 12086-93

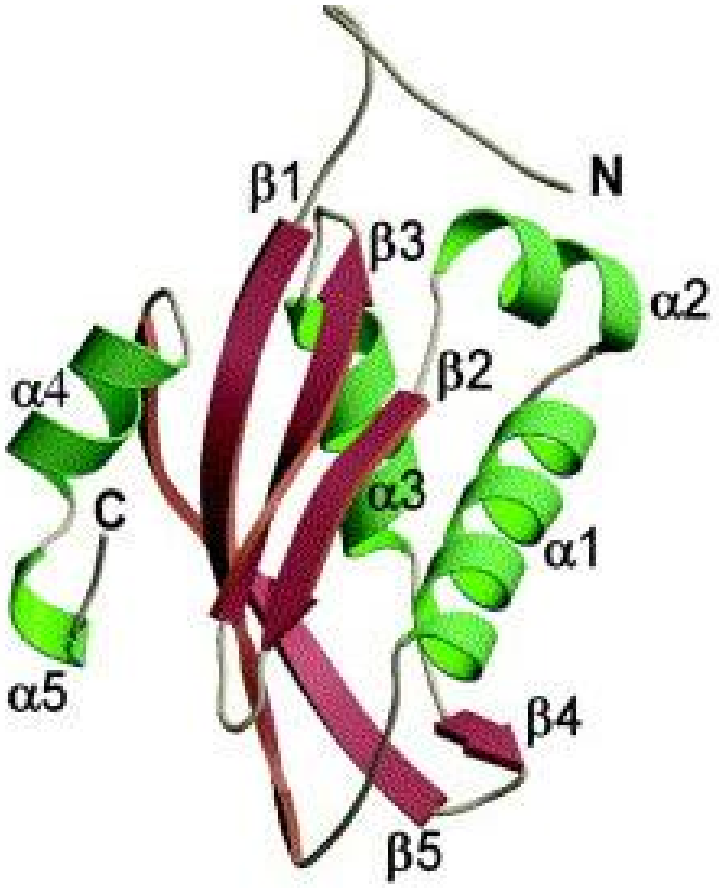


Figure 8: Ribbon structure of RNase P protein Ph1481p

Ribbon representation of the RNase P protein Ph1481p from *P. horikoshii.* Ph1481p is a homologue to human RNase P protein Pop5 and is composed of five antiparallel ß-sheets and five α-helices forming a α/β globular protein.

Figure 8 Reference:

Kawano, S., et al., *Crystal structure of protein Ph1481p in complex with protein Ph1877p of archaeal RNase P from Pyrococcus horikoshii OT3: implication of dimer formation of the holoenzyme.* J Mol Biol, 2006. **357**(2): p. 583-91.

CHAPTER 2: IS ALBA A RIBONUCLEASE P SUBUNIT?

Ellis J.C., Barnes J., Brown J.W., *Is Alba an RNase P subunit?*, Archaea (Manuscript in preparation)

ABSTRACT

It has been suggested that *Alba*, a well-established chromatin protein in Archaea, also has a role in the archaeal RNase P holoenzyme, because a homolog of this protein in humans has been shown to be associated with RNase P activity. Using the same biochemical methods we used previously to show that 4 other proteins homologous to eukaryotic RNase P proteins are *bona fide* RNase P subunits in Archaea, we could *not* detect any association of Mth1483p (the *Alba* homolog in *Methanothermobacter thermoautotrophicus*) with the RNase P holoenzyme. Taken together with recent data from RNase P reconstitution experiments and genetic deletions of chromatin proteins, we find no evidence that *Alba* is an RNase P subunit.

INTRODUCTION

Ribonuclease P is a ribonuclease best known for the processing of pre-tRNA by removing the 5′ leader sequences. [57] It is found in all three Domains of life, and in eukaryotes the nuclear enzyme is distinct from that found in mitochondria and plastids. [58-61] In all characterized instances, RNase P is composed of both RNA and protein subunits. RNase P is best understood in Bacteria, in which it is comprised of a single RNA (*rnp*B, a.k.a. M1) and a single small protein subunit (*rnp*A, a.k.a. C5). [62, 63] In all instances in which it has been tested, bacterial RNase P RNAs are capable *in vitro* (at elevated ionic strength) of cleaving pre-tRNAs in the absence of protein.

In contrast, the eukaryotic nuclear RNase P contains a single RNA, which is a distant homolog of the bacterial RNA, but many proteins. At least 9 proteins in *Homo sapiens* (hPOP1, Rpp29, hPOP5, Rpp20, Rpp30, Rpp21, Rpp40, Rpp25, and Rpp14) are physically associated with RNase P, and 10 proteins in *Saccharomyces cerevisiae* (Pop1-8, Rpp1, and Rpp2). The eukaryotic nuclear RNase P RNAs are absolutely dependent on protein for activity *in vitro* and, at least in *S. cerevisiae*, *in vivo* as well.

RNase P in Archaea is composed of a single RNA and 4 known proteins. The RNA subunit is remarkably similar in secondary structure (less so in sequence) to its bacterial homolog, and in most instances contain all of the structural and sequence elements in the phylogenetically conserved "core" of the bacterial RNA; those that do are likewise able *in vitro* to cleave pre-tRNA in the absence of protein at elevated (*very* elevated) ionic strength. [12] The 4 established RNase P proteins in Archaea are necessary and sufficient for reconstitution of enzymatic activity *in vitro* at moderate ionic

strength. [40, 64] These 4 proteins were originally identified as potential RNase P subunits based on their distant but discernable homology to 4 of the eukaryotic nuclear RNase P proteins: Pop5/hPop5, Pop4/Rpp29, Rpp1/Rpp30 and Rpr2/Rpp21. One of these proteins, the homolog of eukaryotic Rpp1/Rpp30, has a three-dimensional fold that resembles somewhat the conserved structure of the single bacterial RNase P protein, although there is no recognizable sequence similarity between these bacterial proteins and the archaeal/eukaryotic proteins. [25, 26] In addition to these 4 proteins, ribosomal protein L7 improves the thermostability of the reconstituted holoenzyme in *Pyrococcus furiosus*. [39] In addition to its role in the ribosome, L7 in Archaea is a component of box C/D snRNPs; its presence in RNase P *in vivo* has yet to be substantiated.

H. sapiens RNase P protein Rpp25 has no homolog in the *S. cerevisiae* enzyme, nor in the genome, but there is a homolog in Archaea, known as *Alba*. [43] This protein has been extensively characterized in *Sulfolobus acidocaldarius*, in which it is a dimeric, highly basic protein that binds cooperatively and at high density to DNA, inducing negative supercoiling. [65, 66] While the dense coating of Sulfolobus's DNA by Alba and negative supercoiling protects it from nuclease digestion, no significant compaction of genomic DNA has been observed. DNA binding affinity of Alba is regulated by *Sir2,* which deacetylates a key lysine, modulating gene expression. [41, 42, 67, 68] The structure of the N-terminus of *Alba* is similar to the DNA binding domain of DNase I, followed by a region of similarity to initiation factor domain IF3 at the C-terminus. [42] The IF3 RNA binding motif is a component of the YhbY group of RNA binding peptides. Although the role of *Alba* as a DNA binding protein is clear, it has been suggested that it has a dual role in Archaea, one in DNA binding and the other in RNase P, on the basis of

both it's similarity in sequence to the *H. sapiens* RNase P protein Rpp25, and the presence of both DNA- and RNA-binding structural motifs. [43] It has been conjectured that *Alba* may originally have been an RNA binding protein that has retained this function, but subsequently evolved DNA binding properties in Crenarchaea such as *Sulfolobus* (it was not known that *Alba* homologs also exist in Euryarchaea). [43]

One line of evidence in favor of the presence of additional protein(s) in RNase P in Archaea is the buoyant density of the *M. thermoautotrophicus* enzyme in cesium sulfate; at 1.42 and an RNA of 98 kDa, the known proteins (if present in only one copy each per RNA), an estimated 23 kDa of protein remains to be identified. [69]

Is *Alba* a component of the RNase P holoenzyme in Archaea? We examined this question using the same approaches that were used to confirm the presence of the 4 previously identified proteins in the archaeal RNase P, but could find no evidence for the physical association of this protein with the RNase P holoenzyme.

MATERIALS AND METHODS

All methods were substantially the same as previously described for the confirmation of Mth11p, Mth687p, Mth688p and Mth1618p as RNase P protein subunits.

RNase P cleavage assays

Enzyme samples were assayed for RNase P activity in10μL reactions containing 50mM Tris-Cl pH8.0, 20mM $MgCl_2$, 1M NH_4Cl and ca. 2 nCi ^{32}P-labeled pre-tRNA. Reactions were incubated at 50°C for 30 minutes unless otherwise specified, and products were separated by electrophoresis in 8% urea-polyacrylamide gels and visualized by autoradiography.

Production of antisera to Mth1483p

Antiserum against Mth1483p was produced by Cocalico Biologics from recombinant protein provided by Mark Foster, The Ohio State University.

Purification of M. thermoautotrophicus RNase P

1.8 g of *M. thermoautotrophicus* cell paste was ground in dry ice for 30 minutes using a mortar and pestle and resuspended in 5 mLs of TMGN-100 (50mM Tris pH 7.5, 10mM $MgCl_2$, 5% glycerol, 0.1mM DTT, 0.1mM PMSF, and N-100 denotes 100mM NH_4Cl) plus 6 µg/mL DNase I. The suspension was passed three times through a French Press at 20,000psi and centrifuged at 16,000 X g for 30 minutes at 4°C. The supernatant was dialyzed into TMGN-20, and loaded onto a 5mL DEAE-Trisacryl Plus column. After washing with 10mL of TMGN-20, RNase P was eluted with a 20 mL gradient of TMG with 20mM to 1000mM NH_4Cl. Ca. 0.75 mL fractions were collected and 5 µL fractions were assayed for RNase P activity. Active fractions were pooled, and 200uL samples were loaded onto 4.8mL 10-40% glycerol gradients (in TMGN-500 + 0.025% Nonident P-40) for centrifugation at 95,000 X g for 7.5 hours at 4°C in a Sorvall AH-60 swinging bucket rotor. Ca.150 µL fractions were collected from the bottom of the tubes and assayed for RNase P activity.

Western Blots for detection of Mth1483p

Five microliter samples of purified RNase P and 50ng of recombinant Mth1483p were separated by SDS-PAGE in 12.5% gels, and transferred to nitrocellulose membranes. After blocked overnight with 3% non-fat dry milk in TBS (10mM Tris pH

6.5, and 150mM NaCl). Blots were probed with a 1:1000 dilution of anti-Mth1483 serum in TBS + 0.3% Tween-20) for 2 hours at room temperature, washed 3 times with TBS for 10 minutes each, and probed with a 1:100,000 dilution of HRP-conjugated goat anti-rabbit IgG (Supersignal West Pico Rabbit IgG Detection Kit, Pierce #34083) made in TBS + 0.3% Tweeen-20. Blots were washed three times TBS + 1% Tween-20 and once with TBS for 10 minutes each. Membranes were visualized using the SuperSignal reagents according to the manufacturors instructions and images captured using Biomax ML film.

Immunoprecipitation of RNase P using antisera to Mth1483p

For each reaction, 20 mg of protein-A agarose were incubated with 500 µL of immune or preimmune serum overnight at 4°C, then washed twice and resuspended in 1 mL of 200mM sodium borate. 5mg dimethyl pimelimidate was added to each reaction and mixed for 30 minutes at room temperature. Beads were washed with 1 mL of 200mM ethanolamine pH 8.0 and then incubated in 1 mL of the same buffer for 2 hours. The beads were washed three times with 120uL of 10 mM Tris pH___ + 500mM NaCl, three times with 1 mL of TMGN-100 + 0.025% NP-40, and resuspended in 400µL of the same Buffer. 30µL of purified RNase P was added to each reaction, and mixed overnight at 4°C with shaking. The beads were pelleted by centrifugation, and the supernatant ("flow-through") was collected. The beads were washed four times with 1 mL of TMGN-100, and eluted with 3 30sec 20uL washs of TMGN-100 pre-heated to 72°C. 3uL samples of the final bead slurry, flow-through, and elutions were tested for RNase P activity.

RESULTS AND DISCUSSION

The Mth1483p does not co purify with RNase Pactivity.

Mth1483p was detected in western blots of glycerol gradient fractions of partially-purified RNase P from *M. thermoautotrophicus*, but the peaks of RNase P and Mth1483 protein are well-separated, with Mth1483p sedimenting more rapidly even than the RNase P holoenzyme (Fig. 1). We have shown previously that all 4 *bona fide* RNase P proteins copurify exactly with RNase P activity in these gradients. Mth1483p could not be detected in the purified RNase P post-glycerol-gradients, whereas Mth11p and Mth1618p were readily detected in this purified enzyme preparation.

Anti-Mth1483 does not Immunoprecipitate RNase P activity.

RNase P activity was not retained by protein A agarose treated with anti-Mth1483p immune serum (Fig. 2). In contrast, parallel reactions using anti-Mth11p serum efficiently retained RNase P activity compared to the pre-immune sera. We have previously shown that antisera against all 4 *bona fide* RNase P proteins immunoprecipitates RNase P activity from partially purified or purified RNase P preparations from *M. thermoautotrophicus.*

In addition to the results described above, two further lines of evidence do not support the hypothesis that *Alba* homologs are RNase P subunits in Archaea. The RNase P RNA and 4 known RNase P proteins are all necessary and sufficient to reconstitute RNase P activity in both the *M. thermoautotrophicus* and *Pyrococcus horikoshii* systems; neither Mth1483p or PHS053p (the *Alba* homolog in *Pyrococcus horikoshii*) are required for reconstitution of fully active RNase P enzymes *in vitro*. Although reconstitution of the

M. thermoautotrophicus enzyme is not a consistent system (due at least in part to poor solubility of two of the proteins and the requirement to remove the His-tags incorporated into the recombinant system), the addition of Mth1483p in reconstitution assays does not improve activity (V. Gopalan, personal communication). The second line of evidence against a role for *Alba* homologs in Archaea as RNase P subunits is that deletion of the gene for *Alba* (designated *alb*A) in *Methanococcus voltae* does not affect growth, but does affect the overall protein pattern in a way similar to deletion mutants of the other chromatin protein genes *hmv*A and *hst*B, encoding a histone-like protein and a histone, respectively. In contrast, deletion of any of the 10 RNase P proteins in *S. cerevisiae* is lethal.

In conclusion, although the *H. sapiens* homolog of the archaeal chromatin protein *Alba* has been shown to be a subunit of RNase P, the biochemical evidence does not support a second role for *Alba* in the RNase P holoenzyme. Given that other eukaryotic nuclear RNase P enzymes do not contain this protein, it seems more likely that this is a recruitment to the RNase P holoenzyme in a recent (in evolution scales) ancestor of humans rather than an ancient link between chromatin structure and RNA processing.

ACKNOWLEDGEMENTS

We thank Mark Foster, Ventkat Gopalan, and the Fermentation Facility, all at The Ohio State University, for the gifts of recombinant Mth1483 protein, unpublished data, and *M. thermoautotrophicus* cell paste, respectively.

REFERNCES

1. Bothwell, A.L., R.L. Garber, and S. Altman, *Nucleotide sequence and in vitro processing of a precursor molecule to Escherichia coli 4.5 S RNA.* J Biol Chem, 1976. **251**(23): p. 7709-16.
2. Hori, Y., T. Tanaka, and Y. Kikuchi, *In vitro cleavage of Drosophila 2S rRNA by M1 RNA.* Nucleic Acids Symp Ser, 2000(44): p. 93-4.
3. Ushida, C., D. Izawa, and A. Muto, *RNase P RNA of Mycoplasma capricolum.* Mol Biol Rep, 1995. **22**(2-3): p. 125-9.
4. Bothwell, A.L., B.C. Stark, and S. Altman, *Ribonuclease P substrate specificity: cleavage of a bacteriophage phi80-induced RNA.* Proc Natl Acad Sci U S A, 1976. **73**(6): p. 1912-6.
5. Yang, Y.H., et al., *Engineered external guide sequences are highly effective in inducing RNase P for inhibition of gene expression and replication of human cytomegalovirus.* Nucleic Acids Res, 2006. **34**(2): p. 575-83.
6. Gimple, O. and A. Schon, *In vitro and in vivo processing of cyanelle tmRNA by RNase P.* Biol Chem, 2001. **382**(10): p. 1421-9.
7. Buck, A.H., et al., *Structural perspective on the activation of RNAse P RNA by protein.* Nat Struct Mol Biol, 2005. **12**(11): p. 958-64.
8. Haas, E.S., et al., *Structure and evolution of ribonuclease P RNA in Gram-positive bacteria.* Nucleic Acids Res, 1996. **24**(23): p. 4775-82.
9. Walker, S.C. and D.R. Engelke, *Ribonuclease p: the evolution of an ancient RNA enzyme.* Crit Rev Biochem Mol Biol, 2006. **41**(2): p. 77-102.
10. Buck, A.H., et al., *Protein activation of a ribozyme: the role of bacterial RNase P protein.* Embo J, 2005. **24**(19): p. 3360-8.
11. Haas, E.S. and J.W. Brown, *Evolutionary variation in bacterial RNase P RNAs.* Nucleic Acids Res, 1998. **26**(18): p. 4093-9.
12. Pannucci, J.A., et al., *RNase P RNAs from some Archaea are catalytically active.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 7803-8.
13. Li, Y. and S. Altman, *In search of RNase P RNA from microbial genomes.* Rna, 2004. **10**(10): p. 1533-40.
14. Randau, L., M. Pearson, and D. Soll, *The complete set of tRNA species in Nanoarchaeum equitans.* FEBS Lett, 2005. **579**(13): p. 2945-7.
15. Waters, E., et al., *The genome of Nanoarchaeum equitans: insights into early archaeal evolution and derived parasitism.* Proc Natl Acad Sci U S A, 2003. **100**(22): p. 12984-8.
16. Randau, L., et al., *The heteromeric Nanoarchaeum equitans splicing endonuclease cleaves noncanonical bulge-helix-bulge motifs of joined tRNA halves.* Proc Natl Acad Sci U S A, 2005. **102**(50): p. 17934-9.
17. Di Giulio, M., *Nanoarchaeum equitans is a living fossil.* J Theor Biol, 2006.
18. Randau, L., et al., *Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves.* Nature, 2005. **433**(7025): p. 537-41.
19. Brown, J.W., *The Ribonuclease P Database.* Nucleic Acids Res, 1999. **27**(1): p. 314.
20. Harris, J.K., et al., *New insight into RNase P RNA structure from comparative analysis of the archaeal RNA.* Rna, 2001. **7**(2): p. 220-32.

21. Brown, J.W. and E.S. Haas, *Ribonuclease P structure and function in Archaea.* Mol Biol Rep, 1995. **22**(2-3): p. 131-4.
22. Svard, S.G., U. Kagardt, and L.A. Kirsebom, *Phylogenetic comparative mutational analysis of the base-pairing between RNase P RNA and its substrate.* Rna, 1996. **2**(5): p. 463-72.
23. Svard, S.G. and L.A. Kirsebom, *Several regions of a tRNA precursor determine the Escherichia coli RNase P cleavage site.* J Mol Biol, 1992. **227**(4): p. 1019-31.
24. Numata, T., et al., *Crystal structure of archaeal ribonuclease P protein Ph1771p from Pyrococcus horikoshii OT3: an archaeal homolog of eukaryotic ribonuclease P protein Rpp29.* Rna, 2004. **10**(9): p. 1423-32.
25. Takagi, H., et al., *Crystal structure of the ribonuclease P protein Ph1877p from hyperthermophilic archaeon Pyrococcus horikoshii OT3.* Biochem Biophys Res Commun, 2004. **319**(3): p. 787-94.
26. Kawano, S., et al., *Crystal structure of protein Ph1481p in complex with protein Ph1877p of archaeal RNase P from Pyrococcus horikoshii OT3: implication of dimer formation of the holoenzyme.* J Mol Biol, 2006. **357**(2): p. 583-91.
27. Kifusa, M., et al., *Protein-protein interactions in the subunits of ribonuclease P in the hyperthermophilic archaeon Pyrococcus horikoshii OT3.* Biosci Biotechnol Biochem, 2005. **69**(6): p. 1209-12.
28. Hall, T.A. and J.W. Brown, *Interactions between RNase P protein subunits in archaea.* Archaea, 2004. **1**(4): p. 247-54.
29. Kakuta, Y., et al., *Crystal structure of a ribonuclease P protein Ph1601p from Pyrococcus horikoshii OT3: an archaeal homologue of human nuclear ribonuclease P protein Rpp21.* Biochemistry, 2005. **44**(36): p. 12086-93.
30. Stams, T., et al., *Ribonuclease P protein structure: evolutionary origins in the translational apparatus.* Science, 1998. **280**(5364): p. 752-5.
31. Spitzfaden, C., et al., *The structure of ribonuclease P protein from Staphylococcus aureus reveals a unique binding site for single-stranded RNA.* J Mol Biol, 2000. **295**(1): p. 105-15.
32. Kazantsev, A.V., et al., *High-resolution structure of RNase P protein from Thermotoga maritima.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7497-502.
33. Wilson, R.C., et al., *Structure of Pfu Pop5, an archaeal RNase P protein.* Proc Natl Acad Sci U S A, 2006. **103**(4): p. 873-8.
34. Wittmann-Liebold, B., *The Ribosome: Structure, Function, and Evolution,*, in *American Society for Microbiology, Washington, DC,*. 1990. p. 598-616.
35. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.
36. Kuhn, J.F., E.J. Tran, and E.S. Maxwell, *Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein.* Nucleic Acids Res, 2002. **30**(4): p. 931-41.
37. Rozhdestvensky, T.S., et al., *Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea.* Nucleic Acids Res, 2003. **31**(3): p. 869-77.
38. Klein, D.J., et al., *The kink-turn: a new RNA secondary structure motif.* Embo J, 2001. **20**(15): p. 4214-21.

39. Fukuhara, H., et al., *A fifth protein subunit Ph1496p elevates the optimum temperature for the ribonuclease P activity from Pyrococcus horikoshii OT3.* Biochem Biophys Res Commun, 2006. **343**(3): p. 956-64.
40. Kouzuma, Y., et al., *Reconstitution of archaeal ribonuclease P from RNA and four protein components.* Biochem Biophys Res Commun, 2003. **306**(3): p. 666-73.
41. Bell, S.D., et al., *The interaction of Alba, a conserved archaeal chromatin protein, with Sir2 and its regulation by acetylation.* Science, 2002. **296**(5565): p. 148-51.
42. Wardleworth, B.N., et al., *Structure of Alba: an archaeal chromatin protein modulated by acetylation.* Embo J, 2002. **21**(17): p. 4654-62.
43. Aravind, L., L.M. Iyer, and V. Anantharaman, *The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism.* Genome Biol, 2003. **4**(10): p. R64.
44. Sandman, K. and J.N. Reeve, *Archaeal chromatin proteins: different structures but common function?* Curr Opin Microbiol, 2005. **8**(6): p. 656-61.
45. Chou, C.C., et al., *Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 Angstroms.* J Bacteriol, 2003. **185**(14): p. 4066-73.
46. Wang, G., et al., *Crystal structure of a DNA binding protein from the hyperthermophilic euryarchaeon Methanococcus jannaschii.* Protein Sci, 2003. **12**(12): p. 2815-22.
47. Ogasawara, N. and H. Yoshikawa, *Genes and their organization in the replication origin region of the bacterial chromosome.* Mol Microbiol, 1992. **6**(5): p. 629-34.
48. Salazar, L., et al., *Organization of the origins of replication of the chromosomes of Mycobacterium smegmatis, Mycobacterium leprae and Mycobacterium tuberculosis and isolation of a functional origin from M. smegmatis.* Mol Microbiol, 1996. **20**(2): p. 283-93.
49. Hansen, F.G., E.B. Hansen, and T. Atlung, *Physical mapping and nucleotide sequence of the rnpA gene that encodes the protein component of ribonuclease P in Escherichia coli.* Gene, 1985. **38**(1-3): p. 85-93.
50. Hansen, F.G., E.B. Hansen, and T. Atlung, *The nucleotide sequence of the dnaA gene promoter and of the adjacent rpmH gene, coding for the ribosomal protein L34, of Escherichia coli.* Embo J, 1982. **1**(9): p. 1043-8.
51. Panagiotidis, C.A., D. Drainas, and S.C. Huang, *Modulation of ribonuclease P expression in Escherichia coli by polyamines.* Int J Biochem, 1992. **24**(10): p. 1625-31.
52. Dong, H., L.A. Kirsebom, and L. Nilsson, *Growth rate regulation of 4.5 S RNA and M1 RNA the catalytic subunit of Escherichia coli RNase P.* J Mol Biol, 1996. **261**(3): p. 303-8.
53. Ramagopal, S., *Metabolic changes in ribosomes of Escherichia coli during prolonged culture in different media.* Eur J Biochem, 1984. **140**(2): p. 353-61.
54. Feltens, R., et al., *An unusual mechanism of bacterial gene expression revealed for the RNase P protein of Thermus strains.* Proc Natl Acad Sci U S A, 2003. **100**(10): p. 5724-9.

55. Withey, J.H. and D.I. Friedman, *The biological roles of trans-translation.* Curr Opin Microbiol, 2002. **5**(2): p. 154-9.
56. Berkhout, B., et al., *The amino terminal half of the MS2-coded lysis protein is dispensable for function: implications for our understanding of coding region overlaps.* Embo J, 1985. **4**(12): p. 3315-20.
57. Guerrier-Takada, C., et al., *The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme.* Cell, 1983. **35**(3 Pt 2): p. 849-57.
58. Doersen, C.J., et al., *Characterization of an RNase P activity from HeLa cell mitochondria. Comparison with the cytosol RNase P activity.* J Biol Chem, 1985. **260**(10): p. 5942-9.
59. Hollingsworth, M.J. and N.C. Martin, *RNase P activity in the mitochondria of Saccharomyces cerevisiae depends on both mitochondrion and nucleus-encoded components.* Mol Cell Biol, 1986. **6**(4): p. 1058-64.
60. Shevelev, E.L., et al., *Ribonuclease-P RNA gene of the plastid chromosome from Cyanophora paradoxa.* DNA Res, 1995. **2**(5): p. 231-4.
61. Hess, W.R., C. Fingerhut, and A. Schon, *RNase P RNA from Prochlorococcus marinus: contribution of substrate domains to recognition by a cyanobacterial ribozyme.* FEBS Lett, 1998. **431**(2): p. 138-42.
62. Gardiner, K. and N.R. Pace, *RNase P of Bacillus subtilis has a RNA component.* J Biol Chem, 1980. **255**(16): p. 7507-9.
63. Apirion, D. and N. Watson, *A second gene which affects the RNA processing enzyme ribonuclease P of Escherichia coli.* FEBS Lett, 1980. **110**(2): p. 161-3.
64. Boomershine, W.P., et al., *Structure of Mth11/Mth Rpp29, an essential protein subunit of archaeal and eukaryotic RNase P.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15398-403.
65. Biyani, K., et al., *Solution structure, stability, and nucleic acid binding of the hyperthermophile protein Sso10b2.* Biochemistry, 2005. **44**(43): p. 14217-30.
66. Edmondson, S.P., et al., *Characterization of Sac10a, a hyperthermophile DNA-binding protein from Sulfolobus acidocaldarius.* Biochemistry, 2004. **43**(41): p. 13026-36.
67. Marsh, V.L., S.Y. Peak-Chew, and S.D. Bell, *Sir2 and the acetyltransferase, Pat, regulate the archaeal chromatin protein, Alba.* J Biol Chem, 2005. **280**(22): p. 21122-8.
68. Zhao, K., X. Chai, and R. Marmorstein, *Structure of a Sir2 substrate, Alba, reveals a mechanism for deacetylation-induced enhancement of DNA binding.* J Biol Chem, 2003. **278**(28): p. 26071-7.
69. Hall, T.A. and J.W. Brown, *Archaeal RNase P has multiple protein subunits homologous to eukaryotic nuclear RNase P proteins.* Rna, 2002. **8**(3): p. 296-306.
70. Burn, D.M., *The digestive strategy and efficiency of the West Indian manatee, Trichechus manatus.* Comp Biochem Physiol A, 1986. **85**(1): p. 139-42.
71. Rommel, S. and J.E. Reynolds, 3rd, *Diaphragm structure and function in the Florida manatee (Trichechus manatus latirostris).* Anat Rec, 2000. **259**(1): p. 41-51.
72. Edwards, J.E., et al., *Influence of flavomycin on ruminal fermentation and microbial populations in sheep.* Microbiology, 2005. **151**(Pt 3): p. 717-25.

73. Tajima, K., et al., *Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR.* Appl Environ Microbiol, 2001. **67**(6): p. 2766-74.
74. Walsh, C.J., C.A. Luer, and D.R. Noyes, *Effects of environmental stressors on lymphocyte proliferation in Florida manatees, Trichechus manatus latirostris.* Vet Immunol Immunopathol, 2005. **103**(3-4): p. 247-56.
75. Krause, D.O. and J.B. Russell, *How many ruminal bacteria are there?* J Dairy Sci, 1996. **79**(8): p. 1467-75.
76. Sakata, S., et al., *Culture-independent analysis of fecal microbiota in infants, with special reference to Bifidobacterium species.* FEMS Microbiol Lett, 2005. **243**(2): p. 417-23.
77. Wang, M., et al., *T-RFLP combined with principal component analysis and 16S rRNA gene sequencing: an effective strategy for comparison of fecal microbiota in infants of different ages.* J Microbiol Methods, 2004. **59**(1): p. 53-69.
78. Sakamoto, M., et al., *Changes in oral microbial profiles after periodontal treatment as determined by molecular analysis of 16S rRNA genes.* J Med Microbiol, 2004. **53**(Pt 6): p. 563-71.
79. Sakamoto, M., M. Umeda, and Y. Benno, *Molecular analysis of human oral microbiota.* J Periodontal Res, 2005. **40**(3): p. 277-85.
80. Jernberg, C., et al., *Monitoring of antibiotic-induced alterations in the human intestinal microflora and detection of probiotic strains by use of terminal restriction fragment length polymorphism.* Appl Environ Microbiol, 2005. **71**(1): p. 501-6.
81. Rogers, G.B., et al., *Bacterial activity in cystic fibrosis lung infections.* Respir Res, 2005. **6**(1): p. 49.
82. Mengoni, A., et al., *Comparison of 16S rRNA and 16S rDNA T-RFLP approaches to study bacterial communities in soil microcosms treated with chromate as perturbing agent.* Microb Ecol, 2005. **50**(3): p. 375-84.
83. Chin, J., *Intestinal microflora: negotiating health outcomes with the warring community within us.* Asia Pac J Clin Nutr, 2004. **13**(Suppl): p. S24-5.
84. Gong, J., et al., *Diversity and phylogenetic analysis of bacteria in the mucosa of chicken ceca and comparison with bacteria in the cecal lumen.* FEMS Microbiol Lett, 2002. **208**(1): p. 1-7.
85. Kent, A.D., et al., *Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities.* Appl Environ Microbiol, 2003. **69**(11): p. 6768-76.
86. Reynolds, J.E., 3rd and S.A. Rommel, *Structure and function of the gastrointestinal tract of the Florida manatee, Trichechus manatus latirostris.* Anat Rec, 1996. **245**(3): p. 539-58.
87. Gallivan, G.J., J.W. Kanwisher, and R.C. Best, *Heart rates and gas exchange in the Amazonian manatee (Trichechus inunguis) in relation to diving.* J Comp Physiol [B], 1986. **156**(3): p. 415-23.
88. Bergey, M. and H. Baier, *Lung mechanical properties in the West Indian Manatee (Trichechus manatus).* Respir Physiol, 1987. **68**(1): p. 63-75.
89. Flewelling, L.J., et al., *Brevetoxicosis: red tides and marine mammal mortalities.* Nature, 2005. **435**(7043): p. 755-6.

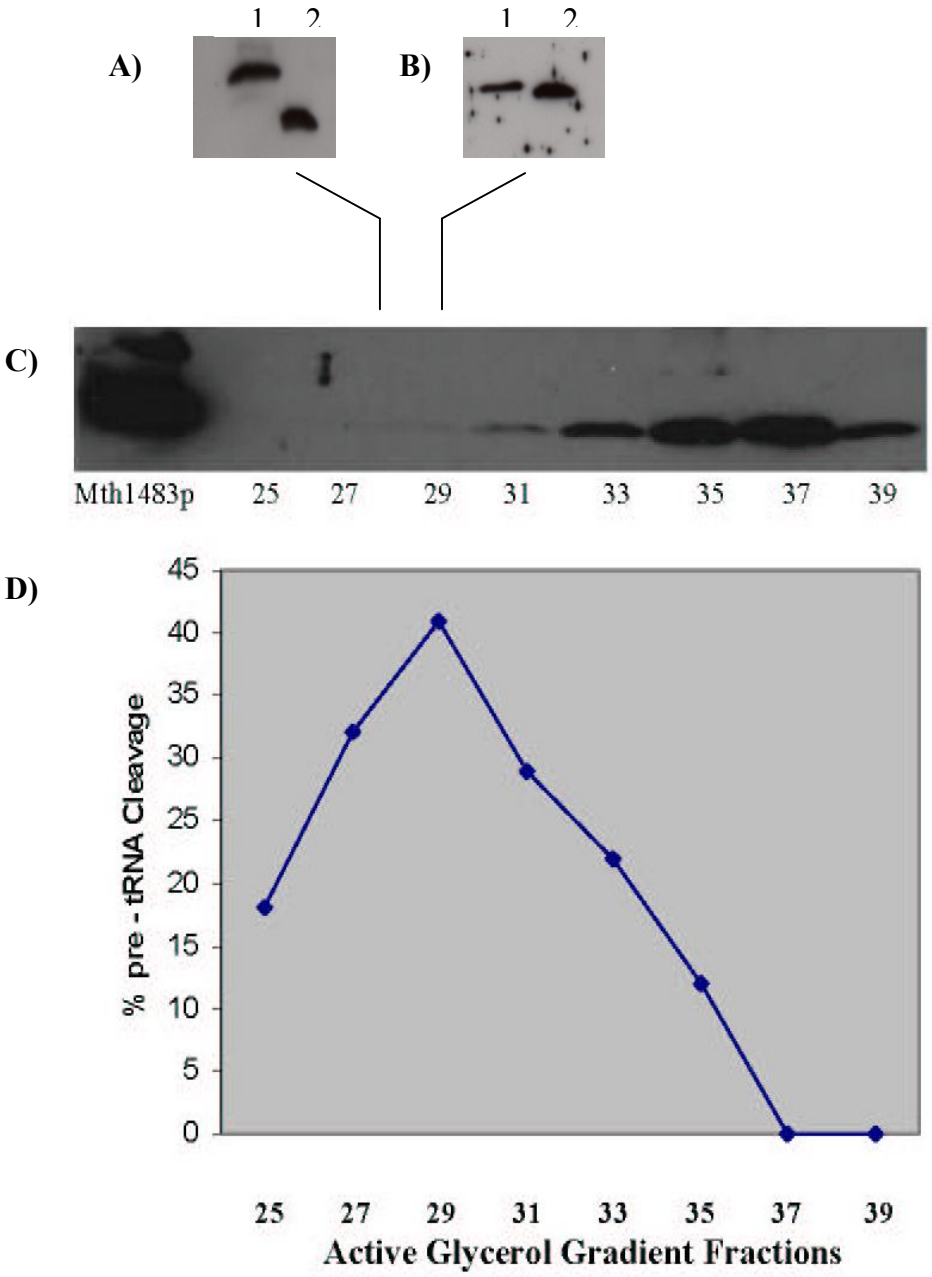


Figure 1: Mth1483p does not copurify with RNase P activity

(A) Western blot of Mth11p. Lane 1: 50ng purified Mth11p with his tag Lane 2: 5µL glycerol gradient fraction 29. (B) Western blot of Mth1618p. Lane 1: 50ng purified Mth1618p with his tag Lane 2: 5µL glycerol gradient fraction 29. (C) 5 µL of glycerol gradient fractions were probed with a 1:1000 of anti-Mth1483p. 50 ng of purified Mth1483p was run as a control (lane 1). (D) The level of RNase P activity in each fraction is shown with the corresponding level of Mth1483p detected in the western blot.

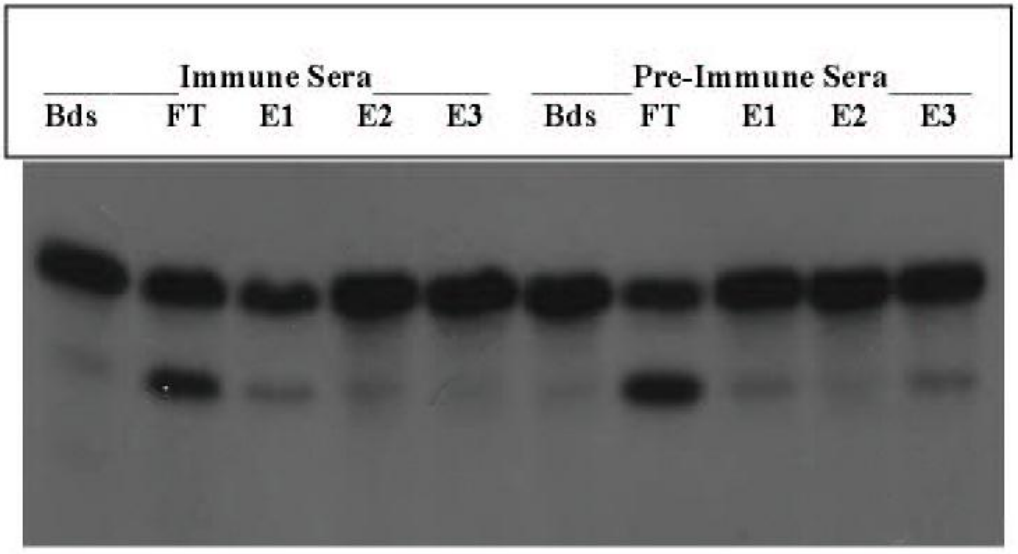


Figure 2: Immunoprecipitation of RNase P activity using anti-Mth1483p

Immune and pre-immune sera to Mth1483p was bound to protein-A agarose beads and mixed with glycerol gradient purified RNase P. Reactions were washed four times and eluted with three times with TMGN-100 heated to 72$_o$C. Beads and elutions were assayed for RNase P activity at 50$_o$C for 30 minutes.

CHAPTER 3: GENES WITHIN GENES WITHIN BACTERIA

Ellis J.C., Brown J.W., *Genes within genes within bacteria*. Trends Biochem Sci. 2003 **28**(10): 521-3

ABSTRACT

A recent paper by Ralph Feltens and colleagues describes an unusual gene structure in species of the genus *Thermus*, in which the *rpmH* (ribosomal protein L34) coding sequence was found to be entirely overlapped by the unusually large *rnpA* (RNase P protein subunit) sequence. Gene overlap is common in viruses, but has not been seen to this extent in any bacterium.

INTRODUCTION

In Bacteria generally, the gene encoding the protein subunit of RNase P (*rnpA*) is located immediately downstream of and in the same orientation as the gene encoding the ribosomal protein L34 (*rpmH*) (see Figure), and the two are located near the origin of replication [1-3]. This co-localization of genes in a wide range of bacterial genomes implies an important linkage in their regulation of expression, but the mechanism of this regulation has not been investigated. These genes in *E. coli* have been demonstrated to be part of the same operon, with two major and one minor promoter upstream of the *rpmH* gene, and two putative transcription termination signals downstream of *rpmH* [3-5]. The levels of expression of the encoded proteins are quite different; the ribosomal protein L34 is, of course, abundant in the cell, whereas the RNase P protein is not [6,7]. L34 is presumably produced in greater abundance because two of the three mRNA species transcribed from the operon do not include the RNase P protein coding region [5]. In addition, the mRNA that includes *rnpA* is produced at much lower levels, and the *rnpA* sequence utilizes codons that are uncommon in *E. coli*, presumably resulting in reduced translation efficiency [3,5].

RpmH and *rnpA* overlap in *Thermus*

A recent paper in the Proceedings of the National Academy of Sciences by Ralph Feltons *et al*. in Roland Hartmann's lab describes the unusual overlapping gene structure of *rpmH* and *rnpA* in Bacteria of the genus *Thermus* (see Figure) [8]. In *Thermus*, these genes begin with start codons separated by only 4 base pairs. The second of these start codons initiates the *rpmH* coding sequence, which is not unusual in any respect. The

upstream start codon initiates the *rnpA* coding sequence, in the same orientation but in the -1 register relative to *rpmH*. The *rnpA* open reading frame continues entirely through *rpmH*, some distance further, and finally includes the sequences homologous to those of the *rnpA* genes of other Bacteria. This overlapping *rnpA* gene results in an unusually long (163 amino acids in the case of *Thermus thermophilus*, compared to the usual ca. 120 amino acids), but functional, RNase P protein. This overlapping arrangement does not result in compaction of the genes, as is commonly the case in the overlapping genes of viral genomes; the regions of homologous sequences occur in the genome much as in other Bacteria, only the site of translation initiation of the *rnpA* gene has "moved" ahead of the upstream gene.

All of the species of *Thermus* that the authors investigated had completely overlapping *rnpA* and *rpmH* genes. The genes all started with the same organization of start codons, separated by 4 base pairs with that of *rnpA* in the lead. Both the *rpmH* and the downstream region of *rnpA* encode typical L34 and RNase P protein sequences, respectively. The length of the *rnpA* sequence varies in different species of *Thermus* in the 'intervening' region, always in multiples of 3 basepairs, and the amino acid sequences encoded in this region are quite variable, implying little or no functional constraint. Likewise, the region of *rnpA* sequence that overlaps *rpmH* seems to be conserved only with respect to the L34 encoded amino acids; the RNase P protein sequence encoded in this region is not otherwise conserved.

Implications for gene expression

The differential expression of *rnpA* and *rpmH* in *Thermus* is predicted to result from at least three different aspects of their unusual gene structure. The first is the

presence of one to four potential rho-dependent transcription termination signals following the *rpmH* sequence. If the ribosome(s) immediately following the RNA polymerase are translated in the *rpmH* reading frame, this would presumably trigger transcriptional termination at these sites, resulting in an mRNA encoding L34 but not RNase P protein. Any ribosomes translating in the *rnpA* reading frame of these terminated mRNAs would reach the end of the RNA without encountering a stop codon, before the functional part of the RNase P protein is reached; these ribosomes would (conceptually) require the tmRNP to direct release of the ribosome and target the non-functional truncated RNase P protein for protease degradation [9]. Secondly, the *rnpA* gene utilizes unusual codons; these unusual codons are primarily found in the region of overlap with the *rmpH* gene. This implies that the rate of translation for *rnpA* is reduced by ribosomal stalling and/or premature disassociation compared to *rpmH*. Thirdly, the distance of the single ribosome binding site (RBS) shared by these two genes is apparently suboptimally close (3 base pairs) to the *rnpA* start codon but optimally spaced (7 base pairs) to the *rpmH* start codon; competition between these start codons presumably favors translation of L34 over the RNase P protein.

Comparison to MS2 coat and lysis protein genes

Although not previously seen in Bacteria, gene overlap of this type is common in viruses. A gene arrangement similar to *rpmH* and *rnpA* in *Thermus* can be found in bacteriophage MS2. In this virus, the lysis protein overlaps (in a different reading frame) with the distal portion of the coat cistron and with the proximal portion of the replicase. The N-terminal 40 amino acids of the unusually-extended lysis protein are not necessary

for functionality and not well conserved among related phages [10]. It seems that the lysis protein overlap with the coat protein cistron is not required for additional coding capacity or additional functionality but couples synthesis of the lysis protein to the coat protein. Therefore, presumably nonfunctional amino acid "extensions" as a mechanism of translational control are found not only in viruses but also in the chromosome of some Bacteria.

Origin of the gene overlap – how could this happen?

The events generating this unusual gene structure are implied in the structure itself. The fundamental difference between these genes in *Thermus* and other Bacteria is the absence of the usual translational start site for the *rnpA*. RNase P is an essential enzyme, and so mutational inactivation of the *rnpA* translational start site would usually be a lethal event. Before the translational start site of *rnpA* could be inactivated in the ancestor of *Thermus*, several conditions must have pre-existed: there must have been a cryptic in-frame translational start site in its modern position upstream of the *rpmH* gene, there must not have been effective stop codons between this cryptic start site and the original *rnpA* sequence, and the arbitrary amino acid sequence extension must not have disabled the RNase P protein. The most difficult of these would seem to be the absence of stop codons in the *rpmH* gene and the intergenic region in the *rnpA* reading frame. However, the high G+C content of the genomes of *Thermus* species (60-69%) dramatically reduces the frequency of AUU, UAG, and UGA codons between genes and in non-coding reading frames. With these preconditions in place, a mutational event inactivating the translational start site of the *rnpA* gene in the ancestor of *Thermus* would

have been tolerated, and selective pressure would result in a re-optimization of the sequences based on the new gene structure.

ACKNOWLEDGEMENTS

Research in the authors lab is supported by NIH grant GM52894 to JWB, and is dedicated to the memory of Dr. Elizabeth Suzanne Haas (1957-2002).

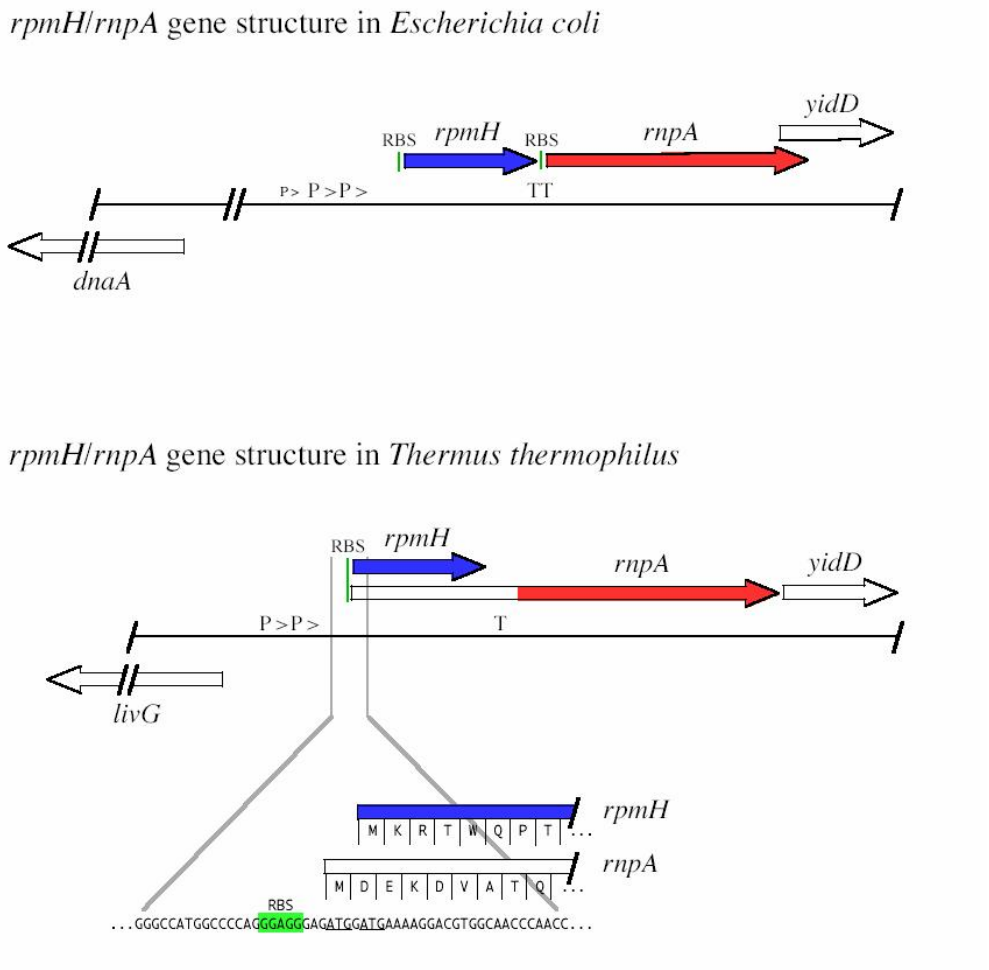


Figure 1: Comparison of the *rpmH/rnpA* gene structure common in Bacteria and in *Thermus*.

The usual gene structure in Bacteria is exemplified by that of *Escherichia coli* (above) [3-5], the overlapping gene structure of *Thermus* species by *Thermus thermophilus* (below) [8]. Promoters, putative transcription terminators, and ribosome binding sequences for *rpmH* (encoding ribosomal protein L34) and *rnpA* (encoding the protein subunit of RNase P) expression are indicated by "P" (large for major promoters, small of minor promoters), "T", and "RBS", respectively. Coding sequences are indicated by large arrows; homologous sequences in *rpmH* and *rnpA* are shaded blue and red, respectively. The region of translational initiation in *T. thermophilus* is expanded below; the start codons for *rpmH* and *rnpA* are underlined.

REFERENCES

1. Ogasawara, N. and Yoshikawa, H.(1992) Genes and their organization in the replication origin region of the bacterial chromosome. *Mol. Microbiol.* 6, 629-634.
2. Salazar, L., *et al*. (1996) Organization of the origins of replication of the chromosomes of *Mycobacterium smegmatis*, *Mycobacterium leprae* and *Mycobacterium tuberculosis* and isolation of a functional origin from *M. smegmatis*. *Mol. Microbiol*. 20, 283-293
3. Hansen, F. G., *et al*. (1985) Physical mapping and nucleotide sequence of the *rnpA* gene that encodes the protein component of ribonuclease P in *Escherichia coli*. *Gene* 38, 85-93.
4. Hansen, F. G., *et al*. (1982) The nucleotide sequence of the *dnaA* gene promoter and of the adjacent *rpmH* gene, coding for the ribosomal protein L34, of *Escherichia coli*. *EMBO J*. 1, 1043-1048.
5. Panagiotidis, C. A. *et al*. (1992) Modulation of ribonuclease P expression in *Escherichia coli* by polyamines. *Int. J. Biochem*. 24, 1625-1631.
6. Dong, H., *et al.* (1996) Growth rate regulation of 4.5 S RNA and M1 RNA the catalytic subunit of *Escherichia coli* RNase P. *J. Mol. Biol*. 3, 303-308.
7. Ramagopal, S. (1984) Metabolic changes in ribosomes of *Escherichia coli* during prolonged culture in different media. *Eur. J. Biochem.* 140, 353-361.
8. Feltens, R., *et al*., (2003) An unusual mechanism of bacterial gene expression revealed for the RNase P protein of *Thermus* strains. Proc. Natl. Acad. Sci. USA 100, 5724-5729.
9. Withey J.H. and Friedman D.I. (2002) The biological roles of trans-translation. *Curr Opin Microbiol.* 5(2),154-159.
10. Berkhout, B. *et al*. (1985) The amino terminal half of the MS2-coded lysis protein is dispensable for function: implications for our understanding of coding region overlaps. *EMBO J*. 4, 3315-3320.

CHAPTER 4: COMPARATIVE ANALYSIS OF RNA SECONDARY STRUCTURE: THE 6S RNA.

Brown, J.W. and Ellis, J.C., *Comparative analysis of RNA secondary structure: The 6S RNA*. Handbook of RNA Biochemistry {Wiley-VCH}, A. Bindereif, R. Hartmann, A. Schön, and E. Westhof, eds.

ABSTRACT

Comparative sequence analysis is the method of choice for determining secondary structure of RNAs. Despite attempts to automate this process, comparative analysis of RNA structure remains predominately a manual process. In this chapter, we describe the process of constructing an initial secondary structure of an RNA by comparative analysis, using the 6S RNA of *Escherichia coli* as an example.

INTRODUCTION

The purpose of this chapter is to provide a 'primer' on the comparative analysis of RNA secondary structure. The emphasis here is on the initial stages of the analysis; in other words, how one goes about creating a working model of the secondary structure *de novo* using the comparative approach. This is a common scenario; you, a student, or coworker, have discovered that an RNA is involved in a biological system under investigation. The sequence of the RNA is determined, usually either from cDNA or from the gene. Or perhaps it is discovered that a region of a messenger RNA or viral RNA is important in some process and it is suspected that the structure of this region is critical for that function. You are interested in obtaining information about the structure of this RNA in order to help guide experiments and to organize data about the RNA. The determination of the three-dimensional structure of the RNA is unlikely to be cost-effective or feasible (certainly not as a first step), but you correctly realize that the single most thermodynamically favorable predicted secondary structure is not going to suffice. How, then, to proceed? Usually, the answer is by creating a secondary structure model based on comparative sequence analysis. The detailed analysis of very high resolution secondary structure, the identification and evaluation of tertiary interactions, and the construction of three-dimensional models based on comparative analysis will not be considered here; these aspects of comparative analysis of RNA structure require specialized experience. Comparative analysis, like X-ray crystallography, is as much art

as science, but the creation of a basic secondary structure is well within the range of a newcomer to the "RNA World", the target audience for this chapter.

The approach taken here is to follow the construction of a basic secondary structure of an example RNA: the 6S RNA. The 6S RNA was discovered in *E. coli* in 1971 (Brownlee, 1971), but it's function remained unknown until very recently (Wassserman & Storz, 2000). The 6S RNA is not essential for viability (Lee, *et al.,* 1985), but accumulates during stationary phase, binds directly and specifically to RNA polymerase, and regulates RNA polymerase function in a growth stage specific manner (Wassserman & Storz, 2000). The secondary structure of the 6S RNA has not been examined in any detail; the existing secondary structure proposed for this RNA was based on a comparison of only the *E. coli* and *P. aeruginosa* sequences (Vogel, *et al*., 1987; Wassarman, *et l*., 1999).

RNA SECONDARY STRUCTURE

What *is* an RNA secondary structure? Although most researchers would agree on simple definitions of primary structure (sequence) and tertiary structure (three-dimensional coordinants), there is a surprising extent of disagreement about *exactly* what RNA secondary structure is, even (perhaps especially!) among established RNA researchers that work with secondary structures on a daily basis (Waugh, *et al*., 2002). At its most basic, however, a secondary structure is a list of adjacent, antiparallel Watson-

Crick (or G○U) base pairs in an RNA chain; these are the pairings for which the rules are clear and that are readily predicted by comparative sequence analysis. Uncertainty about what exactly is 'secondary structure' deal primarily with the distinction between secondary and tertiary interactions. For example, are non-Watson-Crick base pairs other than G○U included? Are isolated base pairs included? What about helical stacks? In the case of a pseudoknot, which helices are considered secondary and which, if any, are tertiary? All of these are subject to some level of disagreement. It is also worth remembering that 'secondary' does not mean two-dimensional; secondary structures contain a plethora of three-dimensional information, beginning with the presumption that the helices are generally A-form in structure. However, in the comparative analysis of secondary structure of an RNA, the basic definition of secondary structure is generally most useful.

Secondary structure can be represented in a variety of ways, but is most often presented as a string of letters, the sequence, twisted around on a page (*i.e.* in two dimensions) such that these antiparallel adjacent interactions can be shown as dashes between each pair of bases. By formal convention, G○U pairs are shown with a hollow dot instead of a dash, and non-Watson-Crick pairings with an closed dot, such as G•A (Leontis & Westhof, 2001). Typically (tRNA is the exception, here) structures are drawn to flow generally clockwise 5′ to 3′. A convenient way to specify whether or not there is

specific evidence for a base pairing is to only put in the dash (or dot, or circle) if there is such evidence.

RNA secondary structure is very specific and highly defined; secondary structure is the central organizing principle in RNA structure. This is a fundamental difference between protein and RNA (and DNA secondary structure, of course). Pragmatically, experiments are almost always developed and results represented in the context of the secondary structure of an RNA.

COMPARITIVE SEQUENCE ANALYSIS

Comparative sequence analysis is the process of extracting information about a macromolecule (in this case RNA) from the similarities and differences between different, but homologous, sequences (for review, see James, *et al*., 1989; Woese & Pace, 1993; Michel & Costa, 1998; Michel, *et al*., 2000). The underlying assumption is that the higher order structure of the molecule is more highly conserved than is the sequence; in other words, the sequence is free to change during evolution as long as the three-dimensional structure is generally maintained. In terms of secondary structure, this means that changes in the identity of a base involved in a pairing should generally be allowed by a compensatory change in its pairing partner so that the ability of the two to form isosteric base pairs is retained. The two bases that pair, then vary together, or *covary*. The work involved in the construction of a secondary structure of an RNA by comparative

analysis is primarily the search for these sequence *covariations*. If sufficient numbers of sequences are available, these covariations can be identified statistically directly from a sequence alignment (Winkler, *et al*, 1990; Chiu & Kolodziejczak, 1991). Comparative analysis, then, is an iterative process in which improvements in the alignment result in additional structural information, which can be used in turn to improve the alignment. Although attempts have been made to automate this process (see, for example, Parsch, *et al*., 2000; Juan & Wilson, 1999; Han & Kim, 1993), with varying levels of success, in practice this is generally still a manual process.

STRENGTHS AND WEAKNESSES OF COMPARATIVE ANALYSIS

Comparative analysis is the "Gold Standard" method for determining secondary structure of RNAs; computational methods for predicting secondary structure are typically validated by comparison with "true" secondary structures as determined by comparative analysis (see, for example, Zuker, *et al*., 1991; Fields & Gutell, 1996). However, other methods for determining secondary structure can be very useful supplements to comparative analysis, or serve as last resort alternatives if comparative analysis is not feasible, for example if few or only one sequence is available for analysis.

A particularly useful supplement to comparative analysis is the genetic analysis of mutation and second-site compensatory mutation; in fact, these methods are formally equivalent, the difference being whether you create the variations or observe them in

nature. This method is typically laborious, and so has not been used generally as an alternative to comparative analysis, but can be especially useful either to confirm the presence of a particular feature of secondary structure (for example, see Haas, *et al*., 1991), or to probe secondary structure than cannot be assessed by the comparative method, such as pairings involving invariant sequences. For instance, the 6S RNA secondary structure used as an example of comparative analysis in this chapter contains a stem/loop in which none of the base pairing are specifically supported by sequence covariations; the paired sequences are invariant among the sequences available. An alternative to obtaining additional 6S RNA sequences in hope of finding covariations in this potential stem/loop would be to make point mutations in this region of the RNA in *E. coli* that affect the function of the molecule, and then make the compensatory change. If the RNA with two substitutions, such that the potential base pairing is maintained, functions better than the RNA with either single-substitution that disrupts the potential pairing, then the pairing is presumed to be legitimate. Genetic analysis has also been used in the absence of comparative data in cases where only a single instance of the functional RNA is known, such as the delta virus ribozyme (Perrotta & Been, 1991).

Another useful supplement to comparative analysis, as we will see, is the prediction of structures thermodynamically. This is, in reality, where secondary structure modeling usually begins. These predictions are steadily improving, especially with the ability to predict a variety of structures near the minimum free energy and assess the frequency that

particular base pairings are predicted in these collections of structures (Jaeger, 1989; Jacobsson & Zuker, 1993; Zuker & Jacobson, 1998). Thermodynamic predictions are routinely used to predict the structures of idiosyncratic elements of structure that appear as insertions in specific instances of an RNA. The danger of thermodynamic prediction is the tendency to consider these structures endings rather than beginnings. A measure of the success of thermodynamic prediction is that the predicted lowest free energy structures contain, on average, about 73 % of base pairs that would exist in a "true" secondary structure determined by comparative analysis (Zuker, *et al*., 1991).

The last commonly used method for assessing secondary structure in RNA is chemical and enzymatic probing. Although these methods have been used extensively in attempt to determine structure, their utility is mostly in the examination of *changes* or *differences* in structure that result from mutation, binding to other molecules, and the like. Chemical and enzymatic probing data are notoriously difficult to judge directly in terms of the secondary structure of the RNA.

COMPARISON WITH OTHER METHODS

Comparative analysis is similar to but more sensitive than genetic experiments because natural selective pressure is more sensitive than our biochemical or genetic methods. Comparative analysis past the initial stages is objective, quantitative and conceptually automatable. Given sufficient numbers of variable sequences, a secondary

structure can be very high resolution, in which every base pair is assessed individually. Only biologically relevant base pairings are identified by comparative analysis. Nevertheless, there are limitations to comparative analysis. The most important of these is that no structure can be assessed in the absence of sequence variation; as a result, the most important aspects of structure, those comprised of the most highly-conserved sequences, are the most difficult to prove by comparative analysis. The initial stage of a comparative analysis, the subject of this chapter, is basically a manual process. No specific information is provided about unique sequences that cannot be meaningfully aligned. Although tertiary interactions can also be detected by comparative analysis (although this typically required large collections of sequences), only base-base interactions in which more than one isosteric possibility is structurally acceptable will be detected. Nevertheless, comparative analysis is certainly the method of choice whenever possible. The list of structures determined definitively by comparative analysis is nearly as long as the list of known RNA types: large and small subunit ribosomal RNAs (Gutell, *et al*., 1994), transfer RNAs (Zachau, *et al*, 1966), RNase P (Harris, *et al*., 2001) and MRP (Schmitt, *et al*, 1993) RNAs, SRP RNA (Larsen & Zwieb, 1991), tmRNA (Kelley, *et al*., 2001), group I (Michel, *et al*., 1990) and II (Michel, *et al*., 1989) introns, nuclear splicing RNAs (for example, Frank, *et al*., 1994), H/ACA (Ganot, *et al*., 1997) and box C/D (Ni, *et al*., 1997) snoRNAs, telomerase RNA (Chen, *et al*., 2000), etc, etc.

DESCRIPTION

Collecting sequence data

The raw material needed to determine the secondary structure of an RNA by comparative analysis is sequence data; more specifically, what is needed is a collection of different but functional and homologous sequences. There are two ready sources for sequences: nature, and GenBank (Wheeler, *et al*., 2000; Benson, *et al*., 2000). The first step, then, is to mine the available databases for homologous sequences. Very often there are sufficient numbers of suitable sequences available for the generation of at least an initial secondary structure. If this is not the case, or if a higher resolution secondary structure is desired, it will be necessary to obtain additional sequences experimentally.

A variety of approaches are needed to identify as many homologous sequences in Genbank as possible. A good starting point is to search the Genbank using BLAST (Altschul, *et al*., 2000) with your sequence of interest. In our example, the *E. coli* 6S RNA sequence (X01238) returned a number of other sequence records containing the *E. coli* 6S RNA:

```
   AE016766.1 Escherichia coli CFT073 section 12 of 18 of the complete genome

   X01238.1   E. coli 6S ribosomal RNA

   AE005521   Escherichia coli O157:H7 EDL933 genome, contig 3 of 3, section

140 of
```

AE000374 Escherichia coli K12 MG1655 section 264 of 400 of the complete genome

U28377.1 Escherichia coli K-12 genome; approximately 65 to 68 minutes

M12965.1 E.coli ssr gene encoding 6S RNA

AP002563.1 Escherichia coli O157:H7 DNA, complete genome, section 14/20

These are all identical to the original sequence, and so of no use to us. Please note that this needn't be the case; for some RNAs there may be useful variants in different strains of the same species. Other sequences obtained in this search were:

AE016988.1 Shigella flexneri 2a str. 2457T section 11 of 16 of the complete genome

AE015303.1 Shigella flexneri 2a str. 301 section 266 of 412 of the complete

AE016844.1 Salmonella enterica subsp. enterica serovar Typhi Ty2, section 11 of

AL627277.1 Salmonella enterica serovar Typhi (Salmonella typhi) strain CT18,

AE008840.1 Salmonella typhimurium LT2, section 144 of 220 of the complete genome

AE008841.1 Salmonella typhimurium LT2, section 145 of 220 of the complete genome

```
AE013931.1   Yersinia pestis KIM section 331 of 415 of the complete genome

AJ414145.1   Yersinia pestis strain CO92 complete genome; segment 5/20
```

The sequences from the two *Shigella flexneri* strains are identical, as are those of the two strains of *Salmonella enterica* and *Yersinia pestis*. The *Salmonella typhimurium* sequences represent the same sequence from the genome sequence, split in two by the separation of sections 144 and 145 of the genome record. This is frequently the case for RNA encoding genes because genome sequences are divided into sections with an eye toward larger "intergenic" regions (spaces between ORFs) that often turn out to be RNA-encoding genes. It is simply a matter of extracting the two fragments of sequence and merging them. For most of the other sequences, the entire sequence can be extracted simply by cutting and pasting from the BLAST results page. Sometimes, however, it is necessary to go to the original sequence record. For example, in the case of the *Yersinia pestis* sequence, the 3′ end of the sequence are different enough than that of *E. coli* that it was not returned in the BLAST alignment and had to be retrieved from the original.

Additional sequences can often be identified by repeating the BLAST searches with the sequences identified in the initial search. In the case of our example, however, a search using the most disparate sequence identified so far, that of *Yersinia pestis*, yielded the same list of sequences.

Another obvious approach is to search using the name of the RNA, but unlike protein-encoding genes, RNA encoding genes (except those of rRNAs and tRNAs) are

often not annotated even in genome sequences. Using "6S RNA" as the search term for our example locates the sequence from *Escherichia coli* (X01238), all of the sequences listed above, and that of *Pseudomonas aeruginosa* (Y00334). However, already we see one of the weaknesses of relying on sequence annotations; the *E. coli* 6S RNA is misannotated the "E. coli 6S ribosomal RNA"! In addition, a number of "6S RNA" sequence annotations are typographical errors where "16S RNA" was meant. A number of other matches are spurious because of the presence of the term "6S" in strain or clone names, enzyme name (e.g. the "6Fe-6S prismane cluster-containing protein"), or other RNAs with the same names (it seems there are different "6S RNA"s in vertebrates and in lambda). Annotations must always be scrutinized critically.

Nevertheless, the identification of the annotated 6S RNA sequence from *Pseudomonas aeruginosa* provides a fresh avenue for the search; a BLAST search using this sequence identified a homolog in the *Pseudomonas syringae* genome (AE016875) (as well as several instances of the *P. aeruginosa* sequence, of course). In addition, a weak match to the *P. aeruginosa* 6S RNA sequence was found in the *Pasteurella multocida* genome (AE006208); this region of the genome sequence was extracted and used, in turn, in a BLAST search that identified a homologous sequence is *Haemophilus influenzae* Rd (U32767). In the cases of both of these sequences, the ends of the RNA are not obvious from sequence similarity, and so a generous amount for sequence was taken from either end.

The 6S RNA is encoded by the *ssrS* gene, and has in the past been referred to as the "ssr RNA" (Lee, *et al*., 1985); a search of the Genbank using these terms did not identify any additional sequences.

A number of complete genome sequences are available for organisms that are related to those for which 6S RNA sequences had been identified but in which homologous sequences had not been found in general BLAST searches of the Genbank. The genomes of all of the gamma proteobacteria in which the 6S RNA had not yet been found were then searched individually from the NCBI genome-specific web pages in hopes of extracting additional sequences. (Phylogenetic information about these organisms can be found on the Taxonomy Browser at http://www.ncbi.nlm.nih.gov/Taxonomy/ {Wheeler, *et al*., 2000}) The *Pseudomonas putida* KT2440 (NC_002947) 6S RNA sequence was identified in a search using the *P. aeruginosa* sequence as the query; it is perhaps surprising that this sequence failed to be identified in the original search of the entire Genbank, but this is not unusual. More surprising still is that no 6S-like sequences could be identified in the complete genomes of *Vibrio cholerae, Vibrio parahaemolyticus* or *Vibrio vulnificus* using the sequence from the closely-related *E. coli* as the query. Nor could a 6S-like sequence be identified in the *Haemophilus ducreyi* complete genome sequence using the *H. influenzae* sequence as query.

Another source of sequences are secondary databases, such as, in this case, the Small RNA Database (http://mbcr.bcm.tmc.edu/smallRNA/) (Gu, *et al*., 1998), the Noncoding

RNAs Database (http://biobases.ibch.poznan.pl/ncRNA/) (Szmanski, *et al*., 2003), or the Washington University Rfam Database (http://rfam.wustl.edu/index.html) (Griffiths-Jones, *et al*., 2003). The first two of these include only the *E. coli* and *P. aeruginosa* sequences, but the Washington University Rfam site contains an alignment of 6S-like sequences from a number of bacterial genomes, including three that were not found in our previous searches: *Shewanella onedensis* (AE015522), *V. vulnificus* (AE016802), and *V. cholera* (AE004317). Using these sequences in turn to search the global Genbank and individual genome sequences using BLAST yielded a sequence from the *Vibrio parahaemolyticus* genome using the *V. vulnificus* (but surprisingly not the *V. cholerae*) sequence as the query.

At this point, 14 presumptive 6S RNA sequences have been identified and extracted from the Genbank. These sequences range from nearly identical (those of *E. coli* and *S. flexneri* differ by only one nucleotide) to less than 50% identical; a reasonable collection to begin a comparative analysis. It is important to have a wide range of sequence variation. The closely related sequences are useful because they are readily aligned and allow the initial identification of structure in the most variable parts of the RNA, but provide no useful information in the conservative regions of the sequence. The distantly related sequences are needed (often at later stages of the analysis) as a source of sequence variation for the analysis of the conservative (and therefore most important) regions of the RNA.

If additional sequences are needed, either because homologous sequences cannot be found by mining sequence databases or to increase the resolution of a secondary structure based on available sequences, they will have to be obtained experimentally. PCR amplification is typically used to obtain these sequences, but because the sequences flanking the gene are unlikely to be conserved, primers for amplification most often are within the gene itself, and so only partial sequences are obtained. Although partial sequences have been very useful in comparative analyses of RNA structure, the entire sequence can usually be obtained using a variety of technologies available in "kit" form. It is important to note that the primer target sequences at either end of a PCR product should *not* be used in a comparative analysis; these sequences are derived from the primers, not the target. A particularly useful approach to collecting large numbers of sequences quickly has been the use of PCR amplification from DNA extracted from complex microbial natural populations, rather than pure cultures (Brown, *et al.*, 1996). The amplification products are populations of sequences, and so must be separated by cloning, but hundreds of sequences can be obtained in a single experiment. The specie from which any particular sequence originates is unknown, but this information is unnecessary for the purposes of comparative analysis; all that matters is that the sequence is a valid sequence. In any case, the phylotype of the sequence itself can be determined after the fact by the construction of phylogenetic trees based on the final sequence alignment.

Thermodynamic predictions

It is useful, early in the process, to have the thermodynamic predictions of the structures of all of the RNA sequences in the collection. These are generated using mfold, most conveniently using the Mfold Web Server (Zuker, 1994) at http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi. For the purposes of initial comparative analysis, the default settings should suffice for most RNAs of reasonable length. If an unmanageable number of structures are predicted, the window parameter can be increased. If only 1 or 2 structures are generated, increase the percent suboptimality parameter to 10. The predicted structures can be downloaded as images for printing, but also download and print the energy table; this represents all of the predicted suboptimal foldings. Consistencies in these folding predictions among the different RNA sequences provide a starting point for comparative analysis.

In the case of the 6S RNAs, mfold consistently predicts pairing of the middle regions (roughly bases 60-130) of the RNAs in a stem/loop, and the two ends (the first and last ca. 20 nt) of the RNA as a terminal helix (see Figs. 1 and 2 for the predicted *E. coli* 6S RNA structures). The interior of this extended stem/loop structure is less consistently predicted. The most common alternatives for the central region of the RNA (between the consistently predicted terminal stem and the medial stem/loop) are base-pairing across this internal region such that the entire RNA would form an extended irregular hairpin, or the presence of local stem/loops on either side of the "conserved" central stem/loop. A

stem/loop on the 3′ side (position ca. 130-150) is predicted more frequently, and the placement of this predicted stem/loop is more consistent, than predictions on the 5′ side.

Initial alignment

A comparative analysis requires that the homologous sequences be aligned; in fact, it is the continuous building and refinement of the alignment that drives the structure analysis. Comparative analysis is an iterative process; additions to or improvements in the alignment result in additional structural information that, in turn, allows the alignment to be refined and provides insight required to add increasingly distantly related sequences to the analysis.

The first step, of course, is to collect all of these sequences into a sequence alignment editor. A variety of alignment editors are available for various computer platforms, and many of them are freely available from the authors. For Windows/PC computers, a particularly useful alignment editor and analysis program, available at no cost, is BioEdit (http://www.mbio.ncsu.edu/BioEdit/). Most commercial DNA manipulation and analysis software packages include an alignment editor. Because you will most often be adding sequences by extracting them from larger (often much larger) sequence records, it is usually most convenient to move them to the alignment editor by simply cutting and pasting. Retyping sequences manually, although it might seem to be a small task, is a last resort; any handtyped sequences will need to be painstakingly checked and rechecked for errors.

Once the sequences are all added to the alignment editor, they will need to be aligned preliminarily. If the sequences are all fairly closely related, this might be easily done "by eye", but generally one would use an automated method, CLUSTAL (Chenna, *et al*., 2003) being the most common method incorporated into most alignment editors. Note that this is your *initial* alignment, not your *final* alignment! Much of the work of a comparative analysis is the iterative improvement of the alignment. Even a novice can usually scan through a preliminary CLUSTAL alignment and find room for improvement. There is a fundamental difference between protein sequence alignments, which are generally based only on some maximizing measure of similarity between all pairs of sequences, and RNA alignments, that are based on the higher-order structure of the molecules. Ultimately, of course, the goal of any sequence alignment is to have homologous residues in alignment, but protein alignments attempt to achieve this by maximizing sequence similarity, because the richness of amino acid variation provides substantially more information on which to base an alignment than do the 4 bases in nucleic acid alignments. On the other hand, protein secondary structure is less informative than the highly organized secondary structures of RNA, which are based on one-to-one interactions between bases, and so RNA alignments are more easily based directly on higher-order structure.

Before proceeding further, it is important to arrange the sequences phylogenetically within the alignment (see Fig 3). The NCBI Taxonomy Browser web site

(http://www.ncbi.nlm.nih.gov/Taxonomy/) is a useful guide to general phylogenetic relationships. In our example, the 6S RNAs of *E. coli* and *S. flexneri* are nearly identical; they should therefore be adjacent in the alignment, as should the two sequences from *Salmonella* species (*S. typhimurium* and *S. enterica*), *Vibrio* species, and *Pseudomonas* species. *E. coli*, *S. flexneri* and *Salmonella* species form a larger cluster and so should be brought together, and likewise all of the sequences from the enteric Bacteria (the species just mentioned and *Vibrio* species) should be clustered. *Haemophilus* and *Pasteurella* are relatives, and so these two sequences belong together as well. It is convenient to have our prototype sequence, that of *E. coli*, at the top of the alignment, with increasingly distant sequences arranged downwards.

Terminal helix (P1a)

Pairing of the sequences near the 5′ and 3′ ends to form a terminal helix is a common element of RNA structure, and is a good starting point in the construction of the secondary structure of an RNA. Assuming that the ends of at least one example of the RNA of interest has been determined experimentally, the identification of a terminal helix allows the prediction of the location of the ends of the remaining RNAs in the alignment. In the case of our example, the 6S RNA, a terminal helix is also consistently predicted thermodynamically (Figs 1 and 2). In fact, all of the sequences in the collection are complementary near the ends, but the length of that complementarity varies somewhat and two sequences contain a bulged "A" interrupting this helix. Alignment of

the nucleotides on either strand of this helix is straightforward, however, on the basis of sequence conservation; only minor alteration of the CLUSTAL alignment is required to bring the bases in each position of the helix into the same columns (Fig 3). When aligned on the basis of sequence similarity, it becomes clear that the variation in helix length results from the addition of 2 bases to the distal ends of the helix (*i.e.* then end of the helix that contains the 5′ and 3′ tails) in *Pseudomonas* and *S. onedensis*. If the alignment is correct, there is a one to one correspondence in pairing partners in columns of the alignment; notice how the helix is opened by one column to accommodate the bulged "A" in two of the *Pseudomonas* sequences in the 3′strand of this helix. To solidify the specific base pairs and their homology among the sequences, a new line in the sequence alignment is added to hold right- and left-facing parenthesis to specify pairing partners (see Fig 3). Additional lines can be added to the alignment for annotations to make it easier to visualize the helices. Once the alignment of this terminal helix (if present) is finalized, the alignment can be trimmed at the ends to match the native ends of any RNAs in which the ends have been determined experimentally. In our example, the ends of the *P. multocida* and *H. influenzae* sequences could not be clearly defined on the basis of sequence conservation relative to either the *E. coli* or *P. aeruginosa* sequences (in which the native ends are known {Brownlee, 1971; Vogel, *et al*., 1987}), but these ends can now be predicted on the basis of the structure, and the alignment trimmed to match. The predicted terminal helix in these organisms is one base pair shorter (at the distal end) than

the other sequences other than those of *Pseudomonas* species and *S. onedensis*. Following the nomenclature used for group I introns and RNase P (Burke, *et al*., 1987; Haas, *et al*., 1994), we will call this helix "P1a"; P for "pairing", "1" because it is the first helix counting from the 5′ end, and "a" because, as we will see, P1 continues after an interruption.

Before moving on, it's important to evaluate in detail the evidence supporting the existence of this helix. The basic bits of evidence upon which secondary structures are built are sequence *covariations*. Two positions in an alignment are said to covary (for the purposes of secondary structure analysis) if *both* positions vary while maintaining the ability to form A-U, G-C, or G○U base pairs. Covariation of two base pairs in a potential helix is generally accepted as proof that the helix exists. In our example, the presence of the terminal helix P1a is supported by sequence covariation in most of the base pairs of the helix with only a few discrepancies. The ends of a helix can be harder to define; what is needed at each end of the helix is sequence covariations supporting pairing on the terminal base pair and clear failure of the adjacent 2 bases to covary arguing against their pairing. Ultimately, one would like to have evidence supporting the pairing of every base pair shown in a secondary structure; one useful way to denote how close you are to this is to only draw the line (or open dot, in the case of G○U pairs, by convention) connecting base pairs in a secondary structure if these positions in the alignment covary, *i.e*. if that individual base pair is supported by sequence covariation (see Fig 4). In our example, we

have covariation supporting the terminal base pair of the conserved structure and the consistent inability of the three 5′ and four 3′ nucleotides to pair, so this end of the helix is well defined, *except* in *Pseudomonas* and *S. onedensis*, in which this helix is potentially lengthened at this end by an additional base pairs and two flanking unpaired nucleotides (Fig 3). Whether or not this extra potential base pair is really paired is not clear, because the 5′ base is a C and the 3′ base always G in these sequences; in the absence of sequence variation, comparative analysis provides no evidence for or against the pairing. At the proximal end of this helix, the last base pair is likewise uncertain; the 5′ base is U or C, the 3′ base is an invariant G. This is consistent with their pairing, but does not constitute specific evidence for it. The penultimate base pair, on the other hand, is supported by covariation; this is typically G=C, but is A-U in *P. putida*. The G○U at this position in *P. syringae* constitutes a covariation neither with G=C nor with A-U, and so is not evidence for or against this pairing. The flanking two bases are often U and G, and so might be thought to pair, but these fail to covary and so should not be included as part of this helix; the 3′ G is invariant, and the 5′ base, although U in most sequences is a G in *P. putida* and an A in *S. onedensis*. There are 3 adjacent unsupported base pairs in the interior of this helix, but they are given provisional acceptance given that they have the potential to base pair and are flanked on both sides by well-supported pairings.

Subterminal helix (P1b)

The sequence immediately interior to the 11-13 base pairings that make up the terminal helix P1a of the 6S RNA cannot pair, but complementarity resumes after only a few bases on either side. The pairing of these sequences is predicted in most of the mfold structures from all of the sequences, although there are usually some idiosyncratic alternatives (see Figs 1 and 2). Adjustment of the alignment of the 5′ region of this potential helix is needed to accommodate an extra nucleotide present only in *V. vulnificus* and *V. parahemolyticus*; assignment of homology is straightforward if you keep conserved purines (G or A) or pyrimidines (U or C) aligned (Fig 3). There is reasonable covariation of all but one of the positions in the potential six base pair helix, which we will call "P1b", with an occasional mismatch and a bulged nucleotide representing the extra nucleotide in two *Vibrio* sequences.

Apical helix (P2a)

In addition to the consistently-predicted terminal helix, all of the 6S RNA sequences are predicted by mfold to have a stem/loop in the middle of the RNA containing some conservative sequence elements: CUCGG on the 5′ side, and CCGAG on the 3′ side (Figs 1 and 2). Attention is also draw to the potential pairing of these sequences because of the presence in most of the 6S RNAs of the conserved tetraloop sequences UNCG or CUYG (GNRA is the other conserved tetraloop motif; Woese & Gutell, 1993), although these would be unusual stem/loops in that the tetraloop sequences are in this case

followed (except in *Y. pestis)* by 1 or 2 extraneous nucleotides before the 3′ strand of the stem (Fig 3). Nevertheless, the alignment of these sequences is straightforward based on sequence conservation. The minor exception is the RNA of *H. influenzae*, which contains extra nucleotides between the tetraloop sequence (CUCG) and the conserved complementary sequences on both sides; these extra nucleotides are generally complementary and so would presumably create a terminal extension to the stem/loop. There are sequence covariations confirming all of the base pairs of the helix, which we will call "P2a", with the exception of the terminal pairing, which is always G=C except in the *H. influenza* sequence in which this is an A•C mismatch in the middle of the extension of this stem

Subapical helices (P2b and P2c)

Flanking the apical helix P1a are highly conserved sequences that are not complementary, and what sequence variation that does occur does not support the specific interaction of these sequences. However, flanking these conserved sequences in turn are variable sequences that are generally complementary (Fig 3). Variation in these regions makes them difficult to assign definitive homologies solely on the basis of sequence, but they are readily aligned, by default, simply based on their conserved distance from the flanking conserved sequences; the only adjustment necessary is the addition of a gap downstream of the apical helix corresponding to an obviously absent U in the *Vibrio* sequences. A similar gap is required in the *S. onedensis* sequence, although

the location of this gap is questionable. This helix is one base pair shorter in the *Pseudomonas* sequences than in the others; this deletion seems to be from the proximal end (furthest from the apex). There are covariations supporting all of the base pairings in this helix, which will be designated "P2b". One RNA, that of *S. onedensis*, has two non-Watson-Crick mispairs in P2b. Although these might seem to argue against the pairing of these sequences, the remaining sequences covary cleanly. More importantly, the mispairs in *S. onedensis* are adjacent G•A/A•G pairs, a three-dimensional motif that is sometimes is seen as an alternative to Watson-Crick pairs in helices and is known not to interrupt the flanking A-form helical structure (Gautheret, *et al*., 1994; Wu & Turner, 1996). Otherwise, there is only a single instance of a mispair between these two sequences (a C•C pairing in *P. putida*). The internal loop between P2b and P2a is nearly symmetrical, and is comprised of highly-conserved sequences, suggesting an important functional role.

Closely flanking P2b are additional complementary sequences. In this case, these sequences are so divergent from one group of sequences to another that alignment based on sequence is impossible (Fig 3). Furthermore, there is significant variation in the spacing between these complementary sequences and the surrounding conserved sequences and helix. Nevertheless, because of the conservation of general complementarity, the presence of this helix in the favorable structures predicted by mfold, and that fact that there is no apparent covariation between either of these sequence regions and anything else in the RNA, we will include this potential helix "P2c" in the

secondary structure on a preliminary basis. After aligning these sequences based on this complementarity, there is good covariation of all of the potential base pairings except that a G•A mispairing is present in the penultimate position of the helix in *E. coli* and its closest relatives (*S. flexneri* and *Salmonella* species).

Potential interior stem/loop (P3)

The secondary structure of the 6S RNA, as we understand it at this point, is an extended terminal helix P1 and an extended apical helix P2 flanking a central loop of as yet undefined structure. Very highly conserved sequences on either side of this loop have the potential to pair to form a nine base pair helix with an single bulged A. Mfold includes this helix in many of the structures it produced. However, the predictions in this region are *not* consistent; a number of equivalently favorable structures are predicted for each sequence, and the pairings predicted are idiosyncratic for each sequence (Figs 1 and 2). Only three sequence variants exist in this region; all three of these changes (an A to U in *S. onedensis* and a G to A in *H. influenzae* and *P. multocida*) disrupt the complemetarity between these sequences, arguing against their pairing (Fig 3).

A search for frequently-occurring tetraloop sequences (UNCG, CUYG, GNRA) (Woese & Gutell, 1990) flanked by complementary sequences, however, reveals an alternative; a potential stem/loop on the 3´ strand of this interior loop. This nine base pair stem would be composed of very highly conserved sequences; the only sequence variation among the potentially paired nucleotides is the change of a conserved C to U in

the terminal base pair in the *Pseudomonas* sequences (Figs 3 and 4). This is consistent with the pairing of this nucleotide with the conserved G opposite, but does not constitute specific evidence for that pairing. In the absence of evidence for any of the pairings in this helix, the conservative approach would be not to propose this helix until additional sequences with variation in this region can be obtained, or a genetic experiment performed. However, two aspects of the potential loop sequences argue for the provisional acceptance of this stem/loop, "P3". First, in the *Pseudomonas* sequence, this loop is a UUCG tetraloop motif, implying a stem/loop structure for at least these RNAs, and by extension the others as well given the conservation of this region. Second is the observation that all of the sequence and length variation in this region is very specifically located in this potential loop (and the closing base pair in *Pseudomonas*), and that variation is consistent with loop structure.

Is there anything else?

At some point in the initial analysis of the secondary structure of your RNA, you will reach a point where no additional structure is obvious. What else can you do in attempt to find structure? Perhaps the most useful approach is to draw all of the RNAs in the secondary structure as it is at this point, and compare them with an eye towards common potential helices; how useful this is likely to be depends on how much structure you've already gleamed; the more you already know the better. Some will find it convenient to draw a single 'reference' structure, and then annotate this with sequence variants.

Another fruitful approach is to go back to mfold and generate another round of structures, using the structure you already know as constraints, *i.e.* force the pairing of all of the helices you're sure of and see what structures are predicted in this context. These structures would then be scrutinized from sequence to sequence in search of commonalities, as before. If you've already identified a large part of the secondary structure, there are likely to be only a small number of favorable structures generated. Another useful approach is to generate a sequence logo from the alignment (a web server for this can be found at http://weblogo.berkeley.edu/logo.cgi) (Schneider & Stephens, 1990). This allows you to consider potential pairing in the context of sequence variation; sequences are expected to pair with other sequences with similar extents of variation. In the case of the 6S RNA, these methods failed to provide any additional insight into the structure, perhaps because so much structure has already been identified.

Where to go from here

Once you are satisfied that you have extracted all of the secondary structure information you can out of your sequences, you have a useful 'working model' for the structure of the RNA. Have you identified all of the base pairing in the RNA? Not likely. Do you have extraneous base pairs in the structure? Probably. What direction do you take from here? In the words of the Cheshire Cat, "That depends a good deal on where you want to get to" (Carroll, 1865). If this working model is sufficient for your needs, then you're finished. If you wish to learn more about the secondary structure of the RNA, or

identify tertiary interactions, then you will need to continue the comparative analysis after additional sequences are obtained.

The choice of where to get additional sequences depends on what you want to learn about the RNA, and how well the initial secondary structure analysis went. Most likely, you will want to know more about both the variable and most highly conserved regions of the RNA, and so you will want to obtain sequences that are closely and distantly related to those already in hand. Sequences similar to those already in hand are typically easy to obtain experimentally, and can often be obtained in large numbers. Distantly-related sequence are much harder to obtain, but are needed to provide details about the regions of the RNA that are most important for function and so are very highly conserved. Thermophiles are a good source of useful sequences; these RNAs typically contain the fewer irregularities than those of mesophiles, and are much better fodder for thermodynamic prediction (Jaeger, *et al*., 1989; Pace, *et al*., 1989; Brown, *et al*., 1993). With new sequences in hand, of course, you have the opportunity to mine the sequence databases again, and potentially identify sequences that were there all along but remained unrecognized.

TROUBLESHOOTING

Comparative analysis is a straightforward process, but as with any approach, it is possible to run into trouble. Below are listed some of the common problems that arise, and how you might try to get around them.

But I only have one sequence!

This *is* a major problem; you can't do a *comparative* analysis with only one sequence. However, this is the usual starting point; you're interested in the structure of a specific RNA from a specific organism, and that's the only one you have in hand. Usually you will be able to get at least one or two additional sequences from the genomes of related species. If data mining fails to yield the sequences you need, you have no choice but to get the sequences experimentally.

I just don't get it – how do I get started once I have some sequences?

If you're having trouble getting started, you can't seem to get a handle on the alignment, then reduce the problem by starting with a smaller collection of very similar sequences that you can align easily by eye. Look at every difference in the sequences – can you find one change that corresponds to another change that means to sequences could remain complementary? Again, start out, as we did in our example, by looking to see if the two ends of the RNA might form a helix. Also be on the lookout for common tetraloop sequences (UUCG, UCCG, GAGA, GAAA, GUGA, GUAA, GCGA, GCAA, CUUG; Woese & Gutell, 1990) flanked by complementary sequences – these are very likely to form stem/loops. Another approach is to start with the single best structure predicted by mfold, and then pick through each helix by comparative analysis to prove or disprove each one. You could then move on to the unique helices predicted in the less favorable structures, or the structures of the other sequences.

Some of my sequences have the most highly conserved sequences, but otherwise can't be aligned.

It is common to have sequences that you can only align to others in the conservative regions of the molecules, at least at first. In our example of the 6S RNAs, the *H. influenzae* and *P. multocida* sequences don't align well to the others at first. This is especially a problem for sequences that are quite different in length that the others; sometimes localizing the sites of the insertions or deletions can be difficult. This is usually best dealt with from both directions: aligning those regions you can to identify structure in common, and dividing the alignment into smaller groups of similar sequences to identify structure in the regions unique to each group. Once some insight on both is obtained, the alignments can be merged on the basis of structure rather than just sequence.

PCR or sequencing artifacts.

It is critically important that the sequences that go into a comparative analysis be valid. If you must enter sequences manually, check them very carefully for errors. The qualities of sequences are only as good as the abilities of the person or machine that did the sequence determination; there is *always* a chance that the sequence is incorrect. Genome sequences are usually reliable, but even here errors occasionally arise. In some cases, it might even be worth your effort to confirm an unusual sequence experimentally. As was seen above, sequence annotations are imperfect; if a sequence doesn't look like

what you expect, it probably isn't what you want, and even if it is you will not (yet) be able to use it. Generally speaking, the more recent the sequence, the less likely it is to contain errors

A common source of problematic sequences is PCR amplification, and there are two commonly seen types of these errors: point "mutations", and chimeras (Wang & Wang, 1996). Point mutations are a problem, but a limited one. These changes will most often appear in the analysis as the occasional mismatch or idiosyncrasy. Chimeric sequences are sequences that have been artificially spliced together during the amplification process; this is a common problem when amplifying genes from DNA extracted from microbial populations rather than pure cultures. Two aspects of chimeric RNAs usually reveal their nature; their failure to conform to long-range structure that is well-maintained among the remaining sequences, or their similarity to one sequence at one end of the RNA but a very different sequence at the other end. Suspected chimeric sequences should, of course, be removed from the analysis.

ACKNOWLEDGEMENTS

Research in the authors lab is supported by NIH grant GM52894 to JWB, and is dedicated to the memory of Dr. Elizabeth Suzanne Haas (1957-2002).

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.,* 215:403-410.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. and Wheeler, D. L. (2000). GenBank. *Nucl. Acids Res.,* 28:15-18.

Brown, J. W., Nolan, J. M., Haas, E. S., Rubio, M. A., Major, F. and Pace, N. R. (1996). Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Acad. Sci. USA*, 93:3001-3006.

Brownlee, G. G. (1971). Sequence of the 6S RNA of *E. coli*. *Nat. New Biol.,* 229:147-149.

Carroll, L. (1865). Alice's Adventures in Wonderland, available from the Gutenberg Project at http://www.cs.indiana.edu/metastuff/wonder/wonderdir.html

Cheena, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G. and Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucl. Acids Res.*, 31:3497-3500.

Chen, J. L. and Pace, N. R. (1997). Identification of the universally conserved core of ribonuclease P RNA. *RNA*, 3:557-560.

Chen, J. L., Blasco, M. A. and Greider, C. W. (2000). Secondary structure of vertebrate telomerase RNA. *Cell*, 100:503-514.

Chiu, D. K. and Kolodziejczak, T. (1991). Inferring consensus structure from nucleic acid sequences. *Comput. Appl. Biosci.*, 7:347-352.

Fields, D. S. and Gutell, R. R. (1996). An analysis of large rRNA sequences folded by a thermodynamic method. *Fold Des.*, 1:419-430.

Frank, D. N., Roiha, H. and Guthrie, C. (1994). Architecture of the U5 small nuclear RNA. *Mol. Cell Biol*., 14:2180-2190.

Ganot, P., Caizergues,-Ferrer, M. and Kiss, T. (1997). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, 11:941-956.

Gautheret, D., Damberger, S. H. and Gutell, R. R. (1994). A major family of motifs involving G•A mismatches in ribosomal RNA. *Nucl. Acids Res*., 242:1-8.

Gilbert, D. (2000). Free software in molecular biology for Macintosh and MS Windows computers. *Meth. Mol. Biol.*, 132:149-184.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S. R. (2003) Rfam: an RNA family database. *Nucl. Acids Res.,* 31:439-441.

Gu, J., Chen, Y. and Reddy, R. (1998). Small RNA Database. *Nucl. Acids Res.*, 26:160-162.

Gutell, R. R., Larsen, N. and Woese, C. R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.*, 58:10-26.

Haas, E. S., Brown, J. W., Pitulle, C. and Pace, N. R. (1994). Further perspective on the catalytic core and secondary structure of ribonuclease P RNA. *Proc. Natl. Acad. Sci. USA*, 91:2527-2531.

Haas, E. S., Morse, D. P., Brown, J. W., Schmidt, F. J. and Pace, N. R. (1991). Long-range structure in ribonuclease P RNA. *Science*, 254:853-856.

Han, K. and Kim H. J. (1993). Prediction of common folding structures of homologous RNAs. *Nucl. Acids Res*., 21:1251-1257.

Harris, J. K., Haas, E. S., Williams, D., Frank, D. N. and Brown, J. W. (2001). New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA*, 7:220-232.

Jacobson, A. B. and Zuker, M. (1993). Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol.*, 221:403-420.

Jaeger, J. A., Turner, D. H. and Zuker, M. (1989). Predicting optimal and suboptimal secondary structure for RNA. *Meth. Enzymol.*, 183:281-306.

James, B. D., Olsen, G. J. and Pace, N. R. (1989). Phylogenetic comparative analysis of RNA secondary structure. *Meth. Enzymol.*, 180:227-239.

Juan, V. and Wilson, C. (1999). RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol*., 289:935-947.

Kelley, S. T., Harris, J. K. and Pace, N. R. (2001). Evaluation and refinement of tmRNA structure using gene sequences from natural microbial communities. *RNA*, 7:1310-1316.

Larsen, N. and Zwieb, C. (1991). SRP-RNA sequence alignment and secondary structure. *Nucl. Acids Res.*, 19:209-215.

Lee, C. A., Fournier, M. J. and Beckwith, J. (1985). *Escherichia coli* 6S RNA is not essential for growth or protein secretion. *J. Bacteriol*., 161:1156-1161.

Leontis, N. B. and Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA.*, 7:499-512.

Michel, F. and Costa, M. (1998). Inferring RNA structure by phylogenetic and genetic analysis. In *RNA Structure and Function.* (eds. R. W. Simons and M. Grunberg-Managa), pp 175-202.

Michel, F. and Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol*., 216:585-610.

Michel, F., Costa, M., Massire, C. and Westhof, E. (2000). Modeling RNA tertiary structure from patterns of sequence variation. *Meth. Enzymol.*, 317:491-510.

Michel, F., Umesono, K. and Ozeki, H. (1989). Comparative and functional anatomy of group II catalytic introns - a review. *Gene*, 82:5-30.

Ni, J., Tien, A. L. and Fournier, M. J. (1997). Small nucleolar RNAs direct site-specific

synthesis of pseudouridine in ribosomal RNA. *Cell*, 89:391-400.

Pace, N. R., Smith, D. K., Olsen, G. J. and James, B. D. (1989). Phylogenetic comparative analysis and the secondary structure of ribonuclease P RNA - a review. *Gene*, 82:65-75.

Parsch, J., Braverman, J. M. and Stephan, W. (2000). Comparative sequence analysis and patterns of covariation in RNA secondary structure. *Genetics*, 154:909-921.

Perrotta, A. T. and Been, M. D. (1991). A pseudoknot-like structure required for efficient self-cleavage of hepatitis delta virus RNA. *Nature*, 350:434-436.

SantaLucia J. Jr. and Turner, D. H. (1996). Solution structure of (rGGCGAGCC)2 from NMR and restrained molecular dynamics. *Biochemistry*, 32:12612-12623.

Schmitt, M. F., Bennett, J. L., Dairaghi D. J. and Clayton, D. A. (1993). Secondary structure of RNase MRP RNA as predicted by phylogenetic comparison. *FASEB J.*, 7:208-213.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.*, 18:6097-6100.

Szmanski, M., Erdmann, V. A. and Barciszewski, J. (2003). Noncoding regulatory RNAs database. *Nucl. Acids Res.*, 31:429-431.

Vogel, D. W., Hartmann, R. K., Struck, J. C., Ulbrich, N. and Erdmann, V. A. (1987). The sequence of the 6S RNA gene of *Pseudomonas aeruginosa*. *Nucl. Acids Res.* 15:4583-4591.

Wang, G. C and Wang, Y. (1996). The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology*, 142:1107-1114.

Wassarman, K. M., Zhang, A. and Storz, G. (1999). Small RNAs in *Escherichia coli*. *Trends Microbiol*., 7:37-45.

Wasserman, K. M. and Storz, G. (2000). 6S RNA regulates *E. coli* RNA polymerase activity. *Cell*, 101:613-623.

Waugh, A., Gendron, P., Altman, R., Brown, J. W., Case, D., Gautheret, D., Harvey, S. C., Leontis, N., Westbrook, J., Westhof, E., Zuker, M. and Major, F. (2002). RNAML: a standard syntax for exchanging RNA information. *RNA*, 8:707-717

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A. and Rapp, B. A. (2000). Database resources of the National Center for Biotechnology Information. *Nucl. Acids Res*., 28:10-41.

Winkler, S., Overbeek, R., Woese, C. R., Olsen, G. J. and Pfluger, N. (1990). Structure detection through automated covariance search. *Comput. Appl. Biosci.*, 6:7-18.

Woese, C. R. and Gutell, R. R. (1990). Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. USA*, 87:8467-8471.

Woese, C.R. and N.R. Pace. (1993). Probing RNA structure, function and history by comparative analysis. In *The RNA World* (ed. R.F. Gesteland and J.F. Atkins), pp 91-117.

Zachau, H. G., Dutting, D., Feldmann, H., Melchers, F. and Karau, W. (1966). Serine specific transfer ribonucleic acids. XIV. Comparison of nucleotide sequences and secondary structure models. *Cold Spring Harb Symp Quant Biol.* 31:417-424.

Zuker, M. (1994). Mfold web server for nucleic acid folding and hybridization prediction. *Nucl. Acids Res.*, 31:3406-3415.

Zuker, M. and Jacobson, A. B. (1998). Using reliability information to annotate RNA secondary structures. *RNA*, 4:699-679.

Zuker, M., Jaeger, J. A. and Turner, D. H. (1991). A comparison of optimal and suboptimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. *Nucl. Acids Res.*, 19:2707-2714.

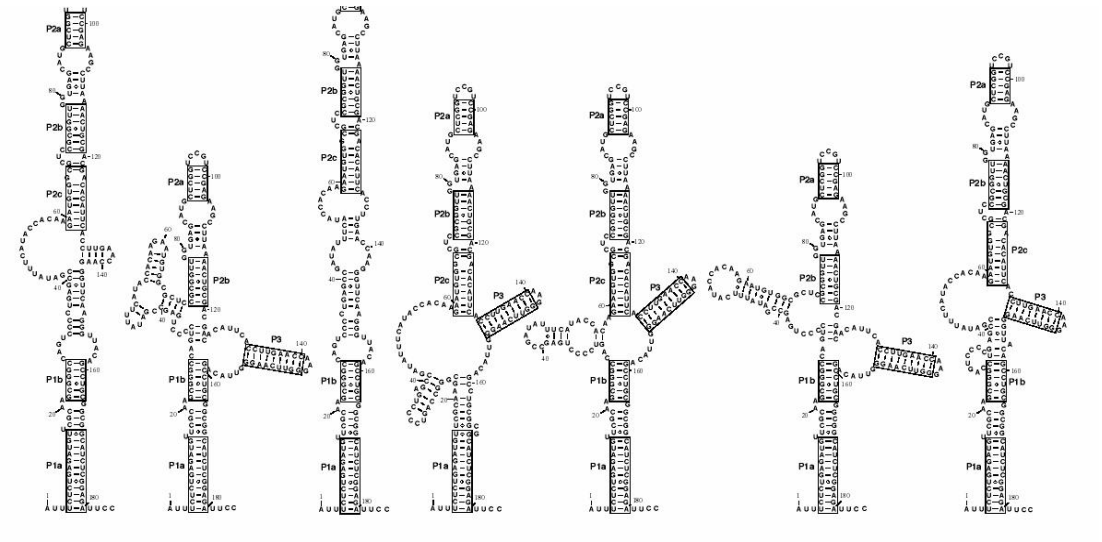


Figure 1: Potential structures of the *E. coli* 6S RNA predicted thermodynamically.

These are all of the structures predicted by mfold using the default parameters; in particular, only structures within 5% of the minimum free energy were allowed, and a window parameter (which defines how dissimilar two structures must be to be considered distinct) of 10 was used. Structures from left to right are from most to least favorable, respectively. Any base pairings in the helices identified in the comparative analysis (see Fig 4) are boxed. Structures were downloaded from the mfold server (Zuker, 1994) as connect (.ct) files, and displayed using LoopDloop (Gilbert 2000).

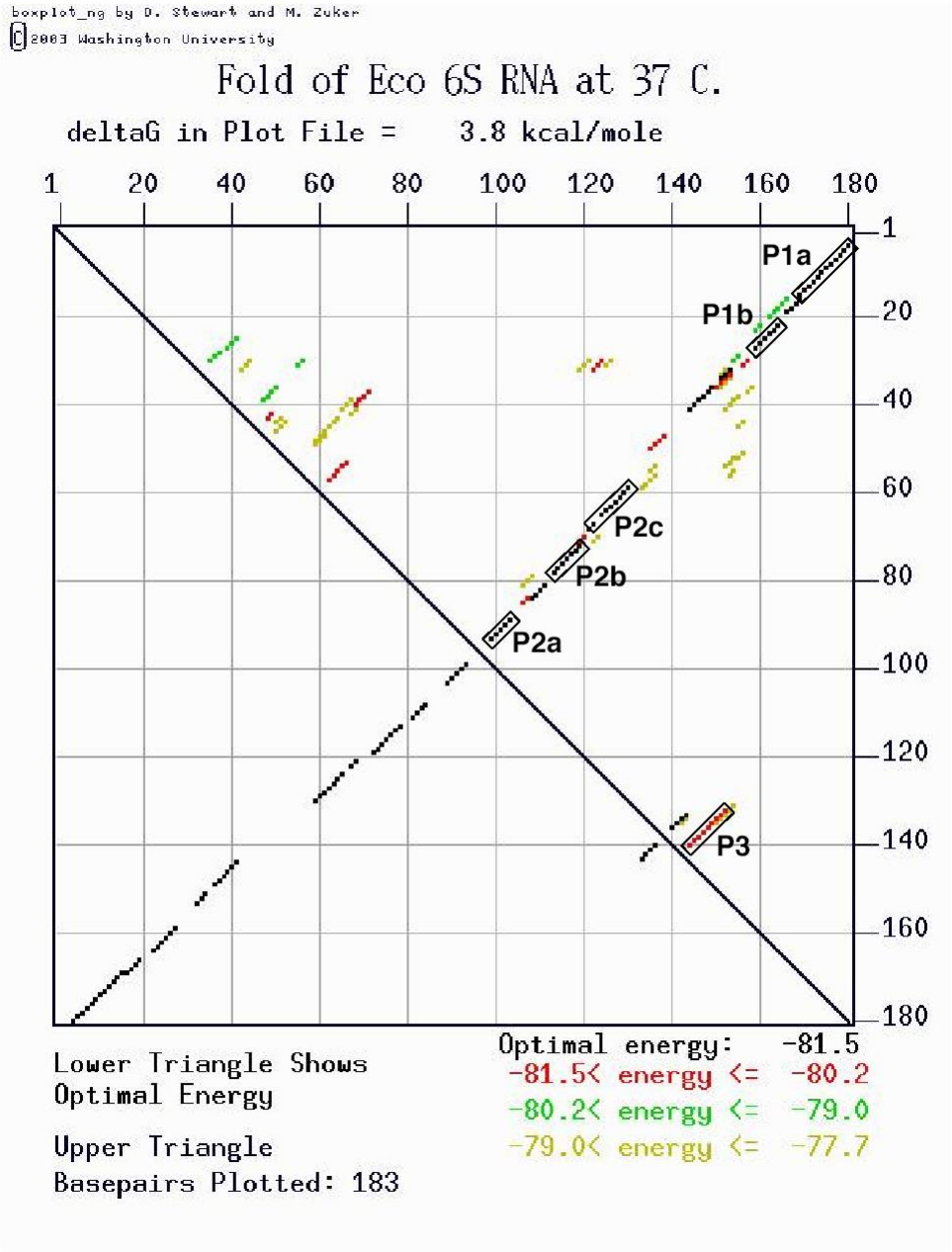


Figure 2: Energy table of the *E. coli* 6S RNA from mfold.

This represents the same structures shown in Fig 1. The X and Y axis each represent the sequence in the 5′ to 3′ direction. Each 'dot' indicates a predicted base pairing; the single best predicted structure (the lefthand-most structure in Fig 1) is shown below the diagonal, all of the predicted structures are show above the diagonal. Helices identified in the comparative analysis (see Fig 4) are boxed above the diagonal. This energy table was generated by the mfold server (Jacobson & Zuker, 1993; Zuker 1994)

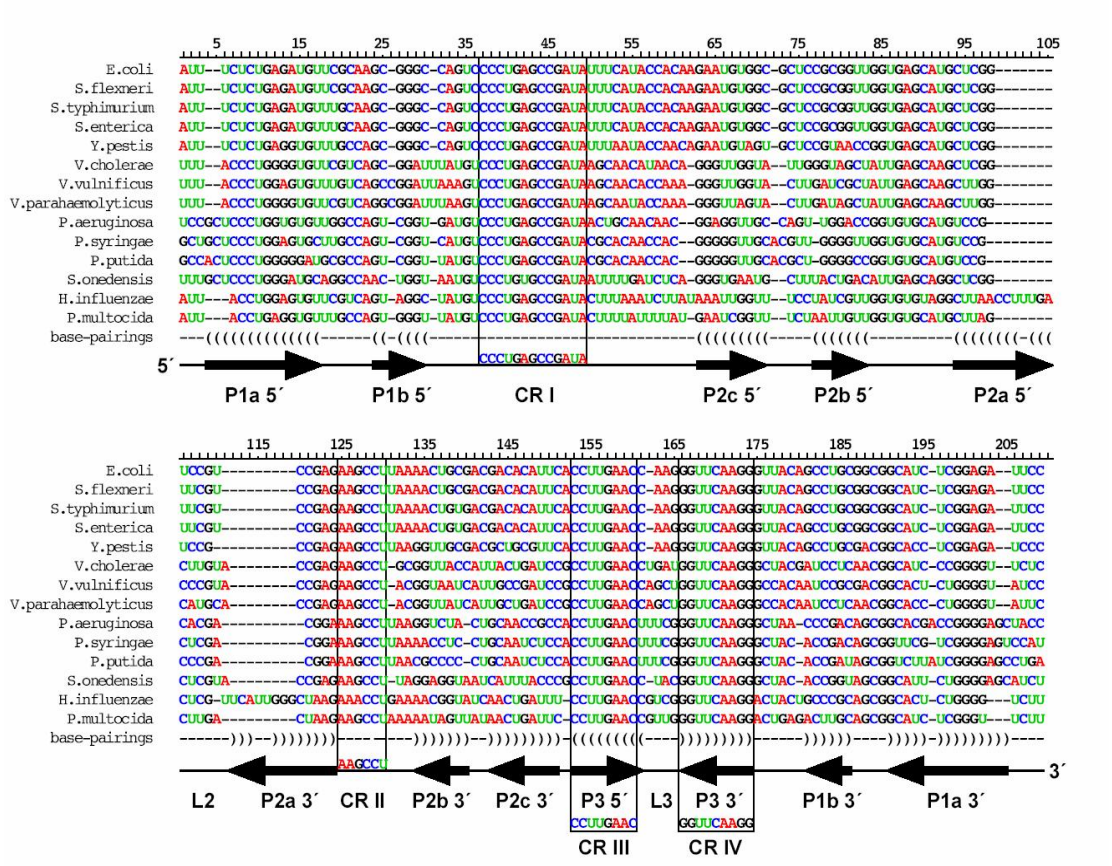


Figure 3: Alignment of 6S RNA sequences following comparative analysis.

Sequences are ordered phylogenetically (see text), and the alignment (*not* any particular sequence) is numbered at the top. The base pairing identified by comparative analysis are defined using parentheses in the last line. The structure is also shown diagrammatically at the bottom; the upstream (5´) and downstream (3´) nucleotides in each helix are shown with arrows. Four regions of absolutely conserved sequence longer than 5 nucleotides are labeled "Conserved Region" (CR) I – IV, as used for the RNase P RNA (Chen & Pace, 1997).

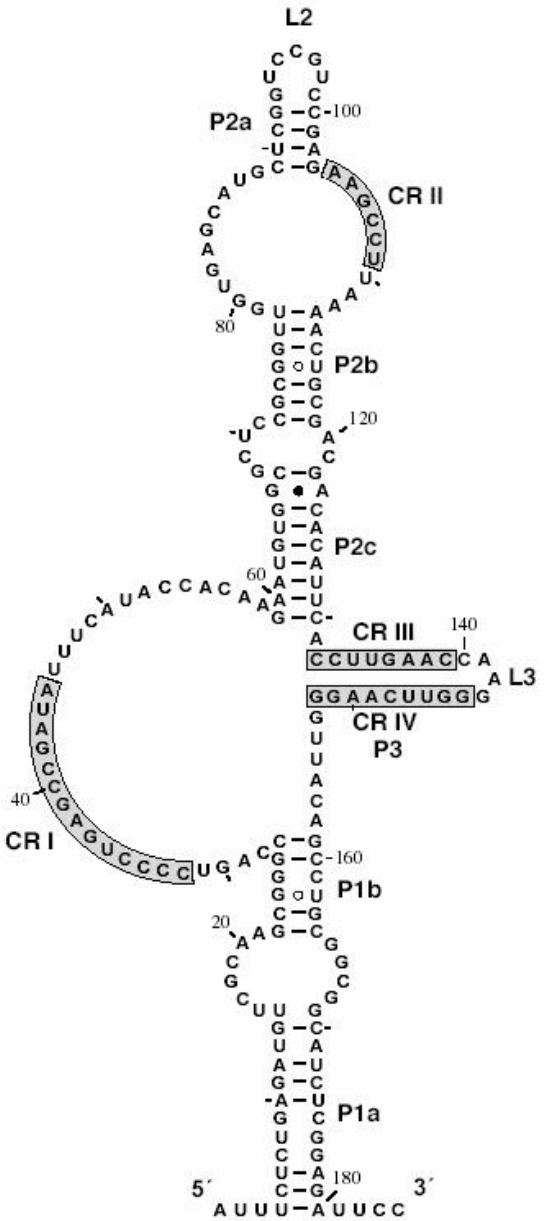


Figure 4: Secondary structure of the *E. coli* 6S RNA.

Helices are labeled as described in the text. Base pairings supported individually by sequence covariation are indicated by the connecting lines or dots; unsupported pairings lack these markers. The sequence is numbered 5′ to 3′ every 20 nucleotides, with a tick mark every 10 nucleotides. Four regions of absolutely conserved sequence longer than 5 nucleotides are boxed and labeled "Conserved Region" (CR) I – IV as in Fig 3. This structure was generated in connect (.ct) format directly from the alignment in Fig 3 using a Hypertalk script, and displayed using LoopDloop.

CHAPTER 5: THE SMALL NUCLEOLAR RIBONUCLEOPROTEIN (snoRNP) DATABASE.

Ellis J.C., Brown J.W., *The snoRNP Database*. RNA (Manuscript in preparation)

ABSTRACT

Small nucleolar ribonucleoproteins (snoRNPs) generally catalyze sequence specific 2'-O- ribose methylation and pseudouridylation of ribosomal RNAs. They are likely evolutionarily ancient, arising before the last common ancestor. Recent advancements in bioinformatics have resulted in new algorithms able to rapidly identify genes encoding the guide RNA components of snoRNPs in genome sequences and in rapidly expanding sequence databases. The snoRNP database at North Carolina State University (www.mbio.ncsu.edu/snoRNP/home.html) is a web-based compilation of snoRNP RNA and protein component sequences. The database currently contains over 6,300 snoRNP RNA sequences and over 300 snoRNP protein sequences from Bacteria, Archaea, and Eukaryotes.

BACKGROUND

The primary function of small nucleolar ribonucleoprotein particles (designated snoRNPs in Eukaryotes and sRNPs in Archaea) is to catalyze the 2'-O-ribose methylation (C/D snoRNPs) and pseudouridylation (H/ACA snoRNPs) of ribosomal RNAs. The RNA component of the snoRNP complex is responsible for specificity. Direct base-pairing between the snoRNA and the target rRNA provides site specificity for the generic catalytic protein subunits.

Ribosomal RNAs are not the only target for snoRNP-directed nucleoside modification; for example, in Archaea, some C/D snoRNPs direct 2′-O-methylation of tRNA [1]. In addition to their primary roles in 2′-O-methylation and pseudouridylation, some snoRNPs perform essential roles in pre-rRNA cleavages, while others function as rRNA chaperons during the assembly of ribosomes [2].

SnoRNPs primarily fall into two major groups based on functional and sequence/secondary structure motifs in their RNAs: boxes C/D and box H/ACA. The primary function of the H/ACA snoRNP complex is to convert uridine ribonucleotides of the pre-rRNA to pseudouridines. The H/ACA small nucleolar motif is composed of a two conserved sequence element the Hinge box (H) between the two stem elements and an ACA sequence at the 3' end of the RNA. In Eukaryotes, the H/ACA snoRNP complex contains four proteins in addition to the RNA: Nhp2p, Nhp10p, Gar1p, and Cbf5p (Dyskerin). The H/ACA sRNP complex in Archaea also contains four proteins: L7 (the Nhp2p homologue), Nhp10p, Gar1p, and Cbf5p.

The C/D box motif in eukaryotic and archaeal snoRNPs is defined by two terminal conserved sequences, box C (RUGAUGA) and box D (CUGA) and two internal C′/D′ motifs. The primary role C/D snoRNPs is the ribose 2'-O- methylation of pre-rRNA. In eukaryotes, the C/D snoRNP complex contains four proteins in addition to the RNA: Nhp2p, Nhp10p, Gar1p, and Cbf5p (dyskerin). The H/ACA sRNP complex in Archaea is also contains four proteins in addition to the RNA: L7, Nhp10p, Gar1p, and Cbf5p.

Recent advancements in RNA-specific search tools (e.g. snoScan and snoFPS) and the accelerating accumulation of sequence data has resulted in the identification of large numbers of novel snoRNP RNA and protein subunit sequences. [3-8]. The snoRNP database has been compiled to facilitate access to this information.

CONTENT OF DATABASE

The snoRNP Database is divided into two main sections: small nucleolar RNAs and small nucleolar proteins. Over 6,300 snoRNP RNA sequences are currently contained in the snoRNP Database at North Carolina State Unvirsity. Upon entering the site all sequences are presented in tabular format and sorted alphabetically by species in which the snoRNA is found. An additional feature of the database is it may also present the data in form format to aid in readability of individual sequences. Researchers may also limit the number and type of snoRNA by using the snoRNP search engine. The snoRNP database search engine allows the user to search based eight search criteria: Type, Genus, Species, Accession_number, Molecular_class, Molecular_family, Citation, and Sequence. The user may also exclude from the search unwanted sequences by using

the drop down menu and choosing “does not contain”. The snoRNP database also contains more than 300 snoRNP protein sequences are. The snoRNP associated proteins are divided into four subsections: archaeal H/ACA sRNA proteins (Nhp10p, Gar1p, Cbf5p), eukaryotic H/ACA snoRNA proteins (Nhp2p, Nhp10p, Gar1p, Cbf5p), archaeal C/D sRNA proteins (Nop56/58p, fibrillarin), and eukaryotic C/D snoRNA proteins (Snu13p, Nop56p, Nop58p, fibrillarin). L7 was not included in the archaeal H/ACA sRNP proteins or the archaeal C/D sRNP proteins because of the uncertainty in separating the bifunctional L7 sRNP/ribosomal sequences from the solely ribosomal L7 sequences. Each group of proteins and a sequence alignment is listed for each subsection.

DATABASE GENERATION

All sequences were initially extracted from the National Center of Biological Institute (NCBI) database using the following search terms: snoRNA, sRNA, snoRNP, small nucleolar RNA, fibrillarin, Nhp10, Gar1p, Cbf5p, Nop56/58p, Snu13p, and Nop58p. Primarily Perl but some C# scripts were used to excise snoRNA sequences, extract the GI and Accession numbers, publications, start and stop sites for each snoRNA sequence, and remove all unrelated sequences. After the small nucleolar associated proteins were extracted, a multiple sequence alignment was generated using CLUSTALW, and unrelated proteins removed iteratively until only sequences with clear homology remained.

ACKNOWLEDGEMENTS

Research on RNase P in the author's laboratory is supported by NIH grant GM52894 and Isis Pharmaceuticals.

REFERENCES

1. Tang, T.H., et al., *Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus.* Mol Microbiol, 2005. **55**(2): p. 469-81.

2. Saez-Vasquez, J., et al., *A plant snoRNP complex containing snoRNAs, fibrillarin, and nucleolin-like proteins is competent for both rRNA gene binding and pre-rRNA processing in vitro.* Mol Cell Biol, 2004. **24**(16): p. 7284-97.

3. Schattner, P., A.N. Brooks, and T.M. Lowe, *The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W686-9.

4. Lowe, T.M. and S.R. Eddy, *A computational screen for methylation guide snoRNAs in yeast.* Science, 1999. **283**(5405): p. 1168-71.

5. McCutcheon, J.P. and S.R. Eddy, *Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics.* Nucleic Acids Res, 2003. **31**(14): p. 4119-28.

6. Accardo, M.C., et al., *A computational search for box C/D snoRNA genes in the Drosophila melanogaster genome.* Bioinformatics, 2004. **20**(18): p. 3293-301.

7. Huttenhofer, A., J. Cavaille, and J.P. Bachellerie, *Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms.* Methods Mol Biol, 2004. **265**: p. 409-28.

8. Ghazal, G., et al., *Genome-wide prediction and analysis of yeast RNase III-dependent snoRNA processing signals.* Mol Cell Biol, 2005. **25**(8): p. 2981-94.

CHAPTER 6: UTILIZATION OF A NOVEL SOFTWARE PACKAGE IN SILICO FOR COMPARATIVE MICROBIAL COMMUNITY ANALYSIS OF THE LARGE INTESTINES OF A COLD-STRESSED AFFLICTED AND AN UNAFFLICTED FLORIDA MANATEE.

Devine A.A, Ellis J.C., Newsome J., Fellner V.,Grunden A.M. *Utilization of a Novel Software Package In Silico© for Comparative Microbial Community Analysis of the Large Intestines of a Cold-Stressed Afflicted and an Unafflicted Florida Manatee*. Nature Biotechnology (Manuscript in preparation)

ABSTRACT

Manatees have been found as far west as Texas and as far North as Rhode Island during the summer months, however, the vast majority are found near southern Florida year round. Florida Fish and Wildlife calculates the number of mature manatees at only 2,181 and they are listed as an endangered species. Manatees are dependent on the microbial communities in their digestive tract not only for nutrients but also buoyancy and as a source of fermentative warmth. During the winter manatees seek warm water refuges, such as natural springs and power plant discharges. The manatees however, must leave these warm environments to forage for food in the colder water, causing significant stress and potential death. Utilizing a novel technology we examined the microbial communities of a cold-stress afflicted manatee and a baseline manatee (one not suffering from cold-stress). We sampled the proximal large intestine, distal large intestine, and mid large intestine of both animals utilizing terminal restriction fragment length polymorphism (T-RFLP) in conjunction with novel algorithms and database called In Silico. With this advanced technology we were able to view significant community shifts,

at the species level, in all three locations of the manatee's large intestines, including the emergence of opportunistic pathogens in the cold-stressed animal. These new insights into Manatee health are being used by the Florida Fish and Wildlife in an effort to restore cold-stressed manatees to full health.

INTRODUCTION

Manatees inhabit warm waters and have extremely slow metabolisms. Manatees subsist primarily on seagrass, and a long transit time allows for its proper digestion by fermentation in the large intestine. [70] Fermentation by-products (heat and gas) are critical for manatee thermoregulation and buoyancy. [71] Large and diverse microbial communities exist in the manatee's digestive tract and are directly linked with the overall health of the animal. Animal intestinal microflora is affected by external and internal factors such as diet, antibiotics, pH, and temperature variation. [72, 73] One such factor for the manatee is cold-stress. [74] Cold-stress syndrome is an accumulation of physiological stresses on manatees caused initially by a drop in water temperature. The drop in water temperature forces manatees to congregate in warm water effluxes. The manatees quickly exhaust the food supply in these areas, forcing them to forage in colder waters where they cannot sustain their energy requirements and subsequently die. It is believed that during the onset of cold-stress, the microbial community of the manatee large intestine changes, resulting in the loss of fermentative microbes responsible for the generation of critical heat and gas. Cold-stress is diagnosed by examining the animal for white skin around the face, flippers and tail, loss of buoyancy, and deep grooves on the manatee's underside from the loss of fat stores. According to the Florida Fish and Wildlife Conservation Commission manatee cold-stress syndrome accounted for 10-30% of all manatee deaths between 2002 and 2003. Because they are endangered and only 2,181 mature manatees are thought to remain, it is important to understand the underlying mechanisms and consequences of cold-stress on these animals. To examine the microbial community of a manatee's intestines for the first time, historical cultivation techniques

and/or cultivation independent methodologies such as 16S rDNA clone library construction would have been laborious. An additional drawback of constructing a clone library is the cost associated with obtaining enough 16S rDNA sequence and likely underestimating the microbial population. [75] To offset the cost of performing direct sequencing and obtaining a more representative sampling, PCR based sequencing independent methods were developed to assess the microbial community profile. One of these sequencing independent methods which has been applied across many research fields is Terminal Restriction Fragment Length Polymorphism (T-RFLP) analysis. [76-84] As with more traditional approaches for studying unculturable microbial diversity, genomic DNA is isolated from the environment and amplified using a fluorescently labeled primer. The amplicons are then digested in separate reactions with a series of restriction enzymes and loaded into a capillary electrophoresis instrument. The fluorescently labeled fragment sizes are determined by capillary electrophoresis and can then be compared to the fragment database generated *in silico* to determine the microbial community. However, we quickly realized at the outset of our research that there were major limitations when using this approach. First database size was a significant limitation. Currently the most commonly cited database, the Phylogenetic Assignment Tool (PAT), only supports 2,000 bacterial species. [85] Additionally, only 27 restriction enzymes are supported by PAT preventing researchers from using optimal restriction enzymes for their research. Furthermore, no rational design utility was available to aid researchers in analyzing the best primers and selecting restriction enzymes for their research needs *in silico*. Rather, one had to rely on trial and error, which is both time consuming and expensive. The absence of tools for rational design of these critical

aspects in conjunction with small fragment databases, and limited restriction enzyme support has limited the utility and applicability of this powerful approach. To address these limitations, an advanced software suite called In Silico comprised of 11,300 sequences and supporting over 1,100 restriction enzymes and their isomers was developed. Here we describe the first application of this advanced technology as a tool to evaluate microbial population shifts between manatees affected by cold-stress and unaffected (baseline) manatees in an effort to help preserve these endangered animals.

RESULTS

The large intestine is responsible for maintaining water balance, absorbing vitamins, and for the absorption of electrolytes. [70, 71, 86] The microbial community associated with the large intestine breaks down the indigestible cellulose and fatty acids generating acetate, propionate, butyrate and other waste products. The waste products of the microflora are then utilized as nourishment by the cells lining the large intestine. This symbiotic relationship provides not only additional nourishment and vitamin supplements to the animal that would have otherwise been lost but in manatees it is also important for the generation of fermentative heat for winter survival and necessary gas production for buoyancy required to surface. [71, 87-89] Because of this important role of the microbial community lining the large intestine to the overall health of manatees, we examined the microbes populating the proximal, mid, and distal large intestine of the cold-stress and baseline animals in order to gain a better understanding of the role microbes play in a cold-stressed manatees and the microbial shift associated with this syndrome.

Microflora of the proximal large intestine

Large microbial pattern shifts were seen between a cold-stressed afflicted and a baseline manatee. The largest difference in identifiable species was observed in the proximal large intestine with more than four times the number of identifiable species, 327 in the cold cold-stressed manatee compared to 73 in the baseline manatee (Figure 1). The disparity in the number of species identified arises because of the large number of unidentifiable fragment patterns (patterns for which there are not any matches in the database) in the baseline manatee compared to the cold-stressed manatee (Figure 2). Since T-RFLP can only identify microbes that have been previously sequenced and incorporated into the database these results indicate that the vast majority of the fragments generated in the baseline manatee are from currently uncultured and unsequenced microorganisms. The majority of the species identified in baseline and cold-stressed manatee's proximal large intestine samples were Unclassified with 36 different species identified comprising 39% and 144 different species comprising 44%, respectively (Figure 3 and 6). Multiples are identical fragment patterns that match several species from different Classes. Unlike other programs that assume all of the species sharing the same exact fragment pattern are present, In Silico compiles them into the Multiple subsection because an accurate determination as to the species presence or absence in the environment can not be made with the chosen restriction enzymes. This Multiple subsection insures high precision and high quality data. For this reason the Multiple subsection comprised a large percentage in the baseline and cold-stressed manatees proximal large intestine 8% and 19%, respectively (Figure 3 and 6). Excluding the unclassified and the multiple subsets the Classes represented by the most unique

species in the proximal large intestine of the baseline manatee are the *gamma-Proteobacteria* (19%) and *Actinobacteria* (10%) (Figure 3). Three species of *Sphingobacteria,* two species of *Actinobacteridae*, but only one *Bacilli* species was detected in the baseline manatee (Figure 3). No *Clostridia* were detected in the baseline manatee. In the cold-stressed manatee the largest Class, excluding the unclassified and the multiple subsets, was *gamma-Proteobacteria* (9%) and the emergence of 16 different *Clostridia* species comprising 4.9% were observed (Figure 6). Also 12 unique species of *Bacilli* or 3.7% of the total number of species observed were detected. Interestingly, in addition to the *Clostridia* and *Bacilli, beta-* and *delta-Proteobacteria* patterns were also only detected in the cold-stressed manatee's but not in the baseline manatee's proximal large intestine.

Microflora of the mid large intestine

Again identifiable species in the cold-stressed manatee's mid large intestine were significantly higher than in the baseline manatee's mid large intestine, 320 to 94 identifiable unique species (Figure 1). As observed in the proximal sampling site, the largest percentage of identified species at the mid large intestine was Unclassified comprising 41% and 47.8%, respectively (Figure 4 and 7). The Multiple subsection also comprised a large percentage in the baseline but even more significantly in the cold-stressed manatee's mid large intestine comprising 7% and 17.8% respectively (Figure 4 and 7). Excluding the unclassified and the multiple subsets, the Classes represented by the most unique species in the mid large intestine of the baseline manatee are *gamma-Proteobacteria* (33%) and *alpha-Proteobacteria* (9%). Noticeably absent from the

baseline manatee's mid large intestine are *Clostridia sp.* and *Bacilli sp.* In the cold-stressed manatee, the largest Class excluding the unclassified and the multiple subsets were *gamma-Proteobacteria* (8.8%) and the emergence of 17 different *Clostridia* species comprising 5.3% were observed (Figure 6). Also 16 unique species of *Bacilli* or 5% of the total number of species identified were also detected. Interestingly, in addition to the *Clostridia* and *Bacilli, beta-*, and *epsilon-Proteobacteria* patterns were also only detected in the cold-stressed manatee's but not the baseline manatee's proximal large intestine. Surprisingly the only archaeal pattern detected was *Methanobrevibacter sp.* and is present in the non-cold-stressed manatee but absent in the cold-stressed manatee.

Microflora of the distal large intestine

The distal sampling sites were the only sites in this study where the baseline manatee appears to have a larger number of identifiable species than the cold-stressed sample, 104 in the baseline compared to 72 patterns in the cold-stressed manatee (Figure 1). The largest percentage of identified species in the baseline and cold-stressed distal large intestine was Unclassified comprising 46% and 50%, respectively (Figure 5 and 8). The Multiple subsection also comprised a large percentage in the baseline but even more significantly in the cold-stressed manatee's mid large intestine comprising 6% and 21% respectively (Figure 4 and 7). Excluding the unclassified and the multiple subsets, the Classes represented by the most unique species in the distal large intestine of the baseline manatee are *gamma-Proteobacteria* (19%) and *Actinobacteria* (9%). For the first time in the distal large intestine of the baseline manatee one *Clostridia sp.* and one *Bacilli sp.* was observed. In the cold-stressed manatee the largest Class excluding the unclassified

and the multiple subsets were *beta-Proteobacteria* (8%) and the emergence of 6 different *Clostridia* species or 8% of identified species in the distal large intestine were observed (Figure 8). However, this is the first instance in the cold-stressed animal, that no *Bacilli* species were detected.

DISCUSSION

Traditionally, microbial communities have been examined using cultivation techniques and later cultivation independent methodologies such as small ribosomal subunit clone library construction. Here we describe the use of a robust high throughput approach utilizing T-RFLP coupled with a software suite called In Silico, composed of a database and novel algorithms. This database and algorithms allow researchers, for the first time, to identify to the species level, entire microbial communities. We applied the software suite to study the manatee's large intestine microbial communities and their shifts in animals suffering from cold-stress syndrome. Cold-stress syndrome accounts for 10-30% of all manatees' deaths according to Florida Fish and Wildlife. Since Florida Fish and Wildlife estimates only 2,181 mature manatees remain, it is important to understand the underlying mechanisms and microbial community shifts associated with this syndrome.

While In Silico's database and advanced algorithms demonstrate profound differences in identifiable microorganisms in the large intestine of the baseline manatee and cold-stressed manatee, it also reveals some commonalities. For instance in the baseline manatee the predominant phylogenetically assigned species belong to the class *gamma-proteobacteria* comprising on average 23.7% of all identifiable species. Based on

the homogeneity of the *gamma-proteobacteria* throughout the large intestine of the baseline manatee and a significant decrease in *gamma-proteobacteria* in the cold-stressed animal, it is likely they play an important role in maintaining their overall health.

Additionally we observed the emergence of opportunistic pathogens in the *Bacilli* and *Clostridia* class. In all samples from the cold-stress manatee, we observed a dramatic increase in diversity of the *Clostridia* class comprising on average 6.1% of all identified species. Additionally, the emergence of the *Bacilli* class in the proximal and mid large intestine of the cold-stressed manatee was also observed with an average of 5% of all identified species. Given that many *Clostridia* and *Bacilli* are opportunistic pathogens and their relative absence in the baseline manatee, we hypothesize that these microbial species are likely having a deleterious effect on the manatee suffering from cold-stress and the reduction/removal of these species is likely important for full recovery. Another commonality among the baseline manatee's large intestine is a large number of unmatched fragments when compared to the cold-stressed manatee (Figure 2). Since T-RFLP can only identify microbes that have been previously sequenced and incorporated into the database these results indicate that the vast majority of fragments generated in the baseline manatee are from currently uncultured and/or unsequenced microorganisms. We hypothesize that these currently uncharacterized microbes are important symbionts of the manatee and warrant further research into their functional role in the overall health of the animal. A better understanding of the natural microbial community would aid scientists in formulating better rehabilitation procedures of manatees suffering from this syndrome.

By utilizing the In Silico software suite we were able to see significant microbial shifts between baseline and cold-stressed manatees. Furthermore, we also observed several commonalities between sampling sites in the baseline manatee as well as commonalities among sampling sites in the cold-stressed animal. It is unlikely that we would have been able to construct clone libraries large enough to see such extensive variation between the two animals at all three sampling sites (proximal, mid, and distal large intestine). For these reasons and because this approach and software suite can be utilized in any environment in which DNA can be extracted, researchers from numerous research fields will be able to utilize this powerful software suit made by In Silico LLC, (www.insilicoinc.com).

MATERIALS AND METHODS

Genomic DNA isolation- Large intestine material was collected from three points, proximal, mid and distal, along the length of two Florida manatees during necropsy. This material was stored at -80ºC until it was ready to be shipped from FWRI to NC State University. Upon receiving the samples they were stored again at -80ºC until further gDNA isolation. Genomic DNA was isolated from 250 mg samples of large intestine material using the MoBio Power Soil™ Kit (MoBio Laboratories Inc, Solana Beach CA) according to the kit protocols. 1µl of each of the gDNA samples was run on a 1% TAE agarose gel to determine the quality of the genomic DNA. The isolated gDNA was kept at -80ºC until it was used in the subsequent PCR reactions.

PCR reactions

The PCR reactions were set prepared as 100µl reactions with a master mix for the bacterial specific reactions of 10µl of 10X Taq buffer (Qiagen, Valencia CA), 0.8µl of dNTP mix (Qiagen, Valencia CA), 1.3µl of 40µM forward primer of either 8F- non labeled or 8F-Hexamide (5'- AGAGTTTGATC(A/C)TGGCTCAG- 3'), 0.5µl of 100µM reverse primer 1492R (5'- GGTTACCTTGTTACGACTT- 3'), 0.5µl of Taq Polymerase (Qiagen, Valencia CA) and 85.9µl of PCR-grade water. For the archaeal specific reactions 10µl of 10X Taq buffer (Qiagen, Valencia CA), 0.8µl of dNTP mix (Qiagen, Valencia CA), 0.5µl of 100µM forward primer either AR109F non labeled or Ar109F-Hexamide (5'- AC(G/T)GCTCAGTAACACGT - 3'), 0.5µl of 100µM reverse primer Ar912R (5'- CTCCCCCGCCAATTCCTTTA - 3'), 0.5µl of Taq Polymerase (Qiagen, Valencia CA) and 86.7µl of PCR grade water. All PCR reactions were carried out in a Bio-Rad Cycler Thermocycler (BioRad Laboratories, Hercules CA) with a thermal profile of 94°C (3mins.), (94°C (1 min.), 50°C (1 min.) and 72°C (2 min.) (25 cycles) with a final extension of 7 min at 72°C. Following the PCR reaction 1µl samples from each reaction were run on a 1% TAE agarose gel and imaged on a BioRad Gel Doc system (BioRad Laboratories, Hercules CA).

PCR cleanup and enzymatic digestion

PCR reactions products were purified using the MoBio UltraClean™ PCR Clean-up kit (MoBio Laboratories Inc, Solana Beach CA) according to kit protocols and following the clean up 1µl samples from each reaction were performed on a 1% TAE agarose gel and imaged on a BioRad Gel Doc system (BioRad Laboratories, Hercules

CA). Three simultaneous 100µl restriction digests were carried out on 15µl aliquots of each PCR product using the restriction enzymes *RsaI*, *HhaI*, and *MspI* from New England Biolabs (New England Biolabs, Ipswich MA). The digests were carried out overnight in a 37°C incubator and subsequently cleaned with the Qiagen Nucleotide Removal Kit (Qiagen, Valencia CA) according to kit protocols with only the elution step was modified, instead of EB buffer supplied with the kit, 50µl of PCR Grade water was used.

TRFLP Analysis

TRFLP analysis was conducted at the Michigan State University Genomics Core facility using an Applied Biosystems Prism 3100 Genetic Analyzer (Applied Biosystems Foster City, CA) according to their protocol. Fragment files returned from MSU were subsequently compressed, removing all unnecessary fragment information. All fragment patterns were then analyzed using a custom designed 16S fragment database.

Clone Library Construction

Pooled PCR samples from both the bacterial and euryarchaeal specific reaction from both genomic DNA samples were cleaned using MoBio's UltraClean™ PCR clean up kit (MoBio Laboratories Inc. Solana Beach, CA) according to the kit protocol. A 4µl aliquot was used in the Topo-TA cloning kit from Invitrogen (Invitrogen Life Technologies, Carlsbad, CA). Cloning and transformation reactions were carried out according to kit protocols. The transformed bacteria were plated on LB-Ampicillin plates according to kit protocols and grown overnight at 37°C. Transformed bacteria were then

transferred to 5ml liquid media supplemented with ampicillin and grown in a 37°C Barnstead Lab-Line shaker incubator overnight. 500µl aliquots were transferred to cryovials containing 500µl of 50% glycerol for frozen stock construction. These stocks were then stored in –80°C freezers until they were cultured for sequencing.

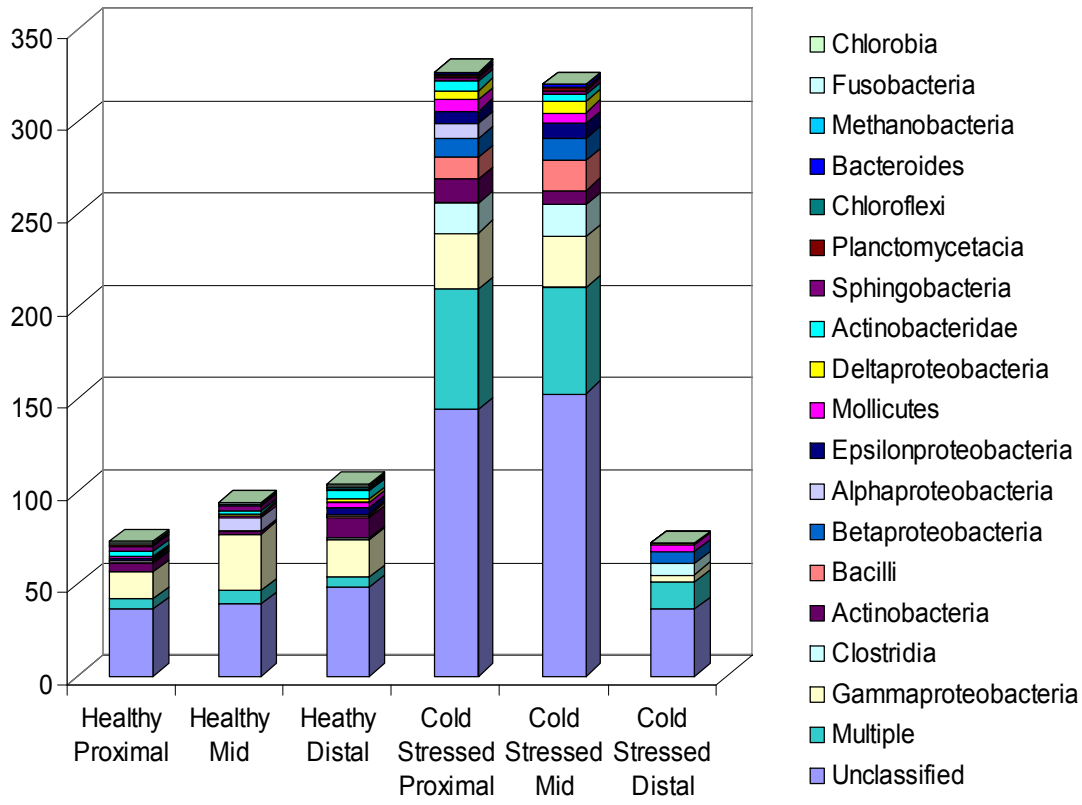


Figure 1: Microflora of baseline and cold-stressed manatees

The microflora of baseline and cold-stressed sampled for the proximal, mid, and distal large intestine of each animal. Microbial communities are phylogeneticly sorted at the class level revealing significant microbial population shifts in the cold-stressed manatee including the emergence of opportunistic pathogens from the *Bacilli* and *Clostridium* classes.

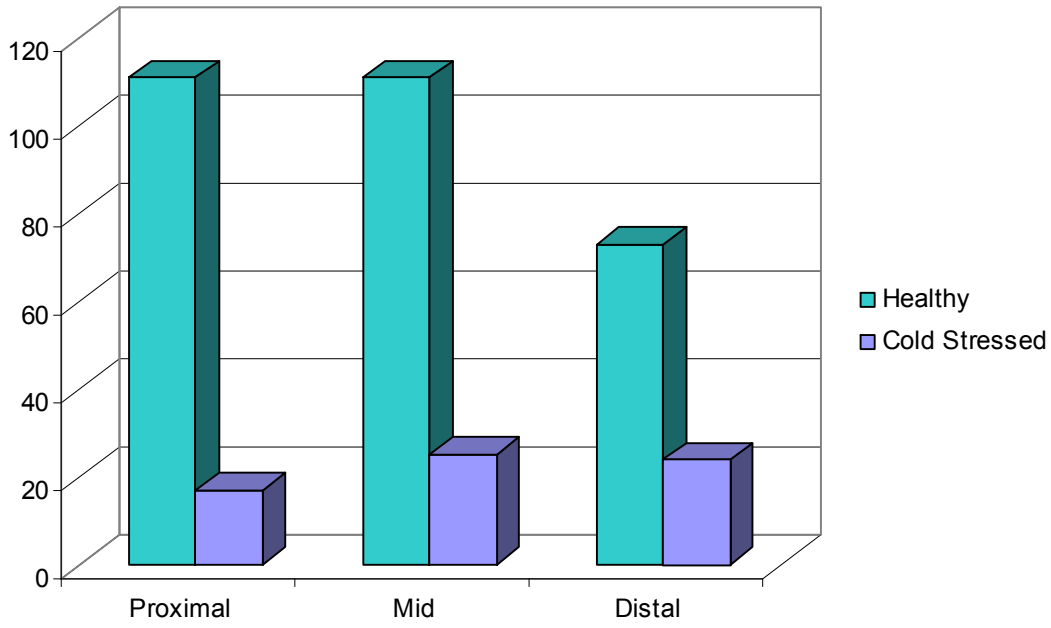
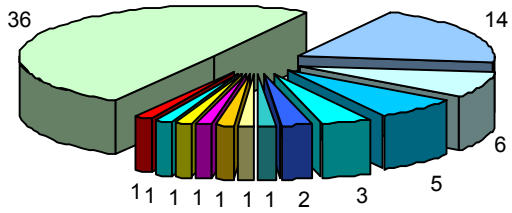


Figure 2: Comparison of unmatched fragment patterns between the baseline manatee and the cold-stressed manatee.

Unmatched Fragment Patterns from the baseline manatee and the cold-stressed manatee. All three sampling sites reveal a significant difference in the number of unidentified fragment patterns. Since T-RFLP can only identify microbes that have been previously sequenced and incorporated into the database these results indicate that the vast majority of the fragments generated in the baseline manatee are from currently uncultured and unsequenced microorganisms.

A)



B)

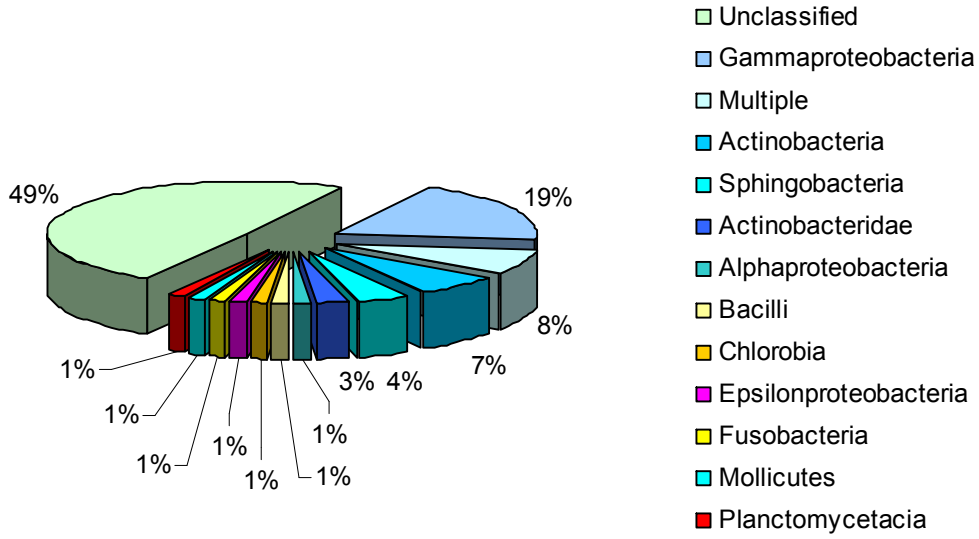
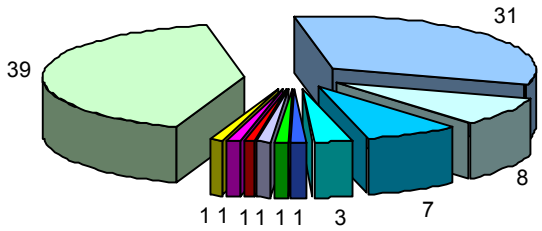


Figure 3: Microflora of the proximal intestine of the baseline manatee

Graphical representation of the baseline manatee's proximal intestine microflora sorted by the number of unique species per class (a) and the percentage of unique species identified (b).

A)



B)

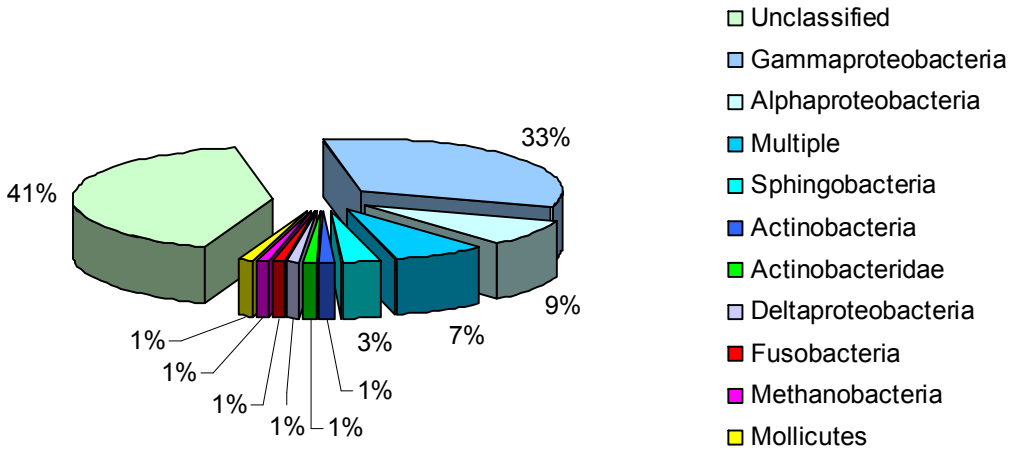


Figure 4: **Microflora of the mid intestine of the baseline manatee**

Graphical representation of the baseline manatee's mid intestine microflora sorted by the number of unique species per class (a) and the percentage of unique species identified (b).

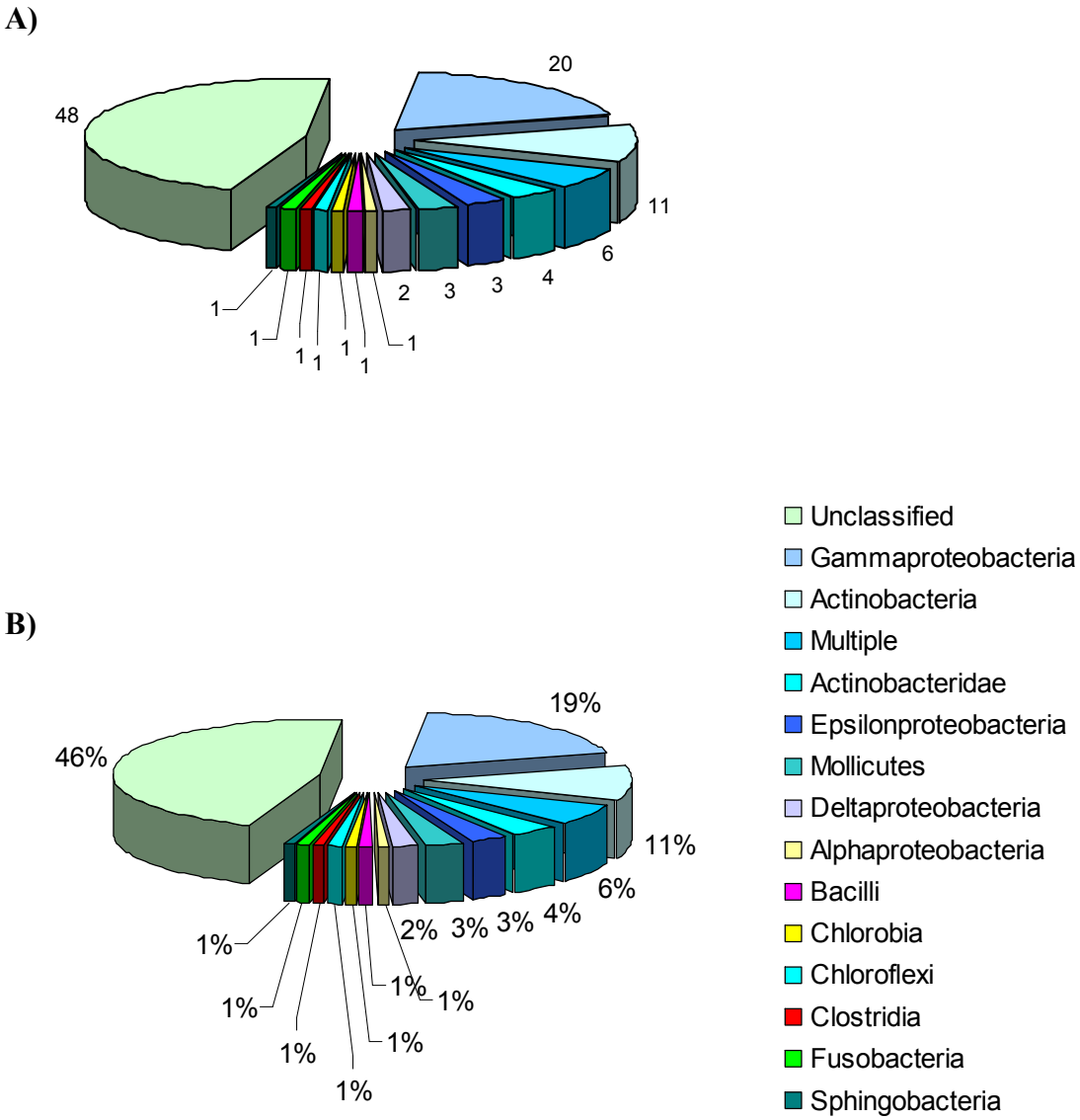


Figure 5: **Microflora of the disal intestine of the baseline manatee**

Graphical representation of the baseline manatee's distal intestine microflora sorted by the number of unique species per class (a) and the percentage of unique species identified (b).

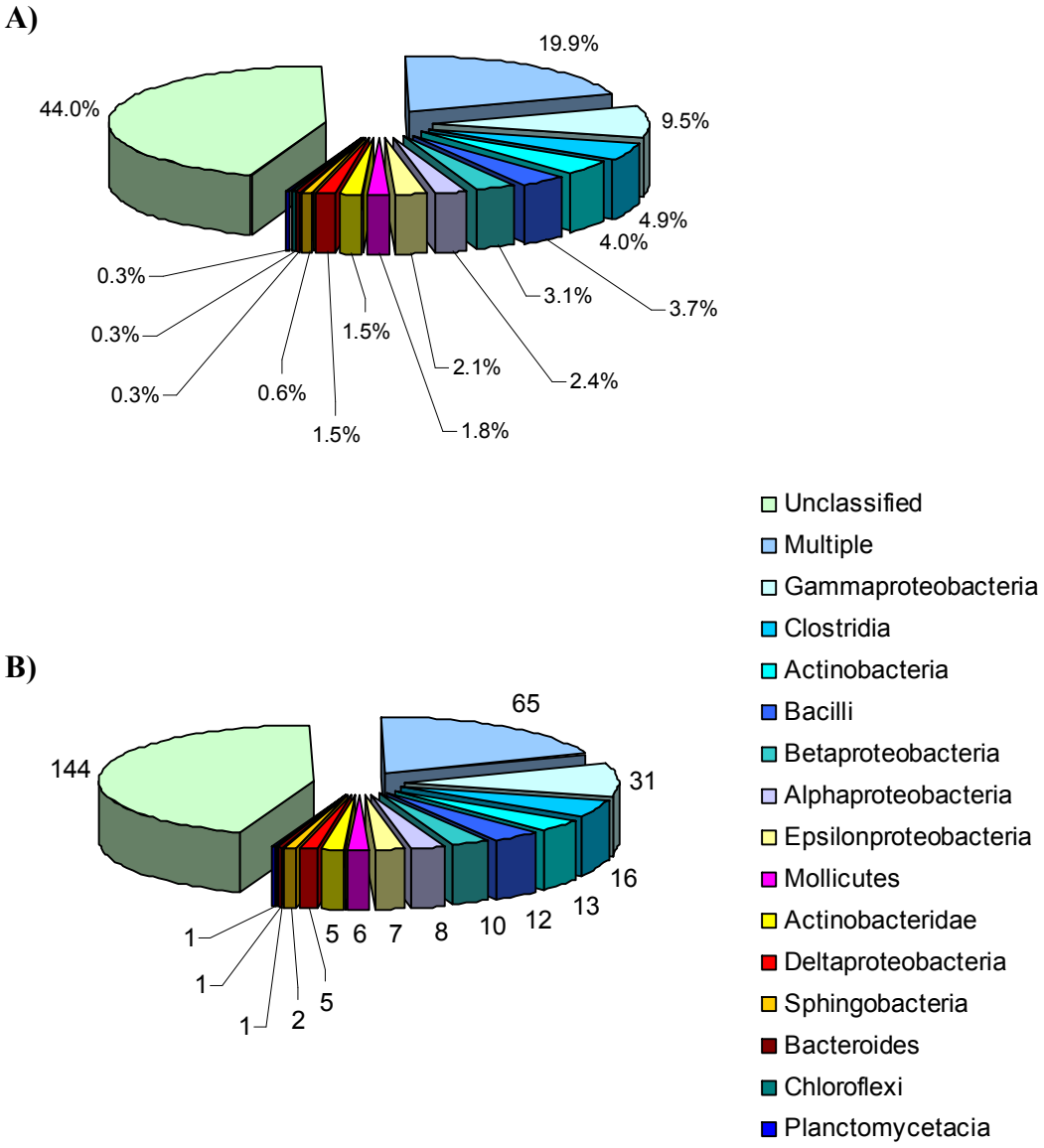
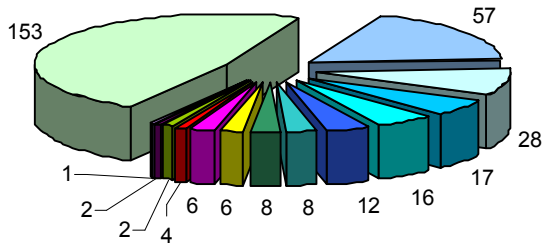


Figure 6: **Microflora of the proximal intestine of the cold-stressed manatee**

Graphical representation of the cold-stressed manatee's proximal intestine microflora sorted by the number of unique species per class (a) and the percentage of unique species identified (b).

A)



B)

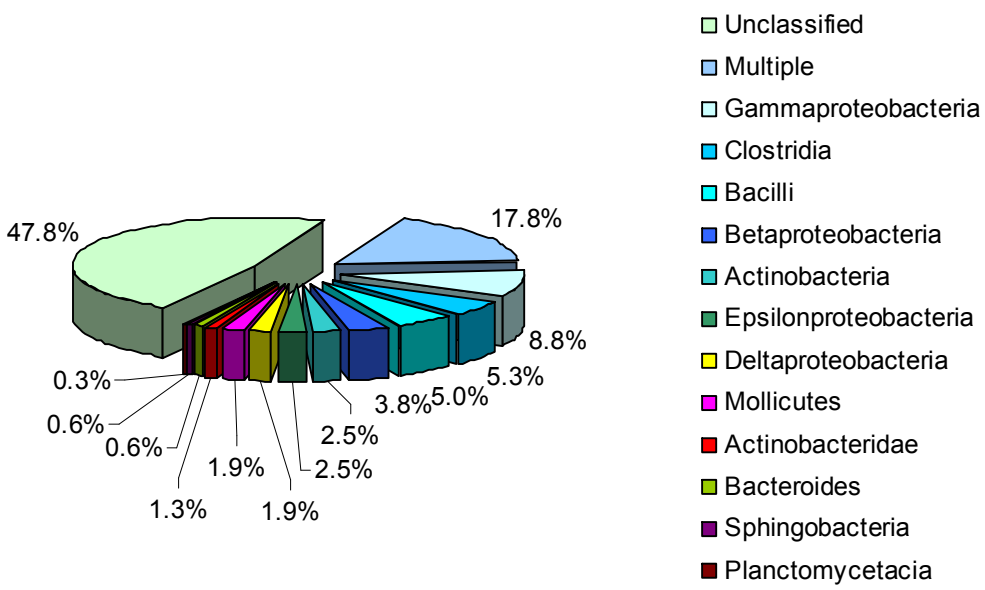
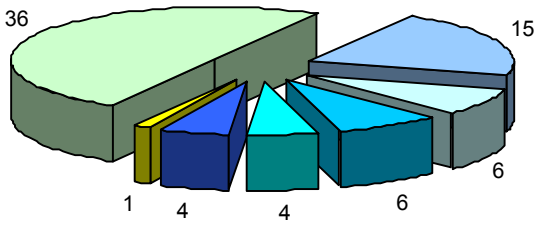


Figure 7: Microflora of the mid intestine of the cold-stressed manatee

Graphical representation of the cold-stressed manatee's mid intestine microflora sorted by the number of unique species per class (a) and the percentage of unique species identified (b).

A)



B)

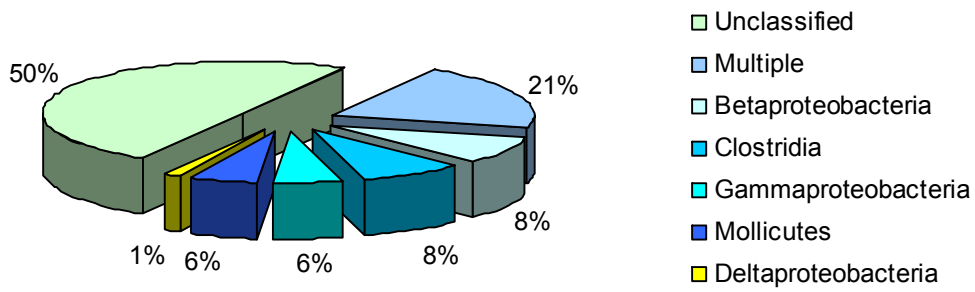


Figure 8: **Microflora of the distal intestine of the cold-stressed manatee**

Graphical representation of the cold-stressed manatee's distal intestine microflora sorted by the number of unique species per class (a) and the percentage of unique species identified (b).

REFERNECES

1. Burn, D.M., *The digestive strategy and efficiency of the West Indian manatee, Trichechus manatus.* Comp Biochem Physiol A, 1986. 85(1): p. 139-42.
2. Rommel, S. and J.E. Reynolds, 3rd, *Diaphragm structure and function in the Florida manatee (Trichechus manatus latirostris).* Anat Rec, 2000. 259(1): p. 41-51.
3. Edwards, J.E., et al., *Influence of flavomycin on ruminal fermentation and microbial populations in sheep.* Microbiology, 2005. 151(Pt 3): p. 717-25.
4. Tajima, K., et al., *Diet-dependent shifts in the bacterial population of the rumen revealed with real-time PCR.* Appl Environ Microbiol, 2001. 67(6): p. 2766-74.
5. Walsh, C.J., C.A. Luer, and D.R. Noyes, *Effects of environmental stressors on lymphocyte proliferation in Florida manatees, Trichechus manatus latirostris.* Vet Immunol Immunopathol, 2005. 103(3-4): p. 247-56.
6. Krause, D.O. and J.B. Russell, *How many ruminal bacteria are there?* J Dairy Sci, 1996. 79(8): p. 1467-75.
7. Sakata, S., et al., *Culture-independent analysis of fecal microbiota in infants, with special reference to Bifidobacterium species.* FEMS Microbiol Lett, 2005. 243(2): p. 417-23.
8. Wang, M., et al., *T-RFLP combined with principal component analysis and 16S rRNA gene sequencing: an effective strategy for comparison of fecal microbiota in infants of different ages.* J Microbiol Methods, 2004. 59(1): p. 53-69.
9. Sakamoto, M., et al., *Changes in oral microbial profiles after periodontal treatment as determined by molecular analysis of 16S rRNA genes.* J Med Microbiol, 2004. 53(Pt 6): p. 563-71.
10. Sakamoto, M., M. Umeda, and Y. Benno, *Molecular analysis of human oral microbiota.* J Periodontal Res, 2005. 40(3): p. 277-85.
11. Jernberg, C., et al., *Monitoring of antibiotic-induced alterations in the human intestinal microflora and detection of probiotic strains by use of terminal restriction fragment length polymorphism.* Appl Environ Microbiol, 2005. 71(1): p. 501-6.
12. Rogers, G.B., et al., *Bacterial activity in cystic fibrosis lung infections.* Respir Res, 2005. 6(1): p. 49.
13. Mengoni, A., et al., *Comparison of 16S rRNA and 16S rDNA T-RFLP approaches to study bacterial communities in soil microcosms treated with chromate as perturbing agent.* Microb Ecol, 2005. 50(3): p. 375-84.
14. Chin, J., *Intestinal microflora: negotiating health outcomes with the warring community within us.* Asia Pac J Clin Nutr, 2004. 13(Suppl): p. S24-5.
15. Gong, J., et al., *Diversity and phylogenetic analysis of bacteria in the mucosa of chicken ceca and comparison with bacteria in the cecal lumen.* FEMS Microbiol Lett, 2002. 208(1): p. 1-7.
16. Kent, A.D., et al., *Web-based phylogenetic assignment tool for analysis of terminal restriction fragment length polymorphism profiles of microbial communities.* Appl Environ Microbiol, 2003. 69(11): p. 6768-76.

17. Reynolds, J.E., 3rd and S.A. Rommel, *Structure and function of the gastrointestinal tract of the Florida manatee, Trichechus manatus latirostris.* Anat Rec, 1996. 245(3): p. 539-58.
18. Gallivan, G.J., J.W. Kanwisher, and R.C. Best, *Heart rates and gas exchange in the Amazonian manatee (Trichechus inunguis) in relation to diving.* J Comp Physiol [B], 1986. 156(3): p. 415-23.
19. Bergey, M. and H. Baier, *Lung mechanical properties in the West Indian Manatee (Trichechus manatus).* Respir Physiol, 1987. 68(1): p. 63-75.
20. Flewelling, L.J., et al., *Brevetoxicosis: red tides and marine mammal mortalities.* Nature, 2005. 435(7043): p. 755-6.

APPENDICES

APPENDIX I: THE ASSOCIATED PROTEINS OF THE RNase P RNA in *Methanocaldococcus jannaschii*

INTRODUCTION

Ribonuclease P (RNase P) is an ancient ribonuclease responsible for the processing of precursor tRNA (pre-tRNA) by removing the 5' leader sequence but also participates in the maturation of other RNAs such as 2S, 4.5S, tmRNA (10S), some polycistronic mRNA, and some viral RNAs. [1-6] Unlike most enzymes, RNase P is composed of both RNA and protein. The RNase P RNA is found in all three domains of life (Bacteria, Archaea, and Eukaryotes) including mitochondria and plastids. Furthermore, it is the RNA, not the protein(s) that catalyze the site specific cleavage reaction and is therefore, by definition, a ribozyme.

As in Bacteria, most archaeal RNase P RNAs are Type A RNAs, resembling bacterial RNase P RNAs in both sequence and structure. However, despite the structural similarity between bacterial RNase P RNA and the archaeal RNA it was widely thought previously that, like eukaryotic RNase P RNA, the archaeal RNase P RNA was absolutely dependent on their associated proteins for catalytic activity. Pannucci et al described for the first time that some archaeal RNase P RNA are catalytically active in the absence of protein, like bacterial RNase P RNAs, but that they required extreme ionic conditions (4 M ammonium acetate, 300 mM MgCl2, 50 mM Tris-Cl). [7] As in bacteria, the protein subunits in the archaeal enzymes seem not to contribute directly to catalysis, but at least predominantly toward stabilization of the superstructure of the RNA subunit.

One of the subsets of archaeal RNase P RNAs which do not display catalytic activity even under high ionic conditions are Type M RNase P RNAs. To date only five archaeal species (*Archaeoglobus fulgidus*, *Methanocaldococcus jannaschii*, *Methanococcus marapaludis*, *Methanothermococcus thermolithotrophicus*, and *Methanococcus vannielii)* have been described with a Type M RNA. [8] Type M RNAs are essentially similar to Type A RNAs but lack two essential substrate recognition elements. One such element found in all other bacterial and archaeal RNase P RNAs is P8. P8 is located in the region of the RNA that forms a highly conserved cruciform consisting of P7, P8, P9, and P10. [9, 10] (Figure 4) P8 recognizes substrate at the T loop of the pre-tRNA. Additionally, in Bacteria P8 stabilizes, via tertiary interactions, P18, P4, and P14. Another structural element, L15, is also missing in Type M RNAs. L15 is distal of P15, and like P8 is also essential for substrate recognition. L15 recognizes and binds the 3'-NCCA tail of the pre-tRNA which is necessary for efficient substrate cleavage. [11, 12] Type M RNase P RNAs, then, have specifically lost all of the regions known to be directly involved in substrate recognition outside of the active site.

The absence of essential RNA elements involved in substrate recognition and the absence of novel structural elements that could replace the functional roles P8 and L15 raises the question, "How does RNase P compensate for the loss of these critical secondary structures in its RNA?" The two most probable scenarios are: 1) One or more of the four known protein homologues associated with other archaeal RNase P RNAs are able to specifically recognize the tRNA substrate, or 2) an additional protein or proteins has/have taken over the functional role of substrate recognition typically associated with

P8 and L15. Here we examine which proteins are associated with the RNase P RNA in *Methanocaldococcus jannaschii*.

RESULTS AND DISCUSSION

Purification of RNase P activity from Methanocaldococcus jannaschii

5g of a *Methanocaldococcus jannaschii* frozen cell pellet was ground in liquid nitrogen with a mortar and pestle. The pellet was resuspended in TMGN-100 (50mM Tris, 10mM MgCl2, and 100mM NH4Cl) and passed through a french press 3 times to ensure complete cell lyses. The solution was cleared by centrifugation at 16,000 X g for 30 min at 4°C. The supernatant was dialyzed overnight in two exchanges of TMGN100 at 4°C. The dialyzed supernatant was passed over a DEAE Sepharose column and fractions collected. Active fractions from the DEAE column were pooled and a Cs_2SO_4 Buoyant Density Gradient Centrifugation was performed. Active fractions were pooled and another Cs_2SO_4 Buoyant Density Gradient Centrifugation performed. Active fractions from the second Cs_2SO_4 purification were pooled and a glycerol gradient centrifugation performed. (Figure 1) Active fractions were loaded onto an SDS-PAGE gel. (Figure 2) Bands from the most active fraction were excised from the SDS-PAGE gel and MALDI-TOF performed to identify proteins that co-purify with RNase P activity. (Figure 3) Most identified proteins were cloned, expressed, and purified as detailed below. Additionally, Western blots and Immunoprecipitation were performed to identify proteins associated with the RNase P holoenzyme.

MJ0212p

MJ0212 was cloned into pBad His A and pET16b expression vectors. Research on MJ0212 homologue, Mth1483p in *Methanothermobacter thermoautotrophicus*, revealed that it was not associated with the RNase P holoenzyme and for this reason no further expression or purifications steps were taken.

MJ0332.1p

MJ0332.1 was cloned into pET16b expression vector. MJ0332.1 has a codon bias that requires the addition of the codon plus vector which encodes for rare tRNAs in *E. coli*. MJ0332.1 was stable in the absence of the codon plus vector however upon addition of this vector the cells lysed and protein expression was not possible. Because of the inability to express this protein synthetic peptides were made and antisera generated. Western blots and immunoprecipitation using the antisera generated from the peptides provided no evidence MJ0332.1p is associated with the RNase P holoenzyme.

MJ0464p

MJ0464 was cloned into pET16b expression vector. Protein expression was optimized in BL21 *E. coli* cells and recombinant protein purified using a Ni^{2+} column. Polyclonal rabbit anti-sera was produced to MJ0464p using the purified recombinant protein. Anti-MJ0464p serum demonstrated the ability to immunoprecipitate RNase P activity compared to pre-immune serum indicating MJ0464p was a component of the RNase P holoenzyme. Western blot analysis proved inconclusive due to low concentrations of MJ0464p protein from the purified whole cell extract.

MJ0494p

MJ0494 was cloned into pET16b expression vector. Protein expression was optimized in BL21 *E. coli* cells and recombinant protein purified using a Ni^{2+} column. Polyclonal rabbit anti-sera was produced to MJ0494p using the purified recombinant protein. Anti-MJ0494p serum demonstrated the ability to immunoprecipitate RNase P activity compared to pre-immune serum indicating MJ0494p was a component of the RNase P holoenzyme. Western blot analysis proved inconclusive due to low concentrations of MJ0494 protein from the purified whole cell extract.

MJ0962p

MJ0962 was cloned into pET16b expression vector. Protein expression was optimizes in BL21 *E. coli* cells and recombinant protein purified using a Ni^{2+} column. Polyclonal rabbit anti-sera was produced to MJ0962p using the purified recombinant protein. Anti-MJ0962p serum demonstrated the ability to immunoprecipitate RNase P activity compared to pre-immune serum indicating MJ0962p was a component of the RNase P holoenzyme. Western blot analysis demonstrated that MJ0962p co-purified with RNase P activity confirming the immunoprecipitation results that MJ0962p is associated with the RNase P holoenzyme.

MJ1128p

MJ1128 was cloned into pBad His A expression vector. Protein expression was optimizes in TOP10 *E. coli* cells and recombinant protein purified using a Ni^{2+} column. Polyclonal rabbit anti-sera was produced to MJ1128p using the purified recombinant

protein. Anti-MJ1128p serum did immunoprecipitate RNase P activity and western blot analysis demonstrated MJ1260p does not co-purify with RNase P activity. These results confirm MJ1260p is not associated with the RNase P holoenzyme.

MJ1139p

MJ1139 was cloned into pET16b expression vector. Protein expression was optimizes in BL21 *E. coli* cells and recombinant protein purified using a Ni^{2+} column. Polyclonal rabbit anti-sera was produced to MJ1139p using the purified recombinant protein. Anti-MJ1139p serum demonstrated the ability to immunoprecipitate RNase P activity compared to pre-immune serum indicating MJ1139p was a component of the RNase P holoenzyme. Western blot analysis proved inconclusive due to low concentrations of MJ1139 protein from the purified whole cell extract.

MJ1260p

MJ1260 was cloned into pBad His A expression vector. Protein expression was optimizes in TOP10 *E. coli* cells and recombinant protein purified using a Ni^{2+} column. Polyclonal rabbit anti-sera was produced to MJ1128p using the purified recombinant protein. Anti-MJ1260p serum did immunoprecipitate RNase P activity and western blot analysis demonstrated MJ1260p does not co-purify with RNase P activity. These results confirm MJ1260p is not associated with the RNase P holoenzyme.

MJ1625p

Despite trying numerous prokaryotic and eukaryotic expression vectors MJ1625 could not be cloned into any expression or non-expression vector. Because of the inability to clone MJ1625, synthetic peptides were made with sequence similarity to the carboxy terminus and amino terminuses and antisera generated to these peptides. Western blots and immunoprecipitation using the antisera generated from the peptides provided no evidence MJ1625p is associated with the RNase P holoenzyme.

REFERENCES

1. Bothwell, A.L., R.L. Garber, and S. Altman, *Nucleotide sequence and in vitro processing of a precursor molecule to Escherichia coli 4.5 S RNA.* J Biol Chem, 1976. **251**(23): p. 7709-16.
2. Hori, Y., T. Tanaka, and Y. Kikuchi, *In vitro cleavage of Drosophila 2S rRNA by M1 RNA.* Nucleic Acids Symp Ser, 2000(44): p. 93-4.
3. Ushida, C., D. Izawa, and A. Muto, *RNase P RNA of Mycoplasma capricolum.* Mol Biol Rep, 1995. **22**(2-3): p. 125-9.
4. Bothwell, A.L., B.C. Stark, and S. Altman, *Ribonuclease P substrate specificity: cleavage of a bacteriophage phi80-induced RNA.* Proc Natl Acad Sci U S A, 1976. **73**(6): p. 1912-6.
5. Yang, Y.H., et al., *Engineered external guide sequences are highly effective in inducing RNase P for inhibition of gene expression and replication of human cytomegalovirus.* Nucleic Acids Res, 2006. **34**(2): p. 575-83.
6. Gimple, O. and A. Schon, *In vitro and in vivo processing of cyanelle tmRNA by RNase P.* Biol Chem, 2001. **382**(10): p. 1421-9.
7. Pannucci, J.A., et al., *RNase P RNAs from some Archaea are catalytically active.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 7803-8.
8. Brown, J.W., *The Ribonuclease P Database.* Nucleic Acids Res, 1999. **27**(1): p. 314.
9. Harris, J.K., et al., *New insight into RNase P RNA structure from comparative analysis of the archaeal RNA.* Rna, 2001. **7**(2): p. 220-32.
10. Brown, J.W. and E.S. Haas, *Ribonuclease P structure and function in Archaea.* Mol Biol Rep, 1995. **22**(2-3): p. 131-4.
12. Svard, S.G., U. Kagardt, and L.A. Kirsebom, *Phylogenetic comparative mutational analysis of the base-pairing between RNase P RNA and its substrate.* Rna, 1996. **2**(5): p. 463-72.
13. Svard, S.G. and L.A. Kirsebom, *Several regions of a tRNA precursor determine the Escherichia coli RNase P cleavage site.* J Mol Biol, 1992. **227**(4): p. 1019-31

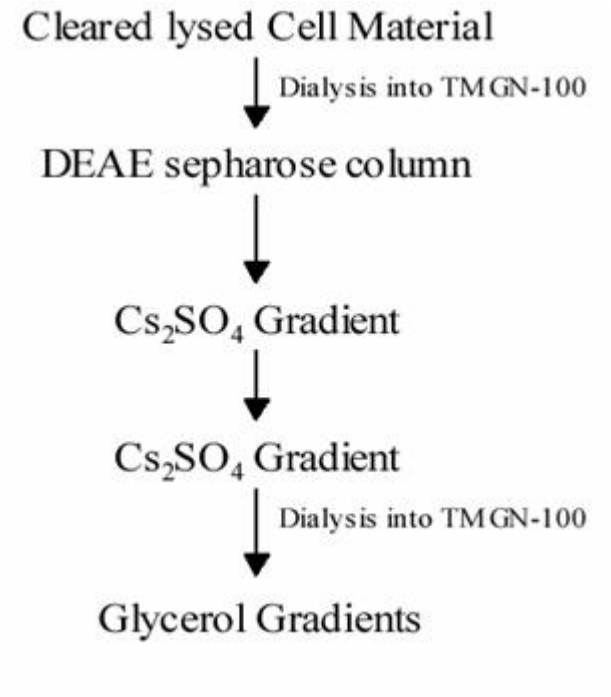


Figure 1: Purification Diagram of RNase P holoenzyme

Diagram outlining the purification steps taken to purify RNase P from the whole cell extracts of *Methanocaldococcus jannaschii.*

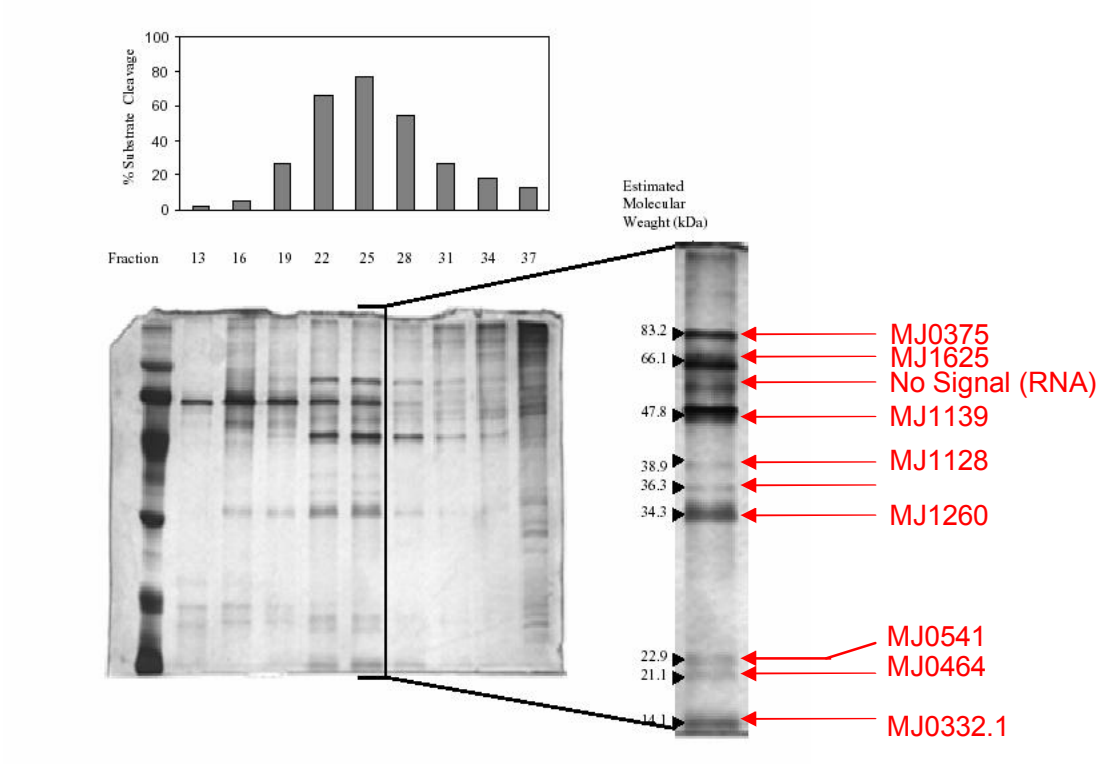


Figure 2: SDS-PAGE gel and corresponding activity

Silver stained SDS-PAGE gel of the glycerol gradient fractions and corresponding activity assay for each fraction. Every third faction was loaded starting with fraction 13 and continuing to fraction 37. The fraction exhibiting the greatest activity has been enlarged to the right with the candidate proteins identified via MALDI-TOF in red.

Figure 2 Reference:

Picture courtesy of Ginger Muse (Unpublished)

Protein Band	Identified Protein	Annotated Protein Function	Number of Peaks Matched	Number of Peaks	% Sequence Covered	Estimated Exp. Mol Weight	Theoretical Molecular Weight (kDa)
2	MJ1625	conserved hypothetical protein	2	5	2	94 - 67	78.60
3	MJ0376	conserved hypothetical protein	4	30	6	94 - 67	86.22
3A	----------	----------	----------	----------	----------	50 - 43	----------
4	MJ1139	RNase P subunit	6	19	31	30	27.33
8	MJ1128	hypothetical protein	2	2	10	14-30	34.4
9	MJ0673	SSU ribosomal protein S8E	4	7	22	20-30	14.52
10	MJ1260	SSU ribosomal protein S6E	2	12	20	20 - 14.4	14.35
11	----------	----------	----------	----------	----------	----------	----------
12	MJ464	RNase P subunit	3	19	32	20 - 14.4	10.87
	MJ0541	nicotinamide-nucleotide adenylyltransferase	4	19	21		19.59
13	MJ464	RNase P subunit	3	14	32	20 - 0	10.87
14	MJ0332.1	hypothetical protein	2	12	20	20 - 0	14.95

Figure 3: MADLI-TOF summary of identified proteins

Summary of MALDI-TOF results from excised bands from figure 2.

Figure 3 Reference:

Picture courtesy of Ginger Muse (Unpublished)

APPENDIX II: PUBLICATIONS

Ellis J.C., Brown J.W., *Genes within genes within bacteria*. Trends Biochem Sci. 2003 **28**(10): 521-3

Devine A.A, Ellis J.C., Newsome J., Fellner V.,Grunden A.M. *Utilization of a Novel Software Package In Silico© for Comparative Microbial Community Analysis of the Large Intestines of a Cold-Stressed Afflicted and an Unafflicted Florida Manatee*. Nature Biotechnology (Manuscript in preparation)

Brown, J.W. and Ellis, J.C., *Comparative analysis of RNA secondary structure: The 6S RNA*. Handbook of RNA Biochemistry {Wiley-VCH}, A. Bindereif, R. Hartmann, A. Schön, and E. Westhof, eds.

Ellis J.C., Devine A.A., Brown J.W., Grunden A.M*, In Silico: A New Software Based Utility for Primer Design and Analysis.* RNA *(Manuscript in preparation)*

Devine A.A, Ellis J.C., Bull L.S., Grunden A.M., *Determining the Microbial Community of Hog Waste Lagoon Systems Utilizing Terminal Restriction Fragment Length Polymorphism Analysis*. Water Research (Manuscript in preparation)

Ellis J.C., Brown J.W.*, The snoRNP Database*. RNA (Manuscript in preparation)

Ellis J.C., Barnes J., Brown J.W., *Is Alba an RNase P subunit?*, Nucleic Acids Research (Manuscript in preparation)