Structure and evolution of ribonuclease P RNA

JW Brown, NR Pace*

Department of Biology, Indiana University, Bloomington IN 47405, USA

(Received 6 November 1990; accepted 12 March 1991)

Summary — Eubacterial RNase P contains a catalytic RNA that cleaves 5' leader sequences from precursor tRNAs. We review the current understanding of RNase P RNA structure and evolution, from the perspective of phylogenetic comparative analysis.

RNase P / phylogenetic analysis / RNA enzyme / RNA structure

Introduction

Ribonuclease P (RNase P) cleaves leader sequences from precursors of tRNA to generate the mature 5' end of the tRNA. In the eubacteria Bacillus subtilis and Escherichia coli, RNase P is composed of a small protein (119 amino acids) and a large RNA (~ 400 nt) [1]. At high salt concentrations in vitro, the RNA alone is an efficient and accurate catalyst; it is the only known RNA that acts as an enzyme in vivo, in the sense that each molecule of RNase P RNA acts on many substrate molecules. The protein component is thought to function only as a local ionic screen to allow the highly negatively charged RNA enzyme and substrate to come into close proximity under physiological conditions [2].

An accurate model for the structure of the eubacterial RNase P RNA is a prerequisite to fruitful studies of the mechanisms of substrate recognition and catalytic activity of this ribozyme. Although the primary structures (nucleotide sequences) of RNAs or their genes are now readily determined, the 3-dimensional foldings of the RNAs that are required for their functions are less accessible. The inference of secondary structure is, therefore, a crucial step in the study of a functional RNA, as subsequent investigations of the RNA are guided by the structure model. The most successful approach to inferring secondary structure in RNAs has been the use of phylogenetic comparative analysis [3]. Comparative analyses have been instrumental in the elucidation of the structures of ribosomal RNAs [4–6], transfer RNAs [7], group I [8–10] introns, small nuclear RNAs [11], and signal recognition particle RNAs [12, 13]. Sequences of genes encoding RNase P RNA from the proteobacteria [14] (‘purple bacteria and relatives’ sensu Woese [15]) and the low-G + C Gram-positives (Bacillus spp) have been used to construct a model for RNase P RNA secondary structure using phylogenetic comparative analysis [16, 17]. These sequences have also been used to test phylogenetic relationships based on comparison of 16S rRNA sequences from within these two eubacterial groups.

Phylogenetic comparative analysis

The essence of the phylogenetic comparative approach for the determination of RNA secondary structure is that potential helices in an RNA molecule (based on the occurrence of complementary sequences) are tested by inspection of the equivalent pairings in homologous RNAs in which the sequences vary. Truly homologous sequences are those with common ancestry and function, and are expected to have similar higher order structure. The occurrence in evolution of compensating base changes (co-variations) that maintain complementarity is evidence for the presence of a helix. If complementarity is not present in homologous sequences from different organisms, the structure is unlikely to exist. Confirmation of sequence pairings by phylogenetic co-variation is analogous to a genetic analysis of mutations and second-site suppressors that maintain complementarity.
RNAse P RNA sequences

The sequences of the RNase P RNA genes from four low-G + C Gram-positive species (Bacillus subtilis [18], Bacillus megaterium [16], Bacillus stearothermophilus [16], and Bacillus brevis [16]), two α-proteobacteria (Agrobacterium tumefaciens [17] and Rhodospirillum rubrum [17]), two β-proteobacteria (Alcaligenes eutrophus [17] and Thiobacillus ferrooxidans [19]), seven γ-proteobacteria (Escherichia coli [20], Salmonella typhimurium [21], Erwinia agglomerans [22], Klebsiella pneumoniae [22], Serratia marcescens [22], Pseudomonas fluorescens [16], and Chromatium vinosum [17]) and one δ-proteobacterium (Desulfovibrio desulfuricans [17]) have been determined (fig 1).

That these genes encode functional RNase P RNA has been confirmed in most cases, except for K pneumoniae, E agglomerans, S marcescens and T ferrooxidans, these RNAs have been tested in vitro and shown to have RNase P enzymatic activity. The genes from K pneumoniae, E agglomerans, and S marcescens were originally identified by their ability to complement a temperature-sensitive mutation in the gene encoding the protein component of RNase P, an unusual complementation assay in which overproduction of the RNA component of RNase P (in a multicopy plasmid) can apparently overcome the defect in the protein component.

Refinement of secondary structure

The current model for the secondary structure of RNase P RNA was originally based on the sequences from six γ-proteobacteria (all close relatives of E coli) and four Gram-positive species of the genus Bacillus [16]. In keeping with the criteria used for the phylogenetic analysis of ribosomal RNA structure [4], co-variation of two base pairings within a potential continuous helix was considered 'proof' of the existence of that helix. The limit of resolution of such a structure model is, therefore, the helix. This resolution of the RNase P RNA model has recently been refined to nearly the base pair level (figs 2 and 3). Evidence in the form of evolutionary co-variation now exists for nearly every base pair in the structure model. Further strengthening of the model comes from the fact that co-variations exist for most of the base pairings among just the proteobacterial sequences, thus eliminating questionable assignments of homologies between the disparate sequences of the proteobacterial and Gram-positive RNase P RNAs.

Two base pairings that were formerly invoked have been removed in the refinement of the structure model. Nucleotides 19/335 (nucleotide numberings used throughout this article are based on the E coli sequence) potentially form a G-U base pair at the end of helix 12–18/336–342; the pairing is no longer included because the nucleotides are invariant and form only a non-canonical pairing. Nucleotides 92/106 potentially form a conserved C-G base pair except in the case of A eutrophus, in which both nucleotides are U. Because U-U pairs in tRNAs are known not to disrupt the structure of a helix, it remains a possibility that nucleotides 92 and 106 interact (by base pairing except in the case of A eutrophus) and stack onto the end of helix 93–96/102–105. This would be especially likely if this helix forms a co-axial stack with either helix 87–91/238–242 or helix 108–110/115–117, thereby internalizing the U-U interaction in the helix. The U at position 91 in A eutrophus is particularly interesting because otherwise invariant C at this position has been suggested to be critical for both catalysis and binding of the protein subunit of the enzyme [23]. In vitro transcripts from the A eutrophus RNase P RNA gene clone, as well as RNase P RNA isolated from cultures, is catalytically active in the RNA-alone, RNase P enzyme assay [17].

Two non-canonical base pairings (11/363 and 305/326) have been added to the structure model. Nucleotides 11 and 363 co-vary as non-canonical A-C or G-A base pairs; this unusual pairing extends helix 1–10/364/373 by a single base pair. Nucleotides 305 and 326 co-vary normally as canonical base pairs in Bacillus, but in the proteobacteria exist as non-canonical A-C or G-A base pairs. The presence of conserved non-canonical pairings most likely represents the requirement of unusual local structure in the helix.

The most significant change in the structure model is in the region of nucleotide 74. In the previous model, nucleotide 74 was paired to nucleotide 247; there is apparent co-variation of these nucleotides between the Gram-positive and the γ-proteobacterial sequences. The alignment of the sequences from the two groups, however, is particularly difficult in this region. The RNase P RNA sequences from A eutrophus and Haloferax volcanii [24], on the other hand, clearly demonstrate co-variation of nucleotide 74 with nucleotide 353, rather than 247. In the refined model, therefore, helix 70–74/353–357 is extended by one base pair at the expense of helix 75–78/243–246. It is possible, by bulging nucleotide U69, to extend the
Fig 2. Secondary structure of RNase P RNA. The structures of the *E. coli* and *B. subtilis* RNase P RNAs are shown according to the refined secondary structure model. Nucleotides are numbered from the 5' end. Canonical pairings are indicated by lines, G-U and A-G/A-C non-canonical pairings are indicated by filled circles (•). Helix 70–74/353–357, which along with helix 12–18/336–342 comprises a pseudoknot, is indicated by brackets and a connecting line.

base pairing of helix 70–74/353–357 by 3 additional base pairs (66–68/358–360). These potential base pairs, however, are composed entirely of invariant nucleotides. Because no co-variations exist to confirm or refute the pairings, and since they would not constitute the direct continuation of a phylogenetically confirmed helix, these potential pairings are not included in the structure model.

**Variation in RNase P RNA structure**

Most of the structural variation in proteobacterial RNase P RNAs are in four helical regions: the distal portion of stem 20–38/43–61, the medial portion stem 142–156/161–176, the medial portion of stem 260–275/279–290, and the whole of stem 344–345. The helix located at position 344–345 in the *Bacillus* sequences (positions 330–367 based on the *B. subtilis* sequence; absent in *E. coli*) has now also been found to be a common (and ancestral) feature of RNase P RNAs of proteobacteria. More significant structural differences exist between the RNase P RNAs of the Gram-positive eubacteria and the proteobacteria. During the evolution of the *Bacillus* RNase P RNAs, helix 260–275/279–290 and nucleotides 183–225, which include helices 184–189/197–202 and 204–211/216–224, were deleted (the sequences from *Thermus aquaticus* and *Deinococcus radiodurans* (ES Haas et al, unpublished data) serve to root the evolutionary changes). The deletion of region 183–225 also occurred, independently, in the β-proteobacterium *A. europaenus*. The insertion of helical elements has occurred in the *Bacillus* sequences at positions 80–81, 122–123, and 301–302. Helices 93–96/102–105 and 108–110/115–117 also contain extra sequences in *Bacillus*. The extra sequence lengths associated with these variable regions of RNase P RNA are likely to be peripheral in the enzyme structure, and it is to be expected that
Fig 3. Variation in RNase P RNA in proteobacteria. The secondary structures of RNase P RNA from a representative of the α (R rubrum), β (A eutrophius), γ (C vinosum) and δ (D desulfuricans) proteobacteria are shown.
Consensus core structure

Because each of the known eubacterial RNase P RNAs exhibits catalytic activity, the portions of the molecule that are present in all of the RNAs must contain the essential recognition and catalytic elements of the enzyme. The 'minimal' structure of known eubacterial RNase P RNAs contains 251 nucleotides, 65 of which are invariant (fig 4). The conserved nucleotides 248 and 249 are of particular interest, because deletions in this region have more severe effects on catalysis than in any other portion of the molecule [25]. Crosslinking experiments using 5'-azidophenacyl-tRNA have identified several sites within the RNase P RNA that are within ~9 Å of the substrate phosphodiester bond, in the region of nucleotides 231, 228, 294, and 332 [26]. All these cross-linking sites are in areas of unknown structure (shown unpaired in the models) which contain several invariant nucleotides, including nucleotides 248 and 249.

Comparison of the previously available γ-proteobacterial sequences with those of Bacillus allowed the construction of a minimal RNase P RNA, minIL RNA, which essentially contains the minimal consensus structure and which is catalytically active [27]. Kinetic anomalies of minIL RNA suggest, however, that some perturbation in structure is present relative to native RNase P RNAs. The design of minIL RNA was particularly challenging in the regions containing group-specific elements that are lacking in the Bacillus structures. In the case of minIL RNA, these helices, including the entire region from 134 to 225, were replaced by the corresponding but much shorter sequences from Bacillus megaterium. Because minIL RNA is otherwise based on the E.coli sequence, the replacement of these elements may have led to alteration of the higher-order structure of the RNA. The use of the sequence from A.eutrophus may allow the construction of a new minimal RNase P RNA which lacks helices 184–189/197–202 and 204–211/216–224, with less perturbation of global structure.

Evolution of RNase P RNA

The phylogenetic diversity of RNase P RNAs is of interest not only for the analysis of structure, but also because of what it reveals about the evolution of this enzyme. Because RNase P is required in all known organisms, it is likely an ancient enzyme. The fact that the catalytic moiety of the enzyme is RNA (at least in eubacteria) may suggest that it is directly descended from the postulated 'RNA world', before the emergence of protein-directed catalysis [28, 29].

Fig 4. Conserved RNase P RNA structure and co-variation evidence. Only nucleotides which are present in all, or all but one, of the sequences are included. Nucleotides which vary are indicated by filled circles (●), invariant nucleotides are indicated in upper case letters (G, A, U, or C), and nucleotides which vary in only one sequence are shown in lower case (g, a, u, or c). Nucleotides which are absent in one sequence are indicated by open circles (○). Heavy lines indicate pairing for which there is co-variation evidence; black indicates canonical or G-U pairings, white indicates conserved non-canonical G-A or A-C pairings. The pairing at position 305/326, indicated with half black/half white (●), is canonical in the Bacillus structures, but non-canonical in the proteobacterial structures. Base pairings which contain only G-C and G-U variants are indicated in gray. Base pairings which lack phylogenetic evidence, yet are included in the model because they represent the simple continuation of a proven helix, are indicated by lines (−).

variably present regions are not directly related to enzyme activity.

Change in sequence length outside of these helices is rare and scattered throughout the molecule. Most of the changes are minor and apparently innocuous, involving single bulged nucleotides in helices or the enlargement or reduction of loops by 1 or 2 nucleotides.
Phylogenetic depth of homologous RNase P RNAs

The proteobacteria and Gram-positive eubacteria represent (along with the cyanobacteria) a single cluster of eubacterial phyla. However, RNase P RNAs which are homologous to those of the characterized eubacterial RNase P RNAs are known to exist in other groups of eubacteria. Genes encoding RNase P RNAs resembling those of the proteobacteria have been cloned and sequenced from *T. aquaticus* and *D. radiodurans*, two members of the ‘radiation resistant micrococci and relatives’ eubacterial phylum (ES Haas et al., unpublished data). In addition, a partially characterized clone encoding catalytically active RNase P RNA has been isolated from *Thermotoga maritima* (ES Haas et al., unpublished data), the most evolutionarily divergent eubacterial relative of the proteobacteria and *B. subtilis* [15]. These organisms span the known diversity of the eubacteria, therefore catalytic RNase P RNA must have been present in the ancestor of the eubacteria.

Nuclear-encoded RNase P of eukaryotes so far examined (yeast, fission yeast, and human) are also associated with RNAs [30–33]. These RNase P RNAs are inactivated by micrococcal nuclease, indicating that the RNA component is essential. However, the isolated eukaryotic RNase P RNAs lack the ability to cleave precursor tRNA in the *in vitro* RNA-alone enzyme assay. In addition, these RNAs seem to be minor components of the enzyme, since the buoyant densities of the RNase P activities are only slightly greater than those of bulk protein in cesium sulfate gradients. No significant similarities between eubacterial and eukaryotic RNase P RNAs have been identified, and higher-order structure models that shape the eukaryotic RNAs into eubacterial-like foldings are not convincing. It therefore remains unclear whether or not nuclear and eubacterial RNase P RNAs are homologous.

Little is known about the RNase P of eukaryotic organelles, which would be expected to be derived from their eubacterial ancestors. The mitochondrial RNase P of *S. cerevisiae* contains an essential RNA component of ≈ 490 nt which is only 13% G + C. Much of this RNA can be lost by fragmentation during purification without complete loss of enzymatic activity [34], although activity is sensitive to micrococcal nuclease, indicating that at least some portion of the RNA is required for function. Although RNase P from chloroplasts has been examined biochemically, the structure of the enzyme is unclear [35].

RNase P activities in the archaeobacteria *Halobacterium volcanii* and *Sulfurolobus solfataricus* are also associated with specific RNAs. In the case of *H. volcanii*, this RNA has been cloned and sequenced, and found to contain significant sequence and structural similarities to those of the eubacteria [24]. Moreover, RNase P activity can be reconstituted from the *H. volcanii* RNase P RNA synthesized *in vitro* and the *B. subtilis* RNase P protein component [24]. It seems, therefore, that the RNase P of *H. volcanii* is homologous to the eubacterial enzyme. The RNA of *S. solfataricus* RNase P has yet to be characterized; it is intriguing that this enzyme, like that of the eukaryotes, has a very low buoyant density in cesium sulfate gradients [36].

**Fig 5.** Comparison of RNase P RNA- and 16S rRNA-derived phylogenetic trees. A) RNase P RNA-derived tree, and B) 16S rRNA-derived tree, compiled from [15]. Similarities in the aligned RNase P RNA sequences were determined pairwise and corrected for unequal evolutionary rate [38]. Trees were then constructed by the least-squares distance-matrix method and tested by bootstrap analysis [39, 40]. All positions in the alignment from nucleotides 1 to 373 were included in the analysis; subsets of the alignment, such as the ‘core’ positions, failed to yield trees which were stable in bootstrap analyses. Evolutionary distance is represented by horizontal line length. Recognized phylogenetic groupings (α, β, γ and δ proteobacteria or Gram-positive) are indicated to the right of the 16S rRNA tree. The placement of the *A. eutrophus* RNase P RNA sequence differs from that obtained by analysis of 16S rRNA sequence and most likely represents an artifact of extremely rapid change in the *A. eutrophus* RNase P sequence (see text).
Evolutionary trees based on RNase P RNA

Although the most useful molecules for determining deep phylogenetic relationships are currently ribosomal RNAs [15], it is important to corroborate phylogenetic trees based on rRNAs with those obtained from the sequences of other genes. Evolutionary trees have been constructed from RNase P RNA sequences [15] (fig 5). The two α- proteobacteria, R rubrum and A tumefaciens, and the seven γ-proteobacteria, E coli, S typhimurium, K pneumoniae, A agglomerans, S marcescens, P fluorescens, and C vinsonus, form distinct groups. The branching order within the γ-proteobacteria is the same as in the 16S rRNA tree, except that P fluorescens and C vinsonus are placed in the RNase P RNA tree, whereas they are closely spaced but separate branches in the 16S rRNA tree. The branch in the RNase P RNA tree that specifically groups these two sequences, however, is quite short and not statistically significant; only 60% of trees generated by bootstrapping specifically group these two sequences. The sequence from the β-proteobacterium, T ferrooxidans is shown to be a deep but specific relative of the γ group, and the sequence from the δ-proteobacterium D desulfuricans defines a branch which diverges before the separation of the α, β and γ sequences. These branching orders are in agreement with the relationships established in the 16S rRNA tree. The placement of the β-proteobacterium A eutrophus, however, is clearly unusual.

In addition to its highly divergent sequence, the A eutrophus RNase P RNA has undergone significant structural change – the deletion of two helices (184–189/197–202 and 204–211/216–224) and 3 adjacent nucleotides (183, 203 and 225), probably representing a single 43-nucleotide deletion. It seems possible, therefore, that in order to compensate for the deletion, the RNase P RNA in A eutrophus has undergone rapid evolutionary change compared to T ferrooxidans and the other proteobacteria, and that the anomalous placement of A eutrophus is an artifact. Rapid rates of sequence evolution have also been observed in the 16S rRNA sequences of some β-proteobacteria [15]. The misplacement of sequences as deep divergences is a well-known effect of unequal evolutionary rates [38]. An algorithm designed to compensate for unequal evolutionary rates [37] fails to change the placement of the A eutrophus sequence: the degree of divergence between RNase P RNA sequences is, however, much greater than those of ribosomal RNAs, for which the algorithm is designed and for which effective correction factors for variable mutation rates have been determined.

Conclusion

Although the secondary structure of eubacterial RNase P RNA is now well established, the tertiary structure remains unsolved. In order to construct a more complex structural model, the higher-order interactions in the folding of RNase P RNA must be identified. The phylogenetic comparative approach can be used to identify tertiary contacts as well as secondary interactions [37]. We believe that it will be possible to infer much about the higher order structure of RNase P RNA as the sequences from more diverse organisms are determined; as yet, RNase P has been examined only in a small fraction of the diversity within the eubacteria.

Acknowledgments

We thank E Haus for critical discussions during the preparation of this manuscript. Figures containing RNA structures were prepared with the help of ‘loopDloop’, a Macintosh computer program by Don Gilbert (Indiana University), whom we thank for allowing us to use the program prior to its release.

References

3 Pace NR, Smith DK, Olsen GI, James BD (1989) Gene 82, 65–75
4 Fox GE, Woese CR (1975) Nature (Lond) 256, 505–507
10 Michel F, Umesono K, Ozeki H (1989) Gene 82, 5–30
26 Burgin AB, Pace NR (1991) EMBO J 9, 4111–4117
27 Washburn DS, Green CJ, Pace NR (1989) Science 244, 1569–1571
29 North G (1987) Nature (Lond) 328, 18–19