
The Ribonuclease P database

James W. Brown*, Elizabeth S. Haas, Donald G. Gilbert¹ and Norman R. Pace¹

Department of Microbiology, North Carolina State University, Raleigh, NC 27695 and ¹Department of Biology, Indiana University, Bloomington, IN 47405, USA

ABSTRACT

The Ribonuclease P Sequence database is a compilation of RNase P sequences, sequence alignments, secondary structures, three-dimensional models, and accessory information. In its initial form, the database contains information on RNase P RNA in bacteria and archaea, and RNase P protein in bacteria. The sequences themselves are presented phylogenetically ordered and aligned. The database also contains secondary structures of bacterial and archaeal RNAs, including specially annotated 'reference' secondary structures of *Escherichia coli* and *Bacillus subtilis* RNase P RNAs, a minimum phylogenetic consensus structure, and coordinates for models of three-dimensional structure.

INTRODUCTION

Ribonuclease P is a ribonucleoprotein endonuclease that cleaves transfer RNA precursors, removing 5' precursor sequences and generating the mature 5' terminus of the tRNA (see reference 1 for review). In Bacteria, in which RNase P is best-studied, the holoenzyme consists of two subunits, an RNA of ca. 400 nucleotides (140kD) and a protein of ca. 120 amino acids (14kD) (2). The RNA subunit is the catalyst; bacterial RNase P RNA is active at high ionic strength *in vitro* in the absence of the protein subunit (3). Understanding the mechanisms of substrate recognition and catalysis by RNase P depends on insight into the structure of the RNA component.

Our knowledge of the higher-order structure of RNase P RNA has come from comparative sequence analysis (4). In this method, homologous sequences from different organisms are compared, and structure is identified on the basis of evolutionary 'covariation' of bases that indicates that those bases interact with one another. Comparative analysis is an iterative process; newly identified structure and newly obtained sequences are used to refine the alignment, which in turn provides further structural information. As the number and variety of sequences in the alignment grows, statistical methods can be used to identify covariation (5). In the case of RNase P RNA, there is specific evidence in the form of sequence covariation for nearly every basepair in the current secondary structure model (Figure) (6). Although the assembly of tertiary structure is just beginning, interactions such as pseudoknots (7, 8), helix stacking (9) and base-triples (10) have already been identified. On a (roughly)

yearly basis since 1991, we have compiled and distributed RNase P RNA sequence alignments and secondary structures in an informal publication, 'The Book of P', in order to provide workers in the field a single source for comparative information. We have now expanded and formalized this publication in the form of the Ribonuclease P Sequence database.

DESCRIPTION

The Ribonuclease P Sequence database is a compilation of RNase P sequences, sequence alignments, secondary structures, three-dimensional models, and accessory information available electronically via Internet (see below for details on accessing the database). In its initial form, the database contains information on RNase P RNA in Bacteria and Archaea (Table), and RNase P proteins in Bacteria; later updates will include information on eucaryal and organellar RNase P RNAs, and other RNase P proteins. The database will be updated periodically as sequences and structural information become available.

The main format of the database is The Book of P. The Book of P contains a list of available RNase P RNA sequences, including references and accession numbers; the sequences themselves are presented phylogenetically ordered and aligned. Individual sequences and sequence alignments are also available in machine-readable form (GenBank format) in the database. The Book of P also contains secondary structures of bacterial and archaeal RNAs, including specially annotated 'reference' secondary structures of *Escherichia coli* and *Bacillus subtilis* RNase P RNAs, and minimum phylogenetic consensus structure. The database includes The Book of P in both Macintosh and Postscript formats. The Macintosh files are in MacDraw II and Microsoft Word formats, readable by a variety of Macintosh graphics and word-processing programs. The secondary structures can be downloaded and modified as desired to produce high-quality publishable figures. Postscript files are provided for those using non-Macintosh computers; these files can be displayed and manipulated using Postscript-compatible software and/or printed using Postscript printers.

AVAILABILITY

The Ribonuclease P Database is accessible by anonymous FTP, directly or via Telenet, Gopher, Fetch, etc, from the IUBIO Biology Archives (<ftp.bio.indiana.edu>); the database resides in

*To whom correspondence should be addressed

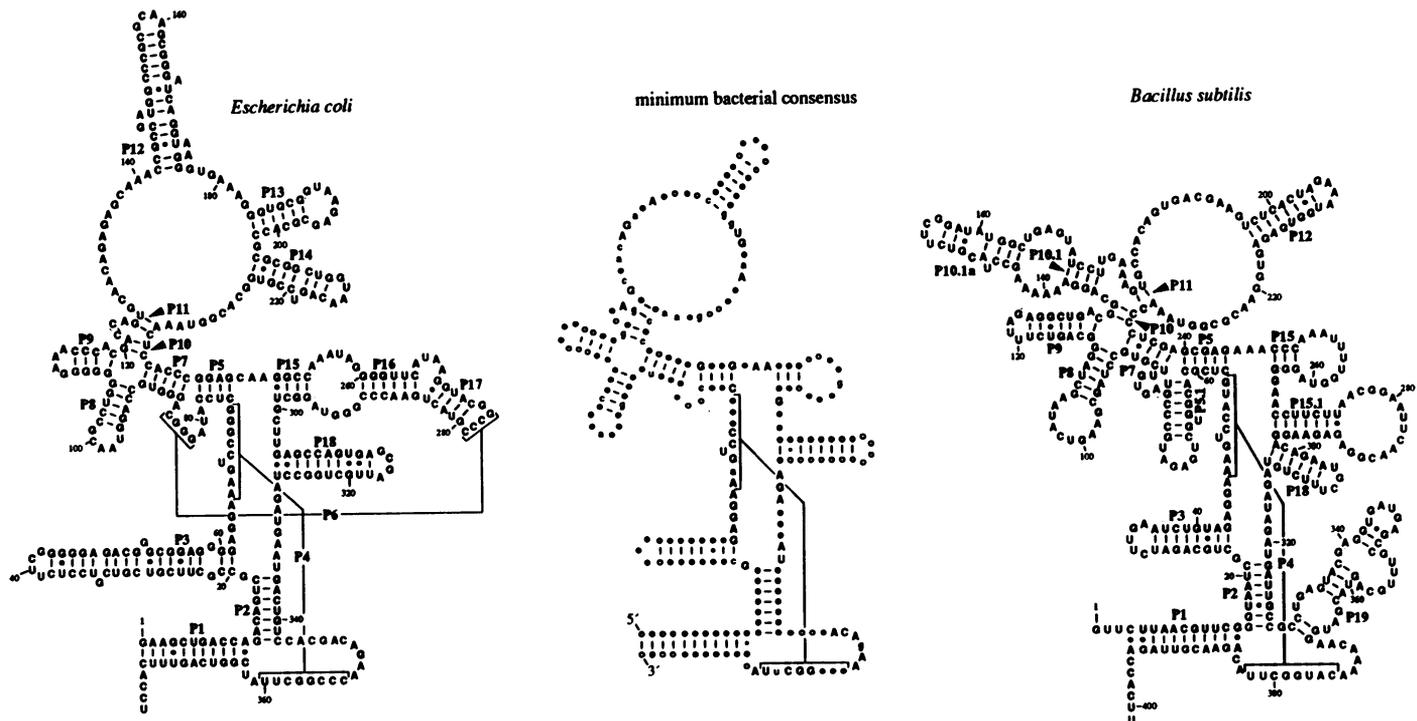


Figure 1. RNase P RNA secondary structure. Secondary structures of *Escherichia coli* and *Bacillus subtilis* RNase P RNAs, and a minimum bacterial consensus structure, are shown. The helix nomenclature used is that developed for group I introns (11) as described for use with RNase P RNA (6). Base-pairings in P4 and P6 are shown as brackets and lines. In the minimum bacterial consensus structure, universally conserved nucleotides are in upper case; those which are at least 80% conserved, but are not invariant, are in lower case. Nucleotides which are not conserved in identity but are present in all sequences are indicated by filled circles; those which are present in at least 80% of sequences, but are absent in at least one, are indicated by open circles. The base-pairings indicated by closed and open dots are, respectively, a conserved non-canonical (G·A or A·C) interaction and a pairing that is frequently G·G (e.g. in *E.coli*) rather than canonical.

the /Molecular Biology/Molbio archive/rnase-p directory. The Book of P is also available on disk for those unable to access the database (please include a formatted 1.44MB Macintosh disk with your request and specify Macintosh or Postscript file formats). The administrator of the RNase P sequence database (J.W. Brown) can be contacted by electronic mail (jwbrown@mbio.ncsu.edu) or by mail at the address given above. Users of the RNase P Sequence Database are encouraged to provide corrections, new information, or other materials for inclusion in the database and the Book of P; unpublished information will be held confidential at the authors request.

ACKNOWLEDGEMENT

Research that led to the establishment of the RNase P Database was supported by NIH grant GM34527 to N.R.P

REFERENCES

- Darr, S. C., Brown, J. W. and Pace, N. R. (1992) *TIBS*, **17**, 178–182.
- Brown, J. W. and Pace, N. R. (1992) *Nucl. Acids Res.*, **20**, 1451–1456.
- Guerrier-Takada, C., Gardiner, K., Marsh, T. L., Pace, N. R. and Altman, S. (1983) *Cell*, **35**, 849–857.
- Pace, N. R., Smith, D. K., Olsen, G. J. and James, B. D. (1989) *Gene*, **82**, 65–75.
- Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. and Stormo, G. D. (1992) *Nucl. Acids Res.*, **20**, 5785–5795.
- Haas, E. S., Brown, J. W., Pitulle, C., and Pace, N. R. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 2527–2531.

- Haas, E. S., Morse, D., Brown, J. W., Schmidt, F. J. and Pace, N. R. (1991) *Science*, **254**, 853–856.
- James, B. D., Olsen, G. J., Liu, J. and Pace, N. R. (1988) *Cell*, **52**, 19–26.
- Brown, J. W., Haas, E. S. and Pace, N. R. (1993) *Nucl. Acids Res.*, **21**, 671–679.
- Brown, J. W., Haas, E. S., Nolan, J. M., Rubio, M. A. and Pace, N. R. (unpublished data)
- Burke, J. M., Belfort, M., Cech, T. R., Davies, R. W., Schweyen, J. R., Shub, D. A., Szostak, J. W. and Tabak, H. F. (1987) *Nucl. Acids Res.*, **15**, 7217–7221.

Table 1. Sequences and structures currently available in the Ribonuclease P sequence database

Bacterial sequences	
Purple bacteria & relatives	
	α sub-division
	<i>Rhodospirillum rubrum</i>
	<i>Agrobacterium tumefaciens</i>
	β sub-division
	<i>Alcaligenes eutrophus</i>
	<i>Thiobacillus ferrooxidans</i>
	<i>Pseudomonas testosteroni</i> (partial)
	γ sub-division
	<i>Escherichia coli</i>
	<i>Salmonella typhimurium</i>
	<i>Klebsiella pneumoniae</i>

Table 1. (cont.)

Bacterial sequences	
	<i>Erwinia agglomerulans</i> <i>Serratia marcescens</i> <i>Pseudomonas fluorescens</i> <i>Chromatium vinosum</i>
δ sub-division	<i>Desulfovibrio</i> <i>desulfuricans</i> <i>Myxococcus xanthus</i> (partial)
Gram positive bacteria	
High G+C subdivision	<i>Streptomyces bikiniensis</i>
Low G+C subdivision	<i>Bacillus brevis</i> <i>Bacillus</i> <i>stearothermophilus</i> <i>Bacillus subtilis</i> <i>Bacillus megaterium</i> <i>Mycoplasma</i> <i>hyopneumoniae</i> <i>Mycoplasma flocculara</i>
Deinococci & relatives	<i>Deinococcus radiodurans</i> <i>Thermus aquaticus</i> <i>Thermus thermophilus</i>
Bacteriodes & relatives	<i>Bacteroides</i> <i>thetaiotaomicron</i> <i>Flavobacterium</i> <i>yabuuchiae</i> (partial)
Thermotogales	<i>Thermotoga maritima</i> <i>Thermotoga neapolitana</i>
Cyanobacteria	<i>Anacystis nidulans</i> (<i>Synechocystis</i> PCC6301) <i>Anabaena</i> sp. PCC 7120 <i>Calothrix</i> sp. PCC7601 <i>Synechocystis</i> sp. PCC6803
Green sulfur bacteria	<i>Chlorobium limicola</i> <i>Chlorobium tepidum</i> (partial)
Archaeal sequences	
Euryarchaeotes	<i>Haloferax volcanii</i>
Crenarchaeotes	<i>Sulfolobus acidocaldarius</i>