

# Evaluation and refinement of tmRNA structure using gene sequences from natural microbial communities

SCOTT T. KELLEY,<sup>1</sup> J. KIRK HARRIS,<sup>1,2</sup> and NORMAN R. PACE<sup>1</sup>

<sup>1</sup>Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309, USA

<sup>2</sup>Graduate Group in Microbiology, University of California, Berkeley, California 94720, USA

## ABSTRACT

DNA harvested directly from complex natural microbial communities by PCR has been successfully used to predict RNase P RNA structure, and can potentially provide an abundant source of information for structural predictions of other RNAs. In this study, we utilized genetic variation in natural communities to test and refine the secondary and tertiary structural model for the bacterial tmRNA. The variability of proposed tmRNA secondary structures in different organisms and the lack of any predicted tertiary structure suggested that further refinement of the tmRNA could be useful. To increase the phylogenetic representation of tmRNA sequences, and thereby provide additional data for statistical comparative analysis, we amplified, sequenced, and compared tmRNA sequences from natural microbial communities. Using primers designed from gamma proteobacterial sequences, we determined 44 new tmRNA sequences from a variety of environmental DNA samples. Covariation analyses of these sequences, along with sequences from cultured organisms, confirmed most of the proposed tmRNA model but also provided evidence for a new tertiary interaction. This approach of gathering sequence information from natural microbial communities seems generally applicable in RNA structural analysis.

**Keywords:** covariation; phylogeny; RNA structure; tmRNA

## INTRODUCTION

Phylogenetic comparative methods have proven effective for inferring the secondary and tertiary structures of RNA molecules (Woese et al., 1980; Williams & Bartel, 1996; Frank & Pace, 1998; Wassarman et al., 1999). Comparative methods use sequences of homologous RNAs from diverse organisms as natural sources of variation to detect correlated change (covariation) between nucleotide positions (Gutell et al., 1992). Such correlated changes are evidence of direct interactions in homologous RNAs. Comparative analysis has been used to establish the structures of several large and small RNA molecules, for instance the rRNAs (Fox & Woese, 1975; Woese et al., 1980; Gutell et al., 1990), tRNA (Gutell et al., 1992), and RNase P RNA (Brown et al., 1996).

Accurate structure predictions using methods of statistical comparative analysis typically require a large number of sequences (>30) from diverse organisms

(Gutell et al., 1992; Akmaev et al., 1999). This extent of data is available for some RNAs, but other stable RNAs such as 6S RNA, OxyS RNA, and CsrB RNA have too few sequences available to analyze with statistical comparative methods (Wassarman et al., 1999). On the other hand, some small RNAs, such as tmRNA, are well represented by 50 or more highly variable sequences that have been used for structure prediction (Williams, 2000; Zwieb & Samuelsson, 2000).

tmRNA, named for its dual tRNA and mRNA roles, is a small RNA so far found exclusively in the Bacteria, and is conserved throughout that phylogenetic domain (Keiler et al., 2000). tmRNA frees ribosomes stalled on mRNAs that lack stop codons (Williams & Bartel, 1996; Williams et al., 1999). tmRNA contains a structural element that mimics a tRNA, for interaction with the ribosome, and a translational open reading frame (mRNA element). Translation of the mRNA element adds a short polypeptide to the incomplete protein that targets it for degradation (Nameki et al., 1999). Recent studies indicate that tmRNA may require an RNA-binding protein (smpB) to function properly as well (Karzai et al., 1999, 2000).

Reprint requests to: Norman R. Pace, Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309, USA; e-mail: Norman.Pace@Colorado.edu.

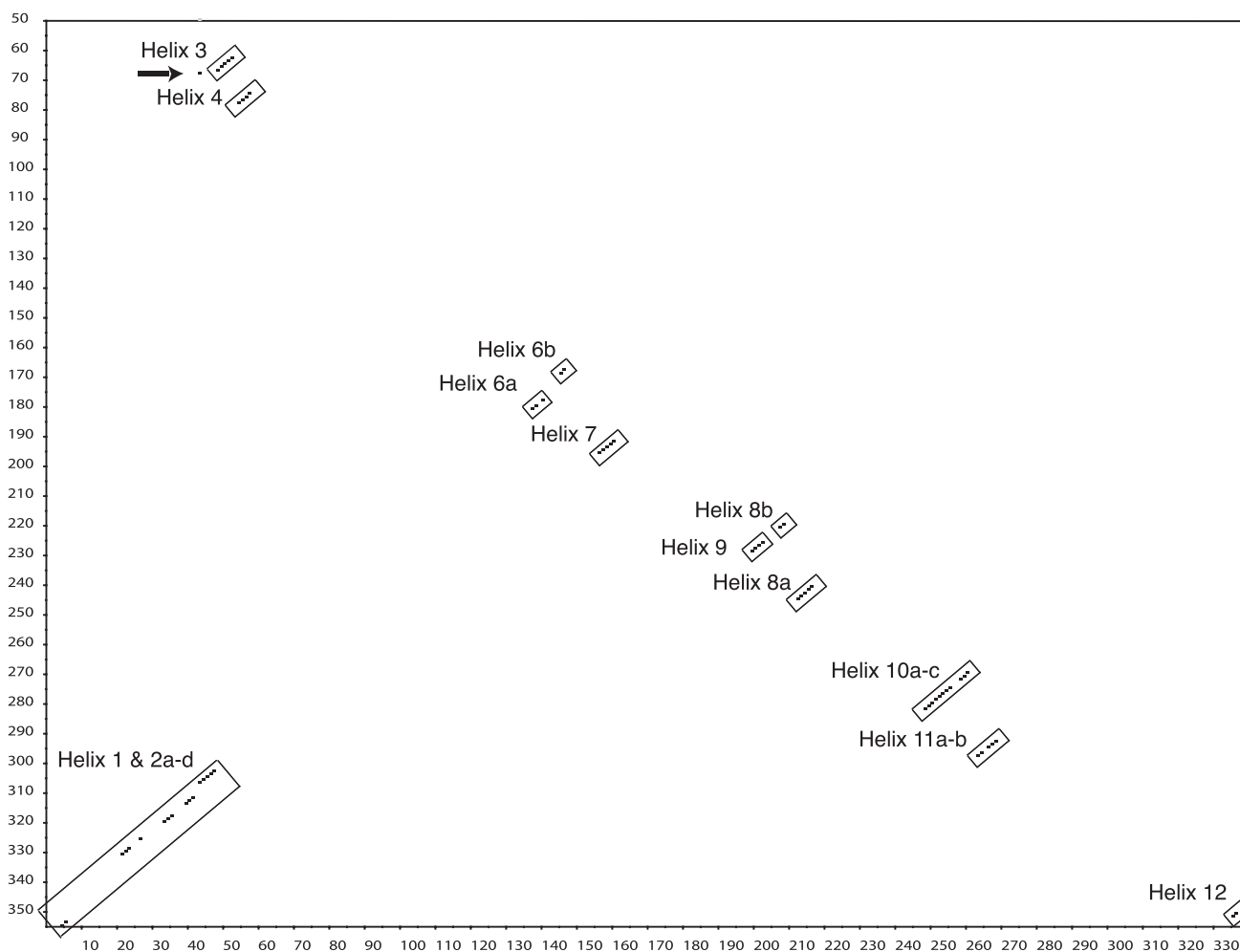


the statistical analyses, we used only the positions of the alignment where there was not a gap in the *Escherichia coli* sequence, as we were testing the validity of this particular model. We included only the most diverse sequences from the natural population analysis (a total of 20 sequences) for the analysis because many differed from other sequences at only a few nucleotide positions (<2% divergence). The alignment used can be downloaded in genbank format by anonymous ftp from ftp://vent.colorado.edu.

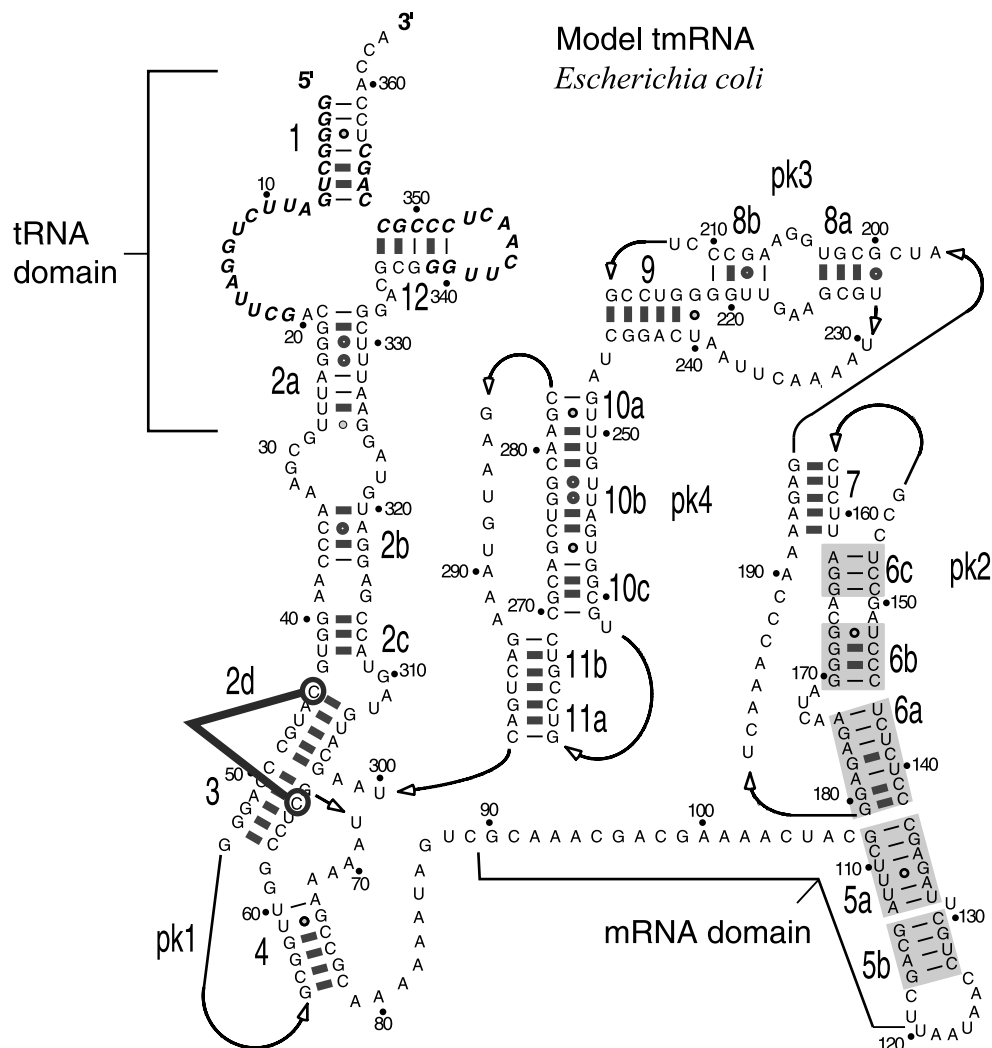
The results of the  $R_{ij}/H_{ij}$  statistical analysis provided strong evidence for most of the helical base pairings in the proposed tmRNA model (Fig. 3). All but four of the proposed helical regions in the secondary structure model had highly significant support with  $H_{ij}$  values exceeding 27 and most exceeding 35 ( $\chi$ -square test,  $P < 0.0001$ ; 9 d.f.), which testifies to the soundness of the basic model. The MI analysis also found statistical support for a majority of the helices in the model, al-

though there was a great deal more noise in the MI analysis than in the  $R_{ij}/H_{ij}$  analysis (data not shown).

The distribution of supported base pairings and proposed base pairs that have no support are summarized in Figure 4. Lack of support for some helical regions does not necessarily mean that the helices do not exist in any particular RNA, but rather that these regions cannot be generalized to tmRNA from all organisms in the data set. In the cases of helical regions with little support (5a, 5b, and 6a–c), we checked whether the failure to predict these regions was due to a lack of variation in the data set, or whether these regions in fact do not occur in all the sequences. Indeed, several sequences had anomalies in sequence regions corresponding to helical regions 5a, 5b, 6a, 6b, and 6c. The alignments of these regions contained gaps, and there was no obvious structure in some of the sequences whereas others contained credible hairpin loops. Highly variable structural elements are common



**FIGURE 3.** Results from the  $R_{ij}/H_{ij}$  statistical analysis of the natural community tmRNA sequence data set along with tmRNA sequences from cultured organisms. The position numbers corresponding to *E. coli* nucleotides are given along both the X and Y axes. The scores presented all exceed 27 (ranging from 27 at the lowest to 120 at the highest) and are significant at least at the  $P < 0.001$  level ( $\chi$ -square test; 9 d.f.). The arrow indicates the predicted tertiary interaction (see text).



**FIGURE 4.** Support for the proposed secondary structure tmRNA model from  $R_{ij}/H_{ij}$  statistical analysis. The thickened bars (or dots for noncanonical pairings) indicate statistically supported covariations, and thin bars (or dots) indicate pairings from the model that were not supported in the analysis. The model is based on the Zwieb et al. (1999) figure for tmRNA *E. coli* secondary structure. The thick line indicates the one significant predicted tertiary interaction ( $H_{ij} > 35$ ;  $P < 0.0001$ ,  $\chi$ -square test; 9 d.f.) The regions shaded in gray denote helices that were not statistically supported in the various analyses reported here. The nucleotides corresponding to the primer sequences are shown in bold italics.

in all RNAs and generally need to be excluded from sequence alignments used for statistical comparative or phylogenetic analyses because they are not identifiable homologs. The  $H_{ij}$  analysis also detected a potential tertiary interaction between nt 44 and 66 (Fig. 4) with an  $H_{ij}$  value exceeding 35 ( $\chi$ -square test,  $P < 0.0001$ ; 9 d.f.).

## DISCUSSION

The collection of sequences from natural microbial communities proved to be a useful approach for testing and refining the structure of tmRNA. Using PCR primers based on gamma proteobacterial tmRNA sequences, we were able to add a significant number of new sequences to the tmRNA collection, more than doubling

the present database of gamma proteobacterial sequences (Williams, 2000). Choice of primers proved to be critical in obtaining genes from natural communities, and we were unable to amplify sequences from bacterial divisions other than the proteobacterial division. Some of these DNA samples had been used in previous studies to amplify 16S ribosomal DNA from a number of bacterial divisions, so our failure to amplify tmRNA sequences from other divisions was not because the samples contained only gamma-proteobacteria (Dojka et al., 2000). Moreover, the primers we designed to amplify other bacterial divisions did not even work on the positive controls (i.e., DNA from cultured organisms belonging to these divisions). We suspect that more sequence data will need to be gathered from cultured members of these other divisions, or from genome se-

quences, so that better primers may be designed to amplify sequences from these groups.

Analysis of the environmental sequences, along with sequences from cultured organisms, provided strong statistical support for the majority of the proposed tmRNA model (Figs. 3 and 4). In addition to testing the secondary structure, we also were able to detect a new well-supported higher order interaction in tmRNA (Fig. 4). The statistical support for this covariation ( $H_{ij} > 35$ ,  $P < 0.0001$ ; Fig. 4) exceeds the values of many of the Watson–Crick pairings in the proposed tmRNA model for the same data set. The interaction occurs between 2 nt apparently involved in base pairs (Fig. 4). Most of the pair-pair combinations were C|G–G|C, with C|G–A|T and T|A–T|A being the next most common sets of combinations. Although interactions between nucleotides involved in separate base pairs are unusual, these types of covariations have been discovered in other RNA molecules (Brown et al., 1996; Kelley et al., 2000). The bases in the potential interaction are separated by half of a helical turn, approximately 17 Å, and the nucleotides of interest may be a good deal closer depending on the local structure (pk1; Fig. 4). Thus, although the three-dimensional structure of this region is unknown, we believe that these base pairs may be close enough to interact directly in some fashion.

We were not able to find evidence for the other recently discovered tertiary interaction between 2 nt in the single-stranded regions of pseudoknots pk3 and pk4 (Felden et al., 2001; Fig. 4). This previously discovered tertiary interaction was identified using 13 beta-proteobacterial sequences and apparently involved two independent changes in the beta-proteobacteria at nucleotide positions 234 and 286 (*E. coli* numbering) from  $A_{234}/A_{286}$  to  $G_{234}/G_{286}$ . However, in the larger data set of tmRNA sequences we used in our analyses, we found no significant covariation between these positions and most of the sequences had either  $A_{234}/A_{286}$  or  $T_{234}/A_{286}$  at these positions.

Although covariation support for most of the model is robust, there were several parts of the secondary structure model that had no support even among the closely related gamma proteobacterial tmRNA sequences. For instance, three of the helices, 5a, 5b, and 6c, had no statistical support in the overall alignment, and 6a and 6b had support of 50% or less of the base pairings in the model (Fig. 4). Most of the sequences we collected had these helical elements, and aligned credibly in these regions, but several of the sequences contained substantial gaps or had sequences in these regions that did not present obvious structure. Of course, the lack of statistical support for these tmRNA helices does not mean they are not part of the *E. coli* tmRNA or other proteobacterial tmRNAs, but rather that these helices are not consistent properties of tmRNA structures from different organisms. The fact that alternative structures may occur even among closely related sequences in-

dicates low levels of evolutionary constraint on this region of the tmRNA secondary structure.

The extensive structural variation in diverse tmRNA sequences, and the sporadic occurrence of some helical regions, has been noted previously (Williams & Bartel 1998; Zwieb et al., 1999). Extensive structural variability occurs even within the gamma-group proteobacterial sequences reported here, confirming the notion that much of the tmRNA structure can vary with little effect on its basic function. Indeed, experimental modifications of the *E. coli* tmRNA have shown that three of the pseudoknots (pk2, pk3, and pk4; Fig. 4) are completely interchangeable and can even be replaced by unstructured regions with no significant loss of function (Nameki et al., 2000). We suspect that these phylogenetically volatile regions of the structure occur on the surface of the functional unit because of space-filling constraints (Burgin et al., 1990), whereas conserved structural elements and sequences are expected to form the functional core of tmRNA. A recent cross-linking study also may shed some light on the variability of tmRNA across different lineages. In this study, the investigators discovered substantial crosslinking between tmRNA and ribosomal protein S1 (Wower et al., 2000). Ribosomal protein S1, however, is not found in all bacterial lineages, specifically the Low G+C group of Gram-positive bacteria. Thus, at least in this group of Bacteria, tmRNA must find an alternative binding site. If the binding site of tmRNA can change radically, as this study suggests, then such extensive variability in tmRNA secondary structure may be expected.

## MATERIALS AND METHODS

### Sample collection and DNA extraction

The clone library designation codes, locales, and sample types collected for DNA extraction were: (1) FS: Fairfax Swamp, Indiana, from sediment; (2) LEM: Lake Lemon, Indiana, from sediment; (3) QL: Queen's Laundry, Yellowstone, from a microbial mat; (4) RCA: rotting cactus from Arizona; (5) VLS: Varsity Pond, University of Colorado at Boulder, from sediment; (6) VLW: Boulder Creek in Boulder, Colorado, from a water sample; and (7) WW: Varsity Pond, University of Colorado at Boulder, from a water sample. DNA was extracted using a bead beating protocol. We suspended 0.5 to 1.0 g of sample in 0.5 mL of an SDS/buffer solution (200 mM Tris, pH 8.0, 20 mM EDTA, 200 mM NaCl, and 2% SDS) and incubated for 20 min at 70 °C. After adding 0.3 g of acid-washed zirconium-silica beads (0.1 mm diameter) and phenol-chloroform, the samples were agitated on a Mini-Beadbeater (Biospec, Inc.) at low speed for 2 min. Nucleic acids were precipitated from the supernatant by addition of 50 µL 3 M sodium acetate, pH 5.2, and 500 µL isopropanol. After a wash with 500 µL 70% ethanol, the samples were resuspended in 50 µL TE, pH 8.0.

## PCR and cloning

Community tmRNA sequences were amplified by PCR using the following general protocol: 1 to 50 ng of DNA in reaction mixtures containing (as final concentrations) 1× PCR buffer II (Perkin Elmer), 2.5 mM MgCl<sub>2</sub>, 200 μM of each deoxy-nucleoside triphosphate, 300 nM of each forward and reverse primer, 2.5 μL DMSO, and 0.025 U of AmpliTaq Gold DNA polymerase (Perkin Elmer) per milliliter. Reaction mixtures were incubated in a Mastercycler Gradient thermal cycler (Eppendorf) at 94 °C for 12 min (for initial denaturation and activation of AmpliTaq Gold), followed by 35 to 40 cycles at 94 °C for 30 s, 55 °C ± 10 °C for 45 s, and 72 °C for 1.5 min, and then by a final extension period of 20 min at 72 °C. We designed primers for tmRNA amplifications using the sequence alignment from the tmRNA database (Williams, 2000) to target specific groups of bacteria, including the gamma, delta, and epsilon proteobacterial groups, and the Low G+C Gram positive division. However, only the primers designed for the gamma proteobacteria amplified tmRNA effectively: SGPROTM1 (5'-GGGGCTGATTCTGGATTCG-3'), and AGPROTM1 (5'-GCTGGSGGGAKTTGAACC-3'). PCR products were cloned with a TOPO TA Cloning Kit following the manufacturer's protocol (Invitrogen Corp.). Plasmid DNAs containing inserts were determined by PCR amplification with T3/T7 primers. Unique inserts were identified by RFLP, using the restriction enzymes *Hin*P1I and *Msp*I (New England Biolabs, Inc.), and sequencing. A detailed description of these methods has been published previously (Dojka et al., 1998). The tmRNA sequences determined were deposited in GenBank under the accession numbers AF389942 to AF389985.

## Alignment and comparative analyses

The new tmRNA sequences were first aligned to the known gamma-proteobacterial sequences using Clustal W (Aiyar, 2000) leaving the original alignment unchanged. We then added the natural community sequences to 87 other tmRNA sequences collected from the internet (Knudsen et al., 2001) and manually refined the alignment by keying on elements of structure using the ARB Program (Strunk & Ludwig, 1999). The basic proposed tmRNA secondary structure model was used as a template to align homologous regions (Williams & Bartel, 1996; Zwieb et al., 1999). All of the comparative analyses were based on multiple alignments that included only the positions that were homologous to the *E. coli* nucleotides. Other sections of the alignments contained numerous gaps and were not considered reliable.

To predict tmRNA structure, we calculated the phylogenetically based R<sub>ij</sub> and H<sub>ij</sub> statistics for the data set using the, appropriately titled, Rij and Hij programs compiled on a PC running Slackware Linux version 3.5 (Akmaev et al., 2000). These statistics incorporate information from the phylogenetic relationships among the sequences and have been shown to be more accurate than standard mutual information methods (Akmaev et al., 1999, 2000). The phylogenetic tree for the 107-sequence tmRNA data set used in the statistical analysis was calculated using the neighbor-joining algorithm in the PHYLIP phylogeny package (Felsenstein, 1993). For given pairs of positions, the R<sub>ij</sub> and H<sub>ij</sub> statistics compare the rates of evolution of each position independently (calculated as the independent likelihood) to the joint rates of evolution (joint

likelihood) of the two positions. If the independent rate of evolution for two positions is very high (low independent likelihoods) but the positions always change together (high joint likelihood) then the values for these statistics will be high. Both statistics calculate the independent likelihoods using the phylogenetic tree, but only H<sub>ij</sub> utilizes the phylogenetic tree to calculate the joint likelihoods, making this statistic much slower to calculate. However, the H<sub>ij</sub> statistic approximates a  $\chi^2$  distribution with nine degrees of freedom, allowing for an assessment of confidence in particular proposed pairings. In the course of the analysis, we utilized the computationally faster R<sub>ij</sub> method to identify the set of initial pairs with high correlation values and then used the computationally intensive H<sub>ij</sub> statistic to determine which of these pairs were statistically significant. We also performed mutual information (MI) analyses using the in-house ALEX program (version 1.0) designed by Dan Frank, which calculated MI values based on previous work (Gutell et al., 1992).

## ACKNOWLEDGMENTS

We thank A. Buck, D. Evans, A. Kazantsev, S. Marquez, J. Spear, V. Thackray, K. Williams, and anonymous referees for useful comments on the manuscript. This research was supported by a grant from the National Institutes of Health to N. R. Pace and a National Research Service Award fellowship to S. Kelley.

Received April 18, 2001; returned for revision

May 10, 2001; revised manuscript received June 15, 2001

## REFERENCES

- Aiyar A. 2000. The use of CLUSTAL W and CLUSTAL X for multiple sequence alignment. *Methods Mol Biol* 132:221–241.
- Akmaev VR, Kelley ST, Stormo GD. 1999. A phylogenetic approach to RNA structure prediction. *ISMB* 10–17.
- Akmaev VR, Kelley ST, Stormo GD. 2000. Phylogenetically enhanced statistical tools for RNA structure prediction. *Bioinformatics* 16:501–512.
- Brown JW, Nolan JM, Haas ES, Rubio MA, Major F, Pace NR. 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.
- Burgin AB, Parodos K, Lane DJ, Pace NR. 1990. The excision of intervening sequences from Salmonella 23S ribosomal RNA. *Cell* 60:405–414.
- Dojka MA, Harris JK, Pace NR. 2000. Expanding the known diversity and environmental distribution of an uncultured phylogenetic division of bacteria. *Appl Environ Microbiol* 66:1617–1621.
- Dojka MA, Hugenholtz P, Haack SK, Pace NR. 1998. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl Environ Microbiol* 64:3869–3877.
- Felden B, Massire C, Westhof E, Atkins JF, Gesteland RF. 2001. Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucleic Acids Res* 29:1602–1607.
- Felsenstein J. 1993. PHYLIP Phylogeny Inference Package. version 3.5c. Department of Genetics, University of Washington, Seattle.
- Fox GE, Woese CR. 1975. The architecture of 5S rRNA and its relation to function. *J Mol Evol* 6:61–76.
- Frank DN, Pace NR. 1998. Ribonuclease P: Unity and diversity in a tRNA processing ribozyme. *Annu Rev Biochem* 67:153–180.
- Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD. 1992. Identifying constraints on the higher-order structure of RNA: Continued development and application of comparative sequence analysis methods. *Nucleic Acids Res* 20:5785–5795.

- Gutell RR, Schnare MN, Gray MW. 1990. A compilation of large subunit 23S-like. Ribosomal RNA sequences presented in a secondary structure format. *Nucleic Acids Res* 18 (suppl.): 2319–2330.
- Karzai AW, Roche ED, Sauer RT. 2000. The SsrA-SmpB system for protein tagging, directed degradation and ribosome rescue. *Nature Struct Biol* 7:449–455.
- Karzai AW, Susskind MM, Sauer RT. 1999. SmpB, a unique RNA-binding protein essential for the peptide-tagging activity of SsrA tmRNA. *EMBO J* 18:3793–3799.
- Keiler KC, Shapiro L, Williams KP. 2000. tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*. *Proc Natl Acad Sci USA* 97:7778–7783.
- Kelley ST, Akmaev VR, Stormo GD. 2000. Improved statistical methods reveal direct interactions between 16S and 23S rRNA. *Nucleic Acids Res* 28:4938–4943.
- Knudsen B, Wower J, Zwieb C, Gorodkin J. 2001. tmRDB tmRNA database. *Nucleic Acids Res* 29:171–172.
- Nameki N, Chattopadhyay P, Himeno H, Muto A, Kawai G. 1999. An NMR and mutational analysis of an RNA pseudoknot of *Escherichia coli* tmRNA involved in *trans*-translation. *Nucleic Acids Res* 27:3667–3675.
- Nameki N, Tadaki T, Himeno H, Muto A. 2000. Three of four pseudoknots in tmRNA are interchangeable and are substitutable with single-stranded RNAs. *FEBS Lett* 470:345–349.
- Strunk O, Ludwig W. 1999. ARB: A software environment for sequence data. <http://www.mikro.biologie.tu-muenchen.de>
- Wassarman KM, Zhang A, Storz G. 1999. Small RNAs in *Escherichia coli*. *Trends Microbiol* 7:37–45.
- Williams KP. 2000. The tmRNA website. *Nucleic Acids Res* 28:168.
- Williams KP, Bartel DP. 1996. Phylogenetic analysis of tmRNA secondary structure. *RNA* 2:1306–1310.
- Williams KP, Bartel DP. 1998. The tmRNA website. *Nucleic Acids Res* 26:163–165.
- Williams KP, Martindale KA, Bartel DP. 1999. Resuming translation on tmRNA: A unique mode of determining a reading frame. *EMBO J* 18:5423–5433.
- Woese CR, Magrum LJ, Gupta R, Siegel RB, Stahl DA, Kop J, Crawford N, Brosius J, Gutell R, Hogan JJ, Noller HF. 1980. Secondary structure model for bacterial 16S ribosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res* 8:2275–2293.
- Wower IK, Zwieb CW, Guven SA, Wower J. 2000. Binding and cross-linking of tmRNA to ribosomal protein S1: on and off the *Escherichia coli* ribosome. *EMBO J* 19:6612–6621.
- Zwieb C, Samuelsson T. 2000. SRPDB signal recognition particle database. *Nucleic Acids Res* 28:171–172.
- Zwieb C, Wower I, Wower J. 1999. Comparative sequence analysis of tmRNA. *Nucleic Acids Res* 27:2063–2071.