RNA (1996), 2:1306–1310. Cambridge University Press. Printed in the USA.

Phylogenetic analysis of tmRNA secondary structure

KELLY P. WILLIAMS and DAVID P. BARTEL
Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA

ABSTRACT

The bacterial tmRNA acts with dual tRNA-like and mRNA-like character to tag incomplete translation products for degradation. Comparative analysis of 17 tmRNA genes (including eight new sequences) has allowed us to deduce conserved features of the tmRNA secondary structure. Except in a segment that includes the first codon of the tag reading frame, tmRNA is highly structured, with four pseudoknots and a total of 11 conserved base pairing regions. The previously identified tRNA minihelix structure is connected by a long base paired region to a large structured domain composed of a pseudoknot, followed by the tag reading frame and a string of three rather similar pseudoknots. The conservation of numerous structural elements among diverse eubacterial species indicates that these elements have important function beyond simply forming an endonuclease-resistant link between the reading frame and the tRNA-like domain.

Keywords: peptide tagging; proteolysis; RNA structure; 10Sa RNA; *trans*-translation

INTRODUCTION

The bacterial 10Sa RNA has been renamed tmRNA to reflect recent appreciation of its combined tRNA-like and mRNA-like properties; it is charged with alanine (Komine et al., 1994; Ushida et al., 1994), yet also bears a short reading frame (Tu et al., 1995; Keiler et al., 1996). It mediates tagging of the incomplete protein product of a ribosome that reaches the end of a truncated mRNA without encountering a stop codon; the peptide tag is recognized subsequently by proteases that degrade the abnormal protein (Keiler et al., 1996). According to the proposed mechanism for tagging (Keiler et al., 1996), tmRNA charged with alanine enters the A site of the stalled ribosome, and the alanine is transferred to the nascent polypeptide; in a remarkable switching event, the tmRNA reading frame replaces the truncated mRNA and polypeptide elongation resumes.

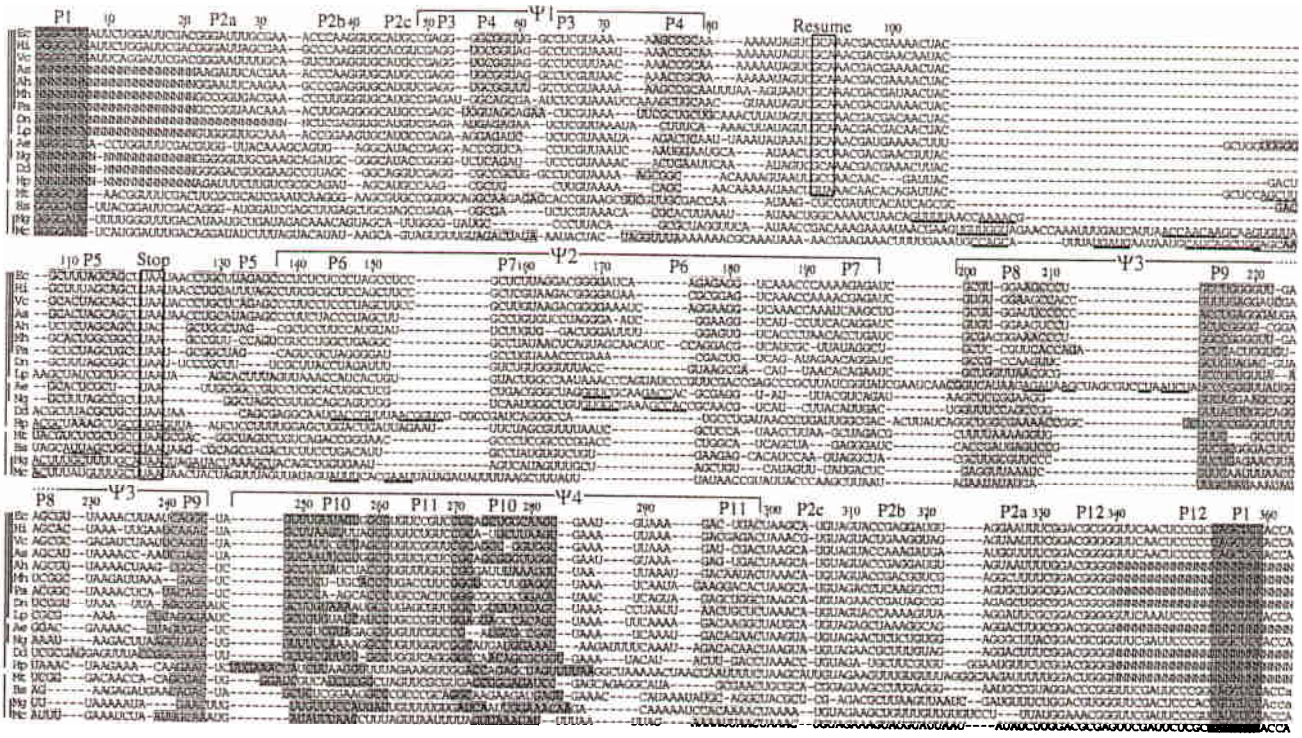
Study of the mechanism of *trans*-translational tagging would be facilitated by knowledge of tmRNA secondary structure, yet recent models conflict and leave at least three-quarters of the 350–400-nt RNA unmodeled. A tRNA minihelix structure involving 35 nt at the termini of the tmRNA has been identified, analogous to the coaxial T stem-loop and CCA-tailed acceptor stem (Komine et al., 1994). Subsequent secondary structure proposals have extended the analogy with tRNA structure in mutually exclusive ways. An analogue of the anticodon stem has been proposed, with the bulk of the RNA replacing the anticodon loop (Ushida et al., 1994). An alternative model proposes analogues of both the anticodon and D stem-loops, with the bulk of the RNA inserted between the acceptor stem and D stem analogues (Felden et al., 1996). We have expanded the list of tmRNA sequences and used comparative analysis to model secondary structure throughout the molecule.

Reprint requests to: David P. Bartel, Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; e-mail: dbartel@wi.mit.edu.

RESULTS AND DISCUSSION

Eight new tmRNA sequences were determined from species that were chosen both to focus attention on the γ3 proteobacteria (which include *Escherichia coli*) and to survey smoothly the rest of the proteobacteria. A convincing primary alignment could be made for the tmRNA sequences from *E. coli* and seven close relatives (top eight lines of Fig. 1A). Helical regions were identified by base covariations according to classical criteria (Noller et al., 1981; James et al., 1988). Including the more distantly related sequences in the alignment allowed identification of additional helices, bringing the number of conserved paired regions to 11 (P1–P4 and P6–P12), 8 of which are organized as four pseudoknots (ψ1–ψ4). An additional pairing (P5) is plausible, particularly in the γ3 proteobacteria. However, the existence of this paired region is not proven by covariation

A



B

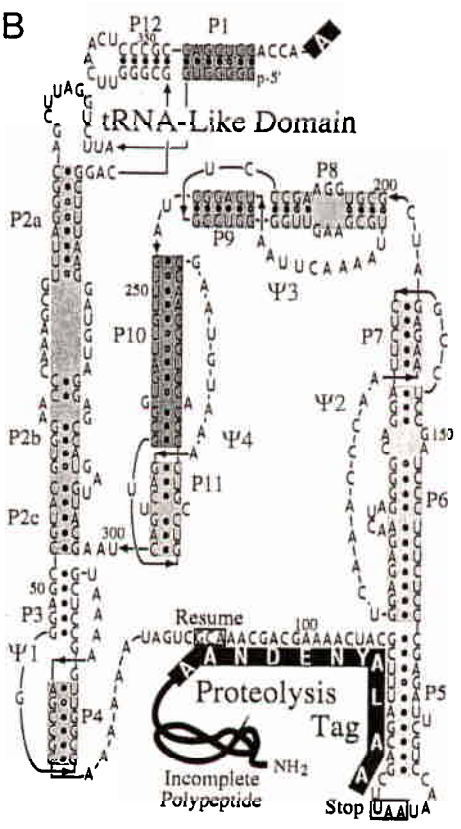


FIGURE 1. tmRNA secondary structure. **A:** Comparative sequence analysis. New tmRNA sequences and previous GenBank entries are shown with deduced base pairing segments in color-coded boxes, each helix generally displayed at the maximal length allowed by canonical base pairs. As discussed in the Materials and methods, primary sequence alignment is considered sound only for the top eight sequences. N indicates bases that were not sequenced. Underlines suggest idiosyncratic pairings that might occupy apparent sequence inserts. The lower-case letters at the end of the Mt and Bs sequences signify unencoded residues that can be added by tRNA nucleotidyltransferase (Ushida et al., 1994). Predicted (known for *E. coli*) resume and stop codons of the tag reading frame are in open boxes. Suspected pairing within unusually long joining regions (black underline) and alternate pairing possibilities within the P5 region (red underline) are indicated. Abbreviations: Ec, *Escherichia coli*; Hi, *Haemophilus influenzae*; Vc, *Vibrio cholerae*; As, *Aeromonas salmonicida*; Ah, *Alteromonas haloplanktis*; Mh, *Marinobacter hydrocarbonoclasticus*; Pa, *Pseudomonas aeruginosa*; Dn, *Dichelobacter nodosus*; Lp, *Legionella pneumophila*; Ae, *Alcaligenes eutrophus*; Ng, *Neisseria gonorrhoeae*; Dd, *Desulfovibrio desulfuricans*; Hp, *Helicobacter pylori*; Mt, *Mycobacterium tuberculosis*; Bs, *Bacillus subtilis*; Mg, *Mycoplasma genitalium*; Mc, *Mycoplasma capricolum*. Phylogenetic relations (indicated by lines at left) can be summarized: proteobacteria of subgroups $\gamma 3$ (Ec, Hi, Vc, As, Ah, Mh, Pa), $\gamma 1$ (Dn), $\gamma 2$ (Lp), β (Ae, Ng), δ (Dd), and ϵ (Hp); Gram-positive bacteria (Mt, Bs, Mg, Mc). **B:** *E. coli* tmRNA secondary structure. The internal loop between P2a and P2b should not be viewed as an analogue of an anticodon loop, because this region is almost completely paired in other species. Charging and tag amino acids are shown in white letters; tilted A refers to the proposal that the uncoded first tag residue comes from the charged 3′ end of the same tmRNA molecule used to translate the rest of the tag (Keiler et al., 1996).

and its precise location would be in doubt in species distant from *E. coli* due to alternate pairing possibilities (Fig. 1A). In the proposed *E. coli* structure (Fig. 1B), 57% of the nucleotides participate in pairing (Watson-Crick or G:U). This percentage approaches those of other stable *E. coli* RNAs (tRNAs, rRNAs, SRP RNA, and ribonuclease P RNA are 55–64% paired; Gutell et al., 1994; Nolan & Pace, 1996).

The P1 and P12 pairings are undisputed analogues of the coaxial acceptor and T stems, with the $tRNA^{Ala}$ identity determinants (Komine et al., 1994). This unit is connected via highly conserved sequences to P2a. The bulk of the RNA exits the other end of the P2 region in another distinctive structural context: P2c is likely to stack coaxially with P3 (intervening bases are never found), and is connected to the most uniform of the pseudoknots, ψ4. Although pairing in the central region of P2 is not uniform, the P2a and P2c end pairings are universal. P2 extends a stem proposed previously by Ushida and coworkers (1994), but directly conflicts with a more recent model, which proposes analogues of both the anticodon and D stem-loops (Felden et al., 1996).

The tag sequence is known by protein sequencing only for *E. coli* (Tu et al., 1995; Keiler et al., 1996); however, the stop codons can be identified in all tmRNAs, and among the proteobacteria, the resume codon is also well aligned, allowing prediction of the tag sequences (Fig. 2). Alignment of the resume codons for the Gram-positive bacteria is not secure; therefore, it is not clear whether the large insertions between P4 and the stop codon in *Mycoplasma genitalium* and *Mycoplasma capricolum* lie upstream of the reading frame or could extend the tag reading frame to twice the usual proteobacterial length. At least some variability is observed in the length of the predicted tags; for example, *Helicobacter pylori* and *Legionella pneumophila* have 4-codon insertions relative to most of the other proteobacteria.

The resume codon resides in the longest unpaired region of the molecule, recalling the unstructured start codon regions of efficiently translated bacterial mRNAs (de Smit & van Duin, 1994). The lack of secondary structure, together with the observed variation in predicted tags, has implications for efforts to engineer the tmRNA-encoded peptide tag: there may be great freedom to alter or insert codons in the unpaired regions of the reading frame (the upstream portion and the loop containing the stop codon). Codon substitution disrupting P5 should be corrected easily by compensatory change. The only serious constraint may be the yet unknown signals for resume codon selection.

```
Ec ANDENY----ALAA
Hi ANDEQY----ALAA
Vc ANDENY----ALAA
As ANDENY----ALAA
Ah ANDDNY----SLAA
Mh ANDENY----ALAA
Pa ANDDNY----ALAA
Dn ANDDNY----ALAA
Lp ANDENFAGGEAIAA
Ae ANDERY----AL-A
Ng ANDETY----ALAA
Dd ANN-DY--DYAYAA
Hp VNNTDYAPAYAKAA
Mt            ...RLAA
Bs            ...ALAA
Mg            ...FAFA
Mc            ...FMFA
```

FIGURE 2. Predicted coded portions of tags. Alignment for proteobacteria, species name abbreviations, and indication of phylogenetic relations are as in Figure 1. Uncertainty in the N-terminal extent for Gram-positive bacteria is indicated by ellipsis. See the Materials and methods for discussion of reading frame identification.

A special feature of tmRNA secondary structure is the incorporation of more than half the nucleotides into pseudoknots. The four pseudoknots are simple, except for branching stems, which might form in sequence insertions that occur infrequently and idiosyncratically within ψ2 and ψ3 (Fig. 1A). Within each pseudoknot, the middle of the three segments that join base pairing segments appears shortest or absent most frequently throughout phylogeny (with some ambiguity for ψ1), suggesting that the other two joining segments form the pseudoknot loops and that RNA enters and exits the pseudoknots at their far ends (Fig. 1b). The three sequential pseudoknots (ψ2-ψ4) generally share other features: (1) the 5′ stem (~8–15 bp) is longer than the 3′ stem (~6 bp); (2) correspondingly, the 3′ loop (~10 nt) is longer than the 5′ loop (~1–3 nt); (3) the downstream loop is adenosine-rich; (4) 2–3-nt spacers separate each pseudoknot from the next base paired element downstream. ψ1 shares features 2 and 3, but its stems are both rather short, and the downstream loop is ~7 nt. Most commonly, the two stems of each pseudoknot abut directly, suggesting coaxial stacking (Puglisi et al., 1990). However, there are many cases where an intervening nucleotide might intercalate between the two stems, which could bend the pseudoknot axis significantly (Shen & Tinoco, 1995). Additional bending could come from the discontinuous pairing observed for many of the stems within pseudoknots. Either mode of bending within these pseudoknots could help accommodate their short loops and also provide flexibility, as might be required for contortions within the ribosome.

The secondary structure of tmRNA will guide future studies of its higher-order structure and of the special interactions that lead to its unique action. Phylogenetic searches for base pairing with ribosomal RNAs may be fruitful. Simple pseudoknots can bind proteins with high affinity and specificity (Tuerk et al., 1992; Ringquist et al., 1995), and specific interaction of a pseudoknot with ribosomal components is thought to promote −1 frameshifting or stop codon suppression on certain viral mRNAs (Gesteland & Atkins, 1996). The string of three roughly similar pseudoknots in tmRNA offers possibilities for cooperative binding.

The tmRNA is reminiscent of the 3′ ends of certain plant viral RNAs (Pleij et al., 1987); for example, the

3′ untranslated region of tobacco mosaic virus (TMV 3′UTR) consists of a string of three pseudoknots leading into a histidylatable tRNA-like domain. The TMV 3′UTR has activity equivalent to a poly(A) tail in stimulating translation from an upstream reading frame in both plant and mammalian cells, suggesting that eukaryotes might recognize (yet unidentified) cellular mRNAs with such 3′ ends (Gallie & Walbot, 1990). Details of function and morphology differ between the tmRNA and TMV 3′UTR, and the TMV 3′UTR function localizes primarily to one of its pseudoknots (Leathers et al., 1993). It is nonetheless an interesting parallel that a pseudoknot string linked to a tRNA-like structure could assist the delivery of reading frames to ribosomes in bacterial tmRNA, plant viral RNAs, and eukaryotic cellular mRNAs.

MATERIALS AND METHODS

Proteobacterial genomic DNAs were prepared from ~10 mg of each freeze-dried type specimen by lysis (20 min incubation at 65 °C in 50 μg/mL proteinase K, 10 mM TrisCl, pH7.5, 5 mM EDTA), incubation with 1% cetyltrimethylammonium bromide (10 min at 65 °C following addition of NaCl to 500 mM), extraction twice with phenol–chloroform, and precipitation with isopropanol (0.6 volumes). Sequences of tmRNA genes were obtained by PCR in 100 μL using 3% of each genomic DNA preparation and either primer set A (5′-GGGCCGACCTGGTT TCGAC and GAGCCGGCGGGAATCGAAC, based on *Alcaligenes eutrophus* tmRNA termini) or E (5′-GGGGCTGATTCT GGATTCGAC and TGGAGCTGGCGGGAGTTGAAC, based on *E. coli* tmRNA termini) for 30 cycles through 96, 46, and 72 °C with Taq DNA polymerase, product purification with QIAquick cartridges, and dye terminator sequencing with both of the primers used in PCR. The sequence from *E. coli* was also confirmed by this method. In some cases, agarose gel purification and reamplification of the initial PCR product was required to obtain high-quality sequence data. By sequencing the entire molecular ensemble of the PCR product rather than an individual clone, we avoided artifacts arising from the copying errors that can occur during PCR; because many repeats of the gene were present in the genomic DNA preparation, any one miscopying event, even occurring during early PCR cycles, should not dominate the product population. ATCC number of specimen, PCR primer set(s) used, and database accession number: 133, E, U68074 (*E. coli*); 33658, E, U68075 (*Aeromonas salmonicida*); 14393, A, U68076 (*Alteromonas haloplanktis*); 49840, E, U68077 (*Marinobacter hydrocarbonoclasticus*); 25330, E, U68078 (*Pseudomonas aeruginosa*); 33152, AE, U68079 (*L. pneumophila*); 19424, A, U68080 (*Neisseria gonorrhoeae*); 27774, E, U68081 (*Desulfovibrio desulfuricans*); 43504, AE, U68082 (*H. pylori*).

Length variations and phylogenetic gaps made *a priori* primary alignment of all available sequences impractical, but the sequences from eight γ proteobacteria were of similar length and exhibited a sufficient degree of primary conservation (top eight sequences of Fig. 1A). An initial computer alignment of these eight sequences underwent a reiterative editing process as described by James et al. (1988), in which secondary structural features were identified by base covariations and the alignment was refined manually, considering both primary and secondary structure, with an effort to avoid gaps. In accordance with the criteria used for rRNA (Noller et al., 1981) and ribonuclease P RNA (James et al., 1988), observation of at least two Watson–Crick covariations within a potential continuous helix was considered proof of the existence of that helix. Established helices were extended across bulges or short internal loops using less stringent criteria (Noller et al., 1981). This allowed the identification of helices P2a, P2c, P4, P6, P7, P9, P10, and P11. No new sequence data were obtained regarding P1 or P12, but they were established previously by covariation among sequences present in the database (Komine et al., 1994).

The nine sequences from more distantly related species were aligned partially with the eight-sequence alignment based on identifications of: (1) short regions of primary sequence homology; (2) the helices found in the eight-species alignment; and (3) the stop codon of the tag reading frame. The new alignment provided the covariations necessary to identify the P2b and P3 helices (which lie within segments anchored by P2c) and the P8 helix (which is anchored by P9). In assigning stop codons, the hydrophobic nature of the predicted C-terminal residues was considered, because this is the defining characteristic of the proteolysis tag in *E. coli* (Parsell et al., 1990; Keiler & Sauer, 1996). A highly conserved 17-nt region, including the known resume codon of *E. coli*, permitted prediction of the resume codons for all the proteobacteria. Because this conservation did not extend to the Gram-positive bacteria, their resume codons remain unassigned.

ACKNOWLEDGMENTS

We thank L. Ziaugra for skilled assistance with sequencing, and J. Theriot for supervising procedures with pathogenic bacteria.

Received September 11, 1996; returned for revision October 17, 1996; revised manuscript received October 21, 1996

REFERENCES

de Smit MH, van Duin J. 1994. Control of translation by mRNA secondary structure in *Escherichia coli*. *J Mol Biol* 244:144–150.

Felden B, Atkins JF, Gesteland RF. 1996. tRNA and mRNA both in the same molecule. *Nature Struct Biol* 3:494.

Gallie DR, Walbot V. 1990. RNA pseudoknot domain of tobacco mosaic virus can functionally substitute for a poly(A) tail in plant and animal cells. *Genes & Dev* 4:1149–1157.

Gesteland RF, Atkins JF. 1996. Recoding: Dynamic reprogramming of translation. *Annu Rev Biochem* 65:741–768.

Gutell RR, Larsen N, Woese CR. 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol Rev* 58:10–26.

James BD, Olsen GJ, Liu JS, Pace NR. 1988. The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme. *Cell* 52:19–26.

Keiler KC, Sauer RT. 1996. Sequence determinants of C-terminal substrate recognition by the Tsp protease. *J Biol Chem* 271:2589–2593.

Keiler KC, Waller PRH, Sauer RT. 1996. Role of a peptide tagging system in degradation of protein synthesized from damaged messenger RNA. *Science* 271:990–993.

Komine Y, Kitabatake M, Yokogawa T, Nishikawa K, Inokuchi H. 1994. A tRNA-like structure is present in 10Sa RNA, a small stable RNA from *Escherichia coli*. *Proc Natl Acad Sci USA* 91:9223–9227.

Leathers V, Tanguay R, Kobayashi M, Gallie DR. 1993. A phylogenetically conserved sequence within viral 3′ untranslated RNA pseudoknots regulates translation. *Mol Cell Biol 13*:5331–5347.

Nolan JM, Pace NR. 1996. Structural analysis of the bacterial ribonuclease P RNA. In: Eckstein F, Lilley DMJ, eds. *Catalytic RNA*. Berlin: Springer-Verlag. pp 109–128.

Noller HF, Kop J, Wheaton V, Brosius J, Gutell RR, Kopylov AM, Dohme F, Herr W, Stahl DA, Gupta R, Woese CR. 1981. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res 9*: 6167–6189.

Parsell DA, Silber KR, Sauer RT. 1990. Carboxy-terminal determinants of intracellular protein degradation. *Genes & Dev 4*:277–286.

Pleij CWA, Abrahams JP, van Belkum A, Reitveld K, Bosch L. 1987. The spatial folding of the 3′ noncoding region of aminoacylatable plant viral RNAs. In: Brinton MA, Rueckert RR, eds. *Positive strand viruses*. New York: Alan R. Liss, Inc. pp 299–316.

Ringquist S, Jones T, Snyder EE, Gibson T, Boni I, Gold L. 1995. High-affinity RNA ligands to *Escherichia coli* ribosomes and ribosomal protein S1: Comparison of natural and unnatural binding sites. *Biochemistry 34*:3640–3648.

Shen LX, Tinoco I Jr. 1995. The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *J Mol Biol 247*:963–978.

Tu GF, Reid GE, Zhang JG, Moritz RL, Simpson RJ. 1995. C-terminal extension of truncated recombinant proteins in *Escherichia coli* with a 10Sa RNA decapeptide. *J Biol Chem 270*:9322–9326.

Tuerk C, MacDougal S, Gold L. 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc Natl Acad Sci USA 89*:6988–6992.

Ushida C, Himeno H, Watanabe T, Muto A. 1994. tRNA-like structures in 10Sa RNAs of *Mycoplasma capricolum* and *Bacillus subtilis*. *Nucleic Acids Res 22*:3392–3396.