Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 3001–3006, April 1996
Biochemistry

Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements

JAMES W. BROWN*, JAMES M. NOLAN†‡, ELIZABETH S. HAAS*, MARY ANNE T. RUBIO†, FRANCOIS MAJOR§, AND NORMAN R. PACE†¶

*Department of Microbiology, North Carolina State University, Raleigh, NC 27695; †Department of Biology and Institute for Molecular and Cellular Biology, Indiana University, Bloomington, IN 47405; and §Departement d'Informatique et de Recherche Operationnelle, Universite de Montreal, and Montreal Joint Center for Structural Biology, Montreal, QC Canada H3C 3J7

Contributed by Norman R. Pace, Indiana University, Bloomington, IN, August 17, 1995

ABSTRACT **PCR amplification of template DNAs extracted from mixed, naturally occurring microbial populations, using oligonucleotide primers complementary to highly conserved sequences, was used to obtain a large collection of diverse RNase P RNA-encoding genes. An alignment of these sequences was used in a comparative analysis of RNase P RNA secondary and tertiary structure. The new sequences confirm the secondary structure model based on sequences from cultivated organisms (with minor alterations in helices P12 and P18), providing additional support for nearly every base pair. Analysis of sequence covariation using the entire RNase P RNA data set reveals elements of tertiary structure in the RNA; the third nucleotides (underlined) of the GNRA tetraloops L14 and L18 are seen to interact with adjacent Watson–Crick base pairs in helix P8, forming A:G/C or G:A/U base triples. These experiments demonstrate one way in which the enormous diversity of natural microbial populations can be used to elucidate molecular structure through comparative analysis.**

Phylogenetic-comparative sequence analysis has proven to be the most generally useful approach for the determination of the higher-order structures of large RNAs (3, 5). In the case of the catalytic RNA subunit of RNase P, a tRNA-processing endonuclease, sequence comparisons have resulted in the formulation of a secondary structure model that engages >60% of the ≈400 nt of the ribozyme in base pairs (for review, see ref. 2). Preliminary models of the tertiary structure of RNase P RNA have been assembled based on the secondary structure and other comparative and biochemical data (6, 7).

The data set used in a phylogenetic-comparative analysis of RNA structure is a collection of differing, but homologous, sequences; covariations in the sequences indicate bases that interact specifically in some way. Typically, the homologous sequences that make up the data set are obtained individually, from pure cultures of selected organisms. Detailed comparative analysis of a large RNA requires, however, a large collection of sequences, hundreds, to detect changes that occur only rarely. The accumulation of many sequences, one-at-a-time from cultivated organisms, becomes a limiting step in the analysis. To facilitate the acquisition of RNase P RNA sequences, we have developed an approach that uses naturally occurring microbial populations as sources of genes. Generally the number of different sequences available for analysis is more important than is knowledge of the specific source of those sequences. Natural populations of microorganisms are highly complex and diverse. For instance, DNA complexity analyses have indicated that woodland soil samples typically contain many thousands of different species (8). The approach we use is based on the polymerase chain reaction (PCR), using oligodeoxynucleotide primers complementary to highly conserved sequences near the ends of known RNase P RNAs and template DNAs purified from natural ecosystems. This laboratory has used (7) similar methods to obtain rRNA genes from the environment.

MATERIALS AND METHODS

Natural Populations Sequences. DNAs were isolated as described (9) from biomass filtered from Indiana University Department of Biology greenhouse pond water, near-shore sediment from Lake Griffy (Bloomington, IN), and "pink filaments" in the 83°C outflow of Octopus Spring (Yellowstone National Park, WY). These community DNAs were used as templates in PCR amplifications with degenerate oligonucleotide primers complementary to highly conserved sequences located near the 5′ and 3′ ends of bacterial RNase P RNA-encoding genes (Fig. 1). Some new RNase P RNA gene sequences arose as contaminants ("volunteer" sequences) in PCRs using known template DNAs. Although of unknown origin, they are authentic RNase P RNAs based on similarity to known RNase P RNAs and proved useful in the structure analysis. PCRs were performed and product DNAs were cloned essentially as described (2). Fragments containing ≈70% of each of the RNase P RNA-encoding genes were amplified by using oligonucleotide primers 59FBam (5′-CGGGATCCGIIGAGGAAAGTCCIIGC-3′; I = inosine) and 347REco (5′-CGGAATTCRTAAGCCGGRTTCTGT-3′; R = A or G) and separated by preparative electrophoresis in 3% agarose gels (NuSieve GTG, FMC BioProducts) after digestion with restriction endonucleases *Eco*RI and *Bam*HI. The diffuse band corresponding to DNA amplification products of ≈300 bp was excised from the gel, ligated into *Eco*RI/*Bam*HI-digested pBluescript KS$^+$ DNA (Stratagene, Inc.), and transformed into *Escherichia coli* DH5αF′. Double-stranded plasmid DNAs were sequenced by the dideoxynucleotide chain-termination method using Sequenase version 2.0 (United States Biochemicals) (10). Clones containing unique RNase P RNA sequences based on sequence data from a single primer were completely sequenced on both strands using M13 universal and reverse primers, 59FBam, 347REco, 174F (5′-AGGGTGAAANGGTGSGGTAAGAG-3′; N = A,

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: To denote interactions between bases we use a slash for canonical base pairs (e.g., G/C and A/U), a dot for noncanonical interactions (e.g., G·U and A·G), and a colon for a triple interaction between a base and a base pair (e.g., A:G/C). Nomenclature of structural elements in RNase P RNA corresponds to the group I intron convention (1), as described (2): P (paired) refers to a helix, numbered according to encounter from the 5′ end; L refers to the loop of particular helices; and J (joining) refers to the nucleotide stretch between particular helices (e.g., J5/6 connects P5 and P6).

‡Present address: Department of Biochemistry, Tulane University School of Medicine, New Orleans, LA 70112.
¶To whom reprint requests should be addressed.

‖The sequences reported in this paper have been deposited in the GenBank data base (accession nos. U28079–U28130).

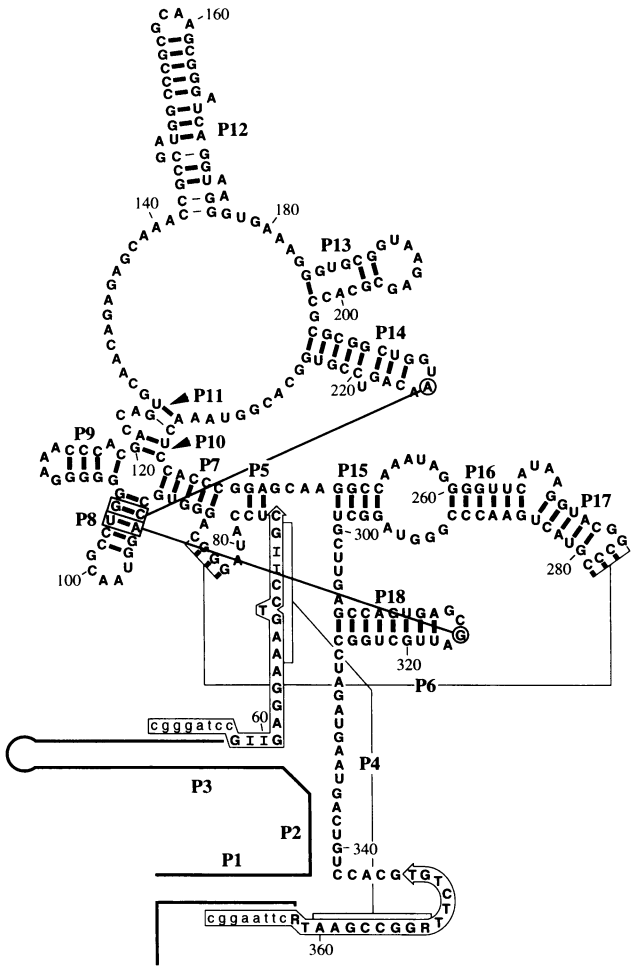

FIG. 1. Comparative analysis of RNase P RNA secondary structure using sequences from mixed naturally occurring microbial populations. The *E. coli* RNase P RNA secondary structure is shown with evidence provided by mutual-information analysis of an alignment containing the 48 natural-population and "volunteer" RNase P RNA sequences obtained in this study and those of previously determined RNase P RNAs. Base pairings marked with thick lines are supported by mutual-information coefficients ($M\{x,y\}$) for a given base that are highest for its pairing partner than any other base; base pairing indicated by thin lines is between highly conserved bases and is supported by individual instances of sequence covariation. Lines connecting circled bases with boxed base pairs indicate base-triples 214:93/105 and 316:94/104 (see text). The amplification primer sequences 59FBam and 347REco (note that the latter is the complement of the sequence present in the RNA) are boxed, with arrowheads indicating the polarity of the primers and nucleotides in lowercase type indicting linker sequences used for cloning. Sequences including and distal to the amplification primers are not obtained by PCR using the specified primers and are indicated by lines based on known structures. The nt 304–305 and 326–327, at the base of P18, which were paired in previous secondary structure models, are shown unpaired; these pairings are inconsistent with the newly determined sequences (text). The medial region of P12 has been rearranged relative to previous secondary structure models, to comply with sequence covariation (see text).

G, C, or T and S = G or C), and 174R (5′-CTCTTACCSCAC-CNTTTCACCCT-3′) oligonucleotide primers. The sequences, alignments, and predicted secondary structures of these RNAs are available from GenBank‖ and the Ribonuclease P Database (11). One sequence that appears to be a PCR-generated chimera (12) and two sequences that were not recognizably related to RNase P RNAs were obtained; these sequences were excluded from the analysis.

Comparative Analysis and Molecular Modeling. Sequences were aligned manually by using conserved structural landmarks as described (13). Phylogenetic trees based on unambiguously homologous nucleotides in the conserved core of the RNA structure were generated by the algorithm of DeSoete (14), using the GDE sequence editor (15). Sequence covariation was analyzed by manual inspection (16, 17) and by the mutual-information algorithm developed by Chiu and Kolodziejczak (18), as implemented by Gutell and coworkers (19), using COVARIATION version 4.0 (available from the Ribonuclease P Database). Alignments for mutual-information analysis contained all available bacterial RNase P RNA sequences except those from the log G+C Gram-positive bacteria. The three-dimensional model of the junction L14–P8–L18 was constructed by using the RNA computer modeling program MC-SYM (20) with measured parameters (21–24).

RESULTS AND DISCUSSION

Based on the collection of RNase P RNA sequences from cultivated organisms, we could identify highly conserved sequences that would be suitable targets for generally applicable PCR primers. Two such primers, shown in Fig. 1, were used in PCR with DNA purified from arbitrarily chosen environmental samples. A total of 52 novel RNase P RNA genes were obtained by PCR amplification and cloning, and the sequences were determined. The new sequences more than double the bacterial RNase P RNA sequence collection available for comparative analysis (11). The phylogenetically conserved sequence and secondary structural core of the RNA is readily apparent in the new sequences; all are of the "ancestral" type exemplified by the *E. coli* version (25). Consequently, the alignment of the sequences was generally straightforward. The degree of variation among the natural-population sequences is similar to that of cultivated species in terms of sequence conservation, and the location and extent of sequence-length variation. The sequences and secondary structure drawings for these and other RNase P RNAs are available electronically at http://JWBrown.mbio.ncsu.edu/RNaseP/. Phylogenetic trees, as well as sequence and structural signatures, were used to determine the phylogenetic affiliations of the new sequences with one another and with sequences from known organisms. Most of the sequences, but not all, could be associated with particular phylogenetic groups on the basis of similarity to previously determined sequences (Table 1). We did not exhaust the diversity in any of these environmental samplings.

Comparative Analysis of Secondary Structure. The new RNase P RNA sequences were scrutinized for covariation of nucleotides to test and refine the structure model. Because the amplified sequences are incomplete, only interactions between nucleotides in the amplified region of the gene (*E. coli* nt 76–346) can be identified in this analysis. This region corre-

Table 1. RNase P RNA sequences obtained from natural microbial populations

Phylogenetic affiliation	DNA source			
	Pond water	Lake sediment	Yellowstone	Volunteer
α-proteobacteria	7	3	—	1
β-proteobacteria	—	—	—	4
γ-proteobacteria	—	1	3	5
Proteobacteria*	2	4	—	2
Cyanobacteria	1	—	—	1
Bacteroids†	1	2	—	2
Planctomycetes	—	4	—	—
Gram-positive	—	—	—	1
Unknown‡	—	3	3	2

*Specific affiliation within the proteobacteria is uncertain.
†*Bacteroides, Flavobacteria*, and relatives.
‡Phylogenetic affiliation within the Bacteria is uncertain.

sponds, however, to the most poorly defined part of the RNA. The few elements of the RNA structure that are excluded by the selection of primer sites are structurally well defined by previous analyses.

It was useful, considering the large number of sequences now available, to assess covariation between bases employing mutual information analysis with the algorithm developed by Chiu and Kolodziejczak (18) and tested by R. Gutell and colleagues (9) (Fig. 2). The mutual-information coefficient, $M(x,y)$, is the sum of the variation at each position minus the degree of variation of the bases taken together; $M(x,y) = H(x) + H(y) - H(x,y)$. Variability of a sequence position, or pair of positions taken together, is described by the entropy term $H = \Sigma_b f_b \ln b$ [where $b \in$ (A,G,U,C,—)]. $M(x,y)$ is greatest when the sequence positions being compared are both highly variable and directly correlated in that variation; i.e, where $H(x)$ and $H(y)$ are large and $H(x,y)$ is small.

As anticipated, the strongest mutual-information correlations correspond to established base pairs in the structure model. In almost every case, the mutual-information coefficient with the highest value for a given sequence position corresponds to its pairing partner, where one exists, in the secondary structure model (2). In a few cases, correlations between paired bases are sufficiently low because of conservation that they are exceeded by local (e.g., nearest neighbor) correlations. Of the 78 bp in the secondary structure model of the sequence span covered by the study, 72 (92%) are confirmed by phylogenetic covariation among the natural-population sequences. Only 6 of the individual base pairs in the model structure are not further supported by the new data set. Four of these unsupported pairings, 3 in P13 and 1 in P5, are not supported due to extreme conservation of the nucleotides; comparative analysis cannot reveal structure in the absence of variation. In two cases, however, the evidence indicates clearly that previously proposed pairings are not likely to occur. The nt 304 and 305, at the proximal end of P18, are much more conserved ($H\{x\}$ = 0.63 and 0.27, respectively) than their previously proposed pairing partners, nt 327 and 326 ($H\{x\}$ = 1.23 and 0.85, respectively), and the identities of the bases are not significantly correlated with their putative pairing partners. These previously proposed pairs were only weakly supported in simpler analyses of much smaller data sets (in one case as a non-Watson–Crick base pair) and now must be considered disproven. These nucleotides are, however, structurally affiliated with P18; RNase P RNAs that lack P18 also lack them, replacing the entire unit with a single nucleotide (2).

The newly obtained sequences also support a minor rearrangement of the medial region of helix P12 in the γ-proteobacterial RNAs (e.g., that of *E. coli*). The previously available sequence collection could not distinguish between two alternative structures: one proposed in previous models, with nt 147–149 paired with nt 166–168, and another in which nt 149–151 are paired with nt 167–169. The latter is clearly supported by sequences from the γ-group of proteobacteria obtained in this study.

RNase P RNA sequences sometimes contain non-Watson–Crick base pairs in helices that are otherwise composed of canonical complements; most often (except for G·U pairs) these are G·A pairs, either alone or adjacent to one another (nearly always GA/GA). These G·A pairs covary with standard Watson–Crick or G·U pairs and appear most frequently in helices that are the most highly variable in sequence (P12, P14, P17, and P18). Similar covariation of GA/GA pairs with Watson–Crick pairs has been observed in small-subunit rRNA

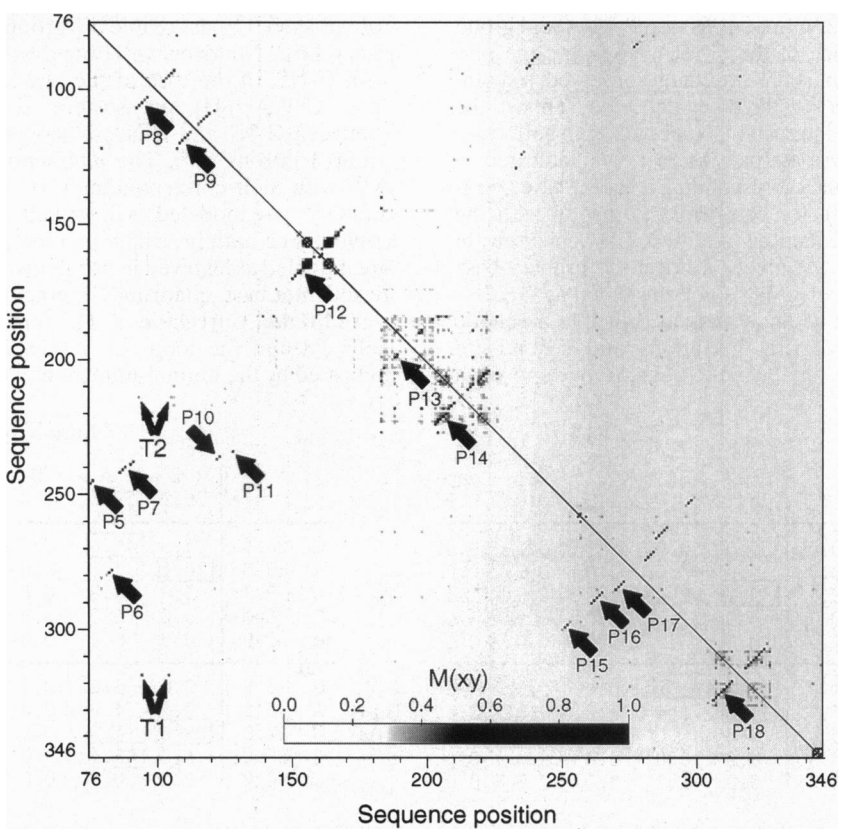

FIG. 2. Mutual information analysis of secondary and tertiary structure. Sequence positions are numbered according to the *E. coli* RNase P RNA on both x and y axes. Pixels at the intersection of two sequence positions (vertical and horizontal) represent nucleotide pairs; the mutual-information coefficient $M(x,y)$ for each pair of bases defines the pixel intensity according to the intensity bar. Correlations that define helices in the secondary structure are numbered P5–P18. Correlations of base-paired nucleotides in P8 with L14 and L18 (i.e., base triples) are indicated by T1 and T2, respectively.

sequences (5). The structure of $(rGGCGAGCC)_2$ derived by NMR indicates that such pairings can occur and lend structural stability in the context of a normal helix (26).

GNRA-Tetraloop:Helix Tertiary Interactions. A few of the mutual-information correlations observed between sequence positions indicate previously unknown base-specific tertiary structure. Two such correlations (Fig. 3). involve nt 214 with the 93/105 bp and nt 316 with the 94/104 bp (Fig. 1). In both of these cases, the third-position purine of a conserved GNRA tetraloop covaries with nucleotides that form Watson–Crick base pairs in helix P8 of the secondary structure. The covariation is particularly striking in the context of the predicted phylogenetic relationships between the sequences; the A:G/C and G:A/U alternatives are phylogenetically dispersed; i.e., the three bases have changed frequently, as a set, among related sequences. The few exceptions to the correlations (including those that lack either P14 or P18) (2, 4) are phylogenetically clustered and likely represent only a small number of evolutionary events (i.e., they are synaptomorphic).

The straightforward interpretation of these mutual-information correlations is that the covarying nucleotides interact directly to form base triples, in which A:G/C and G:A/U are acceptable alternatives. Covariation of this type has been observed previously in group I intron sequences (27) and shown experimentally to indicate direct interactions (28, 29). Isosteric structures have been suggested for A:G/C and G:A/U triples in which the loop purine is hydrogen-bonded to the purine of the Watson–Crick base pair in the minor groove of the helix (29). Consistent with this hypothesis is the observation that A:G·U sets are present for these bases in some RNase P RNAs, but G:G·U does not occur; the tetraloop purine covaries more strongly with the purine position of the base pair than with the pyrimidine position. It has been postulated that an additional base triple might be formed by the adjacent adenine of the GNRA loop and the purine of the 3′-neighboring base pair (27). In both RNase P RNA and group I intron RNAs, the adenine of the GNRA loops and the presumptive base-paired partners are extremely conserved, so comparative analysis provides no direct evidence to support the presence of this additional interaction. Nonetheless, in both types of RNA, where the GNRA:base-pair interaction is indicated by phylogenetic covariation, the corresponding adjacent base pair is conserved and appropriate for base-triple formation with the loop adenine of GNRA sequences (i.e., A:G/C). Conversely, in RNase P RNAs that lack the ability to form the primary base triple, the adjacent base pair in P8 is not conserved as G/C.

Phylogenetic-comparative analysis is in principle a genetic analysis of naturally occurring mutations and second-site intragenic reversions. The phylogenetic-comparative approach can be more sensitive than *in vitro* genetic tests, however, because of the pressure of biological selection. Replacement of L14 GUAA with GUGA, L18 GCGA with GCAA, P8 bp C^{93}/G^{105} with U/A, and/or A^{94}/U^{104} with G/C resulted in no detectable change in behavior in the *in vitro*, RNA-only assay (data not shown). Thus, the tertiary interactions identified by comparative analysis would not have been detected by *in vitro* mutational analysis. Neither of the two models for the global tertiary structure of *E. coli* RNase P RNA (6, 7) predicted the tertiary interactions described here. Helices P14 and P18 both contribute to the global folding stabilities of RNase P RNAs in which they occur (30). Presumably, interactions in addition to the sites of mutation maintain the association of the two helices with the rest of the RNA. The base-triple interactions proposed from the correlation analysis are additionally supported by photoaffinity crosslinking results (M. E. Harris and N.R.P., unpublished data). Moreover, nucleotides in L14 and L18 are resistant to the chemical agents kethoxal and dimethyl sulfate (31), indicating their engagement in structural interactions. Finally, the lengths of P14 and P18 are phylogenetically conserved, consistent with the notion that both ends of these helices interact elsewhere in conserved structure.

Three-Dimensional Interpretation. The region of RNase P RNA containing the L14–P8–L18 tertiary interaction is now sufficiently well-constrained by known interactions to develop an atomic-level model of this domain using the MC-SYM RNA modeling program (20) (Fig. 4). In this model, which is consistent with available comparative, NMR and crystallographic data for GNRA tetraloops (21, 32), helices P14 and P18 approach from opposite directions and are aligned coaxially such that their loop nucleotides interact in opposite orientations with the base-paired purines, in the minor groove of P8. The interaction of the varying A^{214}, which forms a base triple with the G^{105}/C base pair of P8, is modeled as proposed for an A:G/C base triple in group I introns (27, 29): the exocyclic A-N6 forms an H bond with G-N3, and A-N1 pairs with G-N2. In the case of the alternative base triple G:A/U (e.g., G^{316}:A^{94}/U), an isosteric single H-bond interaction (between G-N1 and A-N3) is modeled, also as proposed for group I introns (29). The interactions of invariant A^{215} and A^{317} with their corresponding G/C base pairs (involving G^{95} and G^{106}) are modeled as described above for the A:G/C base triple. Since both invariant adenines in the GNRA loops also are modeled as involved in intraloop A·G pairs, this association results in a base-quadruple interaction.

Additional Correlations. The tertiary interactions between helix P8 and the loops of helices P14 and P18 are clearly indicated by the mutual-information correlations based on the

Base 214

Base 93		%	G 38.1	A 54.8	U 1.2	C 0.0	gap 5.9
	G	0.0	0.0	0.0	0.0	0.0	0.0
	A	0.0	0.0	0.0	0.0	0.0	0.0
	U	38.1	**32.1**	2.4	1.2	0.0	2.4
	C	60.7	5.9	**51.2**	0.0	0.0	3.6
	gap	1.2	0.0	1.2	0.0	0.0	0.0
Base 105							
	G	61.9	7.1	**51.2**	0.0	0.0	3.6
	A	36.9	**31.0**	2.4	1.2	0.0	2.4
	U	0.0	0.0	0.0	0.0	0.0	0.0
	C	0.0	0.0	0.0	0.0	0.0	0.0
	gap	1.2	0.0	1.2	0.0	0.0	0.0

Base 316

Base 94		%	G 36.9	A 58.3	U 0.0	C 0.0	gap 4.8
	G	60.0	4.8	**46.4**	0.0	0.0	4.8
	A	29.8	**28.6**	1.2	0.0	0.0	0.0
	U	8.3	3.6	4.8	0.0	0.0	0.0
	C	6.0	0.0	6.0	0.0	0.0	0.0
	gap	0.0	0.0	0.0	0.0	0.0	0.0
Base 104							
	G	6.0	0.0	6.0	0.0	0.0	0.0
	A	8.3	3.6	4.8	0.0	0.0	0.0
	U	29.8	**28.6**	1.2	0.0	0.0	0.0
	C	56.0	4.8	**46.4**	0.0	0.0	4.8
	gap	0.0	0.0	0.0	0.0	0.0	0.0

FIG. 3. Covariation of the third nucleotides of GNRA tetraloops L14 and L18 with adjacent base pairs in helix P8. Position 214 (the third position in the L14 tetraloop) and position 316 (the third position in the L18 tetraloop) covary with the third (93/105) and second (94/104) base pairs in P8. The frequencies (as percentages) of each base at each position are indicated at the top and right of each table; the frequencies of each pair of bases are shown within the table. The identities of bases 214 and 316 are strongly correlated with bp 93/105 and 94/104, respectively. Boxes indicate the evolutionarily "preferred" sets of bases; e.g., G^{214} with bp U^{93}/A^{105} or A^{214} with bp C^{93}/G^{105}. These covariations correspond to $M(x,y)$ values of 0.36 and 0.39 for position 214 with positions 93 and 105, respectively, and 0.42 for position 316 with both positions 94 and 104.

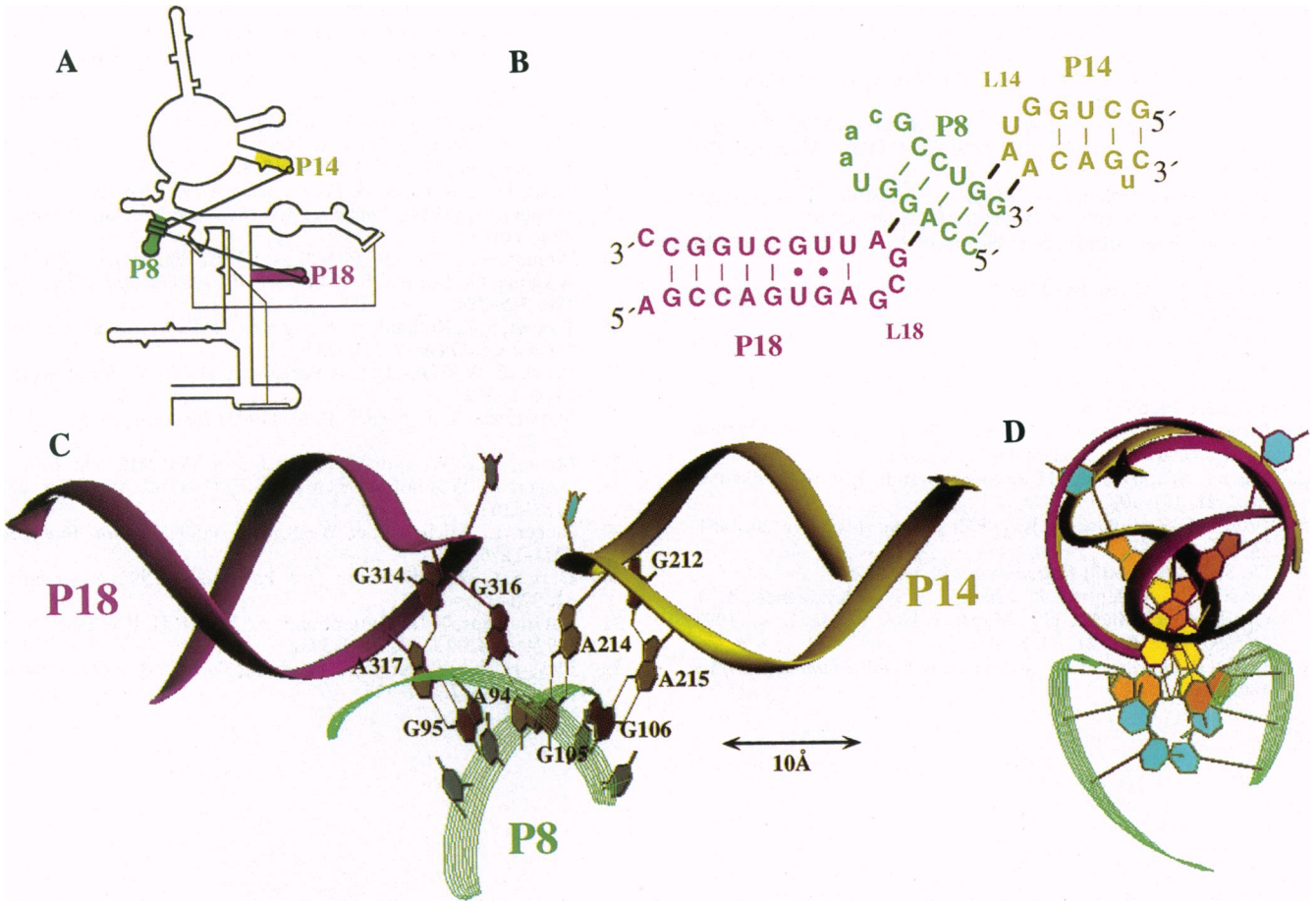

FIG. 4. Proposed base triples formed by the interaction of GNRA tetraloops with Watson–Crick base pairs. The secondary structure of *E. coli* RNase P RNA with the tertiary interactions between the loops of P14 (L14) and P18 (L18) and base pairs in helix P8 identified in this study are shown in *A*. These elements of secondary and tertiary structure are shown in isolation in *B*; the phylogenetically supported tertiary interactions are indicated by heavy lines. Potential additional base triples that because of invariance are not addressed by phylogenetic covariation in RNase P RNA but are implied in other GNRA:helix interactions (27, 29) are also indicated. A three-dimensional model of these elements constructed by using MC-SYM is shown in *C*. The non-Watson–Crick components of the base triples are modeled as associated with the purine of the base pair in the minor groove of the helix (text). Inferred hydrogen bonds are indicated by thin lines. The 3′ ends of P14 and P18 are indicated by arrowheads. An axial view of the structure is shown in *D* to illustrate the coaxial arrangement of P14 and P18.

currently available data set. Other correlations that are weaker, less well-supported phylogenetically, or less interpretable physically also may signal tertiary structure in the RNA. Their validation, however, will require a larger collection of sequences or additional experimental data. These correlations include the structure of the base of P12 and the occurrence of P13 and P14; bp 211/216 with bp 107/119; bases 280 and 281 with bases 81 and 80, respectively; and base 183 with both bases 137 and 140. A few other correlations probably represent local structural effects and synaptomorphies.

Conclusion. For a number of reasons, interactions of tertiary structure are more difficult to identify by comparative analysis than are the base pairs responsible for secondary structure. One reason for this difficulty is that tertiary structure often does not follow the simple rules of secondary structure, the canonical base pairings. Additionally, the occurrences of the base triples in bacterial RNase P RNA, and the other potential tertiary interactions discussed above, are less stringently maintained phylogenetically than are the secondary structural interactions in the core of the RNA. This is also true for similar base triples in group I intron RNAs (27) and, generally, in known tertiary interactions in transfer RNAs and small-subunit rRNAs (5). This variability in tertiary structural elements possibly reflects the dominant role of secondary structure in RNA folding. If tertiary contacts occur, however, the bases involved are generally more conserved than those involved only in secondary structure. Perhaps the pathways leading to "covariation" in tertiary structure are more constrained (fewer permissible intermediates) or complex (substitution of three or more bases) than those resulting in covariation in secondary structure.

Because of the idiosyncratic properties of RNA tertiary structure, interactions that are revealed by sequence comparisons usually are identified in the context of a well-developed model of secondary structure, through the analysis of large sequence data sets. Typically, the accumulation of large sequence collections has been rate-limiting in a comparative analysis. The approach described here—the use of complex natural populations as sources of structural diversity—is a way of rapidly acquiring large sets of homologous sequences.

We thank Dr. Bernadette Pace for the gift of *Thermus aquaticus* DNA polymerase, Drs. Gene Wickham and Sue Barns for biomass samples from Yellowstone National Park hot springs, and Drs. Robin Gutell and David Engelke for useful discussions. This work was supported by grants from the National Institutes of Health and Department of Energy to N.R.P. and a North Carolina Agricultural Research Service Grant to J.W.B.

1. Burke, J. M., Belfort, M., Cech, T. R., Davies, R. W., Schweyen, R. J., Shub, D. A., Szostak, J. W. & Tabak, H. F. (1987) *Nucleic Acids Res.* **15**, 7217–7221.

2. Haas, E. S., Brown, J. W., Pitulle, C. & Pace, N. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2527–2531.
3. Woese, C. R. & Pace, N. R. (1993) in *The RNA World*, eds. Gesteland, R. F. & Atkins, J. F. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 91–117.
4. Pace, N. R. & Brown, J. W. (1995) *J. Bacteriol.* **177,** 1919–1928.
5. Gutell, R. R., Larsen, N. & Woese, C. R. (1994) *Microbiol. Rev.* **58,** 10–26.
6. Harris, M. E., Nolan, J. M., Malhotra, A., Brown, J. W., Harvey, S. C. & Pace, N. R. (1994) *EMBO J.* **13,** 3953–3963.
7. Westhof, E. & Altman, S. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 5133–5137.
8. Torsvik, V., Goksoyer, J. & Daae, F. L. (1990) *Appl. Environ. Microbiol.* **56,** 782–787.
9. Barns, S. M., Fundyga, R. E., Jeffries, M. W. & Pace, N. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1609–1613.
10. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
11. Brown, J. W., Haas, E. S., Gilbert, D. G. & Pace, N. R. (1994) *Nucleic Acids Res.* **22,** 3660–3662.
12. Liesack, W., Weyland, H. & Stackebrandt, E. (1991) *Microbial Ecol.* **21,** 191–198.
13. Brown, J. W. & Pace, N. R. (1992) *Nucleic Acids Res.* **20,** 1451–1456.
14. De Soete, G. (1983) *Psycometrica* **48,** 621–626.
15. Larsen, N., Olsen, G. J., Maidak, B. L., McCaughey, M. J., Overbeek, R., Macke, T. J., Marsh, T. L. & Woese, C. R. (1993) *Nucleic Acids Res.* **21,** 3021–3023.
16. James, B. D., Olsen, G. J. & Pace, N. R. (1989) *Methods Enzymol.* **180,** 227–239.
17. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. (1985) *Prog. Nucleic Acids Res.* and *Mol. Biol.* **32,** 155–215.
18. Chiu, D. K. & Kolodziejczak, T. (1991) *Comput. Appl. Biosci.* **7,** 347–352.
19. Gutell, R. R., Power, A., Hertz, G. Z., Putz, E. J. & Stormo, G. D. (1992) *Nucleic Acids Res.* **20,** 5785–5795.
20. Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E. & Cedergren, R. (1991) *Science* **253,** 1255–1260.
21. Heus, H. A. & Pardi, A. (1991) *Science* **253,** 191–193.
22. Saenger, W. (1984) *Principles of Nucleic Acid Structure* (Springer, New York).
23. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984) *J. Am. Chem. Soc.* **106,** 765–784.
24. Weiner, S. J., Kollman, P. A., Nguyen, D. T. & Case, D. A. (1986) *J. Comput. Chem.* **7,** 230–252.
25. Brown, J. W., Haas, E. S. & Pace, N. R. (1993) *Nucleic Acids Res.* **21,** 671–679.
26. SantaLucia, J. & Turner, D. H. (1993) *Biochemistry* **32,** 12612–12623.
27. Michel, F. & Westhof, E. (1990) *J. Mol. Biol.* **216,** 585–610.
28. Jaeger, L., Westhof, E. & Michel, F. (1991) *J. Mol. Biol.* **221,** 1153–1164.
29. Jaeger, L., Michel, F. & Westhof, E. (1994) *J. Mol. Biol.* **236,** 1271–1276.
30. Darr, S. C., Zito, K., Smith, D. & Pace, N. R. (1992) *Biochemistry* **31,** 328–333.
31. LaGrandeur, T. E., Hüttenhofer, A., Noller, H. F. & Pace, N. R. (1994) *EMBO J.* **13,** 3945–3952.
32. Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994) *Nature (London)* **372,** 111–113.