Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 8467–8471, November 1990
Biochemistry

Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops"

(hairpin loops/comparative analysis/UUCG/CUUG/GCAA)

C. R. WOESE*†, S. WINKER‡, AND R. R. GUTELL*§

*Department of Microbiology, 123 Burrill Hall, University of Illinois, Urbana, IL 61801; ‡Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439; and §Cangene Corporation, 3403 American Drive, Mississauga, ONT L4V 1T4, Canada

Contributed by C. R. Woese, August 6, 1990

ABSTRACT **The four-base loops that cap many double-helical structures in rRNA (the so-called "tetra-loops") exhibit highly invariant to highly variable sequences depending upon their location in the molecule. However, in the vast majority of these cases the sequence of a tetra-loop is independent of its location and conforms to one of three general motifs, GNRA, UNCG, and (more rarely) CUUG. For the most frequently varying of the 16S rRNA tetra-loops, that at position 83 (*Escherichia coli* numbering), the three sequences CUUG, UUCG, and GCAA account for almost all examples encountered, and each of them has independently arisen at least a dozen times. The closing base pair of tetra-loop hairpins reflects the loop sequence, tending to be C·G for UUCG loops and G·C for CUUG loops.**

The prediction of RNA structure from simple principles (e.g., base stacking energies) is an inexact art. Existing methods (1, 2) work acceptably well with simple molecules such as tRNAs, but with large molecules such as the rRNAs their utility is at best limited. However, higher-order structure for large RNAs can readily be inferred by the simple empirical approach of comparative (sequence) analysis, and the detailed secondary structures that now exist for the small- and large-subunit rRNAs attest to the approach's effectiveness (3–6).

Comparative analysis of sequences is obviously not confined to identification of standard secondary structure *per se*. The method in principle can detect any sequence constraints (for which compositional variants are known); it has been used to elucidate some of the "tertiary" interactions in rRNAs (6–10), as well as to define the irregularities, such as "bulged" nucleotides, in secondary structural elements. It also serves effectively as the basis for designing directed mutagenesis experiments that allow structure to be inferred by assessing the functional consequences of changes therein, and it serves as an effective guide to the physical chemist who would determine nucleic acid structure. In the present communication we use comparative analysis to define the constraints on the sequence of the simplest helical structures in rRNAs, the so-called "tetra-loops" (double-stranded stalks capped by a loop of four nucleotides).

Although the finding was never formally published, comparative analysis long ago revealed that the tetra-loops in rRNA are highly constrained in sequence, the vast majority of cases being covered by a very small number of motifs, such as CUUG, UUCG, or GCAA (C.R.W., unpublished lecture¶ and cited in ref. 11). In addition, Tuerk *et al.* (11) have found (C)UUCG(G) tetra-loops to be particularly stable. The collection of small-subunit rRNA sequences is now large enough—i.e., in the range of 500—that the constraints governing the sequences of tetra-loops in this molecule can be defined in some detail. The smaller collection of 23S rRNA sequences is nevertheless large enough to assess the generality of any constraints derived from analysis of 16S rRNA.

Fig. 1 shows a representative (eu)bacterial 16S rRNA secondary structure, that of *Escherichia coli*. Tetra-loops account for about 55% (i.e., 17) of all hairpin loops in this structure, the next most prevalent loop size (13% of the total) being 5 nucleotides. The large-subunit rRNA exhibits a similar pattern, with tetra-loops again being the most prevalent (38% of the total) and penta-loops the next (24%) (12).

Table 1 gives an overall impression of the sequence of the tetra-loops in prokaryotic 16S rRNAs and the variations that occur therein. It is immediately apparent that tetra-loop sequences are highly constrained, as are the evolutionarily permissible changes therein. Of the 16 bacterial tetra-loops listed in Table 1, the dominant sequence of 9 of them fits the general pattern GNRA; and where significant variation in this sequence is encountered, the main alternative (which in almost all cases has arisen independently multiple times) tends to conform to the same pattern. More interestingly, in several cases where the dominant sequence is not of the form GNRA, one of the dominant alternative sequences is. A second sequence motif commonly encountered in 16S rRNA tetra-loops is UUCG (see Table 1). It is the dominant sequence in three of the bacterial cases, and serves as a main alternative in several others. The dominant sequence in all but three of the tetra-loops of Table 1 can be described by either GNRA or UNCG.

To a first approximation archaeal‖ 16S rRNAs show the same tetra-loops as are found in bacterial 16S rRNAs. However, the archaeal 16S rRNA structure lacks four of the loops typical of bacteria and contains one not usually found in bacteria, at position 1135 (see Table 1). [The approximate bacterial homolog of the archaeal position 1135 structure almost always has a loop of five or six nucleotides (5, 6); see Fig. 1.] For all but one of the tetra-loops in Fig. 1, sequence is the same (or of the same general type) in both prokaryotic domains. Variations in the dominant sequences are also similar in the archaea and bacteria. For the lone exception, the loop at position 863, the sequence difference between the archaeal and bacterial versions appears to reflect the composition of the tertiary pairing between positions 866 and 570, which has a different characteristic composition in archaea than in bacteria (7).

Three interrelated factors potentially influence the sequence of a loop: the physical stability of the hairpin structure *per se*, interactions of a loop with other parts of the rRNA molecule (or other molecules), and the degree of selective

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.

†To whom reprint requests should be addressed.

¶Woese, C. R., Oral Presentation, Indiana University Symposium, Sept. 29–Oct. 2, 1985, Bloomington, IN.

‖The terms "archaea" and "bacteria" are used herein in lieu of the more familiar "archaebacteria" and "eubacteria," in keeping with the recently proposed system of organisms based upon the naturally delineated "domains" (13).

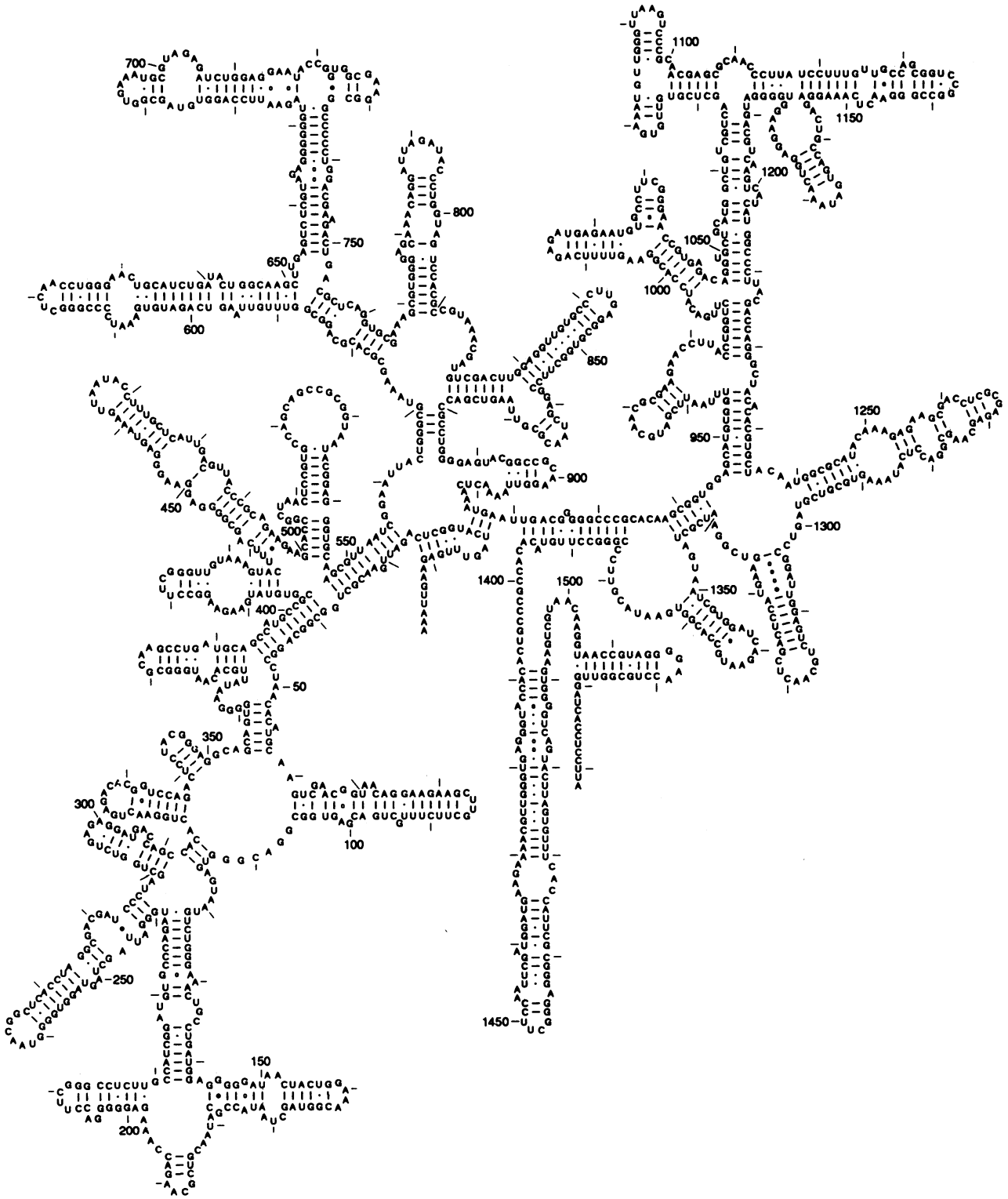

FIG. 1. Secondary structural diagram for a representative bacterial 16S rRNA sequence [*Escherichia coli* (5, 6, 9, 10)]. Every 10th position is marked with a line, every 50th is numbered. Canonical (G-C, C-G, etc.) base pairs are connected by lines, G–U (U–G) pairs by dots, A–G pairs by open circles, and other noncanonical "pairs" (including those with bases not in the normal *anti-anti* configuration) by filled circles (9, 10).

pressure associated with a given sequence. In that loop sequence is, to a first approximation, independent of the loop's location in the overall molecule, and that we have so far failed to detect correlations between (sometimes drastic) sequence changes in a given tetra-loop and changes elsewhere in the 16S rRNA (with the exception of the above-mentioned loop at position 863), we feel that (selection for) stability of the hairpin structure itself is the primary, though not necessarily the only, determinant of a tetra-loop's sequence.

Of the 16S rRNA tetra-loops, the one located at position 83 is perhaps the most interesting and informative. In more than 95% of bacterial examples, this loop comprises four nucleotides, and the sequence of both the loop and its underlying stalk vary frequently (unpublished analysis). [The stalk, whose base is well defined and fixed—by the pairing between

Table 1. Sequence of tetra-loops in prokaryotic 16S rRNAs

Loop position	Domain[a]	Dominant loop sequence	Main alternative sequences	Dominant closing pair	Main alternative closing pair(s)
83			See Tables 2 and 3		
159	B	GAAA 100%	—	G·C 65%	C·G 22%, A·U 11%
	A	GAAA 100%	—	G·C 100%	—
187	B	GCAU[b] 80%	ACAU 8%	C·G 70%	G·C 19%, U·G 7%
208	B	UUCG[c] 40%	UUUA 25%, GCAA 11%	C·G 59%	A·U[d] 25%, G·C 7%
	A	UYCG[e] 52%	AUAU 12%, UCAG 9%	C·G 52%	A·U 27%, U·G 15%
297	B	GAGA >98%	—	U·G 97%	C·G 2.5%
	A	GAGA 77%	GGGA 19%	U·G 100%	—
343	B	UACG >99%	—	C·G >99%	—
	A	UACG 100%	—	C·G 100%	—
380	B	GAAA[f] 64%	GCAA 29%, GGAA 5%	C·G 75%	G·C 25%
	A	GAAA 69%	GCAA 29%	G·C 60%	C·G 36%
420	B	UUCG[g] 79%	UUAG 10%, CUYG 3%	C·G 72%	U·G 28%
727	B	GAAG 86%	GAAA 12%	C·G 96%	G·C 2%
	A	GAAG 86%	GAAA 14%	C·G 100%	—
863	B	UAAC[h] 83%	GAAA 9%, AAAC 6%	C·G 83%	U·G 8%, U·A 8%
	A	GAAG 81%	GAAA 19%	G·C 93%	—
898	B	GCAA 100%	—	C·G 98%	G·C 2%
	A	GCAA 100%	—	C·G 98%	U·G 2%
1013	B	GAGA[i] 82%	GAAA 16%	A·U 76%	G·C 15%, U·G 3%
1029	B	UUCG[j] 75%	GCAA 11%, GAAA 4%	C·G 89%[k]	G·C 7%
1077	B	GUGA 100%	—	C·G 95%	U·A 5%
	A	GUGA 91%	GCGA 7%	U·A 55%	C·G 45%
1135	A	UCCG 49%	UUCG 22%	C·G 97%	U·G 3%
1266	B	GCGA 65%	GUGA 22%, GYAA 12%	C·G 61%	G·C 17%, A·U 10%
	A	GAAA 88%	GAGA 12%	C·G 67%	U·A 33%
1450	B	GCAA 33%	UUCG 15%, GUAA 11%	C·G 81%	U·A 10%
1516	B	GGaa[l] 95%	—	C·G 72%	G·C 28%
	A	GGaa[l] 100%	—	G·C 62%	G·U 31%

[a]A, archaea; B, bacteria (13).
[b]Analysis confined to cases in which stalk has ≈10 pairs (8).
[c]Analysis confined to purple bacteria; too complex otherwise to describe in table.
[d]Closing pair for (all) UUUA loops only.
[e]A few irregular forms encountered (not included in analysis).
[f]Fusobacteria exhibit a loop of five nucleotides, not included in analysis.
[g]The flavobacteria and relatives have a loop of three nucleotides, not included.
[h]Position 866 is involved in a tertiary pseudoknot interaction (7).
[i]A small fraction of loops appear to be closed with noncanonical pairs.
[j]A small fraction of loops are five nucleotides in length.
[k]More than 98% of UUCG loops have a C·G closing pair.
[l]Lowercase a signifies N^6-dimethyladenosine (5, 14).

positions 61–63 and 104–106 (15)—is an irregular helix that shows considerable variation in length (from 24 to 72 nucleotides), in the composition of base pairs, and as to the presence or absence of noncanonical pairs and/or bulged nucleotides (5, 6).] Given this degree of (independent) variation in the overall helix, it is likely that this particular tetra-loop is relatively unconstrained, in the sense of being free of interactions with other parts of the 16S rRNA. If so, the position 83 loop is a good example of a "pure" tetra-loop, one whose sequence is determined solely by internal constraints, rather than by interaction with other elements in rRNA. In further support of this argument we note that in some mitochondrial small-subunit rRNAs the structure in question becomes much larger than the largest known bacterial versions, reinforcing the notion that it is situated unincumbered on the exterior of the small ribosomal subunit (16). [Conceivably the function of this helix is simply to nucleate rRNA folding, as the molecule is being transcribed from its corresponding DNA template. Let it be noted in this context that the helix in question appears particularly stable, as judged by the difficulty usually experienced in sequencing this region of the molecule.]

Tables 2 and 3 show the phylogenetic distribution of the sequence of the position 83 tetra-loop and its (proximal) closing base pair. In 93% of cases, the loop proper has one of three sequences, CUUG (45%), UUCG (36%), or GCAA (13%). To a first approximation the three are more or less evenly distributed phylogenetically, and each of them has arisen independently at least a dozen times. Only 7 other tetra-loop sequences (of the 256 possible) have been observed at position 83, in addition to the tri-loop UUU (which has arisen independently at least seven times), and one example of a penta-loop (see Table 2). Moreover, some of these minor alternative sequences are obvious variations on one of the three principal motifs. For unknown reasons GCAA (and a very small number of GUAAs) are the only variants of the above-discussed GNRA motif encountered in this particular loop; this finding contrasts with the frequent occurrence of other variants, such as GAAA and GYGA, in tetra-loops elsewhere in the molecule (see Table 1). Two other highly variable tetra-loops, at positions 1029 and 1450, also show the same pattern—i.e., almost all of the examples of GNRA found in these two cases are confined to the GYAA pattern (the data of Table 1 show this in part).

It is apparent from Table 3 that the sequence of a tetra-loop influences the composition of the terminal pair in the underlying stalk: The UUCG tetra-loop (at position 83) is almost always associated with a C·G underlying pair, the CUUG loop with a G·C pair, and the GCAA loop usually with an A·U pair. Loop sequence does not have a strong influence on the

Table 2. Sequence of the hairpin loop at position 83 in 16S rRNA

Loop	No. of examples in purple bacteria: α subdivision	β subdivision	γ subdivision	δ subdivision	Loop	No. of examples in Gram-positive bacteria: *Lactobacillus*[a]	Mycoplasma[b]	High G+C	Other[c]	Loop	No. of examples in other bacterial phyla: Flavobacteria[d]	Spirochetes[e]	Thermotogales	Other[f]
UUCG	16 (2)	6	8 (3)	5	UUCG	6 (2)	4 (4)	9 (2)	45	UUCG	11 (2)	3	4	6
CUUG	2	12 (2)	27	5 (2)	CUUG	52 (2)	24 (4)	17 (4)	3 (3)	CUUG	8 (4)	3 (2)	0	0
GCAA	7	3 (3)	0	1	GCAA	0	17 (5)	0	1	GCAA	8 (4)	4 (3)	0	2
CUCG	1	0	0	0	UACG	0	0	0	1	CUCG	1	0	0	0
GUAA	0	0	1	0	GUAA	0	1	0	0	AUUU	0	1	0	0
UUUA	0	0	1	0	AUUA	0	1	0	0	CGUG	0	0	0	1
UUU	1	0	0	0	UUUA	0	3 (3)	0	0	UUCGG	0	1	0	0
					UUUU	0	1	1	1	UUU	0	0	0	1
					UUU	0	2 (2)	4 (2)	1					

Data are presented as the number of examples of each loop sequence, with the minimum estimate of phylogenetically independent occurrences (>1) in parentheses. The data are from the Ribosomal RNA Database Project at the University of Illinois.
[a]Includes relatives such as *Bacillus, Streptococcus*, and others.
[b]Includes walled relatives (17)
[c]Includes *Clostridium, Heliobacterium, Sporomusa* and others, and the fusobacteria.
[d]Includes *Flavobacterium, Flexibacter, Cytophaga, Bacteroides*, and others (14, 18).
[e]Includes spirochetes, treponemes, and leptospiras.
[f]Includes green sulfur and nonsulfur, planctomyces, chlamydia, and deinococcus phyla (14).

composition of the penultimate base pair, however, in that phylogenetic relationship is more evident in the composition of the penultimate pair than in the terminal pair (unpublished observation).

While C·G and G·C pairs account for roughly 25% and 30%, respectively, of all base pairs in a (mesophilic) bacterial 16S rRNA, they account for the vast majority of terminal closing pairs of hairpin loops in general—i.e., about 45% and 40%, respectively. For tetra-loops, C·G closures predominate, accounting for about 60% of cases, while the G·C contribution drops to 20% or less. When the loop sequence is UUCG, the closing pair is C·G in 82% of bacterial 16S rRNAs (not taking into account the tetra-loop at position 83) with U·G pairs accounting for 16%, almost all of the remaining cases. However, the latter are for the most part confined to particular helices in the 16S rRNA molecule. As might be expected, other tetra-loops belonging to the UNCG family are also closed almost exclusively by C·G pairs. Although relatively few UUCG tetra-loops are found in 23S rRNAs, 82% of these have C·G closures. And, as is known from other work (11), (C)UUCG(G) loops seem characteristic of functional RNAs in general.

Loop-specific constraints on the composition of the closing pair for the other principal tetra-loop sequences are not so strict as for UUCG, and they tend to be loop-location specific as well. Except for the tetra-loop at position 83 (where its closing pair is almost always G·C), CUUG tetra-loops are relatively rare in both 16S and the 23S rRNAS: in 23S rRNA the closing pair is G·C in four of five examples (A·U in the remaining one). For 16S rRNA the closing pair is Y·G for the CUUG versions of the loop at position 420 (three phylogenetically independent examples), G·C for the CUUG versions at position 1029 (two phylogenetically independent examples), and G·C and U·A for those at position 208 (two phylogenetically independent examples). With regard to the closing pair for GYAA loops in 16S rRNA (exclusive of the position 83 loop, where GYAA loops are associated mainly with A·U closures), C·G ranks more than 10-fold above all others, A·U and G·C each accounting for about 7%, with other pairings occurring an order of magnitude less frequently than this. However, the A·U closing pair tends to be a significantly higher fraction of the total for those GYAAs in loops that undergo relatively frequent compositional variation. The 23S rRNA molecule shows no particularly strong bias toward any single composition of the closing pair for GYAA tetra-loops.

Table 3. Closing base pair for the position 83 tetra-loops

Loop	No. of examples: Total	Closing pair: G·C	A·U	G·U	C·G	U·A	U·G
UUCG	123	2	0	0	112	1	8
CUUG	153	146	7	0	0	0	0
GCAA	43	4	33	0	6	0	0
GUAA	2	0	2	0	0	0	0
CUCG	2	1	0	0	1	0	0
UACG	1	0	0	0	1	0	0
UUUA	4	1	3	0	0	0	0
UUUU	3	0	0	0	2	1	0
AUUA	1	0	0	0	1	0	0
AUUU	1	0	1	0	0	0	0
CGUG	1	1	0	0	0	0	0
UUU	9	0	0	0	4	5	0

It is apparent that under certain circumstances penta- and tri-loops substitute for tetra-loops. Penta-loops replace the normal 16S rRNA tetra-loops at positions 380, 1029, and 1450 in several major bacterial groups; they also occur as occasional exceptions to tetra-loops elsewhere in the molecule in many bacterial groups (see Table 1). The sequence of these penta-loops often appears derivative of one of the dominant tetra-loop motifs—e.g., CUUGU. The tri-loops that replace tetra-loops occur as rare variants in almost all cases, most having the sequence UUU (with a closing pyrimidine·purine pair) (see Table 3). Their limited and spotty phylogenetic distribution suggests that tri-loops are under negative selection pressure. The only phylogenetically stable tri-loop replacement for a tetra-loop in bacterial 16S rRNA is found at position 420; its sequence is UNU, and it is confined to the flavobacteria and relatives (refs. 14 and 18; C.R.W., unpublished analysis).

Given the exceptional stability of the (C)UUCG(G) tetra-loop (11), this sequence might occur in the nonloop regions of rRNA with lower than random expected frequency, for it could potentially form a structure capable of interfering with normal molecular folding, and so be selected against. We have tallied the occurrence of all sequences of the form XCNNNNGX′ (where X and X′ form a canonical pair) in areas of 16S rRNA that are *not* in tetra- or penta-loop conformation. The sequence XCUUCGGX′ is found only six times in such nonloop regions. While this number of occur-

rences is low, it is by no means exceptionally so, for 80 of the 256 possible loop sequences occur five or fewer times. Although the six occurrences of XCUUCGGX′ are all phylogenetically independent, they are confined to two positions in the molecule (one occurrence at position 137 and five at position 849), and all have the form ACUUCGGU. This restricted distribution is consistent with a weak selective pressure against the occurrence of the sequence XCUUCGGX′ in 16S rRNA.

The authors have benefited from discussion of this problem with Prof. G. J. Olsen. C.R.W.'s contribution and R.R.G.'s initial work on this project have been supported by Grant NSG-7044 from the National Aeronautics and Space Administration. S.W. has been supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, Department of Energy, under Contract W-31-109-Eng-38. The Ribosomal RNA Database Project is funded by the National Science Foundation.

1. Turner, D. H., Sugimoto, N. & Freier, S. M. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17,** 167–192.
2. Jaeger, J. A., Turner, D. H. & Zuker, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 7706–7710.
3. Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., Crawford, N., Brosius, J., Gutell, R. R., Hogan, J. J. & Noller, H. F. (1980) *Nucleic Acids Res.* **8,** 2275–2293.
4. Noller, H. F., Kop, J., Wheaton, V., Brosius, J., Gutell, R. R., Kopylov, A. M., Dohme, F., Herr, W., Stahl, D. A., Gupta, R. & Woese, C. R. (1981) *Nucleic Acids Res.* **9,** 6167–6189.
5. Woese, C. R., Gutell, R. R., Gupta, R. & Noller, H. F. (1983) *Microbiol. Rev.* **47,** 621–669.
6. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. (1985) *Prog. Nucleic Acids Res. Mol. Biol.* **32,** 155–216.
7. Gutell, R. R., Noller, H. F. & Woese, C. R. (1986) *EMBO J.* **5,** 1111–1113.
8. Leffers, H., Kjems, J., Ostergaard, L., Larsen, N. & Garrett, R. A. (1987) *J. Mol. Biol.* **195,** 43–61.
9. Woese, C. R. & Gutell, R. R. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 3119–3122.
10. Gutell, R. R. & Woese, C. R. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 663–667.
11. Tuerk, C., Gauss, P., Thermes, C., Groebe, D. R., Gayle, M., Guild, N., Stormo, G., d'Aubenton-Carafa, Y., Uhlenbeck, O. C., Tinoco, I., Brody, E. N. & Gold, L. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 1364–1368.
12. Gutell, R. R. & Fox, G. E. (1988) *Nucleic Acids Res.* **16,** Suppl., R175–R269.
13. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579.
14. Woese, C. R. (1987) *Bacteriol Rev.* **51,** 221–271.
15. Winker, S., Overbeek, R., Woese, C. R., Olsen, G. J. & Pfluger, N. (1989) *An Automated Procedure for Covariation-Based Detection of RNA Structure* (Argonne Natl. Laboratory, Argonne, IL), Tech. Rep. ANL-89/42.
16. Spencer, D. F., Schnare, M. N. & Gray, M. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 493–497.
17. Weisburg, W. G., Tully, J. G., Rose, D. L., Petzel, J. P., Oyaizu, H., Yang, D., Mandelco, L., Sechrest, J., Lawrence, T. G., van Etten, J., Maniloff, J. & Woese, C. R. (1989) *J. Bacteriol.* **171,** 6455–6467.
18. Woese, C. R., Maloy, S., Mandelco, L. & Raj, H. D. (1990) *Syst. Appl. Microbiol.* **13,** 19–23.