Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 4576–4579, June 1990
Evolution

Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya

(Euryarchaeota/Crenarchaeota/kingdom/evolution)

CARL R. WOESE*†, OTTO KANDLER‡, AND MARK L. WHEELIS§

*Department of Microbiology, University of Illinois, 131 Burrill Hall, Urbana, IL 61801; ‡Botanisches Institut der Universität München, Menzinger Strasse 67, 8000 Munich 19, Federal Republic of Germany; and §Department of Microbiology, University of California, Davis, CA 95616

Contributed by Carl R. Woese, March 26, 1990

ABSTRACT **Molecular structures and sequences are generally more revealing of evolutionary relationships than are classical phenotypes (particularly so among microorganisms). Consequently, the basis for the definition of taxa has progressively shifted from the organismal to the cellular to the molecular level. Molecular comparisons show that life on this planet divides into three primary groupings, commonly known as the eubacteria, the archaebacteria, and the eukaryotes. The three are very dissimilar, the differences that separate them being of a more profound nature than the differences that separate typical kingdoms, such as animals and plants. Unfortunately, neither of the conventionally accepted views of the natural relationships among living systems—i.e., the five-kingdom taxonomy or the eukaryote–prokaryote dichotomy—reflects this primary tripartite division of the living world. To remedy this situation we propose that a formal system of organisms be established in which above the level of kingdom there exists a new taxon called a "domain." Life on this planet would then be seen as comprising three domains, the Bacteria, the Archaea, and the Eucarya, each containing two or more kingdoms. (The Eucarya, for example, contain Animalia, Plantae, Fungi, and a number of others yet to be defined.) Although taxonomic structure within the Bacteria and Eucarya is not treated herein, Archaea is formally subdivided into the two kingdoms Euryarchaeota (encompassing the methanogens and their phenotypically diverse relatives) and Crenarchaeota (comprising the relatively tight clustering of extremely thermophilic archaebacteria, whose general phenotype appears to resemble most the ancestral phenotype of the Archaea).**

Need for Restructuring Systematics

Within the last decade it has become possible to trace evolutionary history back to the (most recent) common ancestor of all life, perhaps 3.5–4 billion years ago (1, 2). Prior to the mid 1970s evolutionary study had for all intents and purposes been confined to the metazoa and metaphyta, whose histories at best cover 20% of the total evolutionary time span. A sound basis for a natural taxonomy was provided in these cases by complex morphologies and a detailed fossil record. The evolution of the microbial world—whose history spans most of the planet's existence—was at that time beyond the biologist's purview, for, unlike their multicellular equivalents, microbial morphologies and other characteristics are too simple or uninterpretable to serve as the basis for a phylogenetically valid taxonomy (3, 4). The sequencing revolution, by making accessible the vast store of historical information contained in molecular sequences (5), has changed all that. As a result, the biologist finds that textbook descriptions of the basic organization of life have become outmoded and so, misleading. The time has come to bring formal taxonomy into line with the natural system emerging from molecular data.

This revision, however, is not accomplished simply by emending the old system. Our present view of the basic organization of life is still largely steeped in the ancient notion that all living things are either plant or animal in nature. Unfortunately, this comfortable traditional dichotomy does not represent the true state of affairs. Thus, as a prerequisite to developing a proper natural system we have to divest ourselves of deeply ingrained, cherished assumptions, as regards both the fundamental organization of life and the basis for constructing a system of organisms. The system we develop will be one that is completely restructured at the highest levels.

Haeckel in 1866 (6) formally challenged the aboriginal plant/animal division of the living world. He recognized that the single-celled forms, the protists, did not fit into either category; they must have arisen separately from both animals and plants. Haeckel saw the tree of life, therefore, as having three main branches, not two. Copeland (7) later split out a fourth main branch, a new kingdom accommodating the bacteria, and Whittaker (8) created a fifth, for the fungi. While Haeckel's original proposal and its two more recent refinements did away with the idea that animal/plant was *the* primary distinction, they left unchallenged the notion that it is *a* primary distinction (by representing it at the highest available taxonomic level). The last of these schemes (Whittaker's), which divides the living world into Animalia, Plantae, Fungi, Protista, and Monera, is the most widely received view of the basic organization of life today (8, 9).

It has been apparent for some time, however, that the five-kingdom scheme (and its predecessors) is not phylogenetically correct, is not a natural system. There are sound logical grounds for presuming that the two eukaryotic microbial taxa (Protista and Fungi) are artificial. It is generally accepted that the metaphyta and metazoa evolved from unicellular eukaryotic ancestors; the extant groups of eukaryotic microorganisms, therefore, comprise a series of lineages some (or many) of which greatly antedate the emergence of the Plantae and Animalia. This is confirmed by the fossil record, wherein recognizable eukaryotic unicells appear about 200 million years before the first primitive algae, and over a billion years before the first animals and higher plants (10). There are thus good reasons in principle to presume that the Protista and perhaps also the Fungi are paraphyletic at best.

More seriously, in giving the kingdom Monera the same taxonomic rank as the Animalia, Plantae, Fungi, and Protista, the five-kingdom formulation ignores the fact that the differences between Monera (prokaryotes) and the four other kingdoms are far more significant, and of a qualitatively different nature, than the differences among these four. In

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.

other words, a primary division of life must lie between the bacteria and the eukaryotic forms; the animal/plant distinction is definitely secondary.

This realization is by no means new. Microbiologists acknowledged it more than 100 years ago (11), and, of course, Chatton (12) codified it with his famous eukaryote–prokaryote proposal, dividing all life into these two primary categories. This view of life, strangely, has coexisted for some time now with the five-kingdom scheme, despite their basic incompatibility, and despite the fact that the evidence overwhelmingly supports the former. However, the eukaryote–prokaryote concept itself has been seriously misunderstood and, consequently, wrongly interpreted.

The problem here arises because the eukaryote–prokaryote concept is fundamentally cytological, and only secondarily, and by inference, phylogenetic. The presumption that the eukaryotic form of cellular organization defines a meaningful phylogenetic unit is a reasonable one; organisms with this cytology are united by possession of a series of complex properties. The same is unfortunately not true of prokaryotes, which are united as a class by their *lack* of the characteristics that define the eukaryotic cell. The definition is consequently a negative one that is empty of meaningful internal phylogenetic information. Microbiologists have long recognized this (even before the articulation of the eukaryotic–prokaryotic concept): e.g., Cohn in 1875: "Perhaps the designation of *Schizophytae* may recommend itself for this first and simplest division of living beings . . . even though its distinguishing characters are negative rather than positive" (11); Pringsheim in 1923: ". . . the possibility of . . . convergent evolution [among bacteria must] be seriously considered" (13); and Stanier in 1971: "Indeed the major contemporary procaryotic groups could well have diverged at an early stage in cellular evolution, and thus be almost as isolated from one another as they are from eucaryotes as a whole" (14).

As the molecular and cytological understanding of cells deepened at a very rapid pace, beginning in the 1950s, it became feasible in principle to define prokaryotes positively, on the basis of shared molecular characteristics. However, since molecular biologists elected to work largely in a few model systems, which were taken to be representative, the comparative perspective necessary to do this successfully was lacking. By default, *Escherichia coli* came to be considered typical of prokaryotes, without recognition of the underlying faulty assumption that prokaryotes are monophyletic. This presumption was then formalized in the proposal that there be two primary kingdoms: *Procaryotae* and *Eucaryotae* (15, 16). It took the discovery of the archaebacteria to reveal the enormity of this mistake.

On the cytological level archaebacteria are indeed prokaryotes (they show none of the defining eukaryotic characteristics), but on the molecular level they resemble other procaryotes, the eubacteria, no more (probably less) than they do the eukaryotes (1, 17). *Procaryotae* (and its synonym *Monera*) cannot be a phylogenetically valid taxon.

Basis for Restructuring

What must be recognized is that the basis for systematics has changed; classical phenotypic criteria are being replaced by molecular criteria. As Zuckerkandl and Pauling (5) made clear many years ago, it is at the level of molecules (particularly molecular sequences) that one really becomes privy to the workings of the evolutionary process. Molecular sequences can reveal evolutionary relationships in a way and to an extent that classical phenotypic criteria, and even molecular functions, cannot; and what is seen only dimly, if at all, at higher levels of organization can be seen clearly at the level of molecular structure and sequences. Thus, systematics in the future will be based primarily upon the sequences, structure, and relationships of molecules, the classical gross properties of cells and organisms being used largely to confirm and embellish these.

It is only on the molecular level that we see the living world divide into three distinct primary groups. For every well-characterized molecular system there exists a characteristic eubacterial, archaebacterial, and eukaryotic version, which all members of each group share. Ribosomal RNAs provide an excellent example (in part because they have been so thoroughly studied). One structural feature in the small subunit rRNA by which the eubacteria can be distinguished from archaebacteria and eukaryotes is the hairpin loop lying between positions 500 and 545 (18), which has a side bulge protruding from the stalk of the structure. In all eubacterial cases (over 400 known) the side bulge comprises six nucleotides (of a characteristic composition), and it protrudes from the "upstream" strand of the stalk between the fifth and sixth base pair. In both archaebacteria and eukaryotes, however, the corresponding bulge comprises seven nucleotides (of a different characteristic composition), and it protrudes from the stalk between the sixth and seventh pair (18, 19). The small subunit rRNA of eukaryotes, on the other hand, is readily identified by the region between positions 585 and 655 (*E. coli* numbering), because both prokaryotic groups exhibit a common characteristic structure here that is never seen in eukaryotes (18, 19). Finally, archaebacterial 16S rRNAs are readily identified by the unique structure they show in the region between positions 180 and 197 or that between positions 405 and 498 (18, 19). Many other examples of group-invariant rRNA characteristics exist; see refs. 2, 18, and 19. [The reader wishing to gain a broader and more detailed appreciation for the molecular definition of the three groups can consult refs. 2, 20, and 21 and the proceedings of the most recent conference on archaebacteria (22).]

Molecular characterizations also reveal that the evolutionary differences among eubacteria, archaebacteria, and eukaryotes are of a more profound nature than those that distinguish traditional kingdoms, such as animals and plants, from one another. This is most clearly seen in the functions that must have evolved early in the cell's history and are basic to its workings. All eubacteria, for example, exhibit nearly the same subunit pattern (in terms of numbers and sizes) in their RNA polymerases; however, this pattern bears little relationship to that seen in either the archaebacteria or the eukaryotes (23). On the other hand, eukaryotes are unique in using three separate RNA polymerase functions (24).

The fossil record indicates that photosynthetic eubacteria (and by inference, therefore, archaebacteria and possibly eukaryotes) were already in existence 3–4 billion years ago (25), so that the evolutionary events that transformed the ancestor common to all life into the individual ancestors of each of the three major groups must have occurred over a relatively short time span early in the planet's history. Both the relatively rapid pace of, as well as the profound changes associated with, this early evolutionary transition argue that this universal ancestor was a simpler, more rudimentary entity than the individual ancestors that spawned the three groups (and their descendants) (26).

Fig. 1 is a universal phylogenetic tree, showing the relationships among the primary groups. The root of the tree is seen to separate the eubacteria from the other two primary groups, making the archaebacteria and eukaryotes specific (but distant) relatives. A relationship between archaebacteria and eukaryotes is not overly surprisingly, for with few exceptions (the rRNA being one) the archaebacterial versions of molecules resemble their eukaryotic homologs more than their eubacterial ones (24, 29, 30). Among the ribosomal proteins there are even cases where the archaebacterial and

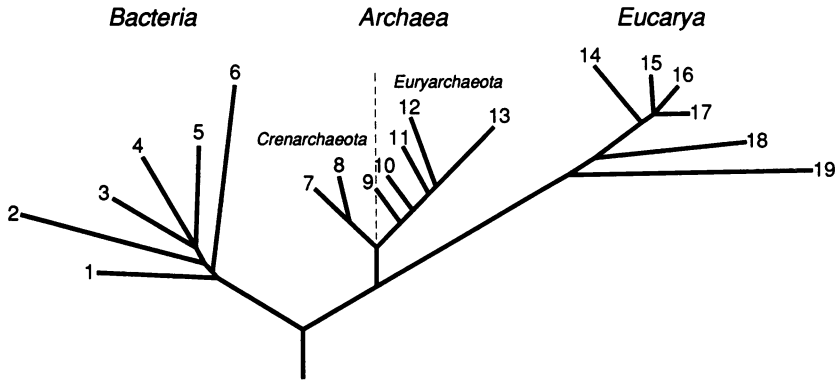

FIG. 1. Universal phylogenetic tree in rooted form, showing the three domains. Branching order and branch lengths are based upon rRNA sequence comparisons (and have been taken from figure 4 of ref. 2). The position of the root was determined by comparing (the few known) sequences of pairs of paralogous genes that diverged from each other before the three primary lineages emerged from their common ancestral condition (27). [This rooting strategy (28) in effect uses the one set of (aboriginally duplicated) genes as an outgroup for the other.] The numbers on the branch tips correspond to the following groups of organisms (2). Bacteria: 1, the Thermotogales; 2, the flavobacteria and relatives; 3, the cyanobacteria; 4, the purple bacteria; 5, the Gram-positive bacteria; and 6, the green nonsulfur bacteria. Archae: the kingdom Crenarchaeota: 7, the genus *Pyrodictium*; and 8, the genus *Thermoproteus*; and the kingdom Euryarchaeota: 9, the Thermococcales; 10, the Methanococcales; 11, the Methanobacteriales; 12, the Methanomicrobiales; and 13, the extreme halophiles. Eucarya: 14, the animals; 15, the ciliates; 16, the green plants; 17, the fungi; 18, the flagellates; and 19, the microsporidia.

eukaryotic homologs have no apparent counterpart among the eubacteria (29, 30).

From a systematic perspective the specific relationship between eukaryotes and archaebacteria does not require taxonomic recognition; these two groups are sufficiently dissimilar, and they diverged so early, that little would be gained by defining a taxon that encompasses both. In other words, the archaebacteria and eukaryotes themselves show the kind of profound molecular differences that distinguish either from the eubacteria.

Proposal for a New Highest Level Taxon

> The only truly scientific foundation of classification is to be found in appreciation of the available facts from a phylogenetic point of view. Only in this way can the natural interrelationships [among organisms] . . . be properly understood. (31)

A phylogenetic system must first and foremost recognize the primacy of the three groupings, eubacteria, and archaebacteria and eukaryotes. These must stand above the conventionally recognized kingdoms, Animalia and the like. This raises the question of whether the term "kingdom" should be used for the taxon of highest rank, with the traditional kingdoms being assigned to a new, lower-level taxon. For two reasons we feel this is not the correct solution: From a scientific perspective, the distinctions among eubacteria, archaebacteria, and eukaryotes are more profound than those customarily associated with kingdoms. Furthermore, two centuries of association of the label "kingdom" with the animals and (green) plants constitutes a tradition that would be most difficult and divisive to change. The most flexible and informative (and least disruptive) approach would appear to be to add a new rank at the top of the existing hierarchy. The name we propose for this new and highest taxon is "domain" (whose Latin counterpart we take to be *regio*). The formal suffix that we would associate with names of domains is *-a*, chosen for its simplicity.

Naming of the individual domains has been guided by several general considerations: (*i*) maintaining appropriate continuity with existing names; (*ii*) suggesting basic characteristics of the group; and (*iii*) avoiding any connotation that the eubacteria and archaebacteria are related to one another, which, unfortunately, is implied by their common names. For the eubacteria the formal name *Bacteria*, based upon a traditional common name for the group, is suggested. The term *Eucarya* derives from that group's common name and captures its defining cytological characteristic—i.e., cells with well-defined encapsulated nuclei. The archaebacteria are called *Archaea* to denote their apparent primitive nature (vis a vis the eukaryotes in particular). The formal names for the domains are simple enough that they can also serve in common usage (note that this requires that "bacteria" be used in a sense that does *not* include the archaea). Additionally, "eukaryotes" will continue to be an acceptable common synonym for the Eucarya. However, we recommend abandonment of the term "archaebacteria," since it incorrectly suggests a specific relationship between the Archaea and the Bacteria.

We will not at this time address the matter of the individual kingdoms within the domains, with the exception of the Archaea. For the others, suffice it to say that there will be numerous kingdoms within each domain, and their formal structuring will require a more detailed analysis than is possible here. We anticipate that such an analysis of the Eucarya will preserve the kingdoms Plantae, Animalia, and Fungi (with the last somewhat restructured to reflect new molecular insights), and will replace Protista with a series of kingdoms corresponding to the various ancient protistan lineages. For the Bacteria, we expect that the majority of the described "phyla" (2) will deserve elevation to kingdom rank.

There are, however, two reasons for suggesting formal names for the kingdoms that constitute the Archaea at this time: One is that the phylogenetic structure of the domain seems relatively simple and well defined at the kingdom level. The other is that the kingdoms within the Archaea have never had appropriate names of any kind.

Phylogenetically the Archaea fall into two distinct groups, two major lineages (refs. 2 and 32; see Fig. 1). One, the methanogens and their relatives, is phenotypically heterogeneous, comprising extreme halophiles, sulfate-reducing species (the genus *Archaeoglobus*), and two types of thermophiles (the genus *Thermoplasma* and the *Thermococcus–Pyrococcus* group), in addition to the three methanogenic lineages (2, 33). The proposed formal name for the methanogens and their relatives is *Euryarchaeota*. For this kingdom we use the common name euryarchaeotes or, more casually, euryotes.

The other archaeal kingdom comprises most of what have been variously called the "thermoacidophiles," "sulfur-dependent archaebacteria," "eocytes," or "extreme thermophiles." It is a physiologically relatively homogeneous group, whose niches are entirely thermophilic (2). Since thermophily is the only general phenotype that occurs on both major branches of the Archaea, it is presumably the ancestral phenotype of the Archaea (2). For this kingdom we suggest the name *Crenarchaeota*. In common usage crenarchaeotes or crenotes would be acceptable.

Definitions

Domain Eucarya [Greek adjective εὖ (good; true in modern common usages); and Greek noun κάρυον (nut or kernel; refers to the nucleus in modern biological usage)]: cells eukaryotic; cell membrane lipids predominantly glycerol fatty acyl diesters; ribosomes containing a eukaryotic type of rRNA (2, 18, 19).

Domain Bacteria [Greek noun βακτήριον (small rod or staff)]: cells prokaryotic; membrane lipids predominantly diacyl glycerol diesters; ribosomes containing a (eu)bacterial type of rRNA (2, 18, 19).

Domain Archaea [Greek adjective ἀρχαῖος (ancient, primitive)]: cells prokaryotic; membrane lipids predominantly isoprenoid glycerol diethers or diglycerol tetraethers; ribosomes containing an archaeal type of rRNA (2, 18, 19).

Kingdom Euryarchaeota (Archaea) [Greek adjective εὐρύς (broad, wide, spacious), for the relatively broad spectrum of niches occupied by these organisms and their varied patterns of metabolism; Greek adjective ἀρχαῖος (ancient, primitive)]: ribosomes containing a euryarchaeal type of rRNA (2, 18, 19).

Kingdom Crenarchaeota (Archaea) [Greek noun κρήνη (spring, fount), for the ostensible resemblance of this phenotype to the ancestor (source) of the domain Archaea; and Greek adjective ἀρχαῖος (ancient, primitive)]: ribosomes containing a crenarchaeal type of rRNA (2, 18, 19).

Conclusion

The system we propose here will repair the damage that has been the unavoidable consequence of constructing taxonomic systems in ignorance of the likely course of microbial evolution, and on the basis of flawed premises (that life is dichotomously organized; that negative characteristics can define meaningful taxonomies). More specifically, it will (*i*) provide a system that is natural at the highest levels; (*ii*) provide a system that allows a fully natural classification of microorganisms (eukaryotic as well as prokaryotic); (*iii*) recognize that, at least in evolutionary terms, plants and animals do not occupy a position of privileged importance; (*iv*) recognize the independence of the lineages of the Archaea and the Bacteria; and (*v*) foster understanding of the diversity of ancient microbial lineages (both prokaryotic and eukaryotic).

The authors acknowledge extremely helpful discussions of this problem with the following individuals: Prof. Dr. J. Poelt, University of Graz; Prof. Dr. F. Ehrendorfer, University of Vienna; Prof. L. Zgusta, University of Illinois; and G. H. Woese. We thank S. Smith for his assistance in producing Fig. 1. The work of C.R.W. in this area is supported by a grant from the National Aeronautics and Space Administration (NSG-7044). The work of O.K. is supported by a grant from the German Collection of Microorganisms, Göttingen, F.R.G.

1. Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R., Magrum, L., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N. & Woese, C. R. (1980) *Science* **209,** 457–463.
2. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
3. van Niel, C. B. (1955) in *A Century of Progress in the Natural Sciences: 1853–1953* (Calif. Acad. Sci., San Francisco), pp. 89–114.
4. Stanier, R. Y. & van Niel, C. B. (1962) *Arch. Mikrobiol.* **42,** 17–35.
5. Zuckerkandl, E. & Pauling, L. (1965) *J. Theor. Biol.* **8,** 357–366.
6. Haeckel, E. (1866) *Generelle Morphologie der Organismen* (Reimer, Berlin).
7. Copeland, H. F. (1938) *Q. Rev. Biol.* **13,** 383–420.
8. Whittaker, R. H. (1959) *Q. Rev. Biol.* **34,** 210–226.
9. Whittaker, R. H. & Margulis, L. (1978) *Biosystems* **10,** 3–18.
10. Knoll, A. H. (1990) in *Origins and Early Evolutionary History of the Metazoa*, eds. Lipps, J. H. & Signor, P. W. (Plenum, New York), in press.
11. Cohn, F. (1875) *Beitr. Biol. Pfl.* **1,** 141–224.
12. Chatton, E. (1938) *Titres et Travoux Scientifiques (1906–1937) de Edouard Chatton* (E. Sottano, Sète, France).
13. Pringsheim, E. G. (1923) *Lotos* **71,** 357–377.
14. Stanier, R. Y. (1971) in *Recent Advances in Microbiology*, eds. Perez-Miravete, A. & Pelaez, D. (Assoc. Mex. Microbiol., Mexico City), pp. 595–604.
15. Murray, R. G. E. (1968) *Spisy Prirodoved. Fak. Univ. J. E. Purkyne Brne* **43,** 249–252.
16. Allsopp, A. (1969) *New Phytol.* **68,** 591–612.
17. Woese, C. R. & Fox, G. E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5088–5090.
18. Woese, C. R., Gutell, R. R., Gupta, R. & Noller, H. F. (1983) *Microbiol. Rev.* **47,** 621–669.
19. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. (1985) *Prog. Nucleic Acid Res. Mol. Biol.* **32,** 155–216.
20. Kandler, O. & Zillig, W., eds. (1986) *Archaebacteria '85* (Fischer, Stuttgart, F.R.G.).
21. Jones, W. J., Nagle, D. P. & Whitman, W. B. (1987) *Microbiol. Rev.* **51,** 135–177.
22. Victoria Meeting on Archaebacteria (1989) *Can. J. Microbiol.* **35** (1).
23. Schnabel, R., Thomm, M., Gerardy-Schahn, R., Zillig, W., Stetter, K. O. & Huet, J. (1983) *EMBO J.* **2,** 751–755.
24. Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H.-P., Lottspeich, F., Garrett, F. A. & Zillig, W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4569–4573.
25. Ernst, W. G. (1983) in *Earth's Earliest Biosphere*, ed. Schopf, J. W. (Princeton Univ. Press, Princeton, NJ), pp. 41–52.
26. Woese, C. R. & Fox, G. E. (1977) *J. Mol. Evol.* **10,** 1–6.
27. Iwabe, N., Kuma, K., Hasegawa, M., Osawa, S. & Miyata, T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 9355–9359.
28. Schwartz, R. M. & Dayhoff, M. O. (1978) *Science* **199,** 395–403.
29. Kimura, M., Arndt, E., Hatakeyama, T., Hatakeyama, T. & Kimura, J. (1989) *Can. J. Microbiol.* **35,** 195–199.
30. Auer, J., Lechner, K. & Böck, A. (1989) *Can. J. Microbiol.* **35,** 200–204.
31. Kluyver, A. J. & van Niel, C. B. (1936) *Zentralbl. Bakteriol. Parasitenkd. Infektionskrankh. Abt. 2* **94,** 369–403.
32. Achenbach-Richter, L., Gupta, R., Zillig, W. & Woese, C. R. (1988) *System. Appl. Microbiol.* **10,** 231–240.
33. Achenbach-Richter, L., Stetter, K. & Woese, C. R. (1987) *Nature (London)* **327,** 348–349.

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).

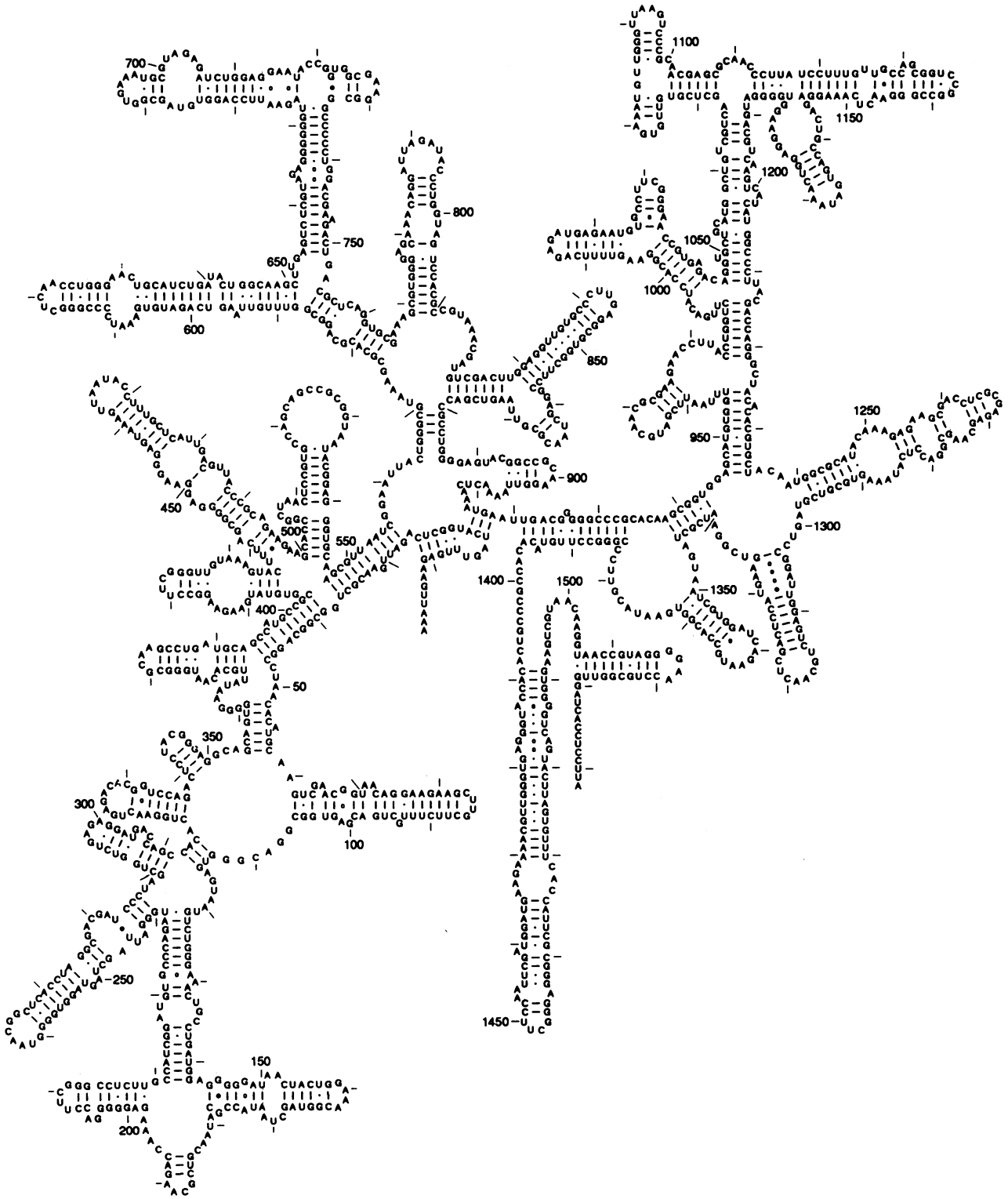

FIG. 1. Secondary structural diagram for a representative bacterial 16S rRNA sequence [*Escherichia coli* (5, 6, 9, 10)]. Every 10th position is marked with a line, every 50th is numbered. Canonical (G-C, C-G, etc.) base pairs are connected by lines, G–U (U–G) pairs by dots, A–G pairs by open circles, 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.