4

Probing RNA Structure, Function, and History by Comparative Analysis

Norman R. Pace and Brian C. Thomas
Departments of Plant and Microbial Biology, and Molecular and Cell Biology
University of California, Berkeley
Berkeley, California 94720

Carl R. Woese
Department of Microbiology
University of Illinois
Urbana, Illinois 61801

Life on this planet is a profusion of incredibly complex systems. From the biologist's perspective it is indeed fortunate that all these systems have sprung from a common ancestor. They have by their nature retained traces of their ancestries, and so are similar—homologous—to a greater or lesser extent. The science of biology has been built from its beginning upon cataloging similarities and differences among different living systems (and various states of the same system), the method that has become known as comparative analysis. It is only through this simple and sometimes tedious approach that biologists could begin to understand the complexity with which they are confronted, could begin to distinguish the important elements from the unimportant, and so reduce living systems to a set of understandable essentials. It is also through such comparisons, through measuring similarity and difference in degree and kind, that biologists have come to learn the genealogical relationships among all organisms.

Despite its compelling utility, comparative analysis has not been a commonly used tool in molecular biology. This is not because comparative analysis is without value at the molecular level: The secondary structures of the ribosomal RNAs, major accomplishments of modern molecular biology, are tribute to the comparative approach (Woese et al. 1980; Noller et al. 1981). The molecular biologist's aversion to comparative analysis would seem to lie in the molecular paradigm itself. Molecular biology arose from chemistry and physics. The entities with which these sciences deal, atoms and their relatively simple chemical combinations, do not have meaningful histories. Consequently, the conceptual and

The RNA World, Second Edition © 1999 Cold Spring Harbor Laboratory Press 0-87969-561-7/99

experimental outlook inherited from chemistry and physics effectively has no comparative, historical dimension. This may also explain why molecular biologists have tended to view evolution (apart from the origin of life issue) as a trivial part of biology, as merely a collection of relatively uninteresting historical accidents: No matter that evolution is the essence, the sine qua non of biology.

Our objective in this chapter is to review the various ways in which the comparative analysis of RNA at the molecular level has contributed and will contribute to biology. Comparative analysis contributes to molecular biology in four important ways: (1) in defining restraints and patterns from which molecular structure can be inferred; (2) as a necessary background for, and adjunct to, experimental analysis of molecular function and structure; (3) as the essence of genealogical analysis and classification; and (4) in providing a general conceptual and organizational framework for much of future biological research.

The changes that have occurred in microbiology over the past 15 years bear witness to the sharpening effect that a phylogenetic/comparative framework has on experimental design and interpretation of results. There can be little doubt that the deluge of genomic sequence information can only be handled effectively in a comparative framework.

COMPARATIVE ANALYSIS AND NUCLEIC ACID STRUCTURE

History: Molecular Biology's Flirtation with Comparative Analysis

The potential of a comparative approach to RNA higher-order structure became evident at the 1966 Cold Spring Harbor Symposium. By that time, the sequences of four tRNAs were known. The sequences reported had not been determined initially with a comparative approach in mind, but the fact that all four tRNA sequences could assume the same "clover leaf" configuration convinced most of those present at the meeting that all tRNAs had a common secondary structure. However, the general principle, that comparative analysis is an effective tool in the analysis of molecular structure, seems not to have been grasped, for the lesson had to be repeated when it came to 5S rRNA structure. Here, too, the initial attempts to determine secondary structure had not involved a systematic, comparative approach. Rather, structures were suggested on the basis of maximizing of base-pairing, for instance, or some other "first principle." The resulting structures tended to be awkwardly complex and unattractive (for review, see Erdmann 1976). A systemic comparative approach was eventually applied in this case, and the true structure of the 5S rRNA then began to emerge (Fox and Woese 1973).

When the sequence of the small subunit rRNA was in the process of being determined, the comparative lesson seemed once more to have been forgotten. Again, the search for secondary structure initially consisted of merely looking for stretches of nucleotides with the potential to form base pairs. The credibility of the resulting structures did not seem to be an issue. There is even an unsubstantiated rumor (worth repeating) that when Brosius, Noller, and their colleagues submitted the first complete 16S rRNA sequence (that of *Escherichia coli*) for publication, one of the reviewers questioned why they had not included the molecule's secondary structure as well! By this time, however, the comparative lesson had to some extent been assimilated, and Noller's group soon teamed up with Woese's group to produce a second sequence, for the specific purpose of using comparisons to identify among the 10,000 or so possible helical elements in any given small subunit rRNA a manageable number of probable ones (Woese et al. 1980). By the time the large subunit rRNA was sequenced (Brosius et al. 1980), it was generally accepted that comparative analysis was essential to determining its secondary structure, and the appropriate experimental steps, determination and comparison of sequences from different organisms, were taken to realize this (Noller et al. 1981).

Using the Comparative Approach to Analyze RNA Structure

Today, comparative analysis has become the method of choice for establishing higher-order structure for large RNAs. Regardless of their other virtues, none of the more direct, physical-chemical approaches can provide so detailed a picture of the relationship among bases in large RNAs.

Comparative analysis starts with the alignment of sets of homologous sequences. In its most stringent definition, homology refers strictly to common ancestry. The simply stated objective of alignment is to juxtapose related sequences so that the homologous residues in each occupy the same column in the alignment. Although there have been significant advances in computer-generated alignment in recent years, the most precise method of alignment is still essentially a manual one (with computer assistance). The reason for this is that a great deal more is involved in alignment than merely shifting nucleotides around until some maximum sequence similarity is achieved. Homology among individual sequence residues is best defined in the context of larger units of homologous structure and function. Therefore, the most useful and generally the most precise alignments are produced by iteratively invoking higher-order structure in the process. A practical condition imposed on sets of aligned sequences is that the sequences are arranged in an order that approximates their phylogenetic relationships to one another. This is the most useful context for comparative analysis and foreshadows a day when phylogenetic structuring will become an integral part of most if not all biological databases.

Given a phylogenetically ordered sequence alignment, one can begin to interpret Nature's evolutionary experimentation. Nature provides the results of those experiments that have worked, those in which molecular function remains optimized (within the limits defined by selective constraints); only rarely are we privy to natural experiments in which function has been adversely altered. Therefore, invariance in compositions of certain residues in a molecule identifies these as being important to the structure and function of the molecule. Conversely, areas of a molecule that frequently vary in composition from organism to organism, especially when length variation also occurs, mark themselves as probably of little or no direct functional significance. Rarely, a particular section of a molecule will become deleted, thereby identifying itself as some kind of functional or structural unit. Residues whose compositions covary (change in concert) must in some way be related, which in almost all cases means in direct physical contact.

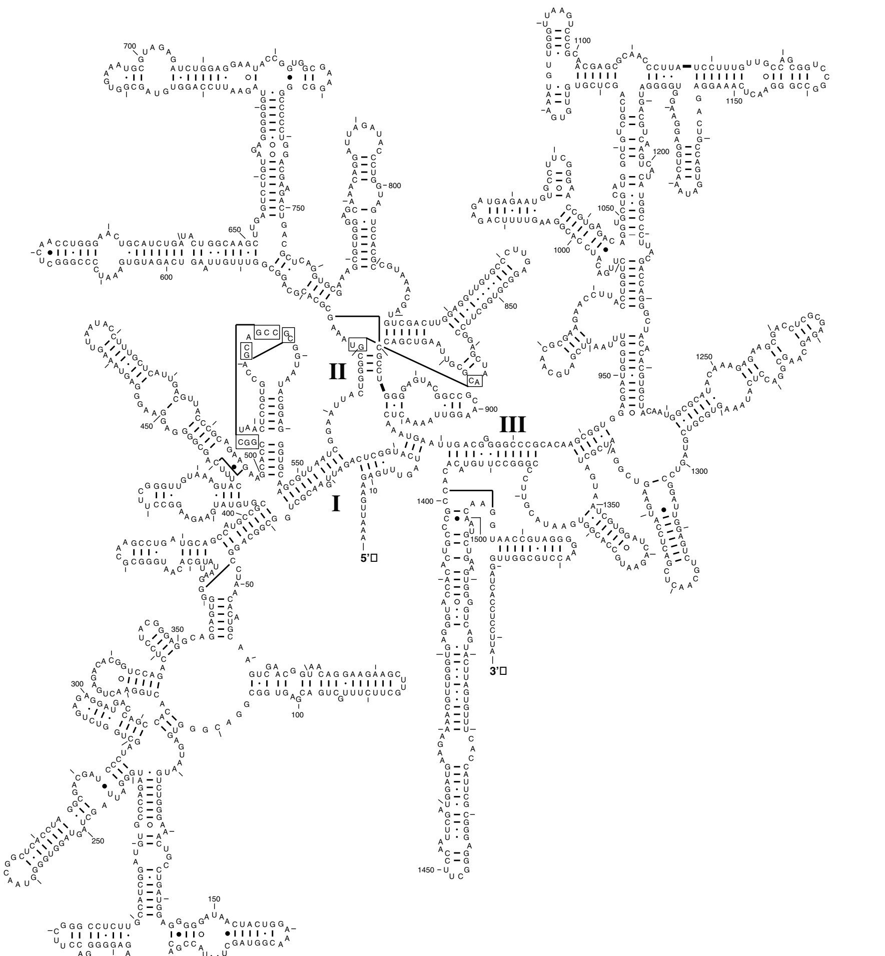
Early comparative studies were limited by difficulties in accumulating large numbers of sequences for consideration. Consequently, comparisons were mainly limited to manual operations. With the expansion of rapid sequencing techniques, large data sets can be accumulated for computer comparisons seeking covariations of sequence or other structural features. Sufficiently large data sets (>100 sequences) can fruitfully be sifted for correlations using computer-based mutual-information algorithms (Gautheret et al. 1995). Homologous genes from different organisms are generally isolated from specific organisms, an arduous task for a thorough analysis. Another way to accumulate a large amount of sequence data for comparative analysis is by use of mixed populations of different organisms, for instance environmental samples, as sources of particular genes. The selected genes are obtained from community DNA by PCR using primers complementary to broadly conserved sequences in the genes (Brown et al. 1996; Hugenholtz et al. 1998). The source organism for a sequence obtained this way would not necessarily be known, but this is generally irrelevant to a comparative structure analysis.

Sequence comparisons reveal more than the simple existence of structural relationships. The pattern of a covariation may suggest the nature of an interaction. For example, if two positions in a sequence covary strictly according to the Watson-Crick rules (and cover all the allowable combinations), one can be fairly certain that they form a bona fide Watson-Crick base pair. On the other hand, a normal pairing geometry is unlikely in those cases exhibiting other than a canonical pattern of covariation. The pattern and frequency of variation at a given position is a far more subtle characteristic of that position than is generally appreciated. It is a refined measure of structural homology, for only if the full structural context, not merely the major contacts, is maintained from one group of organisms to the next, will the pattern (and frequency) of variation be similar in the groups.

Some Specific Examples Involving Ribosomal RNA

A principal use of comparative analysis in the study of rRNA has been to infer secondary structure. As used here, secondary structural elements are contiguous stretches of two or more nucleotides that form canonical (or G:U type) antiparallel pairs with one another. Comparative "proof" of such structures involves finding multiple phylogenetically independent instances in which the compositions of two potentially paired positions change in concert, according to the Watson-Crick pairing rules. Figure 1 gives some idea of what comparative analysis has accomplished in terms of rRNA structure, and the extent of the comparative evidence that supports the various helices (Gutell et al. 1993). About 60% of the nucleotides in the small subunit rRNA are involved in confirmed secondary structure, a proportion comparable to the 55% of nucleotides in tRNA that are known from its crystal structure to be involved in secondary structure (Kim 1979).

Comparative analysis shows that noncanonical pairs occur frequently in rRNAs, and that they occur in a variety of pairing geometries. The U:G type of pair is, of course, the most common of the noncanonical pairs, the next most common being the A:G type (Gutell et al. 1993). From their patterns of variation it is apparent that rRNA contains several different kinds of U:G pairs. One kind exhibits frequent compositional variation over the phylogenetic spectrum, but U:G remains its major composition. Perhaps the most striking example of this kind can be seen in the helix 829-40:846-857 shown in Figure 1. This structure, which typically comprises 10–12 pairs, almost always shows 5 or more U:G pairs in its central section. The helix is highly variable in composition, with U:G pairs often being replaced by other pairs (frequently G:U), and the exact location of the U:G pairs is somewhat variable (Gutell et al. 1993). It seems evident that at least some U:G pairs in rRNA have unique structural or functional significance.



Escherichia coli

Domain: *Bacteria*
Kingdom: *Purple Bacteria*
Order: *gamma*

Figure 1 Higher-order structure of the 16S ribosomal RNA (Gutell 1994). As discussed in the text, this folded structure is based on comparative analysis. Watson-Crick pairs are connected with lines, G:U pairs are connected with filled circles, and A:G pairs are connected with open circles. Bases that are nonjuxtaposed as pairs, but connected, have yet to be proven phylogenetically. Boxed regions pair as indicated by the connecting lines. Every 10th position is indicated with a line and each 50th position is numbered. (Reprinted, with permission, from Gutell at http://pundit.colorado.edu.8080/)

A second kind of U:G pair is identified by the fact that its composition varies only slowly over evolutionary time, and when it does so, it almost always (or always) converts to a C:A pair (Gutell et al. 1993). This pattern of variation suggests a non-Watson-Crick pairing geometry.

Another interesting pattern of covariation is shown by the U:U pair that converts solely to C:C (and vice versa), again suggestive of an atypical pairing geometry (Gutell et al. 1993). Although the pattern of covariation does not allow us to infer the physical conformation of any pair with certainty, it does favor certain possibilities. For example, a U:U pairing that alternates (and is presumably isomorphic) with a C:C pairing could involve nucleotides in a syn-anti conformation. Direct physical measurement is needed to ascertain this.

One of the first unusual higher-order structures in the small subunit rRNA to be uncovered by comparative analysis was the so-called pseudoknot (Woese et al. 1983). Pseudoknot here describes a topology in which a stretch of nucleotides within a hairpin loop pairs with nucleotides external to that loop. Three such pseudoknots involve the pairing of positions 17-20:915-918, 505-507:524-526, and 570-571:865-866 (Fig. 1). In all three instances, the pseudoknot pairings tend to be nearly invariant in composition, suggesting that their three-dimensional context is stringently defined and that these structures are important to ribosome structure/function. Some of the proposed pseudoknot structures have been experimentally tested by mutagenesis: In all cases in which mispairings have been introduced, ribosome function was impaired, if not totally eliminated (Powers and Noller 1991).

Comparative analysis also suggested that the area of the small subunit rRNA between positions 500 and 545 (which contains the second of the three pseudoknot structures) is of major importance because of its highly conserved sequence and overall length (Woese et al. 1983). This particular structure is highly constrained: Of the 46 bases comprising it, only 14 are not known to be involved in secondary or tertiary interactions (see Fig. 1). Mutational analysis and rRNA probing studies show this region to be of functional significance (Noller et al. 1990; Powers and Noller 1991). The fact that the region's known structure is entirely self-contained (confined to these 46 nucleotides) makes it an excellent candidate for isolation and characterization by physical methods.

Another important RNA structural element identified by rRNA sequence comparisons is the so-called tetraloop (Woese et al. 1990b), a loop of four nucleotides that determines (caps) a double-helical stem of two or more base pairs. Such structures account for the majority of all hairpin loops in rRNA (Gutell et al. 1993). The most interesting characteristic of tetraloops is that the sequence in the loop is highly constrained. The naturally occurring examples are confined to an extremely limited subset of the 256 possible permutations of the four nucleotides. The compositions of the first and last bases in the loop are strongly coordinated, and to a lesser extent, the compositions of the inner two bases are as well. Three motifs cover the overwhelming majority of the naturally occurring tetraloops in rRNA; UNCG, CUYG, and GMRA (Woese et al. 1990b). In the small subunit of rRNA, the predominant compositions for each of the three major types are UUCG, CUUG, and GMAA (Woese et al. 1990b). The composition of the terminal base pair in the underlying stem is related to the sequence in the loop, especially for two of the three loop compositions: The UUCG loop strongly favors a C:G closure; CUUG loops are usually closed by a G:C pair, and the closing pair for the GCAA loop, although less constrained, tends to be R:Y (Woese et al. 1990b).

Some of the tetraloops in the small subunit rRNA are invariant or change only slowly in sequence over evolutionary time, whereas others change with remarkable frequency. In both cases, the variation is far from random, with the vast majority of variants tending to conform to one of the above three general types. The most frequently changing of the small subunit rRNA tetraloops is that located at positions 83-86 (Fig. 1). More than 50 phylogenetically independent alterations in loop sequence have been recorded among the Bacteria. The loop has a UUCG, CUUG, or GCAA sequence in 93% of cases (Woese et al. 1990b). Of the remaining 7%, most are themselves tetraloops whose compositions are related to one of the above three. In only 3% or so of cases does the size of the loop vary; and then, by addition or deletion of a single nucleotide (Woese et al. 1990b). The UUCG loop at position 83-86 shows a C:G closing pair 91% of the time, the CUUG loop here closes with a G:C pair in 95% of cases, and 86% of the GCAA loop closures have an R/Y composition (mainly A:U). Solution NMR spectroscopic analysis of two of the characteristic tetraloops, C(UUCG)G and C(GMAA)G (bases outside the parentheses form the closing pair), suggest structures in which the two terminal bases in the loop itself interact to form an unusual, noncanonical pair, which would seem to explain their rather strict covariation (Heus and Pardi 1991; Varani et al. 1991).

Tetraloops of the types described above are not confined to rRNAs; they clearly serve various other functions in the cell. They have been reported to occur in the context of controlling gene expression in T4 phage (Tuerk et al. 1988). Covariation analyses of the group 1 intron (Michel and Westhof 1990) and of ribonuclease P (Brown et al. 1996; Tanner and Cech 1995; Massire et al. 1997) also show that GNRA tetraloops are involved in long-range tertiary interactions. Comparative results are bolstered by a wealth of biochemical data (Jaeger et al. 1994; Murphy and Cech 1994; Pley et al. 1994; Abramovitz and Pyle 1997), and recently the crystal structure of a portion of the group 1 intron provided a direct visualization of such an interaction (Cate et al. 1996). The tetraloop/receptor helix docking interaction is now considered a common structural motif in large RNAs (Costa and Michel 1995; Abramovitz and Pyle 1997).

The Scope and Utility of Comparative Analysis

Comparative sequence analysis utilizes a highly abstract presentation of a molecule, a mere string of symbols representing the four nucleotides. In its own right, comparative analysis reveals only patterns involving these symbols; it says nothing about actual molecular structure (or function). However, when combined with structural knowledge derived from the more direct physical methods of measurement, comparative analysis takes on physical meaning, and so, becomes a more powerful and useful approach to analysis of structure. The potential of this link between comparative analysis and actual structure determination should be fully appreciated and exploited. When the physical chemists pay attention to the results of comparative analysis, it helps them to uncover biologically important and physically interesting molecular structures. What is not generally recognized is that as the physicist's repertoire of biologically important structural motifs increases, the potential for comparative analysis of primary structure to reveal the existence of physical structure increases proportionately. Thus, comparative analysis is increasingly useful as an adjunct to physical and chemical determinations of molecular structure.

One example of the interplay between physical characterizations and comparative analysis is given by work on G:A-type pairs. There are many instances in which otherwise canonical helices contain GpA doublets paired (antiparallel) with GpA, but never ApG paired with ApG. Such a bias among natural RNA indicates a profound difference between the two types of pairing. Based on these comparative data, Turner, Wilson, and their respective colleagues have examined thermodynamic and structural properties of RNA and DNA oligonucleotide duplexes containing the two types of paired doublets (SantaLucia et al. 1990; Li et al. 1991). They found that pairs formed from GpA are as energetically favorable as canonical pairs would be, but that the ApG pairing is unstable. Spectroscopic analysis revealed, moreover, that the more stable GpA pairing has a novel (non-Watson-Crick) geometry.

In a similar way, comparative analysis can serve as a guide in molecular genetic analyses. Much of the molecular biologist's approach to understanding and utilizing molecular structure and function turns upon manipulation of genetic sequences. Since sequence space is enormous, there is no way the molecular biologist can efficiently explore it without some kind of map. Taking random mutational "shots" at the ribosome, for example, is a most unproductive way to attempt to uncover the molecular basis for ribosome function. Comparative analysis can provide an initial guide and give essential clues that turn an intractable task into a manageable one. The following are some examples.

A common need in the study of macromolecules is the identification of features responsible for any activity; for instance, the structural elements involved in catalysis by ribozymes. Sequences responsible for the catalytic activity of the so-called hammerhead self-cleaving RNA (see Chapter 12) could be identified by their conservation throughout a collection of catalytic "satellite" RNAs (and their complements) with otherwise extremely variable sequences (Forster and Symons 1987a). Deletion analysis then demonstrated that the catalytic structure could be reduced to the only common structure in all the RNAs, about 50 nucleotides (Forster and Symons 1987b). Similar sequence comparisons coupled with deletion analyses have revealed functional aspects of self-splicing introns (see Chapter 13).

A more complex example of the identification of minimum catalytic RNA structure is that of ribonuclease P RNA. RNase P is the enzyme that cleaves 5′-leader sequences from precursor forms of tRNA in all cells. RNase P occurs as a complex holoenzyme in vivo, a protein–RNA complex. In vitro, however, at high ionic strength, the bacterial RNA is capable of catalyzing the reaction independently of the protein (Guerrier-Takada et al. 1983); thus, bacterial RNase P is a ribozyme. In contrast, the archaeal and eucaryal RNase P RNAs are inactive in the absence of the protein constituents of the holoenzyme. Substantial variation in sequence and length of RNase P RNAs from different organisms made alignment and comparative structure analysis difficult, but the variations are now reconciled in the context of a universally applicable RNase P RNA secondary structure (Fig. 2). The core bacterial secondary structural elements and several key base identities are also conserved in Archaea and Eucarya. This extent of conservation indicates that the archaeal and eucaryal versions of the RNA remain the catalytic center of RNase P, even though they no longer function independently of protein.

The approximately 130 diverse sequences of bacterial RNA that are now available constitute a comprehensive mutational survey of this cat-

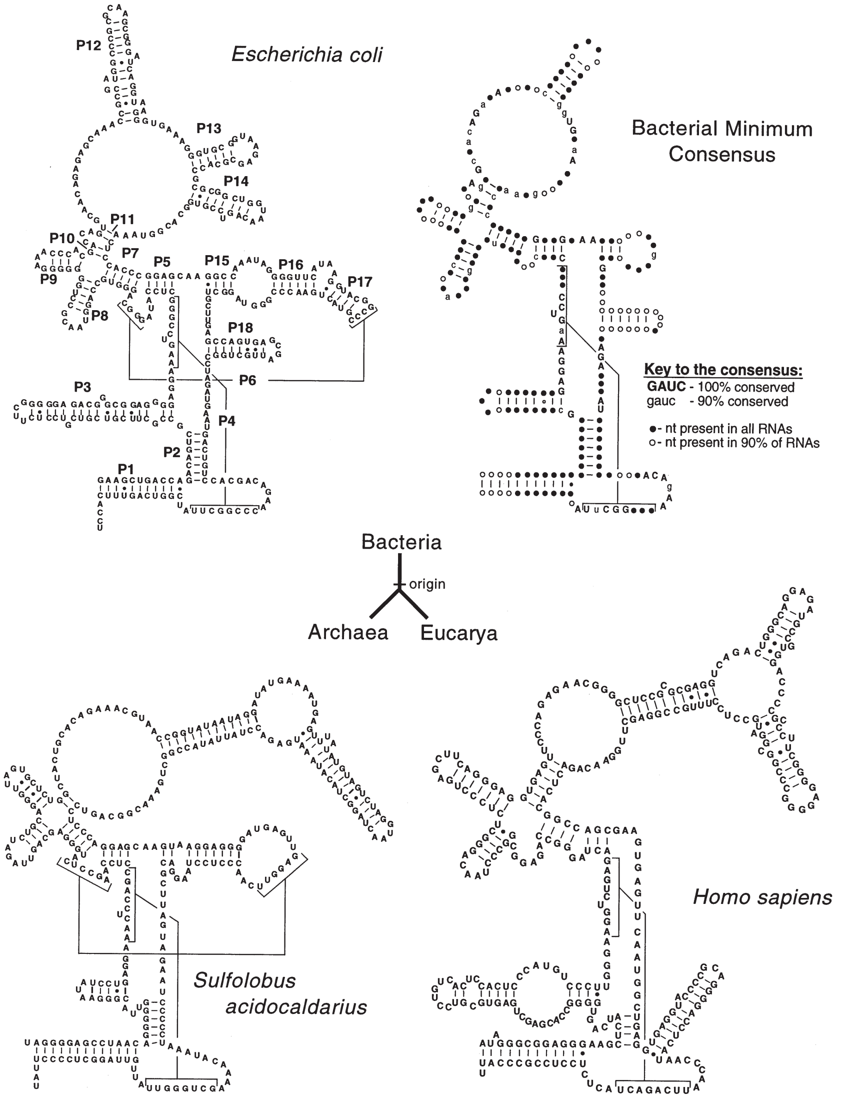


Figure 2 The universality of the core structure of RNase P RNA. Representatives of RNase P RNAs from the three domains of life are shown juxtaposed to a bacterial minimum consensus secondary structure, as described in the text. Helices are numbered for the *E. coli* RNA from 5′ to 3′ and designated with P ("pairing," e.g., P1, P2). Watson-Crick base pairs are represented with a line, and noncanonical base pairs are indicated with a full circle. The phylogenetic relationships of the organisms corresponding to the RNase P RNAs are shown in the central three-domain tree.

alytic RNA. The results are summarized in a "phylogenetic minimum consensus" structure of the bacterial RNase P RNA (Fig. 2), the minimum sequence and length present in each bacterial RNase P RNA. All RNase P RNAs contain sequence-length not present in some other instance of the RNA, so the length of the minimum consensus is only one-half to two-thirds that of the typical native RNA. Nonetheless, since all bacterial RNase P RNAs contain the minimum-consensus structure, it is expected that this structure should contain all the elements required for catalytic activity.

In constructing a phylogenetic-minimum RNase P RNA to test this notion, sequences in one type of the RNA were replaced with the shorter, corresponding sequences from the RNAs of other organisms. An early design was based on the minimum homologous sequence-lengths of *E. coli* and *Bacillus megaterium* RNase P RNAs, and resulted in Min 1 RNA, a highly active, 263-nucleotide RNA. When the RNase P RNA sequence from *Mycoplasma fermentans* became available, the size of the minimum RNase P RNA could be reduced even further. The RNase P RNA subunit from *M. fermentans*, at 276 nucleotides, is the smallest known RNase P RNA to date. This RNA, and those of *Chlorobium limicola* and *Mycoplasma hyopneumonia*, collectively, lack helices previously thought to be universally conserved in bacterial RNase P RNAs (Siegel et al. 1996). Assembly of minimum conserved sequence-lengths resulted in a new RNA, termed Micro P, shown in Figure 3. This 211-nucleotide RNA represents the phylogenetic-minimum RNase P RNA, a structure using only those sequences or secondary structural elements that are present in all RNase P RNAs (Fig. 3). The properties of Micro P RNA are similar to those of Min 1 RNA: It is highly active at a high ionic strength and in the presence of high concentrations of magnesium. A requirement for higher ionic strength than required by the native RNAs reflects a global destabilization of structure by removal of the peripheral helices. Nonetheless, the simplified, phylogenetic-minimum RNA retains all of the structural elements needed for catalysis. It is unlikely that a successful, piecemeal removal of nearly 50% of the molecular length of a native RNA could have been accomplished without approaching the problem from a comparative perspective.

Comparative Analysis as Interpretive Perspective

The interpretations of experimental results in molecular biology are often ambiguous because of the complexity of the systems and the consequent difficulty of differentiating meaningful from trivial information. Results

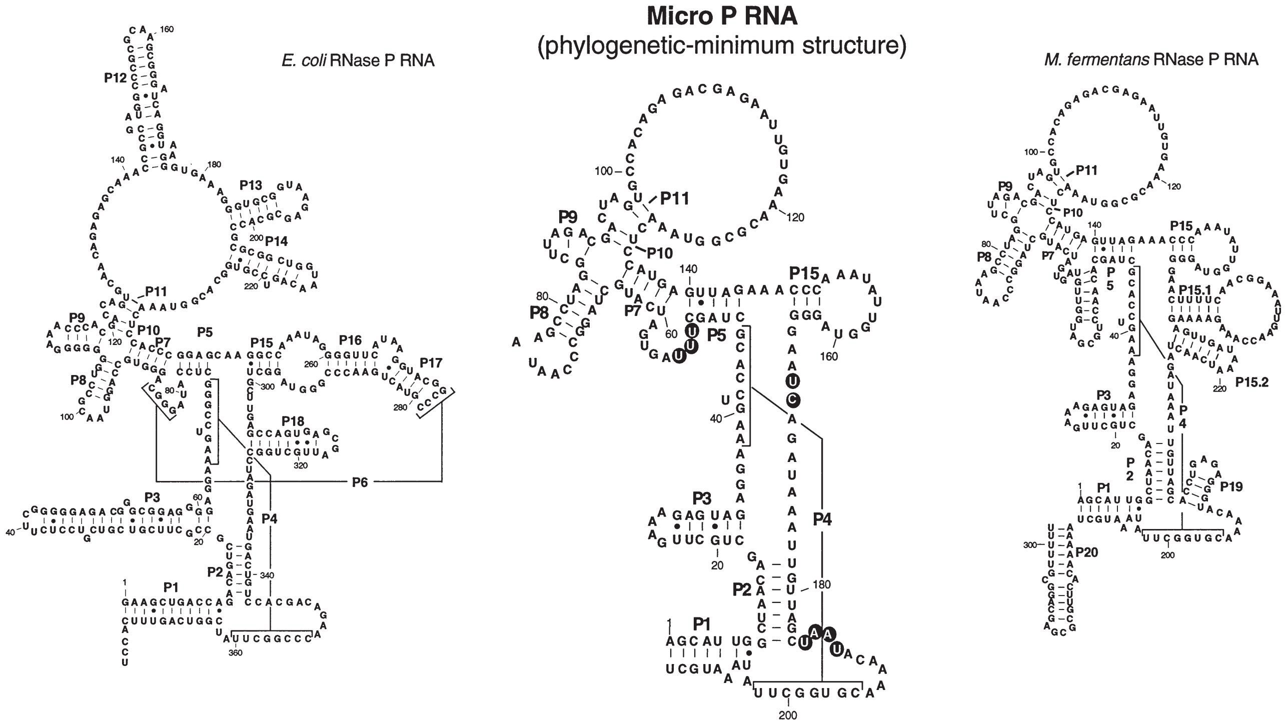


Figure 3 Design of Micro P—the smallest, functional RNase P RNA and a reflection of the phylogenetic-minimum bacterial consensus RNase P RNA. Secondary structures are shown for *E. coli* and *M. fermentans* RNase P RNAs, and the synthetic Micro P RNase P RNA (see text). Filled circles in the Micro P RNA identify nucleotides used to replace helices in the *M. fermentans* RNase P RNA in the design of the Micro P RNA. Base pairs in helix P4 and P6 are indicated with connecting lines. Helix numbering is as in Fig. 2.

derived from chemical cross-linking, footprinting, and nuclease or chemical structure-mapping experiments, for example, contain a wealth of information. However, it is often impossible to distinguish completely the significant data from the background of trivial data (idiosyncratic for the particular organism or an artifact of the particular experimental method) that invariably accompanies them. Such data are commonly presented with little attempt to evaluate their meaning and significance. Comparative studies with homologous molecules from different organisms add depth and precision to the interpretation of these kinds of data. Comparisons identify which elements in the data set correspond to general characteristics, universal properties, at the same time pointing out the possibly trivial results, the idiosyncratic ones.

One example of using the comparative approach in this manner involves a cross-linking analysis to locate the active site of RNase P RNA (Burgin and Pace 1990). In the study, an arylazide photoaffinity cross-linking agent was attached to the 5′-terminal phosphate in tRNA, the phosphate that is acted upon by RNase P. Ultraviolet irradiation cross-links the substrate to RNase P RNA, and sites of cross-linking can be identified by primer extension analysis. Three different RNase P RNAs, from *E. coli, Bacillus subtilis*, and *Chromatium vinosum*, were used in the study. These three RNAs differ extensively in sequence and in the presence or absence of some structural elements. Several nucleotides in each RNA formed cross-links with the photoagent-containing substrate. However, only a subset of these cross-linked nucleotides was found to be common to all three tRNAs. This subset is distributed in the core of all RNase P RNAs known and subsequently was shown to comprise the heart of the RNase P catalytic center (Harris and Pace 1995; Kazantsev and Pace 1998).

Similar site-specific photoaffinity cross-linking methodology with a comparative perspective has been used to identify the global architecture of the RNase P RNA. Strategic placement of photoagents at various sites in the *E. coli* and *B. subtilis* RNase P RNAs, with cross-linking and primer extension analysis, localizes regions of the RNAs relative to the rest of the molecule, and to the substrate. Photoagents at homologous positions in the two RNAs resulted in nearly identical cross-linking patterns, consistent with the notion that these two RNAs contain a common, core tertiary architecture (Fig. 4A, B) (Chen et al. 1998). Molecular modeling of the library of distance constraints has resulted in essentially coincident tertiary structure representations of the RNase P RNA–tRNA complexes (Fig. 4) (Harris et al. 1994; Harris et al. 1997; Chen et al. 1998). The models consist of two juxtaposed, multi-helix domains with the highly con-

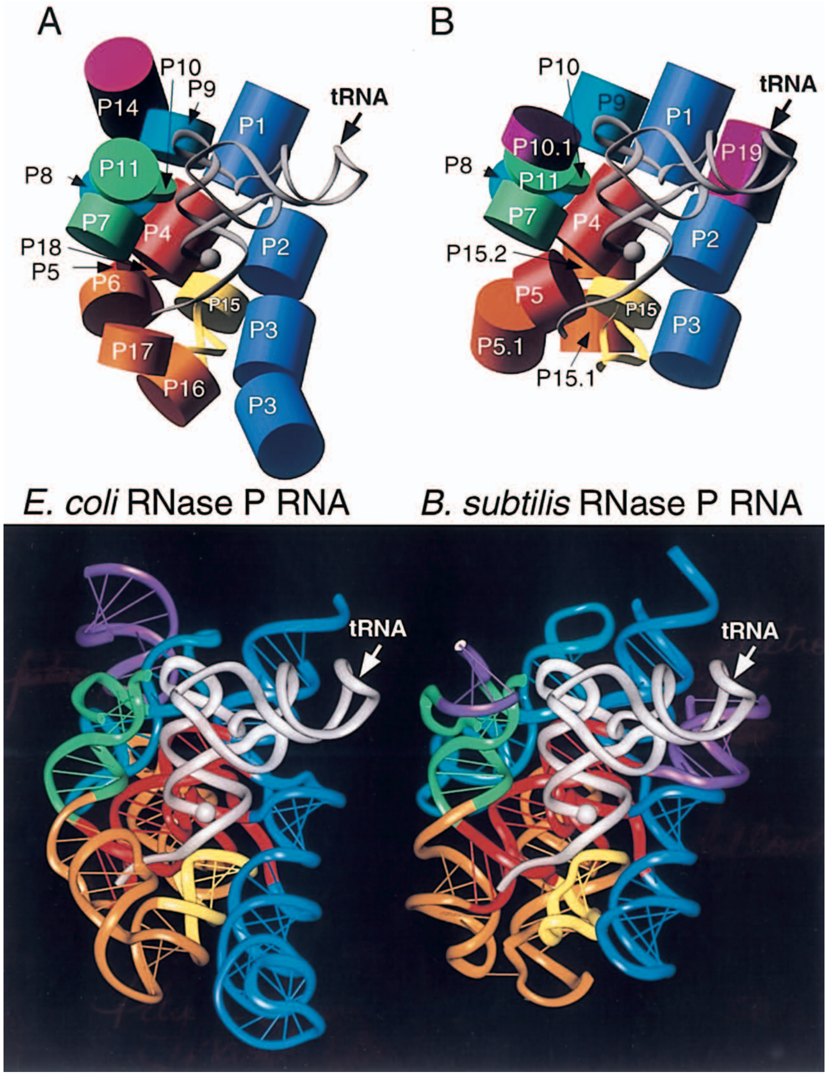


Figure 4 Tertiary structure models of *E. coli* (*A*) and *B. subtilis* (*B*) RNase P RNAs. The upper panel of the figure is a helix-barrel model where each cylindrical barrel presents the location of an A-form RNA helix of appropriate length. tRNA is displayed as a ribbon with a highlighted phosphate at the scissile bond. The lower panel is an extrapolation of the barrel model, showing the entire RNA backbone as a colored ribbon, with bars between ribbons idealizing base pairs. Helices P12, P13, and P14 in the *E. coli* model, and helices P10.1 and P12 in the *B. subtilis* model are not represented in this model due to the lack of experimental data available to position these conserved elements with the same degree of certainty as the rest of the helices shown in the model.

served regions of the molecules at the core of the models and the more variable helical elements near the periphery. Additionally, both models situate the scissile bond in a pre-tRNA immediately adjacent to the universally conserved P4 helix of the RNase P RNA. The agreement between the models of the two bacterial RNase P RNAs validates each of the proposals and testifies to the utility of combining phylogenetic-comparative analysis with biochemical experimentation.

RNA as Historical Record

The availability of rRNA sequences has allowed quantitative analysis of evolutionary relationships. Tapping the phylogenetic information in rRNA has had a revolutionary effect on microbiology, adding an evolutionary dimension where there had been none before. The prior lack of an evolutionary framework in microbiology was the unavoidable consequence of the fact that morphologies and physiologies of microorganisms are too simple and/or unpredictably variable to be of significant value in developing a natural classification. It is no wonder that, using those properties, microbiologists had been unable to produce a valid, natural (phylogenetic) classification from the bacteria. However, as Zuckerkandl and Pauling (1965) pointed out, at the molecular level there is no such problem; molecular sequences are historical records rich in readily interpretable geneaological information. Thus it was possible, by comparison of partial sequences of 16S rRNAs (oligonucleotide catalogs), to make initial sense of microbial genealogical relationships (Fox et al. 1980; Woese et al. 1985). The resulting phylogenetic framework, the natural classification, brings clarity and order to the day-to-day conduct of microbiology. Experimental progress is enhanced; deeper interpretations of results are now possible; new directions of research are indicated. Microbial ecology has come of age: The capacity to identify microorganisms phylogenetically using rRNA sequence-based techniques, even without cultivation (Pace et al. 1985; Pace 1997) and in situ (DeLong et al. 1989), gives microbial ecologists a power they have always lacked for comprehensive analysis of various niches. Not only can organisms be identified with greater precision and characterized in greater depth, but unculturable or previously unrecognized organisms in a particular niche can now be definitely related to those in other niches.

The early studies of microbial phylogeny based on rRNA sequences yielded the remarkable finding that the world of prokaryotes, which all biologists had taken to be phylogenetically unified (monophyletic), was indeed not so. There exist two distinct kinds of prokaryotes, now formally called the domains Archaea and Bacteria (Woese et al. 1990a), that are no more closely related to one another than either is to the eucaryotes. Although the early oligonucleotide cataloging approach could readily define and distinguish three primary groupings of organisms (Archaea, Bacteria, and Eucarya), it could not relate them to one another in any precise way. This joining had to await the technology that permitted determination of complete sequences of rRNAs (or their genes), which in turn allow construction of a universal phylogenetic tree. This tree is shown in Figure 5 (Pace 1997); its root has been inferred by the so-called Dayhoff strategy. In this method, an uprooted tree generated from a set of (homologous) sequences is rooted using a related (or paralogous) sequence. To root a tree that spans all extant life, it is necessary that the gene duplication that originally produced the paralogous genes occurred in the ancestral stem, prior to the initial phylogenetic radiations. The root of the universal tree is seen to lie between the Bacteria on the one hand and the common lineage that the Archaea and Eucarya share, on the other (Gogarten et al. 1989; Iwabe et al. 1989). The universal tree, in other words, predicts that the Archaea are specific relatives of the Eucarya (albeit at a deep level), to the exclusion of the Bacteria. This assessment is now borne out by many macromolecular sequences derived from the accumulating sequences of archaeal genomes (Olsen and Woese 1997). Since there exist characteristic archaeal, bacterial, and eucaryal versions for just about every universal macromolecular function so far characterized in the cell, there can be no doubt that all life on this planet is organized into three, very distinctive groupings. At the levels of the nucleic acid-based information transfer machinery, Archaea and Eucarya resemble one another more closely than either resembles Bacteria.

The Archaea

From a phenotypic perspective, based on relatively few instances of cultivated organisms, the Archaea would seem to be a rather strange and disparate collection. Unlike the Bacteria, they show only a few major phenotypes: the methanogenic; the extremely halophilic; the sulfate-reducing; and the sulfur-“dependent,” extremely thermophilic phenotype (Woese 1987; Winker and Woese 1991). These four phenotypes are sufficiently dissimilar that, although they were known (except for the sulfate-reducing phenotype) before the Archaea were recognized as a phylogenetic unit, their relationship to one another was not recognized (this despite the existence of some molecular evidence suggestive of the relationship; e.g., the unusual archaeal ether-linked lipids [Winker and Woese 1991]). The four main archaeal phenotypes do not define four major taxa of equivalent taxonomic rank, however. As shown in Figure 6, cultivated Archaea comprise two major branches, two "kingdoms," the Euryarchaeota and the Crenarchaeota (Woese et al. 1990a). The Euryarchaeota encompass examples of all four archaeal phenotypes, with methanogens dominating the group. The methanogens comprise three major groups in addition to a very deeply branching lineage represented by the genus

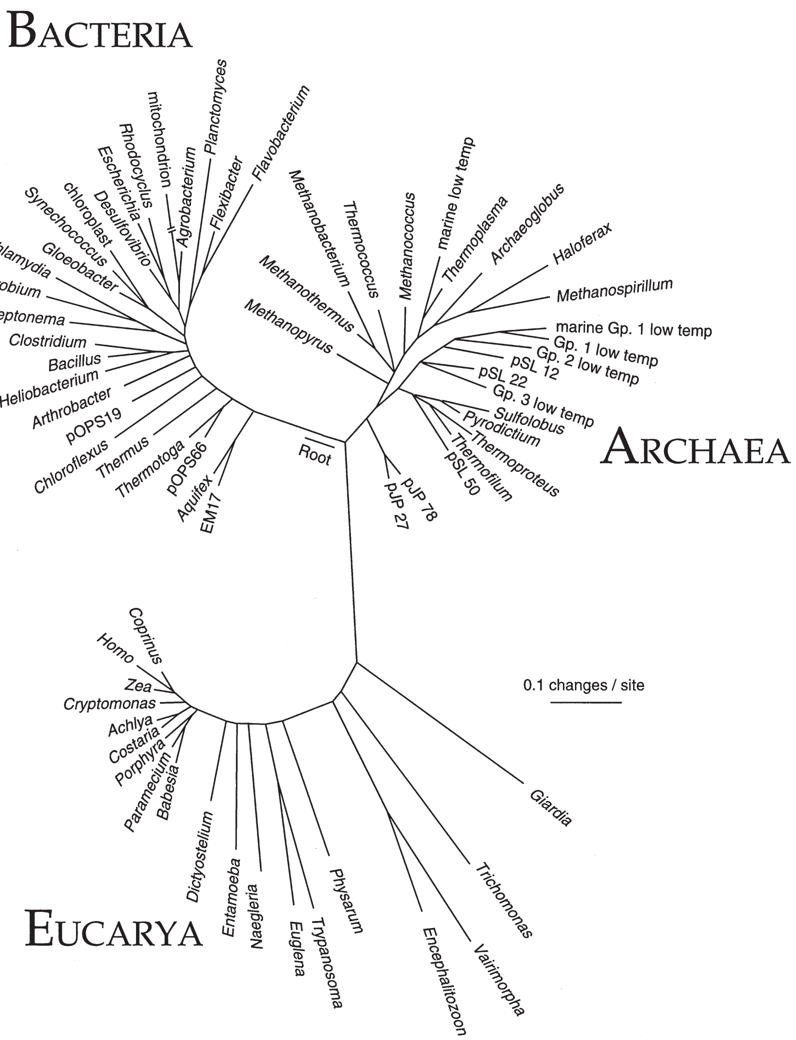


Figure 5 Rooted universal phylogenetic tree based on SSU rRNA sequences. The tree is based on maximum likelihood analysis of 64 rRNA sequences. (Reprinted, with permission, from Pace 1997 [copyright American Association for the Advancement of Science].)

Methanopyrus (Burggraf et al. 1991). Two of the three main methanogen lineages are phenotypically uniform, but the third, the *Methanomicrobiales* lineage, has spawned other phenotypes as well: the extreme halophiles, the sulfate reducers, and perhaps *Thermoplasma*. The remaining euryarchaeal lineage is the phylogenetically compact cluster of species that constitute the *Thermococcales*, a group phenotypically resembling the crenarchaeotes.

Our view of the potential phenotypic diversity of Crenarchaeota has recently expanded substantially with the development of methods for detecting uncultured organisms. These methods are PCR coupled with cloning to obtain naturally occurring rRNA genes and thereby to detect otherwise unknown organisms. Prior to the use of such methods to explore environmental Archaea, the taxon Crenarchaeota was considered phenotypically uniform; all of the sulfur-dependent, thermophilic type. With the application of molecular methods to the study of environmental organisms, however, numerous representatives of mesophilic Crenarchaeota recently have been discovered in marine (DeLong 1992; Fuhrman et al. 1992) and terrestrial (Ueda et al. 1995; Hershberger et al. 1996) environments. None of the low-temperature crenarchaeotes has yet been cultivated, so their metabolic basis remains unknown. The environmental surveys of Archaea additionally have revealed many thermophilic types of Crenarchaeota only distantly related to known organisms, including a third primary branch, Korarchaeota (Barns et al. 1996).

The distribution of phenotypes on the tree of Figure 6 strongly suggests the Archaea to be of thermophilic origin. All cultivated crenarchaeal species are thermophilic, some growing optimally at temperatures above 100°C; all the mesophilic Crenarchaeota originated from lineages that ancestrally were thermophilic. Additionally, the deepest branchings on the euryarchaeal side, the *Thermococcales* and the genus *Methanopyrus*, are thermophilic as well. This is true for the deepest branchings within all the major euryarchaeal sublineages. For the most part, Archaea are anaerobic; ability to grow aerobically, when it occurs, is generally facultative. Similarly, the Archaea are most often chemoautotrophic, particularly the deeply branching lineages. Thermophily, anaerobic growth, and chemoautotrophy, then, can be considered ancestral characteristics of the Archaea.

The surprising and scientifically inviting relationship of the Archaea to the Eucarya (Fig. 5) deserves comment. The sequences of many, but not all, archaeal genes resemble their eucaryotic homologs decidedly more than their bacterial homologs. Ribosomal protein sequences, for instance, are this way (Auer et al. 1989; Ramirez et al. 1989), as are the major subunits of the archaeal RNA polymerase (Pühler et al. 1989), the archaeal

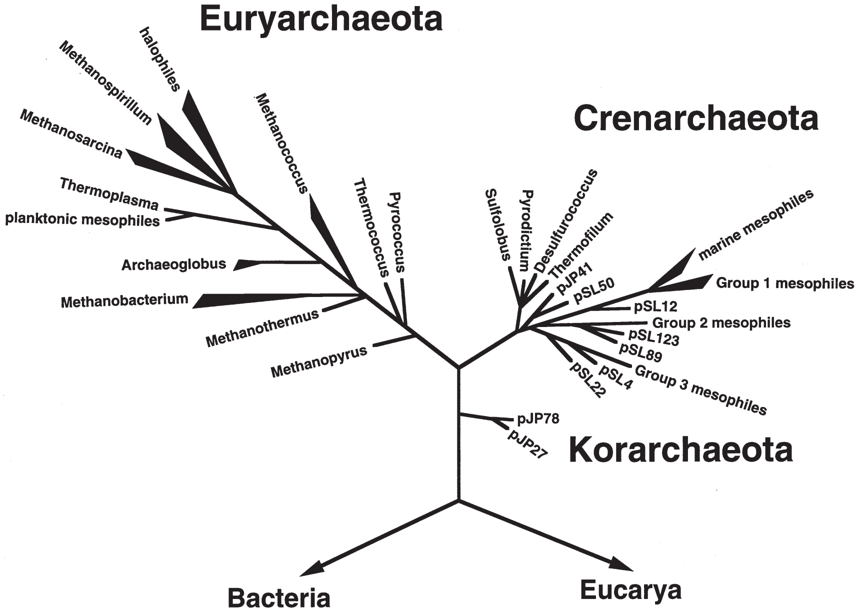


Figure 6 Phylogenetic tree of Archaea. The tree is based on maximum likelihood analysis of selected SSU rRNA sequences. (Figure provided by Scott C. Dawson.)

histone (Sandman et al. 1990), and a chaperone (Trent et al. 1991). Although the archaeal rRNA overall is an exception to this rule (in sequence and secondary structure it most resembles the bacterial type), the most highly conserved positions in the archaeal sequence do show a slight eucaryotic bias (Winker and Woese 1991).

Two examples of the above relationships between archaeal and eucaryal proteins are of particular interest, namely, RNA polymerase and histone. The archaeal RNA polymerase is more similar in sequence to both of the eucaryal polymerases II and III than these two are to each other (Pühler et al. 1989). Moreover, each of the ancestral sequence duplications that make up the archaeal TATA-binding protein are more like both of the corresponding human TATA-binding protein sequences than the human sequences are like one another. The case of the archaeal histone is even more striking. Its sequence is reported to be closer to the sequences of each of the four eucaryal histones, H2a, H2b, H3, and H4, than any of these four is to one another (Sandman et al. 1990). In other words, in these examples, the archaeal version of a molecular type appears to resemble the inferred common ancestor of a eucaryal family of proteins more closely than do any of the extant members of that eucaryal family.

Is this perhaps the tip of an iceberg? Do many families of eucaryotic genes have an archaeal homolog that resembles the (inferred) ancestor of that family more than its extant representatives do? Perhaps we can identify human gene families more readily by comparison with archaeal genes than by comparison of the members of the human gene families.

The Bacteria

Figure 7 shows a bacterial phylogenetic tree inferred from small subunit rRNA sequences. Anyone familiar with classic bacterial taxonomy will see immediately that the groupings defined by molecular sequence analysis bear little relationship to the groupings defined by classic methods (Woese 1987). For instance, microbiologists previously grouped all pho-

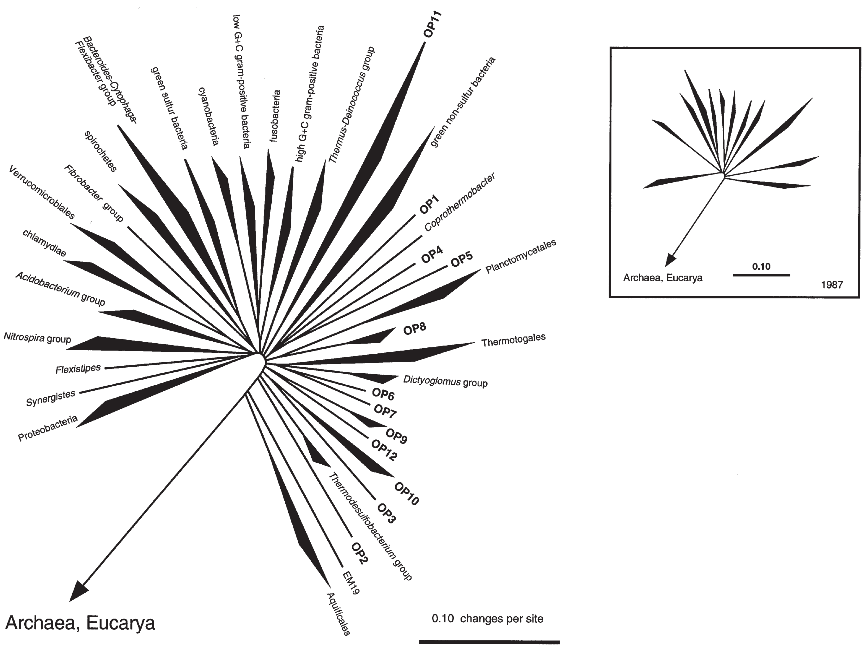


Figure 7 Phylogenetic tree of the Bacteria. The 36 division-level clades of Bacteria are depicted (Hugenholtz et al. 1998). The inset represents the known phylogenetic span of Bacteria in 1987 (Woese 1987), showing the 12 division-level clades known at the time. Filled sectors indicate that several representative sequences fall within the indicated depth of branching. Lines designated by OP represent one or more phylotypes that were identified in Obsidian Pool by means of molecular methods but have not been cultivated. (Reprinted, with permission, from Hugenholtz et al. 1998.)

totrophs into a single taxon, which contained few if any nonphotosynthetic species. In actuality, photosynthetic and nonphotosynthetic phenotypes are often intimately intermixed. Microbiologists formerly used morphology as a primary determinant of classification. Morphology turned out to be an extremely poor indicator of bacterial phylogenetic relationships, although there are a few notable exceptions, such as the spirochetes and the endospore-formers. Properties such as gliding motility, formerly used to group organisms into a few high-level taxa, also are phylogenetically widely dispersed. Gliding organisms, for instance, are intermixed with flagellated forms. The mycoplasmas, rickettsias, and a variety of other pathogens do not warrant the high-level taxonomic distinction previously accorded them on the basis of their parasitic character.

Our understanding of the breadth of bacterial diversity is rapidly changing due to molecular studies of rRNA sequences obtained from natural environments without cultivation. Of the approximately 40 bacterial division-level groups shown in Figure 7, representatives of only about 60% of them so far have been cultivated. Remarkably, since 1987 (Woese 1987), the number of recognized phylogenetic divisions has tripled (Pace 1997).

The distribution of phenotypes in the bacterial tree suggests that the Bacteria, like the Archaea, arose from a thermophilic ancestor: Thermophily is widespread in the bacterial tree, dominant in its deeper branches (Achenbach-Richter et al. 1987; Woese 1987). Moreover, no evidence exists to support the old notion that life began heterotrophically and that, consequently, the first organisms were heterotrophs; if anything, bacterial phylogeny is more consistent with an aboriginal autotrophy (Woese 1987). The fact that the most deeply divergent bacterial and archaeal lineages are thermophilic and chemoautotrophic indicates that the common ancestor of all life was of that nature, thermophilic and chemoautotrophic.

The Eucarya

The textbook picture of eucaryotes emphasizes four great “kingdoms”: animals, plants, fungi, and protists. rRNA sequences tell a rather different story; see Figure 5. The animal, green plant, and fungal kingdoms are seen from that perspective to constitute some of the more superficial branchings on the tree, whereas the protist kingdom is a polyphyletic collection of (little-related) lineages, which together cover the full span of the eucaryal tree. Since mitochondria appear only among the higher branchings in the eucaryal tree, it seems likely that the aboriginal eucaryotic cell possessed none, a conclusion consistent with the fact that the earlier branchings on the tree presumably arose at a time when there was little or no free oxygen on this planet, prior to 2–2.5 billion years ago. Moreover, the eucaryal lineage originated before the ancestors of mitochondria, the α-purple Bacteria and relatives (Yang et al. 1985).

So far, the lower eucaryal branchings are defined only by mesophilic organisms, primarily parasitic species. These offer few clues as to the conditions under which their presumed free-living ancestors flourished, during the early, anaerobic phase in Earth history. Thus, the question of whether the eucaryotes, like the prokaryotes, arose from thermophilic ancestry cannot be addressed.

The Universal Ancestor

Biologists have long believed that all life ultimately arose from a common ancestor. The Universal Ancestor that biologists originally pictured was as simple as possible; it even lacked intermediary metabolism (Oparin 1964). It is only today, with a universal phylogenetic tree and a wealth of diverse sequence data, that we can begin to construct a realistic picture of this entity, one that can be experimentally tested and refined. The Universal Ancestor that we abstract from the genome sequence data is anything but simple, however. If properties that are present in all the modern domains are also in the Universal Ancestor, then the Ancestor would appear highly developed and metabolically rich, basically a modern cell. How else can we explain the spread of so many cellular functions, especially metabolic genes, across the full phylogenetic spectrum?

Yet, the idea that the Ancestor was basically a richly endowed, modern type of cell is not in keeping with the manner in which it appears to have evolved: that is, much more rapidly than have modern cells and in far more drastic ways. This is what would be expected for entities far simpler than modern cells, not ones equally or more complex (Woese 1987). Here, then, is a paradox, one that any consistent theory of the Universal Ancestor must resolve.

Life seems to have started in an RNA World—or at least in a world where polynucleotides played a far more prevalent role than they now do. The dynamic by which such simple living systems evolved must have been unique. It seems logical that the peculiar evolutionary dynamic that characterized the Universal Ancestor and its descent into the primary lineages reflects this earlier evolutionary world, when the cell itself, its basic mechanisms, were still in the throes of developing (Woese 1982, 1998). We suggest that the horizontal (lateral) gene transfer seen in the "modern" world is, as it were, a "background radiation" left over from an era when horizontal, not vertical, gene transfer dominated and defined the evolutionary dynamic (Woese 1998).

In the era of the Universal Ancestor, cells would have been simple and relatively ill-defined—in the number and kind of functions they possessed, in the size of their genomes, in the simplicity and imprecision of their information processing systems, in the sizes and types of proteins they possessed, and in their overall organization (Woese 1967, 1998; Woese and Fox 1977; Woese et al. 1983). Mutation rates were extraordinarily high. Lateral gene transfer was pervasive and pandemic; it applied to all genes and all cellular entities. In a sense there was universal genetic cross-talk among primitive cellular entities. They had yet to evolve to the stage where they were truly idiosyncratic in character. All shared a common evolutionary problem, the evolution of the basic cellular machinery; and its solution was a collective one. Any innovation that occurred in one cell line would readily be shared with other cell lines through lateral gene transfer (Woese 1998). It was in this way, not through vertical inheritance per se, that the basic cellular mechanisms evolved.

The primitive entities, called "progenotes" (Woese and Fox 1977), are taken to differ from one another metabolically. No one of them was sufficiently complex genetically to support a full metabolic capacity. Individual cells each with a specialized metabolic capacity could, however, form a loose-knit community that collectively was metabolically rich. It is such a diverse cellular community, in which individual cells (cell lines) communicated metabolically and genetically, that was the Universal Ancestor—not some particular organism or organismal lineage (Woese 1982, 1998). Modern analogs of this state can be seen in microbial consortia built on interdependent metabolisms.

When levels of lateral gene transfer are sufficiently high, organismal lineages cannot exist. Thus, at the stage of the Universal Ancestor there might have been relatively short-lived cell lines, but no true long-term organismal lineages (Woese 1998). The history of the Universal Ancestor was physical, not genealogical (Woese 1998). Only as cells and their componentry became more complex, idiosyncratic, and integrated did true lineages develop and become relatively refractory to the pervasive lateral gene transfer of the time (Woese 1998). From this point on, the cell (or subsystem) evolves vertically, i.e., has a genealogical history. This development resulted in the establishment of genealogical histories. Not all components in the cell became refractory to lateral gene transfer at the same stage; indeed the genes for some of them, e.g., the aminoacyl-tRNA synthetases and various metabolic enzymes, seem readily transferred laterally even today. However, complex componentries that are tightly integrated into the cell, such as the ribosome and the transcription apparatus, would have been among the first to become refractory to lateral gene transfer (Woese 1998).

One consequence of the early massive lateral gene transfer is that the universal rRNA tree is not a normal organismal tree. Its root and primary branchings are from a time when only a few of the cellular functions were refractory to lateral gene transfer, too few to constitute an organismal lineage. The rRNA tree grew only as increasing numbers of cellular functions became more or less refractory to lateral gene transfer. The rRNA tree became a true organismal one only in its more peripheral branches, as more and more components of the cell became refractory to lateral gene transfer. Thus, the universal tree does not start with an organism of the modern type, representing a specific modern cell. It starts at an earlier stage, near the end of the era of the Universal Ancestor (Woese 1998). This means, fortunately, that the tree provides an evolutionary framework that takes us back into the era when cells were still evolving.

ACKNOWLEDGMENTS

The authors' research activities are supported by grants from the National Science Foundation and the National Aeronautics and Space Agency (C.R.W.), and the National Institutes of Health (B.C.T. and N.R.P.). We thank Drs. Robin R. Gutell, Phil Hugenholtz, and Scott C. Dawson for figures.

REFERENCES

Abramovitz D.L. and Pyle A.M. 1997. Remarkable morphological variability of a common RNA folding motif: The GNRA tetraloop-receptor interaction. *J. Mol. Biol.* **266:** 493–506.

Achenbach-Richter L, Gupta R, Stetter K.O., and Woese CR. 1987. Were the original eubacteria thermophiles? *Syst. Appl. Microbiol.* **9:** 34–39.

Auer J, Lecher, K., and Böck A. 1989. Gene organization and structure of two transcriptional units from *Methanococcus* coding for ribosomal proteins and elongation factors. *Can. J. Microbiol*. **35:** 200–204.

Barns S.M., Delwiche C.F., Palmer J.D., and Pace N.R. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci.* **93:** 9188–9193.

Brosius J., Dull, T.J., and Noller, H.P. 1980. Complete nucleotide sequence of a 23S ribosomal RNA gene from *Escherichia coli. Proc. Natl. Acad. Sci.* **77:** 201–204.

Brown J.W., Nolan J.M., Haas E.S., Rubio M.A.T., Major F., and Pace N.R. 1996. Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Acad. Sci.* **93:** 3001–3006.

Burggraf S., Stetter K.O., Rouviere P., and Woese C.R. 1991. *Methanopyrus kandleri*: An archaeal methanogen unrelated to all other known methanogens. *Syst. Appl. Microbiol.* **14:** 346–351.

Burgin A. and Pace N.R. 1990. Mapping the active site of ribonuclease P RNA using a substrate containing a photoaffinity agent. *EMBO J.* **9:** 4111–4118.

Cate J.H., Gooding A.R., Podell E., Zhou K., Golden B.L., Kundrot C.E., Cech T.R., and Doudna J.A. 1996. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science* **273:** 1678–1685.

Chen J.-L., Nolan J.M., Harris M.E., and Pace N.R. 1998. Comparative photocrosslinking analysis of the tertiary structures of *Escherichia coli* and *Bacillus subtilis* RNase P RNAs. *EMBO J.* **17:** 1515–1525.

Costa M. and Michel F. 1995. Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J.* **14:** 1276–1285.

DeLong E.F. 1992. Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* **89:** 5685–5689.

DeLong E.F., Wickham G.S., and Pace. N.R. 1989. Phylogenetic stains: Ribosomal RNA-based probes for the identification of single cells. *Science* **243:** 1360–1363.

Erdmann V. 1976. Structure and function of 5S and 5.8S RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **18:** 45–90.

Forster A.C. and Symons R.H. 1987a. Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. *Cell* **49:** 211–220.

———. 1987b. Self-cleavage of virusoid RNA is performed by the proposed 5S nucleotide active side. *Cell* **50:** 9–16.

Fox G.E. and Woese C.R. 1973. 5S RNA secondary structure. *Nature* **256:** 505–507.

Fox G.E., Stackenbrandt E., Hespell R.B., Gibson J., Maniloff J., Dyer T.A., Wolfe R.S., Baich 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., and Woese C.R. 1980. The phylogeny of prokaryotes. *Science* **209:** 457–463.

Fuhrman J.A., McAllum K., and Davis A.A. 1992. Novel major archaebacterial group from marine plankton. *Nature* **356:** 148–149.

Gautheret D., Damberger S.H., and Gutell R.R. 1995. Identification of base triples in RNA using comparative sequence analysis. *J. Mol. Biol.* **248:** 27–43.

Gogarten J.P., Kibak H., Dittrich P., Taiz L., Bowman E.J., Bowman B.J., Manolson M.F., Poole R.J., Date T.E., Oshima T., Konishi J., Denda K., and Yoshida M. 1989. Evolution of the vacuolar H$^+$-ATPase: Implication for the origin of eukaryotes. *Proc. Natl. Acad. Sci.* **86:** 9355–9359.

Guerrier-Takada C., Gardiner K., Marsh T., Pace N.R., and Altman S. 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35:** 849–857.

Gutell R.R., Larsen N., and Woese C.R. 1993. Lessons from an evolving ribosomal RNA: 16S and 23S rRNA structure from a comparative perspective. In *Ribosomal RNA Structure, evolution, gene expression and function in protein synthesis* (ed. R.A. Zimmerman and A.E. Dahlberg). Tellford Press, Caldwell, New Jersey.

Gutell R.R. 1994. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucleic Acids Res.* **22:** 3502–3507.

Harris M.E. and Pace N.R. 1995. Identification of phosphates involved in catalysis by the ribozyme RNase P RNA. *RNA* **1:** 210–218.

Harris M.E., Kazantsev A., Chen J.-L., and Pace N.R. 1997. Analysis of the tertiary structure of the ribonuclease P ribozyme-substrate complex by site-specific photoaffinity crosslinking. *RNA* **3:** 561–576.

Harris M.E., Nolan J.M., Malhotra A., Brown J.W., Harvey S.C., and Pace, N.R. 1994. Use of photoaffinity cross-linking and molecular modeling to analyze the global architecture of ribonuclease P RNA. *EMBO J.* **13:** 3953–3963.

Heus H.A. and Pardi A. 1991. Structural features that give rise to the unusual stability of RNA hairpins containing GNRA loops. *Science* **253:** 191–194.

Hershberger K.L., Barns S.M., Reysenbach A.-L., and Pace N.R. 1996. Wide diversity of *Crenarchaeota. Nature* **384:** 420.

Hugenholtz P., Pitulle C., Hershberger K.L., and Pace N.R. 1998. Novel division level bacterial diversity in a Yellowstone hot spring. *J. Bacteriol.* **180:** 366–376.

Iwabe N., Kuma K., Hasegawa M., Osawa S., and Miyata T. 1989. Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci.* **86:** 9355–9359.

Jaeger L., Michel F., and Westhof E. 1994. Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J. Mol. Biol.* **236:** 1271–1276.

Kazantsev A.V. and Pace N.R. 1998. Identification by modification—interference of purine N-7 and ribose 2′-OH groups critical for catalysis by bacterial ribonuclease P. *RNA* **4:** (in press).

Kim S.-H. 1979. Crystal structure of yeast tRNA^{phe} and general structural features of other tRNAs. In *Transfer RNA: Structure, properties, and recognition* (ed. P.R. Schimmel et al.), pp. 83–100. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Li Y., Zon G., and Wilson W.D. 1991. Thermodynamics of DNA duplexes with adjacent G-A mismatches. *Biochemistry* **31:** 7566–7572.

Massire C., Jaeger L., and Westof E. 1997. Phylogenetic evidence for a new tertiary interaction in bacterial RNase P RNAs. *RNA* **3:** 553–556.

Michel F. and Westhof E. 1990. Modeling of the three-dimensional architecture of group I introns based on comparative sequence analysis. *J. Mol. Biol.* **216:** 585–610.

Murphy F.L. and Cech T.R. 1994. GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J. Mol. Biol.* **236:** 49–63.

Noller H.P., Moazed D., Stern S., Powers T., Allen P.N., Robertson J.M., Weiser B., and Triman K. 1990. Structure of rRNA and its functional interaction in translation. In *The ribosome: Structure, function and evolution* (ed. W.E. Hill et al.), pp. 73–92. American Society for Microbiology, Washington, D.C.

Noller H.P., Kop J., Wheaton V., Brosius J., Gutell R., Kopylov A.M., Dohme F., Herr W., Stahl D.A., Gupta R., and Woese C.R. 1981. Secondary structure model for 23S ribosomal RNA. *Nucleic Acids Res.* **9:** 6167–6189.

Olsen G.J. and Woese C.R. 1997. Archaeal genomics: An overview. *Cell* **89:** 991–994.

Oparin A.I. 1964. *The chemical origin of life* (transl. A. Synge). Charles C Thomas, Springfield, Illinois.

Pace N.R. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276:** 734–740.

Pace N.R., Stahl D.A., Lane D.J., and Olsen G.J. 1985. Analyzing natural microbial populations by rRNA sequences. *Am. Soc. Microbiol. News* **51:** 4–12.

Pley H.M., Flaherty K.M., and McKay D.B. 1994. Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature* **372:** 111–113.

Powers T. and Noller H.P. 1991. A functional pseudoknot in 16S ribosomal RNA. *EMBO J.* **10:** 2203–2214.

Pühler G., Leffers H., Gropp P., Palm P., Klenk H.-P., Lottspeich P., Garrett R.A., and Zillig W. 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eucaryotic nuclear genome. *Proc. Natl. Acad. Sci.* **86:** 4569–4573.

Ramirez C., Shimmin L.C., Newton C.H., Matheson A.T., and Dennis P.P. 1989. Structure and evolution of the L11, L1, L10, and L12 equivalent ribosomal proteins in eubacteria, archaebacteria and eukaryotes. *Can. J. Microbiol.* **35:** 234–244.

Sandman K., Krzycki J.A., Dobrinski B., Lurz R., and Reeve J.N. 1990. DNA binding protein HMf, isolated from the hyperthermophilic archaea *Methanothermus fervidus*, is most closely related to histones. *Proc. Natl. Acad. Sci.* **87:** 5788–5791.

SantaLucia J., Jr., Kierzek R., and Turner D.H. 1990. Effects of GA mismatches on the structure and thermodynamics of RNA internal loops. *Biochemistry* **29:** 8813–8819.

Siegel R.W., Banta A.B., Haas E.S., Brown J.W., and Pace N.R. 1996. Mycoplasma fermentans simplifies our view of the catalytic core of ribonuclease P RNA. *RNA* **2:** 452–462.

Tanner M.A. and Cech T.R. 1995. An important RNA tertiary interaction of group II introns is implicated in gram-positive RNase P RNAs. *RNA* **1:** 349–350.

Trent J.D., Nimmersgern E., Wall J.S., Harti F.-U., and Horwich A.L. 1991. A molecular chaperone from a thermophilic archaebacterium is related to the eukaryotic protein t-complex polypeptide-1. *Nature* **354:** 490–493.

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., and Gold L. 1988. CUUCGG hairpins: Extraordinarily stable RNA secondary structures associated with various biochemical processes. *Proc. Natl. Acad. Sci.* **85:** 1364–1368.

Ueda T.Y., Suga Y., and Matsuguchi T. 1995. Molecular phylogenetic analysis of a soil microbial community in a soybean field. *Eur. J. Soil Sci.* **46:** 415–421.

Varani G., Cheong C., and Tinoco I., Jr. 1991. Structure of an unusually stable RNA hairpin. *Biochemistry* **30:** 3280–3289.

Wilson K.S. and Noller H.F. 1998. Molecular movement inside the translational engine. *Cell* **92:** 337–349.

Winker S. and Woese C.R. 1991. A definition of the domains archaea, bacteria and eucarya in terms of small subunit ribosomal RNA characteristics. *Syst. Appl. Microbiol.* **14:** 305–310.

Woese C.R. 1967. *The genetic code: The molecular basis of genetic expression.* Harper and Rowe, New York.

———. 1982. Archaebacteria and cellular origins: An overview. Zentbl. *Bakteriol. Mikrobiol. Hyg. Abt. 1 Orig. C* **3:** 1–17.

———. 1987. Bacterial evolution. *Microbiol. Rev.* **51:** 221–271.

———. 1998. The universal ancestor. *Proc. Natl. Acad. Sci.* **95:** 6854–6859.

Woese C.R. and Fox G.E. 1977. The concept of cellular evolution. *J. Mol. Evol.* **10:** 1–6.

Woese C.R., Kandler O., and Wheelis M.L. 1990a. Toward a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87:** 4576–4579.

Woese C.R., Winker S., and Gutell R.R. 1990b. Architecture of ribosomal RNA: Constraints on the sequence of tetra-loops. *Proc. Natl. Acad. Sci.* **87:** 8467–8471.

Woese C.R., Gutell R., Gupta R., and Noller H.P. 1983. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.* **47:** 621–669.

Woese C.R., Stackenbrandt E., Macke T.J., and Fox G.E. 1985. A phylogenetic definition of the major eubacterial taxa. *Syst. Appl. Microbiol.* **6:** 143–151.

Woese C.R., Magrum L.J., Gupta R., Siegel R.B., Stahl D.A., Kop J., Crawford N., Brosius J., Gutell R., Hogan J.J., and Noller H.P. 1980. Secondary structure model for bacterial 16S robosomal RNA: Phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res.* **8:** 2275–2293.

Yang D., Oyaizu Y., Oyaizu H., Olsen G.J., and Woese C.R. 1985. Mitochondrial origins. *Proc. Natl. Acad. Sci.* **82:** 4443–4447.

Zuckerkandl E. and Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8:** 357–366.

WWW RESOURCE

http://pundit.colorado.edu:8080/RNA/16S/eubacteria.html (eu)BACTERIA 16S rRNA Comparative Structure Database.