MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Dec. 2009, p. 565–576 Vol. 73, No. 4
1092-2172/09/$12.00 doi:10.1128/MMBR.00033-09

Mapping the Tree of Life: Progress and Prospects

Norman R. Pace*

Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado 80309-0347

INTRODUCTION

Gene sequence variation between different organisms provides a metric for biological diversification. Sequence variation can also serve as the basis for inference of the patterns of evolution from precellular life until now. The intent of this article is to assess critically our current understanding of life's phylogenetic diversity on a large scale. My view is from the molecular standpoint, mainly from the perspective of rRNA phylogeny. A molecular perspective on life's diversity and evolution is only now unfolding, and there is much controversy and paradox, only some of which I can address here.

All molecular phylogenetic trees have systematic limitations that cloud our view of the deeper branches in the tree of life (ToL). Consequently, I discuss the building of phylogenetic trees and emphasize the intrinsic limitations of any results. Progress toward assembly of a universal phylogenetic ToL also relies on how comprehensive is our knowledge of the extent and the richness of life's diversity. Therefore, I show how the recent explosion of environmental sequences has heavily influenced the patterns seen in the trees. I conclude that we have in place the outlines of a universal ToL, but the details of the patterns of deep evolution in all the phylogenetic domains remain obscure.

THE PATH TO A SCIENTIFIC ToL

The notion of some sort of connectedness between all of life is ancient. The classical "great chain of being" was an imagined, treelike hierarchy of existence reaching from minerals through simple life to humans and the gods. That, of course, was not an evolutionary model, but treelike portrayals of biological relationships—specific evolutionary models—were proposed by the mid-19th century (58). Some of the ideas of the early evolutionists are still with us. A derivative of one of Ernst Haeckel's 1866 trees (23) is seen in textbooks today as the five-kingdoms model for large-scale evolution. However, where the microbial world fit into the rest of life could not be known. Indeed, the microbial world was barely acknowledged by most biologists of that time, brushed aside as "monera," perhaps nonliving.

Early biologists speculated as to whether life had a common or multiple origins, but there was no way of determining the truth. Discoveries in the first half of the 20th century established that all life, microbes and large organisms, had similar biochemistry. Because of the complexity of life processes and the consequent improbability of their independent evolution in different organisms, the universality of biochemistry indicated that all life had common ancestry. Thus, all life seemed related, but in what way? What are the patterns of the relationships? Evolutionary patterns within animals, plants, and a few other kinds of organisms could be inferred from morphological and developmental properties, but the major kingdoms could not be related to one another. How could one compare an oak tree with a nematode? Moreover, most microbial organisms, in their apparent simplicity, had few criteria for comparison with one another let alone with large organisms. There was no metric, no objective criterion for measuring phylogenetic variation and relating evolutionary histories among disparate organisms.

The development of nucleic acid sequencing technology in the 1960s and 1970s sparked a profound advance in the ways that we can perceive and study microorganisms and probe past evolutionary events (77). Carl Woese's comparative studies of rRNA sequences proved the universal relatedness of all life and, in 1977, put in place the first outlines of a universal sequence-based ToL (73). Known life was seen to fall into one of three phylogenetic domains: *Archaea* (formerly archaebacteria), *Bacteria* (eubacteria), and *Eucarya* (eucaryotes). This three-domain model for the deepest branches in evolution is

* Mailing address: Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309-0347. Phone: (303) 735-1864. Fax: (303) 492-7744. E-mail: nrpace@colorado.edu.

now well grounded by considerable further sequence information and biochemical correlations (50, 74).

Woese's results also opened an entirely new way to understand and classify microbial diversity in the context of phylogenetic sequence comparisons. Prior to the availability of gene sequences, the classification of microbes relied on overt phenotypic traits that were often subjective. In contrast, gene sequence comparisons provided the metric, a natural and objective way to classify organisms based on sequence change. Finally, classification of microbial life could be brought into order.

An important practical corollary of phylogenetic classification by sequence comparisons was that sequences could be used to identify phylogenetically microbes that are otherwise uncharacterized. This opened the door for microbiologists to begin to explore the makeup of the natural microbial world independently of culture by determination of gene sequences obtained directly from the environment through cloning or other means. Only a few microbes can be captured by routine culture, so molecular analyses of environmental sequences have substantially expanded our knowledge of the diversity of microbial life (1, 32, 50, 54).

WHAT IS A MOLECULAR TREE, AND HOW DOES IT RELATE TO A TREE OF ORGANISMS?

A phylogenetic tree is a graphical representation of relationships between organisms or molecules. In a molecular phylogenetic analysis, relationships between orthologous genes from different organisms are elucidated by comparison of their sequences. ("Homologous" genes have common ancestry and are of three kinds: "orthologs" have a common function; "paralogs" arose from a gene duplication and subsequent independent evolution, frequently with different functions; "xenologs" have undergone lateral transfer and so have evolved independently of the cellular line of descent.) Since sequence differences reflect evolutionary variation, the sequences can be used with the techniques of molecular phylogeny to infer maps of the course of evolution, or "phylogenetic trees." Darwin's dream that "our classifications will come to be, as far as they can be so made, genealogies" (12) can now be realized in principle.

However, a sequence-based molecular phylogenetic tree is not exactly what Darwin imagined. Darwin thought of evolution in terms of organisms, not molecules. A molecular tree is not necessarily a tree of organisms because differences between organisms go beyond differences between molecules. Organisms differ not only by evolutionary changes in individual molecules, but also by their overall contents and compositions of genes. Organisms can acquire novel genes by lateral-transfer events or by the propagation of gene families in species-specific ways, through little-known mechanisms. The point is that whole genome sequences of organisms intrinsically are not directly comparable because their genetic compositions are not the same, i.e., not entirely orthologous. Phylogenetic analysis with orthologous genes in different organisms can trace the relationships and evolutionary fate, the line of evolutionary descent, of a particular gene, but only that gene. Large-scale molecular trees based on lines of descent transcend relationships based on organisms. For instance, from the organism view, we may think of the dinosaurs as extinct, but in the phylogenetic view the dinosaur line of descent is with us today in the form of birds. From the phylogenetic standpoint, the organisms that we perceive are not steps in evolution, but rather are dead-end states, transient manifestations of the particular lines of descent.

MAKING MOLECULAR PHYLOGENETIC TREES

Mapping the ToL depends on molecular phylogenetic analyses. The field of molecular phylogeny has blossomed over recent decades, and innumerable texts, reviews, and computer programs are available for instruction and for the practice of molecular phylogeny. Students of rRNA phylogeny have a growing wealth of information to draw upon. Beyond raw sequences available from GenBank (U.S. National Center for Biotechnology Information), which currently holds >2 million small-subunit (SSU) rRNA sequences, the public databases Greengenes (U.S. Department of Energy) (16), Ribosomal Database Project (10), SILVA (52), and others (8) maintain curated rRNA sequence alignments and tools for using the sequences. I treat only elementary issues of molecular phylogeny, with the goal of drawing attention to some aspects of the process that compromise the accuracy with which we can view the deep topology of the ToL with molecular phylogenetic methods.

The process of making a phylogenetic tree is simple in essence. (i) Some collection of sequences is (carefully) aligned to juxtapose homologous residues, nucleobases, or amino acids. (ii) The differences between all pairs of sequences are counted. (iii) Sequence differences between the pairs are fitted to some topology that overall best corresponds to the data. Embedded in this essential simplicity, however, is much computational complexity at each step, and often controversy among phylogeneticists as to how best to conduct and evaluate computations (29).

The sequence alignment process, the pairing of orthologous residues, is critical and receives too little attention. Alignment for deep phylogeny requires not only determination of the register of sequences, but also restriction of less informative elements, such as highly variable sequences for which no alignment is reliably discernible. Incorrectly aligned or evolutionarily randomized (too many unseen changes) sequence elements increase the noise and the uncertainty in calculations.

Beyond the sequence alignment is the count of changes that have occurred between pairs of sequences, made complex by the fact that the number of sequence changes counted is always less than the number that have actually occurred. This is because of some possibility of multiple unseen changes in the past and back-mutations that might cancel changes. The sequence count thus becomes a statistical assessment of the probabilities of "actual" changes based on observed differences. Additionally, the estimation of "actual" sequence change is complicated by different rates of change in different lineages, by base composition effects, and by other factors (28).

For these and other reasons, some fraction of the data interpreted in the calculation of any phylogenetic tree is based on unknowable information. There are a number of ways to estimate unseen changes and to correct for rate variation, but an inevitable conclusion is that the deeper branch points in a

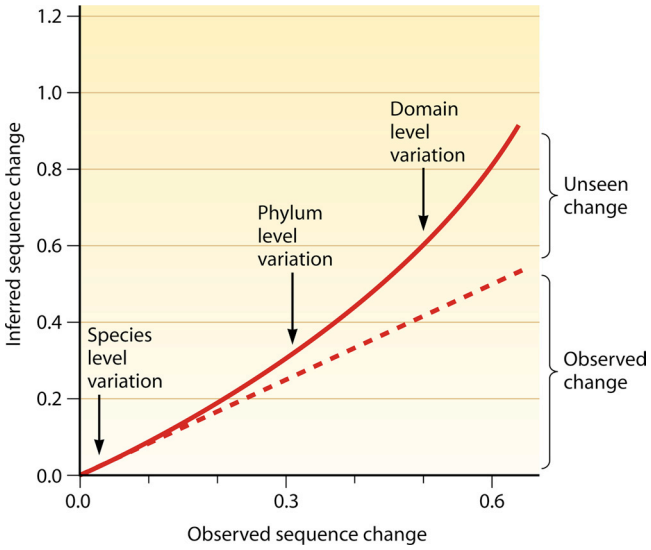


FIG. 1. Sequence uncertainty with depth in a phylogenetic tree. Dashed line, not corrected for unseen changes; solid line, corrected for unseen changes using the following estimation: inferred sequence change (K_{nuc}) = $-3/4$ ln[1 $-$ (4/3)D], where D is the number of changes counted (31).

phylogenetic tree can be based significantly on inferred information. This is illustrated in Fig. 1, which shows the correspondence of counted and inferred sequence changes with depth (sequence difference) in a phylogenetic tree using one of the algorithms for estimation of unseen change (31). As shown in the figure, for close relatives, for instance, at the species or genus level, unseen changes do not matter much. However, SSU rRNA sequences vary by 30 to 35% between the main bacterial phyla and ~50% between the domains. Consequently, any branching orders in phylogenetic trees based on rRNA or other sequences have intrinsic uncertainties that increase nonlinearly with depth in the tree.

The fitting of pairwise sequence differences to a phylogenetically connected matrix, a "phylogenetic tree," can be done in different ways (28). "Evolutionary-distance" methods take inferred sequence change to reflect evolutionary distance and calculate tree topology based on the best fit to the data. "Parsimony" methods calculate ancestral sequences and consider the best trees to be the ones with the fewest sequence changes required for creation of the topology from the modern sequences. "Likelihood" (19) and other statistical methods (30) calculate the probability that the sequence changes observed in the alignment fit the particular tree topology. In the early days of molecular phylogeny, there was much controversy over the "best" method for inference of phylogenies, but at the current stage of development, with compensation for variable rates, variable base compositions, and other potentially confounding factors, these three popular methods generally yield similar (albeit seldom identical) results. Likelihood methods generally are considered the most robust because they are based on statistical assessments. Nonetheless, regardless of the method, the positions of nodes become less resolved with depth in any tree.

A balanced diversity of sequence representation is critical for the most accurate phylogenetic reconstructions (78). Poor sequence representation in phylogenetic calculations results in long line segments connecting nodes in phylogenetic trees and the potential artifact of "long-branch attraction," which can create spurious phylogenetic associations and tree topologies (5). Poor sequence representation of many large relatedness groups currently plagues efforts to infer the deep relationships in the domains.

TESTING TREES

Phylogenetic-tree algorithms always produce an estimate of the "best fit" between the data, the sequence difference counts, and the topology of the tree, but is the result biologically true? The observed optimum topology pertains only to the particular calculation. The specific results of any tree calculation are blurred, not only by the statistical uncertainties mentioned above, but also by the sequences that are included in the calculation, the intrinsically heuristic nature of the computer search for the best tree, and other factors. How, then, can we derive the best biologically relevant tree, the true branch points from which to infer the course of evolution?

There is no single answer to this question. Tests of particular tree topologies include the use of different algorithms for constructing trees, the use of different sequence sets, and biochemical or other correlations. To test any particular tree result, "bootstrap analysis" is commonly conducted (18, 27). In this method, calculations are performed many times with the same sequence set, but each time using only random subsets (e.g., 70%) of residues that comprise individual sequences. Tree nodes that indicate clades, or relatedness groups, are scored by how frequently they are grouped in the different calculations. Even clades with good bootstrap support in a particular analysis, however, are subject to perturbation with different analyses or the incorporation of new sequences. It is critical to view any phylogenetic tree as a tentative model that only more or less accurately reflects the evolutionary history.

WHY rRNA AS THE BACKBONE OF A UNIVERSAL TREE?

No single gene is sufficient for resolution of relationships throughout the ToL simply because no single gene contains sufficient information. Any ultimate ToL will be an amalgam of molecular trees based on many molecules, with the resolution determined by the molecular sequences appropriate for the depth and place in the tree. The question then becomes, what gene or genes are useful as standards for tracing the deep history of all life, for forming the backbone of a universal tree? The constraints on sequences that are useful for building a universal phylogeny are both obvious and subtle (71). An obvious point is that the gene must occur in all organisms. The gene also must reflect the cellular line of descent. That is, the gene must not have undergone lateral transfer between different genetic lineages. Additionally, the gene must be extraordinarily conservative. Accurate alignment of sequences is critical for the most accurate phylogenetic analysis, so conserved sequences and structures are important landmarks for registration of sequence alignments.

More subtly, a high degree of conservation indicates that the gene sequence has not been randomized over the course of

evolution and thereby lost all useful information. For instance, it is common for amino acid sequences of enzymes or regulatory proteins to have diverged to such an extent that they are unrecognizable, except for residues absolutely required for activity. Since a phylogenetic analysis requires an assessment of the number of changes that have occurred between compared sequences, randomized sequences introduce unknowable information and degrade the analysis. Other constraints on the best possible gene for deep phylogeny include the size of the gene, that is, the number of sequence positions available for statistical assessments.

The rRNA genes seem to fit the criteria for developing the first outlines of a universal ToL better than any other genes (71). Carl Woese's choice of SSU rRNA sequences as the basis for a universal phylogeny was prescient. SSU rRNA gene sequences have become the gold standard for microbial identification and the inference of deep evolutionary relationships. rRNA genes occur in all cells and organelles, and the rRNA genes are the most conservative large sequences in nature. For instance, the eucaryotic and bacterial SSU rRNA genes typically have about 50% identity over the alignable lengths of their SSU rRNAs. The rRNA genes have not undergone significant lateral transfer, and the structural properties of the rRNA provide for optimization of alignments. Moreover, at this time, the SSU rRNA databases are the main source of information on environmental microbial diversity.

Beyond the attributes of rRNA sequences for phylogenetic studies and classification, there are limitations. One important limitation is the very conservative nature that makes the rRNA sequence so useful for the inference of deep phylogeny. A consequence of such extreme conservation is that rRNA sequences may not be useful for discrimination of close relatives at the strain or even species level because so few changes in the rRNA sequence have occurred between organisms at those levels of classical taxonomic discrimination. For instance, rRNA sequences do not reliably distinguish humans from mice or *Escherichia coli* from *Shigella dysenteriae*. Molecular discrimination between such closely related organisms requires comparison of less conserved genes than rRNA, in which significant numbers of differences occur among the compared sequences from different organisms. On the other hand, close similarities in rRNA sequences do indicate that these superficially different organisms are similar at the cellular level: cell structure, basal metabolism, etc.

Another limitation on SSU rRNA gene sequences for deep phylogeny is the size of the gene, 1,500 to 2,000 bp, half of which are invariant. Consequently, the information available in the SSU rRNA gene for resolving phylogenetic relationships and branching orders deep in the domains is limited to ~1,000 characters. The amount of information used in a phylogenetic analysis is important because it influences the statistical accuracy of the results; more information, if sound, is always better.

One way to expand the information available for phylogenetic studies is to include other genes, for instance, the large-subunit (LSU) rRNA gene, which is typically about twice the size of the SSU gene. However, compared to SSU sequences, the number of LSU sequences available for inclusion in analyses is currently miniscule. The SILVA database, for instance, currently (2009) contains ~400,000 SSU long reference sequences and only ~15,000 LSU reference sequences. Moreover, the LSU databases currently include only a limited diversity of sequences compared to the SSU sequence collections, because almost all environmental phylogenetic surveys are based on SSU sequences.

Another way to try to expand the information available for phylogenetic reconstructions is to use concatenations or other combinations of multiple gene sequences for the inference of trees. This approach to building phylogenies certainly can be useful for resolving branches in the ToL. However, the use of concatenated gene sets for deep evolution is fraught with considerable uncertainty, such as the accuracy of sequence alignments and the potential inclusion of randomized (highly variable) or nonhomologous sequences. Moreover, sequence representation for diverse organisms is seriously limited. Phylogenetic trees made with concatenated gene alignments generally correspond to the three-domain tree outlined by rRNA sequences, but often with considerable discordance within the domains (7, 9, 13, 20, 75).

Thus, the SSU rRNA does not stand alone for assessment of large-scale phylogenetic relationships and classification. However, the conservative nature of the molecule makes it particularly useful for examining deep relationships. Phylogenetic trees made with sequences of other genes of the central nucleic acid-based information transfer process (e.g., RNA polymerase, DNA polymerases, ribosomal proteins, and protein synthesis elongation factors) are congruent with the rRNA trees (7), so changes in rRNA sequences evidently describe the evolutionary path of the genetic machinery at the very least. It seems likely that this congruence carries over into most other central aspects of cellular organization and function—the "body plan" of the organism, to invoke a usage of the early evolutionists. Metabolic processes that respond to the environment may or may not track with the rRNA and consequently seem more subject to lateral transfer. Practically, the wide use of SSU rRNA sequences for classification and environmental surveys guarantees the continued development of the SSU rRNA database as a reference library.

ENVIRONMENTAL SEQUENCES EXPAND KNOWN DIVERSITY

By the mid-1980s, Woese and colleagues had surveyed by sequence or by oligonucleotide catalog several hundred SSU rRNA sequences, with examples of most of the diverse microbial groups then known (71). All of those sequences were derived from cultured organisms. Also in the mid-1980s, rRNA gene sequences began to be used to survey the phylogenetic makeup of naturally occurring microbial assemblages without the requirement for culture (49, 50). In general, with environmental surveys, rRNA or other genes are isolated, by cloning or by PCR, from DNA isolated directly from environments of interest and sequenced. The results, collections of rRNA gene sequences, are snapshots of the phylogenetic makeup of the natural microbial world.

All molecular environmental surveys so far conducted have been limited in scope and tentative, identifying only the most abundant sequences because of the rich complexity encountered in all environments. Nonetheless, even the early results showed that knowledge of microbial diversity from all three domains based on cultured organisms was and remains seri-

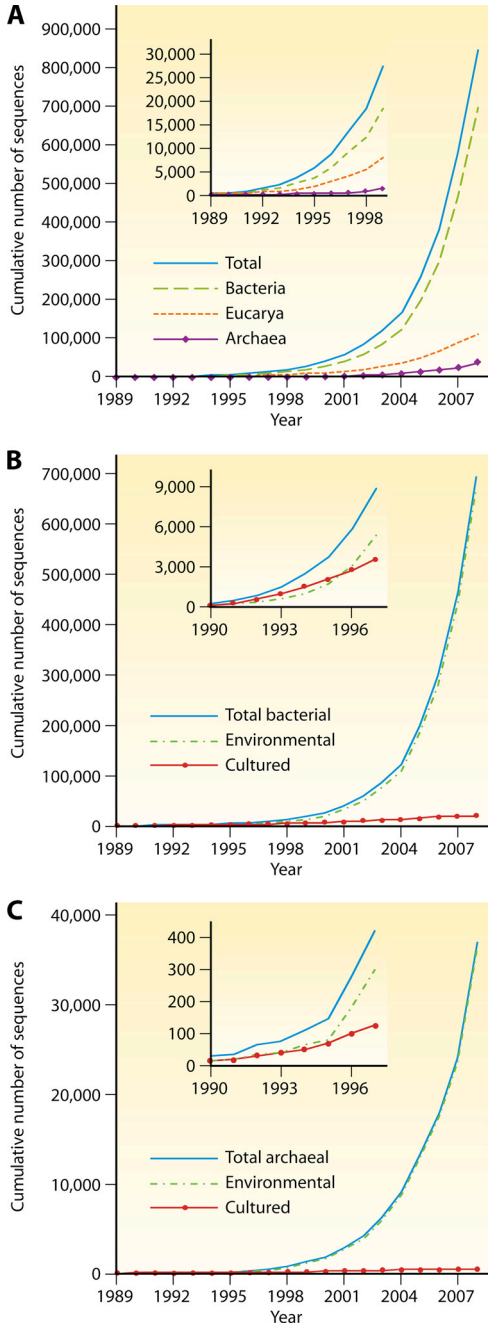


FIG. 2. Chronological accumulation of SSU rRNA sequences. The data are derived from the SILVA 98 SSU Parc database (52) using the EMBL taxonomic designations for the sequences (66). The SILVA SSU Parc database contains rRNA sequences that are 300 or more nucleotides in length and validated as rRNA with RNAmmer (43). (A) Accumulation of total, archaeal, bacterial, and eucaryal SSU sequences. (B) Accumulation of rRNA sequences from cultured and environmental bacteria. (C) Accumulation of rRNA sequences from cultured and environmental archaea.

ously limited. Figure 2 summarizes the chronological accumulation of rRNA sequences (Fig. 2A) and the contributions of sequences from cultured versus environmental organisms (Fig. 2B and C). As shown in the figures, the environmental sequence collection passed that of cultivars in the late 1990s and has since exploded, far outnumbering sequences of cultured organisms in the public databases. Many of the environmental sequences fall into established phyla, but often the environmental sequences identify entirely new major groups of microbial life that are only distantly related to cultivars. These findings expand our perspective on the richness of microbial diversity. Moreover, beyond the abundant sequences detected in any environment is a "rare biosphere" of less common organisms that is only beginning to be plumbed (62).

What can be learned from an environmental rRNA gene sequence? First, the particular phylogenetic type, the "phylotype," is identified as a component of the natural ecosystem. The abundance of the sequence provides some idea of its prevalence in the ecosystem or local assemblage. There is not a direct correspondence between the frequencies of rRNA genes and the frequencies of organisms because different organisms contain different numbers of rRNA genes. *E. coli*, for instance, contains 7 rRNA genes, *Bacillus subtilis* contains 10, and mycobacteria contain only 1, typical of environmental organisms (39). Nonetheless, the sequence census provides some rough assessment of the abundances of otherwise unknown organisms and thereby an idea of their importance to the community.

More or less detail can be inferred about the nature of an environmental organism from the phylogenetics of an environmental sequence. This depends on how closely the sequence is related to that of a characterized organism. There are no formal conventions for extrapolation between rRNA sequence similarity and classical taxonomic definitions. As a rule of thumb, SSU rRNA sequences with ≥97% sequence identity often are taken to represent members of the same "species" (63). Bacterial "species" operationally defined in this manner occupy a large evolutionary space, however. Genomic studies of different isolates of the same described species of bacteria typically find that ~30% of the genes in any particular genome are seen in none of the others; ~30% of each contains genes with no identifiable homologs in the databases (26, 46, 55). Moreover, environmental studies with rapidly evolving sequences to resolve close relationships show that rRNA-defined species occur in nature not as discrete species, but rather as populations, phylogenetic clusters of closely related but not identical organisms (54, 68).

Still, in spite of all the variability, intraspecific genome comparisons typically show a core of ~40% or more of the genes that are common to the various representatives of the species, the core genome of the particular relatedness group. Members of a phylogenetic group at any level are expected to have properties that are common to the group, so if an environmental rRNA sequence is ≥97% identical to that of a characterized organism, considerable information about the known organism can be extrapolated to predict some properties of the environmental organism: the general nature of the cellular machinery and the core metabolic strategies at the very least. Sequences that diverge deeply from known ones in phylogeny provide less descriptive information about the environmental organisms. The sequences do, however, identify targets for further study, if merited. The environmental rRNA sequences then become the basis for molecular tools, such as hybridization probes and PCR primers, with which to pursue those organisms experimentally. Moreover, environmental rRNA se-

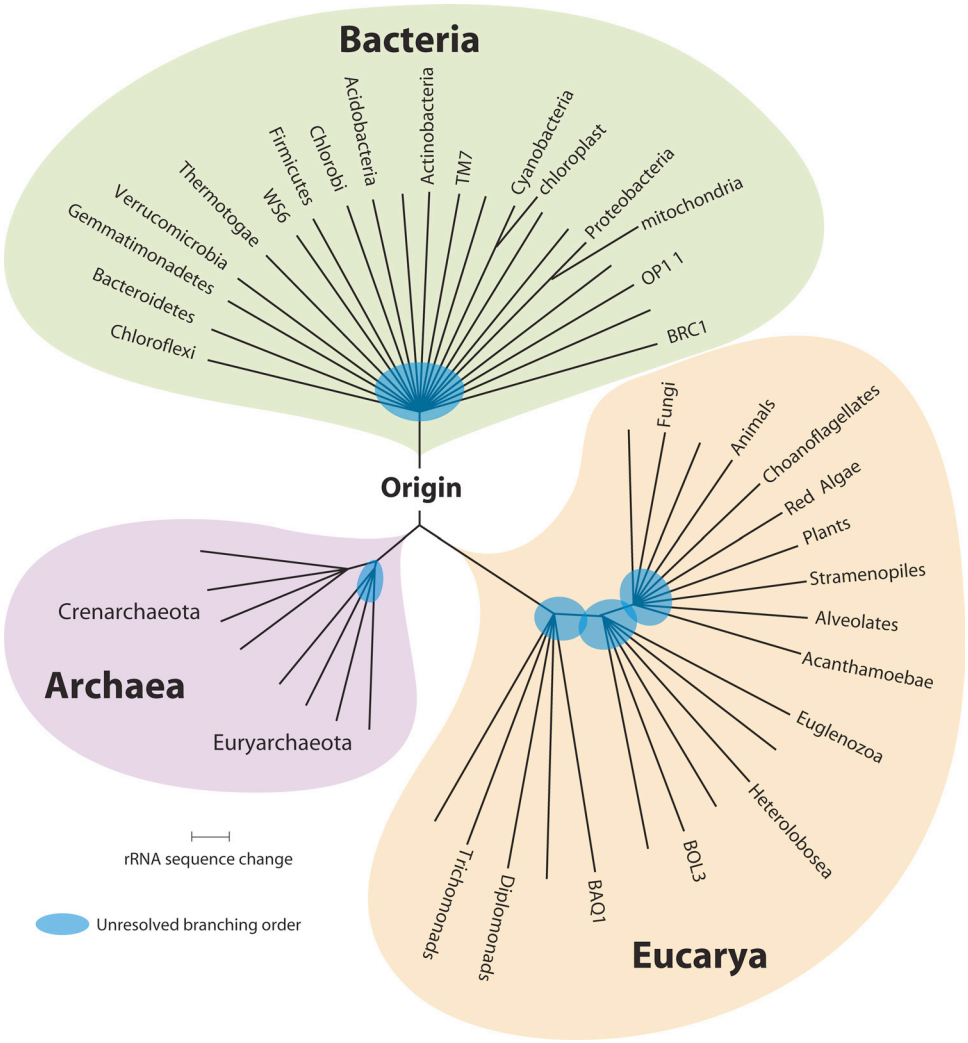


FIG. 3. A molecular ToL based on rRNA sequence comparisons. The diagram compiles the results of many rRNA sequence comparisons. Only a few of the known lines of descent are shown.

quences contribute heavily to the emerging structure of the ToL.

THE OUTLINES OF A UNIVERSAL TREE

Molecular phylogeny is a growing enterprise, but relatively few data are available for many large relatedness groups of organisms in all the domains. Consequently, any branching orders of the main phyla at the base of each of the domains cannot be accurately assessed from currently available information. However, numerous phylogenetic studies of rRNA and other gene products converge to a considerable extent. Figure 3 is a diagram of my assessment of what seems reliable at this time, on a large scale, from the perspective of rRNA sequence variation. Each of the lines of descent that comprise the main phyla of the domains in the figure is itself a complex radiation. The statistical blurring of branching orders at the bases of major radiations is indicated. The topologies of the individual domains are discussed below.

The three-domain topology of the ToL is established by the observation of three fundamentally distinct relatedness groups of organisms; but where is the origin, the "root" of the tree? This could not be established using rRNA sequences. The only way to determine the position of the "root" in a phylogenetic tree is with reference to outgroup sequences, and there are no outgroup sequences in a universal collection such as rRNA sequences. This dilemma was overcome by the use of paralogous sequences that are thought to have derived from gene duplication before the last universal common ancestor (22, 35). Examples of such paralogous gene families are ATP synthase α and β subunits and the protein synthesis factors elongation factor G and EF-Tu. Phylogenetic analysis with sequences of each of these genes analyzed individually produces the three-domain topology, so the duplication must have occurred before the differentiation of the domains. Consequently, each family member sequence can be used to root the tree formed from its paralog. With these and other appropriate paralogous genes, the intersection of the gene family member trees is seen to lie on the line that leads to the bacterial radiation, as illustrated in Fig. 3.

This rooting of the ToL means that the lines of descent that led to archaea and eucaryotes had an early common history independent of the line that resulted in bacterial diversity. This relationship is seen not only by phylogenetic analysis, but also in many cellular properties in which archaea and eucaryotes resemble one another more than either resembles bacteria. For instance, archaea and eucaryotes conduct replicative DNA synthesis using homologous versions of a DNA polymerase complex, whereas bacteria use DNA polymerase III, a different system (45). As another example, archaea and eucaryotes use a TATA-binding protein for transcription initiation, while bacteria employ a σ-protein-dependent mechanism (67). The comparatively close relationship of archaea and eucaryotes does not mean that archaea are rudimentary eucaryotes; vast evolutionary distance separates these two kinds of organisms, and this is reflected in many cellular properties. Archaea, for instance, use only ether-linked lipids in their membranes instead of the ester-linked lipids seen throughout the bacteria and eucaryotes.

The rRNA (and other) sequence comparisons also prove the antiquity of the eucaryotic line of descent. In the absence of lithified fossil evidence, considerable speculation and controversy has accrued with regard to the age of eucaryotes in Earth's history. The conventional wisdom and textbook portrayal would have the emergence of eucaryotes relatively late in the history of life, well after the bacterial and archaeal radiations. The deep separation of the eucaryal and archaeal lines in phylogenetic trees shows, however, that the eucaryotic nuclear line of descent is as old as the archaeal line. These molecular relationships say nothing whatever about the presence or absence of a nuclear membrane in those earliest eucaryotes, but in the light of the molecular relationships, the morphological trait of the nuclear membrane is irrelevant in the determination of the evolutionary history. The sequence comparisons show unequivocally that the eucaryotic nuclear line of descent has been around since the beginning.

The unbranched lines that lead from the common ancestral state to each of the domain level radiations in Fig. 3 presumably represent the period of evolution that preceded the emergence of fully developed cells capable of autonomous propagation and thereby genetic differentiation. Woese has painted that precellular world as a time of communal sharing of genes in far more rampant ways than now (72). Differentiation of the domains would have resulted from evolutionary acquisition of specificity in molecular interactions, which would restrict lateral transfer in phylogenetic space. This could have resulted in a funneling of genetic information into what became the domain relatedness groups. The genetic lines at the bases of the domains could diverge only after the emergence of cellular sophistication sufficient for propagation of independent genetic lines of descent.

The average rates of rRNA sequence change in the three domains have not been constant over evolutionary time. This is seen in rRNA trees as systematically different line lengths for representative sequences of the domains, as illustrated in Fig. 3. Eucaryotic sequences tend to have changed more from the inferred common ancestral sequence than have bacterial sequences, which in turn have changed more than archaeal sequences. The comparatively low rate of change seen among archaeal rRNAs, their low rate of evolution, is the reason that Woese first dubbed them "archaebacteria" (73). Their rRNA genes seemed systematically less changed from the common ancestor, i.e., more primitive, than those of bacteria or eucaryotes.

Exceptions to these generalities in rates of change, as seen in the lengths of line segments in phylogenetic trees, occur in all the domains. Episodes of rapid change in some lineages are evident in the ToL, and sometimes these mark important evolutionary events. Examples of this are seen with the chloroplasts and mitochondria, for instance. Many lines of evidence show that chloroplasts were derived from cyanobacteria and mitochondria from the alphaproteobacteria. As diagrammed in Fig. 3, the organellar rRNA genes appear in tree calculations as long branches compared to their respective parental bacterial branches. The rapidly evolving phases in organellar evolution possibly reflect rapid adaptation to an entirely new environment, that of the host eucaryotic cell.

THE BACTERIAL TREE—STILL EXPANDING

Identification of branching orders among the main phyla within the domains is a way to perceive the course of evolution in each of the domains. It is also a way to structure the classification of the organisms that comprise the domains. There is no sanctioned taxonomy of bacterial phyla, and the ongoing flood of environmental sequences has overwhelmed the accounting (Fig. 2). Woese's early surveys of rRNAs from seemingly diverse bacteria identified 12 main phyla, distinct relatedness groups of organisms by rRNA sequences (71). The number of recognizable bacterial phyla continues to increase due to culture activities and, particularly, environmental rRNA gene surveys. Currently the public databases collectively identify >70 phyla of bacteria, defined as relatedness groups of sequences that have no reliable associations with other phyla in rRNA phylogenetic analyses (11). Table 1 contains a list of all the named (with cultured representation) and some of the "candidate" (not documented by culture) rRNA phyla in use in at least two of the public databases and which are documented by >100 SSU rRNA sequences. The accounting of bacterial phyla, even with currently available sequences, clearly is incomplete. Large numbers of sequences in the databases do not fall into the defined groups and, along with new sequences, will be the fodder for future expansion of our understanding of the bacterial tree.

Only about half of the recognizable bacterial phyla have any cultured representation, and for most of the phyla, that representation is sparse, with only a few cultured examples. Most of the history of microbiology is based on representatives of the few phyla that happen to contain human pathogens and which tend to be readily cultured with classic methods (Table 1). Even sequence representation in the databases is highly skewed and is sparse for most of the bacterial phyla. For instance, Fig. 4 summarizes the distribution of sequences among the dozen most richly covered bacterial phyla; the other ~60 phylum level relatedness groups have comparatively little sequence representation. This skewing of database sequences to only a few phyla possibly reflects environmental abundance, but it seems more likely to be due to limited sampling of diverse environments. For instance, the phylum *Bacteroidetes* is highly represented in the sequence databases (Fig. 4), but most

TABLE 1. Named and candidate rRNA phyla in use in at least two of the public databases[a]

Phylum	Source	Reference
Named		
Acidobacteria		
Actinobacteria[b,c]		
Aquificae		
Bacteroidetes[b,c]		
Chamydiae[b,c]		
Chlorobi[b]		
Chloroflexi[b]		
Chrysiogenetes		
Cyanobacteria[b]		
Deferribacteres		
Dictyoglomi		
Fibrobacteres		
Firmicutes[b,c]		
Fusobacteria[c]		
Gemmatimonadetes		
Lentisphaerae		
Nitrospirae		
Planctomycetes[b]		
Proteobacteria[b,c]		
Spirochaetes[b,c]		
Synergistetes		
Thermi-Deinococci[b]		
Thermodesulfobacteria		
Thermotogae[b]		
Verrucomicrobia		
Candidate		
GN02	Guerrero Negro hypersaline mat	47
OD1	Originally part of OP11 group	24
OP3	Obsidian Pool, hot spring	33
OP9	Obsidian Pool, hot spring	33
OP10	Obsidian Pool hot spring	33
OP11	Obsidian Pool hot spring	33
SR1	Sulfur River cave sediment	2
TG1	Termite group 1[d]	48
TM7	Peat bog	34
WS3	Wurtsmith contaminated aquifer	17
WS6	Wurtsmith contaminated aquifer	17

[a] Acknowledged by two or more databases containing >100 SSU sequences detected in multiple environments. Named, with cultured representation; candidate, not documented by culture.
[b] Determined by early Woese surveys (1987).
[c] Contains pathogens.
[d] Cultivar and genome sequence recently reported (21, 25).

of those sequences are derived from studies of animal feces and have limited diversity. On the other hand, the phylum *Chloroflexi* is represented by comparatively few sequences yet is conspicuous in many environmental settings, for instance, photosynthetic microbial mats worldwide.

Each of the bacterial phyla is itself a branching radiation from the base of the domain. Phylogenetic trees representing the different levels of bacterial diversity, with variable degrees of accuracy, can be downloaded from the public databases. The diversity and richness of the branches among and within the phyla are only beginning to be perceived because natural microbial diversity is so extremely undersampled. Representatives of some phyla, such as *Proteobacteria* (which includes, e.g., *Escherichia* spp., *Rhodobacter* spp., *Rhizobium* spp., *Caulobacter* spp., and *Desulfovibrio* spp.) or *Firmicutes* (which includes *Bacillus* spp., *Clostridium* spp., *Staphylococcus* spp., *Lactobacillus* spp., *Heliobacterium* spp., etc.), are cosmopolitan in the environment and the human experience and express a diversity of metabolisms, with phototrophic, heterotrophic, and autotrophic representatives. Other groups, while geographically distributed, seem more specialized. Representatives of the *Aquificales*, for instance, seem mainly to make a living by hydrogen oxidation, with oxygen or sulfate as the electron acceptor. Although this might seem to be a common physiology, representatives of this phylum so far have been detected only at high temperatures, such as in geothermal springs or hot oil wells. As another example, cyanobacteria seem to be restricted to a phototrophic state, so far as is known. As still other examples, sequences representative of the candidate phyla WS6 and OP11 are widely distributed geographically, but only in anoxic environments. This may indicate rather restricted metabolisms for the kinds of organisms that correspond to the sequences.

The structure of the base of the bacterial rRNA tree—the detail of any branching orders of the bacterial phyla—is not clear at this time. Indeed, the early development of the bacterial phyla may not have been treelike. Rather, it may have been a basal radiation, a "big bang" model for the origin of the bacterial phyla (53). Figure 3 portrays the base of the bacterial tree as a "polytomy," a star radiation clouded by the uncertainties of any estimation at that depth in the bacterial tree. It is common in phylogenetic analyses that particular sequences, for instance, those of some representatives of the *Aquificales* or *Thermotogales*, seem to branch more deeply in the bacterial tree than those of other phyla. This led to the popular notion that such organisms are particularly "primitive" divergences compared to other bacteria. However, the phylogenetic results that indicate greater or lesser depth of branching of the main bacterial phyla are based on only a few SSU rRNA residues and are not seen in analyses using other genes (6, 7, 40). At this

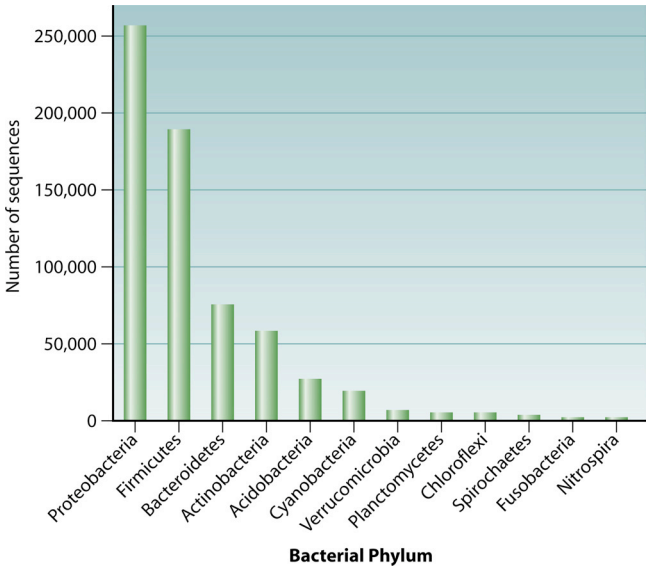


FIG. 4. Distribution of SSU rRNA sequences among the top 12 bacterial phyla. Shown is the SSU rRNA sequence distribution in the SILVA 98 SSU Parc database (52) among the bacterial phyla (Ribosomal Database Project taxonomy) (10) containing the most rRNA sequences.

time, I do not think there is convincing evidence for particularly deeply branching phyla in the bacterial tree.

Any specific relationships among the main bacterial phyla to comprise "superphyla" are unclear at this time, although some results indicate such associations. For instance, some rRNA comparisons conflate the verrucomicrobia, chlamydiae, and planctomycete phyla to indicate deep affiliation, although sequence representation for these little-known groups of organisms is limited (32). Sequence representation in phylogenetic analyses can significantly alter associations of sequences at the base of the bacterial tree, so some phylum shuffling will be inevitable as the databases expand. In general, however, associations seen only in rRNA or other molecular sequence comparisons are subject to uncertainties, as discussed above, and need to be confirmed by other analyses, such as pangenomic comparisons.

THE ARCHAEAL TREE—A WORK IN PROGRESS

The conventional wisdom regarding the phylogeny of archaea is that there are two deeply related taxa, *Crenarchaeota* and *Euryarchaeota*. Based on cultured representation, these organisms collectively have seemed rather simple in their metabolic repertoires. Almost all cultured crenarchaeotes are thermophiles and tend to be lithotrophic, commonly with oxidation or reduction of sulfur. Cultured euryarchaeotes have more varied metabolisms (hence the name "eury-," meaning variable), but most biochemical studies have focused on methanogenesis, a unique property of some archaea. In general, archaea were thought to be restricted to "extreme environments," such as high temperatures and anoxic zones. Culture-independent rRNA sequence surveys have now established, however, that such organisms are ubiquitous in the environment and far more diverse than is represented by cultivars.

Woese identified the taxa *Crenarchaeota* and *Euryarchaeota* in his first assessments of the cultured examples of such organisms (74). However, as additional diverse sequences of cultivars and environmental organisms have poured into the databases (Fig. 2), the coherence of the *Crenarchaeota*/*Euryarchaeota* dichotomy has became questionable. In particular, phylogenetic analyses that incorporate a broad diversity of rRNA sequences, including environmental sequences, break up the coherence of *Euryarchaeota* (56).

Figure 5 illustrates the impact of environmental sequences on the topology of the archaeal rRNA tree. Figure 5A is a tree made with rRNA sequences available in 1993, at the time the *Crenarchaeota*/*Euryarchaeota* dichotomy settled into the conventional wisdom. This tree is consistent with the notion of two main clades of archaea, although the bootstrap values for some deeper branches indicate low confidence in the *Euryarchaeota* clustering. Figure 5B shows an analogous tree of archaeal rRNA sequences but one that includes a broad diversity of current sequences, including environmental sequences. *Crenarchaeota* remains robustly coherent in most trees and includes *Korarchaeota*, previously suggested as a potential third line of archaea (4). However, there is little support for a specific relatedness group that would constitute *Euryarchaeota*. Thus, the base of the archaeal tree seems to be an unresolved polytomy, with any branching orders still to be resolved. The taxon *Euryarchaeota* becomes polyphyletic.

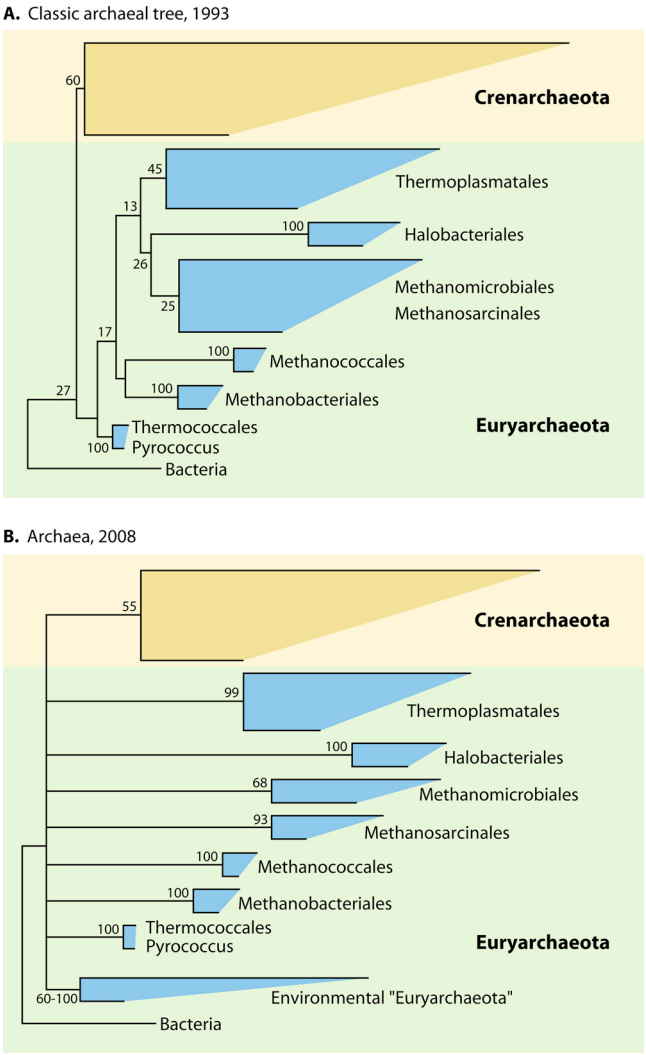


FIG. 5. Archaeal rRNA trees with sequences available in 1993 and 2008. Archaeal SSU rRNA sequences available in 1993 (classic archaeal tree) (A) and in 2008 (B) were used in maximum likelihood bootstrap analysis with RAxML (64) as described previously (56, 57). The boxes represent radiations within the groups, with the long and short dimensions reflecting the line segment lengths within the groups. The sizes of the boxes reflect sequence representation for the groups. The numbers at the base of the boxes are bootstrap percentages. The box labeled Environmental "Euryarchaeota" is not a phylogenetically coherent group.

The taxonomic quandary presented by the expanded archaeal sequence database is more than one of simple classification, what we call organisms. It also influences our perception of relatedness groups among archaea and thereby a proper sense of archaeal diversity. The study of genome sequences that are available for a few representatives of different lines of the basal radiation in Fig. 5B may shed light on the phylogenetic organization. Nevertheless, those organisms with determined genome sequences represent only a small slice of the archaeal diversity identified in the environment. Most of the known archaeal rRNA diversity is represented only by environmental sequences (57). Learning more about this un-

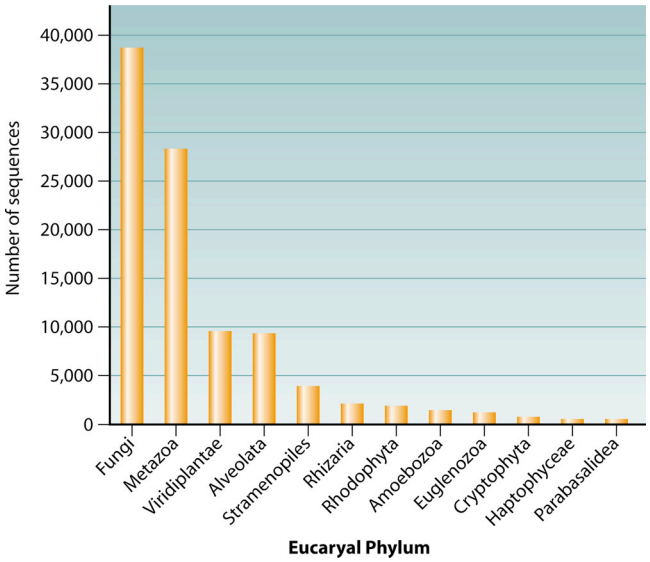


FIG. 6. Distribution of SSU rRNA sequences among the top 12 eucaryal phyla. Shown is SSU rRNA sequence distribution in the SILVA 98 SSU Parc database (52) among the eucaryotic phyla (EMBL taxonomy [66]) containing the most rRNA sequences.

known biology is an obvious challenge to culture (42) and metagenomic (15, 59, 69) efforts.

THE EUCARYAL TREE—ONGOING CONTROVERSY

Phylogenetic studies of eucaryotes have mainly focused on peripheral branches in the overall eucaryal tree, for instance, animals, plants, and their subgroups. The global structure of the eucaryotic tree is not established at this time and is the subject of ongoing controversy. A large obstacle to our perception of the eucaryal rRNA tree is a paucity of data: poor sequence representation from among the many large relatedness groups of organisms that are known from classical studies but are not characterized phylogenetically. Figure 6 shows the sequence distribution of the dozen most represented of the eucaryotic phyla; most eucaryotic rRNA sequences fall into only a few of the groups. Particularly lacking are rRNA (or other) sequences from eucaryotic microbes, which seem to constitute most of the eucaryotic rRNA sequence diversity seen in phylogenetic trees (Fig. 3). Limited sequence representation creates long branches in phylogenetic trees and the consequent susceptibility of the trees to the artifacts of long-branch attraction mentioned above. In any case, limited sequence representation restricts our knowledge of the breadth of eucaryotic phylogenetic diversity.

The first rRNA trees to include those of diverse microbial eucaryotes were assembled by Mitchell Sogin and colleagues (61), and subsequent studies with rRNA sequences have supported the general topology (14, 51). An idealized diagram of the results is incorporated into Fig. 3. The rRNA tree shows an unresolved basal radiation, one line of which radiated, with one line again radiating to form a "crown group" of animals, plants, fungi, and ~10 to 15 other major relatedness groups (41, 61). Branching orders throughout the eucaryal tree are poorly resolved by available data. Specific affiliations of crown group lineages, which might indicate specific branching orders in the crown radiation, are arguable. However, fungi (65, 70), choanoflagellates (38), and the DRIP group (44) have been proposed to be most closely related to the animal line in any crown radiation.

The rRNA model for the structure of the eucaryal tree diagrammed in Fig. 3 is by no means widely accepted as a portrayal of deep eucaryotic evolution. Some models, based on comparisons of some protein sequences and some cellular properties, indicate a basal polytomy of five or six superkingdoms that differentiated into the modern diversity (3, 36, 60). However, there is little agreement as to the composition of the superkingdoms, and not all studies support the superkingdom notion (76). One criticism of the rRNA tree is that the most deeply divergent branches, such as those that lead to diplomonads (e.g., *Giardia* spp.) and trichomonads (e.g., *Trichomonas* spp.), are distant from most reference sequences ("crown group" sequences) in phylogenetic trees and so might represent artifacts of long-branch attraction coupled with accelerated rates of rRNA sequence change. Moreover, some apparent artifacts in rRNA results remain to be resolved. For instance, trees made with rRNA sequences indicate that the lineage represented by microsporidia (obligate intracellular pathogens) is particularly deeply divergent, with diplomonads and trichomonads (Fig. 3) (14, 51). Still, phylogenetic trees made with different protein sequences affiliate microsporidia with fungi (3, 37, 51).

On the other hand, the deeply branching positions of the diplomonad and trichomonad lines seen with rRNA sequences are also indicated by phylogenetic trees derived using various protein sequences (3). Moreover, environmental rRNA sequences have been discovered that diverge from known lineages more or less deeply in the tree and so break up the long branches (14). This makes the potential artifacts of long-branch attraction less likely as explanations for the structure of the base of the eucaryotic tree. Nonetheless, because there is such poor sequence coverage of microbial eucaryotes in general, I think that we currently have little grasp of accurate relationships among the deepest branches in the eucaryotic tree.

PROGRESS AND PROSPECTS

The articulation of an accurate universal ToL, a map of life's evolutionary course, is a lofty goal. Enormous strides in the direction of that goal have been taken as the molecular view of life has developed. The outlines of a universal tree are in place; microbial classification can aspire to a solid foundation based on sequence comparisons; environmental sequences reveal a rich world of unanticipated microbial diversity with significance for the working of ecosystems.

For all the progress, however, these successes also reveal how little we really know about microbial diversity and, consequently, how uncertain is our perception of life's phylogenetic history at the deepest levels. It is only clear at this time that we have merely scratched the surface of an enormous microbial diversity in all the domains, archaeal, bacterial, and eucaryal. To begin to understand the scope of this diversity, continued phylogenetic survey of natural ecosystems has a critical place among the large agendas of the biological

sciences. The results will continue to clarify and confound and bring new insights to our understanding of the global biosphere. Continued discovery of major microbial groups, new arenas for research and resources, seems certain. It also seems certain that future sequence acquisitions will continue to sharpen the molecular view of the deepest branches in the ToL.

The future of microbiology is bright.

ACKNOWLEDGMENTS

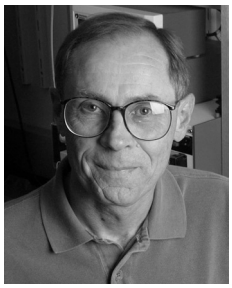
Thanks are due to several colleagues for comments on the manuscript, with special thanks to Charles Robertson for compilations and artwork and to Kirk Harris for artwork.

My research is supported by the Alfred P. Sloan Foundation, the National Institutes of Health, and the National Institute of Occupational Safety and Health.

REFERENCES

1. **Amann, R. I., W. Ludwig, and K. H. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:**143–169.
2. **Angert, E. R., D. E. Northup, A.-L. Reysenbach, A. S. Peek, B. M. Goebel, and N. R. Pace.** 1998. Molecular phylogenetic analysis of a bacterial community in Sulfur River, Parker Cave, Kentucky. Am. Mineralogist **83:**1583–1592.
3. **Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle.** 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290:**972–977.
4. **Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace.** 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. USA **93:**9188–9193.
5. **Bergsten, J.** 2005. A review of long-branch attraction. Cladistics **21:**163–193.
6. **Blank, C. E., S. L. Cady, and N. R. Pace.** 2002. Microbial composition of near-boiling silica-depositing thermal springs throughout Yellowstone National Park. Appl. Environ. Microbiol. **68:**5123–5135.
7. **Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope.** 2001. Universal trees based on large combined protein sequence data sets. Nat. Genet. **28:**281–285.
8. **Cannone, J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell.** 2002. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinform. **3:**2.
9. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283–1287.
10. **Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37:**D141–D145.
11. **Dalevi, D., P. Hugenholtz, and L. L. Blackall.** 2001. A multiple-outgroup approach to resolving division-level phylogenetic relationships using 16S rDNA data. Int. J. Syst. Evol. Microbiol. **51:**385–391.
12. **Darwin, C.** 1882. On the origin of species by natural selection, or the preservation of favored races in the struggle for life, 6th ed. D. Appleton & Co., New York, NY.
13. **Daubin, V., M. Gouy, and G. Perriere.** 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res. **12:**1080–1090.
14. **Dawson, S. C., and N. R. Pace.** 2002. Novel kingdom-level eukaryotic diversity in anoxic environments. Proc. Natl. Acad. Sci. USA **99:**8324–8329.
15. **DeLong, E. F.** 2005. Microbial community genomics in the ocean. Nat. Rev. Microbiol. **3:**459–469.
16. **DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen.** 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. **72:**5069–5072.
17. **Dojka, M. A., P. Hugenholtz, S. K. Haack, and N. R. Pace.** 1998. Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. Appl. Environ. Microbiol. **64:**3869–3877.
18. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783–791.
19. **Felsenstein, J.** 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:**368–376.
20. **Forterre, P., S. Gribaldo, and C. Brochier-Armanet.** 2007. Natural history of the archaeal domain, p. 17–28. *In* R. A. Garrett and H.-P. Klenk (ed.), Archaea: evolution, physiology, and molecular biology. Blackwell, Malden, MA.
21. **Geissinger, O., D. P. Herlemann, E. Morschel, U. G. Maier, and A. Brune.** 2009. The ultramicrobacterium "*Elusimicrobium minutum*" gen. nov., sp. nov., the first cultivated representative of the termite group 1 phylum. Appl. Environ. Microbiol. **75:**2831–2840.
22. **Gogarten, J. P., H. Kibak, P. Dittrich, L. Taiz, E. J. Bowman, B. J. Bowman, M. F. Manolson, R. J. Poole, T. Date, T. Oshima, et al.** 1989. Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86:**6661–6665.
23. **Haeckel, E.** 1866. Generelle Morphologie der Organismen. Verlag Georg Reimer, Berlin, Germany.
24. **Harris, J. K., S. T. Kelley, and N. R. Pace.** 2004. New perspective on uncultured bacterial phylogenetic division OP11. Appl. Environ. Microbiol. **70:**845–849.
25. **Herlemann, D. P., O. Geissinger, W. Ikeda-Ohtsubo, V. Kunin, H. Sun, A. Lapidus, P. Hugenholtz, and A. Brune.** 2009. Genomic analysis of "*Elusimicrobium minutum*," the first cultivated representative of the phylum "*Elusimicrobia*" (formerly termite group 1). Appl. Environ. Microbiol. **75:**2841–2849.
26. **Hiller, N. L., B. Janto, J. S. Hogg, R. Boissy, S. Yu, E. Powell, R. Keefe, N. E. Ehrlich, K. Shen, J. Hayes, K. Barbadora, W. Klimke, D. Dernovoy, T. Tatusova, J. Parkhill, S. D. Bentley, J. C. Post, G. D. Ehrlich, and F. Z. Hu.** 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J. Bacteriol. **189:**8186–8195.
27. **Hillis, D. M.** 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42:**182–192.
28. **Hillis, D. M., J. P. Huelsenbeck, and C. W. Cunningham.** 1994. Application and accuracy of molecular phylogenies. Science **264:**671–677.
29. **Hillis, D. M., C. Moritz, and B. K. Mable.** 1996. Molecular systematics, 2nd ed. Sinauer Associates, Inc., Sunderland, MA.
30. **Holder, M., and P. O. Lewis.** 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. **4:**275–284.
31. **Hori, H., and S. Osawa.** 1979. Evolutionary change in 5S RNA secondary structure and a phylogenic tree of 54 5S RNA species. Proc. Natl. Acad. Sci. USA **76:**381–385.
32. **Hugenholtz, P.** 2002. Exploring prokaryotic diversity in the genomic era. Genome Biol. **3:**reviews0003.1-0003.8.
33. **Hugenholtz, P., C. Pitulle, K. L. Hershberger, and N. R. Pace.** 1998. Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180:**366–376.
34. **Hugenholtz, P., G. W. Tyson, R. I. Webb, A. M. Wagner, and L. L. Blackall.** 2001. Investigation of candidate division TM7, a recently recognized major lineage of the domain *Bacteria* with no known pure-culture representatives. Appl. Environ. Microbiol. **67:**411–419.
35. **Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata.** 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86:**9355–9359.
36. **Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray.** 2005. The tree of eukaryotes. Trends Ecol. Evol. **20:**670–676.
37. **Keeling, P. J., M. A. Luker, and J. D. Palmer.** 2000. Evidence from beta-tubulin phylogeny that microsporidia evolved from within the fungi. Mol. Biol. Evol. **17:**23–31.
38. **King, N., M. J. Westbrook, S. L. Young, A. Kuo, M. Abedin, J. Chapman, S. Fairclough, U. Hellsten, Y. Isogai, I. Letunic, M. Marr, D. Pincus, N. Putnam, A. Rokas, K. J. Wright, R. Zuzow, W. Dirks, M. Good, D. Goodstein, D. Lemons, W. Li, J. B. Lyons, A. Morris, S. Nichols, D. J. Richter, A. Salamov, J. G. Sequencing, P. Bork, W. A. Lim, G. Manning, W. T. Miller, W. McGinnis, H. Shapiro, R. Tjian, I. V. Grigoriev, and D. Rokhsar.** 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. Nature **451:**783–788.
39. **Klappenbach, J. A., P. R. Saxman, J. R. Cole, and T. M. Schmidt.** 2001. rrndb: the Ribosomal RNA operon copy number database. Nucleic Acids Res. **29:**181–184.
40. **Klenk, H. P., T. D. Meier, P. Durovic, V. Schwass, F. Lottspeich, P. P. Dennis, and W. Zillig.** 1999. RNA polymerase of *Aquifex pyrophilus*: implications for the evolution of the bacterial *rpoBC* operon and extremely thermophilic bacteria. J. Mol. Evol. **48:**528–541.
41. **Knoll, A. H.** 1992. The early evolution of eukaryotes: a geological perspective. Science **256:**622–627.
42. **Konneke, M., A. E. Bernhard, J. R. de la Torre, C. B. Walker, J. B. Waterbury, and D. A. Stahl.** 2005. Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature **437:**543–546.
43. **Lagesen, K., P. Hallin, E. A. Rodland, H. H. Staerfeldt, T. Rognes, and D. W. Ussery.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108.
44. **Lang, B. F., C. O'Kelly, T. Nerad, M. W. Gray, and G. Burger.** 2002. The closest unicellular relatives of animals. Curr. Biol. **12:**1773–1778.
45. **Lao-Sirieix, S.-H., V. L. Marsh, and S. D. Bell.** 2007. DNA replication and

cell cycle, p. 93–109. *In* R. Caviccioli (ed.), Archaea: molecular and cellular biology. ASM Press, Washington, DC.
46. **Lefebvre, T., and M. J. Stanhope.** 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. **8:**R71.
47. **Ley, R. E., J. K. Harris, J. Wilcox, J. R. Spear, S. R. Miller, B. M. Bebout, J. A. Maresca, D. A. Bryant, M. L. Sogin, and N. R. Pace.** 2006. Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat. Appl. Environ. Microbiol. **72:**3685–3695.
48. **Ohkuma, M., and T. Kudo.** 1996. Phylogenetic diversity of the intestinal bacterial community in the termite *Reticulitermes speratus*. Appl. Environ. Microbiol. **62:**461–468.
49. **Olsen, G. J., D. J. Lane, S. J. Giovannoni, N. R. Pace, and D. A. Stahl.** 1986. Microbial ecology and evolution: a ribosomal RNA approach. Annu. Rev. Microbiol. **40:**337–365.
50. **Pace, N. R.** 1997. A molecular view of microbial diversity and the biosphere. Science **276:**734–740.
51. **Philippe, H., and A. Germot.** 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol. Biol. Evol. **17:**830–834.
52. **Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner.** 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. **35:**7188–7196.
53. **Puigbo, P., Y. I. Wolf, and E. V. Koonin.** 2009. Search for a 'Tree of Life' in the thicket of the phylogenetic forest. J. Biol. **8:**59.
54. **Rappe, M. S., and S. J. Giovannoni.** 2003. The uncultured microbial majority. Annu. Rev. Microbiol. **57:**369–394.
55. **Rasko, D. A., M. J. Rosovitz, G. S. Myers, E. F. Mongodin, W. F. Fricke, P. Gajer, J. Crabtree, M. Sebaihia, N. R. Thomson, R. Chaudhuri, I. R. Henderson, V. Sperandio, and J. Ravel.** 2008. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. **190:**6881–6893.
56. **Robertson, C. E.** 2008. Ecology, phylogenetics and utrastructure of archaea. University of Colorado, Boulder.
57. **Robertson, C. E., J. K. Harris, J. R. Spear, and N. R. Pace.** 2005. Phylogenetic diversity and ecology of environmental archaea. Curr. Opin. Microbiol. **8:**638–642.
58. **Sapp, J.** 2009. The new foundations of evolution. Oxford University Press, New York, NY.
59. **Schleper, C., G. Jurgens, and M. Jonuscheit.** 2005. Genomic studies of uncultivated archaea. Nat. Rev. Microbiol. **3:**479–488.
60. **Simpson, A. G., and A. J. Roger.** 2004. The real 'kingdoms' of eukaryotes. Curr. Biol. **14:**R693–R696.
61. **Sogin, M. L., J. H. Gunderson, H. J. Elwood, R. A. Alonso, and D. A. Peattie.** 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. Science **243:**75–77.
62. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and J. H. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. USA **103:**12115–12120.
63. **Stackebrandt, E., and B. M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. **44:**846–849.
64. **Stamatakis, A.** 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:**2688–2690.
65. **Steenkamp, E. T., J. Wright, and S. L. Baldauf.** 2006. The protistan origins of animals and fungi. Mol. Biol. Evol. **23:**93–106.
66. **Stoesser, G., W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara, and R. Vaughan.** 2003. The EMBL Nucleotide Sequence Database: major new developments. Nucleic Acids Res. **31:**17–22.
67. **Thomm, M.** 2007. Transcription: mechanism and regulation, p. 139–157. *In* R. Caviccioli (ed.), Archaea: molecular and cellular biology. ASM Press, Washington, DC.
68. **Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D. L. Distel, and M. F. Polz.** 2005. Genotypic diversity within a natural coastal bacterioplankton population. Science **307:**1311–1313.
69. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.
70. **Wainright, P. O., G. Hinkle, M. L. Sogin, and S. K. Stickel.** 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science **260:**340–342.
71. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.
72. **Woese, C. R.** 2000. Interpreting the universal phylogenetic tree. Proc. Natl. Acad. Sci. USA **97:**8392–8396.
73. **Woese, C. R., and G. E. Fox.** 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA **74:**5088–5090.
74. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.
75. **Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin.** 2002. Genome trees and the tree of life. Trends Genet. **18:**472–479.
76. **Yoon, H. S., J. Grant, Y. I. Tekle, M. Wu, B. C. Chaon, J. C. Cole, J. M. Logsdon, Jr., D. J. Patterson, D. Bhattacharya, and L. A. Katz.** 2008. Broadly sampled multigene trees of eukaryotes. BMC Evol. Biol. **8:**14.
77. **Zuckerkandl, E., and L. Pauling.** 1965. Molecules as documents of evolutionary history. J. Theor. Biol. **8:**357–366.
78. **Zwickl, D. J., and D. M. Hillis.** 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. **51:**588–598.



Norman Pace received the Ph.D. degree from the University of Illinois, and has held faculty positions at the National Jewish Hospital and Research Center, the University of Colorado Medical Center, Indiana University, and the University of California, Berkeley. He currently is Distinguished Professor of Molecular, Cellular and Developmental Biology at the University of Colorado, Boulder. Dr. Pace works in two scientific arenas: the first is RNA biochemistry, while on the other hand, Dr. Pace has long worked to develop molecular technology for culture-independent exploration of the natural microbial world. His current efforts range from high-temperature ecosystems to human disease and the indoor environment.