Phylogenetic diversity and ecology of environmental Archaea
Charles E Robertson, J Kirk Harris, John R Spear and Norman R Pace

On the basis of culture studies, Archaea were thought to be synonymous with extreme environments. However, the large numbers of environmental rRNA gene sequences currently flooding into databases such as GenBank show that these organisms are present in almost all environments examined to date. Large sequence databases and new fast phylogenetic software allow more precise determination of the archaeal phylogenetic tree, but also indicate that our knowledge of archaeal diversity is incomplete. Although it is apparent that Archaea can be found in all environments, the chemistry of their ecological context is mostly unknown.

Addresses
Department of Molecular, Cellular, and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

Corresponding author: Pace, Norman R (norman.pace@colorado.edu)

Introduction
Carl Woese first realized that the ribosome, the ubiquitous molecular machine that conducts protein synthesis, offers a way to investigate systematically the relationships between all forms of life. Woese’s approach was to determine the sequences of the RNAs that makes up the ribosome, particularly the small subunit of ribosomal RNA (rRNA). Comparisons of nucleotide sequences of ribosomal genes from different organisms allowed inference of the evolutionary relationships between the organisms: the greater the similarity or difference between the rRNA sequences, the more or less closely related the organisms are. Subsequent work by many investigators formalized the mathematics of sequence comparisons and adopted phylogenetic tree diagrams as the graphical means to display the relationships between sequences (nominally organisms).

Woese’s results using the rudimentary sequencing technology available in the mid-1970s determined that there are three phyla of organisms: Eucarya, Bacteria and Archaea [1]. This sequence-based framework for the description of microbial diversity provided a foundation in the mid-1980s for a significant step in microbial ecology — the culture-independent analysis of rRNA gene sequences from environmental samples [2]. Even early results demonstrated that the microbial world is much larger and more diverse than previously predicted from historical culture studies. Environmental sequences have contributed dramatically to our understanding of archaeal diversity, which continues to expand. Figure 1 shows the accumulation of archaeal rRNA sequences submitted to GenBank. As environmental sequences have accumulated it has become evident that Archaea are a cosmopolitan group that are not limited to ‘extreme’ environments. The expanded sequence collection also affords the opportunity to develop more comprehensive phylogenetic trees than previously possible.

New software has become available that allows phylogenetic trees to be constructed from much larger numbers of RNA sequences than had heretofore been possible. Statistical analysis of phylogenetic trees based on large, predominantly environmentally derived, RNA sequence data sets shows that much of the complex branching traditionally associated with the Archaea is not supported. The complex branching pattern collapses to many branches radiating from single points, known as polytomies or star radiations. Although the wealth of new environmental RNA sequence data show the Archaea to be present in all environments, little progress has been made regarding precisely how the organisms obtain energy from their ecological niches.

Construction of large phylogenetic trees
Various software tools are used to analyze the phylogenetics of rRNA datasets. ARB has become a common phylogenetic software package to use [3]. ARB manages, aligns and annotates sequences as well as managing and printing phylogenetic trees. ARB is often supplemented with additional software packages such as PAUP [4] and more recently MrBayes [5]. PAUP and MrBayes are phylogenetic software tools that are specifically focused on the algorithms used to generate phylogenetic trees.

The most accurate tree-producing algorithms in PAUP and MrBayes are computationally intensive, and therefore are limited in the number of sequences that can be handled. Consequently, compute times are managed by selection of representative sequences that can be used to compute a backbone tree with one of the algorithmic packages. Additional sequences can then be added to the backbone tree using ARB’s ‘parsimony insertion’ feature. Parsimony insertion is an algorithmically simple, and thus
relatively fast, way of adding new taxa to an existing phylogenetic tree. Although expedient, parsimony insertion has risks associated with its use. Owing to the nature of the parsimony insertion algorithm, any attempt at insertion of a sequence to a tree will succeed, even when the sequence is not specifically related to any sequence represented in the tree. This means that some sequences will insert at spurious positions in the backbone tree. Consequently, large phylogenetic trees built by parsimony insertion into small backbone trees run the risk that the more diverse sequences will be positioned incorrectly, which will lead to improper phylogenetic inferences.

Uncertainty in phylogenetic trees
All phylogenetic algorithms have systematic weaknesses. To assess the statistical validity of a particular tree topology, ‘bootstrap analysis’ is commonly used. This technique tests how well particular branching points (nodes) in a tree are supported by the underlying data. In bootstrap analysis, homologous positions within the sequence alignment are randomly resampled and a tree is computed for each modified dataset. The process is repeated many times (at least 100 replicates) and a consensus tree is formulated at the end of the process. Any node that is not adequately supported in the consensus bootstrap tree (we use here 70% for a 95% confidence level [6]) should be eliminated as unsound by the collapse of the associated branch toward the base of the tree.

The repetitive bootstrap analysis process exacerbates the long compute times associated with phylogenetic algorithms. This means that bootstrap analysis has seldom been conducted on large datasets. The most statistically robust method for phylogenetic inference is considered to be ‘maximum likelihood’ (ML) [7]. This method computes the probability that a given tree topology best fits the data. However, ML is computationally demanding, and early versions could handle relatively few sequences. A recent implementation of ML, RAxML [8**], is able to calculate phylogenetic trees from large datasets (hundreds of taxa) using readily available computer hardware in relatively short periods of time (days per tree). This allows a substantial increase in the number of taxa that can be incorporated into backbone trees. The larger the backbone tree, the more likely it is that widely divergent sequences will end up in their most appropriate location in the tree during later parsimony insertion. RAxML is a useful tool for analysis of large numbers of rRNA gene sequences and we use this method to develop the phylogenies discussed below.

Environmental archaeal diversity: collapsing to polytomy
The early overview of archaeal diversity was exemplified by a phylogenetic tree that had two main branches — the Euryarchaeota and the Crenarchaeota. This viewpoint was established in the late 1970s and was based on cultured organisms. The original archaeal trees were generated using less than 20 rRNA gene sequences. Based on the properties of the cultured organisms, it appeared that crenarchaeotes were exclusively high-temperature organisms and that euryarchaeotes, methanogens and halophiles also occupy environments that seem alien to humans. This established the extremophilic archaeal stereotype that persists to the present day. As sequences from environmental samples accumulated it became clear that the archaeal tree was more complicated than expected. The changeover from a picture of archaeal diversity based on cultivars to one based on environmental sequences has accelerated in recent years, as shown in Figure 1. At present, 77% of archaeal rRNA sequences are derived from environmental samples.

Phylogenetic trees of Crenarchaeota and Euryarchaeota based on ~700 full-length rRNA sequences are summarized in Figure 2. The basic form of the two-branch archaeal tree has not changed, even though most of the sequences in the tree are now derived from environmental samples rather than from cultured organisms. After node elimination based on bootstrap analysis (above), some of the basal nodes seen in earlier reports were eliminated. The result is that both Crenarchaeota and Euryarchaeota appear as ‘polytomies’ (star radiations). The results show that the Korarchaeota and the group called FCG, previously thought to be near the base of the archaeal trees, are firmly inside the Crenarchaeota.
position within the tree of the recently discovered Nanoarchaea [9] is unstable in our analysis. The Nanoarchaea tend to associate either at the base of the archaeal tree or are occasionally seen near the Methanopyri of the Euryarchaeota. Our results and the analysis of Brochier et al. [10] led us to place the Nanoarchaea at the base of the Euryarchaeota, as shown in Figure 2b. This conclusion is supported by the presence of a proline tetrad motif that is present in nanoarchaeal and euryarchaeal histone dimers but that is absent from crenarchaeal histone dimers [11]. The halophiles are generally thought to be a sister group to the Methanosarcina, but the bootstrap cutoff of 70% support in RAxML trees does not support this conclusion. The number of singleton sequences shown in the tree is noteworthy. The phylogenetic importance of these sequences will require additional environmental survey data to determine the extent of diversity within these lines of descent. The singletons indicate that fuller perspective on archaeal diversity requires improved sampling statistics; even more sequences are needed.

Archaeal ecology: no longer just extremophiles

Although the phrase ‘archaeal ecology’ is popularly synonymous with ‘extreme’ environments, this is clearly an anthropocentric view. Representatives of Archaea occur everywhere, in samples from ocean water [12], ocean sediments [13], solid gas hydrates [14], tidal flat sediments [15], freshwater lakes [16], soil [17], plant roots [18], peatlands [19], petroleum-contaminated aquifers [20] and the human mouth and gut [21], to cite just a few reports. Archaea commonly occupy a significant fraction of the total microbiota present, typically ~10% of total rRNA phylotypes, and the remainder is composed mainly of bacteria. This proportion of archaeal phylotypes occurs in environments ranging from Yellowstone hot springs [22] to hypersaline mats (JRS and NRP, unpublished) and the deep sea. Archaea can come to dominate total microbial presence in some environments, such as at high temperatures and low pH [23] or in the coldest waters of the deep sea [24]. But what are the ecological roles of such organisms?
Environmental Archaea: possibilities for energy capture

A common theme among known Archaea is use of hydrogen-based energy metabolism. Thus, such organisms are likely to be encountered in anoxic environments, often in syntrophic association with other organisms (for instance engaged in interspecific hydrogen transfer). Beyond such broad generalities, we know little about the physiological properties of Archaea in the environment. Study of the physiological details of environmental organisms and the ecosystem services that they provide is experimentally difficult. Some properties of organisms can be inferred, however, on the basis of their phylogenetic position and the properties of their relatives. Representatives of particular relatedness (phylogenetic groups) are expected to have properties that are common to the entire group. Thus, an environmental rRNA sequence that falls into a clade populated by known methanogens is likely to represent another methanogen in the environment. A caveat to this inferential approach to microbial physiology is the requirement of cultured species that are representative of the particular phylogenetic group so that physiological properties can be determined. In the case of Euryarchaeota there is culture representation for much of the tree. Among the Crenarchaeota, however, only a limited diversity of cultures has emerged, all of which live at high temperatures and are mainly hydrogen-metabolizing. No low-temperature environmental crenarchaeotes has yet been captured for study, so there is essentially no physiological information on this, which is one of the most abundant and well-distributed types of organism on Earth.

Whole-genome approaches to understanding environmental microbial diversity yield an enormity of genetic information [25*]. However, that information is of limited use without correlation to function, roles and niche-occupancy. For instance, Herndl et al. [26**] found that crenarchaeotes fix inorganic carbon at depths of 100 m in the ocean, but what do they use for fuel? Some genomic information has been gathered from a crenarchaeal sponge symbiont Cenarchaeum symbiosum [27] (a representative of the group C1 environmental clade, Figure 2a), but its lifestyle has not been revealed to date. Genome information can shed light on well-characterized metabolic pathways to indicate metabolism that is possibly active, but 25–40% of genes in fully sequenced archaeal genomes are of unknown function. This indicates that potential unknown metabolic pathways will not easily be revealed by current genomic approaches.

Ecology: microbial ecology is chemistry

The databases will continue to swell with environmental sequences. A current challenge to the discipline of microbial ecology is to associate the sequences with the physiological properties of the corresponding organisms and thereby gain insight into their roles in their respective ecosystems. A crucial issue in environmental microbiology will become correlation of different studies to compare particular sequence types in different environmental settings. Traditionally, we have described our large-scale world in terms of place (latitude, longitude, depth, etc.); by medium or matrix (soil or water); by light or dark; by temperature and climate; and by pH or other extremes. The same comparisons have often been used for microbial habitats. Consequently, environmental information deposited with database sequences or even in original publications is commonly inadequate and lacks the chemical information that is essential for correlations between studies. Microbial habitats are complex and are chemically based. Information is needed on potential electron donors and acceptors and on nitrogen sources, etc. Microbial ecology is dictated by the local chemical conditions and such information should become a component of environmental sequence annotation.

Conclusions

The rate of archaeal sequence submission to public sequence databases has increased dramatically in recent years. Most of the new data are rRNA gene sequences derived from environmental samples. Recently developed phylogenetic tree software makes the analysis of large sequence datasets possible with readily available computers. Archaeal phyla, the phylogenetic location of which was previously unstable, assume a fixed position in phylogenetic trees based on large sequence datasets, probably owing to improved sampling statistics. However, the number of sequences in these large trees that do not affiliate with known groups indicates archaeal diversity is not fully sampled. Archaea appear in all environments examined to date. However, in most cases we still have little or no idea of what they are doing in their ecological context. Concerted efforts that utilize genomics and the measurement of appropriate chemistry (with deposition in open databases) as well as development of novel techniques will be required to get to the heart of the question: what are the many kinds of Archaea doing out there anyway?

Acknowledgements

Special thanks to Alexandros Stamatakis for his quick responses to our questions and to our requests for new features in his RAxML software package. This research was supported in part by the NASA Astrobiology Institute.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
  •• of outstanding interest


21. Lepp PW, Brinig MM, Ouermey CC, Palm K, Armitage GC: Methanogenic Archaea and human periodontal disease. Proc Natl Acad Sci USA 2004, 101:6176-6181. To date no evidence exists to show that Archaea, unlike Bacteria and Eucarya, cause human disease. These authors present data that shows Archaea might be indirectly associated with some forms of periodontal disease by way of syntrophic relationships with the bacteria known to cause periodontitis.


26. Herndl GJ, Reinhalter T, Teira E, van Aken H, Veth C, Pernthaler A, Pernthaler J: Contribution of Archaea to total prokaryotic production in the deep Atlantic Ocean. Appl Environ Microbiol 2005, 71:2303-2309. Although the abundance of Archaea below 100 m depth has been known for some time, it was not clear if these organisms were actively metabolizing. These authors show that a high fraction of Archaea in the deep ocean are, in fact, metabolically active and that they might play a significant role in the oceanic carbon cycle.
