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)

Current Opinion in Microbiology 2005, **8**:638–642

This review comes from a themed issue on
Growth and development
Edited by John N Reeve and Ruth Schmitz

1369-5274/$ – see front matter
© 2005 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.mib.2005.10.003

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.

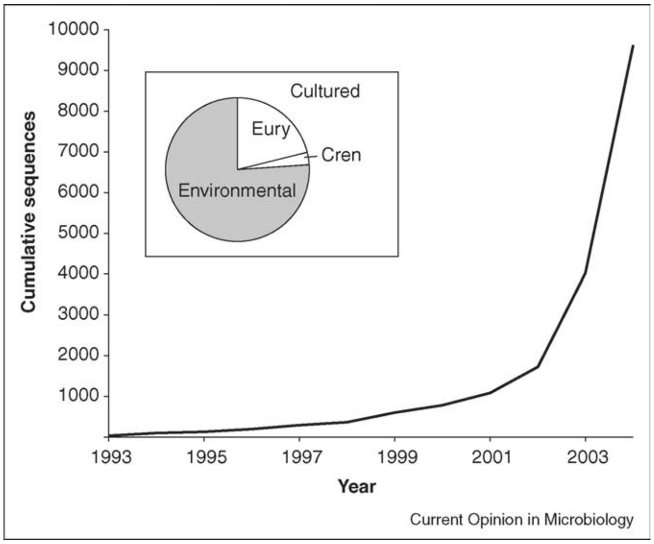
New software has become available that allows phylogenetic trees to be constructed from much larger numbers of RNA sequences than had heretofore been possible. Statistical analysis of phylogenetic trees based on large, predominantly environmentally derived, RNA sequence data sets shows that much of the 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

Figure 1



The cumulative number of archaeal small subunit rRNA sequences submitted to GenBank each year since 1993. Data are obtained from the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov). The inset picture shows the proportion of archaeal rRNA sequences from cultured versus environmental sources. Abbreviations: Cren, Crenarchaeota; Eury, Euryarchaeota.

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.

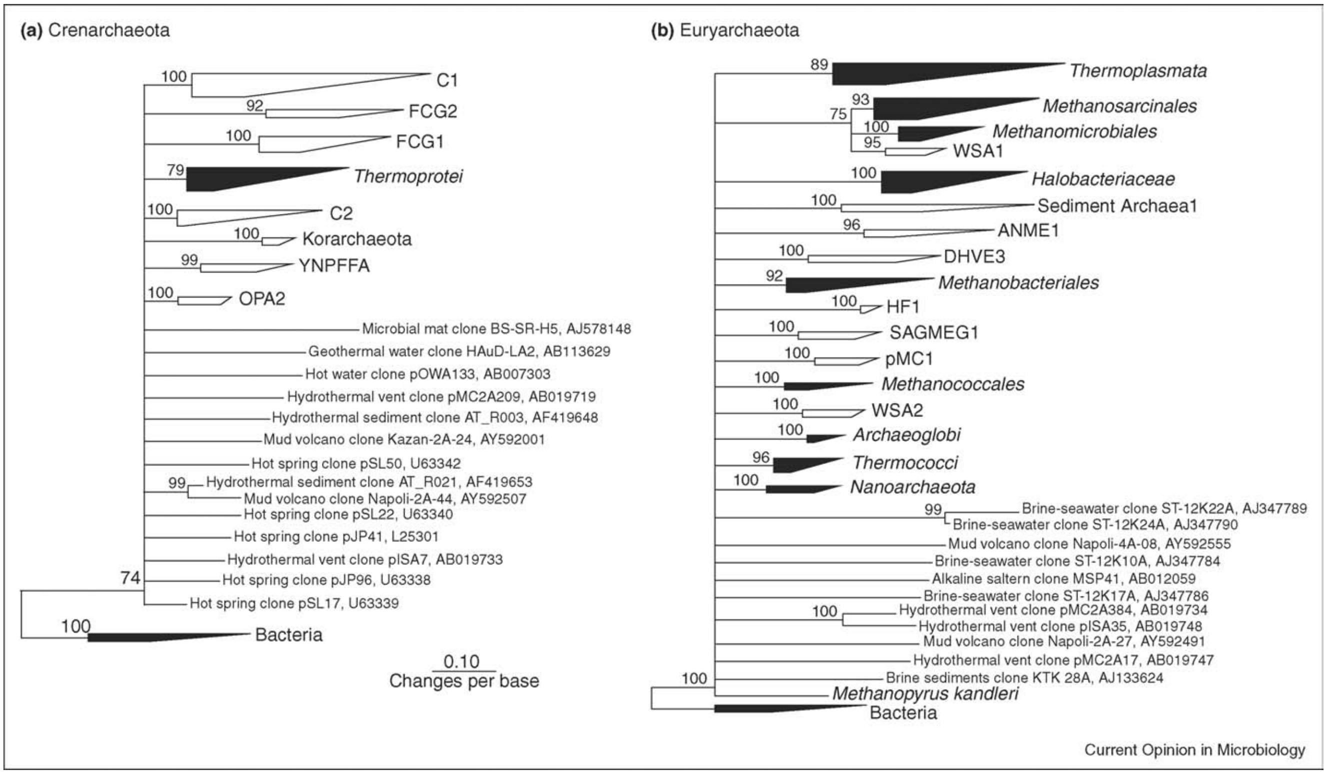
The repetitive bootstrap analysis process exacerbates the long compute times associated with phylogenetic algorithms. This means that bootstrap analysis has seldom been conducted on large datasets. The most statistically robust method for phylogenetic inference is considered to be 'maximum likelihood' (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 extremeophilic 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. The

Figure 2


(a) Crenarchaeota
100
C1
92
FCG2
100
FCG1
79
Thermoprotei
100
C2
100
Korarchaeota
99
YNPFFA
100
OPA2
Microbial mat clone BS-SR-H5, AJ578148
Geothermal water clone HAuD-LA2, AB113629
Hot water clone pOWA133, AB007303
Hydrothermal vent clone pMC2A209, AB019719
Hydrothermal sediment clone AT_R003, AF419648
Mud volcano clone Kazan-2A-24, AY592001
Hot spring clone pSL50, U63342
99
Hydrothermal sediment clone AT_R021, AF419653
Mud volcano clone Napoli-2A-44, AY592507
Hot spring clone pSL22, U63340
Hot spring clone pJP41, L25301
Hydrothermal vent clone pISA7, AB019733
74
Hot spring clone pJP96, U63338
Hot spring clone pSL17, U63339
100
Bacteria
0.10
Changes per base
(b) Euryarchaeota
89
Thermoplasmata
93
Methanosarcinales
75
100
Methanomicrobiales
95
WSA1
100
Halobacteriaceae
100
Sediment Archaea1
96
ANME1
100
DHVE3
92
Methanobacteriales
100
HF1
100
SAGMEG1
100
pMC1
100
Methanococcales
100
WSA2
100
Archaeoglobi
96
Thermococci
100
Nanoarchaeota
99
Brine-seawater clone ST-12K22A, AJ347789
Brine-seawater clone ST-12K24A, AJ347790
Mud volcano clone Napoli-4A-08, AY592555
Brine-seawater clone ST-12K10A, AJ347784
Alkaline saltern clone MSP41, AB012059
Brine-seawater clone ST-12K17A, AJ347786
100
Hydrothermal vent clone pMC2A384, AB019734
Hydrothermal vent clone pISA35, AB019748
Mud volcano clone Napoli-2A-27, AY592491
Hydrothermal vent clone pMC2A17, AB019747
100
Brine sediments clone KTK 28A, AJ133624
Methanopyrus kandleri
Bacteria

Current Opinion in Microbiology

Phylogenetic cladograms generated with the RAxML software from 712 archaeal rRNA sequences that were at least 1250 nucleotides long. **(a)** The Crenarchaeota and **(b)** the Euryarchaeota. The sequences were downloaded from GenBank in February 2005 and were manually aligned in ARB. Three bacterial sequences were used as outgroups. PHYLIP [28] was used to generate 100 bootstrap datasets and to build the consensus tree that resulted from running the 100 datasets through RAxML. Any nodes in the tree that had less than 70% bootstrap support were deleted. Solid colored groups have at least one cultured representative; others are known only from environmental samples. ARB and PHYLIP are commonly used phylogenetic software tools.

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 crenarcheaote 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 crenarcheal 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

1. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms**. *Proc Natl Acad Sci USA* 1977, **74**:5088-5090.

2. Pace NR, Stahl DA, Lane DJ, Olsen GJ: **Analyzing natural microbial populations by rRNA sequences**. *ASM News* 1985, **51**:4-12.

3. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar A, Buchner T, Lai S, Steppi G, Jobb G *et al.*: **ARB: a software environment for sequence data**. *Nucleic Acids Res* 2004, **32**:1363-1371.

4. Swofford D: *PAUP*: Phylogenetic analysis using parsimony (* and other methods)*. Sinauer Associates; 1999.

5. Huelsenbeck JP, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees**. *Bioinformatics* 2001, **17**:754-755.

6. Hillis DH, Bull JJ: **An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis**. *Syst Biol* 1993, **42**:182-192.

7. Holder M, Lewis PO: **Phylogeny estimation: traditional and Bayesian approaches**. *Nat Rev Genet* 2003, **4**:275-284.

8. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for •• maximum likelihood-based inference of large phylogenetic trees**. *Bioinformatics* 2005, **21**:456-463.
The authors compare and contrast RAxML with other commonly used molecular phylogenetic tree building software.

9. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO: **A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont**. *Nature* 2002, **417**:63-67.

10. Brochier C, Gribaldo S, Zivanovic Y, Confalonieri F, Forterre P: **Nanoarchaea: representatives of a novel archaeal phylum or a fast-evolving euryarchaeal lineage related to Thermococcales?** *Genome Biol* 2005, **6**:R42.

11. Cubonova L, Sandman K, Hallam SJ, Delong EF, Reeve JN: •• **Histones in crenarchaea**. *J Bacteriol* 2005, **187**:5482-5485.
The first reported evidence for crenarcheal histones.

12. DeLong EF: **Microbial community genomics in the ocean**. • *Nat Rev Microbiol* 2005, **3**:459-469.
A good overview of the current status of genomics applied to oceanic environmental samples.

13. Knittel K, Losekann T, Boetius A, Kort R, Amann R: **Diversity and distribution of methanotrophic archaea at cold seeps**. *Appl Environ Microbiol* 2005, **71**:467-479.

14. Mills HJ, Martinez RJ, Story S, Sobecky PA: **Characterization of microbial community structure in Gulf of Mexico fas hydrates: comparative analysis of DNA- and RNA-derived clone libraries**. *Appl Environ Microbiol* 2005, **71**:3235-3247.

15. Kim BS, Oh HM, Kang H, Chun J: **Archaeal diversity in tidal flat sediment as revealed by 16S rDNA analysis**. *J Microbiol* 2005, **43**:144-151.

16. Keough BP, Schmidt TM, Hicks RE: **Archaeal nucleic acids in picoplankton from great lakes on three continents**. *Microb Ecol* 2003, **46**:238-248.

17. Ochsenreiter T, Selezi D, Quaiser A, Bonch-Osmolovskaya L, Schleper C: **Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR**. *Environ Microbiol* 2003, **5**:787-797.

18. Simon HM, Dodsworth JA, Goodman RM: **Crenarchaeota colonize terrestrial plant roots**. *Environ Microbiol* 2000, **2**:495-505.

19. Galand PE, Fritze H, Conrad R, Yrjala K: **Pathways for methanogenesis and diversity of methanogenic archaea in three boreal peatland ecosystems**. *Appl Environ Microbiol* 2005, **71**:2195-2198.

20. Kleikemper J, Pombo SA, Schroth MH, Sigler WV, Pesaro M, Zeyer J: **Activity and diversity of methanogens in a petroleum hydrocarbon-contaminated aquifer**. *Appl Environ Microbiol* 2005, **71**:149-158.

21. Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, • Relman DA: **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.

22. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR: **Novel division level bacterial diversity in a Yellowstone hot spring**. *J Bacteriol* 1998, **180**:366-376.

23. Futterer O, Angelov A, Liesegang H, Gottschalk G, Schleper C, Schepers B, Dock C, Antranikian G, Liebl W: **Genome sequence of *Picrophilus torridus* and its implications for life around pH 0**. *Proc Natl Acad Sci USA* 2004, **101**:9091-9096.

24. DeLong EF, Wu KY, Prezelin BB, Jovine RV: **High abundance of Archaea in Antarctic marine picoplankton**. *Nature* 1994, **371**:695-697.

25. Schleper C, Jurgens G, Jonuscheit M: **Genomic studies of • uncultivated archaea**. *Nat Rev Microbiol* 2005, **3**:479-488.
The current status of archaeal genomics, as understood from environmental samples.

26. Herndl GJ, Reinthaler 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.

27. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, Swanson RV: **Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum***. *J Bacteriol* 1998, **180**:5003-5009.

28. Felsenstein J: **PHYLIP phylogeny inference package**. *Cladistics* 1989, **5**:164-166.