J Mol Evol (2002) 54:539–547
DOI: 10.1007/s00239-001-0054-5

JOURNAL OF MOLECULAR EVOLUTION

Using Homolog Groups to Create a Whole-Genomic Tree of Free-Living Organisms: An Update

Christopher H. House,[1] Sorel T. Fitz-Gibbon[2]

[1] Penn State Astrobiology Research Center and Department of Geosciences, Pennsylvania State University, 212 Deike Building, University Park, PA 16802, USA
[2] IGPP Center for Astrobiology and Department of Microbiology and Molecular Genetics, University of California, Los Angeles, CA 90095-1489, USA

Received: 30 May 2001 / Accepted: 10 October 2001

Abstract. Genomic trees have been constructed based on the presence and absence of families of protein-encoding genes observed in 27 complete genomes, including genomes of 15 free-living organisms. This method does not rely on the identification of suspected orthologs in each genome, nor the specific alignment used to compare gene sequences because the protein-encoding gene families are formed by grouping any protein with a pairwise similarity score greater than a preset value. Because of this all inclusive grouping, this method is resilient to some effects of lateral gene transfer because transfers of genes are masked when the recipient genome already has a homolog (not necessarily an ortholog) of the incoming gene. Of 71 genes suspected to have been laterally transferred to the genome of *Aeropyrum pernix,* only approximately 7 to 15 represent genes where a lateral gene transfer appears to have generated homoplasy in our character dataset. The genomic tree of the 15 free-living taxa includes six different bacterial orders, six different archaeal orders, and two different eukaryotic kingdoms. The results are remarkably similar to results obtained by analysis of rRNA. Inclusion of the other 12 genomes resulted in a tree only broadly similar to that suggested by rRNA with at least some of the differences due to artifacts caused by the small genome size of many of these species. Very small genomes, such as those of the two *Mycoplasma* genomes included, fall to the base of the Bacterial domain, a result expected due to the substantial gene loss inherent to these lineages. Finally, artificial "partial genomes" were generated by randomly selecting ORFs from the complete genomes in order to test our ability to recover the tree generated by the whole genome sequences when only partial data are available. The results indicated that partial genomic data, when sampled randomly, could robustly recover the tree generated by the whole genome sequences.

Key words: Homologs — Tree of life — Genome — Archaea — Bacteria

Introduction

Evolutionary relationships and significant evolutionary events can be studied using whole genome sequences by, for example, the building of genomic trees using methods based on the presence and absence of genes in each genome. Several different methods for the generation of trees using gene content have been developed (Fitz-Gibbon and House 1999; Montague and Hutchison 2000; Snel et al. 1999; Tekaia et al. 1999). These different processes for building genome trees can be divided into two broad categories: those based on the presence and absence of suspected ortholog pairs or the "Ortholog method" (e.g., Snel et al. 1999) and those based on the presence and absence of gene families or the "Homolog method" (e.g., Fitz-Gibbon and House 1999).

Correspondence to: C.H. House; *email:* chouse@geosc.psu.edu

These two distinct methods of genome tree building have advantages and disadvantages. The ortholog method seems to be quite effective at recovering an overall average of the different phylogenetic histories for the genes in the genomes studied. Further, because in this method "evolutionary distance" is based on the proportion of orthologs shared between two genomes divided by the size of the smallest genome of the two, the method is resistant to artifacts caused by differing genome size. However, the method can in principle be greatly influenced by lateral gene transfer (as recently transferred genes will appear as orthologs) and by the loss of shared genes or the duplication of unshared genes (Eisen 2000).

In contrast, the homolog method can be adversely affected by greatly reduced genomes, but is resistant to influences of many cases of lateral gene transfer because only the transfer of novel gene families can influence the observed distribution of characters in the tree building process. Furthermore, the expansion of certain gene families through duplication, or the reduction of the size of certain gene families through limited gene loss, has no influence on the tree building process because all gene families are treated equally regardless of their size.

Since the publication presenting the homolog method of genome tree construction (Fitz-Gibbon and House 1999), several more complete genome sequences have become available with which to test and evaluate this method of genomic tree construction, as well as with which to continue to explore genome evolution in a greater diversity of species. The dataset used in this analysis includes genome sequences from 27 species, including 15 free-living taxa. Here, using these genome sequences, we present tree construction using all 27 taxa, tree construction using only the 15 free-living taxa, an analysis of the influence of certain laterally transferred genes, and an evaluation showing that the method is often able to robustly recover the genome tree topology when only partial protein data sets are available.

Materials and Methods

For this analysis, we used all of the published complete genome sequences available at the time (Alm et al. 1999; Andersson et al. 1998; Blattner et al. 1997; Bult et al. 1996; Cole et al. 1998; Consortium 1998; Deckert et al. 1998; Fleischmann et al. 1995; Fraser et al. 1997; Fraser et al. 1995; Fraser et al. 1998; Goffeau et al. 1997; Himmelreich et al. 1996; Kaneko et al. 1996; Kawarabayasi et al. 1999; Kawarabayasi et al. 1998; Klenk et al. 1997; Kunst et al. 1997; Nelson et al. 1999; Parkhill et al. 2000; Read et al. 2000; Smith et al. 1997; Stephens et al. 1998; Tomb et al. 1997; White et al. 1999), plus the soon to be published genome data for the free-living crenarchaeon *Pyrobaculum aerophilum* (Fitz-Gibbon et al. submitted). Trees were constructed both for all 27 taxa and for just the 15 free-living taxa.

Construction of Data Matrices. As previously described (Fitz-Gibbon and House 1999), the first step in this analysis groups proteins based on pairwise sequence similarity. Comparisons were done using the FASTA3 software (Pearson and Lipman 1988), comparing each protein sequence in turn to each of the 27 databases of all protein sequences for each organism. The proteins were grouped if any pairwise similarity score was greater than a preset *z*-score (Pearson 1995; Pearson and Lipman 1988) regardless of the length of the matching region or the relative lengths of the proteins. FASTA3 *z*-scores are based on an extreme value distribution and scaled to a mean of 50 and a standard deviation of 10. The presence or absence of each protein group was scored for each genome to construct the data matrix for phylogenetic analysis. By grouping all recognizable members of gene families into the same group (even protein sequences linked via another intermediate protein sequence or via a fused multidomain protein), protein families of varying sizes among the genomes do not influence our phylogenetic analysis. This process of clustering sequences collapses traditional gene families into a larger group reducing some signal, but cannot create novel homoplasy in the data matrix. The data matrices are available on the World Wide Web at http://www.pyrobaculum.geosc.psu.edu/data/jme1/treedata.html.

Phylogenetic Analysis. Parsimony and distance analyses were performed using PAUP v.4.0b (Sinauer Associates) for a series of data matrices derived using the following *z*-score cut-offs: 150, 170, 190, 200, 300, 500. Bootstrap scores and consistency indices were calculated using PAUP v.4.0b. The consistency index for all characters on a tree is the minimum possible tree length divided by the observed tree length (Farris 1989). The decay index (also called Bremer support) is defined as the number of additional steps required to collapse the branch in question (Bremer 1988) and was calculated using AUTODECAY v.4.0 and PAUP v.4.0b.

Analysis of Partial Genomes. Each of the 15 complete genome sequences of free-living organisms was used in this analysis. First, proteins from each genome were randomly selected until the selected proteins represented 2/3 of the genome's protein dataset. Using these artificial partial datasets, a genomic tree was constructed via the homolog method of genome tree construction. This process was repeated so that 100 different partial datasets were formed producing 100 genomic trees. From these 100 trees, the number of times a certain node of the tree was preserved was determined. A majority consensus tree was formed showing the frequency with which each node was preserved among the 100 trees constructed. The whole process was repeated to study the consequence of genome tree construction using partial genomes with 1/2, 1/3, and 1/4 of the genes present in the whole genome sequences.

Results and Discussion

Genome Tree Construction—Free Living Organisms

A genome tree of 15 free-living taxa was constructed using the homolog method (Fig. 1). For the most part, the result is well supported as shown by the high bootstrap values and high decay indices (Fig. 1B) and by the consistency of the resulting tree across a wide range of different z-score cut-offs (Fig. 1C). The tree shown in Fig. 1A is very similar to that formed based on the small subunit rRNA (Woese 1987). The result is also consistent with the published 11 taxa tree produced using the same method (Fitz-Gibbon and House 1999). The principal difference between these results is the inclusion of four additional genomes.

Two of the four additional genomes were from taxa that are members of archaeal groups represented by other

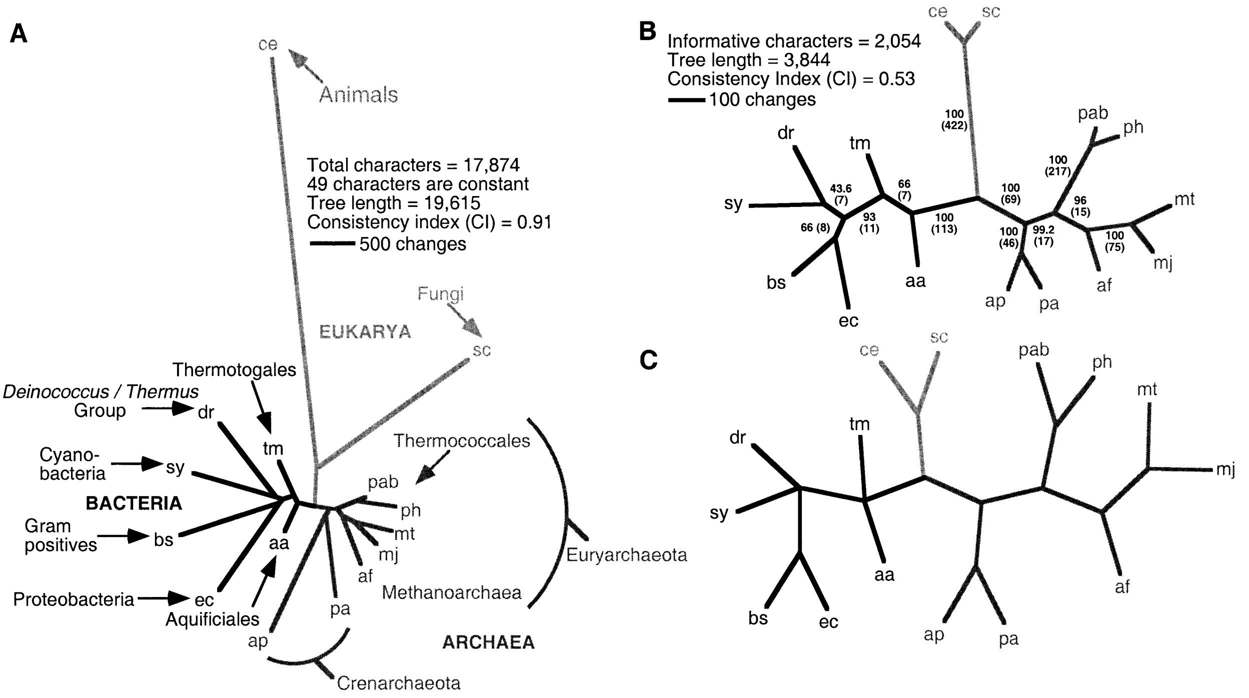

Fig. 1. Phylogenetic analysis of 15 free-living microorganisms, *Aeropyrum pernix* K1 (ap), *Aquifex aeolicus* (aa), *Archaeoglobus fulgidus* (af), *Bacillus subtilis* (bs), *Caenorhabditis elegans* (ce), *Dienococcus radiodurans* (dr), *Escherichia coli* (ec), *Methanobacterium thermoautotrophicum* (mt), *Methanococcus jannaschii* (mj), *Pyrobaculum aerophilum* (pa), *Pyrococcus abyssi* (pab), *Pyrococcus horikoshii* (ph), *Saccharomyces cerevisiae* (sc), *Synechocystis* sp. (sy), and *Thermotoga maritima* (tm), using the presence/absence of families of protein-encoding genes in each genome as characters. **(A)** The single most parsimonious phylogram (unrooted), produced by use of a *z*-score cutoff of 170 as the criterion for single linkage clustering of homologous protein groups. **(B)** The same phylogram, plotted with only the parsimony-informative characters. Because 15,771 of the 17,874 protein groups identified at this z-score cutoff are unique to a single taxon (and therefore phylogenetically uninformative), the long terminal branches evident in **A** are not present in **B.** Values obtained from 500 bootstrap replicates are listed together with Decay Indices (in parentheses). **(C)** The consensus topology for all of the tested z-score cutoffs in the range 150–500 for both maximum parsimony and distance (neighbor joining) methods. Branch and bound searches were used in this case.

taxa resulting in additional representation of the Crenarchaeota by *Aeropyrum pernix,* and of the Thermococcales by *Pyrococcus abyssi.* In both of these cases, the new taxa included in the analysis were placed on the genomic tree clustered with the other members of its rRNA taxonomic group. Another of the additional taxa was *Caenorhabditis elegans,* which clustered with the other Eukarya as expected. The placement of these three taxa with their expected relatives suggests that at some significant level the genomic content of closely related species (as discerned by rRNA sequences) are similar.

An additional genome from *Deinococcus radiodurans* was also added. In this case, the taxon belongs to a bacterial group not represented on the tree published by Fitz-Gibbon and House (1999). In the new analysis shown in Fig. 1, *Deinococcus radiodurans* clusters weakly with the cyanobacterial species *Synechocystis* sp. Although this pairing is not statistically well supported in the genomic tree, the pairing of the *Deinococcus/Thermus* group with the Cyanobacteria is supported by an insertion or deletion in the Hsp40 gene (Gupta 1998). Generally, the support of resolving the relationships of the Gram positives, Proteobacteria, the Cyanobacteria, and the *Deinococcus/Thermus* group is very difficult in this analysis suggesting that these taxa either have undergone significant lateral gene transfer, significant gene loss, or that these taxa were part of a rapid bacterial diversification.

Genome Tree Construction—All Organisms

In order to investigate further the use of homologous protein families as phylogenetic characters, we have also built a genomic tree using the 27 complete genome sequences available at the time of analysis (Fig. 2). Although the result is broadly similar to that shown in Fig. 1 (and to the rRNA tree), the tree is in detail rather poor with the relationships of many of the non-free-living taxa not properly resolved. Specifically, the method has correctly paired *Haemophilus influenzae* with its close relative *Escherichia coli* and also united the two Spirochetes (*Borrrelia burgdorferi* and *Treponema pallidum*) together, but we assume it has failed to present the proper phylogeny by failing to unite the *Mycoplasma* with other gram positives or to unite the Proteobacteria together in a clade. It appears that the incorrectly placed taxa are misplaced due to an artifact caused by small genome size. In principle, the homolog method of whole-genome tree construction could be influenced by substantial gene

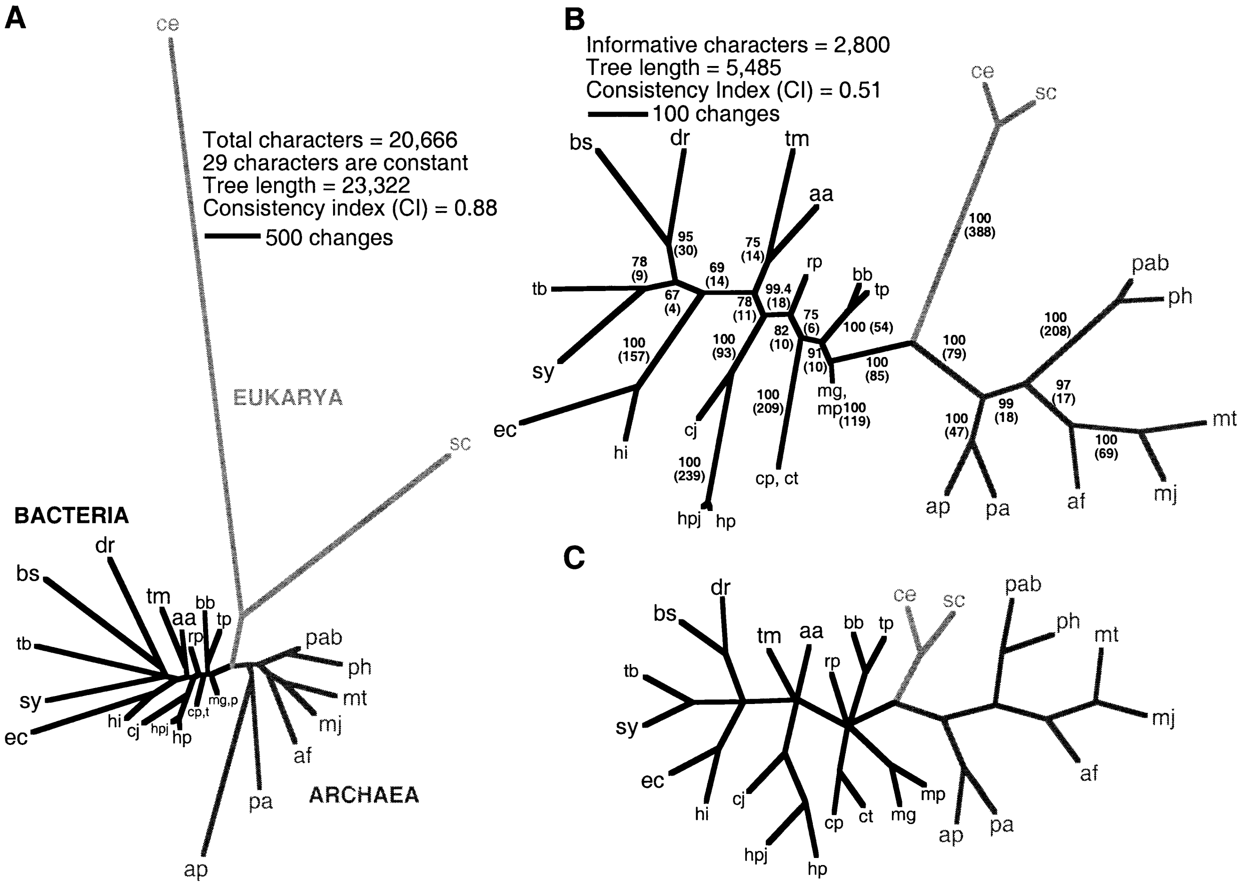

Fig. 2. Phylogenetic analysis of 27 microorganisms, *Aeropyrum pernix* K1 (ap), *Aquifex aeolicus* (aa), *Archaeoglobus fulgidus* (af), *Bacillus subtilis* (bs), *Borrelia burgdorferi* (bb), *Caenorhabditis elegans* (ce), *Campylobacter jejuni* (cj), *Chlamydia pneumoniae* (cp), *Chlamydia trachomatis* (ct), *Dienococcus radiodurans* (dr), *Escherichia coli* (ec), *Haemophilus influenzae* (hi), *Helicobacter pylori* (hp), *Helicobacter pylori* (hpj), *Methanobacterium thermoautotrophicum* (mt), *Methanococcus jannaschii* (mj), *Mycobacterium tuberculosis* (tb), *Mycoplasma genitalium* (mg), *Mycoplasma pneumoniae* (mp), *Pyrobaculum aerophilum* (pa), *Pyrococcus abyssi* (pab), *Pyrococcus horikoshii* (ph), *Rickettsia prowazekii* (rp), *Saccharomyces cerevisiae* (sc), *Synechocystis* sp. (sy), and *Thermotoga maritima* (tm), *Treponema pallidum* (tp), using the presence/absence of families of protein-encoding genes in each genome as characters. Non-free-living taxa are shown with smaller font. **(A)** The single most parsimonious phylogram (unrooted), produced by use of a *z*-score cutoff of 170 as the criterion for identification of homologous protein groups. **(B)** The same phylogram, plotted with only the parsimony-informative characters. Because 17,837 of the 20,666 protein groups identified at this *z*-score cutoff are unique to a single taxon (and therefore phylogenetically uninformative), the long terminal branches evident in **A** are not present in **B.** Values obtained from 1000 bootstrap replicates are listed together with Decay Indices (in parentheses). **(C)** The consensus topology for all of the tested *z*-score cutoffs in the range 150–500 for both maximum parsimony and distance (neighbor joining) methods. A heuristic search was required with this large number of taxa.

loss from a specific genome. This particular artifact can easily account for the placement of reduced genomes (such as *Mycoplasma*) at the base of the Bacteria. One interesting relationship found in Fig. 2, but not easily explained via gene loss, is the placement of *Mycobacterium tuberculosis* with the cyanobacterium *Synechocystis* sp. Results from looking at several of the more parsimonious trees formed when the *Mycobacterium* genome is added to the analysis of free-living-taxa suggest that this genome has an affinity with both *Synechocystis* and *Deinococcus* as several combinations of these three taxa are found. Perhaps the High GC Gram positive bacteria belong to a clade containing the cyanobacteria and the Deinococcus/Thermus group rather than to a clade containing the Low GC Gram positive bacteria. This relationship is consistent with that found for α^{70}-type sigma factors of group 1 and group 2 during initial investigations (Gruber and Bryant 1998).

In the analysis presented in Fig. 2, the only non-free-living genomes larger than the smallest genome of a free living taxon, the 1.5 Mb genome of *Aquifex aeolicus,* are the genomes of *Campylobacter jejuni, Helicobacter pylori,* and *Mycobacterium tuberculosis.* As just discussed, *Mycobacterium tuberculosis* may not belong in a united clade of gram positive bacteria, and so its misplacement is not confirmed at present. *Campylobacter jejuni* and *Helicobacter pylori* will join the other Proteobacteria (*Escherichia coli* and *Haemophilus influenzae*) if either of two changes is made to the analysis. First, they will be united with the other Proteobacteria if there is no smaller genome than 1.6 Mb (removing *Aquifex aeolicus* from the analysis). This solution works because it removes other small genomes that could attract the genomes in question. However, we believe that the relatively small genomes of *Campylobacter jejuni* and *Helicobacter pylori* are not representative of the early Proteobacteria (Fitz-Gibbon and House 1999) rendering such solutions unsatisfactory until a greater diversity of Proteobacteria can be included in the analysis. Second, *Campylobacter jejuni* and *Helicobacter pylori* will unite with the other Proteobacteria if a stepwise matrix is used to count gene family loss less than gene family gain with losses counted as one step and gains counted as five steps. This has the effect of somewhat reversing the artifact caused by a reduced genome by not penalizing gene loss as severely as lateral gene transfer. The majority of our consistent characters in the tree shown in Fig. 1, however, are novel gene family gains, and therefore, such a stepwise matrix has unintentional consequences on the tree building process and is not an adequate solution to the problem presented by reduced genomes. One possible solution may lie in using a tree building process in which gene family gain and gene family loss are counted equally for consistent characters, but gains are counted more severely than losses when characters are inconsistent (implying that the there has been either gene loss or lateral gene transfer).

Generally, the results shown in Fig. 2 indicate that our homolog method of whole genome tree construction can be adversely affected by derived genomes that are not typical of the ancestral state of the group they are representing. This result is by no means surprising (Wiley 1981) and only indicates that care must be taken during taxon selection if a phylogenetically correct tree of life is to be constructed using the presence and absence of protein families. Without further specific knowledge of genome evolution, we recommend general exclusion of non-free-living taxa, resulting in a pruned dataset where *Aquifex aeolicus* has the smallest genome containing only about 1500 protein-encoding genes (Deckert et al. 1998).

Comparison of Our Phylogeny with Ones Obtained Using Other Genomic Methods

Several different methods for the generation of "trees of life" using gene content have been proposed (Fitz-Gibbon and House 1999; Snel et al. 1999; Tekaia et al. 1999). Each of these systems produces a tree with three domains of life robustly separated. Furthermore, each method seems to correctly capture additional features of microbial phylogenetics. However, some of the details of each tree differ, suggesting that the methods are sensitive to different influences.

Snel et al. (1999) have presented a whole genomic tree constructed based on an "evolutionary distance" calculated as the proportion of orthologs shared between two genomes divided by the size of the smallest genome of the two. This method of tree building is quite effective at clustering the members of the Proteobacteria, the Low GC Gram positive bacteria, and the Spirochaetales, but fails to form a monophyletic clade for all of the Gram positive bacteria just as our method fails to do so (discussed above). The original published tree presented using this method found *Aquifex* as a sister taxon with *Synechocystis* (Snel et al. 1999). However, when more taxa were included in an updated version of the tree, *Aquifex* was found to fall at the base of the Bacteria (Huynen et al. 1999) as it does in rRNA trees (Stetter 1996) and in the whole genomic tree presented here (Fig. 1). Finally, the published updated tree finds *Thermotoga maritima* as a sister taxon to the Spirochaetales rather than near the base of the tree as suggested by rRNA (Woese 1987) and by our analysis (Fig. 1).

Two additional methods of whole genome tree construction based on a hierarchical classification system have been presented by Tekaia et al. (1999). In both trees presented using this system, *Aquifex aeolicus* is not at the base of the Bacteria as suggested by rRNA, but rather clustered with *Rickettsia* and *Chlamydia.* It is possible that *Aquifex* is misplaced upon the rRNA tree due to long branch attraction toward the root of the Bacteria and that *Aquifex* is misplaced toward the root of our tree due to its relatively small genome size. Recent conventional phylogenetic analysis of 39 conserved proteins, however, revealed a deeply branched position for *Aquifex* and *Thermotoga* (Hansmann and Martin 2000), suggesting that *Aquifex* is correctly placed on our tree in Fig. 1 and misplaced on the tree presented by Tekaia et al. (1999) possibly due to the inclusion of non-free-living taxa.

Another genomic method is the analysis of combined datasets containing large numbers of conserved proteins. For example, Hansmann and Martin (2000) compiled a dataset of 39 conserved proteins analyzed the dataset phylogenetically taking special note of the significant influence that the inclusion or exclusion of highly variable sites has on the resulting tree. Although they do not present a single tree, their results taken as a whole generally place *Aquifex* and *Thermotoga* as deeply branched, as does our result. However, other relationships within the Bacteria are not consistently resolved by Hansmann and Martin (2000), making a comparison difficult and suggesting that even a large number of proteins may be insufficient to resolve phylogenetic relationships between the bacterial orders.

Brown et al. (2001) have also recently presented a phylogenetic tree based on the combined phylogenetic analysis of a large dataset of proteins. In this work, they present two different reconstructions. Their initial result has 23 proteins and places the Spirochaetales as the earliest diverging bacterial lineage, while placing *Aquifex* and *Thermotoga* in a derived clade including the Proteobacteria and Cyanobacteria. Their second tree is based on the same dataset after the removal of nine genes that the authors believe have been laterally transferred (Brown et al. 2001). The new result shows *Aquifex* and *Thermotoga* as early diverging lineages, while many of the other relationships within the Bacteria are also altered from their placement in the first tree. Besides the placement of *Aquifex* and *Thermotoga* at the base of the Bacteria, the second tree presented by Brown et al. (2001) has some additional features in common with our result. For example, the second tree presented by Brown et al. (2001) unites *Deinococcus* with *Synechocystis* and does not place the High GC Gram positive bacteria (e.g., *Mycobacterium tuberculosis*) in a monoplyletic clade with Low GC Gram positive bacteria. Further, Brown et al. (2001) find that in some of their trees built by neighbor joining the High GC Gram positive bacteria are a sister group to the Cyanobacteria/Deinococcus clade, which is a similar arrangement to that suggested by our analysis of the *Mycobacterium* genome discussed above. Beyond these similarities, our results differ from those obtained by Brown et al. (2001) when it comes to the arrangement of the orders within the bacterial domain as this part of the tree is hard to resolve for any of the techniques compared. In fact, Brown et al. (2001) present several alternative arrangements found within their neighbor joining trees for these relationships.

In general, when methods of whole genomic tree construction are compared, it is clear that the different methods produce trees that differ in subtle details of topology within the Bacteria. However, in contrast, most of the methods produce the same general topology for the Archaea including the separation of the domain into the two kingdoms of the Crenarchaeota and Euryarchaeota and the pairing of the methanogens together. There is an interesting relationship found within the Archaea both by Hansmann and Martin (2000) and by Brown et al. (2001) that differs from our results. The results presented by both of these papers place *Pyrococcus* as a sister taxon to the methanogens, whereas our analysis places *Archaeoglobus* as the sister taxon to the methanogens. With more research, this surprising difference may help to critically evaluate these various methods of tree building. *Archaeoglobus* is micromethanogenic (Stetter 1988) and contains many of the biochemical pathways found within methanogens (Klenk et al. 1997). Future work may reveal that the unique gene families shared by *Archaeoglobus* and the methanogens are not really synapomorphies, but rather homoplasies that were transferred, indicating that the placement found by our whole genome method was altered by the genetic transfer of the particular biochemistry shared by these three organisms. Alternatively, future work may find that *Archaeoglobus* belongs in a clade uniting it with the methanogens based on vertical decent, that the unique gene families it shares with the methanogens are synapomorphies, and that the protein trees are incorrect due to one of the many artifacts that can adversely effect sequence analysis.

Table 1. Results from the investigation of phylogenetic inconsistency induced by suspected lateral gene transfers from Bacteria to the genome of *Aeropyrum pernix* (ap) without and with the genome of *Pyrobaculum aerophilum* (pa) included in the analysis

155	genes pass "test" for lateral gene transfer from a bacterium to ap without pa included in the analysis (112 of these are cases in which the gene family represented by the gene was already present in ap)
71	genes pass "test" for lateral gene transfer from a bacterium to ap with pa included in the analysis (45 of these are cases in which the gene family represented by the gene was already present in ap)
26	of these 71 genes are members of 24 unique parsimony informative gene families on the free-living genomic tree presented in Figure 1 (3 of these are genes from a single gene family that was already present in ap)
20	of these 26 genes represent 18 unique gene families that are present in ap, but absent in pa suggesting the lateral gene transfer may have induced homoplasy in the Crenarchaeota (3 of these are genes from a single gene family that was already present in ap)
7	of these 20 genes represent 7 unique gene families whose presence in the Archaea appears to be the result of a single gene transfer to ap from a bacterium

Investigation of Phylogenetic Inconsistency Induced by "Suspected" Lateral Gene Transfers from Bacteria to the Genome of Aeropyrum pernix

Faguy and Doolittle (1999) have suggested that many cases of lateral gene transfer from the Bacteria to the genome of *Aeropyrum pernix* (an Archaea) can be identified by using BLAST to search each *Aeropyrum pernix* gene against a library of genes for other organisms (Bacteria, Archaea, and Eukaryotes) using 50 BLAST bits as the minimum score that can be considered a match. Faguy and Doolittle claim that if the highest bacterial match is 10 BLAST bits higher than the highest non-*Aeropyrum* archaeal match, then it is probably a case of lateral gene transfer (Faguy and Doolittle 1999). We decided to use this criterion to investigate the effect laterally transferred genes have on our own genome tree analysis even though this test for lateral gene transfer is probably quite prone to false positives.

The results are shown in the Table 1. Without the genome *Pyrobaculum aerophilum,* this method identifies 155 genes suspected to have been transferred into the *Aeropyrum* genome. However, this number drops to 71 when *Pyrobaculum aerophilum* is included in the analysis, demonstrating that the results are highly dependent on how many archaeal genomes have been sequenced and included in the analysis. This dependence on taxa sampling results from the fact that the methodology used to identify genes suspected of having been transferred assumes that a transfer has occurred if the strongest match is to the Bacteria even though an even stronger archaeal match may exist outside of the taxa sampled. For example, Faguy and Doolittle (1999) identify the nitrate reductase of *Aeropyrum pernix* as a clear example of a lateral gene transfer from the Bacteria because its strongest match is to the Bacteria. However, there is a similar nitrate reductase in the archaeal genome of *Pyrobaculum aerophilum* (Ladner submitted) making the exact history of this gene unclear. In principle, the known distribution of this archaeal nitrate reductase can be explained by: (1) this gene having been present in the common ancestor of the Crenarchaeota and the Bacteria, (2) having been transferred into the Crenarchaeota prior to the divergence of *Pyrobaculum* and *Aeropyrum,* or (3) having been transferred twice (e.g., once from the Bacteria into one of these crenarchaeal lineages and then a second time between *Pyrobaculum* and *Aeropyrum*). This case demonstrates the limitations of such an analysis and the general difficulties in identifying clear cases of lateral gene transfer. It seems, therefore, that this type of analysis would overestimate the frequency of lateral gene transfer. Nevertheless, we used these results as a rough proxy for the frequency of lateral gene transfer between domains in order to investigate the degree to which lateral gene transfer between distantly related organisms can adversely affect our character set used for tree construction.

Of the 71 genes passing the test and therefore "suspected" to be cases of lateral gene transfer to the *Aeropyrum pernix* genome from the Bacteria, 45 are from gene families already present in the *Aeropyrum pernix* genome and, therefore, cases in which the possible lateral gene transfer would not have altered the character set because the character set is constructed at the gene family level. Furthermore, only 20 of the 71 "transferred" genes are members of gene families (18 gene families in all) present in *Aeropyrum pernix* but absent in the genome of *Pyrobaculum aerophilum,* cases in which a lateral gene transfer may have induced homoplasy in the Crenarchaeota when our character set was constructed. Of these 20 genes, only seven genes are cases in which the gene family was otherwise absent from all of the Archaea on the tree. The characters represented by these seven gene families would be consistent with the tree topology shown in Fig. 1 if it had not been for the gene transfer event. Therefore, in these seven cases, it appears as though a single gene transfer event has induced novel homoplasy in our character data set used for tree building. For the cases where the gene family is present in other Archaea studied besides *Aeropyrum,* the archaeal distribution of the gene families could be explained most parsimoniously in three cases with parallel gene loss rather than lateral gene transfer, in five cases, equally with parallel gene loss as with lateral gene transfer, and in only three cases, with lateral gene transfer rather than parallel gene loss. So, these results suggest that while only seven cases of lateral gene transfer to *Aeropyrum* induced novel homoplasy in our character matrix, perhaps an additional eight cases of lateral gene transfer to *Aeropyrum* would have induced novel homoplasy had that character otherwise been consistent in the Archaea. In either case, this is a small number relative to the decay index of 46 supporting the node uniting *Aeropyrum* with *Pyrobaculum,* indicating significantly more lateral transfers of gene families from the Bacteria would be necessary to alter the placement of *Aeropyrum* on our tree.

In general, in our investigation of phylogenetic inconsistency induced by "suspected" lateral gene transfers, we found that identifying cases of lateral gene transfer from Bacteria to the genome of *Aeropyrum pernix* was difficult. We also find that by building our character matrix with gene families rather than orthologs, some cases of lateral gene transfer will be masked. Furthermore, the level of inter–domain transfer does not seem to be great enough to impact the topology of our tree. This result does not, however, address the level of intra–domain transfer, which is expected to be more problematic.

Analysis of Partial Genome Sequences Using Homolog Method of Tree Construction

In order to evaluate the ability of the homolog method to recover a tree topology when only partial genomic data is available, we conducted an analysis using partial data sets randomly generated (see methods) from the 15 free-living genomes used in the tree shown in Fig. 1. Figure 3A shows the results of 100 replicate analyses of partial genomes consisting of two-thirds of the original protein-encoding genes. Generally, the plurality consensus tree formed from the 100 replicates is similar to the tree shown in Fig. 1 demonstrating that the homolog method of tree construction can recover the phylogenetic information from partial genome datasets. Further, almost all of the relationships found within the Archaea are preserved in all 100 replicate experiments. The exception is the relationship of *Archaeoglobus* to that of *Pyrococcus.* In 85 of the 100 replicates, this relationship was the same as shown in Fig. 1A. However, in the remaining 15 replicates, the two branches were swapped. In contrast, the relationships of the bacterial taxa to each other are not preserved in all 100 replicates, but rather in only a plurality of the replicates. The pairing of *Escherichia coli* with *Bacillus subtilis* was the most reproducible feature found within the Bacteria appearing in 88 of the 100 replicates.

The analysis of datasets containing a half, a third, and a quarter of the genes present in the whole genome dataset is shown in Fig. 3B. The tree topology has changed with the pairing of *Aquifex* and *Thermotoga* as less genome coverage is used. Perhaps as fewer genes are included in the analysis, the broad similarity of these two genomes overcomes specific synapomorphies shared between *Thermotoga* and the other Bacteria. Surprisingly, many of the relationships found to be robust in the original analysis continue to be found with a high degree of reproducibility even when only a quarter of the protein-encoding genes from each genome are being used. This suggests that it may be possible to use inexpensive low

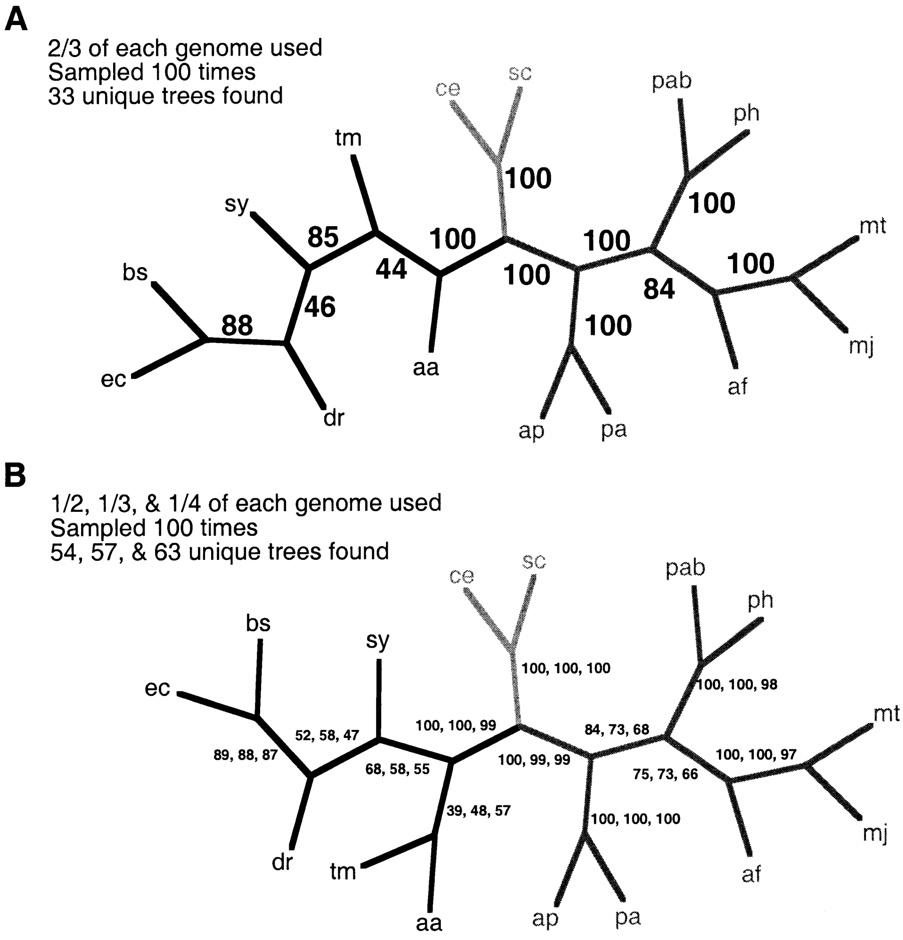

Fig. 3. Results from plurality consensus tree construction using randomly sampled fractionations (2/3, 1/2, 1/3, and 1/4) of each genomic set of predicted proteins for the 15 free-living taxa analyzed in Fig. 1. The numbers at each node indicate the number of times out of 100 replicates that the particular node was preserved.

coverage genome sequencing to survey a wide range of microorganisms in order to produce a genomic Tree of Life that has the taxonomic coverage found only for small subunit rRNA trees.

Conclusions

We conclude that a robust tree can be constructed using groups of homologs as phylogenetic characters, that this method is more resistant to the influences of lateral gene transfer than similar ortholog-based tree building methods, and that partial genome data may be of use in the tree building process (at least when random sampling is invoked). With respect to the phylogenetic relationships studied with our genomic analysis, we have the most confidence in: (1) the separation of the Bacteria and Archaea into separate domains, (2) the separation of the Archaea into two kingdoms—the Crenarchaeota and Euryarchaeota, and (3) the pairing of the *Pyrococcus* species together as well as the pairing of the methanogens studied together. We have slightly less confidence in the placement of *Aquifex* near the base of the Bacteria and in the placement of the Eukarya outside of the Archaea due to the possibility that the loss of gene families in these groups have moved their position in our tree toward the root. Finally, we have low confidence about the exact relationship of the Proteobacteria, the Cyanobacteria, and the Gram positive bacteria as this is a difficult area of the tree to resolve. Future studies will be designed to directly investigate these relationships on the whole genomic scale.

Acknowledgments. We thank Bruce Runnegar (University of California, Los Angeles) for useful discussions. We also thank two anonymous reviewers for useful suggestions. This work was supported by a grant to Jeffrey H. Miller from the National Institutes of Health (GM57917), by the Penn State Astrobiology Research Center and the UCLA Center for Astrobiology, NASA Astrobiology Institute. IGPP publication # 5666.

References

Alm RA, Ling LS, Moir DT, King BL, Brown ED, Doig PC, Smith DR, Noonan B, Guild BC, deJonge BL, et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori.* Nature 397:176–180

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UC, Podowski RM, Näslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140

Blattner FR, Plunkett Gr, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Bremer K (1988) The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nature Genet 28:281–285

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science 273:1058–1073

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE, et al. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393:537–544

Consortium (1998) Genome sequence of the nematode *C. elegans:* a platform for investigating biology. Science 282:2012–2018

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M, et al. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature 392:353–358

Eisen JA (2000) Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol 3:475–480

Faguy DM, Doolittle WF (1999) Lessons from the *Aeropyrum pernix* genome. Curr Biol 9:R883–R886

Farris JS (1989) The retention index and the rescaled consistency index. Cladistics 5:417–419

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Fitz-Gibbon ST, House CH (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res 27: 4218–4222

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature 390:580–586

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, et al. (1998) Complete genome sequence of *Treponema pallidum,* the syphilis spirochete. Science 281:375–388

Goffeau A, Authora A, Authorb B (1997) The yeast genome directory. Nature 387:5

Gruber TM, Bryant DA (1998) Characterization of the group 1 and group 2 sigma factors of the green sulfur bacterium *Chlorobium tepidum* and the green non-sulfur bacterium *Chloroflexus aurantiacus.* Arch Microbiol 170:285–296

Gupta RS (1998) Protein Phylogenies and Signature Sequences; A Reappraisal of Evolutionary Relationships among Archaebacteria, Eubacteria, and Eukaryotes. Microbiol Mol Biol R 62:1435–1491

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res 24:4420–4449

Hansmann S, Martin W (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Micr 50:1655–1663

Huynen M, Snel B, Bork P (1999) Lateral Gene Transfer, Genome Surveys, and the Phylogeny of Prokaryotes. Science 286:1443a

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109–136

Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jinno K, Takahashi M, Sekine M, Baba S, Ankai A, et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 6:83–101, 145–152

Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, et al. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 5:55–76

Kawarabayasi Y, Hino Y, Horikawa H, Yamazaki S, Haikawa Y, Jinno K, Takahashi M, Sekine M, Baba S, Ankai A, et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res 6:83–101, 145–152

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature 390:364–370

Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessiaeres P, Bolotin A, Borchert S, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature 390:249–256

Montague MG, Hutchison CA (2000) Gene content phylogeny of herpesviruses. PNAS 97:5334–5339

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson LD, Nelson WC, Ketchum KA, et al. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima.* Nature 399:323–329

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, et al. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature 403:665–668

Pearson WR (1995) Comparison of methods for searching protein sequence databases. Protein Sci 4:1145–1160

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448

Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K, et al. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res 28:1397–1406

Smith DR, Doucette-Stamm LA, Deloughery C, Lee H, Dubois J, Aldredge T, Bashirzadeh R, Blakely D, Cook R, Gilbert K, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J Bacteriol 179:7135–7155

Snel B, Bork P, Huynen MA (1999) Genome phylogeny based on gene content. Nature Genet 21:108–110

Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L, Mitchell W, Olinger L, Tatusov RL, Zhao Q, et al. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis.* Science 282:754–759

Stetter KO (1996) Hyperthermophilic procaryotes. FEMS Microbiol Rev 18:149–158

Stetter KO (1988) *Archaeoglobus fulgidus* gen. nov., sp. nov.: a new taxon of extremely thermophilic archaebacteria. Syst Appl Microbiol 10:172–173

Tekaia F, Lazcano A, Dujon B (1999) The genomic tree as revealed from whole proteome comparisons. Genome Res 9:550–557

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Gill S, Dougherty BA, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature 388:539–547

White O, Eisen JA, Heidelberg JF, Hickey EK, Peterson JD, Dodson RJ, Haft DH, Gwinn ML, Nelson WC, Richardson DL, et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286:1571–1577

Wiley EO (1981) Phylogenetics: The Theory and Practice of Phylogenetic Systematics. Wiley, New York

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271