Universal trees based on large combined protein sequence data sets

James R. Brown, Christophe J. Douady, Michael J. Italia, William E. Marshall & Michael J. Stanhope

Universal trees of life based on small-subunit (SSU) ribosomal RNA (rRNA) support the separate mono/holophyly of the domains Archaea (archaebacteria), Bacteria (eubacteria) and Eucarya (eukaryotes) and the placement of extreme thermophiles at the base of the Bacteria[1–4]. The concept of universal tree reconstruction recently has been upset by protein trees that show intermixing of species from different domains[5,6]. Such tree topologies have been attributed to either extensive horizontal gene transfer[7] or degradation of phylogenetic signals because of saturation for amino acid substitutions[8]. Here we use large combined alignments of 23 orthologous proteins conserved across 45 species from all domains to construct highly robust universal trees. Although individual protein trees are variable in their support of domain integrity, trees based on combined protein data sets strongly support separate monophyletic domains. Within the Bacteria, we placed spirochaetes as the earliest derived bacterial group. However, elimination from the combined protein alignment of nine protein data sets, which were likely candidates for horizontal gene transfer, resulted in trees showing thermophiles as the earliest evolved bacterial lineage. Thus, combined protein universal trees are highly congruent with SSU rRNA trees in their strong support for the separate monophyly of domains as well as the early evolution of thermophilic Bacteria.

Data generated by genomic sequencing projects from a wide variety of species now allow for the assembly of concatenated or combined protein sequence data sets to reconstruct the universal tree of life. Phylogenies based on such data sets are potentially more robust and representative of the evolutionary relationships among species because the number of phylogenetically informative sites and sampled gene loci are greatly increased. The main principle behind combining data is that it allows for the amplification of phylogenetic signal, and increased resolving power, when the signal is masked by homoplasy (similarities in amino acids for reasons other than inheritance) among the individual gene data sets. Such protein data sets have helped resolve evolutionary relationships among photosynthetic bacteria[9] and eukaryotic protists[10].

Table 1 • Proteins included in concatenated alignments, number of residues and support for domain monophyly in individual protein trees

Cellular function	Protein	Amino acids[a]	Support for domain monophyly[b]		
			Archaea	Bacteria	Eucarya
Translation	alanyl-tRNA synthetase	502	100	–	100
	aspartyl-tRNA synthetase[c]	249	–	100	100
	glutamyl-tRNA synthetase[c]	188	50 (–)	100	100
	histidyl-tRNA synthetase	166	–	–	100(93)
	isoleucyl-tRNA synthetase	552	–	–	–
	leucyl-tRNA synthetase[c]	358	–	100	100
	methionyl-tRNA synthetase	306	–	–	99
	phenylalanyl-tRNA synthetase β subunit	177	–	–	100
	threonyl-tRNA synthetase	305	–	– (34)	100
	valyl-tRNA synthetase	538	–	–	100
	initiation factor 2[c]	337	–	100	100
	elongation factor G[c]	536	64(87)	100	100
	elongation factor Tu[c]	340	– (42)	100	100
	ribosomal protein L2[c]	192	46(–)	100	100
	ribosomal protein S5[c]	131	46(19)	100	100(99)
	ribosomal protein S8[c]	118	–	100	100
	ribosomal protein S11[c]	110	–	100	100
	aminopeptidase P	95	–	–	–
Transcription	DNA-directed RNA polymerase β chain[c]	537	99(78)	100	100
DNA replication	DNA topoisomerase I[c]	236	–	100	100
	DNA polymerase III subunit[c]	194	46(49)	100	100(95)
Metabolism	signal recognition particle protein[c]	298	71(39)	100	100
	rRNA dimethylase	126	–	–	100(98)
	full alignment length[d]	6,591			
	truncated alignment length[e]	3,824			

[a]Length of alignments after removing ambiguously aligned regions. [b]Percentage occurrence of monophyletic nodes in 100 bootstrap replicated data sets of protein distance/NJ and MP methods (in parentheses, where MP values differ from those of the NJ consensus tree). Dash indicates nodes that were not monophyletic. [c]Proteins included in both the full and truncated alignments. [d]Length of multiple sequence alignment, which includes all proteins, used to produce phylogeny in Fig. 1. [e]Length of multiple sequence alignment, which excludes proteins in which Bacteria were not monophyletic, used to produce phylogeny in Fig. 2.

Anti-Microbial Bioinformatics Group, GlaxoSmithKline,1250 South Collegeville Road, UP1345 P.O. Box 5089, Collegeville, Pennsylvania 19426-0989, USA. Correspondence should be addressed to J.R.B. (e-mail: James_R_Brown@gsk.com).

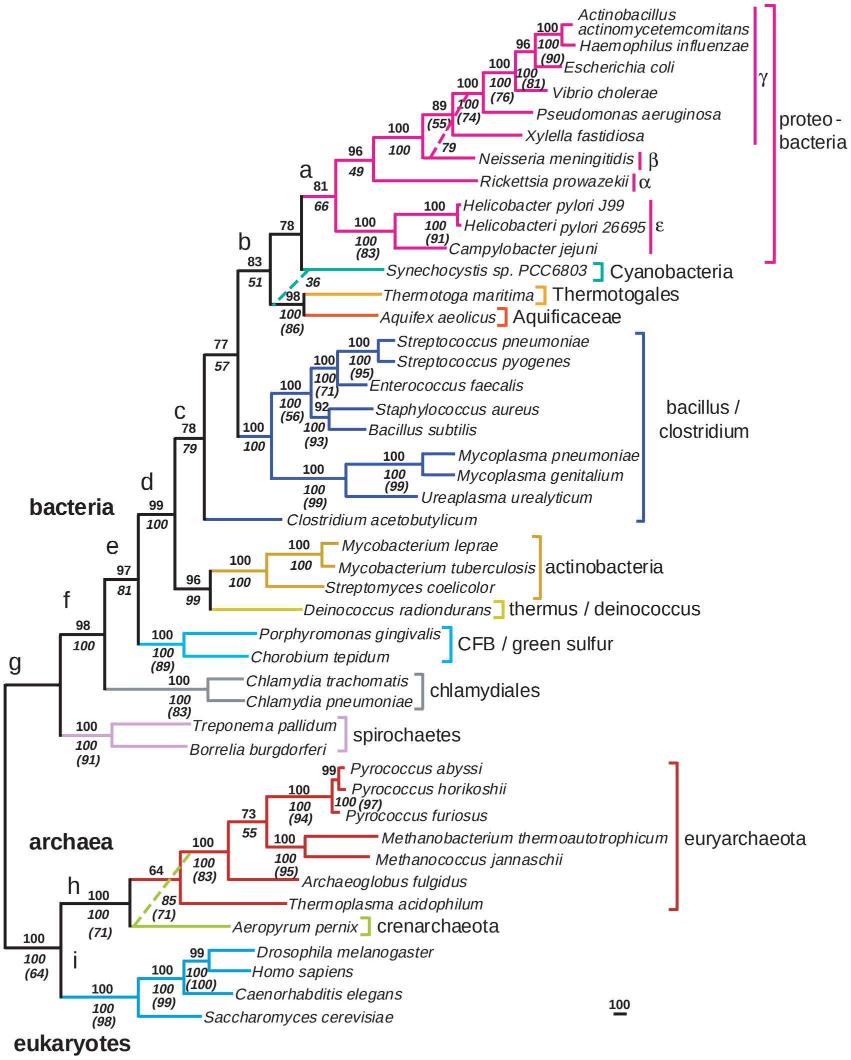


Fig. 1 Minimal length MP universal trees based on 23 combined protein data sets. Spirochaetes are placed as the lowest branching Bacteria. Numbers along the branches show percentage occurrence of nodes in 1,000 bootstrap replicates of MP (ref. 29) (plain text) and NJ (ref. 28) (italicized text) analyses or greater than 50% of 1,000 QP steps of maximum-likelihood[30] analysis (parentheses). Dashed lines indicate occasional differences in branching orders in NJ trees. Trees constructed by the ME method[28] are not shown but had topologies identical to those of NJ trees. Letters indicate nodes tested in the VLB simulations (see Fig. 3). Scale bar represents 100 amino acid residue substitutions.

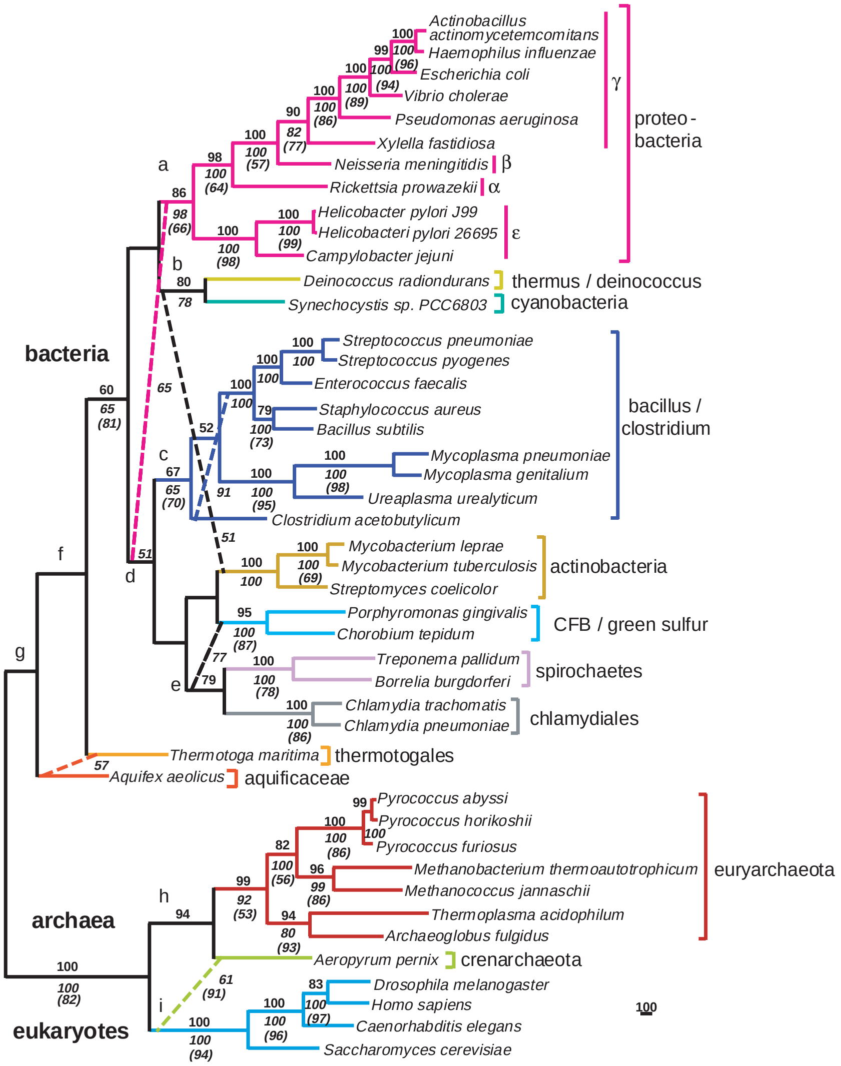
By comparing the open reading frames from finished or nearly complete genomic sequences of 45 species of Bacteria, Archaea and Eucarya, we identified 23 proteins conserved across all species and, therefore, highly suited for constructing universal trees (Table 1). We used database homology searches and individual phylogenetic trees to confirm that all proteins were orthologous (proteins were not duplicated in any organism and direct ancestral descent throughout all species could be inferred). According to the endosymbiosis hypothesis, eukaryotic mitochondria-targeted proteins encoded in the nuclear genome can be bacteria-like because they were secondarily transferred from the bacterial progenitor of the mitochondria[11]. Therefore, when eukaryotes had both mitochondria- and cytoplasm-targeted versions of the same protein, we used only the latter because cytoplasmic versions probably best represent the evolution of the eukaryotic nuclear genome. Individual protein families were first computer aligned and then we manually refined the alignments. We removed poorly conserved regions in individual protein alignments (characterized by gaps of variable length) before concatenation, which resulted in a final data set of 6,591 aligned amino acids from 45 species. To our knowledge, this is the largest protein data set used for universal tree reconstruction.

Phylogenetic trees constructed by maximum-likelihood quartet puzzling (QP), maximum parsimony (MP), minimum evolution (ME) and neighbor-joining (NJ) methods have highly congruent topologies. Similar to universal rRNA trees, all combined protein data set phylogenetic trees strongly support the monophyly of the three domains (Fig. 1). On average, archaeal and eukaryotic species are slightly more similar to each other than either is to Bacteria. However, we cannot confirm that Archaea and Eucarya share a more recent common ancestor because the tree is unrooted. Within each domain, the branching order of most nodes is well supported by bootstrap replications (>70%). Although only a few genomes of Archaea and eukaryotes have been completely sequenced, branching orders of those species are consistent with contemporary views of organism evolution. In eukaryotes, Pseudocoelomata (*Caenorhabditis elegans*) and Coelomata (*Homo sapiens* and *Drosophila melanogaster*) cluster together with fungi (*Saccharomyces cerevisiae*) as an outgroup. In the Archaea, the deepest branch point separates the kingdoms Euryarchaeota and Crenarchaeota, although the latter kingdom is represented by only a single species, *Aeropyrum pernix*.

In the Bacteria, the major subdivisions of *Bacillus*/*Clostridium* (low G+C Gram positives), spirochaetes and proteobacteria are strongly supported as being monophyletic, as postulated by the universal rRNA trees[3]. However, a major departure is the placement of spirochaetes (*Treponema pallidum* and *Borrelia burgdorferi*) instead of thermophiles (*Aquifex aeolicus* and *Thermotoga maritima*) as the first bacterial branch. Although the basal position of spirochaetes is incompatible with hypotheses about the thermophilic origins of life for prokaryotes, it is consistent with hypotheses of extensive horizontal gene transfer between spirochaetes and Archaea. For example, Archaea and spirochaetes share novel class I type lysyl-tRNA synthetase to the exclusion of eukaryotes and most other bacteria[12].

Our examination of the individual gene trees reveals other instances in which the domains are not monophyletic, implicating possible horizontal gene transfer (Table 1). For example, phenylalanyl-tRNA synthetase β-subunit has been previously indicated to be a specific transfer from the Archaea to spirochaetes[13]. Histidyl-tRNA synthetase trees show the spirochaetes as well as *Porphyromonas gingivalis*, *Helicobacter pylori*, *Xylella fastidiosa*, and *Clostridium acetobutylicum* clustering with eukaryotes[14], and the isoleucyl-tRNA synthetases of 14

Fig. 2 Minimal length MP universal tree based on 14 combined protein data sets, with 9 horizontal gene transfer proteins removed. Tree indicates thermophiles as the basal group in Bacteria. Methods and labeling are the same as Fig. 1.



bacterial species from different groups (spirochaetes, Chlamydiales, Actinobacteria, *Thermus/Deinococcus* and the Cytophaga-Flexibacter-Bacterioides or CFB group) seem to have been exchanged with eukaryotes[15]. Valyl-tRNA synthetase was exchanged between *Richettsia prowazekii* and the Archaea[16]. Interestingly, none of the 23 individual protein trees indicates that the hyperthermophilic bacteria *T. maritima* and *A. aeolicus* exchange genes with either eukaryotes or the Archaea.

To construct a universal tree based on genes directly descended from the last common ancestor, we removed proteins that were possibly affected by horizontal gene transfers between the domains from the combined protein alignment. Among the protein trees examined, most of the horizontal gene transfers probably occurred between Bacteria and eukaryotes or between Bacteria and Archaea. Although horizontal gene transfers between eukaryotes and Archaea cannot be eliminated, such instances would not be detected in the present collection of completed eukaryotic genomes because only crown eukaryotes are represented[17]. Therefore, we judged proteins to be candidates for horizontal gene transfer if individual protein trees did not depict the Bacteria as a monophyletic group. We removed nine horizontal transferred proteins from the combined protein data set, which resulted in a truncated alignment of 3,824 amino acids (Table 1).

In contrast to the combined alignment of 23 proteins, phylogenetic trees based on the alignment of 14 proteins agree with universal rRNA trees in the placement of the hyperthermophilic species *A. aeolicus* and *T. maritima* as the lowest branching bacterial lineages, whereas spirochaetes are the more derived group (Fig. 2). All phylogenetic methods applied to this data set support the early evolution of hyperthermophiles, with bootstrap values for the node separating hyperthermophiles from the rest of the Bacteria ranging from 60% to 81%. The agreement between the data set that excludes horizontal transferred genes (truncated protein tree) and the rRNA tree in the placement of extreme thermophiles as the basal lineage in the Bacteria lends further support to the theory that life evolved at high temperatures[2–4]. However, many questions about the origin of life at high temperatures, such as the overall viability and stability of extracellular biochemical reactions[18] and RNA molecules[19], remain unresolved.

Although truncated protein data set phylogenies constructed by the MP method depict three monophyletic domains, NJ, ME, and QP methods show the Archaea as a paraphyletic group with *A. pernix*, a Crenarchaeota, branching at the base of eukaryotes. This branching pattern is probably an artifact of phylogenetic reconstruction known as long branch attraction and not evidence for the eocyte hypothesis, which suggests that eukaryotes evolved from the Crenarchaeota[20]. As genome sequences from other species of Crenarchaeota and early branching eukaryotes become available, phylogenetic analyses of combined protein data sets should help resolve the evolutionary relationships between early eukaryotes and the Archaea.

To determine the impact of potential saturation for multiple reversible amino acid substitutions on the topology derived from the truncated protein data set, we used a weighted parsimony method in which all possible amino acid replacement pairs are handicapped by their tendency for superimposed changes. We derived the amino acid substitution weights from consistency index values (a measure of the degree of homoplasy) obtained from individual trees based on alignments in which only a single pair of amino acid substitutions could occur, and all other states were coded as missing data[21]. We empirically estimated weights based on the degree of saturation for each of the 190 possible pairwise amino acid substitutions. The resulting weighted parsimony tree has the same topology as the respective unweighted MP tree, although bootstrap supporting values for all nodes are lower. Thus, the truncated protein data set tree topology is robust even after a rigorous correction for amino acid saturation.

Variable length bootstrap (VLB) analysis[22] suggests that the size of the amino acid alignment is an important factor in improving bootstrap support and reliability for nodes in trees constructed from the full and the truncated protein alignment

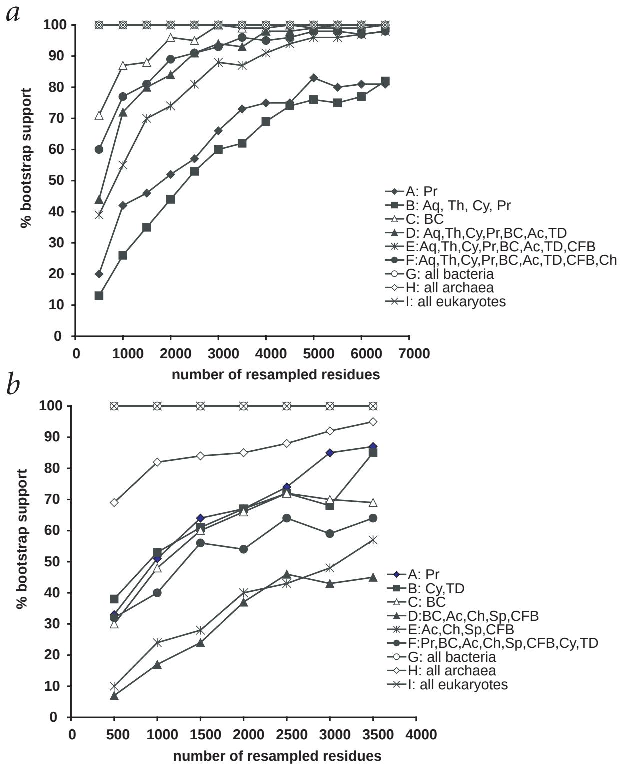


Fig. 3 VLB analysis[22] showing support for major nodes in the MP universal trees. Letters correspond to nodes for trees in Figs. 1 and 2. Actinobacteria (Ac), Aquificaceae (Aq), *Bacillus/Clostridium* (BC), CFB/green sulfur (CFB), chlamydia (Ch), cyanobacteria (Cy), proteobacteria (Pr), spirochaetes (Sp), Thermotogales (Th) and *Thermus/Deinococcus* (TD). ***a***, VLB for tree based on 23 combined proteins (Fig. 1). ***b***, VLB for tree based on 14 combined proteins (Fig. 2).

(Fig. 3*a* and 3*b*, respectively). Bootstrap support values for major nodes in the full tree exceed 70% only if more than about 3,000 residues are used in the phylogenetic reconstruction, which is 6 times longer than any single protein alignment (Table 1). Nodes weakly supported by bootstrapping show increasing levels of support improvement as alignments grow longer. Furthermore, the positive trend lines indicate that further sequence data could improve bootstrap values for all nodes. Although it may be difficult to find more universally conserved genes, VLB analysis indicates that further sequence data will help resolve branch points within the Bacteria. Expanded data sets of broadly found proteins in the Bacteria, which are likely to be more numerous than universally conserved proteins, could be used to provide more statistically robust trees of bacterial groups. Combined with the results of this study, which supports a rooting in either Thermotogales or Aquificaceae, a more comprehensive picture of bacterial evolution could emerge through the use of bacteria-specific combined protein data sets.

Coevolution could also cause the concordance between the truncated protein and rRNA trees because many proteins used in this phylogenetic analysis are involved in protein synthesis, the same general pathway as SSU rRNA. However, subtle but important differences between the protein trees here and rRNA trees indicate independent evolution. High G+C and low G+C Gram-positives are not collectively monophyletic as previously reported for rRNA and other molecular markers[23]. The clustering of Chlamydiales, CFB and spirochaetes together is novel relative to rRNA trees[3].

Both full and truncated protein trees have important implications with respect to recent discussions about the role of horizontal gene transfer in genome evolution[7]. Early radiation of thermophiles in the diversification of the Bacteria, indicated by our analysis, may have contributed to claims of extensive horizontal gene transfer between Bacteria, particularly *T. maritima*, and Archaea[24]. Genes found only in thermophilic Bacteria and Archaea are just as likely to be shared symplesiomorphies, which were later lost in other bacterial species. More widely conserved genes from deep branching bacterial species might find best matches to archaeal genes in general homology searches of databases, although careful phylogenetic analyses would show otherwise[25].

Truncated protein trees, in particular those based on the NJ and ME methods, show a fundamental division in the Bacteria where, after diverging from hyperthermophiles, Proteobacteria split from all other bacteria. Furthermore, within the Proteobacteria, the earliest diverged group is the α subdivision, represented by *R. prowazekii*, from which the endosymbiont progenitor of the mitochondria probably evolved[16]. The early emergence of α-Proteobacteria suggests that endosymbiotic relationships between eukaryotes and bacteria could have occurred early in cellular evolution, perhaps shortly after the divergence of the domains Bacteria, Archaea and eukaryotes. As bacterial species were evolving, they could have shared genes with early eukaryotes either directly or through secondary transfers with free-living relatives of endosymbionts. The net result would be the seemingly extensive exchange of genes between eukaryotes and many diverse and currently distantly related groups of bacteria.

The proposed divergence between Proteobacteria and Gram-positive bacteria has important implications for the discovery and development of antibiotics. Key targeted organisms that can coinfect humans occupy very distal branches in the bacterial clade (the Proteobacterium *Haemophilus influenzae* versus the Gram-positive coccal species *Staphylococcus aureus* and *Streptococcus pneumoniae*). The wide evolutionary distances between pathogenic species need to be carefully considered when broad-spectrum anti-microbial targets are being selected.

Phylogenetic analysis of combined protein data sets represents an important approach in the use of genome sequence data to address evolutionary questions. Although horizontal gene transfer has probably played a critical role in building genomic diversity, we have shown that genomes have retained sufficient phylogenetic signal for reconstruction of robust universal trees, which can provide important insights into the pattern of early cellular evolution.

Methods

Identification of orthologous proteins. We collected sequences from complete or nearly complete public genomes. We identified orthologous proteins with a relational bacterial genomic database, which contains an array of protein-by-protein sequence similarity scores (smallest sum probabilities) across multiple genomes as calculated by BLASTP v2.0 (ref. 26). Using the smallest complete genome, that of *Mycoplasma genitalium*, as the driver query, we built individual protein data sets on the basis of significant homology to the driver protein (E-value ≤0.00001) and the occurrence of the protein in all sampled genomes. We collected 65 protein families and constructed an unambiguous orthologous protein data set by reducing the number of protein families until there was only a single occurrence of any particular protein in the data set. Then we visually evaluated multiple sequence alignments and phylogenies to confirm orthologous relationships. Note that the final number of 23 orthologous protein families cannot be considered to be a definitive list of all evolutionarily conserved proteins. The criterion for inclusion of a particular protein in this analysis is the absolute presence of a single orthologous copy in all 45 genomes and, because a few genomes were not completely sequenced, it is possible that there are additional universally conserved genes.

Phylogenetic analyses. We used the program CLUSTALW v1.7 (ref. 27) with default settings to align individual protein data sets. Then we manually refined multiple sequence alignments with the program SEQLAB of the GCG v10.0 software package (Genetics Computer Group). From the alignments, we removed regions with residues that could not be unambiguously aligned or that contained insertions or deletions. Customized computer scripts concatenated aligned protein data sets before phylogenetic analyses.

We used NJ, MP, QP and ME methods to construct phylogenetic trees. NJ trees were based on pairwise distances between amino acid sequences using the programs NEIGHBOR and PROTDIST (Dayhoff option) of the PHYLIP 3.6 package[28]. We used the programs SEQBOOT and CONSENSE to estimate the confidence limits of branching points from 1,000 bootstrap replications. We used the program FITCH with negative branch lengths, global rearrangements and five randomizations of species input order to construct ME trees. We used the software package PAUP4.0b5 (ref. 29) for MP as well as VLB analyses. The numbers and lengths of minimal trees were estimated from 100 random sequence additions, and confidence limits of branch points were estimated by 1,000 bootstrap replications.

We constructed ML tree topologies with the software PUZZLE 4.0 (ref. 30), using 1,000 puzzling steps, the JTT substitution matrix, estimation of rate heterogeneity using the gamma distribution model with 8 rate categories, and the α-parameter estimation from the data set.

For the full protein alignment, a single minimal length MP tree was recovered, 66,717 steps with a consistency index (CI) of 0.4731 and a retention index (RI) of 0.5902. For the truncated protein alignment, there were four minimal length MP trees with overall similar topologies except for rearrangements of *C. acetobutylicum* within the *Bacillus*/*Clostridium* clade and different branching orders of Actinobacteria and green sulfur/CFB with respect to chlamydia and spirochaetes. Truncated protein MP trees were 34,439 steps long with CI and RI values of 0.4924 and 0.6342, respectively. Multiple sequence alignments and phylogenetic trees are available upon request.

Acknowledgments

We thank the GlaxoSmithKline Bioformatics Division for their support of this study.

Received 27 March; accepted 18 May 2001.

1. Fox, G.E., Magrum, L.J., Balch, W.E., Wolfe, R.S. & Woese, C.R. Classification of methanogenic Bacteria by 16S ribosomal RNA characterization. *Proc. Natl. Acad. Sci. USA* **74**, 4537–4541 (1977).
2. Woese, C.R. Bacterial evolution. *Microbiol. Rev.* **51**, 221–271 (1987).
3. Olsen, G.J., Woese, C.R. & Overbeek, R. The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.* **176**, 1–6 (1994).
4. Pace, N.R. Origin of life—facing up to the physical setting. *Cell* **65**, 531–533 (1991).
5. Golding, G.B. & Gupta, R.S. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol. Biol. Evol.* **12**, 1–6 (1995).
6. Brown, J.R. & Doolittle, W.F. Archaea and the prokaryote-eukaryote transition. *Microbiol. Mol. Biol. Rev.* **61**, 456–502 (1997).
7. Doolittle, W.F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2128 (1999).
8. Forterre, P. & Philippe, H. Where is the root of the universal tree of life? *BioEssays* **21**, 871–879 (1999).
9. Xiong, J., Inoue, K. & Bauer, C.E. Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*. *Proc. Natl. Acad. Sci. USA* **95**, 14851–14856 (1998).
10. Baldauf, S.L., Roger, A.J., Wenk-Siefert, I. & Doolittle, W.F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290**, 972–977 (2000).
11. Gray, M.W. The endosymbiont hypothesis revisited. *Int. Rev. Cytol.* **141**, 233–357 (1992).
12. Ibba, M., Morgan, S., Curnow, A.W., Pridmore, D.R., Vothknecht, U.C., Gardner, W., Lin, W., Woese, C.R. & Söll, D. A euryarchaeal lysyl-tRNA synthetase: resemblance to class I synthetases. *Science* **278**, 1119–1122 (1997).
13. Teichmann, S.A. & Mitchison, G. Is there a phylogenetic signal in prokaryote proteins? *J. Mol. Evol.* **49**, 98–107 (1999).
14. Bond, J.P. & Francklyn, C. Proteobacterial histidine-biosynthetic pathways are paraphyletic. *J. Mol. Evol.* **50**, 339–347 (2000).
15. Brown, J.R., Zhang, J. & Hodgson, J.E. A bacterial antibiotic resistance gene with eukaryotic origins. *Curr. Biol.* **8**, R365–R367 (1998).
16. Kurland, C.G. & Andersson, S.G.E. Origin and evolution of the mitochondrial proteome. *Microbiol. Mol. Biol. Rev.* **64**, 786–820 (2000).
17. Chihade, J.W., Brown, J.R., Schimmel, P. & Ribas de Pouplana, L. Origin of mitochondria in relation to evolutionary history of eukaryotic alanyl-tRNA synthetase. *Proc. Natl. Acad. Sci. USA* **97**, 12153–12157 (2000).
18. Miller, S.L. & Lazcano, A. The origin of life—Did it occur at high temperatures? *J. Mol. Evol.* **41**, 689–692 (1995).
19. Galtier, N., Tourasse, N. & Gouy, M. A nonhyperthermophilic common ancestor to extant life forms. *Science* **283**, 220–221 (1999).
20. Rivera, M.C. & Lake, J.A. Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**, 74–76 (1992).
21. Hassanin, A. Lecointre, G. & Tillier, S. The evolutionary signal of homoplasy in protein-coding gene sequences and its consequences for a priori weighting in phylogeny. *C.R. Acad. Sci.* **321**, 611–620 (1998).
22. Springer, M.S., Amrine, H.M., Burk, A. & Stanhope, M.J. Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. *Syst. Biol.* **48**, 65–75 (1999).
23. Shah, H.N., Gharbia, S.E. & Collins, M.D. The Gram stain: a declining synapomorphy in an emerging evolutionary tree. *Rev. Med. Microbiol.* **8**, 103–100 (1997).
24. Nelson, K.E. *et al.* Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**, 323–329 (1999).
25. Logsdon, J.M., Jr. & Faguy, D.M. Evolutionary genomics: *Thermotoga* heats up lateral gene transfer. *Curr. Biol.* **9**, R747–R751 (1999).
26. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
27. Thompson, J.D., Higgins, D.G. & Gibson, T. J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalities and weight matrix choice., *Nucleic Acids Res.* **22**, 4673–4680 (1994).
28 Felsenstein, J. *PHYLIP* (*Phylongeny Inference Package*), version 3.6. (Department of Genetics, University of Washington, Seattle, 2000.)
29. Swofford, D.L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. (Sinauer Associates, Sunderland, MA, 1999).
30. Strimmer, K. & von Haeseler, A. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969 (1996).