Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 9355–9359, December 1989
Evolution

Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes

NAOYUKI IWABE*, KEI-ICHI KUMA*, MASAMI HASEGAWA†, SYOZO OSAWA‡, AND TAKASHI MIYATA*

*Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, Japan; †Institute of Statistical Mathematics, Minatoku, Tokyo 106, Japan; and ‡Department of Biology, Faculty of Science, Nagoya University, Nagoya 464-01, Japan

Communicated by Motoo Kimura, August 22, 1989

ABSTRACT **All extant organisms are thought to be classified into three primary kingdoms, eubacteria, eukaryotes, and archaebacteria. The molecular evolutionary studies on the origin and evolution of archaebacteria to date have been carried out by inferring a molecular phylogenetic tree of the primary kingdoms based on comparison of a single molecule from a variety of extant species. From such comparison, it was not possible to derive the exact evolutionary relationship among the primary kingdoms, because the root of the tree could not be determined uniquely. To overcome this difficulty, we compared a pair of duplicated genes, elongation factors Tu and G, and the α and β subunits of ATPase, which are thought to have diverged by gene duplication before divergence of the primary kingdoms. Using each protein pair, we inferred a composite phylogenetic tree with two clusters corresponding to different proteins, from which the evolutionary relationship of the primary kingdoms is determined uniquely. The inferred composite trees reveal that archaebacteria are more closely related to eukaryotes than to eubacteria for all the cases. By bootstrap resamplings, this relationship is reproduced with probabilities of 0.96, 0.79, 1.0, and 1.0 for elongation factors Tu and G and for ATPase subunits α and β, respectively. There are also several lines of evidence for the close sequence similarity between archaebacteria and eukaryotes. Thus we propose that this tree topology represents the general evolutionary relationship among the three primary kingdoms.**

Based on comparison of the small rRNAs, Woese and colleagues (1–3) proposed that there are two fundamentally different groups of bacteria, eubacteria and archaebacteria, and that, with eukaryotes, they constitute the three primary kingdoms of life. Although the existence of the archaebacterial urkingdom is accepted by many biologists, the classification is still a matter of controversy: Lake and colleagues (4, 5) argued that archaebacteria are paraphyletic; sulfobacteria (eocytes) are more closely related to eukaryotes than to other archaebacteria, whereas halobacteria are more closely related to eubacteria than to other archaebacteria. Phylogenetic trees based on the small and large rRNAs (2, 3, 6), 5S rRNA (7), and the RNA polymerases (8), however, support the monophyletic view of the archaebacteria originally proposed by Woese and colleagues (1–3).

The evolutionary relationship of the three primary kingdoms is another crucial problem that remains unanswered. There are several reports that, in some RNA and protein species, archaebacteria are much more similar in sequence to eukaryotes than to eubacteria. These include 5S rRNA (7, 9, 10), elongation factors Tu (11) and G (12) (EF-Tu and EF-G), large subunit of DNA-dependent RNA polymerase (8, 13), and several ribosomal proteins (14, 15). The α and β subunits of *Sulfolobus* ATPase (16, 17) also bear closer resemblance in sequence to eukaryotic counterparts than to eubacterial ones. However, a close similarity in sequence does not necessarily mean a close relatedness in phylogeny, unless similar rates of evolution for different lineages can be postulated. There is no *a priori* reason to believe that the three primary kingdoms have evolved with similar rates.

A phylogenetic tree inferred from a comparison of a single RNA or protein species from a variety of organisms of the primary kingdoms is generally unrooted, because one cannot determine the root, the universal ancestor from which all extant life ultimately diverged. This difficulty, however, could be overcome by inferring a composite phylogenetic tree from a comparison of a pair of duplicated genes that exist in all extant life. It is reasonable to consider that this gene duplication occurred prior to the divergence of the primary kingdoms and thus the root of the inferred composite tree could be unambiguously set at a point where the two genes diverged by gene duplication. The universal ancestor and the evolutionary relationship of the three primary kingdoms are subsequently determined from the composite phylogenetic tree. Pairs of genes for EF-Tu and EF-G and for the α and β subunits of F_1-ATPase, F_1-α and F_1-β, are examples of such duplicated genes. Each pair shows extensive sequence similarity for species of the three primary kingdoms.

On the basis of the composite trees for EF-Tu and EF-G and for F_1-α and F_1-β, we report that archaebacteria are phylogenetically more closely related to eukaryotes than to eubacteria. Judging from the strong sequence similarities between archaebacteria and eukaryotes found in 5S rRNA, RNA polymerases, and ribosomal proteins, the phylogenetic relationship presented here would represent the general evolutionary relationship among the three primary kingdoms.

METHODS

Phylogenetic Tree. Alignment of amino acid sequences was carried out as described (18). The number of amino acid substitutions per site or evolutionary distance between sequences of extant species was measured by calculating the proportion of amino acid difference, K, between the sequences compared and by correcting K for multiple substitutions by using $k = -\ln(1 - K)$ (19); positions where gaps are present in any one of the aligned sequences were excluded from the analysis. Based on the evolutionary distance matrix, a phylogenetic tree was inferred by the neighbor-joining method (20).

Reliability of the Tree. To obtain the reliability of the inferred phylogenetic tree, the bootstrap method (21) was applied. The bootstrap resamplings were repeated 1000 times, and for each of the resamplings a tree was inferred by the neighbor-joining method. The bootstrap probability that a particular tree topology occurs during the resamplings was evaluated.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: EF-Tu and EF-G, elongation factors Tu and G, respectively; LDH, lactate dehydrogenase; MDH, malate dehydrogenase.

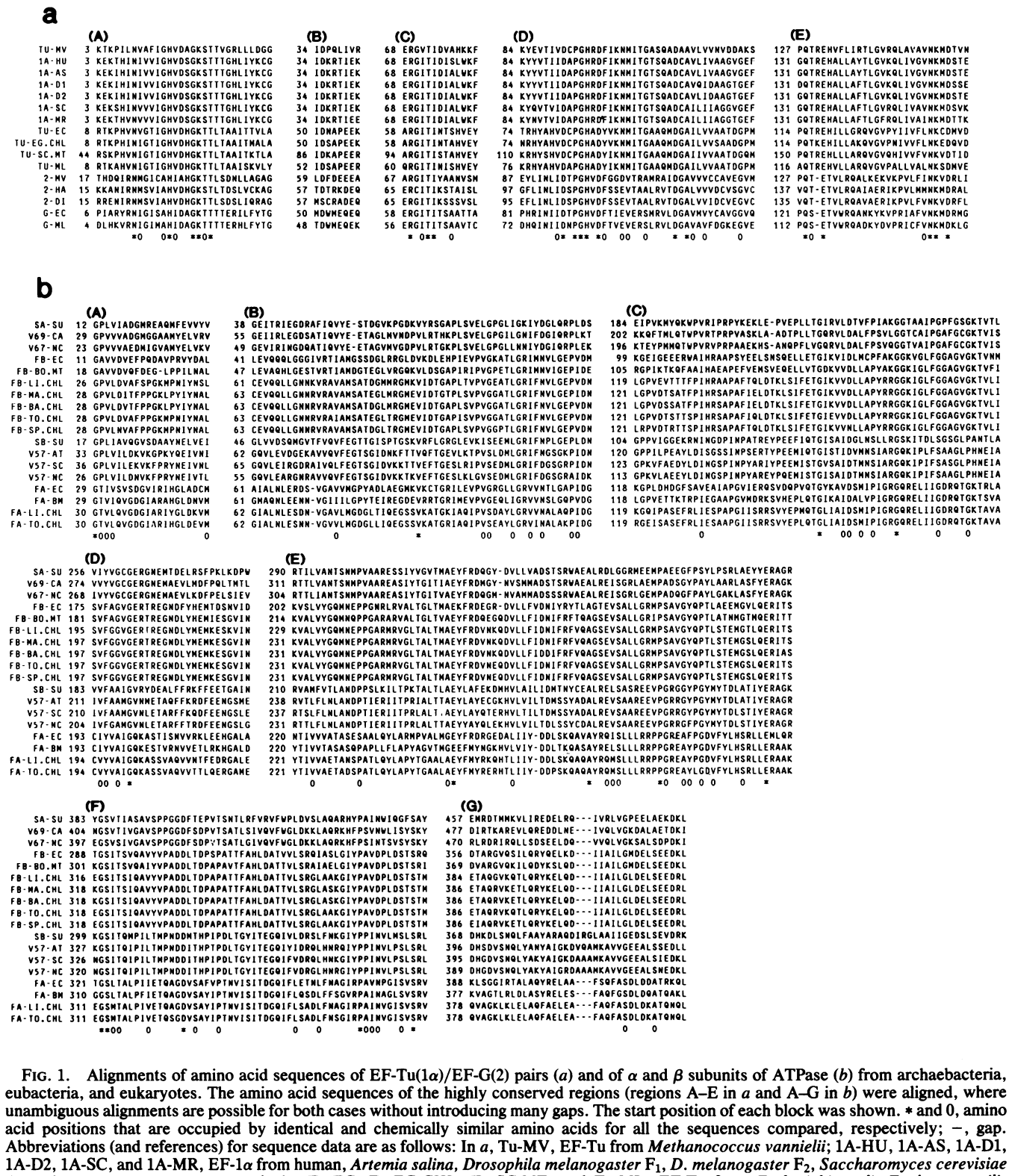


FIG. 1. Alignments of amino acid sequences of EF-Tu(1α)/EF-G(2) pairs (*a*) and of of α and β subunits of ATPase (*b*) from archaebacteria, eubacteria, and eukaryotes. The amino acid sequences of the highly conserved regions (regions A–E in *a* and A–G in *b*) were aligned, where unambiguous alignments are possible for both cases without introducing many gaps. The start position of each block was shown. * and 0, amino acid positions that are occupied by identical and chemically similar amino acids for all the sequences compared, respectively; −, gap. Abbreviations (and references) for sequence data are as follows: In *a*, Tu-MV, EF-Tu from *Methanococcus vannielii*; 1A-HU, 1A-AS, 1A-D1, 1A-D2, 1A-SC, and 1A-MR, EF-1α from human, *Artemia salina, Drosophila melanogaster* F_1, *D. melanogaster* F_2, *Saccharomyces cerevisiae* A, and *Mucor racemosus*, respectively; Tu-EC, Tu-EG.CHL, Tu-SC.MT, and Tu-ML, EF-Tu from *Escherichia coli, Euglena gracilis* chloroplast DNA, *S. cerevisiae* nuclear DNA-coded mitochondrial isozyme, and *Micrococcus luteus*, respectively; 2-MV, 2-HA, and 2-DI, EF-2 from *M. vannielii* (12), hamster, and *Dictyostelium discoideum* (22), respectively; G-EC and G-ML, EF-G from *E. coli* and *Micrococcus luteus*, respectively. In *b*, SA-SU, *Sulfolobus acidocaldarius* α subunit of ATPase; V69-CA and V67-NC, vacuolar large subunits of ATPase from carrot and *Neurospora crassa*, respectively; FB-EC, FB-BO.MT, FB-LI.CHL, FB-MA.CHL, FB-BA.CHL, FB-TO.CHL, and FB-SP.CHL, β subunits of F_1 ATPase from *E. coli*, bovine nuclear DNA-coded mitochondrial isozyme, liverwort chloroplast DNA, maize chloroplast DNA, barley chloroplast DNA, tobacco chloroplast DNA, and spinach chloroplast DNA, respectively; SB-SU: *S. acidocaldarius* β subunit of ATPase (17); V57-AT, V57-SC, and V57-NC, small subunits of vacuolar ATPase from *Arabidopsis thaliana, S. cerevisiae* (23), and *N. crassa*, respectively; FA-EC, FA-BM, FA-LI.CHL, and FA-TO.CHL, α subunits of F1-ATPase from *E. coli, Bacillus megaterium* (24), liverwort chloroplast DNA, and tobacco chloroplast DNA, respectively. Sequence data not referenced were taken from the National Biomedical Research Foundation data base (release 20.0) and GenBank data base (release 59.0); sequence data were checked with their original data. The single-letter amino acid code is used.

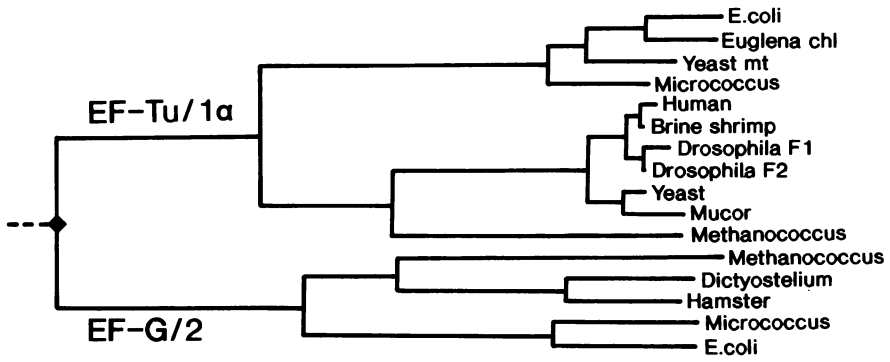


FIG. 2. Composite phylogenetic tree inferred from a simultaneous comparison of EF-Tu(1α) and EF-G(2) from archaebacterium, eubacteria, and eukaryotes. Based on the alignment of Fig. 1*a*, this tree was constructed. The deepest root was arbitrarily chosen at a point between the two clusters corresponding to the different proteins. chl, Chloroplast DNA-coded gene product; mt, nuclear gene-coded mitochondrial isozyme.

RESULTS

Composite Phylogenetic Tree of EF-Tu and EF-G. The amino acid sequences of EF-Tu and EF-G from an archaebacterium were aligned with those of eubacterial homologs and eukaryotic homologs (EF-1α and EF-2) for highly conserved regions (regions A–E), covering a total of 120 amino acids, where an unambiguous alignment is possible without introducing many gaps (Fig. 1*a*). On the basis of the alignment, the distance matrix was calculated and the composite phylogenetic tree of the EF-Tu(1α)/EF-G(2) pair was inferred by the neighbor-joining method. It is evident that EF-Tu(1α) and EF-G(2) diverged by gene duplication; they have extensive sequence similarity and these genes are present in close proximity to each other on archaebacterial and eubacterial genomes (12, 25–27). It is reasonable to assume that this gene duplication occurred before divergence among the primary kingdoms, because the two genes exist in all the urkingdoms. This is consistent with the inferred composite tree, in which there exist two major clusters corresponding to the respective two genes. Thus the deepest root of the composite tree could unambiguously be placed at a point between the two major clusters.

Fig. 2 shows the composite tree of the EF-Tu and EF-G pair. This tree is composed of two universal trees for the different proteins *Methanococcus vannielii*, an archaebacterium, is more closely related to eukaryotes than to eubacteria for both the proteins. This result indicates that close sequence similarities between *M. vannielii* and eukaryotes for EF-Tu (11) and EF-G (12) represent their close evolutionary relatedness.

Because distantly related sequences were compared, only limited amino acid positions are able to align. Thus it remains possible that the inferred tree of Fig. 2 is an artifact that was realized by chance. To disclose this, the bootstrap resamplings were carried out. The evolutionary relationship among the primary kingdoms shown in Fig. 2 is statistically significant; the bootstrap probabilities of occurrence of the tree topology that archaebacteria and eukaryotes share a cluster with eubacteria as an outgroup are 0.96 for EF-Tu and 0.79 for EF-G. The probabilities of alternative tree topologies are very small (see Fig. 4).

Composite Phylogenetic Tree of the α and β Subunits of ATPase. A similar analysis was performed for α and β subunits of ATPase, which apparently diverged by gene duplication before separation of the primary kingdoms. Fig. 1*b* shows alignment of amino acid sequences of the α and β subunits of ATPase from *Sulfolobus*, an archaebacterium, with those of the α and β subunits of F_1-ATPase from eubacterial origins and their eukaryotic counterparts, the large and small subunits of vacuolar ATPase. These proteins are strongly conserved in sequence and the alignment of Fig. 1*b* covers almost the entire regions. Fig. 3 shows the composite phylogenetic tree inferred from the alignment of Fig. 3. It is evident that, for both the subunits, *Sulfolobus* is shown to be more closely related to eukaryotes than to eubacteria. The bootstrap resamplings did not realize any alternative tree topology (see Fig. 4).

The initiator and elongator methionine tRNAs probably diverged by gene duplication before divergence of the three primary kingdoms, because both RNAs exist in all the primary kingdoms. A similar analysis reproduced essentially the same tree topology as those of the above two cases (data

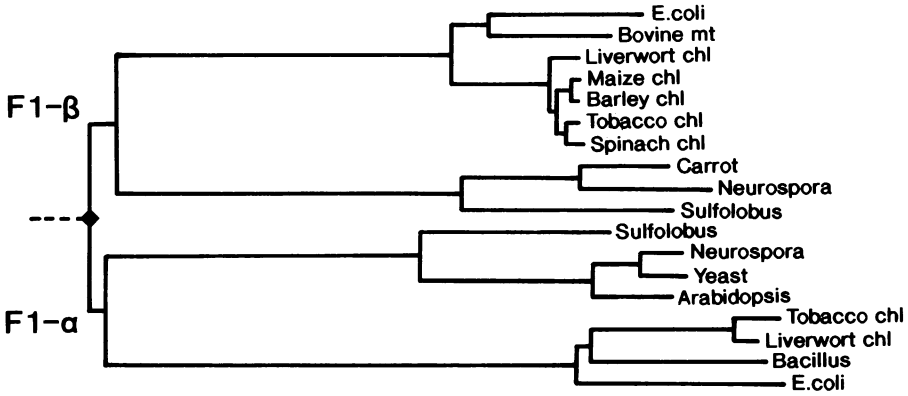


FIG. 3. Composite phylogenetic tree inferred from a simultaneous comparison of α and β subunits of F_1-ATPase from eubacteria and their archaebacterial and eukaryotic homologs. This tree was constructed based on the alignment of Fig. 1*b*. The deepest root was arbitrarily chosen at a point between the two clusters corresponding to the different proteins. chl, Chloroplast DNA-coded gene product; mt, nuclear gene-coded mitochondrial isozyme.

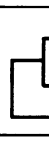
GENE	TREE 1	TREE 2	TREE 3
			
EF-Tu	0.96	0.03	0.01
EF-G	0.79	0.21	0
ATPase F1-β	1.0	0	0
ATPase F1-α	1.0	0	0
tRNA Met-E	0.55	0.33	0.12
tRNA Met-I	0.50	0.41	0.09

FIG. 4. Bootstrap probabilities of the three possible tree topologies among the three kingdoms. A, archaebacteria; EK, eukaryotes; EB, eubacteria; O, outgroups; tRNAs Met-E and Met-I, elongation and initiation Met tRNAs, respectively.

not shown), although the tree topology is statistically not significant in both the tRNAs, probably because of the limited length of the sequences.

The above results were summarized in Fig. 4. From these results we conclude that, according to the molecular phylogenies based on the two elongation factors and two subunits of ATPase, archaebacteria are phylogenetically more closely related to eukaryotes than to eubacteria. Interestingly, relatedness in phylogeny positively correlates with similarity in sequence in the four proteins analyzed. Furthermore, in the phylogenetic trees of Figs. 2 and 3, the average branch lengths from the universal ancestor to the extant species appear not to differ significantly between different urkingdoms, suggesting, on average, similar evolutionary rates for different kingdoms.

The α and β subunits of ATPase show the reverse relationship between archaebacterium and eubacteria in the composite tree of Fig. 3; i.e., the archaebacterial α and β subunits are clustered with the eubacterial β and α subunits, respectively. The bootstrap resampling revealed that this tree topology is highly likely with the bootstrap probability of 0.74 (Fig. 5), suggesting that so-called *Sulfolobus* ATPase α subunit (β subunit) is the archaebacterial homolog of the eubacterial F_1-ATPase β subunit (α subunit), from an evolutionary viewpoint.

DISCUSSION

Simultaneous comparisons of EF-Tu and EF-G from archaebacteria, eubacteria, and eukaryotes and also of α and β subunits of ATPase from the three primary kingdoms have shown that archaebacteria are phylogenetically more closely related to eukaryotes than to eubacteria. In these proteins the close similarity in sequence positively correlates with the close relatedness in phylogeny. Also 5S rRNA, RNA polymerases, and several ribosomal proteins from archaebacteria show sequence similarities to eukaryotes much more strongly than to eubacteria. Therefore, the phylogenetic relationship among the primary kingdoms presented here would be generally correct and would not be restricted only to the cases specific for the four protein species.

Based on the strong sequence similarity of 5S RNA, Hori and Osawa (7, 9, 10) proposed that archaebacteria (metabacteria) and eukaryotes are sister urkingdoms. Cavalier-Smith (28) also proposed a similar hypothesis on the basis of the similarities between them at the molecular and cellular levels. These hypotheses are consistent with the phylogenetic tree proposed here. According to the phylogenetic trees of Figs.

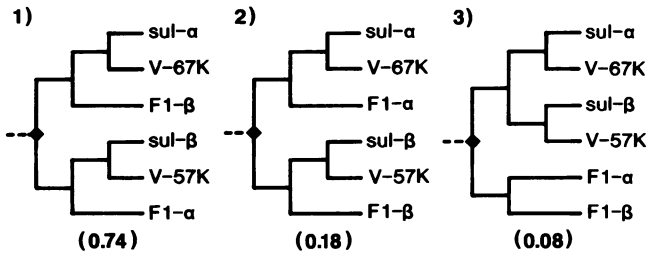


FIG. 5. Phylogenetic relationship of the α and β subunits of *Sulfolobus* ATPase with the eubacterial and eukaryotic counterparts. sul-α and sul-β, α and β subunits of *Sulfolobus* ATPase, respectively; F1-α and F1-β, α and β subunits of eubacterial F_1-ATPase, respectively; V-67K and V-57K, 67-kDa and 57-kDa subunits of eukaryotic vacuolar ATPase, respectively. The bootstrap probabilities (indicated in parentheses) correspond to the three tree topologies.

2 and 3, both *M. vannielii*, a methanobacterium, and *Sulfolobus acidocaldarius*, a sulfobacterium, are more closely related to eukaryotes than to eubacteria, which apparently contradicts with the phylogenetic tree (eocyte tree) proposed by Lake (4, 5); according to the eocyte tree, *M. vannielii* should be more closely related to eubacteria rather than to eukaryotes.

The phylogenetic relationship presented here should be confirmed by composite trees of other protein sequences. A pair of proteins, lactate dehydrogenase (LDH) and malate dehydrogenase (MDH), may be one of them. These proteins show a marked sequence similarity in the nucleotide binding domain (29). It is highly likely that both genes diverged by gene duplication prior to the divergence of the three primary kingdoms. Fig. 6 shows the composite phylogenetic tree inferred from a simultaneous comparison of MDH and LDH sequences. Unfortunately no archaebacterial sequence is available at present.

The phylogenetic position of *Thermus flavus*, a eubacterium, is surprising (Fig. 6). The MDH sequence of *T. flavus* is much more similar to that of a eukaryote than those of eubacteria (30) and it behaves like an archaebacterium on this phylogenetic tree. There may be several interpretations for this observation. (*i*) Horizontal gene transfer of the MDH gene occurred from a eukaryote or archaebacterium to *T. flavus*. (*ii*) *T. flavus* is an archaebacterium but not a eubacterium. (*iii*) Eubacteria are composed of a large number of distinct groups and a group to which *T. flavus* belongs may be more closely related to archaebacteria and eukaryotes than to other eubacterial groups. If so, archaebacteria are likely to have diverged from an ancestral eubacterium resembling *T. flavus*. Determination of the amino acid sequence of *T. flavus* LDH as well as those of archaebacterial MDH and LDH will clarify this problem.

There are other proteins that are expected to have diverged by gene duplication before the separation of the three primary kingdoms. Initiation factor 2 (31) and LepA protein (the signal peptidase I) (32) show sequence similarities to each other as well as with elongation factors. Valyl-tRNA synthetase and isoleucyl-tRNA synthetase also show an extensive sequence similarity (33).

Finally, two examples that appear to contradict to our conclusion will be discussed. The sequence similarity of both the small and large rRNAs appears to be higher between archaebacteria and eubacteria than between archaebacteria and eukaryotes (1–3, 6). In the light of our evolutionary tree, it is possible that mutational changes have accumulated more rapidly in the eukaryotic rRNAs than in the archaebacterial and eubacterial rRNAs shortly after the separation of archaebacteria and eukaryotes.

The strong sequence divergence of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) from an archaebacte-

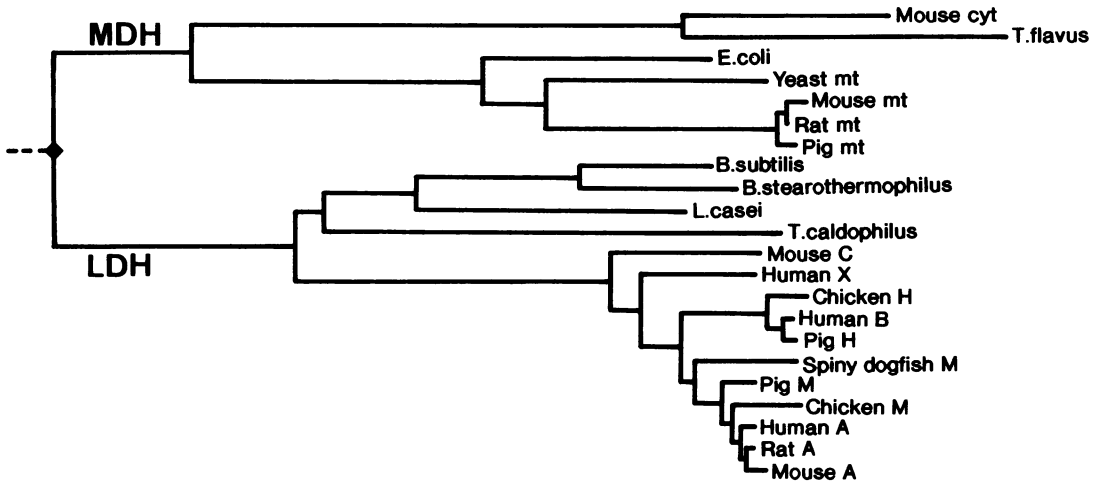


FIG. 6. Composite phylogenetic tree inferred from a simultaneous comparison of MDH and LDH. cyt and mt, nuclear gene-coded cytoplasmic and mitochondrial isozymes, respectively. Sequence data were taken from the National Biomedical Research Foundation data base (release 20.0) and GenBank data base (release 59.0). The sequence data used were checked with their original data.

rium *Methanothermus fervidus* may be a matter of much debate. The similarity between the archaebacterial GAPDH and eubacterial or eukaryotic counterpart is much less than that between the latter two (34). A plausible explanation, which reconciles with our phylogenetic tree, may be that a gene duplication event occurred in GAPDH locus before the separation of the three urkingdoms and one of the duplicates functions only in archaebacteria and the remaining functions in both eubacteria and eukaryotes.

We thank Drs. R. Doolittle and M. Bulmer for discussion and Prof. I. Takeuchi for permitting us to use the *Dictyostelium* EF-2 sequence before publication. This work was supported in part by grants from the Ministry of Education, Science and Culture of Japan.

1. Woese, C. R. & Fox, G. E. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5088–5090.
2. Woese, C. R. & Olsen, G. J. (1986) *Syst. Appl. Microbiol.* **7,** 161–177.
3. Woese, C. R. (1987) *Microbiol. Rev.* **51,** 221–271.
4. Lake, J. A., Henderson, E., Oakes, M. & Clark, M. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3786–3790.
5. Lake, J. A. (1988) *Nature (London)* **331,** 184–186.
6. Gouy, M. & Li, W.-H. (1989) *Nature (London)* **339,** 145–147.
7. Hori, H. & Osawa, S. (1987) *Mol. Biol. Evol.* **4,** 445–472.
8. Pühler, G., Leffers, H., Gropp, F., Palm, P., Klenk, H.-P., Lottspeich, F., Garrett, R. A. & Zillig, W. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 4569–4573.
9. Hori, H. & Osawa, S. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 381–385.
10. Osawa, S. & Hori, H. (1979) in *Ribosomes, Structure, Function and Genetics*, eds. Chambliss, G., Craven, G. R., Davies, J., Davis, K., Kahan, L. & Nomura, M. (University Park Press, Baltimore), pp. 333–355.
11. Lechner, K. & Böck, A. (1987) *Mol. Gen. Genet.* **208,** 523–528.
12. Lechner, K., Heller, G. & Böck, A. (1988) *Nucleic Acids Res.* **16,** 7817–7826.
13. Berghöfer, B., Kröckel, L., Körtner, C., Truss, M., Schallenberg, J. & Klein, A. (1988) *Nucleic Acids Res.* **16,** 8113–8128.
14. Matheson, A. T., Yaguchi, M., Balch, W. E. & Wolfe, R. S. (1980) *Biochim. Biophys. Acta* **626,** 162–169.
15. Kimura, M., Arndt, E., Hatakeyama, T., Hatakeyama, T. & Kimura, J. (1989) *Can. J. Microbiol.* **35,** 195–199.
16. Denda, K., Konishi, J., Oshima, T., Date, T. & Yoshida, M. (1988) *J. Biol. Chem.* **263,** 6012–6015.
17. Denda, K., Konishi, J., Oshima, T., Date, T. & Yoshida, M. (1988) *J. Biol. Chem.* **263,** 17251–17254.
18. Miyata, T., Hayashida, H., Kikuno, R., Toh, H. & Kawade, Y. (1985) in *Interferon 6*, ed. Grosser, I. (Academic, London), pp. 1–30.
19. Kimura, M. (1983) in *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, England), pp. 55–97.
20. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
21. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
22. Toda, K., Tasaka, M., Mashima, K., Kohno, K., Uchida, T. & Takeuchi, I. (1989) *J. Biol. Chem.*, in press.
23. Nelson, H., Mandiyan, S. & Nelson, N. (1989) *J. Biol. Chem.* **264,** 1775–1778.
24. Brusilow, W. S. A., Scarpetta, M. A., Hawthorne, C. A. & Clark, W. P. (1989) *J. Biol. Chem.* **264,** 1528–1533.
25. Post, L. E. & Nomura, M. (1980) *J. Biol. Chem.* **255,** 4660–4666.
26. Zengel, J., Arch, R. H. & Lindahl, L. (1984) *Nucleic Acids Res.* **12,** 2181–2192.
27. Ohama, T., Yamao, F., Muto, A. & Osawa, S. (1987) *J. Bact.* **169,** 4770–4777.
28. Cavalier-Smith, T. (1987) *Ann. N.Y. Acad. Sci.* **503,** 17–54.
29. Rossmann, M. G., Liljas, A., Bränden, C.-I. & Banszak, L. J. (1975) in *The Enzymes*, ed. Boyer, P. D. (Academic, New York), 3rd Ed. Vol. II, pp. 61–102.
30. Nishiyama, M., Matsubara, N., Yamamoto, K., Iijima, S., Uozumi, T. & Beppu, T. (1986) *J. Biol. Chem.* **261,** 14178–14183.
31. Sacerdot, C., Dessen, P., Hershey, J. W. B., Plumbridge, J. A. & Grunberg-Manago, M. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 7787–7791.
32. March, P. E. & Inouye, M. (1985) *J. Biol. Chem.* **260,** 7206–7213.
33. Jordana, X., Chatton, B., Paz-Weisshaar, M., Buhler, J.-M., Cramer, F., Ebel, J. P. & Fasiolo, F. (1987) *J. Biol. Chem.* **262,** 7189–7194.
34. Fabry, S. & Hensel, R. (1988) *Gene* **64,** 189–197.