Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 7749–7754, July 1996
Evolution

The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny

SANDRA L. BALDAUF*†, JEFFREY D. PALMER‡, AND W. FORD DOOLITTLE*

*Canadian Institute for Advanced Research and Department of Biochemistry, Dalhousie University, Halifax, NS B3H 4H7, Canada; and ‡Department of Biology, Indiana University, Bloomington, IN 47405

Communicated by John C. Avise, University of Georgia, Athens, GA, April 12, 1996 (received for review December 12, 1995)

ABSTRACT **The genes for the protein synthesis elongation factors Tu (EF-Tu) and G (EF-G) are the products of an ancient gene duplication, which appears to predate the divergence of all extant organismal lineages. Thus, it should be possible to root a universal phylogeny based on either protein using the second protein as an outgroup. This approach was originally taken independently with two separate gene duplication pairs, (*i*) the regulatory and catalytic subunits of the proton ATPases and (*ii*) the protein synthesis elongation factors EF-Tu and EF-G. Questions about the orthology of the ATPase genes have obscured the former results, and the elongation factor data have been criticized for inadequate taxonomic representation and alignment errors. We have expanded the latter analysis using a broad representation of taxa from all three domains of life. All phylogenetic methods used strongly place the root of the universal tree between two highly distinct groups, the archaeons/eukaryotes and the eubacteria. We also find that a combined data set of EF-Tu and EF-G sequences favors placement of the eukaryotes within the Archaea, as the sister group to the Crenarchaeota. This relationship is supported by bootstrap values of 60–89% with various distance and maximum likelihood methods, while unweighted parsimony gives 58% support for archaeal monophyly.**

The use of primordially duplicated proteins to root the tree of life was pioneered by Gogarten *et al.* (1) for the catalytic and regulatory subunits of the V- and F-type ATPases and by Iwabe *et al.* (2) for the elongation factors Tu (EF-Tu) and G (EF-G). These analyses divide all living organisms into two clades, one consisting of the true bacteria, or eubacteria, and the other consisting of the archaeons and eukaryotes. The ATPase analyses are now complicated by the discoveries of eukaryotic/archaeal (V-type) ATPases in some eubacteria and of a eubacterial (F-type) ATPase in at least one archaebacterium, raising the possibility of multiple horizontal gene transfers (3). The elongation factor analyses have also been criticized, for the limited size of the homologous region, for alignment errors, and for inadequate taxonomic sampling [two eukaryotes, two eubacteria, and one archaebacterium for EF-G; and a few additional animal/fungal and organellar sequences for EF-Tu (4)].

We have reanalyzed the elongation factor rooting with the much larger and more broadly representative data base now available for both proteins. This joint analysis now includes three to five representatives of each of the two archaeal kingdoms and a broad sampling of both eukaryotes and eubacteria. We have also modified and expanded the alignment of Iwabe *et al.* (2) using consensus sequences and crystal structures for both proteins. We find that rooted phylogenies for both elongation factors strongly support the Iwabe/Gogarten rooting for the universal tree.

These analyses address a second major issue, the origin of eukaryotes. Most data argue for a monophyletic Archaea composed of two kingdoms, crenarchaeotes and euryarchaeotes, with eukaryotes and eubacteria each arising separately (5–9). However, Lake *et al.* (10, 11) have argued for a polyphyletic Archaea, with a paraphyletic Euryarchaeota giving rise to eubacteria (the photocyte hypothesis) and a monophyletic Crenarchaeota arising with eukaryotes (the eocyte hypothesis). These hypotheses were originally based on ribosome morphology (10) and were later supported by Lake's analyses of small subunit rRNA (11). However, these small subunit rRNA analyses have been challenged (12). In addition, the ribosome data have been criticized as artifactual (13) and are further challenged by the presence of "eocyte-specific characters" in two taxa now known to be euryarchaeotes (*Thermoplasma* and *Thermococcus*; refs. 5–9). Our results with a combined EF-Tu/EF-G data set strongly reject both the original and rerooted (14) forms of the photocyte hypothesis, but they support the sisterhood of crenarchaeotes and eukaryotes.

METHODS

Sequences were initially aligned by computer using the Genetics Computer Group (Madison, WI) program PILEUP (15) with default gap penalties. Minor modifications were made by eye to minimize within-kingdom insertion/deletion events, and the alignments were used to construct separate, kingdom-specific consensus sequences for both proteins. For EF-Tu, sites were scored as conserved if 85% of all eukaryotes or all eubacteria or 80% of all archaebacteria shared the same amino acid at a given position; for EF-G, a value of 85% conservation was used for all three kingdoms. More stringent criteria (85–90% identity for eubacteria and eukaryotes, respectively) were required if the variation was found primarily in early branching lineages, and a 5–10% lower stringency was used if the variation was mostly restricted to a single phylum. Use of stringent criteria to construct the consensus sequences avoided the necessity of using a phylogenetic method for this purpose, which might bias the results of subsequent analyses. The consensus sequences covered a total of 59%, 65%, and 71% of all positions of EF-G for archaebacteria, eubacteria, and eukaryotes, respectively, and ≈75% of all positions for all taxa for EF-Tu. The Dayhoff PAM 250 matrix (16) was used to define conservative substitutions.

These consensus sequences and crystallographically determined secondary structures (17, 18) and the structure-based alignment of Ævarsson (19) were then used to refine the alignment among kingdoms and to identify regions of likely homology. Priority was given to placing gaps at the edges of structural elements or within loops connecting elements. Segments of the alignment were accepted as homologous if their

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, elongation factor Tu; EF-G, elongation factor G.

†e-mail: sbaldauf@is.dal.ca.

consensus sequences were readily alignable throughout and bordered by well-conserved sequence in all kingdoms. The "Oryzadopsis" EF-G is a composite of partial cDNA sequences from the angiosperms *Oryza* and *Arabidopsis* and includes 634 out of 824 total amino acid positions. (Sequences and alignments are available upon request from S.L.B.)

All trees were constructed from amino acid sequences, with some results also tested using first and second codon position nucleotides. Parsimony analyses used PAUP version 3.1.1 (20). Shortest tree searches consisted of 50 replicates of random sequence addition with TBR (tree-bisection-reconnection) branch swapping. Bootstrap analyses used 100–500 replicates of a single round of random addition each. For these (parsimony) analyses only, the joint data set was augmented with an additional 173 EF-Tu-specific and 300 EF-G-specific positions, which were scored as missing data for each other. This increased the resolution in terminal clades, thus greatly reducing analysis times. This use of missing data should not affect the branching order in deeper lineages (21).

Distance analyses used the neighbor-joining algorithms of PHYLIP version 3.5c (22) and MEGA 1.0 (23). PHYLIP analyses consisted of 100 or 500 bootstrap replicates using two different substitution matrices—the Dayhoff PAM250 and the George–Hunt–Barker chemical index (22). MEGA analyses consisted of 500 bootstrap replicates using a gamma distribution to correct for rate variation among sites. The gamma variable alpha was estimated as described (24) using parsimony trees (20) and was found to be ≈0.7 for both proteins with various combinations of taxa.

Maximum likelihood analyses used the program PROTML version 2.2 (25). The combined data set was analyzed by the RELL (resampling of estimated log-likelihood) bootstrap method using the 1000 best trees found from an exhaustive search of a partially constrained starting tree. Bootstrap and SE values were also determined from individual EF-Tu and EF-G data sets using fully resolved trees. All PROTML analyses used the Jones–Taylor–Thornton (JTT; ref. 26) substitution matrix. Maximum likelihood analyses were also performed using the PAML version 1.1 program CODEML (27). Fully resolved trees were analyzed using the JTT matrix and an eight-component gamma model with the starting alpha value set at 0.7.

All nucleotide-level analyses used the PHYLIP version 3.5C (22) programs for maximum likelihood (DNAML) or distance (DNADIST). DNA maximum likelihood analyses were done for the combined data only and consisted of 50 bootstrap replicates using a transition/transversion ratio of 1.0, empirical base frequencies, and global branch rearrangement to search for the best tree. The EF-G nucleotide data was analyzed by distance using Jukes–Cantor weighting with 100 bootstrap replicates and trees constructed by neighbor-joining (22).

RESULTS

Root of the Universal Tree. The alignment of consensus sequences, which covers the entire GTP-binding domain of both proteins, confirms the homology of the amino termini of EF-Tu and EF-G (Fig. 1). These domains contain seven colinear blocks of conserved, alignable sequence interspersed with more rapidly evolving regions of variable size and uncertain alignment. In all cases, the latter regions correspond to variable-sized loops predicted to lie at the surface of the protein (Fig. 1; refs. 17 and 18). Altogether, 158 amino acid positions of likely homology between EF-Tu and EF-G (overlining in Fig. 1), corresponding to ≈60–75% of the GTP-binding domains of both proteins (Fig. 1), were identified. Of these, 102 positions were identified as alignable with strong confidence (underlining in Fig. 1).

Our alignment (Fig. 1) differs from that of Iwabe *et al.* (2) in that we include three additional blocks of apparent homology (F–H), for which weak sequence similarity is strengthened by secondary structure data (17–19). In addition, Iwabe *et al.* (2) aligned our positions 30–37 of archaeal and eukaryotic EF-Tu with positions 56–63 of all other sequences. This is probably incorrect, because it is a much poorer match and also requires two additional, large gaps. Our alignment also differs in placing a single amino acid gap at position 4 of EF-G and in placing the single amino acid gap of region E at position 146 instead of 133. In the latter case, results were tested with both versions of the alignment (see below). Our alignment uses the predicted G and H regions of Ævarsson but differs from his alignment of region F in that he aligned our positions 158–170 of EF-Tu with positions 160–173 of EF-G (19).

All analyses of the joint EF-Tu/EF-G data set support the Gogarten/Iwabe rooting of the universal tree (1, 2), by con-

```
            1            A             30  31                         55  56        B/C          79         80          D             105
Tu-K [4+]  K•HiNiVVIGHVDSGKSTtTGHLIYKCGGI [D•R•IEKFEKEaaEmGKGSFKYAWV] LDKLKAERERGITIDIaLWKFET• [ 0 ] KY••TiIDAPGHRDFIKNMITGTSQA
   A [3+]  KPHlNlv•IGHVDHGKST•vGRLLyd•G•i [•e••i••••EEa•••GK•sF•FAw•] mDrLKEERERGVTId•a••kFET• [ 0 ] •Y•fTIIDAPGHRDFVKNMITGASQA
   B [9 ]  KPHvNiGTIGHVDHGKTTLTAAIT••La•• [••**•••••y•--------------] ID•APEEkaRGITIn•aHVEYeT• [ 0 ] •RHYAHVDCPGHADYvKNMITGAAQM
             ^^^β1^^^       ~~~~~~~~αA~~~~~~  ~~   .........            ..........  ^^^^^β2^^^^ .        ^^^β3^^^    ~~~αB~~~~
 G-K [17]  NIR-NMSVIAHVDHGKSTLTDSLv••AGII [---a•••AG•-------------•Rf] TDTR•DEQER•ITIKSTgiSlyyE [14+] •FLINLIDSPGHVDFSSEVTAALRVT
   A [18]  qiR-NIGI•AHvDHGKTTLSDnLLAgaGmi [---S••lAG•-------------•L•] LDy•e•EQ•RGITi•AAN•Sm•H• [ 4 ] •YLINLIDTPGHVDFgG•VTRamR•i
   B [6+]  k•R-NIGI•AHIDAGKTTTTERILfYTG•• [••iGEV•eG--------------•AT] •DWM•QE•ERGITITaAa•t•fW• [ 0+] ••rINIIDTPGHVDFTiEVERSmRVL
             ^^ ^β1^^^       ~~~~~~αA~~~~~~   .........           ...  ..........     ^^^β2^^          ^^β3^^         ~~~~αB~~~~

      106  D   117 118      127 128            E              155            158     F      173         175  G  186           189  H   200
Tu-K  D•AiLii•a••G [EFEAGISKdG] ---QTREHALLAfTLGVKQmIV••NKMD --    [8+] yeEI•K-EvS•ylKKv Gy [4+] vpFvPISGw•G DN [15+] LleALD•-i••P
  AC               [EfEAGMS•EG] ---
  AE  DAAvLVVsA••G [----##----] v.pQTrEH•fLArTlGi•qliVAvNKMD --    [8+] y•ev••*•••K•Lk•• Gf [2+] ••FiPvSAw•G dN [15 ] L•E•LD•-•••P
   B  DGAILVVsAaDG [----------] PMPQTREHiLLArQVGVP•IvVFlNK•D --    [6+] LLeLVEmEvReLLs•Y •f [3+] •PiirGSAL•A LE [11+] LmdavD••iP•P
       ^^^β4^^                    ~~~~~~αC~~~~~~~   ^^^β5^^        ~~   ~~~~~~αD~~~~~~~~~           ^β6^ ~~αE ~~        ~~~αF~~~~
 G-K  DGALVVVDCv•G [----------] VCVQTETVLRQAlaERik-Pv••iNKvD R•  --  lLEL••-••Ee•y•sF -- [28+ GtVaFGSGLHG -- 112+] lLeMi••HLPSP
   A  DG•iVVvDAVEG [----------] vM•QTETVlRQaL•E•Vk-PvLFINKvD RL  --  I•ELk*-t•qEm••R• -- [29+ G•V•FGSA••n --  39 ] iLDM•ik••PnP
   B  DG•v•VF•a••G [----------] VePQSEtVWRQAdkY•VP-RI•FvNKMD r•  --  GADFy•-••••i••rL -- [96+ •PV•CGSA•KN --   5 ] lLDAVv••LP•P
       ^^β4^^                     ~~~~~~αC~~~~~    ^^β5^^              ~~~ ~αD~~~~~          ^^β6^                ~~~αE~~~
```

FIG. 1. Consensus alignment of the amino termini of EF-Tu and EF-G for eukaryotes (K), Archaea (A), and eubacteria (B). Universal or highly conserved positions are shown in uppercase letters; positions with only conservative substitutions are shown in lowercase letters corresponding to the most common amino acid found at that position (16). Capital letters above the alignment indicate blocks of homology; blocks A–E correspond to regions A–E of Iwabe *et al.* (2). The Crenarchaeota (AC) and Euryarchaeota (AE) are indicated separately for positions 118–130; the symbols ## correspond to the sequence GE in *Thermoplasma acidophilum* and AKS in *Methanococcus vanellii.* Homologous positions are indicated by overlining, with the conservative, 102-position core further indicated by underlining. Nonconservative but universally present sites are indicated by •, gaps by –, and positions that are not present in all taxa are indicated by *. Unalignable regions are indicated by numbers corresponding to their overall length; a number followed by + indicates the minimum size of a region that differs in length among taxa. Secondary structure elements are indicated below the alignment by ^ for beta strands, and ~ for alpha helices. The trypsin-sensitive effector region, for which the EF-G structure is unknown, is indicated under the alignment by ·. Bracketed regions were omitted from all analyses.

sistently placing the Archaea together with the eukaryotes to the exclusion of all eubacteria (Fig. 2). Maximum parsimony (20) and neighbor-joining analyses (22) of the conservatively defined 102-site data set gave 87–93% bootstrap support for this rooting for the EF-Tu subtree and 81–95% support for the EF-G subtree (Fig. 2). Similar results were found with the variation of the 102-site alignment (ref. 2, see above) modified at positions 133–146 (85–94% for EF-Tu and 88–95% for EF-G). Consistently stronger values were found for the larger, 158-site data set (97–98% for EF-Tu and 95–96% for EF-G). This rooting for the universal tree was also supported by high bootstrap values (91–99% with EF-Tu and 94–97% for EF-G) when three-way analyses were performed using another member of the GTPase-protein superfamily (28), protein synthesis initiation factor 2, as an additional outgroup (data not shown).

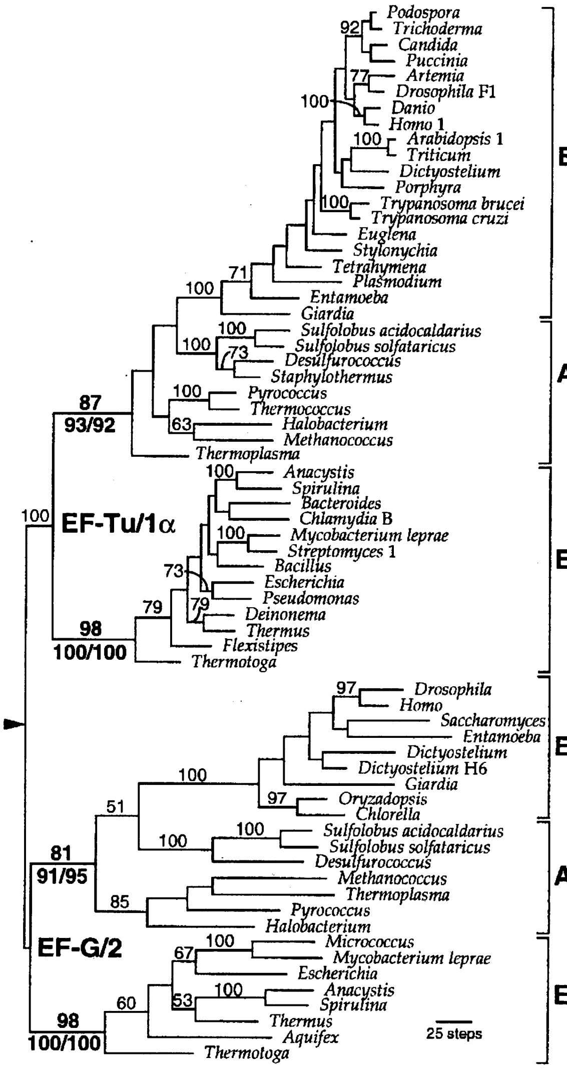

FIG. 2. A joint EF-Tu/EF-G protein phylogeny roots the universal tree between eubacteria and Archaea/eukaryotes. The tree shown is one of six shortest trees derived by parsimony analysis. Branches are drawn to scale (see bar). Brackets on the right indicate eukaryotic (euk), archaeal (arc), and eubacterial (eub) derived sequences. The arrow indicates the proposed point of gene duplication. The tree is 3416 steps long and has a consistency index of 0.580, exclusive of uninformative characters, and a retention index of 0.756. Parsimony bootstrap values >50% are indicated above the branches. Bootstrap values for distance analyses using the substitution matrices of Dayhoff and George–Hunt–Barker are indicated in that order below the branches of primary interest only. The branch connecting the EF-Tu and EF-G subtrees is short due to the fact that it is based on the 102 shared amino acid positions only, while the terminal branches are based on 173 and 300 additional sites for EF-Tu and EF-G, respectively. Alternative trees at this length place *Pyrococcus* and *Thermococcus* EF-Tu as a separate branch from *Halobacterium* and *Methanococcus* EF-Tu, and switch the branching position of the Gram-positive bacteria with that of the proteobacteria EF-Tus.

Origin of Eukaryotes. Both protein-specific subtrees of the joint analysis of Fig. 2 support, albeit weakly, the crenarchaeotes as the sister group to eukaryotes. To further investigate this issue, individual EF-Tu and EF-G data sets were analyzed more thoroughly. These data sets were restricted to regions for which homology could be determined with confidence based on the alignment of domain-specific consensus sequences, and consisted of 295 amino acid positions for EF-Tu and 382 positions for EF-G. Most analyses of both data sets weakly support the crenarchaeotes as the sister group to eukaryotes, with bootstrap values of 35–52% for EF-Tu and 50–60% for EF-G. Consistently stronger support was found with maximum likelihood analyses. These showed both data sets as supporting their specific topologies in Fig. 2 over ones showing monophyletic Archaea by 84% bootstrap (0.935 SEs) for EF-G and 89–93% bootstrap (1.15–1.39 SEs) for EF-Tu. The two data sets were then combined for further analysis into a single data set totaling 677 sites and consisting of all taxa sequenced for both proteins.

With the exception of parsimony, all analyses of this combined amino acid data set support a paraphyletic Archaea with the crenarchaeotes as the sister group to eukaryotes (Fig. 3). Strongest support came from maximum likelihood analyses of amino acids, which found 89% bootstrap support for this affiliation (Fig. 3). Only a single tree supporting monophyletic Archaea was found within one SE of the best tree (Fig. 3), and this tree also required placement of *Halobacterium* as the deepest branch among euryarchaeotes. The topology shown in Fig. 3 was favored over an otherwise identical tree showing monophyletic Archaea by a difference in log-likelihood of 5.7, corresponding to a confidence level of 72% by the Kishino–Hasegawa test (SE = 5.3; ref. 25). Similarly, a 6.03 difference in log-likelihood between these trees was found by maximum likelihood analysis using a gamma distribution to correct for rate variation among sites. Neighbor-joining analyses using gamma-corrected distances gave moderate support for

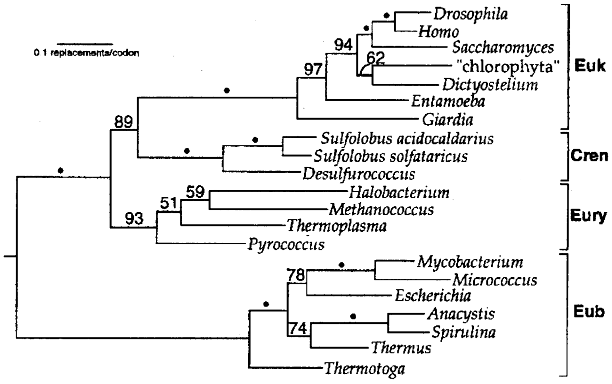

FIG. 3. A Combined EF-Tu/EF-G phylogeny places the origin of eukaryotes within the Archaea. The tree shown is the single best tree derived by maximum likelihood analysis of combined EF-Tu and EF-G sequences. Analyses utilized a semiconstrained starting tree; constrained nodes are indicated by ·. Branches are drawn to scale as indicated by the scale bar; numbers above the nodes indicate bootstrap values. Brackets on the right are as in Fig. 2, except that the Archaea are divided into crenarchaeotes (cren) and euryarchaeotes (eury). The tree shown has a log-likelihood of −16054.19, while an identical tree with monophyletic Archaea has a log-likelihood of −16059.9 (25). When a gamma correction is used (27), these two topologies have log-likelihoods of −14189.26 and −14195.29, respectively. "Chlorophyta" is a composite of the *Arabidopsis* EF-Tu and *Chlorella* EF-G sequences. The tree is rooted based on the data presented in Fig. 2.

paraphyly (79% bootstrap), while neighbor-joining analyses without this correction gave weak support (60–65% bootstrap). Unweighted parsimony analysis, on the other hand, weakly supported monophyly (58% bootstrap). Limited jackknife analyses (29) showed that while the distance methods were fairly robust to the taxonomic composition of the data set, parsimony was not, with bootstrap values ranging from 69% support for monophyletic Archaea to 89% support for paraphyly, depending on which taxa were used. Analyses of first and second codon nucleotide positions by maximum likelihood, which did not require the use of constraints, also supported paraphyly (66% bootstrap).

DISCUSSION

Root of the Universal Tree. The strong similarity in the amino termini of EF-Tu and EF-G, both in terms of structure and of sequence (Fig. 1), supports a relationship of homology—i.e., they appear to be the products of an ancient gene duplication. In fact, secondary structure comparisons strongly suggest that this homology also includes the second domain of both proteins (18, 19). Thus, ≈40% of EF-G appears to be homologous to 75% of EF-Tu (19). The antiquity of this duplication is shown by the presence of the EF-Tu- and EF-G-encoding genes in all major groups of organisms (Fig. 2). In fact, these genes may be the products of a tandem duplication, as they follow one another in the *str* operon (30). This arrangement occurs in both Archaea and eubacteria (30), a fact that further weakens the possibility of deep paralogy confusing these analyses.

Phylogenetic analyses of the aligned amino-termini of EF-Tu and EF-G, including a broad representation of taxa for all three domains of life for both proteins, places the root of the universal tree between the eubacteria and the Archaea/eukaryotes. These results are supported by high bootstrap values for both proteins by all methods used. This rooting of the universal tree is also supported by a recent analysis of another ancient gene duplication, involving tRNA synthetases (6). Our analyses strengthen and expand the findings of Iwabe *et al.* (2) by including a large, broadly representative group of taxa, correcting and extending the EF-Tu/EF-G alignment, and examining the results with multiple methods of analysis.

Striking similarity between archaeal and eukaryotic sequences, sometimes including large insertions/deletions (1, 8, 31, 32), is also seen with 5S rRNA (7), nearly all ribosomal proteins with known homologs from all three domains (30), the largest and second largest subunits of RNA polymerase (8, 32), protein synthesis initiation factor 2 (33), and the key recombination protein RecA (34). Archaea also contain homologs of the eukaryote TATA-binding protein and transcription factors TFIIB and BRF and have eukaryote-like promoters, which interact efficiently with eukaryotic transcription factors *in vitro* (35). Archaea, like eukaryotes, have short tRNA introns, most of which are located one nucleotide 3′ of the anticodon (36), uniformly add the 3′-terminal CCA to their tRNAs posttranscriptionally (37), and use the protein fibrillarin and a U3-like RNA in rRNA processing (38, 39). Both groups probably also use family B-type DNA polymerases for replication, whereas eubacteria use this enzyme exclusively for repair (40).

Thus, most aspects of archaeal DNA, RNA, and protein synthesis resemble those of eukaryotes. The most notable exceptions to this are 16S and 23S rRNA, for which Archaea are closer to eubacteria in sequence (5, 41). We suggest that this reflects high rates of rRNA sequence evolution in eukaryotes rather than a fundamental difference in the phylogeny of rRNA and protein genes. A similar situation is seen with the isoleucyl-tRNA synthetases, which also show greater overall sequence similarity between Archaea and eubacteria, but which unite Archaea and eukaryotes when analyzed phylogenetically, using an outgroup (6). In principle, then, if outgroup sequences were available for the rRNAs, they should also root these trees in the same place as the gene duplications (Fig. 2).

Widely conflicting results to both our trees (Fig. 2) and trees based on rRNA are found with glutamine synthetase (42), glutamate dehydrogenase (43), and the 70-kDa heat-shock proteins (44). However, phylogenetic trees based on these proteins tend to be deeply incongruent with each other as well, supporting various combinations of paraphyletic or even polyphyletic eubacteria and/or Archaea. This is more consistent with independent horizontal gene transfers or comparisons of paralogous sequences (45). One or both of these explanations now appears to be the case with nitrogenase (46) and with glyceraldehyde-3-phosphate dehydrogenase (47), which was originally interpreted as an exception to the universal rooting presented here (Fig. 2).

Origin of Eukaryotes. The elongation factor data now comprise a sufficiently large data set—i.e., they are broadly sampled in eukaryotes, eubacteria, and both branches of Archaea—that we may begin to test the relationship between Archaea and eukaryotes using these sequences. We find that a combined data set of both proteins most consistently supports the crenarchaeotes as the sister group to eukaryotes, albeit at varying levels of confidence depending on the method of analysis. Strongest support comes from maximum likelihood analyses of amino acids (Fig. 3) and from neighbor-joining analyses of gamma-corrected distances (89% and 79% bootstrap, respectively). Recent simulations suggest that these methods are probably the most robust at large evolutionary distances, since they are least affected by rate variation among sites and among lineages (48, 49). Nonetheless, while our results are certainly suggestive, they should be interpreted with caution, as the data set is still limited in terms of taxon sampling, especially among Archaea and eukaryotes. Thus, long branch effects (50) and other artifacts could still be problematic and difficult to detect, as suggested by the lack of consistent results among methods. Greater resolution of this issue will require the development of other molecular data sets as well as further development of the elongation factor data.

However, in terms of the euryarchaeotes, our combined EF-Tu/EF-G data strongly support their monophyly (Fig. 3), as do almost all other relevant molecular phylogenetic data (5–9). Thus, the photocyte hypothesis, which postulates paraphyletic euryarchaeotes in either its original (11) or rerooted (14) form, can be soundly rejected.

The results of our analyses of EF-G contradict those of Creti *et al.* (9), who found 63–99% bootstrap support for monophyletic Archaea versus our 50–84% support for paraphyly. Our analyses differ from theirs in that we have largely used amino acids rather than nucleotides and included a broader taxonomic sampling. However, the primary difference appears to be the fact that we did not use roughly 30% of the Creti *et al.* alignment (169 amino acid positions; ref. 9), because we do not find evidence clearly supporting the homology of these regions among the major groups. Consistent with this, even when we analyze our character set at the nucleotide level using the smaller taxon sample of Creti *et al.* (9), we find only 61% bootstrap support for monophyly with first and second codon positions and 58% support with second positions alone. If we expand the data set to include the same taxa as in Fig. 2, these values fall to 50% bootstrap for monophyly versus paraphyly. We suggest that the use of ambiguously aligned regions for phylogeny is questionable, since it cannot be assumed that the sequences found in these regions in different taxa are homologous, and phylogenetic results based on these data can be easily influenced by any bias in the method of sequence alignment.

Undoubtedly the most widely used and influential source of evidence for archaeal monophyly has been ribosomal RNA. It is important to realize, however, that the level of support for monophyly in the latest and in some senses most comprehensive analyses of both small (51) and large (52) subunit rRNA is only moderate. Furthermore, the strongest support for monophyly comes from parsimony analyses, which is also the only method supporting monophyly with our data. Given the uneven rates of evolution evident in both data sets (Fig. 3; refs. 51 and 52), it should be noted that parsimony is especially prone to fail in this situation (48). Therefore, we suggest that, just as the exceptionally long eukaryotic branch distorts the (midpoint) rooting of the entire rRNA tree (see *Results*), the exact placement within the tree of this long branch may also be incorrect. In particular, the rRNA placement of eukaryotes could reflect spurious attraction (11, 50) between the two longest branches on the tree (leading to eukaryotes and eubacteria), which would force together the short branches leading to crenarchaeotes and euryarchaeotes and thus artifactually give archaeal monophyly. Likewise, the grouping of crenarchaeotes and euryarchaeotes in our parsimony analyses of elongation factor data could be the result of the same phenomenon.

Other kinds of data have of course been brought to bear on the issue of archaeal monophyly. Supporting monophyly are (*i*) individual phylogenetic analyses of isoleucyl tRNA synthetase (6), 5S rRNA (7), and RNA polymerase largest subunit (8), (*ii*) a shared split in their RNA polymerase largest subunit genes (*rpo*A′A″; ref. 8), and (*iii*) the exclusive presence of isoprenoid ether-linked lipids in their membranes (53). No single analysis is fully compelling: the phylogenetic studies suffer from poor taxon sampling, and the RNA polymerase results could be rationalized by a fusion of linked genes early in eukaryote evolution. The differences in membrane lipids are indeed striking. However, these must have arisen through a transition state in which both lipid types were maintained in the same cell, and, thus, all scenarios require only loss(es) of an ancestrally present redundant lipid type.

Analyses supporting a crenarchaeote/eukaryote clade include phylogenetic treatments of 5S rRNA (54), of unambiguously aligned regions of RNA poymerase subunits B/β' (32), and of EF-Tu using paralinear metrics (55). Again, these studies can be criticized, for instance, for poor taxonomic representation. Other characters supporting archaeal paraphyly, such as a uniquely shared lack of a tRNA alanine gene in the rRNA intergenic spacer and a transcriptionally and usually physically unlinked 5S rRNA gene (56), can vary even among closely related taxa. Possibly the single strongest character supporting this relationship is a 7–11 amino acid insertion in the GTPase domain of EF-Tu, shared by crenarchaeotes and eukaryotes (Fig. 1; ref. 14), although the facts are more complex than originally presented, since this region has sustained additional short insertions/deletions in the euryarchaeotes (Fig. 1, legend)

Thus, although we find strong support for the Gogarten/Iwabe rooting of the universal tree, the sporadic nature of the data on the origin of eukaryotes makes it difficult to derive a consensus at this time. Nonetheless, the EF-Tu/EF-G analyses we present here do support a crenarchaeote origin for the eukaryotes (paraphyletic Archaea), as well as speaking strongly in favor of monophyletic euryarchaeotes (thus against the "photocyte" grouping). In supporting a crenarchaeote/eukaryote clade, our data differ from most rRNA phylogenies. It is difficult to imagine that the two classes of molecules have different evolutionary histories. Substantial new data, perhaps from genome sequencing projects, may resolve this disagreement.

Note added in proof. After submission of this manuscript we learned that T. Hashimoto and M. Hasegawa (57) have also found strong support for the sisterhood of Crenarchaebacteria and eukaryotes by maximum likelihood analysis of EF-Tu and EF-G.

We wish to thank C. Delwiche, R. Hoeh, S. Kumar, A. Roger, and Z. Yang for helpful discussion and assistance and A. Ævarsson, S. Barns, J. Czworkowski, C. Delwiche, D. Edgell, P. Keeling, H. P. Klenk, P. Kuhlman, and J. Logsdon for critical reading of the manuscript. This work was supported in part by an A. P. Sloan postdoctoral fellowship to S.L.B., Medical Research Council Grant MT4467 to W.F.D., and National Institutes of Health Grant GM-35087 to J.D.P.

1. Gogarten, J. P., Kibak, H., Dittrich, P., Taiz, L., Bowman, E. J., Bowman, B. J., Manolson, M. F., Poole, R. J., Date, T., Oshima, T., Konishi, J., Denda, K. & Yoshida, M. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 6661–6665.
2. Iwabe, N., Kuma, K.-I., Hasegawa, M., Osawa, S. & Miyata, T. (1989) *Proc. Natl. Acad. Sci. USA* **86**, 9355–9359.
3. Hilario, E. & Gogarten, J. P. (1993) *Biosystems* **31**, 111–119.
4. Forterre, P., Benachenhou-Lahfa, N., Confalonieri, F., Duguet, M., Elie, C. & Labedan, B. (1993) *BioSystems* **28**, 15–32.
5. Olsen, G. J., Woese, C. R. & Overbeek, R. (1994) *J. Bacteriol.* **176**, 1–6.
6. Brown, J. R. & Doolittle, W. F. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 2441–2445.
7. Hori, H. & Osawa, S. (1987) *Mol. Biol. Evol.* **4**, 445–472.
8. Klenk, H.-P., Palm, P. & Zillig, W. (1994) *Syst. Appl. Microbiol.* **16**, 638–647.
9. Creti, R., Ceccarelli, E., Bocchetta, M., Sanangelantoni, A. M., Tiboni, O., Palm, P. & Cammarano, P. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 3255–3259.
10. Lake, J. A., Henderson, E., Clark, M. W., Scheinman, A. & Oakes, M. I. (1986) *Syst. Appl. Microbiol.* **7**, 131–136.
11. Lake, J. A. (1988) *Nature (London)* **331**, 184–186.
12. Gouy, M. & Li, W.-H. (1989) *Nature (London)* **339**, 145–147.
13. Stoffler, G. & Stoffler-Mailicke, M. (1986) *Syst. Appl. Microbiol.* **7**, 123–130.
14. Rivera, M. C. & Lake, J. A. (1992) *Science* **257**, 74–76.
15. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12**, 387–395.
16. Dayhoff, M. O., Eck, R. V. & Park, D. M. (1972) *Atlas of Protein Sequence and Structure* (Nat. Biomed. Res. Found., Silver Spring, MD).
17. Czworkowski, J., Wang, J., Steitz, T. A. & Moore, P. B. (1994) *EMBO J.* **13**, 3661–3668.
18. Clark, B. F. C., Kjeldgaard, M., la Cour, T. F. M., Thirup, S. & Nyborg, J. (1990) *Biochim. Biophys. Acta* **1050**, 203–208.
19. Ævarsson, R. (1995) *J. Mol. Evol.* **41**, 1096–1104.
20. Swofford, D. L. (1993) PAUP: Phylogenetic Analysis Using Parsimony (Illinois Natural History Survey, Champaign, IL), Version 3.1.
21. Maddison, W. P. (1993) *Syst. Biol.* **42**, 576–581.
22. Felsenstein, J. (1993) PHYLIP: Phylogeny Inference Package (Department of Genetics, Univ. of Washington, Seattle), Version 3.5c.
23. Kumar, S., Tamura, K. & Nei, M. (1993) MEGA: Molecular Evolutionary Genetics Analysis (The Pennsylvania State University, University Park, PA), Version 1.0.
24. Nei, M., Chakraborty, R. & Fuerst, P. A. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 4164–4168.
25. Adachi, J. & Hasegawa, M. (1992) MOLPHY: Programs for Molecular Phylogenetics—PROTML: Maximum Likelihood Inference of Protein Phylogeny (Institute of Statistical Mathematics, Tokyo), Vol. 1.
26. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comput. Appl. Biosci.* **8**, 275–282.
27. Yang, Z. (1995) PAML: Phylogenetic Analysis by Maximum Likelihood (Institute of Molecular Evolution and Genetics, The Pennsylvania State University, University Park, PA), Version 1.1.
28. Bourne, H. R., Sanders, D. A. & McCormick, F. (1991) *Nature (London)* **349**, 117–127.
29. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22**, 521–565.
30. Matheson, A. T., Auer, J., Ramirez, C. & Bock, A. (1990) in *The Ribosome Structure, Function and Evolution*, ed. Hill, W. E. (Am. Soc. Microbiol., Washington, DC), pp. 617–633.
31. Baldauf, S. L. (1990) Ph. D. thesis (Univ. of Michigan, Ann Arbor, MI).
32. Iwabe, N., Kuma, K.-i., Kishino, H., Hasegawa, M. & Miyata, T. (1991) *J. Mol. Evol.* **32**, 70–78.
33. Klenk, H.-P., Baldauf, S. L., Keeling, P. J., Doolittle, W. F. & Zillig, W. (in press) *Syst. Appl. Microbiol.*

34. Clark, A. J. & Sandler, S. J. (1994) *Crit. Rev. Microbiol.* **20,** 125–142.
35. Baumann, P., Qureshi, S. A. & Jackson, S. P. (1995) *Trends Genet.* **11,** 279–283.
36. Kaine, B. P. (1989) *J. Mol. Evol.* **23,** 248–254.
37. Langer, D. & Zillig, W. (1993) *Nucleic Acids Res.* **21,** 2251.
38. Reiter, W. D., Hudepohl, U. & Zillig, W. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 9509–9513.
39. Potter, S., Durovic, P. & Dennis, P. P. (1995) *Science* **268,** 1056–1060.
40. Forterre, P., Bergerat, A., Gadelle, D., Elie, C., Gottspeich, F., Confalonieri, F., Duguet, M., Holmes, M. & Dyall-Smith, M. (1994) *Syst. Appl. Microbiol.* **16,** 746–758.
41. Leffers, H., Kjems, J., Ostergaard, L., Larson, N. & Garrett, R. A. (1987) *J. Mol. Biol.* **195,** 43–61.
42. Brown, J. R., Masuchi, Y., Robb, F. T. & Doolittle, W. F. (1994) *J. Mol. Evol.* **38,** 566–576.
43. Benachenhou-Lahfa, N., Forterre, P. & Labedan, B. (1993) *J. Mol. Evol.* **36,** 335–346.
44. Gupta, R. S. & Golding G. B. (1993) *J. Mol. Evol.* **37,** 573–582.
45. Doolittle, R. F., Feng, D. F., Anderson, K. L. & Alberro, M. R. (1990) *J. Mol. Evol.* **31,** 383–388.
46. Chien, Y.-T. & Zinder, S. H. (1994) *J. Bacteriol.* **176,** 6590–6598.
47. Martin, W., Brinkmann, H., Savonna, C. & Cerff, R. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 8692–8696.
48. Kuhner, M. K. & Felsenstein, J. (1994) *Mol. Biol. Evol.* **11,** 459–468.
49. Tateno, Y., Takezaki, N. & Nei, M. (1994) *Mol. Biol. Evol.* **11,** 261–277.
50. Felsenstein, J. (1978) *Syst. Zool.* **27,** 401–409.
51. Barnes, S., Delwiche, C. F., Palmer, J. D. & Pace, N. R. (1996) *Proc. Natl. Acad. Sci. USA*, in press.
52. De Rijk, P., Van de Peer, Y., Van den Broeck, I. & De Wachter, R. (1995) *J. Mol. Evol.* **41,** 366–375.
53. Gambacorta, A., Trincone, A., Nicolaus, B., Lama, L. & De Rosa, M. (1994) *Syst. Appl. Microbiol.* **16,** 518–527.
54. Wolters, J. & Erdmann, V. A. (1989) *Can. J. Microbiol.* **35,** 43–51.
55. Lake, J. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1455–1459.
56. Garrett, R. A., Dalgaard, J., Larsen, N., Kjems, J. & Mankin, A. S. (1991) *Trends Biochem. Sci.* **16,** 22–27.
57. Hashimoto, T., & Hasegawa, M. (1996) *Adv. Biophys.* **32,** 73–120.