Archaea Sister Group of Bacteria? Indications from Tree Reconstruction Artifacts in Ancient Phylogenies

Henner Brinkmann and Hervé Philippe

Phylogénie et Evolution Moléculaires (UPRES-A 8080 CNRS), Université Paris-Sud, Orsay, France

The 54-kDa signal recognition particle and the receptor SRα, two proteins involved in the cotranslational translocation of proteins, are paralogs. They originate from a gene duplication that occurred prior to the last universal common ancestor, allowing one to root the universal tree of life. Phylogenetic analysis using standard methods supports the generally accepted cluster of Archaea and Eucarya. However, a new method increasing the signal-to-noise ratio strongly suggests that this result is due to a long-branch attraction artifact, with the Bacteria evolving fastest. In fact, the Archaea/Eucarya sisterhood is recovered only by the fast-evolving positions. In contrast, the most slowly evolving positions, which are the most likely to retain the ancient phylogenetic signal, support the monophyly of prokaryotes. Such a eukaryotic rooting provides a simple explanation for the high similarity of Archaea and Bacteria observed in complete-genome analysis, and should prompt a reconsideration of current views on the origin of eukaryotes.

Introduction

The global picture emerging from the analysis of ribosomal RNA (rRNA) is that all extant life forms belong to one of three distinct groups, called domains (Woese, Kandler, and Wheelis 1990): Bacteria (B), Archaea (A), and Eucarya (E). Although a possible outgroup for the extant life forms cannot exist a priori, a universal tree of life can be rooted using paralogous relatives (ancient duplication) (Schwartz and Dayhoff 1978). The data sets used for this purpose have to fulfill the following conditions: (1) both copies must be present in all three domains, and (2) they must remain alignable without ambiguity. Up to now, only six proteins have been found to be useful for rooting the tree of life: the translation elongation factors EF-Tu/1α-EF-G/2 (Iwabe et al. 1989; Baldauf, Palmer, and Doolittle 1996), the V- and F-type ATPases (Gogarten et al. 1989), the two amino-acyl tRNA synthetase pairs Val/Ile and Trp/Tyr (Brown and Doolittle 1995; Brown et al. 1997), carbamoyl phosphate synthetase (CPS) (Lawson, Charlebois, and Dillon 1996), and signal recognition particle (SRP) proteins (Gribaldo and Cammarano 1998). Archaea and Eucarya are always each other's closest relatives, with Bacteria emerging first. This bacterial rooting is generally considered to be quite sound (Brown and Doolittle 1997; Olsen and Woese 1997).

Nevertheless, it is surprising that such ancient relationships as that between the three domains could be inferred by molecular phylogenies without major difficulty, whereas much more recent events, e.g., the position of amphioxus (Naylor and Brown 1997) or the monophyly of rodents (Philippe 1997), cannot be recovered despite the use of the complete mitochondrial genome (about 3,500 amino acids). In fact, all phylogenetic markers used to root the tree of life are highly saturated mutationally (Philippe and Forterre 1999). During the several billion years of the evolution of life, multiple substitutions would be expected. Thus, this high level of noise should have erased most, if not all, of the ancient phylogenetic signal. Therefore, the question arises: how could it be possible that such ancient relationships can be recovered? A potential answer (Philippe and Forterre 1999) is that the inferred topology, (B,(A, E)), is the result of a tree reconstruction artifact that is known as the long-branch attraction (LBA) phenomenon (Felsenstein 1978). This artifact is known to seriously influence numerous molecular phylogenies (for review see Philippe and Laurent 1998). Moreover, the efficiency of current tree reconstruction methods strongly relies on an appropriate model of sequence evolution (Lockhart et al. 1996; Sullivan and Swofford 1997). Since this condition is not fulfilled, the reliability of ancient phylogenies remains limited (Philippe and Laurent 1998). The rooting of the universal tree of life is thus far from being firmly established, and a more reliable answer should come from technical improvements and/or from the analysis of additional ancient duplicated genes.

The signal recognition particle (SRP) plays an important role in cotranslational targeting to the rough endoplasmic reticulum by (1) binding the signal sequences of the nascent polypeptide, (2) pausing the protein biosynthesis, and (3) docking the complex, ribosome, nascent polypeptide, and SRP to the heterodimeric SRP receptor (SRα, SRβ, docking protein) (Rapoport, Rolls, and Jungnickel 1996). The two subunits SRP54 and SRα share a homologous region of about 300 amino acids, corresponding to a GTPase domain (Valent et al. 1995). Bacteria and Archaea also possess an SRP system, although it is less complex. They contain at least an SRP-RNA, as well as homologs of the two components SRP54 and SRα, named, in case of the Bacteria, *ffh* (fifty-four homolog) and *fts*Y (essential for cell division), respectively (Pohlschroder et al. 1997). Bacteria possess an additional gene, named *flh*F, located in a large motility operon (Ge and Charon 1997). The FlhF protein is essential for the biogenesis of the flagella in *Bacillus subtilis* (Carpenter, Hanlon, and Ordal 1992). Neverthe-

Key words: Archaea, Bacteria, eukaryotes, molecular phylogeny, long-branch attraction artifact, signal recognition particle.

Address for correspondence and reprints: Hervé Philippe, Phylogénie et Evolution Moléculaires, Bâtiment 444, Université Paris-Sud, 91405 Orsay cedex, France. E-mail: herve.philippe@bc4.u-psud.fr.

Mol. Biol. Evol. 16(6):817–825. 1999

less, the *flh*F gene is not present in all Bacteria, e.g., it has no equivalent in the genome of *Escherichia coli.*

In this paper, we present a phylogenetic analysis of the SRP gene family. Our results confirm that the duplication leading to the SRP54 and SRα genes occurred before the last universal common ancestor (LUCA). Trees based on SRP54 using the classical reconstruction methods provided robust support for the sister group relationship between Archaea and Eucarya. However, we have developed a new method that increases the signal-to-noise ratio for ancient events. The evolution of the phylogenetic signal, starting with the most conserved positions, was studied while adding more and more fast-evolving characters. Using the slowly evolving positions, a eukaryotic rooting was found, whereas the bacterial rooting only appeared once the faster-evolving characters, i.e., those containing much more noise due to multiple substitutions, were added. In consequence, the (B,(A,E)) topology is very likely the result of an LBA artifact, and the correct topology is probably (E,(A,B)) for SRP54.

Materials and Methods

Data Set

All sequences homologous to the SRP gene family were identified by a BLAST search (blast@ncbi.nlm.nih.gov) and were automatically retrieved with the programs blast2retp and retp2ali (P. Lopez, personal communication). The alignment of these sequences was first carried out with the program CLUSTAL W (Thompson, Higgins, and Gibson 1994) and visually refined with the help of the ED program of the MUST package, version 1.0 (Philippe 1993). Partial or redundant sequences, as well as those with sequencing errors, were discarded. Genome projects provided numerous additional sequences that have been manually retrieved at the following web sites: www.tigr.org (*Streptococcus pneumoniae* and *Thermotoga maritima*), www.pandora.cric.com *(Clostridium acetobutylicum*), www.pseudomonas.com *(Pseudomonas aeruginosa*) and www.sanger.ac.uk (*Streptomyces coelicolor*). Since for the receptor gene only four complete eukaryotic sequences (one mammal, one nematode, and two fungi) were found, a nearly full length artificial plant sequence was created by concatenating cDNAs (expressed sequence tag, EST) from rice and cabbage, the sequence was named *Oryza*/*Brassica*.

A preliminary phylogenetic analysis of the complete data set (92 sequences) demonstrated that the bacterial sequences clearly separated into three monophyletic groups (Ffh, FtsY, and FlhF). To reduce computing time and to have a similar taxonomic sampling for each domain (about six sequences), 40 sequences were chosen. Reducing the number of species used can appear counterintuitive when dealing with a phylogenetic question for which LBA is suspected. Indeed, using a large number of species may help to reduce the LBA artifact (Hendy and Penny 1989; Graybeal 1998), but adding fast-evolving lineages can increase this artifact (Kim 1996). We therefore selected the slowest-evolving species within well-sampled groups. Moreover, the addition of a taxon is more helpful if it breaks the long branch (Hendy and Penny 1989; Graybeal 1998), and, unfortunately, the discarded sequences are mainly bacterial ones and do not allow us to break the long branches at the bases of Bacteria and Eucarya. Finally, since only Bacteria are well sampled, using all the available species in the S-F method overweighs this group in the estimation of variability, and we therefore chose a similar taxonomic sampling for the seven groups (about six species). Only 187 unambiguously alignable amino acid positions were used.

Phylogenetic Analysis

Phylogenetic trees were constructed with maximum-likelihood (ML), maximum-parsimony (MP), and distance-based methods (neighbor joining, NJ) with the programs PROTML (Adachi and Hasegawa 1996) version 2.3, PAUP (Swofford 1993) version 3.1.1, and NJ in the MUST package (Philippe 1993), version 1.0, respectively. The distances were computed with the substitution model of Kimura (1983). MP trees were obtained by 100 random-addition heuristic search replicates, and ML trees were obtained by the quick-add OTUs search with the JTT model of amino acid substitution, retaining the 5,000 top-ranking trees (options -jf -q -n 5000). Bootstrap proportions (Felsenstein 1985) were calculated by the analysis of 1,000 replicates for MP and NJ. For ML, bootstrap proportions were computed by using the RELL method (Kishino, Miyata, and Hasegawa 1990) because of computing time limitations. The saturation level was estimated by comparing the number of observed differences with the number of substitutions inferred by either MP or PROTML (Philippe et al. 1994), using the programs TREEPLOT and COMP_MAT of the MUST package (Philippe 1993).

Rate Variation Within and Among Sites

Among-site rate variation can strongly influence tree reconstruction methods (Yang 1996). But a further complication is possible: according to the covarion model (Fitch and Markowitz 1970), the evolutionary rate of a given position can be different in different taxonomic groups. Testing the homogeneity of the evolutionary rate of each position in the seven taxonomic groups (three domains for the two genes plus the bacterial FlhF) using the method of Lopez, Forterre, and Philippe (1999) showed few positions to be significantly heterogeneous. Thus, classical methods dealing with among-site rate variation could be used. The α parameter of the gamma distribution was estimated with the programs PAML (Yang 1997) and GZ-gamma (Gu and Zhang 1997).

The S-F Method

A simple method to deal with among-site rate variation was derived from the H-P method (Lopez, Forterre, and Philippe 1999). Aligned sequences were divided into seven groups, each containing about six species: three domains for the two genes plus the bacterial FlhF. Using PAUP, the number of changes per position

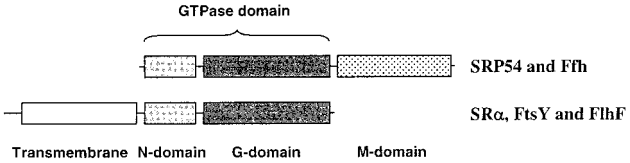

FIG. 1.—Schematic presentation of the SRP54 and SRα proteins.

within each group was calculated. The evolutionary rate of a given position was estimated as the sum of the numbers of steps for this position within the seven groups. A submatrix S_n (S = slowly evolving) contained all of the positions for which the total number of steps was below *n*. A complementary matrix F_n (F = fast-evolving) contained all of the positions for which the number of steps was above *n*. Thirteen matrices $S_0, \ldots, S_n, \ldots, S_{12}$ were constructed, allowing us to study the evolution of the phylogenetic inference as more and more fast-evolving characters were added. A simple quantification of this evolution was obtained by calculating the Bremer Decay Index (BDI) (Bremer 1988). The BDI of a group is equal to the difference of the number of steps of the MP tree in which this group is not monophyletic and the number of steps of the MP tree in which this group is monophyletic. The BDI was calculated with PAUP (10 random additions of taxa and the TBR branch-swapping option).

The S-F method differs from the H-P method in the way of selecting slow-evolving positions. In the S-F method, a position is selected depending on its evolutionary rate, evaluated as the sum of the number of steps within predefined groups. In the H-P method, all positions are selected, but character states are replaced by question marks within groups that display too many substitutions (see Lopez, Forterre, and Philippe [1999] for a detailed description).

Results and Discussion

Phylogenetic Analysis of the SRP-Type GTPase Domain

The SRP54 and SRα proteins have almost the same length (about 500 amino acids), but in SRP54, the GTPase domain is located N-terminal, whereas it is C-terminal in SRα (fig. 1). The C-terminal region of SRP54, called the M-domain, binds the mRNA and the signal sequence (Oh et al. 1996) and, although well conserved, it has no counterpart in SRα. The N-terminal domain of SRα is much more variable, with membrane anchoring being the only function described (Zelazny et al. 1997), and also has no equivalent in SRP54. The homologous region between the two paralogous proteins includes essentially the N domain (about 80 amino acids), which probably has a regulatory function, and the G domain (about 200 amino acids), with the GTPase activity. Only the more conserved portion (G domain) of the GTPase region was used, in order to reduce ambiguity in the alignment, leading to 187 positions. In contrast to the elongation factors (Baldauf, Palmer, and Doolittle 1996), the use of the 3D structure (Freymann et al. 1997; Montoya et al. 1997) did not improve the alignment of the G domain.

The ML tree based on the complete alignment (fig. 2) showed three discrete groups, corresponding to SRP54 (upper part), FlhF (center), and SRα (lower part). The duplication that led to SRP54 and SRα very likely occurred before the separation of the three domains. For both genes, the usual Archaea/Eucarya sister group relationship was recovered, with the Bacteria emerging first. The statistical support for these two nodes was rather high, with bootstrap values ranging from 63% to 100%. For both genes, the Archaea were paraphyletic, although supported only by low bootstrap values. In contrast, the monophylies of both Bacteria and Eucarya were always supported by bootstrap values close to 100%. Similar results have recently been obtained by Gribaldo and Cammarano (1998), with the notable difference that the FlhF sequences were not used.

The branch of the bacterial FlhF group was clearly the longest, showing its very high evolutionary rate. The position of the FlhF sequences depended on the tree reconstruction method used and remained uncertain. In contrast to the ML analysis, two very highly conserved positions (numbers 119 and 277 in the human SRP54) suggested a specific relationship between FlhF and both archaeal and eukaryotic SRαs to the exclusion of the functional equivalent, the bacterial receptor FtsY. In contrast, no conserved position supported an FtsY/SRα relationship. The bacterial ortholog of the archaeal and eukaryotic receptor SRαs could not be determined with confidence, and it cannot be excluded that FlhF was originally the bacterial receptor and was later functionally replaced in Bacteria by FtsY after a gene duplication. This ambiguity precluded a rooting of the tree of life based on the receptor sequences.

Intradomain phylogenies were generally correctly recovered, especially for closely related organisms. For example, the monophylies of fungi, Metazoa, crenotes, and high-G+C gram-positive bacteria were obtained, as was the sister group relationship between chloroplast and cyanobacterial SRP54 sequences (fig. 2). However, some incorrect results were also noted, e.g., paraphyly of euryotes for SRP54, whereas they were monophyletic for SRα. Several bacterial groups known to be monophyletic were also not recovered when the complete data set (92 sequences) was used, although the Bacteria remained clearly monophyletic (data not shown). The widely accepted sister group relationship of fungi and Metazoa (Cavalier-Smith 1993) was recovered with the SRP54 sequences, albeit with low statistical support. In contrast, in the case of the SRα sequences, Metazoa and plants formed a robust group (bootstrap value of 100%) to the exclusion of fungi.

This last result is likely due to an LBA artifact (fig. 3*A*). First, SRα evolved faster than SRP54 (compare the branch lengths within fungi and Metazoa in fig. 2). Second, for both genes, the fungi evolved faster, as demonstrated by their long branches in figure 2 and, more significantly, by the fact that the distances between two ascomycetes were similar to that between Metazoa and plants. The fast-evolving fungal SRα sequences

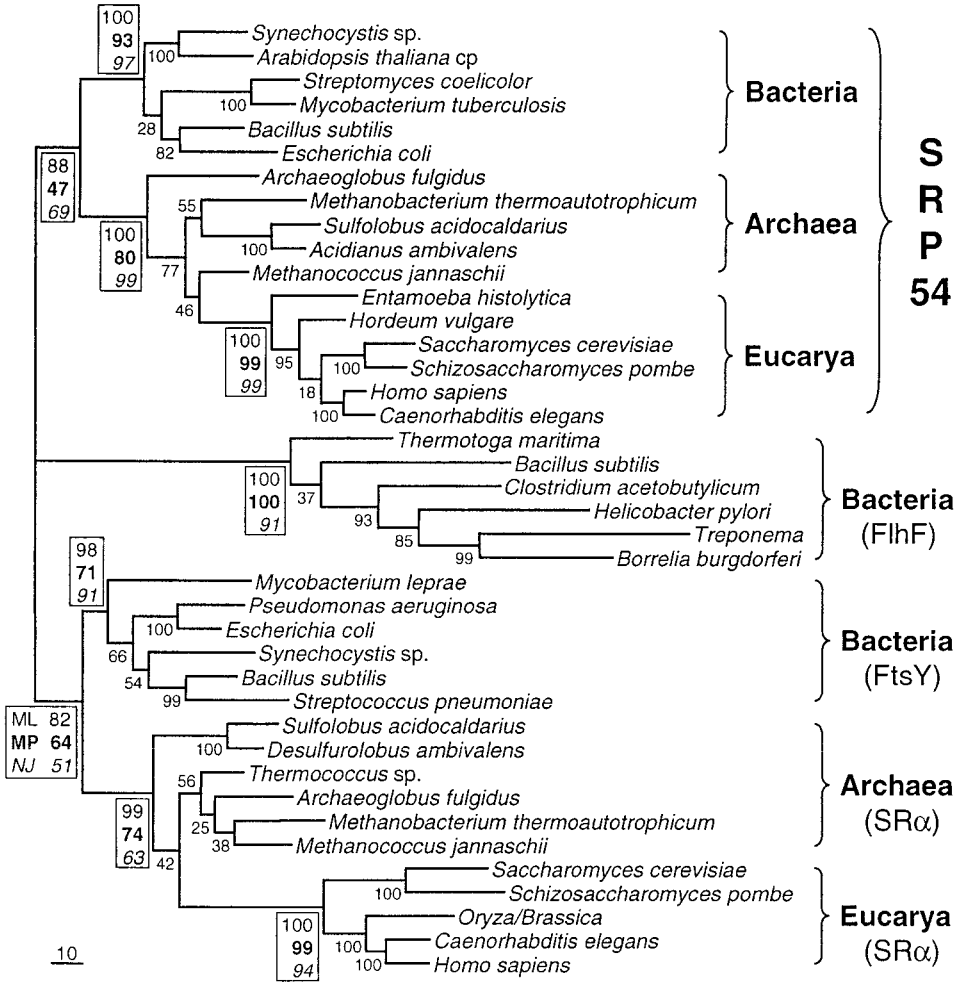

FIG. 2.—Phylogenetic tree based on the complete data set (40 sequences and 187 positions) of SRP-type GTPase-domain homologs. The tree was reconstructed using an ML method. The ML bootstrap values are shown for all nodes, and the bootstrap values of MP and NJ analysis are only displayed for deep nodes. The scale bar corresponds to 10 substitutions for 100 amino acids.

emerged first because they were attracted by the long branch of the outgroup (fig. 3*A*). ML was the only method able to correctly locate fungi and Metazoa in the case of the slowly evolving SRP54 but not in the case of the fast evolving SRα sequences. Because differences in evolutionary rates can mislead the ML method in the reconstruction of such recent events, the sister group relationship between Archaea and Eucarya found for SRP54 may also be the result of an LBA artifact (fig. 3*B*), similar to the one observed between plants and Metazoa in the case of SRα.

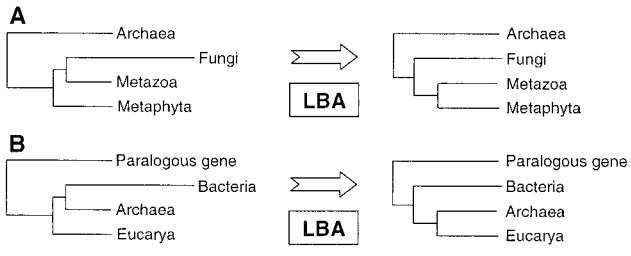

FIG. 3.—Schematic illustration of an LBA artifact.

Saturation and a New Method to Increase the Signal-to-Noise Ratio

If most of the variable positions in the SRP54/SRα genes have undergone multiple substitutions since the last common ancestor, the signal-to-noise ratio should be low, and a tree reconstruction artifact becomes very likely. In order to test this assumption, the level of mutational saturation was studied. Most of the pairwise comparisons were part of a large plateau (fig. 4, top), where about 100–150 observed differences corresponded to about 150–450 inferred substitutions, indicating that for several species pairs there were at least up to three times more inferred substitutions than observed differences.

The high level of saturation, and thus a possible tree reconstruction artifact, is essentially due to fast-evolving positions, and only slowly evolving positions may contain reliable information for rooting the tree of life (Forterre et al. 1992). Methods that take into account rate variations among sites could thus improve the reliability of tree reconstruction. Unfortunately, the estimation of the α parameter of the gamma law varied from

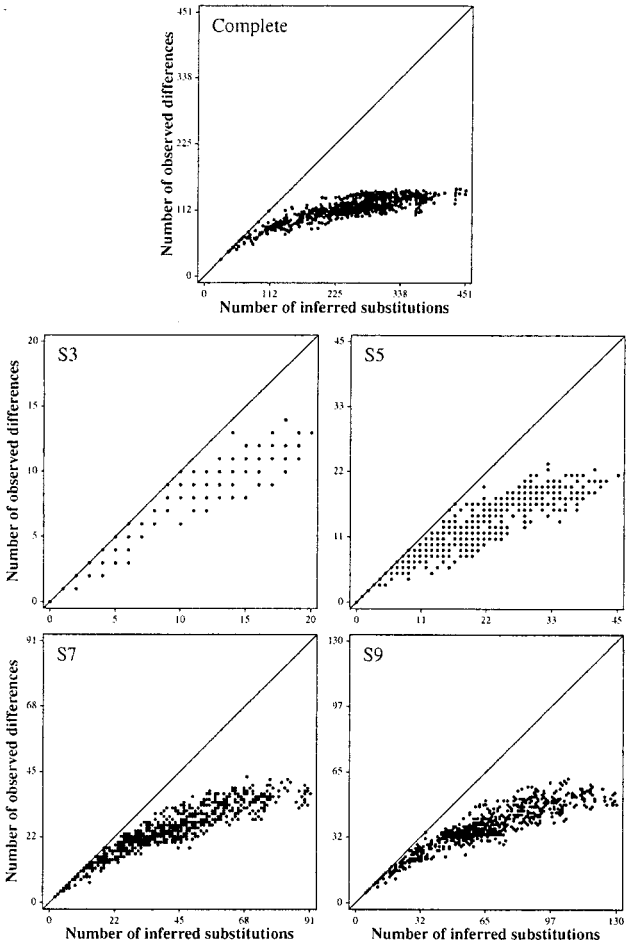

FIG. 4.—Estimations of the mutational saturation of the complete SRP-type GTPase-domain data set and of four S matrices (S_3, S_5, S_7, S_9). The Y-axis shows numbers of observed differences for pairwise comparisons. The X-axis shows estimations of the number of substitutions for the same pairs by the MP method. Ideal data points corresponding to completely unsaturated comparisons would all be on the diagonal.

0.5 to 20 depending on the method (Gu and Zhang 1997; Yang 1997) and/or the set of sequences used, thus excluding the use of standard methods.

Lopez, Forterre, and Philippe (1999) have developed a method, called H-P, that generates less-saturated data sets. This method is especially designed to handle the covarion model of evolution (Fitch and Markowitz 1970), whereby the evolutionary rate of a given position is heterogeneous, i.e., differs in different taxonomic groups. However, only 2 out of 187 positions were found to be significantly heterogeneous, considerably less than in the case of EF-1α (70 out of 158) (Lopez, Forterre, and Philippe 1999). The H-P method was thus not suitable for the SRP phylogeny, but could be easily adapted to homogeneous data sets.

The principle of the new method is to select positions that minimize the lengths of the terminal branches in order to increase the signal-to-noise ratio for the internal branches. The evolutionary rate of a given position was estimated as the sum of the number of steps within the seven monophyletic groups (SRP54 and SRα for three domains plus FlhF). This estimation depends only on the intradomain phylogenies and is totally independent of the interdomain relationships. The circularity of the classical successive-weighting approach (Farris 1969) is thus avoided, and this constitutes a major advantage of the new method. Thirteen data sets, named S_0, S_1, . . . , S_{12}, were generated by selecting only positions for which the total number of steps was below a given threshold. As expected, the level of saturation increased with the threshold, as shown for matrices S_3, S_5, S_7, and S_9 (fig. 4). The matrix S3 was close to the ideal case in which all points are on the diagonal, corresponding to a data set for which, at each position, no more than one substitution occurred. Under this condition, all tree reconstruction methods will infer the correct topology (Swofford et al. 1996). For matrices S_7 and S_9, the presence of a plateau indicated a high level of saturation. Unfortunately, the S_3 matrix, which had a very good signal-to-noise ratio, contained only a few positions (30) and thus a rather limited signal, since the length of the MP tree was only 44 steps. In contrast, the MP tree of the much more saturated matrix S_9 had considerably more steps (506 steps for 86 positions), corresponding to more information but also to much more noise. Because the addition of fast-evolving positions can mask the phylogenetic signal, we studied its evolution via an MP reconstruction using the BDIs of selected nodes for the data sets with increasing variability.

Phylogenetic Inference and Evolutionary Rate per Site

Since we were interested in the relationship between the three domains, the BDIs of the three possible groups, A-B, A-E, and B-E, for the SRP54 gene were studied and compared with the BDIs of undisputedly monophyletic groups, i.e., FlhF and the eukaryotic SRP54 and SRα (fig. 5). The support for these three latter groups increased continuously with the threshold (fig. 5*A*). This could be due to the fact that a phylogenetic signal was added. It is possible that the LBA artifact also favors these groupings, because their monophylies are favored by the attraction of all the other taxa, which clearly display long branches (Siddall 1998). Therefore, the increase in BDI (fig. 5*A*) for indisputable ancient nodes could be due, for slowly evolving positions, to the addition of phylogenetic signal and, for fast-evolving positions, to the addition of noise that by chance supports their monophyly. Accordingly, the amount of support for the three deep nodes did not rise proportionally to the total number of steps (fig. 5*B*). In fact, the ratio between the sum of the BDIs for the three nodes and the total number of steps decreased quite significantly (fig. 5*C*). For example, for thresholds of 1, 6, and 12 (11, 196, and 877 total steps, respectively), the BDIs for FlhF were 3, 9, and 16, respectively. In conclusion, the fraction of the total number of steps supporting this monophyletic group was very high in the case of S_1 (27%) and then dramatically dropped (5% for S_6) to 2% for S_{12}. In agreement with the saturation analysis, the addition of positions with more than about six changes mainly represented the addition of noise as far as deeply branching nodes were concerned.

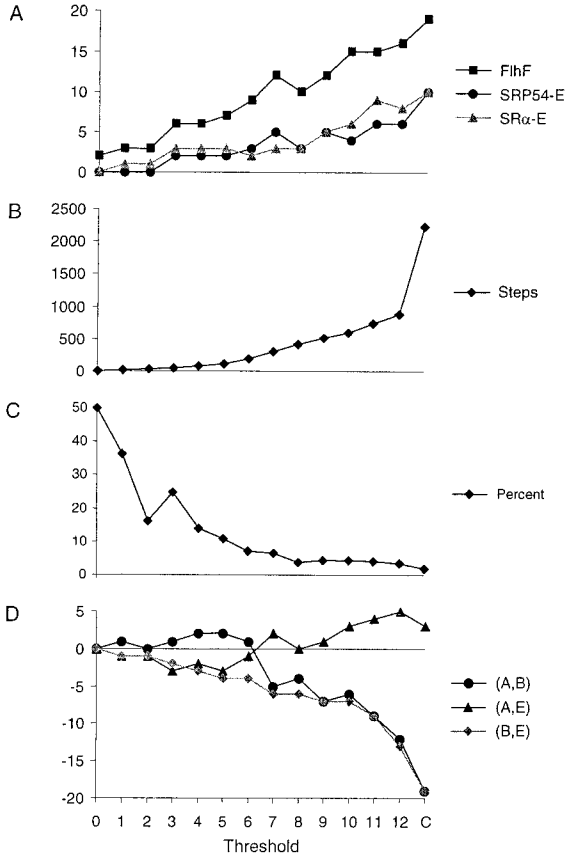

FIG. 5.—Evolution of the BDI. In all four cases, the numbers 0–12 on the X-axis correspond to the matrices S_0–S_{12}, whereas *C* represents the complete data set. *A* and *D*, The Y-axis corresponds to the number of steps that support monophyly (positive values) or nonmonophyly (negative values) of a given group. *B*, The numbers of steps of the most-parsimonious trees for the different data sets are shown. *C*, The percentages of the BDI steps for FlhF, SRP54-E, and SRα-E compared with the total number of steps are shown.

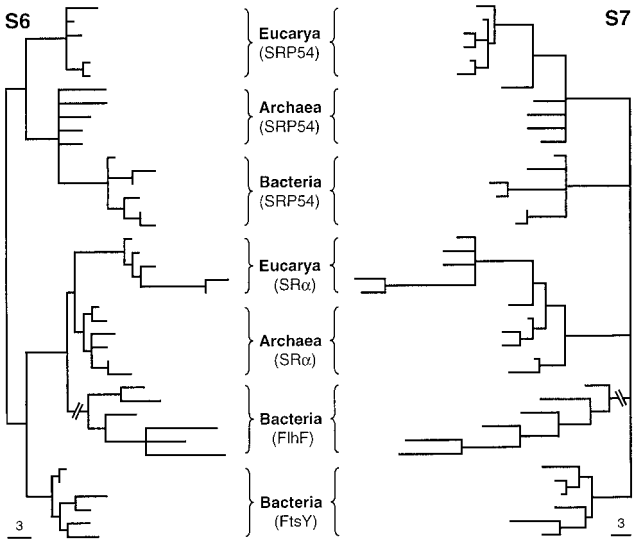

FIG. 6.—Strict consensus of MP trees based on the matrices S_6 (left) and S_7 (right), encompassing the transition. The names of the organisms were omitted for simplification. For space reasons, the branch at the base of FlhF has been shortened and corresponds to 19 (22) substitutions for the S_6 (S7) matrix. The scale bar corresponds to three steps.

A B-E monophyly was never supported by the SRP54 sequences. Corresponding BDI values were always negative and continuously decreased with the addition of further characters (fig. 5*D*). Interestingly, the support for A-E and A-B monophylies showed a more complex evolution. In the most slowly evolving matrices (S_0–S_6), the A-E monophyly was rejected and an A-B monophyly was supported. In the fast-evolving matrices(S_7–S_{12}), the reverse situation arose, with constantly rising positive support in favor of the A-E sister group relationship and more and more support against the A-B sister group. The topology recovered for the complete data set (fig. 1) was thus due to the influence of noisy positions.

The transition between the matrices S_6 and S_7 required further examination. If an A-E relationship is correct, one would expect the 13 positions with exactly seven changes to provide strong support in its favor. However, these 13 positions did not support the monophyly of A-E (BDI = 0) and, indeed, contained almost no ancient signal, since the BDIs for SRP54, FlhF, bacterial and eukaryotic SRP54, and SRα were 0, −1, 0, 1, −1, and −1, respectively. In fact, these positions only contained a phylogenetic signal for much more recent events, e.g., the monophyly of fungi and of Metazoa for SRα. Although containing very limited ancient phylogenetic signal, these fast-evolving positions rejected an A-B monophyly (BDI = −2). The transition was thus the result of the rejection of an A-B monophyly rather than support for the A-E monophyly.

The Transition Is Due to an LBA Artifact

Since the A-E sister group relationship appeared only after the addition of fast-evolving positions, a tree reconstruction artifact was likely responsible for it. The two strict consensuses of the MP trees based on the matrices S_6 (fig. 6, left; 53 positions, 1,728 trees with 196 steps) and S_7 (fig. 6, right; 66 positions, 88 trees with 304 steps) have, as expected, strikingly different topologies. First, let us describe the SRP54 part of the tree. In both cases, Archaea were paraphyletic, but they were either close to Bacteria (S_6 tree) or close to Eucarya (S_7 tree). In the S_6 tree, eukaryotic sequences were the most slowly evolving, whereas Bacteria evolved the fastest, i.e., about two times faster than Archaea and three times faster than Eucarya. In contrast, in the S_7 tree, the eukaryotic sequences were the fastest-evolving ones, with a branch length that was about 1.5 times longer than for Bacteria and Archaea, both evolving at similar rates. The fast evolutionary rate of the bacterial SRP54 sequences (S_6 tree) led to their attraction by the long branch of the outgroup when noisy data were added (S_7 tree). An LBA phenomenon (fig. 3*B*) is the best explanation for the transition between the two topologies. In agreement with this, the other parts of the tree were more doubtful for the matrix S_7: the long branch of FlhF emerged earlier, and the monophyly of the SRP54 was not recovered.

All of this confirmed that the matrix S_7 contains less-reliable phylogenetic signal than the matrix S_6.

This LBA artifact did not appear for the very slowly evolving positions, because very few multiple substitutions occurred (Swofford et al. 1996). The fast-evolving positions, in contrast, strongly rejected both the A-B and the B-E relationships in a similar way, since their BDIs varied from −6 to −12 in the matrices S_7–S_{12}, respectively. Despite the strong rejections of A-B and B-E, these positions did not support a sister group relationship of Archaea and Eucarya, since the BDI for A-E always remained close to 0 (between −1 and 2). Theoretically, the fast-evolving positions (more than seven changes), which were not expected to contain any information for deep nodes, should display BDIs close to 0 for A-B, A-E, and B-E. Thus, the strong rejection of the A-B and B-E groups was due to the higher evolutionary rate of Bacteria, which moved this group away from the others. This explanation was consistent with the fact that the BDIs for A-B and B-E were almost identical for the matrices S_9–S_{12}, whereas they were quite different for matrices S_0–S_6 (fig. 5*D*).

The present analysis suggested that the prokaryotes are monophyletic and that this monophyly is not usually recovered because of the LBA phenomenon. A paradox became apparent when looking at the branch lengths on figure 2, because the Bacteria, which probably evolved the fastest, displayed branch lengths similar to those of Archaea and even shorter than those of Eucarya. However, trees inferred from highly saturated data tend to show similar branch lengths, thus shortening the longest branches (Philippe and Laurent 1998). This paradox is due to the inefficiency of tree reconstruction algorithms in detecting multiple substitutions and estimating the real branch lengths. Therefore, the high evolutionary rate of bacterial SRP54 sequences observed with the slowly evolving positions (fig. 6) was likely correct.

Why Should Bacteria Evolve Much Faster?

The analyses of the SRP54/SRα pair discussed here and of the elongation factors (Lopez, Forterre, and Philippe 1999) suggest that these genes evolve significantly faster in Bacteria. This could also be true for numerous genes of the transcriptional and translational apparatus (unpublished observations). The higher similarity between Archaea and Eucarya observed in these genes (Brown and Doolittle 1997) is presumably the result of the maintenance of the ancestral state and can therefore not be used as a proof for a sister group relationship between Archaea and Eucarya. The global acceleration of the translational machinery of Bacteria as compared with that of Archaea is also reflected in the branch lengths of the rRNA tree.

As first noticed by Dickerson (1971), the physical interactions between a protein and other cellular constituents (proteins or nucleic acids) constrain its evolution. If one or more of these interactions disappear, the corresponding constraints will be removed and the sequences will evolve faster. Therefore, we suggest that the reason for the acceleration of some constituents in Bacteria lies in the much simpler structure of their transcriptional apparatus (especially RNA polymerases) and, to a lesser extent, also of their translational apparatus. If the eukaryotic rooting is correct, characters shared between Eucarya and Archaea, particularly the complex structure of the RNA-polymerase and the TATA-box-binding protein as the central regulatory element of transcription, are ancestral. Bacteria have thus undergone a serious simplification of their transcriptional apparatus, leading to an acceleration of the evolutionary rate of the remaining proteins that had functional constraints removed. An even further simplification is observed today in mitochondria (Cermakian et al. 1997), in which a nuclear-encoded single-subunit enzyme related to the bacteriophage T7/T3 RNA polymerase has replaced the ancestral bacterial multisubunit polymerase. As expected, all the genes of the mitochondrial translational apparatus, especially the rRNA, evolve much faster than their bacterial counterparts (Gray and Spencer 1996), because these organelles have lost their original complexity through an ongoing streamlining. Interestingly, the nonhomologous replacement of the plastid polymerase by a copy of the mitochondrial T7/T3 RNA polymerase is underway in angiosperms (Gray and Lang 1998), indicating also a strong selective pressure toward simplification, at least in chloroplasts.

Eukaryotic Rooting and the Nature of the Last Universal Common Ancestor

The eukaryotic rooting of the tree of life has very important implications. It best explains the presence of much more bacterial-like than eukaryotic-like genes in Archaea (Koonin et al. 1997), which is difficult to explain by the current scenarios that assume a bacterial rooting and that involve lineage fusion or horizontal transfers. The explanation that these genes were simply inherited from the common prokaryotic ancestor of Archaea and Bacteria is quite straightforward. As discussed above, the higher similarity between Archaea and Eucarya that is observed mainly for proteins of the transcription and translation machinery is most likely due to an acceleration of their evolutionary rate in Bacteria.

A fascinating consequence of the eukaryotic rooting is the possibility that the LUCA (also named cenancestor) was a eukaryotic-like organism instead of a prokaryotic one, as is almost universally accepted today. However, LUCA must have been quite different from extant eukaryotes. In fact, all present eukaryotes are the descendants of a massive radiation that occurred after the mitochondrial endosymbiosis (Philippe and Adoutte 1998). The acquisition of mitochondria and the adaptation to oxygen have profoundly modified the former anaerobic eukaryotic organisms. The hypothesis that LUCA was a eukaryotic-like DNA-based organism is indeed very consistent with the existence of an RNA world, which is a widely accepted intermediate state in the evolution of cellular life. As thoroughly discussed (Jeffares, Poole, and Penny 1998; Poole, Jeffares, and Penny 1998), the numerous types of RNA sequences present in eukaryotes could have been directly inherited from a complex RNA-based organism. A prokaryotic rooting requires that these RNA sequences either have been lost independently in Bacteria and Archaea or

have been recently acquired in Eucarya, but these two hypotheses are less parsimonious. A eukaryotic LUCA also supports the intron-early theory (Gilbert, Marchionni, and McKnight 1986), which has been seriously challenged (Rzhetsky et al. 1997; but see also Long et al. 1998). An intensive exon shuffling could thus have occurred in a very early stage of organismal evolution, generating numerous universal proteins composed of several domains. For example, a copy of the SRP-type GTPase domain could have been fused via exon shuffling to the M-domain to yield the SRP54 gene and another copy to a transmembrane module to yield the SRα gene. An important implication of a eukaryotic LUCA is that prokaryotes derive from eukaryotes. A massive reduction of the genome of the common prokaryotic ancestor must have happened, probably as a result of selection for rapid reproduction (r selection; Pianka 1970). A new field of study could thus be the eukaryote-to-prokaryote transition (see the thermoreduction hypothesis; Forterre 1995).

Conclusions

The phylogenetic analysis of the SRP54 sequences is the first study of anciently duplicated genes that suggests a sister group relationship between Archaea and Bacteria, thus proposing the prokaryotes to be a natural group. The identity values between the paralogous genes of the SRP54/SRα pair are significantly higher than those for the other markers used for the rooting of the tree of life: 25% for ATPase, 30% for elongation factors, 30% for Ile-/Val-tRNA synthetase, 33% for CPS, and 38% for SRP54/SRα. The SRP54/SRα gene pair has the closest outgroup, thus reducing the LBA artifact, which is the prime obstacle in establishing very ancient events. However, the S-F method was necessary to increase the signal-to-noise ratio and thus to eschew the LBA artifact, since the A-E cluster is found with the complete data set (fig. 2), whereas the A-B clade is only recovered with the most slowly evolving positions in the case of SRP54 (fig. 6). A similar analysis of the elongation factors demonstrates that the fast-evolving positions provide the strongest support for the A-E grouping, suggesting its artifactual nature. Unfortunately, there is not enough remaining signal in the slow-evolving positions, yielding slight support for the artifactual A-E grouping rather than the A-B one (500 vs. 502 steps) (Lopez, Forterre, and Philippe 1999). The discrepancy between the two genes may be due to the fact that in the latter case, the outgroup is more distantly related, which enhances the LBA phenomenon. We are currently performing an analysis of all anciently duplicated genes using the S-F and H-P methods in an attempt to confirm or reject the eukaryotic rooting of the tree of life.

Acknowledgments

We thank Patrick Forterre, Jacqueline Laurent, Hervé Le Guyader, Philippe Lopez, William Martin, David Moreira, and Miklòs Müller for critical reading of the manuscript and helpful comments. We acknowledge Philippe Lopez for his help in statistical analyses and Karine Budin for the drawings.

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. Comput. Sci. Monogr. **28**:1–150.

BALDAUF, S. L., J. D. PALMER, and W. F. DOOLITTLE. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. Proc. Natl. Acad. Sci. USA **93**:7749–7754.

BREMER, K. 1988. The limits of amino acid sequences data in angiosperm phylogenetic reconstruction. Evolution **42**:795–803.

BROWN, J. R., and W. F. DOOLITTLE. 1995. Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc. Natl. Acad. Sci. USA **92**: 2441–2445.

———. 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. **61**:456–502.

BROWN, J. R., F. T. ROBB, R. WEISS, and W. F. DOOLITTLE. 1997. Evidence for the early divergence of tryptophanyl- and tyrosyl-tRNA synthetases. J. Mol. Evol. **45**:9–16.

CARPENTER, P. B., D. W. HANLON, and G. W. ORDAL. 1992. flhF, a Bacillus subtilis flagellar gene that encodes a putative Gtp-binding protein. Mol. Microbiol. **6**:2705–2713.

CAVALIER-SMITH, T. 1993. Kingdom Protozoa and its 18 phyla. Microbiol. Rev. **57**:953–994.

CERMAKIAN, N., T. M. IKEDA, P. MIRAMONTES, B. F. LANG, M. W. GRAY, and R. CEDERGREN. 1997. On the evolution of the single-subunit RNA polymerases. J. Mol. Evol. **45**:671–681.

DICKERSON, R. E. 1971. The structures of cytochrome c and the rates of molecular evolution. J. Mol. Evol. **1**:26–45.

FARRIS, J. 1969. A successive approximations approach to character weighting. Syst. Zool. **18**:374–385.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **40**:783–791.

FITCH, W. M., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

FORTERRE, P. 1995. Thermoreduction, a hypothesis for the origin of prokaryotes. C. R. Acad. Sci. III **318**:415–422.

FORTERRE, P., N. BENACHENHOU-LAHFA, F. CONFALONIERI, M. DUGUET, C. ELIE, and B. LABEDAN. 1992. The nature of the last universal ancestor and the root of the tree of life, still open questions. Biosystems **28**:15–32.

FREYMANN, D. M., R. J. KEENAN, R. M. STROUD, and P. WALTER. 1997. Structure of the conserved GTPase domain of the signal recognition particle. Nature **385**:361–364.

GE, Y., and N. W. CHARON. 1997. Identification of a large motility operon in Borrelia burgdorferi by semi-random PCR chromosome walking. Gene **189**:195–201.

GILBERT, W., M. MARCHIONNI, and G. MCKNIGHT. 1986. On the antiquity of introns. Cell **46**:151–153.

GOGARTEN, J. P., H. KIBAK, P. DITTRICH et al. (13 co-authors). 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. Proc. Natl. Acad. Sci. USA **86**:6661–6665.

GRAY, M. W., and B. F. LANG. 1998. Transcription in chloroplasts and mitochondria: a tale of two polymerases. Trends Microbiol. **6**:1–3.

GRAY, M., and D. SPENCER. 1996. Organellar evolution. Pp. 109–126 *in* D. ROBERTS, P. SHARP, G. ALDERSON, and M. COLLINS, eds. Evolution of microbial life. Cambridge University Press, Cambridge, England.

GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. **47**:9–17.

GRIBALDO, S., and P. CAMMARANO. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J. Mol. Evol. **47**:508–516.

GU, X., and J. ZHANG. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106–1113.

HENDY, M., and D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. **38**:297–309.

IWABE, N., K. KUMA, M. HASEGAWA, S. OSAWA, and T. MIYATA. 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86**:9355–9359.

JEFFARES, D. C., A. M. POOLE, and D. PENNY. 1998. Relics from the RNA world. J. Mol. Evol. **46**:18–36.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. **45**:363–374.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KISHINO, H., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny, and the origin of chloroplats. J. Mol. Evol. **31**:151–160.

KOONIN, E. V., A. R. MUSHEGIAN, M. Y. GALPERIN, and D. R. WALKER. 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. Mol. Microbiol. **25**:619–637.

LAWSON, F. S., R. L. CHARLEBOIS, and J. A. DILLON. 1996. Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. Mol. Biol. Evol. **13**:970–977.

LOCKHART, P. J., A. W. LARKUM, M. STEEL, P. J. WADDELL, and D. PENNY. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA **93**:1930–1934.

LONG, M., S. J. DE SOUZA, C. ROSENBERG, and W. GILBERT. 1998. Relationship between ''proto-splice sites'' and intron phases: evidence from dicodon analysis. Proc. Natl. Acad. Sci. USA **95**:219–223.

LOPEZ, P., P. FORTERRE, and H. PHILIPPE. 1999. A method for extracting ancient phylogenetic signal: the rooting of the universal tree of life based on elongation factors. J. Mol. Evol. (in press).

MONTOYA, G., C. SVENSSON, J. LUIRINK, and I. SINNING. 1997. Expression, crystallization and preliminary X-ray diffraction study of FtsY, the docking protein of the signal recognition particle of E. coli. Proteins **28**:285–288.

NAYLOR, G. J., and W. M. BROWN. 1997. Structural biology and phylogenetic estimation. Nature **388**:527–528.

OH, D. B., G. S. YI, S. W. CHI, and H. KIM. 1996. Structure of a methionine-rich segment of Escherichia coli Ffh protein. FEBS Lett. **395**:160–164.

OLSEN, G. J., and C. R. WOESE. 1997. Archaeal genomics: an overview. Cell **89**:991–994.

PHILIPPE, H. 1993. MUST, a computer package of management utilities for sequences and trees. Nucleic Acids Res. **21**: 5264–5272.

———. 1997. Rodent monophyly: pitfalls of molecular phylogenies. J. Mol. Evol. **45**:712–715.

PHILIPPE, H., and A. ADOUTTE. 1998. The molecular phylogeny of Eukaryota: solid facts and uncertainties. Pp. 25–56 *in* G. COOMBS, K. VICKERMAN, M. SLEIGH, and A. WARREN, eds. Evolutionary relationships among Protozoa. Chapman and Hall, London.

PHILIPPE, H., and P. FORTERRE. 1999. The rooting of the universal tree of life is not reliable. J. Mol. Evol. (in press).

PHILIPPE, H., and J. LAURENT. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. **8**:616–623.

PHILIPPE, H., U. SÖRHANNUS, A. BAROIN, R. PERASSO, F. GASSE, and A. ADOUTTE. 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. J. Evol. Biol. **7**:247–265.

PIANKA, E. 1970. On r- and K-selection. Am. Nat. **104**:592–597.

POHLSCHRODER, M., W. A. PRINZ, E. HARTMANN, and J. BECKWITH. 1997. Protein translocation in the three domains of life: variations on a theme. Cell **91**:563–566.

POOLE, A. M., D. C. JEFFARES, and D. PENNY. 1998. The path from the RNA world. J. Mol. Evol. **46**:1–17.

RAPOPORT, T. A., M. M. ROLLS, and B. JUNGNICKEL. 1996. Approaching the mechanism of protein transport across the ER membrane. Curr. Opin. Cell. Biol. **8**:499–504.

RZHETSKY, A., F. J. AYALA, L. C. HSU, C. CHANG, and A. YOSHIDA. 1997. Exon/intron structure of aldehyde dehydrogenase genes supports the ''introns-late'' theory. Proc. Natl. Acad. Sci. USA **94**:6820–6825.

SCHWARTZ, R. M., and M. O. DAYHOFF. 1978. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science **199**:395–403.

SIDDALL, M. E. 1998. Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. Cladistics **14**:209–220.

SULLIVAN, J., and D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. **4**:77–86.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1.1. Illinois Natural History Survey, Champaign.

SWOFFORD, D., G. OLSEN, P. WADDELL, and D. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. HILLIS, C. MORITZ, and B. MABBLE, eds. Molecular systematics. Sinauer, Sunderland, Mass.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

VALENT, Q. A., D. A. KENDALL, S. HIGH, R. KUSTERS, B. OUDEGA, and J. LUIRINK. 1995. Early events in preprotein recognition in E. coli: interaction of SRP and trigger factor with nascent polypeptides. EMBO J. **14**:5494–5505.

WOESE, C. R., O. KANDLER, and M. L. WHEELIS. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87**:4576–4579.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11**:367–370.

———. 1997. Phylogenetic analysis by maximum likelihood (PAML). Version 1.3. Department of Integrative Biology, University of California at Bekerley.

ZELAZNY, A., A. SELUANOV, A. COOPER, and E. BIBI. 1997. The NG domain of the prokaryotic signal recognition particle receptor, FtsY, is fully functional when fused to an unrelated integral membrane polypeptide. Proc. Natl. Acad. Sci. USA **94**:6025–6029.

MASAMI HASEGAWA, reviewing editor

Accepted February 24, 1999