Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 11295–11300, September 1998
Evolution

Genetic code origins: tRNAs older than their synthetases?

LLUÍS RIBAS DE POUPLANA, ROBERT J. TURNER, BRIAN A. STEER, AND PAUL SCHIMMEL*

The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037

Contributed by Paul Schimmel, July 20, 1998

ABSTRACT We present a phylogenetic analysis to determine whether a given tRNA molecule was established in evolution before its cognate aminoacyl-tRNA synthetase. The earlier appearance of tRNA versus their metabolically related enzymes is a prediction of the RNA world theory, but the available synthetase and tRNA sequences previously had not allowed a formal comparison of their relative time of appearance. Using data recently obtained from the emerging genome projects, our analysis points to the extant forms of lysyl-tRNA synthetase being preceded in evolution by the establishment of the identity of lysine tRNA.

The hypothesis of an RNA world postulates that self-replicating RNA molecules preceded the use of DNA and proteins, and that this world existed before the appearance of the universal ancestor of the extant tree of life (1). The existence of an RNA world has been supported by the biochemical characterization of catalytic RNA molecules, either from contemporary metabolic pathways or after *in vitro* selection of RNA ribozymes (2–6). Viral RNA genomes and the role or tRNA-like structures in viral replication are also indicative of the ancestral existence of an RNA world (7). A more direct proof of an RNA world could come from the direct comparison of the evolutionary time of appearance of protein and RNA molecules involved in a universal metabolic pathway. If this analysis was possible, then the RNA world theory would predict that the moment of appearance of the RNA component would precede the appearance of the protein elements involved in the same reaction. Here we present a phylogenetic analysis that suggests that, in an RNA-protein interaction essential for the elucidation of the genetic code, the RNA molecule is ancestral to its associated enzyme.

Aminoacyl-tRNA synthetases (aaRSs) evolved as two distinct classes (I and II), each containing 10 enzymes (8–14). Each aaRS is responsible for establishing the genetic code by specifically aminoacylating only its cognate tRNA isoacceptors, thereby linking an amino acid with its corresponding anticodon triplets. Because the aminoacylation of tRNA establishes the genetic code, a strong coevolution exists between the enzymes and their cognate tRNAs (15). The aminoacylation reaction precedes the first split of the tree of life, resulting in almost invariable conservation of aaRSs and their cognate tRNAs in all living organisms (16).

The strict conservation of aaRS and tRNA sequences across the whole phylogenetic tree prevented the analysis of initial events in the evolution of the system, because no sequences exist from precursors of the extant aaRSs. Without this kind of sequence information, the relative age of the duplications that gave rise to the current set of aaRSs could not be calculated. Moreover, the relative time of appearance of aaRSs and tRNAs could not be analyzed, because no extant organism are known presently where earlier, simpler sets of aaRS or tRNAs are used. As a result, it has not been possible to calculate whether the final evolutionary events that gave rise to modern aaRSs had taken place after the time when tRNAs had already evolved.

This situation changed with the sequencing of the genome of the archaebacterium *Methanococcus jannaschii* (17) and with the exponential growth of sequence data from other genome sequencing projects. In an initial analysis, *M. jannaschii*'s genome was found to lack an ORF coding for a canonical class II LysRS. Two reports by Ibba *et al.* (18, 19) established that the aminoacylation of $\mathrm{tRNA}^{\mathrm{Lys}}$ in a subset of archaebacteria (i.e., *M. jannaschii*, a member of the euryarchae) and bacteria of the spirochete group (i.e., *Treponema pallidum* and *Borrelia burgdorferii*) appears to be catalyzed by a class I-type LysRS. This is the first example of a class switch by an aaRS.

The origin of this new enzyme must lie, presumably, within the set of duplication events that gave rise to the rest of class I aaRSs. However, its distribution within the phylogenetic tree (it is present in a limited number of archaeal and bacterial species) can be initially explained by three different evolutionary models (Fig. 1). A first possibility would be a late duplication event from a class I aaRS in one of these branches, followed by horizontal gene transfer. Another potential model would require a late duplication event that, independently, gave rise to two different class I LysRSs in a subgroup of archaea and of bacteria. Finally, the observed distribution also can be explained by an early duplication event, at the base of the phylogenetic tree, which produced a class I LysRS that later was conserved only in limited groups of organisms, while the majority of species adopted a class II LysRS. The later scheme of events would imply the coexistence of class I and II LysRS enzymes in an organism ancestral to all existing species (Fig. 1). The phylogenetic relationships between class I LysRS sequences and the rest of class I aaRSs would be different in each model. As a result, cladistic analysis can be used to test each of the three possible evolutionary schemes (Fig. 1).

The third evolutionary model would make possible, for the first time, the use of phylogenetic methods to determine the relative age of an aaRS and its cognate tRNA isoacceptors. If $\mathrm{tRNA}^{\mathrm{Lys}}$ preceded the appearance of LysRS (whether class I or II), then the preservation of the genetic code would require the emerging enzymes to recognize the existing $\mathrm{tRNA}^{\mathrm{Lys}}$. These lysine tRNAs would have remained phylogenetically related in extant organisms independently of the type of LysRS used to aminoacylate them.

In this paper we report that phylogenetic methods point to the newly found class I LysRSs constituting a monophyletic group in the context of other class I aaRSs. That is, these class I LysRSs are more related to each other than to the rest of the enzymes in the class. More detailed analysis of closely related sequences points to a relationship between class I LysRS and CysRS, ArgRS, and GluRS. Through the analysis of the phylogenetic relationship between class I LysRS and the rest of class I enzymes, we conclude that the distribution of LysRS

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: RS, tRNA synthetase; aaRS, aminoacyl RS.

*To whom reprint requests should be addressed. e-mail: Schimmel@scripps.edu.

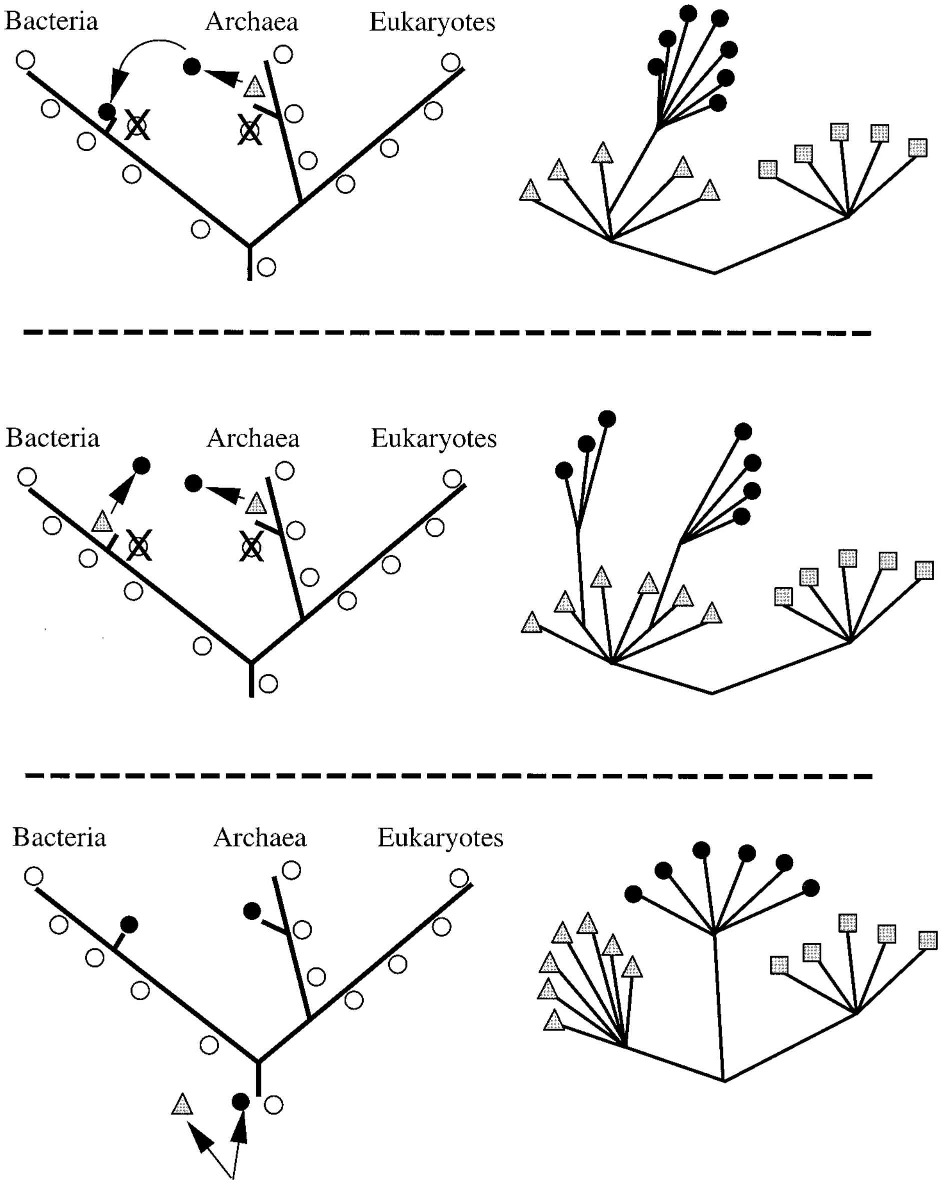

FIG. 1. (*Left*) The three possible schemes that may have given rise to the extant distribution of class I LysRS enzymes within the phylogenetic tree. (*Right*) The predicted phylogenetic relationships that would be expected between class I LysRS sequences and their closest class I aaRSs. (*Top*) A late duplication event followed by horizontal transfer. (*Middle*) Two independent duplication and gene replacement events. (*Bottom*) Two ancestral enzymes coexist initially, and they displace each other in different groups. ○, Class II LysRS. ●, Class I LysRS. ▲ ■, Class I synthetases closest to class I LysRS. ⊗, Displacement of class II LysRS.

in the phylogenetic tree is not caused by horizontal gene transfer. Thus, the ancestor of class I LysRS seems to have coexisted with the ancestral class II LysRS at the root of the tree of life.

The ancestral coexistence of two different types of synthetases that catalyze the same reaction makes possible (through the analysis of the sequences of the corresponding tRNAs) the testing of the ancestral origin of a tRNA with respect to its cognate enzymes. Our evolutionary analysis of tRNA^{Lys} sequences from the bacterial and archaeal branches of the phylogenetic tree suggests a single origin for this molecule. This origin is independent of the enzyme used to charge tRNA^{Lys} in any given organism. Thus, the identity of tRNA^{Lys} appears to have been established before the nature of the enzyme that reacts with it.

MATERIALS AND METHODS

All tRNA and aaRS sequences were obtained from GenBank (20). The tRNA^{Lys} gene sequences of *T. pallidum* and *B. burgdorferii* were extracted from their respective genomes with the program TRNASCAN (21).

Sequence alignments were done with CLUSTALW (22). The alignment of tRNAs was done with and without the anticodon sequences and checked for consistency with other data. The alignments of class I aaRSs were carried out with a variety of gap opening and gap extension penalties and were inspected visually to ensure the proper alignment of the conserved sequence motifs of the family (11). Several analyses were carried out with different sets of sequences. To establish the position of class I LysRS within the whole group of class I enzymes, three independent analyses were performed. First, a set of sequences from all available species was used to analyze the relationship of the class I LysRS with the rest of the class I enzymes. Second, species-specific analyses were done for five species found to contain a class I type LysRS (*T. pallidum, B. burgdorferi, M. jannaschii, Archeoglobus fulgidus,* and *Pyrococcus horikoshii*). This analysis was carried out to test relationships independently in each of these species. Finally, to analyze with more sensitivity the relationship between class I LysRS and its closest enzymes within its class, new alignments and phylogenies were constructed with those class I sequences having the highest sequence similarity to class I LysRS (ArgRS, CysRS, and GluRS).

All phylogenetic analysis was done by parsimony methods (PROTPARS and DNAPARS) (23), which were later confirmed by distance methods (KISTCH) (23). The soundness of the alignments used for the analysis was tested by bootstrap analysis (typically 100 replicates). Heuristic searches (usually 50 cycles) were used to maximize the space searched by the maximum parsimony algorithm.

RESULTS

The analysis of all class I tRNA synthetases for each of five different species (*T. pallidum, B. burgdorferi, M. jannaschii, A. fulgidus,* and *P. horikoshii*) consistently placed the class I LysRS sequence outside the large hydrophobic group (IleRS, ValRS, LeuRS, and MetRS), and closer to ArgRS, CysRS, and GluRS (Fig. 2). With minor variations these relationships were maintained in the trees built with the combined set of sequences from all species, in which class I LysRS sequences behaved as a monophyletic group (data not shown).

To increase the quality of the sequence alignments and to analyze more sensitively the relationships between ArgRS, CysRS, GluRS, and LysRS, all sequences available for these enzymes were used to generate sequence alignments and evolutionary relationships. LysRS sequences, once again, behaved as a monophyletic group more closely related to CysRS (Fig. 3). These relationships were confirmed by distance methods, which strengthened the monophily of the LysRS sequence cluster (data not shown).

The strong clustering of class I LysRS and, more importantly, the strong clustering of the related class I enzymes,

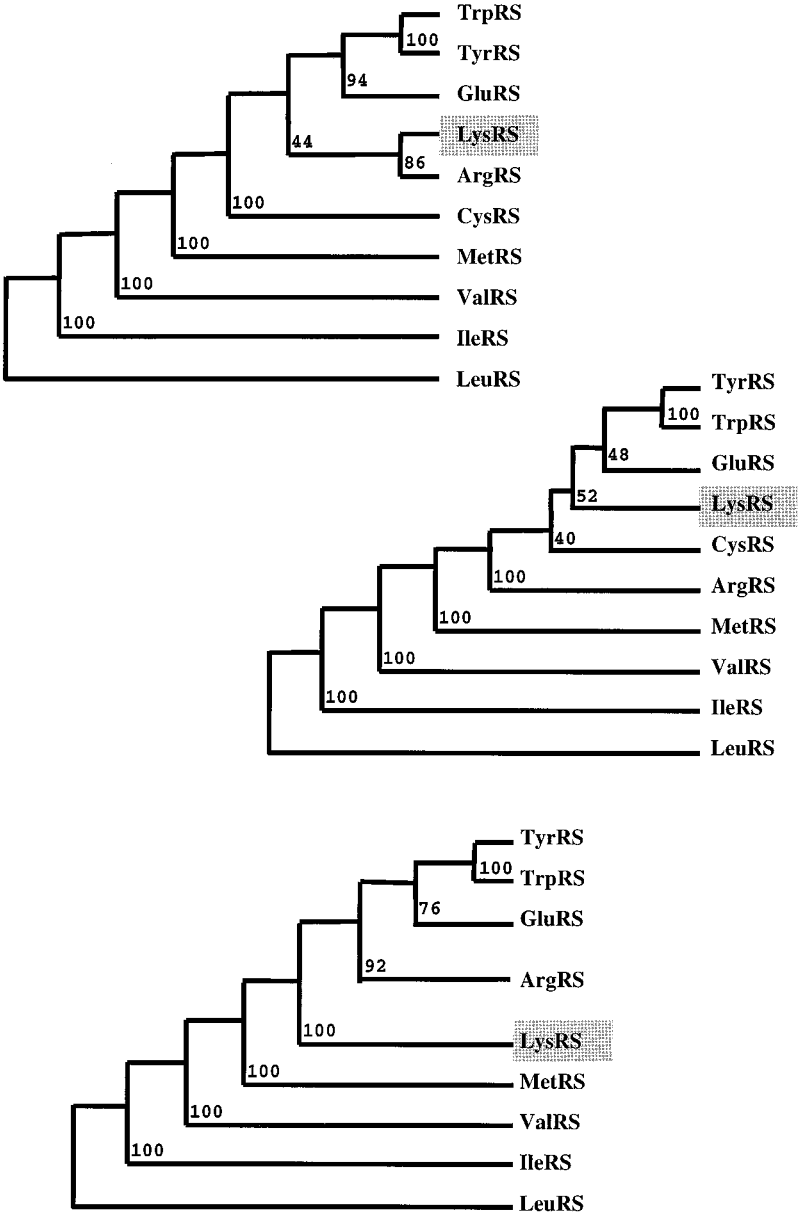

FIG. 2. Evolutionary relationships obtained for class I LysRS in the context of the sequences of all other class I aaRS. The analyses of five species that contain a class I *lysS* gene are shown. Numbers at branches correspond to bootstrap frequencies obtained from 100 replicates. (*Top*) Tree obtained with *A. fulgidus* class I aaRS sequences. (*Middle*) Tree obtained with class I aaRS sequences from *B. burgdorferi*, *T. pallidum,* and *P. horikoshii* (bootstrap frequencies correspond to the tree obtained with *T. pallidum* sequences). (*Bottom*) Tree obtained with *M. jannaschii* class I aaRS sequences.

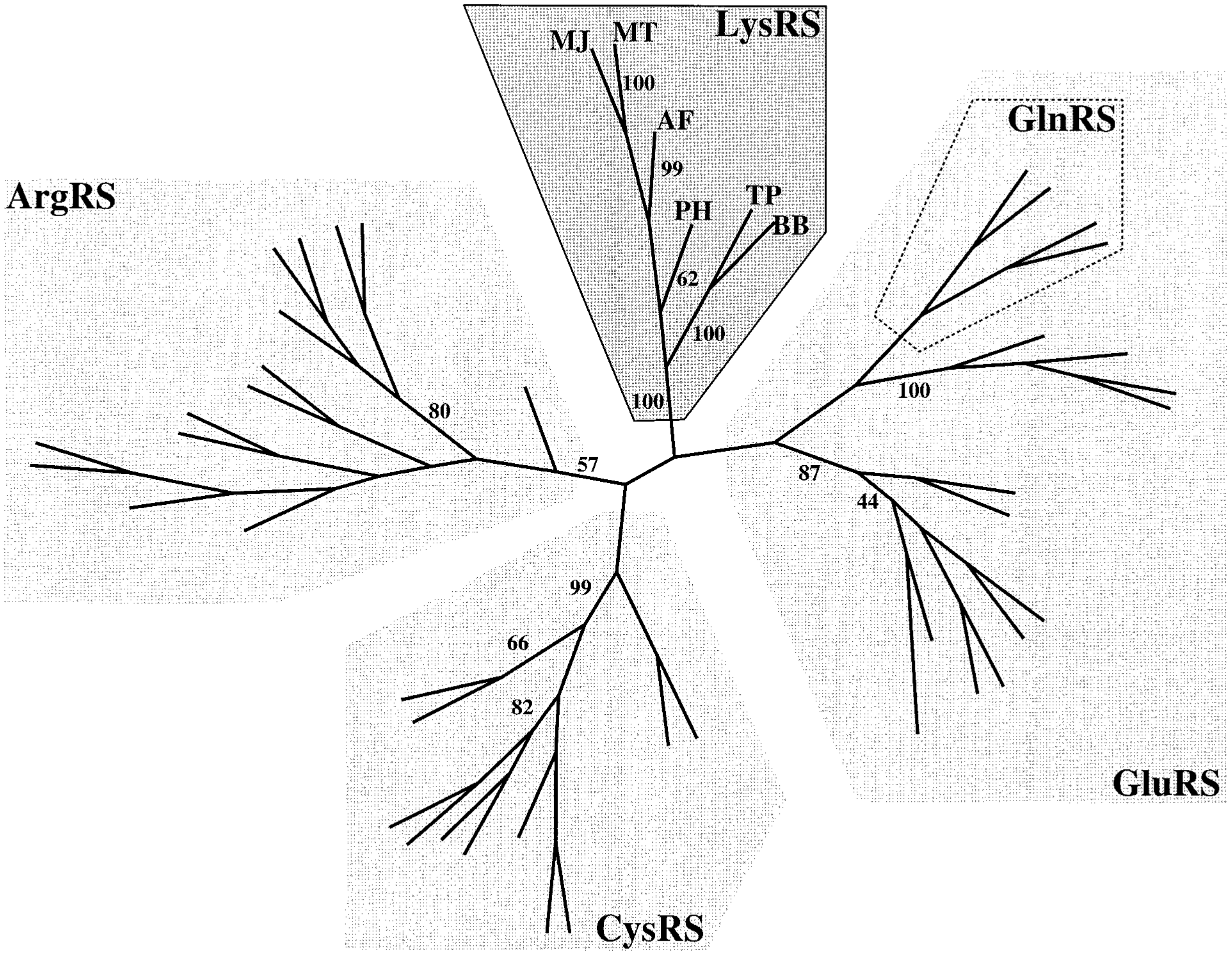

FIG. 3. Unrooted maximum parsimony tree for all known class I LysRS sequences in the context of the sequences of CysRS, ArgRS, GluRS, and GlnRS from several bacterial, archaeal, and eukaryotic organisms. Numbers at nodes correspond to bootstrapping frequencies for 100 different trees. AF, *A. fulgidus*; BB, *B. burgdorferi*; TP, *T. pallidum*; MJ, *M. jannaschii*; MT, *M. thermoautotrophicum*.

implies that the origin of the class I LysRS group is not the result of a late gene duplication event. The deep rooting of the class I LysRSs suggests that they share a common ancestor from which they evolved before the first split of the evolutionary tree (Fig. 1, *Bottom*).

In contrast to the existence of two ancient forms of LysRS that originated from the two different classes, a phylogeny of bacterial and archaea tRNA^{Lys} (including those of *T. pallidum*, *B. burgdorferi,* and archaeal organisms that use a class I LysRS), in the context of sequences from all 20 tRNA types from *Escherichia coli* (including *E. coli* tRNA^{Lys}, charged by the class II LysRS) showed a strong clustering of tRNA^{Lys} sequences (Fig. 4). This clustering of tRNA^{Lys} sequences is not dependent on, or biased by, the anticodon sequences (data not shown).

DISCUSSION

The discovery of a class I LysRS enzyme in a limited set of organisms that occupy seemingly distant positions in the tree of life begs the question of the origin of this enzyme, and of the evolution of the LysRS-tRNA^{Lys} metabolic interaction. Our results suggest that all class I LysRS sequences share a common evolutionary ancestor, which existed before the bacteria-archaea evolutionary split (Fig. 1, *Bottom*).

A priori, the sequence distribution found for class I LysRS also could have been explained by a late duplication of a class I aaRS followed by a horizontal gene transfer event to a second group of species (19) (Fig. 1, *Top*). Similarly, two independent duplication and gene replacement events also could account for the present situation (Fig. 1, *Middle*). However, the phylogenetic relationships that would result from these kind of events would produce evolutionary trees with different connectivities to those found in our study. This difference becomes apparent when the relationships between the sequences of GluRS and GlnRS (a well documented case of parallel gene transfer; ref. 24) are compared with the quite different relationships of class I LysRS with ArgRS, GluRS, and CysRS (Fig. 3).

The nature and present distribution of the ancestor of class I LysRS is a question that remains to be solved. Clearly, the evolutionary scheme favored by our results (Fig. 1, *Bottom*) suggests that class I LysRS, or its ancestral enzyme, should have a wide distribution in the phylogenetic tree, because it must have been present at the root of the tree, and it is now found in two distant clusters of organisms. Moreover, one of these clusters (spirochetes) is placed in a rather late position in the 16S RNA-derived tree (25), suggesting that the gene for the class I LysRS in these organisms also should be found in other protists.

We do not have a satisfactory explanation for the absence of a close relative to the gene for class I LysRS in other bacteria. However, the relative time of appearance of the different bacterial groups is still a matter of debate (26). Spirochetes form a large, and highly evolved, group, which includes organisms with very different metabolic and ecological characteristics and that display a high level of divergence from the rest of bacterial species (25). Possibly spirochetes, as a group, had

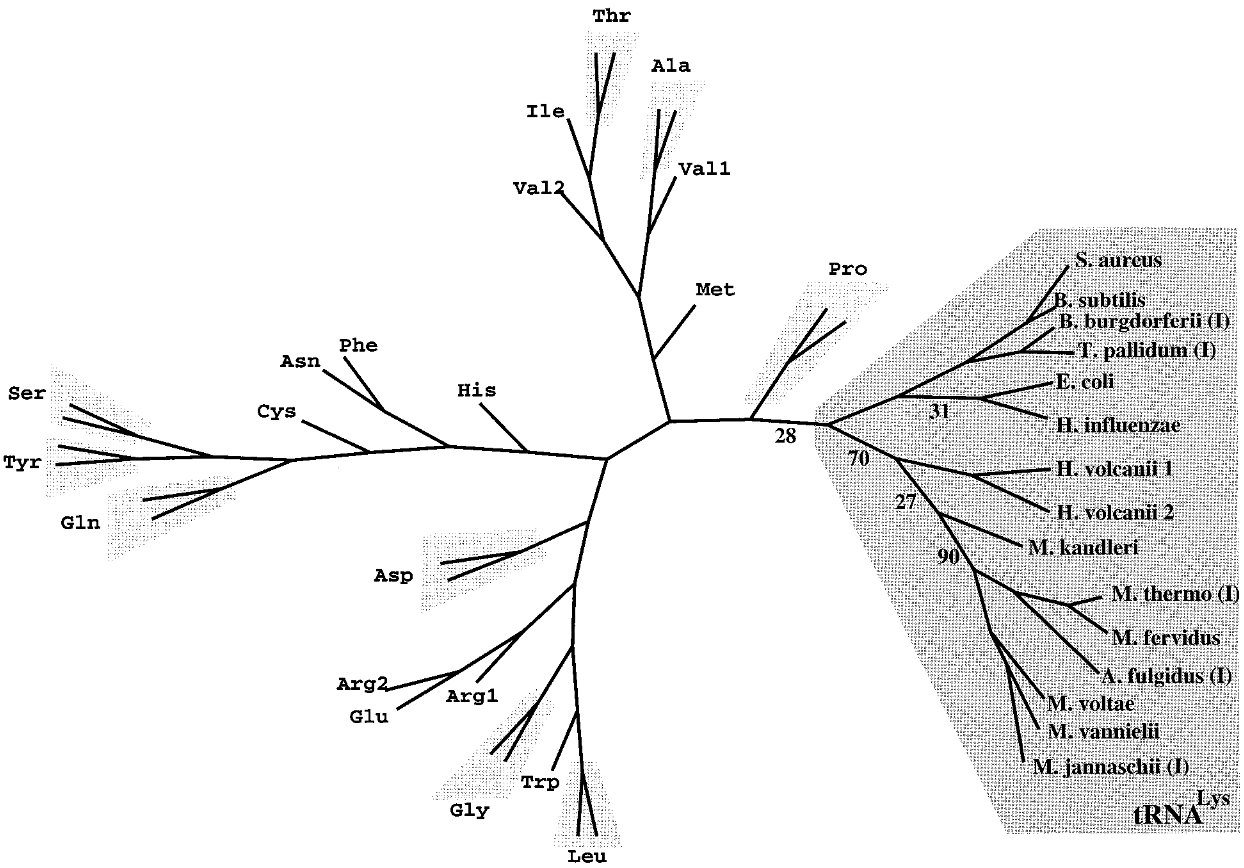

FIG. 4. Unrooted maximum parsimony tree of several archaeal and bacterial tRNA^{Lys} sequences, in the context of sequences from all 20 tRNA types from *E. coli*, aligned without their anticodon triplets (tRNAs other than tRNA^{Lys} are indicated by the three-letter code of their corresponding amino acid). Numbers at nodes of the tRNA^{Lys} branches correspond to bootstrapping frequencies for the branch closest to the number (calculated from 100 different trees). The tRNA^{Lys} sequences, charged by class I or II LysRS, are boxed, and those known to be charged by a class I LysRS are marked (I). *S. aureus*, *Staphylococcus aureus*; *B. subtilis*, *Bacillus subtilis*; *H. influenzae*, *Haemophilus influenzae*; *H. volcanii*, *Halobacterium volcanii*; *M. kandleri*, *Methanopyrus kandleri*; *M. thermo.*, *Methanobacterium thermoautotrophicum*; *M. fervidus*, *Methanothermus fervidus*; *M. voltae*, *Methanococcus voltae*; *M. vannielii*, *Methanococcus vannielii.*

a more primitive origin than that inferred from 16S RNA sequences. In this hypothetical situation, an early branching event giving rise to the spirochete group would explain the limited distribution of the *lysS* gene. This gene may have been lost in the main protist branch, which gave rise to the majority of bacterial species.

From the analysis of tRNA sequences it is clear that the extant tRNA^{Lys} sequences behave as a monophyletic group. Thus, the identity of this molecule appears to have been also established before the bacteria-archaea evolutionary split. As a result, we suggest that the evolution and definition of modern LysRSs was a process that took place around a pre-existing molecule, namely tRNA^{Lys}. Consistent with this conclusion, Ibba *et al.* (18, 19) reported that the class I LysRS from *B. burgdorferii* can efficiently aminoacylate *E. coli* tRNA^{Lys} (normally a substrate for a class II LysRS) (19). Given that class I and class II enzymes approach the acceptor helix from opposite sides (14, 16), we suspect that the fine structure ancestral helix determinants for charging were different for the two classes of enzymes.

Our results are consistent with the hypothesis that the emerging tRNA synthetases adapted to an already established tRNA^{Lys}, and thus also consistent with predictions that tRNAs preceded their synthetases (12, 27). The mechanism of aminoacylation of this primordial tRNA^{Lys} in the absence of its extant cognate enzymes may have involved a catalytic RNA (2, 28, 29).

We thank professors E. Alvarez-Buylla, W. F. Doolittle, and C. Woese, and Dr. J. Chihade for helpful comments and discussions. This work was supported by Grant GM23562 from the National Institutes of Health. R.J.T. holds a postdoctoral fellowship from the American Cancer Society (1996–99), and B.A.S. has a postdoctoral fellowship from the National Research Council of Canada (1996–99).

1. de Duve, C. (1991) *Blueprint for a Cell: The Nature and Origin of Life* (Neil Patterson, Burlington, NC).
2. Cech, T. R. & Bass, B. L. (1986) *Annu. Rev. Biochem.* **55,** 599–629.
3. Altman, S., Baer, M. F., Bartkiewicz, M., Gold, H., Guerrier-Takada, C., Kirsebom, L. A., Lumelsky, N. & Peck, K. (1989) *Gene* **82,** 63–64.
4. Noller, H. F., Hoffarth, V. & Zimniak, L. (1992) *Science* **256,** 1416–1419.
5. Szostak, J. W. (1993) *Nature (London)* **361,** 119–120.
6. Santoro, S. W. & Joyce, G. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 4262–4266.
7. Weiner, A. M. & Maizels, N. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7383–7387.
8. Webster, T., Tsai, H., Kula, M., Mackie, G. A. & Schimmel, P. (1984) *Science* **226,** 1315–1317.
9. Hountondji, C., Dessen, P. & Blanquet, S. (1986) *Biochemie* **68,** 1071–1078.
10. Ludmerer, S. W. & Schimmel, P. (1987) *J. Biol. Chem.* **262,** 10807–10813.
11. Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. (1990) *Nature (London)* **347,** 203–206.
12. Nagel, G. M. & Doolittle, R. F. (1995) *J. Mol. Evol.* **40,** 487–498.
13. Brown, J. R. & Doolittle, W. F. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 2441–2445.
14. Cusack, S. (1997) *Curr. Opin. Struct. Biol.* **6,** 881–889.

15. Ribas de Pouplana, L., Frugier, M., Quinn, S. & Schimmel, P. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 166–170.
16. Moras, D. (1992) *Trends Biochem. Sci.* **17,** 159–164.
17. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D., *et al*. (1996) *Science* **273,** 1017–1140.
18. Ibba, M., Morgan, S., Curnow, A. W., Pridmore, D. R., Vothknecht, U. C., Gardner, W., Lin, W., Woese, C. R. & Soll, D. (1997) *Science* **278,** 1119–1122.
19. Ibba, M., Bono, J. L., Rosa, P. A. & Soll, D. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 14383–14388.
20. Benson, D., Boguski, M., Lipman, D. J. & Ostell, J. (1994) *Nucleic Acids Res.* **22,** 3441–3444.
21. Lowe, T. M. & Eddy, S. R. (1997) *Nucleic Acids Res.* **25,** 955–964.
22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
23. Felsenstein, J. (1993) *Phylogeny Inference Package*, Department of Genetics, Univ. of Washington, Seattle.
24. Lamour, V., Quevillon, S., Diriong, S., N′Guyen, V. C., Lipinski, M. & Mirande, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 8670–8674.
25. Paster, B. J., Dewhirst, F. E., Weisburg, W. G., Tordoff, L. A., Fraser, G. J., Hespell, R. B., Stanton, T. B., Zablen, L., Mandelco, L. & Woese, C. R. (1991) *J. Bacteriol.* **173,** 6101–6109.
26. De Rijk, P., Van de Peer, Y., Van den Broeck, I. & De Wachter, R. (1995) *J. Mol. Evol.* **41,** 366–375.
27. Woese, C. (1967) *The Genetic Code* (Harper and Row, New York).
28. Piccirilli, J. A., McConnell, T. S., Zaug, A. J., Noller, H. F. & Cech, T. R. (1992) *Science* **256,** 1420–1424.
29. Illangasekare, M., Sanchez, G., Nickles, T. & Yarus M. (1995) *Science* **267,** 643–647.