Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4443–4447, July 1985
Evolution

Mitochondrial origins

(*Agrobacterium tumefaciens*/*Pseudomonas testosteroni*/α purple bacteria/16S rRNA sequence/evolution)

D. YANG*, Y. OYAIZU*, H. OYAIZU*, G. J. OLSEN†, AND C. R. WOESE*‡

*Department of Genetics and Development, University of Illinois, 515 Morrill Hall, 505 South Goodwin Avenue, Urbana, IL 61801; and †Department of Biology, Indiana University, 142 Jordan Hall, Bloomington, IN 47405

Communicated by Ralph S. Wolfe, March 11, 1985

ABSTRACT **The 16S ribosomal RNA sequences from *Agrobacterium tumefaciens* and *Pseudomonas testosteroni* have been determined to further delimit the origin of the endosymbiont that gave rise to the mitochondrion. These two prokaryotes represent the α and β subdivisions, respectively, of the so-called purple bacteria. The endosymbiont that gave rise to the mitochondrion belonged to the α subdivision, a group that also contains the rhizobacteria, the agrobacteria, and the rickettsias—all prokaryotes that have developed intracellular or other close relationships with eukaryotic cells.**

The question of mitochondrial origins has on one level been settled. Mitochondria arose as bacterial endosymbionts within some ancestral type of eukaryotic cell (1–3). The question has now become the nature, the phylogenetic origin, of the endosymbiont(s). The proper comparative analyses of macromolecular sequences should provide the answer.

Cytochrome *c* comparisons ostensibly localize the mitochondrial origin rather precisely. Mitochondrial cytochromes *c* are of the medium subunit type (4). This, and the large subunit type appear confined to a particular group of purple photosynthetic bacteria and their relatives—now called the α subdivision of the purple bacteria (4–6). The remaining subunit type, small (which is presumably the ancestral type), is found in other eubacterial groups, but not among the α purple bacteria (4). Mitochondrial cytochromes *c* are also relatively close to the bacterial medium subunit cytochromes in sequence (4). However, the gene for mitochondrial cytochrome *c* is located in the eukaryotic nucleus, not in the mitochondrial genome, which makes the assumption that cytochrome *c* genealogy represents mitochondrial genealogy somewhat questionable. Additional evidence is needed if the case for mitochondrial origins is to be a strong one.

Ribosomal RNA genes do reside in the mitochondrial genome (1). Unfortunately, mitochondrial rRNA sequences do not readily localize the origin of the mitochondrion precisely—for the reason that these RNAs have changed drastically over their evolutionary course. They are very different from all other rRNA sequences and remarkably different from one another as well (2, 3, 7). So far all that has been concluded from their analysis is that they belong to the general eubacterial line of descent, that they represent a very deep branching therefrom (2, 7), and that they are probably polyphyletic in origin (7, 8).

The plant mitochondrial rRNAs are an exception to the above in that they are not so highly idiosyncratic; they are obviously eubacterial in kind, *almost* typically so (3, 34). Thus, if the appropriate eubacterial rRNA sequences were available for comparison, the origin of the plant mitochondrial rRNA could be localized to one of the eubacterial "phyla" (9), which, in turn, might serve to localize the origin of the remaining mitochondria. Since cytochrome *c* analyses suggest the mitochondrion to have arisen from the purple bacterial "phylum," we have sequenced the 16S rRNA genes from *Agrobacterium tumefaciens* and *Pseudomonas testosteroni*, organisms that represent the α (6) and β (10) subdivisions, respectively, of the purple bacteria. The *Escherichia coli* 16S rRNA (11) is representative of the remaining, γ, subdivision (12).

MATERIALS AND METHODS

Strains. *Agrobacterium tumefaciens* DSM 30105 was a generous gift of the German Collection of Microorganisms (6). Our isolate of *Pseudomonas testosteroni* ATCC 11996, the type strain, was maintained for some time (as strain KS 0043) in the culture collection of K. Komagata (Tokyo).

Growth of Organisms and Isolation of Nucleic Acids. The organisms were grown under standard conditions and harvested in logarithmic phase (6, 9). Nucleic acids, RNA and DNA, were isolated by standard procedures (13, 14).

Cloning. The rRNA genes of *A. tumefaciens* were cloned as a 9-kilobase-pair (kb) partial *Eco*RI restriction fragment in the λgt*WES*·λB system, using standard methods (15, 16). From the original clone, two subfragments, approximately 0.6 and 2.0 kb, which together covered the entire 16S rRNA gene, were produced by complete *Eco*RI digestion and inserted into the phage M13 (mp8 and mp9) vectors for use as sequencing templates (17). [The smaller clone covers the 16S rRNA gene from just before the 5′ terminus to the *Eco*RI site ending at position 679, in *E. coli* numbering (11). The larger clone starts from this site and ends well into the 23S rRNA gene.]

The corresponding *P. testosteroni* rRNA gene was initially cloned as a partial *Sau*3A restriction fragment, of about 13 kb, in the *Bam*HI site of λL47.1 (18). Subcloning in phage M13 utilized, as above, the *Eco*RI site (that ends at position 679), to give two fragments covering the entire 16S rRNA gene. In addition, several subclones of these two were created in the phage M13 system—running from (*i*) upstream of the 5′ end to position 187 (*Bgl* II); (*ii*) position 182 (*Bgl* II) to 679 (*Eco*RI); (*iii*) position 674 (*Eco*RI) to 1359 (*Sau*3A); and (*iv*) position 1356 (*Sau*3A) to 1530 (*Sau*3A).

Sequencing Methods. The 2′,3′-dideoxynucleotide chain termination method (19) was used throughout. Synthesized strands were labeled by the inclusion of deoxyadenosine 5′-[α-[^{35}S]thio]triphosphate (20). Two types of G sequencing reactions were routinely employed, one normal, the other in which dGTP was replaced by dITP (dideoxy GTP being used to terminate chain growth) (21). (The method affords better resolution in regions where G bands are otherwise collapsed.) The M13 priming site (17) as well as specific priming sites within the rRNA gene, for which primers were synthesized (most at the University of Illinois DNA Synthesis Facility),

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.

Abbreviation: kb, kilobase pair(s).

‡To whom reprint requests should be addressed.

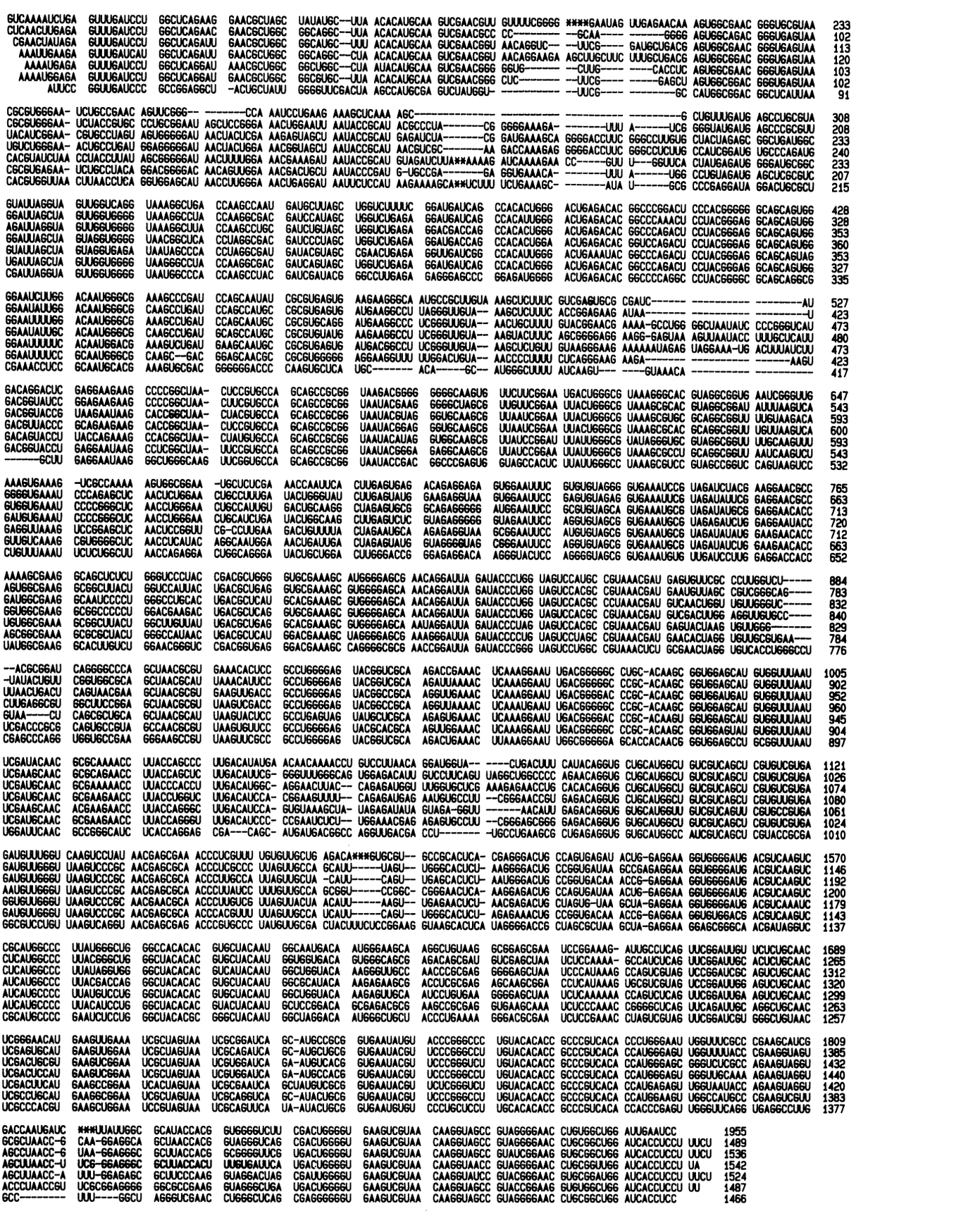

FIG. 1. Alignment of seven 16S rRNA-like gene sequences. The sequences, in order of listing, are from the following: wm, wheat mitochondrion (3); At, *A. tumefaciens*; Pt, *P. testosteroni*; Ec, *E. coli* (11); Mc, *Mycoplasma capricolum* (23); An, *Anacystis nidulans* (24); and Mv, *Methanococcus vannielii* (25). To make presentation more concise three large, idiosyncratic insertions (replaced by asterisks) have been omitted from the mitochondrial sequence (3), as have several smaller idiosyncratic regions (similarly indicated) from two of the other sequences. [The sequence of *M. capricolum* has been corrected in about five positions, where we feel certain, from oligonucleotide analysis (26), that the published version is incorrect (unpublished analysis).] Each sequence is broken into lines of 100 *E. coli* positions. Numbering and 10-nucleotide spacings reflect the *E. coli* sequence as well (11).

were used. Details of the internal primer system will be published elsewhere.

In determining the *A. tumefaciens* sequence, the following sequence fragments were read from the indicated priming sites (arrows show the direction of reading, the number in parentheses gives the approximate location of the internal priming site, and "M13" in parentheses means the usual M13 priming site was used): 5′→145 (M13); 10←235 (270); 15←330 (340); 35→335 (010); 265→590 (255); 280←505 (520); 360←679 (M13); 535→679 (520); 674→940 (M13); 674←895 (920); 810→1085 (795); 935←1090 (1100); 1085→1360 (1050); 1080←1370 (1400); 1250→1450 (1240); 1300←1520 (1540); 1415→3′ (1400).

In determining the *P. testosteroni* sequence, the following sequence fragments were generated from the indicated priming sites: 5′←186 (M13); 183→450 (M13); 275→510 (260); 350←510 (520); 350←679 (M13); 370→490 (340); 535→679 (520); 674→880 (M13); 680←900 (920); 840→1095 (800); 890←1090 (1100); 1085→1320 (1050); 1090←1350 (1390); 1170←1359 (M13); 1356→1533 (M13); 1465→3′ end (1400).

Sequence Alignment and Tree Construction. The sequences were aligned by the procedure of Woese *et al.* (22)—i.e., an initial alignment based upon obvious sequence homology was refined by use of the known secondary structural features of the molecule.

Analyses were limited to those regions of the aligned sequences in which structural homology was evident among all sequences being compared. The calculation of sequence homologies, the conversion to estimates of evolutionary distance, and the inference of the tree most consistent with these data were performed as described previously (2).

RESULTS AND DISCUSSION

Fig. 1 shows an alignment of the 16S rRNA-like sequences from (*i*) the plant mitochondrion; (*ii–iv*) representatives of the α, β, and γ subdivisions of the purple bacteria—i.e., *A. tumefaciens*, *P. testosteroni*, and *E. coli*, respectively (6, 10–12); (*v*) *Mycoplasma capricolum* (23), which is phylogenetically a Gram-positive eubacterium (26, 27) (see Fig. 2); (*vi*) *Anacystis nidulans* (24); and (*vii*) the archaebacterium *Methanococcus vannielii* (25).

The upper right triangle of data in Table 1 is a similarity matrix based upon those positions in the Fig. 1 alignment that are represented in all of the first six sequences. The plant mitochondrial 16S rRNA-like sequence is clearly closest to its counterparts among the purple bacteria, closest of all to the α subdivision representative, from *A. tumefaciens*. However, the converse does not hold: *All* the eubacterial sequences in Fig. 1 are closer to one another than any is to the plant mitochondrial sequence. There are two possible explanations of the low overall homology the plant mitochondrial sequence shows with the eubacterial sequences. Either the mitochondrial sequence branched deeply and the similarity to *A. tumefaciens* is a coincidence, or the mitochondrial sequence is specifically related to the *A. tumefaciens* sequence and has diverged from all the sequences due to a more rapid accumulation of mutations. Comparisons with the archaebacterial sequence support this latter explanation; although all the eubacterial sequences are about equally similar to the archaebacterial sequence, the mitochondrial sequence is significantly less similar (data not shown), indicating a higher rate of divergence of the mitochondrial lineage.

Table 1. Homology matrix for sequences of Fig. 1

	wm	Ag	Pt	Ec	Mc	An
wm	—	48	38	35	34	34
Ag	46 (26)	—	55	57	52	53
Pt	35 (14)	48 (6)	—	61	52	52
Ec	33 (7)	52 (14)	58 (33)	—	48	52
Mc	33 (21)	47 (14)	47 (22)	44 (9)	—	50
An	29 (17)	44 (14)	44 (18)	44 (21)	44 (22)	—

Data are analyzed in three ways: (*i*) Upper right triangle. Only positions represented in the mitochondrial and all five eubacterial sequences are considered; for convenience of analysis all positions of constant composition among these six are excluded; similarity is expressed as percentage of total positions (in this case 617) in which the given pair have the same composition. (*ii*) Lower left triangle. Additionally removed from consideration are those positions of constant composition among six of the seven sequences in the Fig. 1 alignment; similarity is again percentage of total positions (in this case 491). (*iii*) Lower left triangle, values in parentheses. Number of positions (set of 617) in which composition is the same in *and* unique to a given pair of sequences (exclusive of the archaebacterial sequence). Abbreviations are as in Fig. 1.

The relationships among sequences are more concisely summarized as a phylogenetic tree. Fig. 2 presents a phylogeny inferred from the small subunit rRNA sequences presented in Fig. 1, as well as those of *Bacillus subtilis* (28), and representative animal (29), fungal (7), and ciliate (30) mitochondria. The tree groups the mitochondrial sequences together, and specifically with the *A. tumefaciens* sequence. Testing different combinations of these sequences leads us to several generalizations regarding the tree: The inclusion of the more rapidly evolving—i.e., nonplant—mitochondrial sequences "pushes" the mitochondrial lineage closer to the root of the tree, decreasing the apparent affinity to the *A. tumefaciens* sequence in the tree shown. The tree is insensitive to the choice or omission of archaebacterial representation. The branching order in the tree remains unchanged if a eukaryotic sequence is added and the mouse mitochondrial sequence is omitted. The great divergence of the mouse mitochondrial sequence from all other sequences makes its branching location the least certain.

It is instructive to analyze the alignment of Fig. 1 further to reveal the source of the asymmetry in the initial analysis of sequence similarities (Table 1). Its branch length in Fig. 2 suggests that the plant mitochondrial rRNA has evolved more rapidly than its bacterial counterparts—but nowhere near as rapidly as the other mitochondrial rRNAs. Rapidly evolving rRNAs show a pronounced tendency to vary from sequence patterns common to all normal rRNAs (22, 31). This characteristic is clearly evident in the plant mitochondrial sequence (and, of course, far more pronounced in the other mitochondria). To give some examples, the plant sequence contains large idiosyncratic insertions in the regions 80–90, 1130–1140, and 1445–1455 (3), and lacks the structure in the 143–179 region characteristic of all eubacterial, archaebacterial, and eukaryotic rRNAs (22, 32). More subtle idiosyncrasy occurs in the secondary structural loops located at positions 297–300 and 618–622 (see Fig. 1), which otherwise conform to a common pattern in all eubacteria and archaebacteria (22). Single nucleotide idiosyncrasies can be seen—e.g., at positions 1381 and 1413. [In the last three examples the common pattern is defined not simply by the sequences of Fig. 1 but by the 400 or so bacterial 16S rRNAs that have been partially sequenced by the oligonucleotide cataloging method (9, 27).]

The analysis in the upper triangle of Table 1 corrects for this idiosyncrasy to a slight extent in confining itself to those positions that are represented in all eubacterial (and the plant mitochondrial) sequences. A more extensive correction for idiosyncrasy would be to additionally eliminate from the analysis of the Fig. 1 alignment all positions wherein one sequence differs from a pattern that is common to all the rest (including the archaebacterial sequence). Of the additional 125 or so positions thereby excluded, 50 are from the plant mitochondrial sequence! This correction, shown in the lower left triangle of Table 1, significantly enhances the apparent

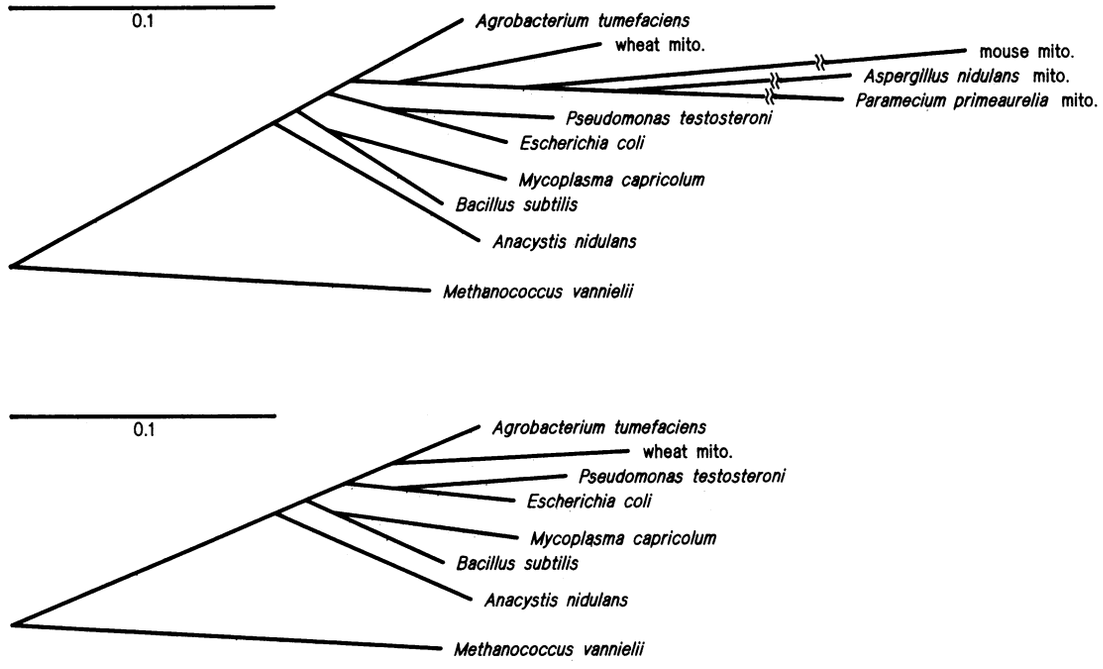

FIG. 2. Phylogenetic trees of small subunit rRNA sequences. The evolutionary distances (estimated number of mutational events per sequence position) between the nodes of the trees are reflected in their horizontal separation. The upper tree includes mitochondrial sequences representing all of the eukaryotic kingdoms. Tree construction is based upon positions 6–63, 105–122, 240–254, 272–290, 310–315, 339–405, 499–587, 665–739, 761–825, 874–993, 1045–1069, 1183–1245, 1292–1420, and 1480–1534 (*E. coli* numbering). The branches to the mouse, *Aspergillus nidulans*, and *Paramecium primeaurelia* mitochondrial sequences have been shortened (note the line breaks) by 0.4, 0.2, and 0.2 unit, respectively. Due to the potential errors in the inference of the mouse mitochondrial branch location we do not attach significance to the difference between the mitochondrial branching order and that for the corresponding nuclear-defined rRNAs [in which the metazoan lineage diverges prior to the separation of the plant, fungal, and ciliate lineages (unpublished analysis)]. The lower tree was calculated without the nonplant mitochondrial sequences to eliminate the distortion introduced by the faster-evolving sequences and, thereby, more accurately illustrate the affiliation of the *A. tumefaciens* and mitochondrial lineages.

closeness between the plant mitochondrial and *A. tumefaciens* sequences. While the latter, as expected, remains the closest of all to the plant sequence, the converse now almost holds—i.e., the plant mitochondrial sequence is as close to the *A. tumefaciens* sequence as any of the others are, except the *E. coli* sequence. However, the other eubacterial sequences still remain further from the sequence of the plant mitochondrion than they are from each other.

Another way to bring out mitochondrial genealogy is to focus on (nonidiosyncratic) derived characters, traits that have arisen in, and so are characteristic of, particular groupings. In this case the alignment is scored for positions in which composition is not only *common* to a given pair of sequences, but *unique* to that pair as well (within the eubacteria). The result is shown in parentheses in the lower left triangle of Table 1. (As above, the analysis involves only those positions that are represented in all eubacterial sequences.) Now, not only is the *A. tumefaciens* sequence the closest to that of plant mitochondrion, but the converse holds as well.

The failure of previous analyses to cluster mitochondrial rRNA sequences specifically with one another and with the purple photosynthetic bacteria (then represented by *E. coli*) appears to be an artifact of the methodology and the data then available (2, 7, 8). Several features distinguish the present analysis from the earlier ones: (*i*) The earlier studies lacked as specific a mitochondrial relative as *A. tumefaciens* and as slowly diverging a mitochondrial sequence as that of the plant mitochondrion, thereby requiring the analysis to resolve smaller differences in relatedness against a larger background of sequence divergence; i.e., the earlier studies looked at smaller differences between larger numbers. (*ii*) The increased understanding of eubacterial 16S rRNA primary and secondary structure (22) has led to improved sequence alignments. (*iii*) The inclusion of animal mitochondrial and eukaryotic sequences in the same tree exaggerated the systematic errors in the tree inference methodology that is caused by widely varying evolutionary rates.

Thus, the mitochondrial rRNA analysis is now in accord with the conclusions from cytochrome *c* sequence analysis (4, 33). Mitochondria from all eukaryotic kingdoms appear to have originated from the α subdivision of the purple eubacterial "phylum."

The α subdivision of the purple bacteria contains several types of organisms whose common characteristic is intimate, usually intracellular, association with eukaryotic cells—the rhizobacteria, agrobacteria and, a recently recognized addition (unpublished observation), the rickettsias. It would be of interest to know whether the mitochondrial ancestor belonged to the same subgroup of the α subdivision, α-2 (6), as do these other organisms. The answer may be found in sequencing rRNA representatives of the three known subgroups of the α subdivision in an attempt to refine further the placement of the mitochondrial ancestor.

We are grateful to Prof. Kenneth Bott for permission to use the sequence of *B. subtilis* rRNA prior to its publication. This work is supported by grants to C.R.W. from the National Aeronautics and Space Administration (NSG-7044) and the National Science Foundation (DEB-81 07061). H.O. was supported by a fellowship from the Toyobo Biotechnology Foundation. G.J.O. was supported by National Institutes of Health Grant AM-34527 to Prof. N. R. Pace.

1. Gray, M. W. & Doolittle, W. F. (1982) *Microbiol. Rev.* **46**, 1–42.
2. McCarroll, R., Olsen, G. J., Stahl, Y. D., Woese, C. R. & Sogin, M. L. (1983) *Biochemistry* **22**, 5858–5868.
3. Spencer, D. F., Schnare, M. N. & Gray, M. W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 493–497.

4. Dickerson, R. E. (1980) *Nature (London)* **283,** 210–212.
5. Woese, C. R., Gibson, J. & Fox, G. E. (1980) *Nature (London)* **283,** 212–213.
6. Woese, C. R., Stackebrandt, E., Weisburg, W. G., Paster, B. J., Madigan, M. R., Fowler, V. J., Hahn, C. M., Blanz, P., Gupta, R., Nealson, K. H. & Fox, G. E. (1984) *Syst. Appl. Microbiol.* **5,** 315–326.
7. Kuntzel, H. & Kochel, H. G. (1981) *Nature (London)* **293,** 751–755.
8. Gray, M. W., Sankoff, D. & Cedergren, R. J. (1984) *Nucleic Acids Res.* **12,** 5837–5852.
9. Woese, C. R. (1985) in *The Evolution of Prokaryotes*, eds. Schleifer, K.-H & Stackebrandt, E. (Academic, London), in press.
10. Woese, C. R., Weisburg, W. G., Paster, B. J., Hahn, C. M., Tanner, R. S., Krieg, N. R., Koops, H.-P., Harms, H. & Stackebrandt, E. (1984) *Syst. Appl. Microbiol.* **5,** 327–336.
11. Brosius, J., Palmer, J. L., Kennedy, J. P. & Noller, H. F. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 4801–4805.
12. Woese, C. R., Weisburg, W., Hahn, C. M., Paster, B., Zablen, L. B., Lewis, B. J., Macke, T., Ludwig, W. & Stackebrandt, E. (1985) *Syst. Appl. Microbiol.* **6,** in press.
13. Marmur, J. (1961) *J. Mol. Biol.* **3,** 208–218.
14. Woese, C. R., Sogin, M., Stahl, D. A., Lewis, B. J. & Bonen, L. (1976) *J. Mol. Evol.* **7,** 197–213.
15. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
16. Leder, P., Tiemeier, D. & Enquist, L. (1977) *Science* **196,** 175–177.
17. Messing, J. (1983) *Methods Enzymol.* **101,** 20–78.
18. Loenen, W. A. M. & Brammar, W. J. (1980) *Gene* **20,** 249–259.
19. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
20. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3963–3965.
21. Mills, D. R. & Kramer, F. R. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 2232–2235.
22. Woese, C. R., Gutell, R. R., Gupta, R. & Noller, H. F. (1983) *Microbiol. Rev.* **47,** 621–669.
23. Iwami, M., Muto, A., Yamao, F. & Osawa, S. (1984) *Mol. Gen. Genet.* **196,** 317–322.
24. Tomioka, N. & Sugiura, M. (1983) *Mol. Gen. Genet.* **191,** 46–50.
25. Jarsch, M. & Böck, A. (1985) *Syst. Appl. Microbiol.*, in press.
26. Woese, C. R., Maniloff, J. & Zablen, L. B. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 494–498.
27. Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R., Magrum, L., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N. & Woese, C. R. (1980) *Science* **209,** 457–463.
28. Green, C. J., Stewart, G. C., Hollis, M. A., Vold, B. S. & Bott, K. F. (1985) *Gene*, in press.
29. Van Etten, R. A., Walberg, M. W. & Clayton, D. A. (1980) *Cell* **22,** 157–170.
30. Seilhamer, J. J., Olsen, G. J. & Cummings, D. J. (1984) *Jour. Biol. Chem.* **259,** 5167–5172.
31. Woese, C., Ludwig, W. & Stackebrandt, E. (1985) *J. Mol. Evol.*, in press.
32. Gutell, R. R., Weiser, B., Woese, C. R. & Noller, H. F. (1985) *Prog. Nucl. Acids Res. Mol. Biol.*, in press.
33. Schwartz, R. M. & Dayhoff, M. O. (1978) *Science* **199,** 395–403.
34. Bonen, L., Cunningham, R. S., Gray, M. W. & Doolittle, W. F. (1977) *Nucleic Acids Res.* **4,** 633–671.