Volume 12 Number 14 1984 Nucleic Acids Research

On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA

Michael W.Gray[1,*], David Sankoff[2] and Robert J.Cedergren[3]

[1]Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, [2]Centre de Recherche de Mathématiques Appliquées and [3]Departement de Biochimie, Université de Montréal, Montréal, PQ H3C 3J7, Canada

Received 8 May 1984; Accepted 2 July 1984

ABSTRACT

To probe the earliest evolutionary events attending the origin of the five known genome types (archaebacterial, eubacterial, nuclear, mitochondrial and plastid), we have analyzed sequences corresponding to a ubiquitous, highly conserved core of secondary structure in small subunit rRNA. Our results support (i) the existence of three primary lineages (archaebacterial, eubacterial, and nuclear), (ii) a specific eubacterial ancestry for plastids and mitochondria (plant, animal, fungal), and (iii) an endosymbiotic, evolutionary origin of the two types of organelle from within distinct groups of eubacteria (blue-green algae (cyanobacteria) in the case of plastids, nonphotosynthetic aerobic bacteria in the case of mitochondria). In addition, our analysis suggests (iv) a biphyletic origin of mitochondria, with animal and fungal mitochondria branching together but separately from plant mitochondria, and (v) a monophyletic origin of plastids. The method described here provides a powerful and generally applicable molecular taxonomic approach towards a global phylogeny encompassing all organisms and organelles.

INTRODUCTION

Molecular taxonomy ultimately seeks to establish evolutionary connections that encompass not only the three primary kingdoms (archaebacteria, eubacteria, and eukaryotes) into which all organisms can be divided [1], but in addition the mitochondrion and plastid of the eukaryotic cell, both of which contain distinctive genetic systems [2]. In theory, evolutionary relationships among the five known genome types (archaebacterial, eubacterial, nuclear, mitochondrial, plastid) can be elucidated by comparative analysis of the sequences of functionally equivalent informational macromolecules encoded by each of the five genomes. At present, the only molecules known to be of this type are the ribosomal and transfer RNA components of the translation systems found in prokaryotes (archaebacteria and eubacteria) and in the cytoplasm, mitochondria, and plastids of eukaryotes.

Sequences of tRNA and 5S rRNA have been widely used in the construction of phylogenetic trees [3-6]. However, for making global evolutionary connections, these macromolecules have certain limitations. For example, 5S

rRNA has not been found in mitochondria except in those of higher plants [7], and mitochondrial tRNAs, at least those in mammals, are sufficiently peculiar in structure that meaningful alignments with more conventional tRNA sequences cannot be generated by the usual methods [8].

Here, we present an approach to a global phylogeny using a data set derived from small subunit (SSU) rRNA sequences. Our choice of this informational macromolecule is based on (i) the universal occurrence of functionally equivalent and evolutionarily homologous SSU rRNA genes in eubacteria, archaebacteria, mitochondria, plastids, and nucleus; (ii) the recognition [9] that SSU rRNA contains a universally conserved core of secondary structure (probably defining its basic function in protein biosynthesis) that permits a very accurate alignment of the primary sequences of homologous regions [10], while at the same time providing a data set large enough to reduce the danger of misleading statistical fluctuations in the distribution of mutations; and (iii) the availability of a sufficient number of SSU rRNA sequences spanning the five known genome types to permit analysis of the early evolutionary events attending the origin of these genomes.

Molecular-based phylogenies are not easily and unequivocally determined from sequence data; indeed, different phylogenies can be inferred from the same data set depending on the reconstruction method used [4]. Furthermore, different phylogenies for a given group of organisms can be obtained by a single method applied to data derived from the sequences of different genes [4]. Clearly, then, great care must be taken in the selection of data sets, in their analysis, and in the evolutionary interpretation given to the results. In this paper, we make use of a methodology that minimizes the subjectivity normally inherent both in the selection of data sets and in the reconstruction technique employed. Our method is based on an interactive strategy, previously described by us [4], for inferring evolutionary relationships from nucleic acid sequence data.

RATIONALE AND STRATEGY OF ANALYSIS

Selection of the Data Set

At first view, comparison of SSU rRNA sequences would seem complicated by the marked variability in size and/or base composition of the homologous molecules. For example, known SSU rRNA sequences range in length from 953 bp in human mitochondria [11] to 1955 bp in wheat mitochondria [12], and in G+C content from 22.5% in yeast mitochondria to 55.8% in maize chloroplasts [9]. Although a potentially informative collection of diverse SSU rRNA sequences

now exists, the valid alignment of these structurally heterogeneous sequences presents a serious problem. Fortunately, comparative analysis has provided considerable insight into the probable architecture of SSU rRNAs, and has revealed a pattern in which relatively conserved stretches of primary sequence and/or secondary structure alternate with variable domains that differ markedly in size, base composition, and potential secondary structure [13,14]. This allows the recognition and selection of homologous regions of SSU rRNA for phylogenetic analysis, and provides a strategy for generating reliable primary sequence alignments of these homologous regions, based on secondary structure considerations.

For the purposes of the present analysis, we have divided structural domains in SSU rRNA into three categories, which we designate "U" (universally conserved), "S" (semi-conserved), and "V" (variable, or non-conserved). Fig. 1 illustrates the interspersion of these three types of sequence in the secondary structure model of Escherichia coli 16S rRNA. U segments define a highly conserved secondary structure core, identified by Stiegler et al. [9], that is ubiquitous among SSU rRNAs. This "universal core" is composed of nine non-contiguous segments of primary sequence that are held together by short- and long-range hydrogen bonding interactions and that can assume an overall secondary structure containing the same helical elements in all cases. Alignment of homologous segments of the core is relatively easy, since each residue has a precisely defined secondary structure position, and a number of invariant and semi-invariant residues serve as additional landmarks. Relatively few insertions/deletions have to be invoked, and these can be placed with a high degree of certainty.

S segments are more restricted in occurrence, being confined more or less to specific lineages, although they are not necessarily found within all SSU rRNA sequences of any one lineage, and may occasionally be found in more than one major group. Although conservation of primary sequence is lower than in the case of U segments, alignment of homologous S segments according to secondary structure is still quite reliable. The S regions in Fig. 1, for example, are highly conserved in primary sequence and potential secondary structure within the eubacterial/plastid/plant mitocondrial grouping, and even in a comparison of E. coli 16S and wheat mitochondrial 18S rRNA sequences, a one-to-one alignment (without any insertions/deletions) is possible within the C regions (which are 77% identical between these two rRNAs).

V segments differ markedly in length, primary sequence, and /or potential secondary structure, even within a given lineage, and may be

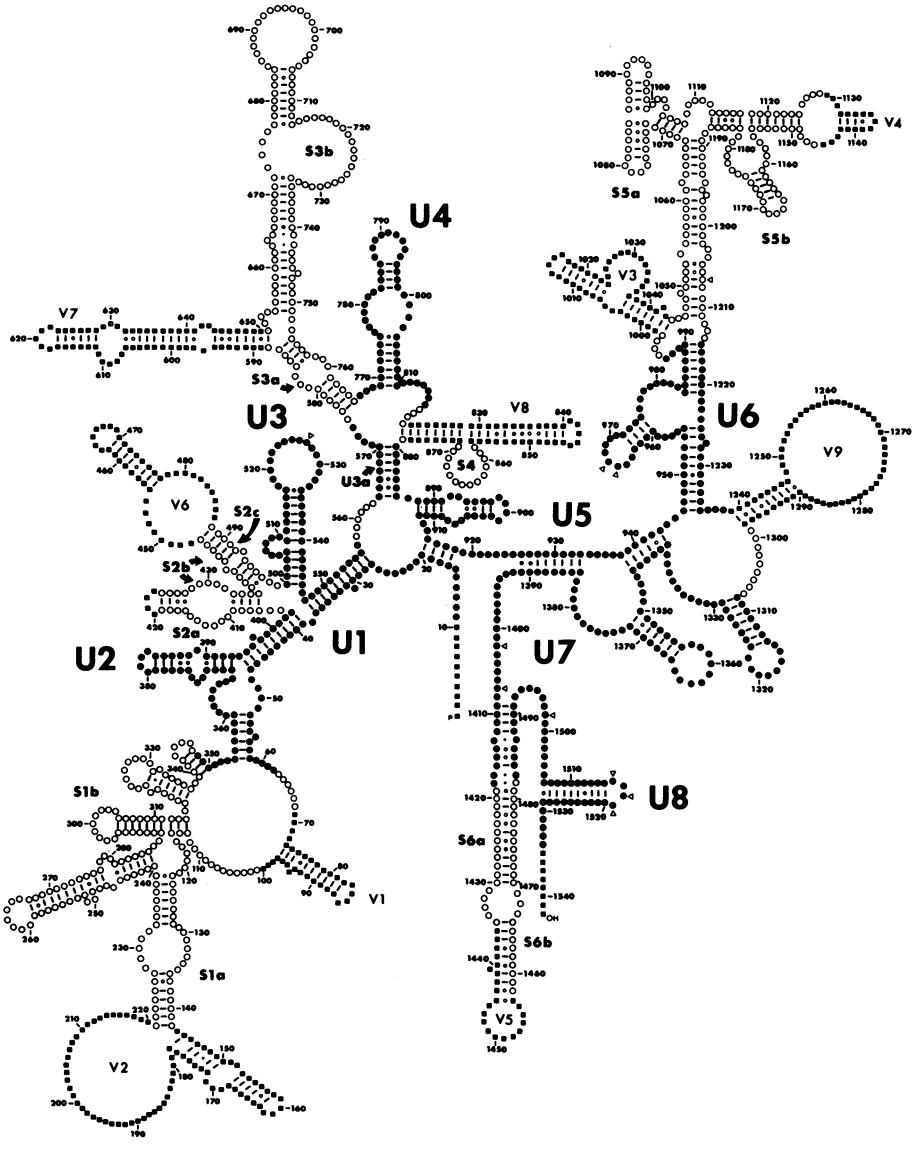


Figure 1. Representation of the universal ("U", ●), semi-conserved ("S", ○), and variable ("V", ▣) regions in SSU rRNAs. The figure is based on the eubacterial (E. coli) 16S rRNA secondary structure model recently published by Woese et al. [49]. Base pairs are denoted by — (G-C, A-U), ○ (G○U), and • (G•A), and the positions of modified nucleosides in E. coli 16S rRNA are indicated by ▽. U, S and V regions are defined in the text.

recognizably similar only in very closely related sequences. For example, V1 (Fig. 1) is the same length (32 nucleotides) and sequence in E. coli [15,16] and Proteus vulgaris [17] 16S rRNAs, but is a different length (14 nucleotides) and sequence in Anacystis nidulans 16S rRNA [18]. Between E. coli 16S and wheat mitochondrial 18S rRNAs, the V regions either differ markedly in length and/or potential secondary structure (V1-V6) or show low conservation of primary sequence (25-50%) where there is a reasonable correspondence in potential secondary structure (V7-V9). Except in cases of very close relationship, therefore, it is not possible to produce meaningful alignments of the primary sequences of corresponding V regions.

Because of their ubiquitous distribution, the U segments are ideally suited for inferring global evolutionary connections among all organisms and organelles. The data set analyzed here consists of regions U1 - U8 (Fig. 1; we exclude U3a because of its short length), corresponding to a total of 519 residues in E. coli 16S rRNA (Table 1). A further 12 positions are considered to be deletions in the E. coli sequence relative to one or more other SSU rRNA sequences, so that 531 positions altogether are evaluated. To date, 20 complete SSU rRNA sequences (1 archaebacterial, 3 eubacterial, 4 plastid, 1 plant mitochondrial, 2 fungal mitochondrial, 4 mammalian mitochondrial, 5 eukaryotic nuclear) have been published, and these all contain the complete universal core. The primary references to the 16 sequences we use here, and the positions in each one corresponding to the individual U segments, are compiled in Table 1.

To confirm branching order within the eubacterial lineage, an additional data set encompassing segments S3a + V7 + S3b (see Fig. 1) was compiled from 3 eubacterial and 4 organellar (3 plastid and 1 plant mitochondrial) sequences. This data set corresponds to 191 consecutive positions in E. coli 16S rRNA (residues 575-765) and exhibits considerably less primary sequence conservation than the U segments. For example, this region is only 60% identical in primary sequence between E. coli 16S and wheat mitochondrial 18S rRNAs, compared with 87.5% sequence identity between these two molecules in the combined U regions.

The Methodology

Our goal is to find the minimal mutation phylogeny -- the most parsimonious evolutionary explanation of the data sequences. However, no method exists, or is likely to be developed, that can unequivocally determine the minimal tree with reasonable computational effort when many sequences (e.g., the 16 considered here) are involved. Nonetheless, we can use

Table I. Positions of universal (U) regions in small subunit rRNA sequences

Sequence[a]			U1	U2	U3	U4	U5	U6	U7	U8
HVO	A	16S	6- 57	327- 382	437- 495	705- 756	823- 936	1163-1186	1256-1366	1415-1466
ECO	E	16S	11- 62	348- 403	500- 557	766- 817	880- 993	1215-1238	1308-1418	1483-1534
PVU	E	16S	11- 62	347- 402	499- 556	764- 815	878- 991	1217-1240	1310-1420	1485-1536
ANI	E	16S	11- 62	315- 368	443- 517	709- 760	824- 937	1158-1181	1251-1361	1428-1479
MAI	C	16S	11- 62	318- 373	447- 504	713- 764	828- 941	1161-1184	1254-1365	1431-1482
TOB	C	16S	12- 63	319- 372	446- 503	712- 763	826- 939	1160-1183	1253-1364	1427-1478
CRE	C	16S	13- 64	326- 379	453- 510	720- 771	826- 937	1159-1182	1256-1367	1434-1485
EGR	C	16S	13- 64	342- 397	471- 528	737- 787	844- 957	1167-1190	1259-1370	1435-1486
WHT	M	18S	19- 70	416- 471	547- 604	811- 862	922-1035	1585-1608	1677-1787	1899-1950
SCE	M	15S	432- 483	766- 821	1062-1120	1279-1330	1393-1506	1681-1704	1810-1920	2009-2059
ALS	M	15S	11- 62	308- 362	453- 511	688- 739	811- 924	1085-1108	1178-1288	1383-1434
HUM	M	12S	653- 702	811- 866	872- 929	1054-1105	1133-1247	1345-1367	1393-1502	1548-1599
MOU	M	12S	198- 247	357- 412	416- 473	597- 648	677- 791	893- 915	942-1050	1096-1147
RAT	M	12S	1396-1347	1239-1184	1179-1122	996- 945	917- 803	706- 684	656- 548	501- 450
SCE	N	18S	6- 58	420- 474	546- 604	976-1017	1091-1205	1437-1460	1534-1644	1735-1786
XEN	N	18S	6- 57	433- 487	560- 618	995-1046	1122-1236	1467-1490	1566-1676	1771-1822

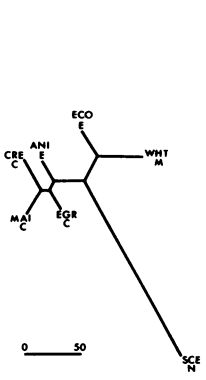
[a]Abbreviations and primary references: A, archaebacterial; E, eubacterial; C, chloroplast; M, mitochondrial; N, nuclear; HVO, Halobacterium volcanii [40]; ECO, Escherichia coli [15,16]; PVU, Proteus vulgaris [17]; ANI, Anacystis nidulans [18]; MAI, maize, Zea mays [41]; TOB, tobacco, Nicotiana tabacum [42]; CRE, Chlamydomonas reinhardtii (C, [43]); EGR, Euglena gracilis (C, [44]); WHT, wheat, Triticum aestivum (M, [12]); SCE, yeast, Saccharomyces cerevisiae (M, [45]; N, [46]); ALS, Aspergillus nidulans (M, [47]); HUM, human, Homo sapiens (M, [11]); MOU, mouse, Mus musculus (M, [29]); RAT, rat, Rattus norvegicus (M, [30]); XEN, Xenopus laevis (N, [48]). The numbering of the ECO E 16S sequence follows that of Woese et al. [49]. A 10-nucleotide stretch in U-4, missed during the original sequencing of the SCE N 18S sequence [46] (cf. [48]), has been taken from the revised sequence [50]. This stretch corresponds to positions 986-995 in the revised sequence (posiions 775-784 in the ECO E 16S sequence; Fig. 1) and inserts between positions 985 and 986 of the SCE N 18S sequence, as originally published and cited in the above table.

established biological facts to impose constraints on the set of possible solutions to the problem, and hence reduce it to the equivalent of finding the minimal tree for a much smaller, and manageable, set of sequences.

The basic method is straightforward. A number "N" of aligned sequences, each containing "n" nucleotides, is provided as input to a computer program. The program generates all possible phylogenetic histories (trees) of "N" sequences, and for each tree it calculates the most parsimonious assignment of ancestral sequences according to the method of Sankoff and Rousseau [19]. It then outputs the best, or best few, solutions.

While the time requirements for ancestral sequence construction grow only linearly with "n" (the length of the sequences) and linearly as well with "N" (the number of sequences), the total number of trees that have to be generated and evaluated grows exponentially with "N". Thus, it is not feasible to calculate solutions even for N=9 or 10 if "n" is at all large. To circumvent this difficulty, we impose biologically unexceptionable constraints, for example, the constraint that all eukaryotes (i.e., the nuclear component) group more closely together than any of them does with any other organism. The basic program can then be adjusted so that when it generates all possible trees, it disregards those not satisfying one or more of the imposed constraints. Thus, when N=10-15, the idea is to impose enough constraints of this type to reduce the computational effort to the equivalent of that which would be required if N=7.

There are not sufficient general constraints of the form "all eukaryotes group together" to be able to achieve the necessary reduction in computation for a data set as large as ours, and so it is necessary to add more detail to the constraints. For example, within an assigned grouping of various chloroplasts and blue-green algae, we would like to be able to assert (if true) that chloroplasts are more similar among themselves than any are to any cyanobacterium, and even to state which chloroplasts are most closely related. To derive this information, we carry out some preliminary runs of the program, including only those sequences within the group in question, plus one or a few external sequences, under the hypothesis (again fairly unexceptionable) that the external sequences contribute relatively little information about the relationships among the within-group sequences, except for how the group is connected to the larger phylogeny. This latter information is determined by the positioning of the one or more external sequences in the preliminary run. It is important to note that this methodology in itself provides no information on the location of the root (earliest ancestor) in a phylogeny.



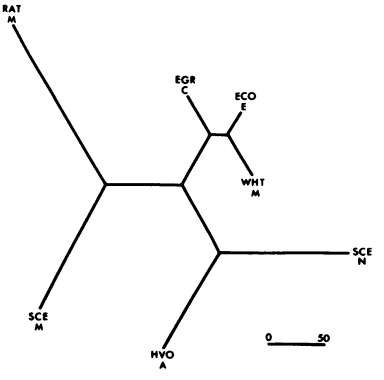


Figure 2 (left). Schematic representation of minimal mutation tree demonstrating phylogenetic relationships within the chloroplast/cyanobacterium grouping. Branch lengths are proportional to the weighted mutational distance between two end points, as indicated by the scale. Organisms/organelles are designated as in Table I.
Figure 3 (right). Phylogenetic tree showing the relationships among representatives of the five identified genome types (archaebacterial, A; eubacterial, E; nuclear, N; chloroplast, C; mitochondrial, M). Further details are given in Fig. 3 and Table I.

RESULTS

To determine relationships within the chloroplast-cyanobacterium grouping, a minimal mutation tree was generated, situating the root of this subtree with respect to eubacteria (E. coli) and the nuclear component of eukaryotes (yeast). The wheat mitochondrial sequence was also included, in view of its demonstrated strong resemblance to eubacterial/plastid sequences [12]. The results (Fig. 2) show Anacystis nidulans branching off before the chloroplasts diverge from one another. Among the next best trees was one indicating that Anacystis nidulans and Euglena chloroplast had a common evolutionary history not shared by the other chloroplasts; however, this analysis was decidedly inferior to the one depicted in Fig. 2, requiring 402 vs. 398 mutations. Notably, wheat mitochondrion groups with E. coli in the subtree of Fig. 2, and this grouping recurred in all trees that have close-to-optimal mutational counts.

A second preliminary analysis was designed to probe relationships among mitochondrial sequences, and to situate the archaebacterium (Halobacterium volcanni) with respect to the eukaryotic nuclear and eubacterial branches. The resulting tree (Fig. 3) shows the halobacterial sequence decidedly closer, measured along the branches of the tree, to the eukaryotic (yeast) nuclear sequence (209 differences) than to any of the other sequences (257-395

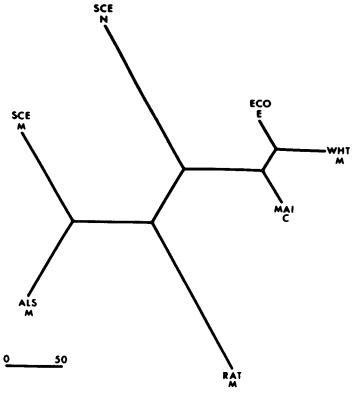


Figure 4. Phylogenetic tree illustrating relationships within the eubacterial / plastid / mitochondrial lineage. Further details are given in Fig. 2 and Table I.

differences), an affinity that was consistently observed in subsequent analyses involving other combinations of sequences. Note, however, that this result does not preclude the possibility that the root of the tree should be located on the branch between the earliest eukaryotic (nuclear) node and the origin of the halobacter branch. Again, the wheat mitochondrial sequence groups with the E. coli sequence, clearly within the eubacterial lineage (cf. position of the Euglena chloroplast sequence); it does not, however, group with the other (fungal and animal) mitochondrial sequences.

To confirm the relationship among the mitochondrial sequences, we replaced the halobacterial and Euglena chloroplast sequences with those of Aspergillus nidulans mitochondrion and maize chloroplast. This tree (Fig. 4) depicts the close affinity between the fungal mitochondrial sequences, and the continued separate grouping of the wheat mitochondrial sequence with those of E. coli and maize plastid.

Although other combinations did not always yield consistent phylogenies, the following generalities obtained: the archaebacterial sequence was always closer to the eukaryotic nuclear-encoded sequences than to any others; the mitochondrial sequences (other than that of wheat) consistently clustered together; and the relative positions of the E. coli, Anacystis nidulans, and wheat mitochondrial sequences varied. Thus, in formulating the constraints for our final assessment of trees containing all sequences simultaneously, we left the wheat mitochondrial, Anacystis, and Euglena chloroplast sequences free to attach themselves wherever they fit best. We did, however, feel justified in grouping the remaining mitochondrial sequences as in Fig. 3 and 4. The remaining three chloroplast sequences were grouped together, as determined earlier (Fig. 2), with that of tobacco placed on a branch with the

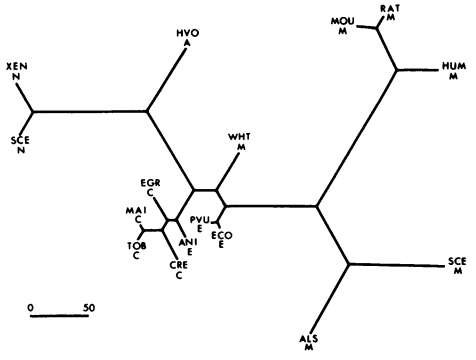


Figure 5. Global phylogenetic tree showing evolutionary relationships among 16 organisms/organelles spanning the five known genome types. Further details are given in Fig. 2 and Table I.

very similar maize chloroplast sequence, while the archaebacterial sequence was grouped with the eukaryotic nuclear sequences. Finally, the P. vulgaris sequence was grouped with the highly similar E. coli one. Thus, the seven mutually unconstrained lines in our final 16-sequence analysis were: (i) maize, tobacco, and Chlamydomonas chloroplasts; (ii) Euglena chloroplast; (iii) Anacystis nidulans; (iv) other eubacteria (E. coli and P. vulgaris); (v) eukaryotic nuclear component (yeast and Xenopus) + archaebacterium; (vi) wheat mitochondrion; (vii) other mitochondria (Aspergillus, yeast, rat, mouse, man).

In the final tree (Fig. 5), it can be seen that the branching between the animal and fungal mitochondria involves the largest distances and least precision. To further verify this grouping, we repeated the analysis, removing the constraint that these two groups be closely related. To keep the computational effort within reasonable limits, we added the constraint that Euglena chloroplast group with the other chloroplasts, an affinity clearly justified by the analysis of Fig. 2. This computation gave a tree (not shown) having a topology identical to that of Fig. 5.

DISCUSSION

Our data set differs from those of two other groups that have used SSU rRNA sequences for phylogenetic analysis. Given the additional insights provided by the study of Stiegler et al. [9], we include three regions (U2, U4, and part of U7) not considered by Küntzel and Köchel [20] in their analysis, and we exclude regions used by them extending beyond the boundaries (as defined in Fig. 1 and Table 1) of U1, U3, U5, U6, and U8. In a more recent study, McCarroll et al. [21] utilized either complete SSU rRNA sequences, in which case many insertions/deletions had to be assumed in aligning highly divergent regions of primary structure, or a combination of

semi-conserved and conserved regions, which included the "universal" regions we use (except U1 and U8) but extended beyond these. Since our interest was in the very earliest evolutionary events, we judged it preferable to use only those regions of most stable sequence in SSU rRNA. Because our analysis thus focuses strictly on a ubiquitous, highly conserved core of primary and secondary structure, it is likely that little or no change in primary sequence alignment will be necessary to accommodate additional diverse SSU rRNA sequences, regardless of their phylogenetic position. It should therefore be possible to incorporate new results readily into existing trees determined by the methodology presented here.

There are no topological inconsistencies between the tree in Fig. 5 and those in [21]. There are some metric differences, however, notably in the placement of the archaebacterial sequence. This is likely due to two factors: (i) the use by McCarroll *et al.* [21] of semi-conserved and variable regions, which show a relatively greater proximity between eubacteria and the halobacterium than between the latter and eukaryotes; and (ii) their optimization criterion for trees, which matches tree path lengths between sequences to the actual distances between pairs of sequences. Hence, branches in the middle of the tree are counted inordinately often, since they are on many paths, in contrast to branches at the extremities, which are on relatively few paths.

Although one might expect some variation as new data are added, there are a number of reasons for believing that the global tree presented in Fig. 5 is basically correct. First, this particular solution constitutes the most parsimonious interpretaion of the data and is generally consistent with the results of McCarroll *et al.* [21], who used different criteria and a different data set. Second, when the data are input under altered constraints, the same solution is obtained. And third, application of the same treeing method to the S3a + V7 + S4b data set for eubacterial/plastid/plant mitochondrial sequences yields a phylogeny (not shown) fully consistent with that obtained when the U data set from these sequences is analyzed.

The set of SSU rRNA sequences analyzed here includes two new ones, those of the cyanobacterium, *Anacystis nidulans* [18], and the mitochondrion of wheat (*Triticum aestivum*, a higher plant) [12], that are of prime importance in evaluating the question of the evolutionary origin of organelles. A major conclusion of our study is that all published chloroplast and mitochondrial sequences cluster within the eubacterial lineage, reinforcing the concept of an endosymbiotic, eubacterial, evolutionary origin for plastids and

mitochondria [2,10]. However, in agreement with much existing data, plastids and mitochondria appear to have been derived from distinct types of eubacteria [2,22], since the plastid sequences branch specifically with the Anacystis sequence, whereas the mitochondrial sequences branch with those of E. coli and P. vulgaris.

Within the plastid/cyanobacterial lineage, Euglena chloroplast branches early, whereas Chlamydomonas and higher plant chloroplasts diverge later from a common ancestor. This order is consistent with the fact that rRNA gene organization in Euglena chloroplast DNA is different from that in Chlamydomonas and maize and tobacco chloroplast DNAs, with tandemly repeated rRNA genes in the former but inverted rRNA repeats in the latter group [23]. Of particular note in this lineage, however, is the fact that the Anacystis branch is clearly the most ancient of all, diverging before the plastid sequences separate from one another. This result has been confirmed by analysis of the S3a + V7 + S3b data set (not shown). This phylogeny differs from ones based on T_1 oligonucleotide catalogue data [24,25], and also from one we have derived from 5S rRNA sequences (data not shown), both of which show Euglena chloroplast to be the most ancient line of descent. While the present analysis clearly supports a monophyletic rather than polyphyletic origin of plastids, confirmation of this conclusion will require a larger and more broadly representative set of SSU rRNA sequences (including additional diverse cyanobacterial ones). The contrasting phylogenies obtained for the plastid/cyanobacterial grouping by us and by others most likely reflect differences in the data sets used and/or the method of data manipulation, and provide an example of the caution that must be exercised in such analyses.

In all trees in which both appear, fungal and animal mitochondria branch together, suggesting they are of monophyletic origin. This conclusion is in agreement with the results of McCarroll [21], but contrasts with the results of Küntzel and Köchel [20], whose analysis suggested a biphyletic origin of animal and fungal mitochondria. A monophyletic origin is consistent with the simpler genetic codes (expanded codon recognition patterns) shared by animal and fungal mitochondria, as well as their common use of UGA to code for tryptophan (reviewed in [2,26]). On the other hand, the separate branching of plant (wheat) mitochondria is suggestive of a separate evolutionary origin of plant and animal/fungal mitochondria, and this view is supported by observations strongly suggesting that UGA is not used as a tryptophan codon in plant mitochondria [27,28]. It remains to be seen whether the plant mitochondrial branch will eventually coalesce with the fungal/animal

mitochondrial branch as additional SSU rRNA sequences become available and can be included in the analysis. At present, the persistent separation of these two branches is striking. The emerging phylogeny of mitochondria has an interesting parallel with 5S rRNA-based phylogenies of eukaryotic nuclear genomes [5,6], which show an early divergence within the nuclear lineage leading to plants and algae on the one hand, and to fungi, protozoa, and metazoa on the other.

Another notable feature of Fig. 5 is the fact that animal and fungal mitochondrial sequences have diverged to a much greater extent than the wheat mitochondrial sequence since their separation from a remote common ancestor. Apparently, plant mitochondrial DNA does not sustain the unusually high rate of sequence divergence that is characteristic of fungal and especially animal mitochondrial genomes (cf. [26]). (As one example of the faster evolutionary clock in animal mitochondria, we note that the mitochondrial SSU rRNA genes of mouse [29] and rat [30] differ at 29 out of 515 positions (5.6%) within the universal core, whereas their nuclear-encoded counterparts [31,32] are completely identical within the core in these two rodents.)

Our results bear on the question of whether the simplified codon recognition pattern of animal and fungal mitochondria is a "frozen relic" of the proposed archetypal code [33] that is supposed to have preceded the universal code, or whether it represents a reversion from the codon-anticodon interactions of the universal code to a less complex set of interactions in the immediate common ancestor of fungal and animal mitochondria. The phylogeny determined here strongly suggests the latter. Starting at the node from which nuclear, archaebacterial, and eubacterial lineages diverge (Fig. 5), and proceeding into the eubacterial lineage, it can be seen that animal/fungal mitochondria branch after plastids/Anacystis nidulans, and from the same node as E. coli and P. vulgaris. In view of the fact that cytoplasmic, eubacterial, archaebacterial [34] and plastid [35] translation systems all appear to utilize the universal code, this must have been the code used by the last common ancestor of all organisms (cf. [36]).

In animal and fungal mitochondria, an expanded codon recognition pattern is correlated with the lack of modification of a uridine residue in the Wobble position of certain tRNAs [37,38]; thus, loss of genetic information encoding modification enzymes could represent an important mechanism in the reversion (in terms of codon-anticodon interactions) of the universal code to a simpler one. Such loss may have accompanied the evolutionary "streamlining" of fungal and particularly animal mitochondrial genomes, suggesting that tRNAs in the

progenote contained modifications that were subsequently eliminated from fungal and animal mitochondrial tRNAs. This is in line with a proposal by Reanney [39] that the lesser content and variety of modified nucleosides in eubacterial tRNAs, compared to their eukaryotic cytoplasmic (nuclear-encoded) counterparts, is a consequence of the elimination (within the eubacterial lineage) of genes for all modifying enzymes other than those obligatorily needed for translation, a process that we would suggest has been carried to an extreme in animal and fungal mitochondria. In view of the branching position of wheat mitochondria in the phylogeny of Fig. 5, determination of the modification status of plant mitochondrial tRNAs will be important in evaluating this proposal.

Finally, the order of divergence of the archaebacterial, nuclear, and eubacterial lineages remains unsettled. Although in preliminary trees the archaebacterial sequence was found substantially closer to its nuclear than its eubacterial counterpart, in the final tree (Fig. 5) the H. volcanii universal core is about equidistant from those of X. laevis nuclear-encoded 18S rRNA (197 mutations) and E. coli 16S rRNA (206 mutations), but considerably closer to each of the latter than they are to one another (271 mutations between X. laevis nuclear and E. coli cores). The final tree (Fig. 5) supports the view [21,40] that archaebacterial, eukaryotic nuclear, and eubacterial SSU rRNA sequences define three separate lineages, with the archaebacterial universal core being the closest of the three to the ancestral universal core common to all.

ACKNOWLEDGEMENTS

We gratefully acknowledge the continuing financial support provided by the Natural Sciences and Engineering Research Council (N.S.E.R.C.) and Medical Research Council (M.R.C.) of Canada.

*To whom reprint requests should be sent

REFERENCES

[1] Woese, C.R. and Fox, G.E. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 5088-5090.
[2] Gray, M.W. and Doolittle, W.F. (1982) Microbiol. Rev. 46, 1-42.
[3] Cedergren, R.J., Sankoff, D., LaRue, B. and Grosjean, H. (1981) CRC Crit. Rev. Biochem. 11, 35-104.
[4] Sankoff, D., Cedergren, R.J. and McKay, W. (1982) Nucleic Acids Res. 10, 421-431.
[5] Huysmans, E., Dams, E., Vandenberghe, A. and De Wachter, R. (1983) Nucleic Acids Res. 11, 2871-2880.

[6] Küntzel, H., Piechulla, B. and Hahn, U. (1983) Nucleic Acids Res. 11, 893-900.
[7] Leaver, C.J. and Gray, M.W. (1982) Annu. Rev. Plant Physiol. 33, 373-402.
[8] Cedergren, R.J. (1982) Can. J. Biochem. 60, 475-479.
[9] Stiegler, P., Carbon, P., Ebel, J.-P. and Ehresmann, C. (1981) Eur. J. Biochem. 120, 487-495.
[10] Gray, M.W. (1983) BioScience 11, 693-699.
[11] Eperon, I.C., Anderson, S. and Nierlich, D.P. (1980) Nature 286, 460-467.
[12] Spencer, D.F., Schnare, M.N. and Gray, M.W. (1984) Proc. Natl. Acad. Sci. U.S.A. 81, 493-497.
[13] Cox, R.A. and Kelly, J.M. (1982) Biochem. Soc. Symp. 47, 11-48.
[14] Maly, P. and Brimacombe, R. (1983) Nucleic Acids Res. 11, 7263-7286.
[15] Brosius, J., Palmer, M.L., Kennedy, P.J. and Noller, H.F. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 4801-4805.
[16] Carbon, P., Ehresmann, C., Ehresmann, B. and Ebel, J.-P. (1979) Eur. J. Biochem. 100, 399-410.
[17] Carbon, P., Ebel, J.P. and Ehresmann, C. (1981) Nucleic Acids Res. 9, 2325-2333.
[18] Tomioka, N. and Sugiura, M. (1983) Mol. Gen. Genet. 191, 46-50.
[19] Sankoff, D. and Rousseau, P. (1975) Math. Programming 9, 240-246.
[20] Küntzel, H. and Köchel, H.G. (1981) Nature 293, 751-755.
[21] McCarroll, R., Olsen, G.J., Stahl, Y.D., Woese, C.R. and Sogin, M.L. (1983) Biochemistry 22, 5858-5868.
[22] Doolittle, W.F. (1980) Trends in Biochem. Sci. 5, 146-149.
[23] Bohnert, H.J., Crouse, E.J. and Schmitt, J.M. (1982) in Parthier, B., and Boulter, D. (eds.) Encyclopedia of Plant Physiology, Vol. 14B, Springer, Heidelberg, pp. 475-530.
[24] Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Leuhrsen, K.R., Chen, K.N. and Woese, C.R. (1980) Science 209, 457-463.
[25] Doolittle, W.F. and Bonen, L. (1981) Ann. N.Y. Acad. Sci. 361, 248-256.
[26] Gray, M.W. (1982) Can. J. Biochem. 60, 157-171.
[27] Fox, T.D. and Leaver, C.J. (1981) Cell 26, 315-323.
[28] Hiesel, R. and Brennicke, A. (1983) EMBO J. 2, 2173-2178.
[29] Van Etten, R.A., Walberg, M.W. and Clayton, D.A. (1980) Cell 22, 157-170.
[30] Kobayashi, M., Seki, T., Yaginuma, K. and Koike, K. (1981) Gene 16, 297-307.
[31] Raynal, F., Michot, B. and Bachellerie, J.-P. (1984) FEBS Lett. 167, 263-268.
[32] Torczynski, R., Bollon, A.P. and Fuke, M. (1983) Nucleic Acids Res. 11, 4879-4890.
[33] Jukes, T.H. (1983) J. Mol. Evol. 19, 219-225.
[34] Dunn, R., McCoy, J., Simsek, M., Majumdar, A., Chang, S.H., RajBhandary, U.L. and Khorana, H.G. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 6744-6748.
[35] McIntosh, L., Poulsen, C. and Bogorad, L. (1980) Nature 288, 556-560.
[36] Crick, F.H.C. (1968) J. Mol. Biol 38, 367-379.
[37] Heckman, J.E., Sarnoff, J., Alzner-DeWeerd, B., Yin, S. and RajBhandary, U.L. (1980) Proc. Natl. Acad. Sci. U.S.A. 77, 3159-3163.
[38] Roe, B.A., Wong, J.F.H., Chen, E.Y., Armstrong, P.W., Stankiewicz, A., Man, D.-P. and McDonough, J. (1982) in Slonimski, P., Borst, P. and Attardi, G. (eds.) Mitochondrial Genes, Cold Spring Harbor Laboratory,

New York, pp. 45-49.
[39] Reanney, D.C. (1974) J. Theor. Biol. 48, 243-251.
[40] Gupta, R., Lanter, J.M. and Woese, C.R. (1983) Science 221, 656-659.
[41] Schwarz, Z. and Kössel, H. (1980) Nature 283, 739-742.
[42] Tohdoh, N. and Sugiura, M. (1982) Gene 17, 213-218.
[43] Dron, M., Rahire, M. and Rochaix, J.-D. (1982) Nucleic Acids Res. 10, 7609-7620.
[44] Graf, L., Roux, E., Stutz, E. and Kössel, H. (1982) Nucleic Acids Res., 10, 6369-6381.
[45] Sor, F. and Fukuhara, H. (1980) C. R. Acad. Sc. Paris Série D. 291, 933-936.
[46] Rubtsov, P.M., Musakhanov, M.M., Zakharyev, V.M., Krayev, A.S., Skryabin, K.G. and Bayev, A.A. (1980) Nucleic Acids Res. 8, 5779-5794.
[47] Köchel, H.G. and Küntzel, H. (1981) Nucleic Acids Res. 9, 5689-5696.
[48] Salim, M. and Maden, B.E.H. (1981) Nature 291, 205-208.
[49] Woese, C.R., Gutell, R., Gupta, R. and Noller, H.F. (1983) Microbiol. Rev. 47, 621-669.
[50] Mankin, A.S., Kopylov, A.M. and Bogdanov, A.A. (1981) FEBS Lett. 134, 11-14.