J Mol Evol (1988) 28:98–112

Journal of Molecular Evolution

The Evolutionary Relationships among Known Life Forms

Robert Cedergren,[1] Michael W. Gray,[3] Yvon Abel,[2] and David Sankoff[2]

[1] Département de biochimie, [2] Centre de recherches mathématiques, Université de Montréal, Montréal, Québec H3C 3J7, Canada
[3] Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Summary. Sequences of small subunit (SSU) and large subunit (LSU) ribosomal RNA genes from archaebacteria, eubacteria, and the nucleus, chloroplasts, and mitochondria of eukaryotes have been compared in order to identify the most conservative positions. Aligned sets of these positions for both SSU and LSU rRNA have been used to generate tree diagrams relating the source organisms/organelles. Branching patterns were evaluated using the statistical bootstrapping technique. The resulting SSU and LSU trees are remarkably congruent and show a high degree of similarity with those based on alternative data sets and/or generated by different techniques. In addition to providing insights into the evolution of prokaryotic and eukaryotic (nuclear) lineages, the analysis reported here provides, for the first time, an extensive phylogeny of the mitochondrial lineage.

Key words: rRNA — Evolution — Sequence comparison — Parsimony — Bootstrap

Introduction

Determining the evolutionary relationships among life forms involves the compilation and analysis of common and unique traits. A set of organisms that share many features may justifiably be considered to have arisen from a more recent common ancestor than those that share only a limited number of these features. Classically, phylogenetic analysis has been performed on what may be called phenotypical data, i.e., morphological, chemical, metabolic, or behavioral, and has given rise to the discipline of numerical taxonomy in order to weigh, compare, and rationalize these data (Sokal and Sneath 1963; Sneath and Sokal 1973). Beginning over 20 years ago (Zuckerkandl and Pauling 1965; Fitch and Margoliash 1967), the advent of protein and nucleic acid sequences (genotypical data) provided evolutionists with a new type of database and a further stimulus to study the phylogeny of organisms. Although this sequence-based taxonomy greatly improves the resolution of inferred organismal relationships, it is not without some shortcomings. These include the assumption that one gene sequence is the only such sequence in a population and that it is a faithful representative of the entire genome (Rothschild et al. 1986).

The relative value of different data for phylogeny determinations is a rather complicated and disputed question (cf. Ruvolo and Smith 1986). It is our opinion, however, that the ideal data set would come from sequences of entire genomes. Currently, data of this scope are available for only a few bacteriophage, viral, and organellar (mitochondrial and plastid) genomes, although with continuing rapid developments in DNA sequencing, complete genome sequences of prokaryotic organisms will undoubtedly become available in the near future.

In the absence of sequences of entire genomes, considerable effort has been invested in the analysis of single gene sequences, such as those of proteins (Nei and Kowhn 1983), analyzed in the form of the amino acid or the nucleotide alphabet, 5S RNA (Huysmans and de Wachter 1986b; Hori and Osawa 1987), small subunit ribosomal RNA (McCarroll et al. 1983; Gray et al. 1984; Pace et al. 1986; Sogin et al. 1986a; Field et al. 1988), and transfer RNA (Cedergren et al. 1981). Early work using sequences

Offprint requests to: R. Cedergren or D. Sankoff

established much of the methodology and showed a remarkable similarity between morphometric (phenotype-derived) and sequence-based (genotype-derived) phylogenies (Dayhoff 1972). More recently, and particularly among prokaryotes, unexpected groupings and divisions have been observed. RNA sequence data have thus separated prokaryotes into the archaebacteria and eubacteria, which together with the eukaryotic nucleus define three primary lines of descent of known life on earth (Woese and Fox 1977a). Also, relationships among eubacteria have been completely redefined, earlier phylogenies having been based too heavily on cellular metabolism (Woese 1987). In the light of molecular comparisons, for example, photosynthesis is seen to be a property of organisms in a number of distinct phyla (Woese et al. 1985).

Over the past several years, we have worked extensively with the sequences of small subunit ribosomal RNA (SSU rRNA) genes (Gray et al. 1984). We present here a comparison of the phylogenetic tree inferred from a greatly expanded SSU rRNA database with a parallel tree based on a large number of sequences corresponding to the large subunit rRNA (LSU rRNA). Our analysis has been made possible by major methodological improvements, including the use of a refined version of our previous algorithm, its implementation on a CRAY supercomputer, and the addition of statistical criteria to evaluate the significance of various aspects of tree topologies.

Database

Given our goal of determining a global phylogeny comprising the three primary lines mentioned above and including the eukaryotic organelles (mitochondria and chloroplasts), few gene sequence databases fulfill the requirement that the gene in question be encoded in all of the genomes under consideration. Because 5S RNA genes are not present in mitochondrial genomes other than those of plants (Spencer et al. 1981), and because it appears that some mtDNAs do not contain a full set of tRNA genes (Suyama 1986; Gray and Boer 1988), only the LSU and SSU rRNA genes appear to be ubiquitous. Moreover, as we and others have previously noted, tRNA is too short to determine the desired global phylogeny (Gray et al. 1984). However, even SSU and LSU rRNA sequences are not perfect: an alignment problem is posed by their variable lengths, resulting from insertions or deletions during the evolutionary history of the genes in different taxa. Alignment of nucleotide sequences is not trivial, because unlike proteins, these informational macromolecules are constructed from only four monomeric units, which often creates many competing plausible alignments (Cedergren et al. 1981). Improper alignment may lead to very different, if not false, tree topologies (Feng and Doolittle 1987). In order to avoid this possible source of error, our database consists solely of sequence segments that correspond to the most highly conserved portions of the RNAs. This conservatism is evaluated using both primary and secondary structure determinants (Gray et al. 1984). It is these selected regions that correspond to the highly conserved "core" of functional SSU and LSU rRNA molecules.

The data set in the case of these two rRNAs therefore contains very few insertions and deletions; those that are included are easily dealt with because both the primary and secondary structure are available to guide alignment. An added data management advantage of using selected conserved regions is that the addition of new sequences is unlikely to significantly affect previously aligned sequences, which is not the case when less highly conserved regions are used.

Figure 1 shows representations of the secondary structures of *Escherichia coli* SSU (16S) and LSU (23S) rRNAs, with core regions that constitute our SSU and LSU databases being shaded. The figure legend gives the specific *E. coli* sequence coordinates of these universal regions. Table 1 is a listing of the organisms and organelles for which complete rRNA sequences are known (76 SSU, 41 LSU), together with the appropriate literature citation(s). For each SSU and LSU sequence, the core secondary structure was constructed and the sequence positions corresponding to these indicated in Fig. 1 were selected and aligned, the alignment following naturally from the secondary structure. Although our selection eliminates from the analysis a substantial portion of the available data, we believe that this is more than compensated for by the unambiguous quality of the alignment of the retained portion. The entire database is available from the authors and was supplied to the reviewers.

Finally, in evaluating two independent data sets (i.e., LSU and SSU), as we do in the present analysis, we can assess the congruency of the two phylogenetic topologies as an internal check on our methodology. This can be considered a step toward the ultimate goal of determining phylogeny not on the basis of a single gene, but rather on the basis of the entire genome.

Phylogenetic Methodology

There are three types of purely data-analytic problems to be faced in phylogenetic inference from aligned sequences. The first problem is one of *va-*

Table 1. List of organisms

Organism	SSU references	LSU references
Archaebacteria		
Desulfurococcus mobilis	—	Leffers et al. 1987
Halobacterium cutirubrum	*	—
Halobacterium halobium	*	Mankin and Kagramanova 1986
Halococcus morrhua(e)	*	Leffers et al. 1987
Halobacterium volcanii	*	Woese, personal communication
Methanobacterium formicicum	*	—
Methanobacterium hungatei	*	—
Methanobacterium thermoautotrophicum	—	Leffers et al. 1987
Methanococcus vannielii	*	Jarsch and Böck 1985
Sulfolobus solfataricus	*	Woese, personal communication
Thermoproteus tenax	*	—
Chloroplasts		
Chlamydomonas eugametos (chlorophyte)	Lemieux, personal communication	Lemieux, personal communication
Chlorella ellipsoidea (chlorophyte)	—	Yamada and Shimaji 1987
Chlamydomonas reinhardtii (chlorophyte)	*	—
Euglena gracilis (euglenoid flagellate)	*	—
Zea mays (maize)	*	Edwards and Kössel 1981
Marchantia polymorpha (liverwort)	Ohyama et al. 1986	Ohyama et al. 1986
Nicotiana tabacum (tobacco)	*	Takaiwa and Sugiura 1982
Eubacteria		
Anacystis nidulans	*	Kumano et al. 1983; Douglas and Doolittle 1984
Agrobacterium tumefaciens	*	—
Bacteroides fragilis	*	—
Bacillus stearothermophilus	—	Kop et al. 1984
Bacillus subtilis	*	Green et al. 1985
Chlamydia psittaci	Weisburg et al. 1986	—
Desulfovibrio desulfuricans	*	—
Escherichia coli	*	Brosius et al. 1980, 1981
Flavobacterium heparinum	*	—
Heliobacterium chlorum	*	—
Mycoplasma capricolum	*	—
Mycococcus xanthus	*	—
Mycoplasma strain PG50	*	—
Pseudomonas testosteroni	*	—
Proteus vulgaris	*	—
Rochalimaea quintana	Weisburg et al. 1985	—
Mitochondria		
Animal		
Bos taurus (ox)	*	Anderson et al. 1982
Pan troglodytes (common chimpanzee)	Hixson and Brown 1986	—
Drosophila yakuba (fruit fly)	*	Clary and Wolstenholme 1985
Gorilla gorilla (lowland gorilla)	Hixson and Brown 1986	—
Homo sapiens (human)	*	Eperon et al. 1980
Locusta migratoria (locust)	—	Uhlenbusch et al. 1987
Aedes albopictus (mosquito)	—	HsuChen et al. 1984
Mus musculus (mouse)	*	Van Etten et al. 1980
Pan paniscus (pygmy chimpanzee)	Hixson and Brown 1986	—
Pongo pygmaeus (orangutan)	Hixson and Brown 1986	—
Rattus norvegicus (rat)	*	—
Xenopus laevis (frog)	*; Dunon-Bluteau and Brun 1986	—
Fungal		
Aspergillus nidulans	*	Netzker et al. 1982
Saccharomyces cerevisiae	*	Sor and Fukuhara 1983
Schizosaccharomyces pombe	Wolf, personal communication	Lang et al. 1987
Plant		
Zea mays (maize)	*	Dale et al. 1984
Oenothera berteriana (evening primrose)	*	Manna and Brennicke 1985
Glycine max (soybean)	Grabau 1985	—
Triticum aestivum (wheat)	*	D.F. Spencer, unpublished

Table 1. Continued

Organism	SSU references	LSU references
Protist		
Chlamydomonas reinhardtii (chlorophyte)	P.H. Boer, unpublished	P.H. Boer, unpublished
Paramecium primaurelia (ciliate)	*	Seilhamer et al. 1984
Paramecium tetraurelia (ciliate)	*	Seilhamer et al. 1984
Tetrahymena pyriformis (ciliate)	Schnare et al. 1986b	—
Nucleocytoplasmic		
Animal		
Artemia salina (brine shrimp)	*	—
Caenorhabditis elegans (nematode)	Ellis et al. 1986	Ellis et al. 1986
Homo sapiens (human)	*	Laudien Gonzalez et al. 1985
Mus musculus (mouse)	*	Hassouna et al. 1984
Rattus norvegicus (rat)	*	Chan et al. 1983; Hadjiolov et al. 1984
Xenopus laevis (frog)	*	Ware et al. 1983
Fungal		
Neurospora crassa	Sogin et al. 19[illegible]6b	—
Saccharomyces carlbergensis	—	Veldman et al. 1981
Saccharomyces cerevisiae	*	Georgiev et al. 1981
Plant		
Zea mays (maize)	*	—
Oryza sativa (rice)	*	Takaiwa et al. 1985
Glycine max (soybean)	*	—
Protist		
Acanthamoeba castellanii (amastigote amoeba)	Gunderson and Sogin 1986	—
Achlya bisexualis (oomycete)	Gunderson et al. 1987	—
Chlamydomonas reinhardtii (chlorophyte)	Gunderson et al. 1987	—
Crithidia fasciculata (trypanosoid flagellate)	Schnare et al. 1986a	Spencer et al. 1987
Dictyostelium discoideum (slime mold)	*	—
Euglena gracilis (euglenoid flagellate)	Sogin et al. 1986a	—
Euplote aediculatus (ciliate)	Sogin et al. 1986c	—
Ochromonas danica (chrysophyte)	Gunderson et al. 1987	—
Oxytricha nova (ciliate)	*	—
Paramecium tetraurelia (ciliate)	Sogin and Elwood 1986	—
Physarum polycephalum (slime mold)	—	Otsuka et al. 1983
Plasmodium berghei (sporozoan)	Gunderson et al. 1986	—
Prorocentrum micans (dinoflagellate)	Herzog and Maroteaux 1986	—
Stylonychia pustulata (ciliate)	*	—
Tetrahymena thermophila (ciliate)	*	—
Trypanosoma brucei (trypanosoid flagellate)	Sogin et al. 1986a	—
Vairimorpha necatrix (microsporidian)	Vossbrinck et al. 1987	—

Asterisks indicate that the sequence and reference are included in Huysmans and de Wachter (1986a) and dashes indicate that the sequence used is either unknown or, if known, was not in this study

lidity. What optimization criterion will best assure that we select the true evolutionary history out of all the myriad combinatorial possibilities: (1) a minimum of inferred mutation (parsimony); (2) maximum likelihood under some probabilistic model; or (3) least squares fit to a distance matrix? The second problem is computational *feasibility*. The optimality criteria to be satisfied in evolutionary inference lead to problems as difficult computationally as the NP-complete class of problems, if not worse (Sankoff 1987). The third problem is that of *reliability*. Assuming that all methods will give at least partially erroneous results at least some of the time, how can we determine which parts of a reconstructed phylogeny are most likely to be correct and which parts are only slightly better than, or even just as good as, one or more other configurations? Furthermore, methods for assessing the reliability or statistical meaningfulness of results may themselves require computationally expensive resampling schemes.

The choice of an optimality criterion has been the subject of much controversy (Fitch 1977; Farris 1983; Felsenstein 1983a; Lake 1987). Probabilistic models of sequence evolution lead naturally to maximum likelihood or least squares criteria, but little is known about the sensitivity of these methods to

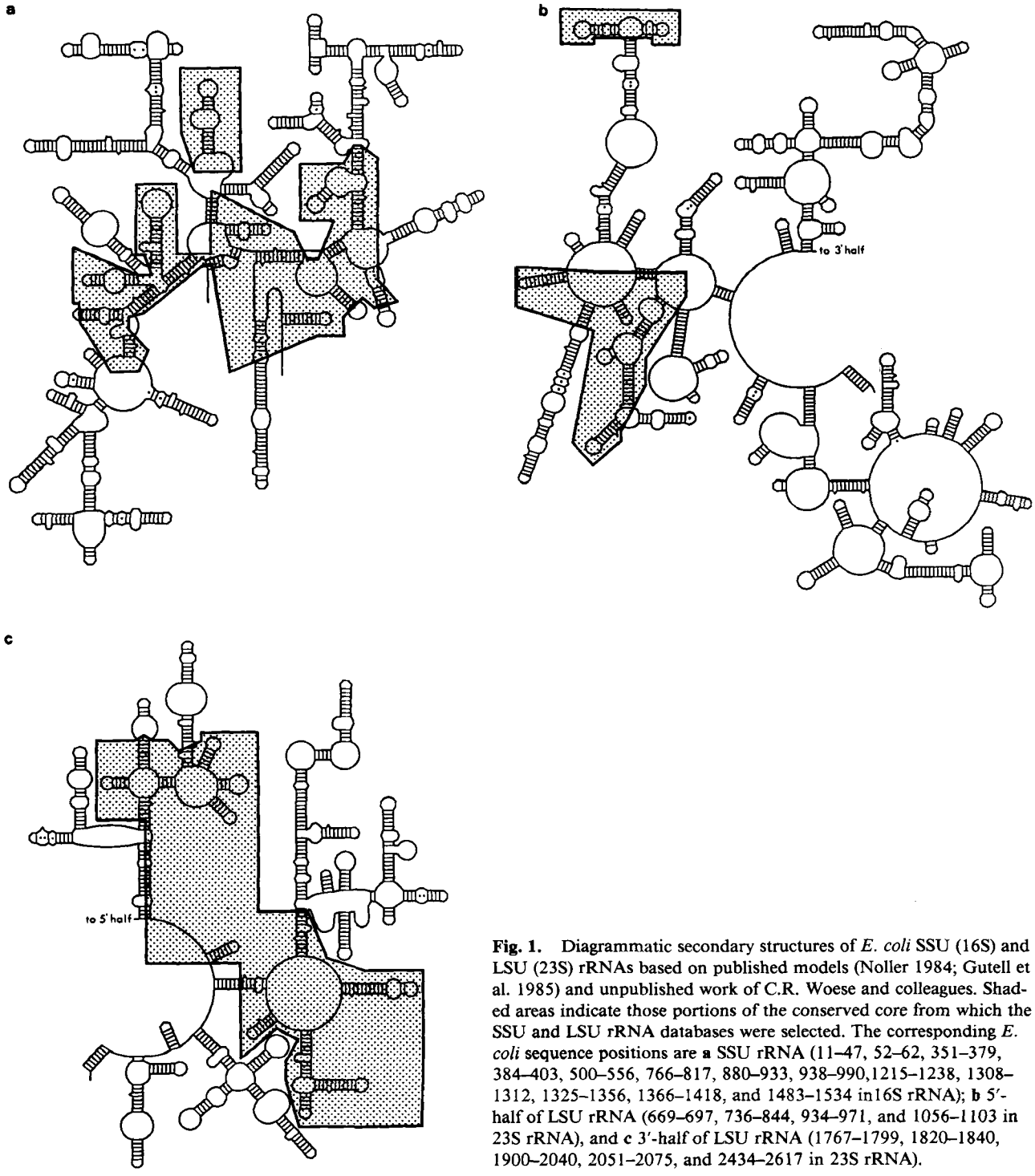


Fig. 1. Diagrammatic secondary structures of *E. coli* SSU (16S) and LSU (23S) rRNAs based on published models (Noller 1984; Gutell et al. 1985) and unpublished work of C.R. Woese and colleagues. Shaded areas indicate those portions of the conserved core from which the SSU and LSU rRNA databases were selected. The corresponding *E. coli* sequence positions are **a** SSU rRNA (11–47, 52–62, 351–379, 384–403, 500–556, 766–817, 880–933, 938–990,1215–1238, 1308–1312, 1325–1356, 1366–1418, and 1483–1534 in16S rRNA); **b** 5′-half of LSU rRNA (669–697, 736–844, 934–971, and 1056–1103 in 23S rRNA), and **c** 3′-half of LSU rRNA (1767–1799, 1820–1840, 1900–2040, 2051–2075, and 2434–2617 in 23S rRNA).

breakdown in such assumptions as constant mutation rates at given sequence positions or independence of mutation processes at different positions (Golding 1983). On the other hand, parsimony is model-free, which is sometimes a disadvantage and at other times an advantage. There is no general model, at least in the molecular evolution context, for generating data from an arbitrary phylogenetic tree, such that the most parsimonious tree tends to be the true tree. However, given that all such models necessarily contain highly restrictive assumptions that almost certainly break down repeatedly over the course of evolution, the fact that parsimony results in the most economical reconstruction of mutational history, with no assumptions and with the minimum of coincidence and unobserved changes, makes it highly attractive. Furthermore, with many data sets (those in which a particular tree configuration is most strongly inherent), parsimony tends to select the same tree as maximum likelihood does

(Felsenstein 1983a). There are trees, however, where a probabilistic model will generate data that will "fool" the parsimony criterion in a predictable way (Felsenstein 1983b; Lake 1987). These trees characteristically contain a number of pairs of evolutionarily closely related species where one member of each pair has undergone rapid evolution and the other has remained relatively unchanged. In some of these cases the parsimony criterion may mistakenly group all the rapidly evolving species together and all the conservative ones together. This "long branches attract" bias is a hazard in using the parsimony criterion. Thus, because the procedure we have developed is based on parsimony, we must take special measures to avoid artifactual grouping of long branches (M.W. Gray et al., unpublished; also see Olsen 1988). In cases where the topology is in doubt, we resort to a test of "invariants" that evaluates possible trees joining four species according to two measures (described by Cavender and Felsenstein 1987 as well as by Lake 1987) that are insensitive to branch length distortions.

The Molecular Cladistics Problem

First, a formal statement of the problem of finding the most parsimonious unrooted tree in molecular evolution studies is presented. We are given N aligned nucleotide sequences of length n. With RNA data, each position s(k, i) for k = 1, . . . , N and i = 1, . . . , n is drawn from the alphabet {A, C, G, U, –} where "–" represents a term deleted from the k-th sequence (or inserted in some other sequences). We wish to find the unrooted tree T* with N terminal nodes labeled 1, . . . , N that has minimal length (or cost). The length of any tree T is defined as the sum of r(i), over all sequence positions i = 1, . . . , n, where r(i) is the minimal number of branches in T with two different alphabet letters assigned to the nodes at each end, given that the N terminal nodes are assigned letters according to s(k, i), k = 1, . . . , N. For a *given* tree T, the r(i), as well as the optimal nonterminal node assignments, can be found in time proportional to N by dynamic programming (Fitch 1971; Hartigan 1973; Sankoff and Rousseau 1975; Sankoff and Cedergren 1983).

Turning to the question of feasibility, when branching from (inferred) ancestral nodes is allowed, all tree optimization problems become computationally intractable as the number of species increases. There are a variety of ways of confronting this fact of NP-completeness, and our approach is to combine several of these strategies, including the use of the supercomputer, in proportions particularly appropriate to our specific goal, that of inferring the panevolutionary tree based on rRNA sequence data.

For moderate N, say N = 10 or 11, "brute-force" methods can be used to solve the parsimony problem on a supercomputer, even with n = 100 or 1000. Our program examines all and only the $(2N - 5)!/2^{N-3}(N-3)!$ possible unrooted binary trees with N terminal nodes, because it is known that the minimum tree length must occur among binary trees. For each tree, the dynamic programming algorithm is executed to find r(i) for each of the n sequence positions i = 1, . . . , n, and that tree T*, which minimizes the sum of these r(i), is retained.

The use of the CRAY XMP with a completely vectorizable version of the dynamic programming core of our FORTRAN program increases computing speed by a factor of 100 over a CYBER 855. Because some of the computation for one tree is generally pertinent to the next one being examined, carry-over of partial results speeds up the search by another factor of about 4. Depending upon how good an initial "guess" at the best tree is, branch-and-bound techniques also speed up the search, at least by a factor of about 2, but generally by much more.

Local Optimization Using Temporary Constraints

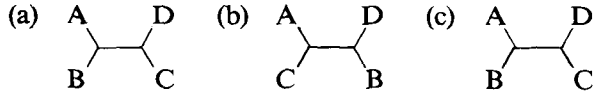
Efforts to push the technology of exhaustive searching to higher values of N will eventually reach the point of diminishing returns, if not at N = 11 or 12, then certainly at N = 15 or 20, barring lucky initial guesses in the branch-and-bound context. It is not the dynamic programming that is responsible for the complexity of the parsimony search, because this only takes time proportional to N, but rather the generation of an exponentially growing number of trees for increasing N.

Thus, for large numbers of organisms, in practice 50 or 100, we resort to the following iterative method whereby only a section of the tree is optimized (using the parsimony criterion) at a time, with the rest of the structure being temporarily constrained. At any stage, we have a currently best tree, T′. We then identify a connected fragment of T′, which itself is a small (N = 10) binary tree, t′, whose own terminal nodes include some nonterminal nodes (and possibly some terminal nodes) of T. The configuration of T′ outside of t′ is held fixed while we search for the most parsimonious tree configuration, t″, to replace t′. This uses basically the same method as the exhaustive search described in the previous section, but for each candidate small tree t being examined, the dynamic programming for each sequence position is carried out over the larger tree consisting of T′ without t′ but including t. The minimizing t, say t = u′, replaces t′, thus correcting T′ to a new current optimum, T″. Then another small tree, t″, is identified in T″, where t″ generally over-

laps somewhat with u', and the process is repeated until no fragment in the tree structure remains that can be improved. This, of course, may be only a local minimum, but is far more likely to be a global minimum than the results of other methods such as nearest-neighbor interchange (Moore et al. 1973), which is similar to our method but with a fragment size of only N = 4.

The Bootstrap

Concerning the reliability question, tree construction methods generally output binary, or fully resolved, trees where all nonterminal nodes (except the "root" if there is one), i.e., the inferred ancestral nodes, are at the intersection of three branches. (In rooted trees, each node, except the root, is connected by one branch to its immediate ancestor, and, except for terminal nodes, connected by two branches to its two descendant nodes. In unrooted trees, a branch has no specified ancestor–descendant orientation, so that there is no such distinction among the three branches meeting at a nonterminal node.) As mentioned above, however, the data may not really support all aspects of the branching structure equally well. For example, an inferred tree may be as in (a),



but the data may be equally consistent with (b). The optimization method is nevertheless constrained to pick one branching sequence, either (a) or (b). To represent the fact that we have no indication which pair among A, B, C, and D is most closely related, it would be preferable to select a "less-resolved" tree, such as (c). In other words, we would like to know which branches of an inferred tree are meaningful, and which branches we should delete from the tree (such as the interior branch not connected to any terminal nodes in our example). In our procedure, we do this systematically through a statistical technique called the "bootstrap" (Diaconis and Efron 1983; Felsenstein 1985).

Although our exhaustive evaluation approach is computationally expensive, it leads to one striking economy shared by no other method. With most statistical methods, application of the bootstrap for assessing the significance of results requires that the same analysis be carried out hundreds or thousands of times; in contrast, our approach allows the incorporation of the bootstrap *with no significant additional computing time requirements over the original analysis.*

Recall that each of the N aligned sequences has n positions. Before the search for the most parsimonious tree begins, we draw a random sample, *with replacement,* of size n, from the set {1, . . . , n}. We denote by n(i, 1) the number of times position i is chosen, for each i = 1, . . . , n. This is the first bootstrap sample. A second sample n(i, 2), i = 1, . . . , n is then drawn from (1, . . . , n) and so on, until 100 such bootstrap samples are in hand.

Now, in the course of evaluating a particular tree T as a candidate for the most parsimonious tree, suppose position i contributes a cost of r(i) to the total cost of the tree. Then it is considered to contribute n(i, j)r(i) to the total cost of the same tree T as a candidate for the best tree representation of the j-th bootstrap sample. The calculations for these bootstrap samples do not take very much time because they use the same dynamic programming results as the original sequence. Furthermore, it is no more complicated to store the most parsimonious tree for each of the 100 bootstrap samples than the most parsimonious tree for the original sequence.

After all trees have been examined, the information contained in the 100 bootstrap trees is used to test the branches of the most parsimonious tree T* (derived from the original sequence data). Each branch corresponds to a division of the N terminal nodes into two sets, those closest to one end of the branch versus those closest to the other end. Thus, we can verify easily whether a certain branch of T* is also in one of the bootstrap trees even if the latter has a very different overall structure from T*.

If a given branch of T* is also contained in many of the 100 bootstrap trees, then this branch may be considered well supported by the data. A branch of T* that appears rarely among the bootstrap trees should be omitted, and its two ends amalgamated, resulting in a node of degree 4, or more. The cutoff point is a question of some controversy. If we require that a branch be present in more than one-half of the bootstrap trees, then we can be sure that the set of such branches is consistent, i.e., that it will always be possible to build a tree out of these branches. This consistency condition is not generally true if some weaker criterion is used to accept branches, such as their being contained in at least one-third of the bootstrap trees. On the other hand, Felsenstein (1985) would require that 95% of the bootstrap trees contain a branch before it is accepted.

In our local optimization of tree fragments using a temporary constraint, we carry out the bootstrap analysis for each fragment t after the best tree has been established. Because these fragments overlap, each of the branches to be validated is generally tested several times, i.e., against 200 or 300 bootstrap trees. Moreover, the fact that we can handle a reasonably large fragment means that each branch validated by the bootstrap has been tested against a vast number of alternative topologies. With other

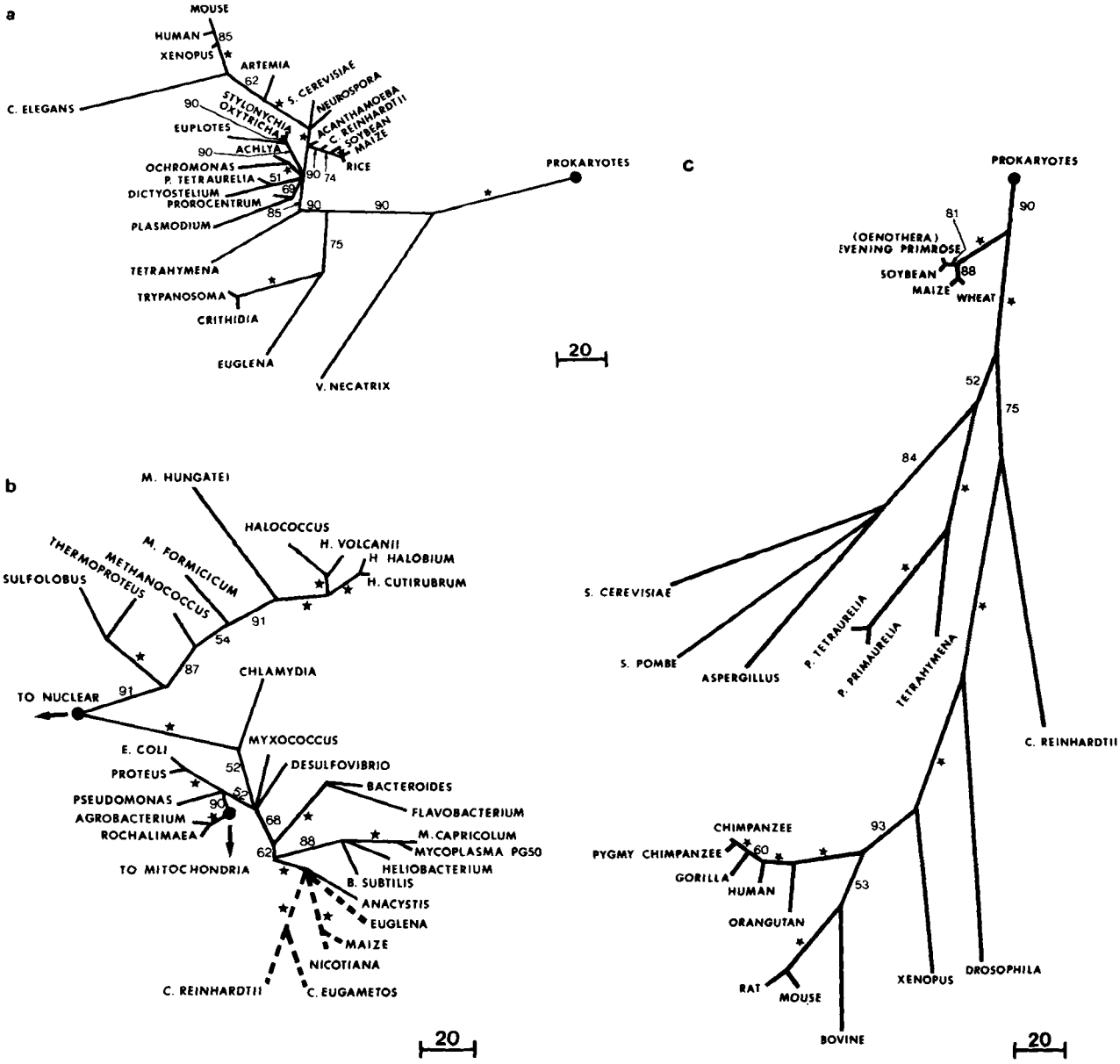


Fig. 2. The SSU tree. Branch lengths are proportional to the inferred mutational distance (number of mutations, as indicated by the scale). Numerals on branches refer to the number of times a particular branch was found in 100 bootstrap samples. The star symbol (*) indicates that greater than 95% of bootstrap runs contained this branch. **a** Nucleocytoplasmic sequences; **b** archaebacteria and eubacteria (chloroplast sequences are indicated by dashed lines); and **c** mitochondria. These trees use the same scale and a composite figure may be constructed by superimposing the appropriate solid circles that lead to the different lineages.

methods, such as "nearest-neighbor interchange," each branch competes with only a small number of alternatives, and a bootstrap validation by this approach is more likely to be artifactual.

Once we have obtained the final bootstrapped tree, dynamic programming is used again to establish the ancestral sequences as well as the lengths of the branches.

Congruency of the Phylogenetic Trees

Figure 2 shows the tree inferred from the SSU rRNA database, which comprises a total of 76 sequences: 9 archaebacterial, 15 eubacterial, 26 eukaryotic nuclear, 5 chloroplast, and 21 mitochondrial. This collection represents a broad selection of prokaryotic and eukaryotic taxa. Note that although the parsimony criterion does not bear on the determination of the position of the tree root (the proto-organism, or progenote), it is reasonable to assume that it falls somewhere on a branch leading to one of the three primary lineages (archaebacterial, eubacterial, eukaryotic nuclear). The tree is drawn to scale, i.e., branch lengths are proportional to the number of inferred mutations. Particularly evident are the relatively long branches in the mitochondrial subtree and leading to *Caenorhabditis elegans* in the eukaryotic nuclear lineage.

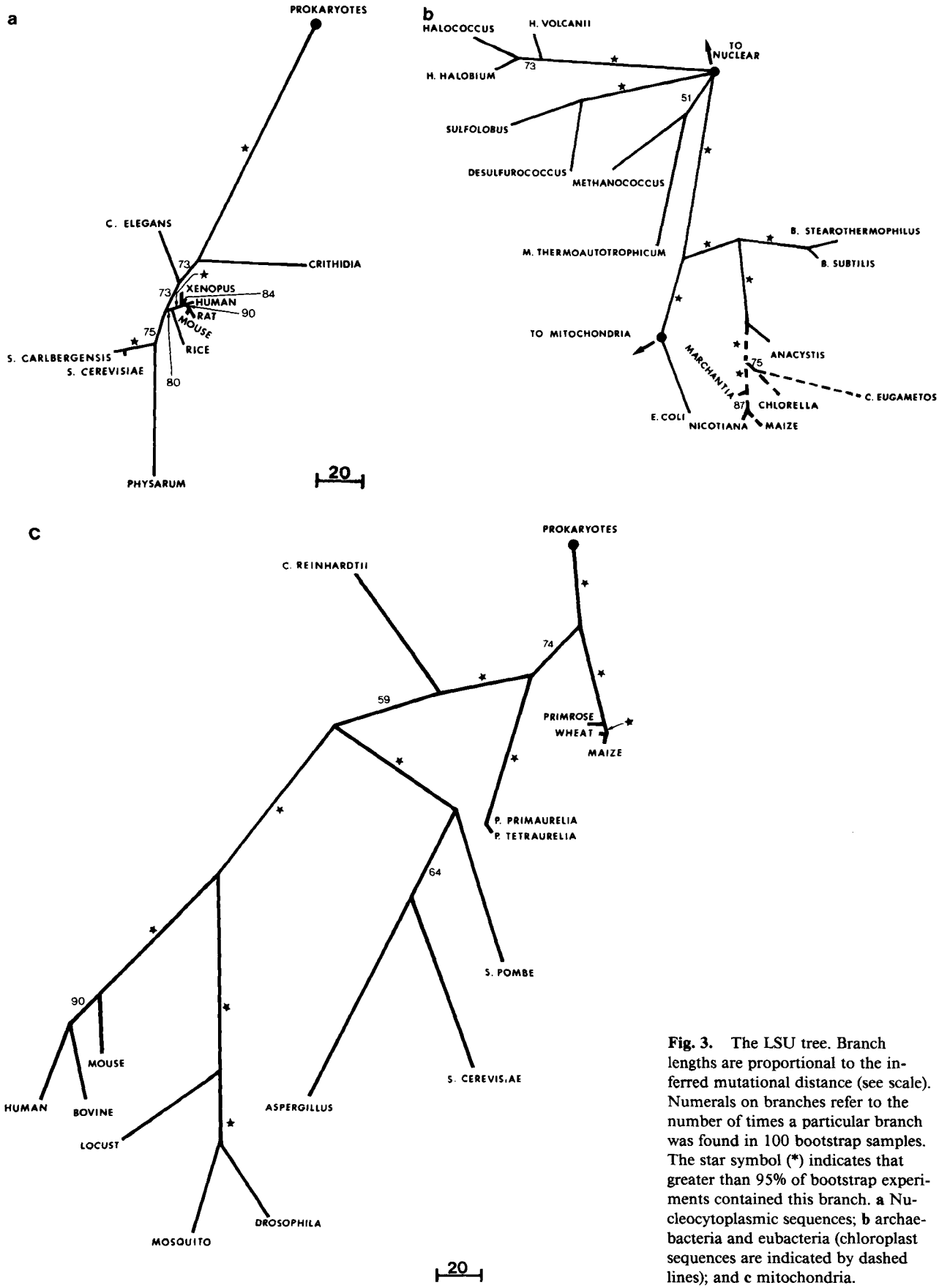


Fig. 3. The LSU tree. Branch lengths are proportional to the inferred mutational distance (see scale). Numerals on branches refer to the number of times a particular branch was found in 100 bootstrap samples. The star symbol (*) indicates that greater than 95% of bootstrap experiments contained this branch. **a** Nucleocytoplasmic sequences; **b** archaebacteria and eubacteria (chloroplast sequences are indicated by dashed lines); and **c** mitochondria.

In establishing the final form of the tree, we have taken into account the results of the bootstrap analysis. The chosen branches correspond to those that are found in more than 50% of the 100 trees analyzed by bootstrapping. Less well-separated branches (occurring in 50% or less of the trees) have been collapsed to a single node even though branching is indicated by the parsimony criterion. Such is the case within the chloroplast/cyanobacteria grouping, where, in the optimal tree, a deep branching for *Euglena* was observed. However, such a branching topology is not statistically verified, at least with the present data set, and all but the two plant and two algal chloroplast sequences are attached directly to the common root. Other collapsed nodes can be seen within fungal mitochondria, fungal nuclear, and protist nuclear sequences, and near the root of the eubacterial subtree.

The SSU rRNA tree confirms, at bootstrap levels of 91%, 100%, and 100%, respectively, the proposal of three primary lines of descent represented by the archaebacterial, eubacterial, and eukaryotic (nuclear) lineages. In our analysis, archaebacterial sequences are all more closely related to each other than to either eubacterial or eukaryotic sequences, with the root of the achaebacterial subtree approximately equidistant from the deepest eubacterial and eukaryotic nodes. If the root of the global tree is relatively close to the point of connection of the three primary lineages, some support can be given to the notion that archaebacteria, demonstrating lower mutation rates, most closely resemble the progenote (Woese 1987; Lake 1988).

The LSU rRNA tree is shown in Fig. 3. This tree is constructed from the sequences of 41 organisms/organelles (7 archaebacterial, 4 eubacterial, 10 eukaryotic nuclear, 5 chloroplast, and 15 mitochondrial). Again, statistically unreliable branches have been collapsed. Such is the case for different archaebacterial groupings, whose interrelationships cannot be completely determined according to the bootstrap analysis. Nevertheless, the three archaebacterial groupings are significantly separated from both the eukaryotic nuclear lineage (at a bootstrap level of 100%) and eubacteria (100%). Note that the relationships among fungal mitochondrial sequences, not defined by the SSU rRNA data set, can be defined with the LSU rRNA sequences (at the 70% level).

Although some inconsistencies in fine structure are evident, the SSU and LSU trees are remarkably congruent in their global groupings. Both trees confirm the eubacterial origin of the eukaryotic organelles, with chloroplasts emanating from within a grouping that includes the cyanobacteria, and mitochondria emerging from the so-called purple bacteria (Figs. 2B and 3B). Neither tree offers any evidence for a bi- or polyphyletic origin of chloroplasts, although in view of the fact that few cyanobacterial/plastid rRNA sequences have been determined, we do not exclude this possibility (see above).

In the SSU tree, mitochondria are seen to originate from within the alpha subdivision of the purple bacteria (Fig. 2B), confirming the conclusion of Yang et al. (1985). Our calculations show that the SSU tree has 1704 inferred mutations. Branching all mitochondria at the base of the eukaryotic subtree raises the mutation value to 1744. Placing the chloroplast grouping at the eukaryotic origin produces a tree with 1766 mutations. Surprisingly, plant mitochondria do not branch with green algal (chlorophyte) mitochondria, but rather as a cluster near the mitochondrial root in both subtrees. Because both nuclear (Fig. 2A) and chloroplast (Figs. 2B and 3B) phylogenies place plants and green algae on the same branch, a dichotomy exists with respect to plant mitochondria. The possible causes and significance of this anomaly will be discussed elsewhere, in connection with a more detailed analysis of mitochondrial phylogeny (M.W. Gray et al., unpublished).

Within the mitochondrial lineage, fungi and protozoa branch together in the SSU tree but separately in the LSU tree. In the LSU tree, the branching order among the fungal mitochondrial sequences is consistent with a phylogeny determined using tRNA sequences (Cedergren and Lang 1985). In the SSU tree, the bootstrap does not validate any structure internal to the fungal mitochondrial group. Indeed, it rejects a *Saccharomyces cerevisiae* versus (*Schizosaccharomyces pombe* + *Aspergillus nidulans*) branching produced by the treeing algorithm.

In spite of the smaller number of organisms in the LSU tree, the eubacterial subtrees in Figs. 2 and 3 are in reasonably good agreement. A fundamental divergence between gram-negative species on one hand and gram-positive/cyanobacterial species on the other is observed.

Among archaebacteria, the two trees differ more by the degree of resolution of nodes rather than by different topologies. Because the organisms represented in the two trees are not identical, slight differences in the configuration of methanogens could be an artifact.

The two eukaryotic nuclear subtrees display one major and some minor differences. Perhaps the most important overall difference between the two trees concerns the nematode, *C. elegans*, which branches as an animal in the SSU tree but as a protist in the LSU tree. The rapid mutation rate of *C. elegans* rRNA genes, as manifested by the long branch length, may be pertinent here, because the test of "invariants" (described by M.W. Gray et al., unpublished) does not confirm the results shown in Fig. 2A, but suggests instead that *C. elegans* branches closer to

the protists than *Artemia* in the SSU tree. Additional LSU sequences should clarify this point. Other differences between the SSU and LSU phylogenies involve the position of the yeast subtree with regard to the subtree containing plants, and the position of *Physarum* (a protist), which branches with fungi in the LSU tree, rather than nearer *Crithidia.* In these two cases, the test of "invariants" shows that this is not an artifact of branch lengths.

All in all, and of major significance, the two trees yield essentially identical global phylogenies defining the evolutionary relationships among the principal phyla. Although we do not claim or even think that all explicit relationships in these two trees are correct, the congruence between the SSU and LSU trees supports most of the major groupings shown in Figs. 2 and 3.

Comparing Phylogenies

The trees presented here permit the evaluation of previously inferred phylogenies. Firstly, with regard to the tripartite theory of life forms (archaebacteria, eubacteria, and eukaryotes) advanced by Woese and Fox (1977a) and coworkers (Woese and Olsen 1986), we are in agreement. Our trees do not support the alternative hypotheses of Lake et al. (1984, 1985) that subdivide archaebacteria into eocytes (that are treed together with eukaryotes), methanogens, and halobacteria (that are branched with eubacteria) (Lake 1988). Even in the LSU tree, where the topology within archaebacteria is not statistically validated, no alternate topologies were observed that would be in agreement with predictions based on Lake's model. The test of "invariants," however, does not distinguish clearly between the two hypotheses. For both the SSU and LSU trees, Lake's or Woese's model can be seen to be preferred *depending on the choice of the particular eukaryotic and eubacterial representatives.* Whether archaebacteria are more eubacterial or eukaryotic cannot be determined from our data; we can only assert that archaebacteria differ fundamentally from both. We note, however, that both Woese (1987) and Lake (1988) are in general agreement on the archaebacterial nature of the progenote (Woese and Fox 1977b), as we are.

The prokaryotic phylogeny presented here corresponds closely to published phylogenies based on SSU (Pace et al. 1986; Woese 1987) and LSU (archaebacteria, Leffers et al. 1987) rRNA sequences; it also agrees to a large extent with 5S rRNA-derived phylogenies (Willekens et al. 1986; Hori and Osawa 1987). Our SSU tree differs from that of Woese (1987) in the position of the bacteroides–flavobacteria: we place together with gram-positive bacteria/cyanobacteria, whereas Woese (1987) proposes two separate "superphyla" containing the gram-positive bacteria/cyanobacteria/purple bacteria and green sulfur bacteria/bacteroides–flavobacteria. In addition, in 5S trees the cyanobacteria/chloroplast lineage diverges prior to the separation of gram-positive bacteria and the purple bacteria. In our SSU tree and other rRNA sequence-based trees, the reverse is seen.

The archaebacterial phylogeny we suggest here is almost identical to others determined by other methods (Willekens et al. 1986; Hori and Osawa 1987; Leffers et al. 1987; Woese 1987). In contrast, the SSU trees of Lake (1988) and Wolters and Erdmann (1986) would place *Sulfolobus/Thermoproteus* on the eukaryotic branch; however, the 5S RNA data presented by Wolters and Erdmann (1986) are more in agreement with the topology presented here.

The topology of the eukaryotic nuclear SSU tree is in very good agreement with recent results from other laboratories. In particular, the *Vairimorpha necatrix* (microsporidian) sequence defines the deepest known branching in the nuclear tree (Vossbrinck et al. 1987), with the euglenoid (*Euglena*)/trypanosoid (*Trypanosoma, Crithidia*) divergence the next deepest (Sogin et al. 1986a). The branching order of later diverging protists is less consistent. The groupings *Euplotes/Stylonychia/Oxytricha* and *Achyla/Ochromonas* agree with the results of Sogin and co-workers (Elwood et al. 1985; Sogin et al. 1986c; Gunderson et al. 1987). However, these workers place *Paramecium tetraurelia* and *Tetrahymena thermophila* in a single branch together with other ciliates [*Euplotes, Stylonychia,* and *Oxytricha* (Sogin and Elwood 1986; Sogin et al. 1986c)], a branch that also includes the dinoflagellate, *Prorocentrum micans* (Gunderson et al. 1987). Additionally, Gunderson et al. (1986) find that *Dictyostelium discoideum* and *Plasmodium berghei* each branches early (and separately) from the backbone of the nuclear tree. As indicated in Fig. 2, bootstrap analysis for the SSU tree shows that our data do not warrant a high degree of branching resolution (i.e., binary branches) within the Protista. This is undoubtedly due to the fact that our global database is composed of a fewer number of sequence positions that are less divergent than the solely eukaryotic nuclear database of Sogin and co-workers.

The co-branching of the chlorophyte and higher plant SSU rRNA sequences confirms the relationship noted by Gunderson et al. (1987) and is consistent with traditional phylogenies that place chlorophytes and higher plants together (Ragan and Chapman 1978). A notable feature of this particular branch is the inclusion of the amoeboid protozoan, *Acanthamoeba castellanii.* If supported by other data, this relationship (first suggested by the work of

Schnare 1984) would represent the first strong evolutionary connection between a multicellular eukaryotic group and a specific nonphotosynthetic protist.

Conclusions

The database and analytical techniques described here together comprise a powerful approach in the evaluation of evolutionary relationships, not only among all living organisms, but within the mitochondrial and chloroplast lineages, and between these organelles and the three primary lines of descent (archaebacteria, eubacteria, and the eukaryotic nucleus). The utility of this technique is exemplified in this paper by the first extensive phylogeny of mitochondria. Assessment of evolutionary relationships within this lineage is complicated by the enormous diversity evident in patterns of mitochondrial genome organization and expression in different eukaryotic phyla, and in the very different rates at which homologous mtDNA-encoded genes diverge in sequence (M.W. Gray et al., unpublished). Our use of a sequence database drawn from "core" regions of secondary structure makes it possible to probe global evolutionary relationships that include such divergent lineages.

In the present analysis, phylogenies have been determined in parallel from separate SSU and LSU rRNA databases. The use of two such databases provides an important methodological check: as discussed earlier, the correspondence between the SSU and LSU phylogenies, especially in their global groupings, is very encouraging. Although the number of available LSU rRNA sequences, especially eubacterial and nucleocytoplasmic, is still rather limited, we should soon be in a position to carry out a systematic analysis of any persistent differences in fine structure between the two trees. For example, the effect of different combinations of sequences in determining the final topology, especially within the mitochondrial lineage, could be investigated. Such additional tests of the methodology, in concert with the statistical evaluation of tree topology described here, are important steps in our goal of describing a rigorously objective approach to determining panevolutionary phylogenies.

By their nature, our SSU and LSU databases are constrained to include only the most conservative regions of primary sequence. For that reason, our method is ideal for probing distant relationships, but is perhaps less satisfactory for determining close relationships, particularly when the sequences involved are slowly diverging. This may account for some of the lack of resolution seen among the Protista in the nuclear lineage of our SSU tree. Even so, there is a notable degree of agreement between the two phylogenetic trees presented here and those previously published for archaebacteria, eubacteria, and eukaryotes. In particular, in connection with the debate between supporters of the "archaebacterial tree" (Pace et al. 1986) and the "eocyte tree" (Lake 1988), our results support the archaebacterial tree.

In the analysis of rRNA data in this paper, we have used the maximum parsimony technique. We are well aware of the concern about the effect of differing rates of sequence evolution in the derivation of phylogenetic trees using different methods, including maximum parsimony (cf. Lake 1987; Olsen 1988). Wherever this effect may have led to artifactual results in our trees, we have applied the Cavender and Felsenstein (1987) and Lake (1987) tests of tree invariants. In general, these tests corroborate our analysis, showing little if any tree distortion caused by unequal evolutionary rates. Aside from analytical considerations, we emphasize the congruence between the two phylogenetic trees presented here and their agreement with previously published ones, determined by different methods and/or using a different subset of SSU or LSU rRNA sequence information. We submit that this argues against any major spurious artifacts in the topologies presented.

In any discussion about the significance of a given phylogenetic relationship, a paramount consideration is the nature of the data used to build the phylogeny and the selection of those data. This is particularly true in nucleic acid comparisons, because tree topologies can be affected by sequence alignments (Feng and Doolittle 1987; Lake 1988). We have discussed this question previously and have presented a statistically valid method for aligning RNA sequences (Sankoff and Cedergren 1973). However, aside from the validity of the alignment procedure itself, it is important to recognize that rRNA molecules contain conserved, semiconserved, and variable regions of primary sequence, and that the inclusion in an alignment of less strongly conserved regions of structure *that may not in fact be demonstrably similar* could well affect the resulting topology. Our approach to data selection, the rationale for which has been outlined in detail previously (Gray et al. 1984), eliminates potential alignment problems. In this context, we maintain that although it may no longer be "acceptable to throw sequences through any available tree-building method and to publish the results" (Penny 1988), neither is it acceptable to throw sequence data of questionable quality and/or alignment through even the most sophisticated algorithm.

Finally, as Olsen (1988) has recently asserted, "there is no substitute for raw data: more infor-

mation will always yield more reliable phylogenetic inferences." The ability of our procedure to handle a large number of sequences, and to accommodate new ones as they become available, should make it a valuable addition to the approaches currently available for exploring global evolutionary relationships.

Acknowledgments. We thank Drs. P.H. Boer, D.F. Spencer, C. Lemieux, M. Sogin, C.R. Woese, and K. Wolf for providing unpublished sequence data. We are also grateful to C.R. Woese and R.R. Gutell for providing the skeleton secondary structures used in Fig. 1. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada to R.C., D.S., and M.W.G., all of whom are Fellows of the Canadian Institute for Advanced Research.

References

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683–717

Brosius J, Dull TJ, Noller HF (1980) Complete nucleotide sequence of a 23S ribosomal RNA gene from *Escherichia coli.* Proc Natl Acad Sci USA 77:201–204

Brosius J, Dull TJ, Sleeter DD, Noller HF (1981) Gene organization and primary structure of a ribosomal RNA operon from *Escherichia coli.* J Mol Biol 148:107–127

Cavender JA, Felsenstein J (1987) Invariants of phylogenies: simple case with discrete states. J Classif 4:57–71

Cedergren R, Lang BF (1985) Probing fungal mitochondrial evolution with tRNA. BioSystems 18:263–267

Cedergren RJ, LaRue B, Sankoff D, Grosjean H (1981) The evolving tRNA molecule. CRC Crit Rev Biochem 11:35–104

Chan Y-L, Olvera J, Wool IG (1983) The structure of rat 28S ribosomal ribonucleic acid inferred from the sequence of nucleotides in a gene. Nucleic Acids Res 11:7819–7831

Clary DO, Wolstenholme DR (1985) The ribosomal RNA genes of *Drosophila* mitochondrial DNA. Nucleic Acids Res 13: 4029–4045

Dale RMK, Mendu N, Ginsburg H, Kridl JC (1984) Sequence analysis of the maize mitochondrial 26S rRNA gene and flanking regions. Plasmid 11:141–150

Dayhoff MO (1972) Atlas of protein sequence and structure, vol 5. National Biomedical Research Foundation, Washington DC

Diaconis P, Efron B (1983) Computer-intensive methods in statistics. Sci Am 248:116–130

Douglas SE, Doolittle WF (1984) Complete nucleotide sequence of the 23S rRNA gene of the cyanobacterium, *Anacystis nidulans.* Nucleic Acids Res 12:3373–3386

Dunon-Bluteau D, Brun G (1986) The secondary structures of the *Xenopus laevis* and human mitochondrial small ribosomal subunit RNA are similar. FEBS Lett 198:333–338

Edwards K, Kössel H (1981) The rRNA operon from *Zea mays* chloroplasts: nucleotide sequence of 23S rDNA and its homology with *E. coli* 23S rDNA. Nucleic Acids Res 9:2853–2869

Ellis RE, Sulston JE, Coulson AR (1986) The rDNA of *C. elegans*: sequence and structure. Nucleic Acids Res 14:2345–2364

Elwood HJ, Olsen GJ, Sogin ML (1985) The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates *Oxytricha nova* and *Stylonychia pustulata.* Mol Biol Evol 2: 399–410

Eperon IC, Anderson S, Nierlich DP (1980) Distinctive sequence of human mitochondrial ribosomal RNA genes. Nature 286:460–467

Farris JS (1983) The logical basis of phylogenetic analysis. In: Plonick NI, Funk VA (eds) Advances in statistics, vol 2. Columbia University Press, New York, pp 7–36

Felsenstein J (1983a) Statistical inference of phylogenies. Roy Stat Soc, Series A 146:246–272

Felsenstein J (1983b) Inferring evolutionary trees from DNA sequences. In: Weir BS (ed) Statistical analysis of DNA sequence data. Marcel Dekker, New York, pp 133–150

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351–360

Field KG, Olsen GJ, Lane DJ, Giovannoni SJ, Ghiselin MT, Raff EC, Pace NR, Raff RA (1988) Molecular phylogeny of the animal kingdom. Science 239:748–753

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specified tree topology. Syst Zool 20: 406–416

Fitch WM (1977) On the problem of generating the most parsimonius tree. Am Nat 111:223–257

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees: a method based on mutational distances as estimated from cytochrome *c* sequences is of general applicability. Science 155:279–284

Georgiev OI, Nikolaev N, Hadjiolov AA, Skryabin KG, Zakharyev VM, Bayev AA (1981) The structure of the yeast ribosomal RNA genes. 4. Complete sequence of the 25S rRNA gene from *Saccharomyces cerevisiae.* Nucleic Acids Res 9: 6953–6958

Golding GB (1983) Estimation of DNA and protein sequence divergence: an examination of some assumptions. Mol Biol Evol 1:125–142

Grabau EA (1985) Nucleotide sequence of the soybean mitochondrial 18S rRNA gene: evidence for a slow rate of divergence in the plant mitochondrial genome. Plant Mol Biol 5: 119–124

Gray MW, Boer PH (1988) Organization and expression of algal (*Chlamydomonas reinhardtii*) mitochondrial DNA. Philos Trans R Soc Lond B, in press

Gray MW, Sankoff D, Cedergren RJ (1984) On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit rRNA. Nucleic Acids Res 12:5837–5852

Green CJ, Stewart GC, Hollis MA, Vold BS, Bott KF (1985) Nucleotide sequence of the *Bacillus subtilis* ribosomal RNA operon, *rrnB.* Gene 37:261–266

Gunderson JH, Sogin ML (1986) Length variation in eukaryotic rRNAs: small subunit rRNAs from the protists *Acanthamoeba castellanii* and *Euglena gracilis.* Gene 44:63–70

Gunderson JH, McCutchan TF, Sogin ML (1986) Sequence of the small subunit ribosomal RNA gene expressed in the bloodstream stages of *Plasmodium berghei*: evolutionary implications. J Protozool 33:525–529

Gunderson JH, Elwood H, Ingold A, Kindle K, Sogin ML (1987) Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes. Proc Natl Acad Sci USA 84:5823–5827

Gutell RR, Weiser B, Woese CR, Noller HF (1985) Comparative anatomy of 16-S-like ribosomal RNA. Prog Nucleic Acids Res Mol Biol 32:155–216

Hadjiolov AA, Georgiev OI, Nosikov VV, Yavachev LP (1984) Primary and secondary structure of rat 28S ribosomal RNA. Nucleic Acids Res 12:3677–3693

Hartigan JA (1973) Minimum mutation fits to a given tree. Biometry 29:53–65

Hassouna N, Michot B, Bachellerie J-P (1984) The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. Nucleic Acids Res 12:3563–3583

Herzog M, Maroteaux L (1986) Dinoflagellate 17S rRNA sequence inferred from the gene sequence: evolutionary implications. Proc Natl Acad Sci USA 83:8644–8648

Hixson JE, Brown WM (1986) A comparison of the small ribosomal RNA genes form the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications. Mol Biol Evol 3:1–18

Hori H, Osawa S (1987) Origin and evolution of organisms as deduced from 5S ribosomal RNA sequences. Mol Biol Evol 4:445–472

HsuChen C-C, Kotin RM, Dubin DT (1984) Sequences of the coding and flanking regions of the large ribosomal subunit RNA gene of mosquito mitochondria. Nucleic Acids Res 12: 7771–7785

Huysmans E, de Wachter R (1986a) Compilation of small ribosomal subunit RNA sequences. Nucleic Acids Res 14:r73–r118

Huysmans E, de Wachter R (1986b) The distribution of 5S rRNA sequences in phenetic hyperspace. Implications for eubacterial, eukaryotic, archaebacterial and early biotic evolution. Endocyt Cell Res 3:133–155

Jarsch M, Böck A (1985) Sequence of the 23S rRNA gene from the archaebacterium *Methanococcus vannielii*: evolutionary and functional implications. Mol Gen Genet 200:305–312

Kop J, Wheaton V, Gupta R, Woese CR, Noller HF (1984) Complete nucleotide sequence of a 23S ribosomal RNA gene from *Bacillus stearothermophilus*. DNA 3:347–357

Kumano M, Tomioka N, Sugiura M (1983) The complete nucleotide sequence of a 23S rRNA gene from a blue-green alga, *Anacystis nidulans*. Gene 24:219–225

Lake JA (1987) A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. Mol Biol Evol 4:167–191

Lake JA (1988) Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. Nature 331: 184–186

Lake JA, Henderson E, Oakes M, Clark MW (1984) Eocytes: a new ribosome structure indicates a kingdom with a close relationship to eukaryotes. Proc Natl Acad Sci USA 81:3786–3790

Lake JA, Clark MW, Henderson E, Fay SP, Oakes M, Scheinman A, Thornber JP, Mah RA (1985) Eubacteria, halobacteria and the origin of photosynthesis: the photocytes. Proc Natl Acad Sci USA 82:3716–3720

Lang BF, Cedergren R, Gray MW (1987) The mitochondrial genome of the fission yeast, *Schizosaccharomyces pombe*. Sequence of the large-subunit ribosomal RNA gene, comparison of potential secondary structure in fungal mitochondrial large-subunit rRNAs and evolutionary considerations. Eur J Biochem 169:527–537

Laudien Gonzalez I, Gorski JL, Campen TJ, Dorney DJ, Erickson JM, Sylvester JE, Schmickel RD (1985) Variation among human 28S ribosomal RNA genes. Proc Natl Acad Sci USA 82:7666–7670

Leffers H, Kjems J, Ostergaard L, Larsen N, Garrett RA (1987) Evolutionary relationships amongst archaebacteria. A comparative study of a sulphur-dependent extreme thermophile, an extreme halophile and a thermophilic methanogen. J Mol Biol 195:43–61

Mankin AS, Kagramanova VK (1986) Complete nucleotide sequence of the single ribosomal RNA operon of *Halobacterium halobium*: secondary structure of the archaebacterial 23S rRNA. Mol Gen Genet 202:152–161

Manna E, Brennicke A (1985) Primary and secondary structure of 26S ribosomal RNA of *Oenothera* mitochondria. Curr Genet 9:505–515

McCarroll R, Olsen GJ, Stahl YD, Woese CR, Sogin ML (1983) Nucleotide sequence of the *Dictyostelium discoideum* small-subunit ribosomal ribonucleic acid inferred from the gene sequence: evolutionary implications. Biochemistry 22:5858–5868

Moore GW, Goodman M, Barnabas J (1973) An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. J Theor Biol 38:423–457

Nei M, Kowhn RK (1983) Evolution of genes and proteins. Sinauer Associates, Sunderland MA

Netzker R, Köchel HG, Basak N, Küntzel H (1982) Nucleotide sequence of *Aspergillus nidulans* mitochondrial genes coding for ATPase subunit 6, cytochrome oxidase subunit 3, seven unidentified proteins, four tRNAs and L-rRNA. Nucleic Acids Res 10:4783–4794

Noller HF (1984) Structure of ribosomal RNA. Annu Rev Biochem 53:119–162

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S-i, Inokuchi H, Ozeki H (1986) Complete nucleotide sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Plant Mol Biol Reporter 4:148–175

Olsen GJ (1987) The earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol LII:825–839

Otsuka T, Nomiyama H, Yoshida H, Kukita T, Kuhara S, Sakaki Y (1983) Complete nucleotide sequence of the 26S rRNA gene of *Physarum polycephalum*: its significance in gene evolution. Proc Natl Acad Sci USA 80:3163–3167

Pace NR, Olsen GJ, Woese CR (1986) Ribosomal RNA phylogeny and the primary lines of evolutionary descent. Cell 45: 325–326

Penny D (1988) What was the first living cell? Nature 331:111–112

Ragan MA, Chapman DJ (1978) A biochemical phylogeny of the protists, Academic Press, New York

Rothschild LJ, Ragan MA, Coleman AW, Heywood P, Gerbi S (1986) Are rRNA sequence comparisons the Rosetta stone of phylogenetics? Cell 47:640

Ruvolo M, Smith TF (1986) Phylogeny and DNA–DNA hybridization. Mol Biol Evol 3:285–289

Sankoff D (1987) Computational complexity and cladistics. In: Hoenigswald HM, Wiener LF (eds) Biological metaphor and cladistic classification. University of Pennsylvania Press, Philadelphia, pp 269–280

Sankoff D, Cedergren R (1973) A test for nucleotide sequence homology. J Mol Biol 77:159–164

Sankoff D, Cedergren R (1983) Simultaneous comparison of three or more sequences related by a tree. In: Sankoff D, Kruskel JB (eds) Time warps, string edits, and macromolecules: the theory and practices of sequence comparison. Addison-Wesley, Reading, pp 253–263

Sankoff D, Rousseau P (1975) Locating the vertices of a Steiner tree in an arbitrary metric space. Math Program 9:240–248

Schnare MN (1984) Ribosomal RNA structure and evolution revealed by nucleotide sequence analysis. Thesis, Dalhousie University, Halifax, Nova Scotia

Schnare MN, Collings JC, Gray MW (1986a) Structure and evolution of the small subunit ribosomal RNA gene of *Crithidia fasciculata*. Curr Genet 10:405–410

Schnare MN, Heinonen TYK, Young PG, Gray MW (1986b) A discontinuous small subunit ribosomal RNA in *Tetrahymena pyriformis* mitochondria. J Biol Chem 261:5187–5193

Seilhamer JJ, Gutell RR, Cummings DJ (1984) *Paramecium*

mitochondrial genes. II. Large subunit rRNA gene sequence and microevolution. J Biol Chem 259:5173–5181

Sneath PH, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

Sogin ML, Elwood HJ (1986) Primary structure of the *Paramecium tetraurelia* small-subunit rRNA coding region: phylogenetic relationships within the Ciliophora. J Mol Evol 23: 53–60

Sogin ML, Elwood HJ, Gunderson JH (1986a) Evolutionary diversity of eukaryotic small-subunit rRNA genes. Proc Natl Acad Sci USA 83:1383–1387

Sogin ML, Miotto K, Miller L (1986b) Primary structure of the *Neurospora crassa* small subunit ribosomal RNA coding region. Nucleic Acids Res 14:9540

Sogin ML, Swanton MT, Gunderson JH, Elwood HJ (1986c) Sequence of the small subunit ribosomal RNA gene from the hypotrichous ciliate *Euplotes aediculatus*. J Protozool 33:26–29

Sokal RR, Sneath PH (1963) Principles of numerical taxonomy. Freeman, San Francisco

Sor F, Fukuhara H (1983) Complete DNA sequence coding for the large ribosomal RNA of yeast mitochondria. Nucleic Acids Res 11:339–348

Spencer DF, Bonen L, Gray MW (1981) Primary sequence of wheat mitochondrial 5S ribosomal ribonucleic acid: functional and evolutionary implications. Biochemistry 20:4022–4029

Spencer DF, Collings JC, Schnare MN, Gray MW (1987) Multiple spacer sequences in the nuclear large subunit ribosomal RNA gene of *Crithidia fasciculata*. EMBO J 6:1063–1071

Suyama Y (1986) Two-dimensional polyacrylamide gel electrophoresis analysis of *Tetrahymena* mitochondrial tRNA. Curr Genet 10:411–420

Takaiwa F, Sugiura M (1982) The complete nucleotide sequence of a 23-S rRNA gene from tobacco chloroplasts. Eur J Biochem 124:13–19

Takaiwa F, Oono K, Iida Y, Sugiura M (1985) The complete nucleotide sequence of a rice 25S.rRNA gene. Gene 37:255–259

Uhlenbusch I, McCracken A, Gellissen G (1987) The gene for the large (16S) ribosomal RNA from the *Locusta migratoria* mitochondrial genome. Curr Genet 11:631–638

Van Etten RA, Walberg MW, Clayton DA (1980) Precise localization and nucleotide sequence of the two mouse mitochondrial rRNA genes and three immediately adjacent novel tRNA genes. Cell 22:157–170.

Veldman GM, Klootwijk J, de Regt VCHF, Planta RJ, Branlant C, Krol A, Ebel J-P (1981) The primary and secondary structure of yeast 26S rRNA. Nucleic Acids Res 9:6935–6952

Vossbrinck CR, Maddox JV, Friedman S, Debrunner-Vossbrinck BA, Woese CR (1987) Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. Nature 326: 411–414

Ware VC, Tague BW, Clark CG, Gourse RL, Brand RC, Gerbi SA (1983) Sequence analysis of 28S ribosomal DNA from the amphibian *Xenopus laevis*. Nucleic Acids Res 11:7795–7817

Weisburg WG, Woese CR, Dobson ME, Weiss E (1985) A common origin of Rickettsiae and certain plant pathogens. Science 230:556–558

Weisburg WG, Hatch TP, Woese CR (1986) Eubacterial origin of Chlamydiae. J Bacteriol 167:570–574

Willekens P, Huysmans E, Vandenberghe A, de Wachter R (1986) Archaebacterial 5S ribosomal RNA: nucleotide sequence in two methanogen species, secondary structure models, and molecular evolution. Syst Appl Microbiol 7:151–159

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese CR, Fox GE (1977a) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090

Woese CR, Fox GE (1977b) The concept of cellular evolution. J Mol Evol 10:1–6

Woese CR, Olsen GJ (1986) Archaebacterial phylogeny: perspectives on the urkingdoms. Syst Appl Microbiol 7:161–177

Woese CR, Debrunner-Vossbrinck BA, Oyaizu H, Stackebrandt E, Ludwig W (1985) Gram-positive bacteria: possible photosynthetic ancestry. Science 229:762–765

Wolters J, Erdmann VA (1986) Cladistic analysis of 5S rRNA and 16S rRNA secondary and primary structure, the evolution of eukaryotes and their relation to archaebacteria. J Mol Evol 24:152–166

Yamada T, Shimaji M (1987) An intron in the 23S rRNA gene of the *Chlorella* chloroplasts: complete nucleotide sequence of the 23S rRNA gene. Curr Genet 11:347–352

Yang D, Oyaizu Y, Oyaizu H, Olsen GJ, Woese CR (1985) Mitochondrial origins. Proc Natl Acad Sci USA 82:4443–4447

Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–366

Received March 29, 1988/Revised July 15, 1988