Long-Branch Attraction and the rDNA Model of Early Eukaryotic Evolution

John W. Stiller and Benjamin D. Hall

Departments of Botany and Genetics, University of Washington

Phylogenetic analyses of ribosomal RNA genes have become widely accepted as a framework for understanding broad-scale eukaryotic evolution. Nevertheless, conflicts exist between the phylogenetic placement of certain taxa in rDNA trees and their expected position based on fossils, cytology, or protein-encoding gene sequences. For example, pelobiont amoebae appear to be an ancient group based on cytologic features, but they are not among the early eukaryotic branches in rDNA analyses. In this report, the derived position of pelobionts in rDNA trees is shown to be unreliable and likely due to long-branch attraction among more deeply branching sequences. All sequences that branch near the base of the tree suffer from relatively high apparent substitution rates and exhibit greater variation in ssu rDNA sequence length. Moreover, the order of the branches leading from the root of the eukaryotic tree to the base of the so-called ''crown taxa'' is consistent with a sequential attachment, due to ''long-branch'' effects, of sequences with increasing rates of evolution. These results suggest that the basal eukaryotic topology drawn from rDNA analyses may be, in reality, an artifact of variation in the rate of molecular evolution among eukaryotic taxa.

Introduction

Current views of eukaryotic evolution rely heavily on evidence from DNA sequence-based phylogenies, particularly on analyses of small-subunit ribosomal RNA genes (ssu rDNA). Trees based on ssu rDNA sequences (see Sogin 1997 for recent review) indicate that a number of unrelated protist lineages, including parasitic groups that lack mitochondria, diverged sequentially after the origin of eukaryotes. These trees also suggest that relatively late in the course of eukaryotic evolution, an apparent explosion of diversity occurred, leading to the so-called ''crown taxa'' (Knoll 1992; Sogin 1997). This crown radiation includes all multicellular organisms and several diverse protist groups.

Although rDNA analyses have provided valuable insights into evolutionary relationships among eukaryotes, a number of authors have questioned whether the rDNA model of evolution may be biased due to differences in G+C content or variation in evolutionary rates (Loomis and Smith 1990; Siddall, Hong, and Desser 1992; Hasegawa and Hashimoto 1993; Philippe and Adouette 1998; Hirt et al. 1999). Moreover, data from independent sources, both molecular and cytological, are in conflict with the positions of a number of protist groups on the rDNA tree (e.g., Loomis and Smith 1990; Baldauf and Doolittle 1997; Germot, Philippe, and Le Guyader 1997; Hirt et al. 1999). One such group comprises the pelobiont amoebae.

The Pelobiontida are composed of free-living amoeboflagellates that occur widely in anoxic and microoxic sediments. Pelobionts lack mitochondria and Golgi bodies and have an extremely simple endomembrane system (Brugerolle 1991; Simpson et al. 1997). They are defined taxonomically by a unique flagellar root architecture that sets them apart from all other eukaryotes and which, combined with their lack of mitochondria and nonparasitic habit, led to the proposal that pelobionts were the first lineage to diverge from the common ancestor of all eukaryotes (Cavalier-Smith 1991; Simpson et al. 1997). In phylogenetic analyses using ssu rDNA sequences, however, the pelobiont *Mastigamoeba balamuthi* (under the synonym *Phreatamoeba balamuthi*) does not emerge near the base of the eukaryotic tree but, rather, branches just before the ''crown'' radiation (Hinkle et al. 1994).

In contrast, recent analyses of the gene encoding the largest subunit of DNA-dependent RNA polymerase II (*RPB*1) from a different pelobiont, *Mastigamoeba invertens* (Stiller, Duffield, and Hall 1998), place it among the most basal eukaryotic taxa. Moreover, unlike sequences from parasitic amitochondriate protists that also branch near the base of *RPB*1 trees, the *Mastigamoeba* gene does not exhibit strong indications of ''long-branch attraction'' (Stiller, Duffield, and Hall 1998) that could lead to an artificially deep branching position (Felsenstein 1978; Hendy and Penny 1989). In an attempt to reconcile the discordant views of the pelobiont origin provided by cytology and *RPB*1 on the one hand and the single published pelobiont ssu rRNA gene on the other, we isolated the rDNA sequence from *M. invertens.*

We examined these and a broad array of eukaryotic rDNA sequences for evidence of long-branch artifacts that might help to explain conflicts between the rDNA and other lines of evidence regarding the evolutionary history of certain eukaryotic taxa. Our results suggest that the basal topology of the rDNA tree can be fully explained by the effects of long-branch attraction among more rapidly evolving eukaryotic sequences.

Materials and Methods

DNA Isolation and Sequencing

Mastigamoeba invertens was obtained from the American Type Culture Collection. Cultures were grown

Abbreviation: ssu rDNA, small-subunit ribosomal RNA gene.

Key words: amitochondriate amoebae, evolution, long-branch attraction, *Mastigamoeba,* phylogenetic, ssu rRNA.

Address for correspondence and reprints: John W. Stiller, Departments of Botany and Genetics, University of Washington, Box 357360, Seattle, Washington 98195. E-mail: stiller@u.washington.edu.

Mol. Biol. Evol. 16(9):1270–1279. 1999

and DNA was recovered as previously described (Stiller, Duffield, and Hall 1998). *Mastigamoeba* ssu rDNA was PCR-amplified using the general eukaryotic primers NS-1 and NS-8 (White et al. 1990); the single band present in all amplifications was column purified (Qiagen, Chatworth, Calif.). This band was digested with three restriction enzymes, *Cfo*I, *Hin*fI, and *Msp*I, to verify the presence of a homogeneous PCR product. The purified fragment then was completely sequenced in complementary directions using the ABI Prism dye-terminator cycle sequencing system (PE Applied Biosystems, Foster City, Calif.). The amplified PCR fragments were also cloned with the TOPO/TA kit (Invitrogen, San Diego, Calif.). To look for small sequence heterogeneity undetected by restriction digests, four different plasmids containing the *Mastigamoeba* rDNA sequence were column purified and sequenced with the ABI automated system.

Phylogenetic Analyses

Alignments of 35 eukaryotic and 3 archaeal sequences, which take into consideration rRNA secondary structure, were retrieved from the ribosomal database project (www.cme.msu.edu/RDP; Maidak et al. 1997) and subsequently aligned by eye with the sequence isolated from *M. invertens.* Large insertions in individual sequences and regions of ambiguity in the total alignment were removed before performing phylogenetic analyses (alignment available at http:///www.faculty.washington.edu/stiller/rDNA). In addition, portions of the alignment that were ambiguous in only one or a few taxa (usually one or both of the Microsporidia) or in the archaebacterial outgroup sequences but could be aligned reliably for all other sequences were retained with the region of ambiguity removed from the individual sequence(s). These positions were encoded as missing data and were not taken into account in our analyses of long-branch attraction. The resulting alignment included 1,172 positions including gaps.

Maximum-likelihood (ML) estimates of substitution parameters were made with the program PUZZLE (Strimmer and von Haeseler 1996) assuming a mixed model for variation in nucleotide substitution rate. First, the percentage of invariable positions (18%) was calculated and supplied as the rate probability for constant sites. Variable rates then were estimated using a discrete approximation to a γ distribution model with six rate substitution categories and an α parameter (0.66, SE = 0.04) calculated from the data set. These discrete substitution rate categories, as well as the proportion of invariable sites, were incorporated into ML tree construction invoking the C-option in DNAml (PHYLIP; Felsenstein 1989) to supply the relative rate and proportion of the total sites for each category. Three subreplicates were executed, each with random addition of sequences and global rearrangements. One thousand maximum-parsimony bootstrap replicates were performed at default settings in PAUP 3.1.1 (Swofford 1993) using 10 random-addition subreplicates per bootstrap round. After examination of the alignment for indicators of long-branch attraction (see below), phylogenetic analyses were repeated with a subalignment of 26 eukaryotic rDNA sequences using the programs described above to estimate parameters and construct trees.

Analyses of Long-Branch Attraction Indicators

Four different analyses were conducted, each independent of a predetermined tree topology, to ascertain whether characteristics of individual sequences within the rDNA alignment made their position in phylogenetic reconstructions unreliable. First, a χ^2 test for deviant nucleotide composition was performed using the ML program PUZZLE. Next, the number of unique substitutions was tabulated for each sequence, i.e., instances in which a nucleotide at a given position in one sequence was different and invariable in all other sequences, including the archaebacterial outgroups. We also counted such unique substitutions if they were found in all members of a terminal lineage that was defined clearly by multiple taxonomic criteria. These substitutions were averaged over all members of the taxon. We consider these terminal taxa to be animals, apicomplexans, ciliates, diplomonads, fungi, green plants, kinetoplastids (*Euglena* + *Trypanosoma*), microsporidians, stramenopiles, and red algae. Other well-defined terminal taxa (dinoflagellates, trichomonads, haptophytes, etc.) are represented by a single sequence in our data set.

Regression and variance analyses were performed with the computer package RASA (Lyons-Weiler, Hoelzer, and Tausch 1996), which examines each sequence in an alignment for an excess of cladistic signal (positions that could be considered uniquely shared-derived characters in phylogenetic analyses) against a background of total phenetic sequence similarity. Regression and variance plot analyses were used both to assess the total signal content present in the data set and to identify individual sequences that are most likely to be affected by long-branch attraction.

Finally, random sequences were constructed using the program MacClade (Maddison and Maddison 1992), each with the average nucleotide composition for eukaryotic rDNA genes estimated by PUZZLE. Ten different sequences were used to replace the archaebacterial outgroups in parsimony bootstrap analyses of 100 replicates per sequence, with 10 random additions per replicate, to empirically determine which of the eukaryotic sequences had a tendency to attract ‘‘long branches’’ (Graham 1997). In addition, ‘‘long-branch’’ rDNA sequences identified by RASA analyses were removed from the alignment one by one and replaced by 10 random sequences, and parsimony analyses (one with each random sequence) were performed with each new data set.

The problematical rDNA sequences identified using these four methods were then tested individually to determine where they joined the tree in the absence of other rDNA ‘‘long branches.’’ Finally, separate parsimony analyses were performed to examine the behavior of 100 individual random sequences (10 jumbled subreplicates with each) when all rDNA ‘‘long branches’’ were removed from the alignment.

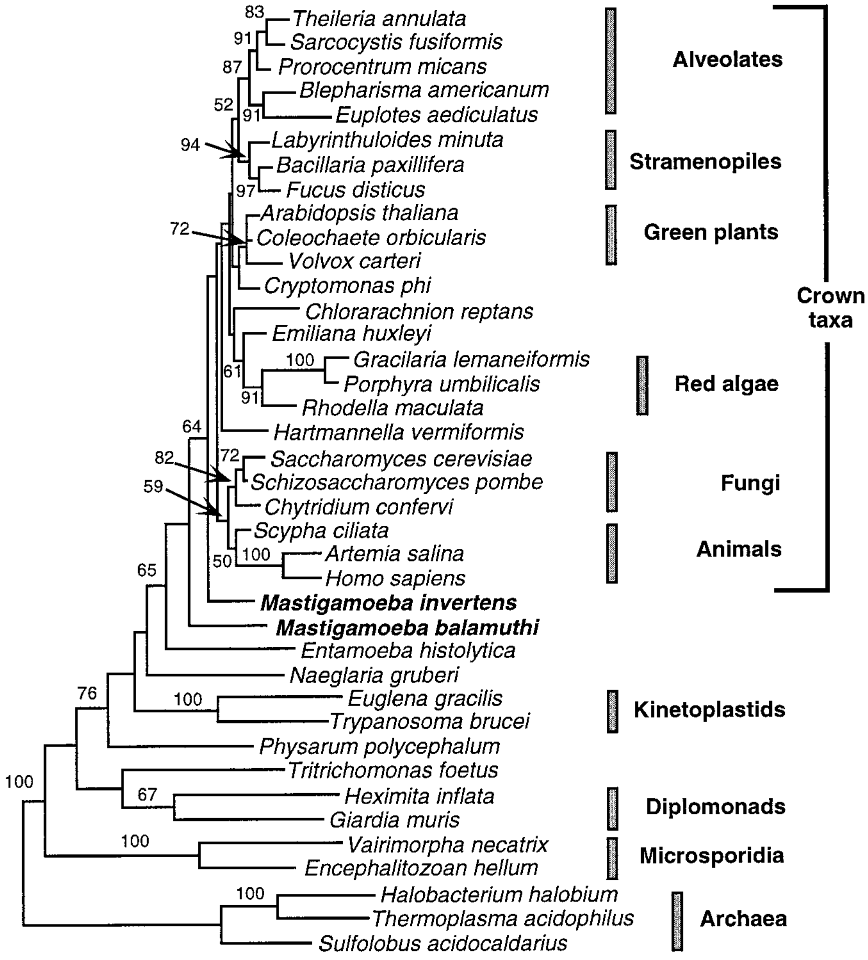

FIG. 1.—Maximum-likelihood tree with branch lengths. This was the most likely tree from three random-addition replicates using DNAml (PHYLIP), with substitution parameters calculated by PUZZLE (see *Materials and Methods*). Parsimony bootstrap support values greater than 50% are shown.

Results and Discussion

Using universal eukaryote-specific ssu rDNA PCR primers, a single fragment of approximately 1.8 kb was amplified from *M. invertens* DNA (GenBank accession number AF153206). Restriction digests of this band produced unambiguously discrete banding patterns with all enzymes tested (data not shown), indicating that little sequence heterogeneity is present among rDNA copies in *M. invertens.* Three individual nucleotide differences were found among the four copies that were sequenced, each occurring in only one of the clones examined. The consensus of the four sequences was used in all further analyses.

The *M. invertens* sequence is typical for a eukaryotic ssu rRNA gene with regard to size (1,748 bp) and G+C content (45%). The absence of large indels allows the *M. invertens* sequence to be aligned unambiguously with the conserved core regions from most other eukaryotic rDNA sequences. Maximum-likelihood phylogenetic analyses (fig. 1) yielded a topology in general agreement with the typical rDNA model of eukaryotic evolution that has emerged from a number of independent studies that used different data sets and various methods (e.g., Hinkle et al. 1994; Kumar and Rzhetsky 1996; Sogin 1997). Also in agreement with previous analyses, the *Mastigamoeba* sequences branch close to the base of the ''crown'' radiation, not near the eukaryotic origin.

While both pelobionts emerge in the same region of the tree, they do not group together. Rather, *M. balamuthi* branches just before the common node that supports *M. invertens* and the ''crown'' eukaryotes. Although trees constraining a sister relationship between the pelobionts cannot be rejected by Templeton (1983) or Kishino-Hasegawa (1989) tests ($P = 0.36$ and $P = 0.24$, respectively), the internode separating the two is reasonably well supported in parsimony bootstrap analyses, compared with many other internal nodes (fig. 1).

The failure of the two *Mastigamoeba* sequences to form a clade suggested to us that one or both of them may branch incorrectly in rDNA-based phylogenetic analyses. The most striking difference between the two *Mastigamoeba* sequences is that the gene from *M. invertens* is approximately 1.8 kb in length, a size typical for most crown eukaryotes, whereas the sequence from *M. balamuthi* exceeds 2.7 kb due to several long insertions relative to other eukaryotic genes.

The *M. balamuthi* insertions occur in regions of the rRNA gene that often contain insertions or deletions

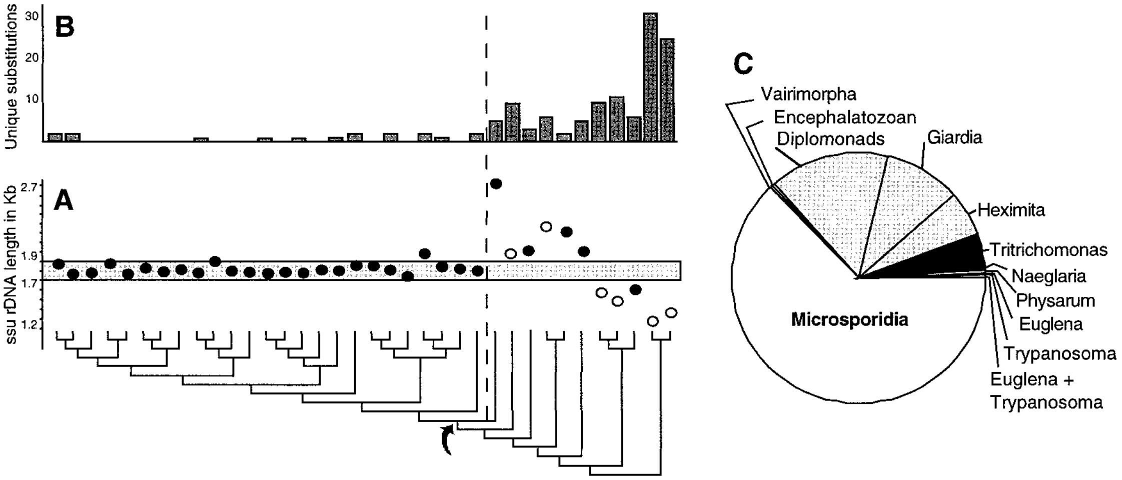

FIG. 2.—Distribution of sequence lengths and long-branch indicators mapped onto the tree shown in figure 1 (with archaean branches pruned). *A*, Total length in kilobases, prior to alignment editing, of the rDNA sequences analyzed. Open circles represent sequences that failed a 5% χ^2 test for average eukaryotic base composition. The arrow indicates the node that separates the two *Mastigamoeba* sequences. *B*, Number of unique substitutions found in each sequence. *C*, Distribution of partitions in which random sequences were attracted to a specific taxon in 1,000 bootstrap replicates (10 random sequences × 100 replicates each).

when comparisons are made between distantly related eukaryotes (Hinkle et al. 1994). When large-scale eukaryotic phylogenetic analyses are performed, these more variable regions typically are removed (as in this study), with the implicit assumption that major expansions and contractions of total rRNA are not directly tied to a systematic difference in the pattern and rate of substitutions in more conserved core regions of the gene. The failure of the two *Mastigamoeba* sequences to branch together, however, raises the possibility that the presence of these insertions might be correlated with an increased substitution rate in *M. balamuthi* ssu rDNA that causes it to be drawn away from *M. invertens* and toward longer branches at the base of the tree.

Further examination shows that size variation is not distributed randomly on the ssu rDNA tree. The node that separates the two pelobiont sequences also divides the tree between genes that are approximately 1.8 kb in length and those that are substantially larger or smaller (fig. 2*A*). Nearly all sequences belonging to "crown" eukaryotes are between 1.7 and 1.9 kb long. In contrast, sequences that cluster near the tree's base range between 1.2 and 2.7 kb in length, and, more significantly, none of them fall into the size range of typical "crown" eukaryotic genes. These observations led us to examine the entire rDNA data set to see whether this greater variation in sequence length was associated with a tendency toward long-branch attraction.

Analyses of Long-Branch Attraction

All sequences that deviate from the ML estimate of average eukaryotic nucleotide composition branch near the base of the tree (fig. 2*A*). Biased nucleotide composition has been suggested as a possible cause for artificially deep branches in rDNA trees (Loomis and Smith 1990; Hasegawa and Hashimoto 1993) and probably contributes to an incorrect placement of certain sequences. Nevertheless, nearly half of the deep-branching sequences, including the *M. balamuthi* gene itself, do not deviate significantly from typical eukaryotic nucleotide frequencies. Differences in base composition, therefore, do not appear to explain an incorrect placement of either pelobiont sequence, nor do they represent a ubiquitous source of long-branch attraction at the base of the rDNA tree.

There are clearly a larger number of unique substitutions at otherwise invariable positions in all rDNA sequences that are substantially larger or smaller than the typical 1.8-kb length (fig. 2*B*). Some increase in unique substitutions can be expected near the base of the tree, even in correctly placed branches, due to longer evolutionary histories; however, the numbers of unique substitutions in the microsporidian ($\mu = 29$), diplomonad ($\mu = 10$), and entamoeban ($\mu = 9$) sequences are larger even than those in archaebacteria ($\mu = 7$). In addition, there is an obvious disparity in the typical number of unique substitutions between sequences that lie on opposite sides of the node that separates the two *Mastigamoeba* sequences (fig. 2*B*). It is highly improbable that a protracted period of evolution separated all of the lower branches of the tree from the terminal radiation, particularly given the relative branching positions of the two pelobiont sequences. It is far more likely that the greater numbers of unique substitutions at conserved positions, along with the accompanying presence of large insertions and deletions, reflect increased rates of sequence evolution in all of the basal branches of the rDNA tree. This hypothesis was further examined using statistical analyses of the distribution of apparently synapomorphic sites, as well as through the behavior of randomly generated sequences in phylogenetic analyses of the rDNA data set.

Analyses of apparently synapomorphic sites (Lyons-Weiler, Hoelzer, and Tausch 1996) indicate that phylogenetic signal is limited in the total rDNA data set and that the tendency toward long-branch attraction is a feature of all deeply branching sequences. In regression

Table 1
Results of Relative-Synapomorphy and Random-Sequence Analyses

Sequences Removed/Number Analyzed	t_{RASA}	df	*P*	Longest Branch(es)	Random Sequence Attracted (Frequency)
None/39	0.3596	699	NS	Microsporidia	Microsporidia (0.70)
Archaea/36	1.028	591	NS	Microsporidia	Microsporidia (0.88) *Heximita* (0.12)
Microsporidia/34	−7.901	524	NS	Diplomonads	Diplomomads (0.55) *Tritrichomonas* (0.45)
Diplomonads/32	−2.119	461	NS	*Tritrichomonas*	*Tritrichomonas* (0.90)
Tritrichomonas/31	0.7450	431	NS	*Euglena*	*Physarum* (0.65) Kinetoplastids (0.20) *Naeglaria* (0.15)
Euglena/30	−0.9808	402	NS	*Physarum* *Trypanosoma* *Naeglaria*	*Physarum* (0.50) *Trypanosoma* (0.40) *Naeglaria* (0.10)
Physarum/29	−0.7776	374	NS	*Trypanosoma* *Naeglaria*	*Trypanosoma* (1.0)
Trypanosoma/28	−5.541	347	NS	*Naeglaria*	*Naeglaria* (1.0)
Naeglaria/27	2.24	321	0.01	*Entamoeba*	*Entamoeba* (0.94) *M. balamuthi* (0.06)
Entamoeba/26	4.541	296	<0.001	*Mastigamoeba balamuthi*	*M. balamuthi* (0.70) Rhodophyta (0.20) *Euplotes* (0.05) *Chlorarachnion* (0.05)

NOTE.—Support for significant difference of t_{RASA} from the null hypothesis of a random relationship between cladistic signal and phenetic similarity among sequences which is approximated by Student's *t* distribution (Lyons-Weiler, Hoelzer, and Tausch 1996). NS = not significant; df = degrees of freedom. The ''longest branch'' is the most skewed sequence in the RASA taxon variance analysis. Several sequences are indicated when they are close to equally skewed.

analyses of signal content, a null hypothesis of a random relationship between cladistic signal and phenetic similarity among all sequences cannot be rejected (table 1). Taxon variance analysis (fig. 3*A*) identifies the sequences previously shown to contain a disproportionate number of unique substitutions as having an excess of cladistic variance and therefore being prone to attracting long branches (Lyons-Weiler and Hoelzer 1997). Furthermore, when these sequences are removed sequentially from the alignment in the order of their skewness in taxon variance plots and the regression is recalculated with each reduced data set, phylogenetic signal content cannot be distinguished from a random distribution (table 1) until all sequences that branch below *Entamoeba* (fig. 1) are removed.

The data are judged to contain significant phylogenetic signal when *Entamoeba* and/or *M. balamuthi* are included (table 1); however, both sequences show excessive cladistic variance (fig. 3*A*). Although the *M. balamuthi* sequence is the least skewed of the rDNA genes of atypical sizes, it has clear long-branch indications in a taxon variance analysis when all of the more deviant long-branch sequences are excluded (fig. 3*B*).

An empirical tendency to attract long branches was tested by replacing the archaebacterial outgroups with randomly generated sequences of average eukaryotic rDNA base composition. These random sequences usually attach to the tree on the branch leading to the Microsporidia (fig. 2*C* and table 1). Furthermore, the Microsporidia attract random sequences most of the time, even when archaebacteria are included in the analyses (table 1). Diplomonads and *Tritrichomonas* also show a significant attraction to random sequences when all of the eukaryotic sequences are included in the analysis (fig. 2*C*). With the exception of *Entamoeba* and *M. balamuthi,* all of the potential long-branch sequences identified in taxon variance plots attracted a random sequence in at least one of the analyses using the entire eukaryotic alignment. In contrast, random sequences never attached to any of the more typically sized genes that characterize the ''crown'' radiation.

Perhaps more significantly, as each longest branch was removed successively and replaced by 10 different random sequences, in a large majority of phylogenetic analyses the random sequences attached to the point on the tree that had been vacated by the long-branch rDNA sequence (table 1). Beginning with the archaeal outgroups, this general correlation between the branching position of the most basal rDNA sequence and one generated randomly was repeated up to the removal of *Entamoeba*; however, the pattern did not hold when *M. balamuthi* was replaced (see below). Moreover, in no case did the random sequences attach to the tree's center of gravity, as determined by midpoint rooting, which might be expected if evolutionary signal and not rate variation was responsible for the long basal branches. This close correlation between the identification of long branches in taxon variance plots and the empirical tendency to behave as a random sequence in phylogenetic analyses casts doubt on the hypothesis that the topology of deep branches on the rDNA tree is based on evolutionary signal.

Analyses with Long Branches Removed

With all long branches identified by taxon variance and random sequence additions removed, the two *Mastigamoeba* species branch together in both ML and parsimony analyses. Although bootstrap values are not ro-

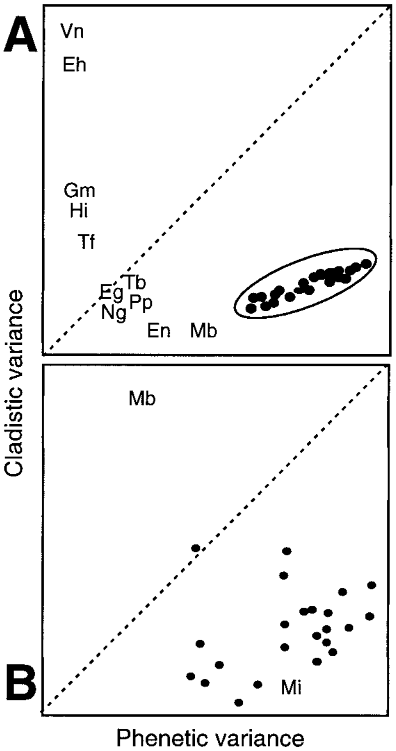

FIG. 3.—RASA taxon variance analyses. *A,* Plot of the 36-sequence eukaryotic data set. Each point represents one of the sequences in the alignment after pairwise comparisons with all other sequences. The variance in the number of uniquely shared positions (or apparently unique synapomorphies) for a given sequence is plotted against the variance in overall sequence similarity. Those sequences with more variance in apparently synapomorphic sites than is expected based on their overall pairwise similarities to all other sequences are likely to be prone to long-branch attraction. The typically sized genes from so-called "crown" eukaryotes are encircled, and sequences with excessive cladistic variance are identified by initials: Eg = *Euglena,* Eh = *Encephalatozoan,* En = *Entamoeba,* Gm = *Giardia,* Hi = *Heximita,* Mb = *Mastigamoeba balamuthi,* Ng = *Naeglaria,* Pp = *Physarum,* Tb = *Trypanosoma,* Tf = *Tritrichomonas,* and Vn = *Vairimorpha. B,* Taxon variance plot with long branches (except for *M. balamuthi* = Mb) removed. Mi = *Mastigamoeba invertens.*

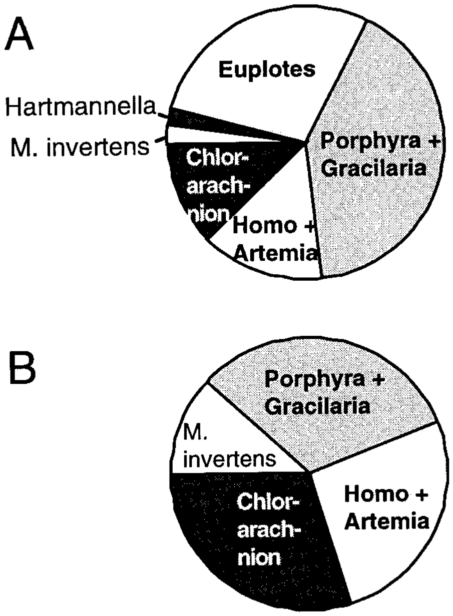

FIG. 4.—Long-branch analyses performed with most or all long-branch sequences removed from the alignment. *A,* Distribution of attraction of 100 random sequences in parsimony analyses. *B,* Distribution of attachment points for each long-branch rDNA sequence when other rDNA long branches and random sequences were excluded from the analysis. Analyses were carried out with *M. balamuthi,* along with all sequences that branched more deeply, excluded from the alignment.

bust, there is no indication that the grouping of the two *Mastigamoeba* sequences is the result of long-branch attraction. Random sequence and taxon variance analyses identify *M. balamuthi* as a potential long branch; however, *M. invertens* ssu rDNA shows no indication of long-branch tendencies (figs. 2 and 3*B*). Moreover, with *M. balamuthi* removed from the alignment, random sequences substituted in its place typically are not attracted to *M. invertens* (fig. 4*A*). Finally, when each of the long-branch rDNA sequences is tested individually with this reduced alignment, they join the tree at most of the same positions as do random sequences (fig. 4*B*) and not to the *M. invertens* branch in a majority of cases. It appears, therefore, that a monophyletic *Mastigamoeba* is recovered in spite of long-branch attraction associated with the *M. balamuthi* sequence that disrupts the relationship between the two pelobionts when still longer branches are present.

Because all more putatively ancient sequences behave as "long branches" and attach sequentially to the *M. balamuthi* sequence, a reliable rooting of the tree cannot be inferred. It is, therefore, difficult to determine what significance can be assigned to the recovery of a monophyletic pelobiont clade. Based on our results, however, rejection of pelobiont amoebae as both primitively amitochondriate and as the most ancient of extant eukaryotic lineages (e.g., Hinkle et al. 1994) is premature. Although pelobionts may indeed have lost mitochondria and be highly derived due to a secondary adaptation to low-oxygen habitats, the evidence from rDNA analyses that led to this conclusion is unreliable.

A Consistent Problem with rDNA?

A survey of previous analyses that have examined putatively ancient eukaryotes indicates that a difference in variation between crown and basal sequences is a consistent feature of rDNA trees. The variance among sequences that emerge outside the crown is always greater, and the difference is highly significant in most cases (table 2). Clearly, there is not a perfect correlation between a deviation from the "typical" 1.8-kb size and emergence at the base of the rDNA tree (Hinkle et al. 1994). For example, there are numerous animal genes more than 2 kb in length that do not cluster with the basal sequences. It is also true that many animal sequences are problematical in phylogenetic analyses (Maley and Marshall 1998) and that bilaterian metazoans do tend to be attracted toward the base of the tree when among-site rate variation is not taken into account (Van de Peer and de Wachter 1997). Whether by accident or design, however, more typically sized animal rDNA genes tend to be the ones sampled (table 2), es-

Table 2
One-Tailed Variance Ratio Test for Lengths of Crown Versus Basal rDNA Sequences in Previously Published Trees

	Crown Taxa			Basal Taxa			*F*-Test
Tree	*n*	Mean	SD	*n*	Mean	SD	*P*
Sogin et al. (1989)	9	1,827	93.7	5	1,824	471.7	0.0001
Hinkle and Sogin (1993)	10	1,845	157	10	2,029	210.7	0.19
Leipe et al. (1993).	9	1,799	19.7	11	1,893	307	$<$0.00001
Gunderson et al. (1995)	13	1,797	25.7	22	1,674	321.3	$<$0.00001
Pawlowski et al. (1996)	15	1,835	149.9	14	2,273	536.9	0.00002
Kumar and Rzhetsky (1996) . .	179	1,784	98.9	28	1,882	330.1	$<$0.00001

Note.—Sequences are scored as ‘‘crown’’ or ‘‘basal’’ based on the designations of Sogin (1997). The exception is *Plasmodium*, which is counted as a basal taxon in the Kumar and Rzhetsky (1996) data set because it does not branch with other apicomplexans but, rather, emerges in the cluster below the crown. *Plasmodium* is considered a crown taxon in all other trees in which it appears. The *F*-test provides a level of significance (*P*) for rejection of the hypothesis that the variances of the two groups of sequences are equal.

pecially when smaller data sets are examined (e.g., Sogin et al. 1989; Leipe et al. 1993; Pawlowski et al. 1996).

A correlation between deviance from the more typical rDNA size for ‘‘crown’’ eukaryotes and an increased substitution rate may help to explain the enigmatic behavior of certain protist sequences in rDNA phylogenies. For example, although they are clearly representative of apicomplexan protists, ssu rRNA genes from the parasitic genus *Plasmodium* tend to be drawn away from other apicomplexan sequences (Van de Peer and de Wachter 1997); frequently, they emerge among more basal rDNA branches in global eukaryotic phylogenetic analyses. Unlike rRNA genes from other apicomplexans, and alveolate protists in general, those in *Plasmodium* are atypically large, generally well over 2 kb in length.

Because they are characterized by a calcareous test, foraminiferans have a well-documented fossil record dating back to the early Cambrian (540 MYA) (Culver 1991). Since fossils of several ‘‘crown’’ taxa are at least twice as old (Xiao, Zhang, and Knoll 1998), the Foraminifera might be expected to nest well within the ‘‘crown’’ taxa or to branch recently from some earlier-diverging protist lineage; however, foraminiferan sequences are found among the more basal branches in rDNA analyses (Pawlowski et al. 1996). They are also some of the largest ssu rDNA sequences on record, ranging between 2.8 and 3.3 kb (Pawlowski et al. 1996). Furthermore, their detailed fossil record permits comparative rate calibrations among foraminiferan rDNA sequences. These analyses show enormous variation in evolutionary rate and indicate that some sequences have diverged up to 100 times faster than typical eukaryotic ssu rRNA genes (Pawlowski et al. 1997).

Mycetozoans branch neither together nor among the ‘‘crown’’ taxa in rDNA analyses; however, phylogenetic analyses based on protein-encoding genes indicate that the slime molds are a monophyletic group nested within the ‘‘crown taxa’’ (Loomis and Smith 1990; Baldauf and Doolittle 1997). *Physarum* ssu rDNA, which branches more deeply (fig. 1), is 1.96 kb in length, while the shorter (1.87 kb) sequence from *Dictyostelium* typically emerges closer to the crown radiation (e.g., Hinkle and Sogin 1993; Liepe et al. 1993; Gunderson et al. 1995). The size differences between these slime mold rRNA genes, along with their respective branching positions in phylogenetic analyses, are consistent with a hypothesized correlation between size variation and an increased rate of sequence evolution.

The Microsporidia are an example of a group in which extreme reduction in rDNA size is correlated with a deep but apparently incorrect position in rDNA phylogenies. Microsporidian rRNA genes typically are 200–300 bp smaller than archaebacterial sequences. Therefore, even if small rDNA size is considered an ancestral state for eukaryotes, the microsporidian genes must have undergone substantial length reductions during their evolution. Although they branch consistently near the base of rDNA trees, a growing body of evidence from other molecular, cytologic, and biochemical data suggests that microsporidia are closely related to fungi and have undergone extreme diminution due to their intracellular parasitic habit (Germot, Philippe, and Le Guyader 1997; Hirt et al. 1999).

These empirical observations do not conclusively establish a correlation between greater size variation and increased substitution rates that lead to consistent long-branch attraction at the base of the rDNA tree. Indeed, without an a priori knowledge of the branching pattern of early eukaryotes, a formal test of this hypothesis as it relates to early eukaryotic evolution may not be possible. The proposed correlation might be tested, however, using the well-sampled rDNA sequences and well-defined evolutionary relationships among metazoans. If a connection were to be found in animals between greater size variation and a tendency to disrupt known phylogenetic relationships, it would lend support to the proposition that long-branch attraction is largely responsible for the rDNA basal tree topology.

Cause or Effect?

If such a correlation does exist, large insertions or deletions in rRNA genes could be either the cause or a consequence of an increased rate of sequence evolution. The incorporation or loss of large sequence blocks might have increased selection for compensatory changes to enhance and tighten rRNA secondary structure. This seems a reasonable explanation for the apparent rate differences found between the two *Mastigamoeba* sequences. If insertions and deletions consistently result in an increase in evolutionary rate throughout the rRNA gene, their presence could account for the observed relationship between increased size variance and apparent substitution rates in deeply branching sequences (fig. 2).

In most cases, however, it is more likely that rDNA size variation is a reflection of a generally increased rate of molecular evolution, caused by independent and unrelated factors in different lineages. Asexuality and population bottlenecks (both of which have been proposed to explain increased rDNA substitution rates in symbiotic organisms; Lutzoni and Pagel 1997; Lambert and Moran 1998), relaxed selection on rDNA primary and/or secondary structure, and positive selection for sequence changes all could result in associated increases in insertion, deletion, and individual base substitution events. It is probably not coincidental that many of the most deeply branching rDNA sequences are from parasitic organisms that are likely to be subject to these kinds of effects.

Accelerated rates of sequence evolution that are due to generally altered selection regimes or population-level parameters can be expected to affect the entire genome. For this reason, the problem of how to interpret the inferred root of the eukaryotic tree is not unique to rDNA data sets. Artifactual or unreliable deep-branching sequences have been demonstrated in phylogenetic analyses based on protein-encoding genes as well, including sequences from some of the same taxa that are problematic in rDNA trees (e.g., Baldauf and Doolittle 1997; Germot, Philippe, and Le Guyader 1997; Stiller, Duffield, and Hall 1998; Hirt et al. 1999). Indeed, the rooting of eukaryotic trees based on molecular characters may be complicated by long-branch attraction in most or all cases.

A Crown or an Artificial Tail?

The analyses presented here suggest that the so-called ‘‘crown taxa’’ are a group of eukaryotes in which rDNA has undergone a more normal mode of evolution, while the long branches that precede it represent the artificial clustering of more rapidly evolving sequences. When Knoll (1992) coined the term ‘‘crown’’ eukaryotes, he also observed that this apparently recent radiation in rDNA phylogenies might be due to a reduced substitution rate in these taxa relative to more deeply branching organisms. More recently, several authors have argued that conflicts between phylogenies based on different molecular data sets, along with the poor resolution among eukaryotic groups provided by these analyses, are due to the fact that most major lineages diverged from each other over a very short period (e.g., Phillipe and Adoutte 1998).

Kumar and Rzhetsky (1996) analyzed over 200 eukaryotic ssu rRNA genes, incorporating corrections for biased base compositions, rate variation, and nonindependence of substitutions across sites, and determined that most deep branches of the rDNA tree are unreliable. They depicted the rDNA tree with a large polytomy at the base that included, with the exception of diplomonads, the same groups identified in this study as having abnormally long branches. With no statistical support and with strong indications of long-branch attraction, the depiction of a well-resolved branching topology among basal eukaryotic rDNA sequences is misleading. Any of these putatively basal eukaryotes could well be most closely related to taxa in the rDNA crown.

Regardless of their true evolutionary affiliations, there is no reliable evidence to support the rDNA model that these taxa evolved long before the common ancestor of the crown group. In contrast, recent fossil evidence (Zhu and Chen 1995; Xiao, Zhang, and Knoll 1998), phylogenetic analyses of certain protein-encoding genes (Keeling and Doolittle 1996; Budin and Philippe 1998; Stiller, Duffield, and Hall 1998; Hirt et al. 1999), and several independent molecular clock calibrations (Feng, Cho, and Doolittle 1997; Wright and Lynn 1997) all are consistent with a crown radiation that began much closer to the common origin of eukaryotes than is indicated by the rDNA tree.

Given that the archebacterial outgroup attaches to precisely the same point on the tree as do random sequences, the inferred root of the eukaryotic tree must be considered unreliable and probably the result of long-branch attraction. When the tree is rooted in this manner, the cluster of long-branch sequences becomes inverted, resulting in the familiar rDNA model of eukaryotic evolution. If this interpretation is correct, the historical pattern of ancient eukaryotic evolution has been obscured by long-branch artifacts.

Conclusions

The analyses we performed do not prove that the basal topology of the rDNA tree is inaccurate and due only to long-branch attraction. Conceivably, each of our individual observations could be reconciled with an evolutionary explanation of that topology. For the rDNA tree, or any phylogenetic hypothesis, to be useful as a scientific premise, however, it must be significantly better than the null hypothesis that inferred evolutionary relationships are due either to random effects or to a consistent bias in the data. What our analyses do show is that the branching order from the crown to the base of the rDNA tree can be explained entirely by the sequential attachment of longer and longer branches in the absence of any evolutionary signal. Moreover, a ‘‘long-branch hypothesis’’ better explains the positions of several taxa that appear to branch incorrectly on the rDNA tree. In other words, the null hypothesis cannot be rejected and is actually more consistent with the data than is the alternative hypothesis of an rDNA basal topology inferred from evolutionary signal.

Before the advent of molecular phylogenetic analyses, there were few characters available for inferring relationships among major groups of eukaryotes. As a result, the eukaryotic tree was often depicted as a largely unresolved radiation from an unknown ancestor (Keeton 1980). Large-scale phylogenetic analyses based on rDNA sequences have provided a topology of inferred relationships that is a useful working hypothesis to be tested against independent data sets.

The sequencing of multiple genes from diverse eukaryotic and prokaryotic organisms has resulted in many conflicts with the rDNA hypothesis of cellular evolution. In order to conform to the rDNA model, lateral gene transfer now is often invoked to explain the lack of harmony between different gene trees; this explanation has even led to the proposal that genetic transfer predominated over geneological descent during the early stages of cellular evolution (Woese 1998). Given that the widely accepted rDNA model of early eukaryotic evolution may be an artifact of uneven substitution rates, it seems premature to make such broad judgments about ancient evolutionary events on the basis of conflicting gene trees. It may be more prudent to first determine the ultimate reliability of the trees themselves.

Acknowledgments

We thank S. Edwards, S. Graham, A. Knoll, and R. Olmstead for critical review of the manuscript, E. Duffield for culturing efforts, and B. Oxelman for helpful suggestions.

LITERATURE CITED

Baldauf, S. L., and W. F. Doolittle. 1997. Origin and evolution of the slime molds (Mycetozoa). Proc. Natl. Acad. Sci. USA **94**:12007–12012.

Brugerolle, G. 1991. Flagellar and cytoskeletal systems in amitochondrial flagellates: Archamoeba, Metamonada and Parabasala. Protoplasma **164**:70–90.

Budin, K., and H. Philippe. 1998. New insights into the phylogeny of eukaryotes based on ciliate Hsp70 sequences. Mol. Biol. Evol. **15**:943–956.

Cavalier-Smith, T. 1991. Archamoebae: the ancestral eukaryotes? Biosystems **25**:25–38.

Culver, S. J. 1991. Early Cambrian Foraminifera from West Africa. Science **254**:689–691.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **25**: 401–410.

———. 1989. PHYLIP—phylogenetic inference package (version 3.2). Cladistics **5**:164–165.

Feng, D.-F., G. Cho, and R. F. Doolittle. 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. USA **94**:13028–13033.

Germot, A., H. Philippe, and H. Le Guyader. 1997. Evidence for loss of mitochondria in Microsporidia from a mitochondrial-type HSP 70 in *Nosema locustae.* Mol. Biochem. Parasitol. **87**:159–168.

Graham, S. 1997. Phylogenetic analysis of breeding-system evolution in heterostylous monocotyledons. Ph.D. dissertation, University of Toronto, Canada.

Gunderson, J., G. Hinkle, D. Leipe, H. G. Morrison, S. K. Stickel, D. A. Odelson, J. A. Breznak, T. A. Nerad, M. Müller, and M. L. Sogin. 1995. Phylogeny of trichomonads inferred from small-subunit rRNA sequences. J. Eukaryot. Microbiol. **42**:411–415.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature **326**:411–414.

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

Hinkle, G., D. D. Leipe, T. A. Nerad, and M. L. Sogin. 1994. The unusually long small subunit ribosomal RNA of *Phreatamoeba balamuthi.* Nucleic Acids Res. **22**:465–469.

Hinkle, G., and M. L. Sogin. 1993. The evolution of the Vahlkampfiidae as deduced from 16s-like ribosomal RNA analysis. J. Eukaryot. Microbiol. **40**:599–603.

Hirt, R. P., J. M. Logsdon, B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. USA **96**:580–585.

Keeling, P. J., and W. F. Doolittle. 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol. Biol. Evol. **13**:1297–1305.

Keeton, W. T. 1980. Biological science. 3rd edition. W. W. Norton and Company, New York.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29**:170–179.

Knoll, A. H. 1992. The early evolution of eukaryotes: a geological perspective. Science **256**:622–627.

Kumar, S., and A. Rzhetsky. 1996. Evolutionary relationships of eukaryotic organisms. J. Mol. Evol. **42**:183–193.

Lambert, J. D., and N. A. Moran. 1998. Deleterious mutations destabilize ribosomal RNA in endosymbiotic bacteria. Proc. Natl. Acad. Sci. USA **95**:4458–4462.

Leipe, D. D., J. H. Gunderson, T. A. Nerad, and M. L. Sogin. 1993. Small subunit rimosomal RNA$^+$ of *Hexamita inflata* and the quest for the first branch of the eukaryotic tree. Mol. Biochem. Parisitol. **59**:41–48.

Loomis, W. F., and D. W. Smith. 1990. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. Proc. Natl. Acad. Sci. USA **87**:9093–9097.

Lutzoni, F., and M. Pagel. 1997. Accelerated evolution as a consequence of transitions to mutualism. Proc. Natl. Acad. Sci. USA **94**:11422–11427.

Lyons-Weiler, J., and G. A. Hoelzer. 1997. Escaping from the Felsenstein zone by detecting long branches in phylogenetic data. Mol. Phylogenet. Evol. **8**:375–384.

Lyons-Weiler, J., G. A. Hoelzer, and R. J. Tausch. 1996. Relative apparent synapomorphy analyis (RASA). I. The statistical measurement of phylogenetic signal. Mol. Biol. Evol. **13**:749–757.

Maddison, W. P., and D. R. Maddison. 1992. MacClade: analysis of phylogeny and character evolution. Version 3.0. Sinauer, Sunderland, Mass.

Maidak, B. L., G. J. Olsen, N. Larsen, R. Overbeek, M. J. McCaughey, and C. R. Woese. 1997. The RDP (Ribosomal Database Project). Nucleic Acids Res. **25**:109–111.

Maley, L. E., and C. R. Marshall. 1998. Evolution: the coming of age of molecular systematics. Science **279**:505–506.

Pawlowski, J., I. Bolivar, J. Fahrni, T. Cavalier-Smith, and M. Gouy. 1996. Early origin of Foraminifera suggested by SSU rRNA gene sequences. Mol. Biol. Evol. **13**:445–450.

Pawlowski, J., I. Bolivar, J. F. Fahrni, C. de Vargas, M. Gouy, and L. Zaninetti. 1997. Extreme differences in rates of molecular evolution of Foraminifera revealed by comparison of ribosomal DNA sequences and the fossil record. Mol. Biol. Evol. **14**:489–505.

Phillipe, H., and A. Adoutte. 1998. The molecular phylogeny of Eukaryota: solid facts and uncertainties. Pp. 25–56 *in* G. H. Coombs, K. Vickerman, M. A. Sleigh, and A. Warren, eds. Evolutionary relationships among protozoa. Chapman and Hall, London.

SIDDALL, M. E., H. HONG, and S. S. DESSER. 1992. Phylogenetic analyses of the Diplomonadida (Wenyon, 1926) Bergerolle, 1975: evidence for heterochrony in protozoa and against *Giardia lamblia* as a ''missing link.'' J. Protozool. **3**:361–367.

SIMPSON, A. G. B., C. BERNARD, T. FENCHEL, and D. J. PATTERSON. 1997. The organisation of *Mastigamoeba schizophrenia* n. sp.: more evidence of ultrastructural idiosyncrasy and simplicity in pelobiont protists. Eur. J. Protistol. **33**:87–98.

SOGIN, M. L. 1997. History assignment: when was the mitochondrion founded. Curr. Opin. Genet. Devel. **7**:792–799.

SOGIN, M. L., J. H. GUNDERSON, H. J. ELWOOD, R. A. ALONSO, and D. A. PEATTIE. 1989. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia.* Science **243**:75–77.

STILLER, J. W., E. C. S. DUFFIELD, and B. D. HALL. 1998. Amitochondriate amoebae and the evolution of DNA-dependent RNA polymerase II. Proc. Natl. Acad. Sci. USA **95**:11769–11774.

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

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

TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37**:221–244.

VAN DE PEER, Y., and R. DE WACHTER. 1997. Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site variation in 18s rRNA. J. Mol. Evol. **45**: 619–630.

WHITE, T. J., T. BRUNS, S. LEE, and J. TAYLOR. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pp. 315–322 *in* M. A. INNIS, D. H. GELFAND, J. J. SNINSKY, and T. J. WHITE, eds. PCR protocols: a guide to methods and applications. Academic Press, San Diego.

WOESE, C. 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA **95**:6854–6859.

WRIGHT, A. D. G., and D. H. LYNN. 1997. Maximum ages of ciliate lineages estimated using a small subunit rRNA molecular clock: crown eukaryotes date back to the paleoproterozoic. Arch. Protistenkd. **148**:329–341.

XIAO, S., Y. ZHANG, and A. H. KNOLL. 1998. Three-dimensional preservation of algae and animal embryos in a Neoproterozoic phosphorite. Nature **391**:553–558.

ZHU, S., and H. CHEN. 1995. Megascopic multicellular organisms form the 1700-million-year old Tuanshanzi formation in the Jixian area, north China. Science **279**:620–6222.

MASAMI HASEGAWA, reviewing editor

Accepted June 3, 1999