Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 580–585, January 1999
Evolution

Microsporidia are related to Fungi: Evidence from the largest subunit of RNA polymerase II and other proteins

ROBERT P. HIRT*†, JOHN M. LOGSDON, JR.†‡, BRYAN HEALY*, MICHAEL W. DOREY‡, W. FORD DOOLITTLE‡, AND T. MARTIN EMBLEY*§

*Department of Zoology, The Natural History Museum, London SW7 5BD, United Kingdom; and ‡Canadian Institute for Advanced Research, Program in Evolutionary Biology and Department of Biochemistry, Halifax, NS, B3H 4H7, Canada

Edited by Estella B. Leopold, University of Washington, Seattle, WA, and approved December 1, 1998 (received for review August 12, 1998)

ABSTRACT **We have determined complete gene sequences encoding the largest subunit of the RNA polymerase II (RBP1) from two Microsporidia, *Vairimorpha necatrix* and *Nosema locustae*. Phylogenetic analyses of these and other RPB1 sequences strongly support the notion that Microsporidia are not early-diverging eukaryotes but instead are specifically related to Fungi. Our reexamination of elongation factors EF-1α and EF-2 sequence data that had previously been taken as support for an early (Archezoan) divergence of these amitochondriate protists show such support to be weak and likely caused by artifacts in phylogenetic analyses. These EF data sets are, in fact, not inconsistent with a Microsporidia + Fungi relationship. In addition, we show that none of these proteins strongly support a deep divergence of Parabasalia and Metamonada, the other amitochondriate protist groups currently thought to compose early branches. Thus, the phylogenetic placement among eukaryotes for these protist taxa is in need of further critical examination.**

Microsporidia are highly specialized eukaryotic unicells, living only as obligate intracellular parasites of other eukaryotes (1). They lack the mitochondria and peroxisomes typical of most eukaryotes. Thus, in 1983 Cavalier-Smith (2) included the Microsporidia with other amitochondriate protists, Parabasalia (e.g., *Trichomonas*), Metamonada (e.g., *Giardia*), and Archamoebae (e.g., *Entamoeba*) in the kingdom Archezoa. These protists were presumed to have diverged from other eukaryotes before the acquisition of mitochondria and were suggested as the earliest eukaryotic lineages.

Phylogenetic trees based initially on small-subunit ribosomal RNA (SSUrRNA) (3) and then on protein translation elongation factor (EF-1α and EF-2) (4, 5) sequences showed that Microsporidia indeed diverged early, along with the Parabasalia and Metamonada. Thus, these data apparently confirmed the archezoal hypothesis in general and what we call the "Microsporidia-early" hypothesis in particular—Archamoebae were eliminated from the Archezoa when their SSUrRNAs neither placed them together nor as early branches (6).

However, more recent trees constructed from tubulin (7, 8) and Hsp70 data (9, 10) placed Microsporidia in the eukaryotic "crown," favoring a position within, or as the sister group to, Fungi. Although the support for the "Microsporidia + Fungi" hypothesis (or M + F) from the Hsp70 data is not compelling [when judged by maximum-likelihood (ML) difference tests or other support values] (9, 10), support from α-tubulin is firm (7). If Microsporidia are truly related to Fungi, then their simplified cell structures and small genomes are degenerate features permitted, or perhaps encouraged, by their parasitic lifestyles.

Here we report sequences for the largest subunit of RNA polymerase II (RPB1) from two Microsporidia, *Vairimorpha necatrix* and *Nosema locustae*. Several phylogenetic methods applied to these and other RPB1 sequences strongly support a close relationship of Microsporidia and Fungi. A reanalysis of the apparently conflicting EF data show that the support that these sequences lend to a deeply diverging Microsporidia is weak and attributable to artifacts. Furthermore, the EF-1α gene from the microsporidian *Glugea plecoglossi* (5) carries an insertion encoding 11 amino acids that is otherwise only found in, and is diagnostic for, the EF-1α genes of Fungi and Metazoa (animals) (11), which form a clade based on other criteria (6, 11).

An alternative hypothesis to reconcile the apparently conflicting gene trees and to allow Microsporidia to retain their early status is that they might have borrowed genes from host genomes (6). We conclude that such chimeric theories are not necessary because there is no significant conflict between these proteins concerning the position of Microsporidia. Where a convincing phylogenetic signal is present, it relates them to Fungi.

Numerous recent reports indicate that the Microsporidia, Metamonada, and Parabasalia possess nuclear genes of α-proteobacterial provenance whose products normally (in mitochondriate eukaryotes) serve mitochondrial functions; the presence of such genes indicates mitochondrial loss from these protists (9, 10, 12–16). Although the ancestral presence of mitochondria does not itself preclude a deep divergence of any of these taxa, our analyses of RPB1 and reanalyses of EF data (in addition to challenging the deep placement of Microsporidia) show that the support for an early divergence for *Trichomonas* and *Giardia* is also weaker than generally supposed. The inferred early branching positions of Metamonada and Parabasalia largely depend on a single data set, i.e., SSUrRNA. The Microsporidia thus may be the only member of the (now former) Archezoa (*sensu* Cavalier-Smith, ref. 17) about whose phylogenetic position we now have confidence based on multiple molecular data sets.

MATERIALS AND METHODS

Isolation of Genomic Clones. Two sets of oligonucleotide primers—RPB1-F1 (cgGACTTYGAYGGNGAYGARATG-

This paper was submitted directly (Track II) to the *Proceedings* office.

Abbreviations: BP, bootstrap proportions; CSR, constant-site removal; ISR, invariant-site removal; CTD, C-terminal domain; EF, elongation factor (protein translation); LBP, local bootstrap proportions; ML, maximum likelihood; MP, maximum parsimony; RPB1, RNA polymerase II largest subunit; SSUrRNA, small subunit ribosomal RNA; M + F, Microsporidia + Fungi; FSR, fast-site removal.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF060234 and AF061288).

†R.P.H. and J.M.L. contributed equally to this work.

§To whom reprint requests should be addressed at: Department of Zoology, Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom. e-mail: tme@nhm.ac.uk.

A)/R1 (CCCGCKNCCNCCCATNGCRTGRAA) (codons = capital letters) and RPB1-F2 (cgcgATHGASYACIGCNGTNAARAC)/R2 (ccggGTCATYTGNGTNGCNGGYTC) amplified ≈1.1-kbp and 1.2-kbp fragments of RPB1 from *V. necatrix* genomic DNA (9). Amplicons were used as probes in Southern blots to confirm the source of the single-copy *V. necatrix* RPB1 gene and to screen a *V. necatrix Eco*RI genomic library (9). Two *Eco*RI fragments were cloned to obtain the full-length RPB1 gene that was sequenced on both strands by primer walking.

Oligonucleotide primers RPB1-F4 (CTACGTGGCAAGYTNATGGG)/RPB1-R3 (AGACCTTCACGNCCNWCCAT) were used to amplify a ≈1,500-bp fragment of RPB1 from *N. locustae* genomic DNA. Nested amplification with primers RPB1 F3 (GCATTCGATGGCGAYGARATG)/RPB1-R3 resulted in a ≈1-kbp fragment. This was used as a probe to isolate clones from a *N. locustae Sau*IIIA partial-digest genomic library (18). One clone containing the entire RPB1 gene was completely sequenced on both strands.

Sequence Alignments. The inferred amino acid sequences of the two microsporidian RPB1 sequences were aligned to published RPB1, RPA1 (RNA polymerase I largest subunit), and RPC1 (RNA polymerase III largest subunit) sequences, by using CLUSTALW Version 1.7 (19) with further manual adjustments considering previous alignments (20). Positions that could not be aligned unambiguously, insertions/deletions, or missing data present in three or more taxa, were removed from phylogenetic analysis by using the mask facility in GDE Version 2.2 (21). Alignments of elongation factor EF-1α and EF-2 protein sequences were supplied by M. Hasegawa (Institute Statistical Mathematics, Tokyo) and were those that previously supported the Microsporidia-early hypothesis (4, 5). DNA sequences were aligned to the existing protein alignments by using PUTGAPS (J. MacInerney, Natural History Museum, London). All alignments are available from R.P.H. (e-mail: rch@nhm.ac.uk).

Testing for Potential Amino Acid and Nucleotide Compositional Biases. Amino acid and/or nucleotide compositional biases in molecular data can distort phylogenetic analyses, causing taxa that share similar compositions to cluster together irrespective of their true relationships (22–24). To investigate whether compositional biases were potentially influencing tree topologies, we used MOLPHY Version 2.3 (25) or SPECTRUMPPC (26) to make trees based solely on distances calculated from amino acid and nucleotide frequencies of variable sites. We also compared amino acid and nucleotide frequencies by using a 5% χ^2 test in PUZZLE Version 4.0 (27).

Phylogenetic Analyses of Protein Sequences. Protein ML trees were inferred with PROTML by using the JTT-F substitution model in MOLPHY Version 2.2 (28). Bootstrap support was estimated from 100 resampled data sets (SEQBOOT, PHYLIP Version 3.52c), and a heuristic search on each one was carried out. A full likelihood analysis identified the ML tree from each search, and a majority consensus tree was calculated from all 100 ML trees. Support for conflicting hypotheses of relationships also was investigated by using the Kishino–Hasegawa test (29) to compare differences in log-likelihood for different trees. The trees were generated by using constrained analyses in PAUP* Version 4.0 d64 (D. L. Swofford, personal communication) to identify maximum parsimony (MP) trees for different hypotheses of relationship; these trees then were supplied as user trees to PROTML for likelihood calculations.

The Influence of Site-by-Site Rate Variation on Hypothesis Testing in Protein ML Analyses. Failure to correct for site-by-site rate variation can lead to the wrong tree being selected from molecular data (30). For example, methods of analysis that assume that all sites are free to vary will undercorrect for change when sites that cannot change, termed invariant (30), are present. The magnitude of this error can be expected to be particularly severe for very dissimilar sequences, as might be expected if these sequences diverged a long time ago (30–33). All of the data sets we investigated contained sites that are constant in all taxa in the alignment and thus potentially invariant. ML models in PROTML have no site-by-site rate correction and so may be susceptible to this source of error. We therefore investigated the effect of reducing site rate heterogeneity for PROTML analyses by editing the data to remove the category of fastest evolving sites (fast-site removal, FSR) or a fraction of inferred invariant sites (invariant-site removal, ISR; ref. 30) before phylogenetic analysis. The fraction of invariant sites was estimated by using a two-site rate (either variable or invariant) ML model with the JTT-F substitution matrix in PUZZLE Version 4.0. For FSR we used PUZZLE Version 4.0 to estimate a discrete γ distribution for each data set (comprising one invariant-site rate and eight variable-site rates) under the JTT-F model. Because tree topology can affect these calculations, the rates for the RPB1 data set were calculated over two MP trees (themselves constrained to represent either the M + F or the Microsporidia-early hypothesis). Sites in the fastest rate category common to both trees were excluded from the PROTML phylogenetic analyses. Rate categories for the EF-1α and EF-2 protein data sets were calculated by using published ML trees (which placed Microsporidia early) (4, 5).

Analysis of DNA Sequences. Mutational saturation of sequences by superimposed nucleotide substitutions can mask historical signal, causing severe problems for phylogenetic inference (24, 34). We investigated whether any of our DNA sequences were affected by this problem by plotting transitions against transversions for all pairwise comparisons and all codon positions; clustering of points indicates potential saturation (24). These analyses suggested that codon position 3 is saturated in all of the data sets (position 3 also showed the most extreme base composition variation), so we excluded it from phylogenetic analyses.

Because analyses of nucleotide compositions (see above) indicated base composition heterogeneity between DNA sequences for all of our data sets, we used a phylogenetic method, LogDet/Paralinear distances, which is reported to be able to recover the correct tree under such conditions (22, 23). However, like PROTML discussed above, the LogDet/Paralinear distance method does not incorporate a correction for site-by-site rate variation but also assumes that all sites are free to vary. We therefore investigated the effects on tree topology of reducing rate heterogeneity by removing different fractions of constant sites before analysis (constant-site removal, CSR; ref. 33). In addition, we used an ML method in PAUP* to estimate the proportion of sites actually free to vary across our alignments (30, 35). These sites are henceforth referred to as the variable sites in LogDet analyses and they include a small fraction (typically 7% or less depending on the data set) of the sites observed as constant for our taxon sampling. All distance trees were constructed by using minimum evolution, and the data were bootstrapped 1,000 times.

RESULTS AND DISCUSSION

Microsporidian RPB1 Genes and Inferred Protein Sequences. The RPB1 genes for *V. necatrix* and *N. locustae* include ORFs, uninterrupted by introns, of 1,606 codons (4,818 bp) and 1,554 codons (4,662 bp), respectively. Southern blot analysis indicated the presence of a single copy of the RPB1 gene in the *V. necatrix* genome (data not shown). Conserved RPB1 domains (A–H) were present in both microsporidial sequences as were the majority of amino acids conserved in all published RPB1 sequences, particularly those in the zinc-binding domain (ref. 36 and references therein). Interestingly, the C-terminal domains (CTD) of the microsporidial sequences contain characteristic heptapeptide repeats: 17 YSPTSPT repeats in *V. necatrix* and 13 YSPTSPA repeats in *N. locustae*. These repeats also occur in fungal, animal, plant, and some protist RPB1 protein sequences (36, 37) but are absent from the protists *Trichomonas*, *Trypanosoma,* and *Giardia* (38) as well as from the red algae *Bonnemaisonia* and *Porphyra* (39). Thus, the presence of these CTD repeats is consistent with the inferred fungal relationship (see below). The CTD has been implicated in the processing of spliceosomal introns from pre-mRNA (40), so it is interesting that spliceosomal snRNAs have now been reported from both *V. necatrix* and *N. locustae* (ref. 18 and references therein), and a spliceosomal intron has recently been discovered in the microsporidian *Encephalitozoon cuniculi* (41). With the exception of red algae, the known presence of spliceosomal introns in eukaryotes (ref. 42; J.M.L., unpublished data) is perfectly correlated with the possession of CTD heptapeptide repeats (37, 39).

RPBI Protein and DNA Sequences Support a Relationship Between Microsporidia and Fungi. The protein ML tree (Fig. 1) placed the two Microsporidia together with the fungal RPB1 sequences (M + F) with strong bootstrap support in all analyses. Use of the Kishino–Hasegawa (29) test to compare the statistical significance of different trees also supports M + F. In fact, the one tree we found that could not be rejected at the 0.05 level and that placed Microsporidia early also placed Fungi as the next deepest branch. Such a deep position for Fungi conflicts with analyses of several proteins (11) and SSUrRNA (6) that support a Fungi + Metazoa relationship. Indeed, we also recovered a Fungi (+ Microsporidia) relationship with Metazoa in our protein ML tree, albeit with weak support. When Fungi are constrained with Metazoa (as most data would have them), trees in which Microsporidia branched before *Trichomonas*, *Giardia,* or *Trypanosoma* could all be rejected at the 0.05 level.

Reduction of rate heterogeneity between sites did not reduce support for M + F with protein ML (Fig. 1; data not shown), whereas removal of fast-evolving sites dramatically increased bootstrap support from 41% to 77% for M + F in MP analyses. This last result is consistent with the hypothesis that parsimony is particularly sensitive to long-branch effects caused by unequal substitution rates (43). Analyses of RPB1 DNA sequences also support a relationship between Microsporidia and Fungi. Application of the LogDet transformation to variable sites at coding positions 1 + 2 produced a tree (not shown) where M + F was supported with BP 74%, rising to BP 93% in the absence of the outgroup.

In summary, a relationship between Microsporidia and Fungi is strongly supported by our analyses of the RPB1 data sets, and this relationship cannot be attributed to shared amino acid or nucleotide biases, mutational saturation, or long-branch effects. Because RPB1 appears robust in supporting M + F, we were interested in how inferences from the elongation factors, EF-1α and EF-2, which apparently support Microsporidia-early, would stand up in the face of the same analyses.

Do Elongation Factors Support M + F or Microsporidia-Early? Our ML analysis of the EF-2 protein sequences gave a tree similar to the one recently published for the same data set (4) where the microsporidian *Glugea plecoglossi* is a long branch at the base of the eukaryote clade. However, bootstrap support (with resampled datasets) for *Glugea* as first branch was only 54%, and support was further reduced to 33% when invariant sites were removed. This is in striking contrast to published local bootstrap support (75%) from the same data set for the basal position of *Glugea* (4). However, local bootstrap support can only be interpreted as bootstrap probabilities of a particular internal branch when the other parts of the tree are correct (25), an assumption that is likely not met for these data.

We surmised that a common amino acid bias with some outgroups may be influencing the observed deep position of the long *Glugea* branch relative to other eukaryotes, because the amino acid frequency tree for the aligned sequences grouped *Glugea* with the outgroup Archaea *Sulfolobus* and *Methanococcus*. To investigate whether base-compositional effects and/or long-branch attraction was contributing to the observed deep position for *Glugea*, we removed the archaeal outgroup sequences. We also removed the category of fastest evolving sites because these are expected to contribute most to any long-branch effect (33). Consistent with our hypothesis that *Glugea* is branching deep because of artifact, ML and MP analyses both recovered M + F (Fig. 2*A*), albeit with weak bootstrap support (ML = 29%, MP = 40%).

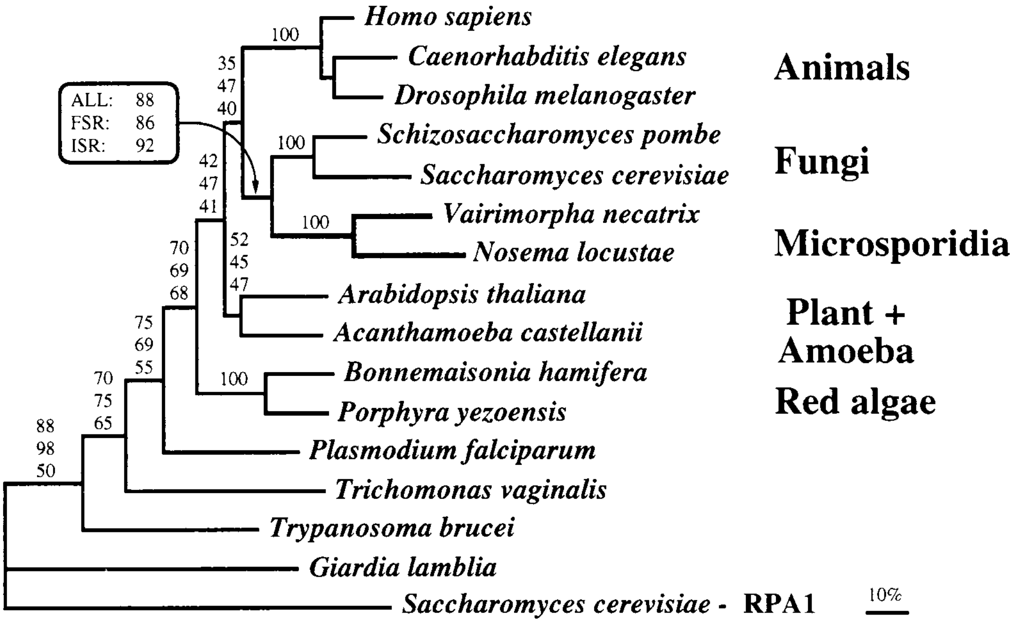

FIG. 1. Phylogenetic analyses of RPB1. Effects on bootstrap support of ISR or FSR were considered by using protein ML. The tree shown is the ML consensus tree topology from analysis of 760 aligned positions for 15 RPB1 sequences and 1 outgroup RPA1 sequence. Values report bootstrap support from ML analyses for all 760 sites (ALL sites), 669 sites (FSR, where the fastest evolving sites common to the ML tree and a tree where Microsporidia are at the base of the eukaryotes were removed), and 645 sites (ISR). Where only a single bootstrap value is shown, support was 100% in all analyses. The scale bar represents 10% estimated sequence divergence under the JTT-F model.

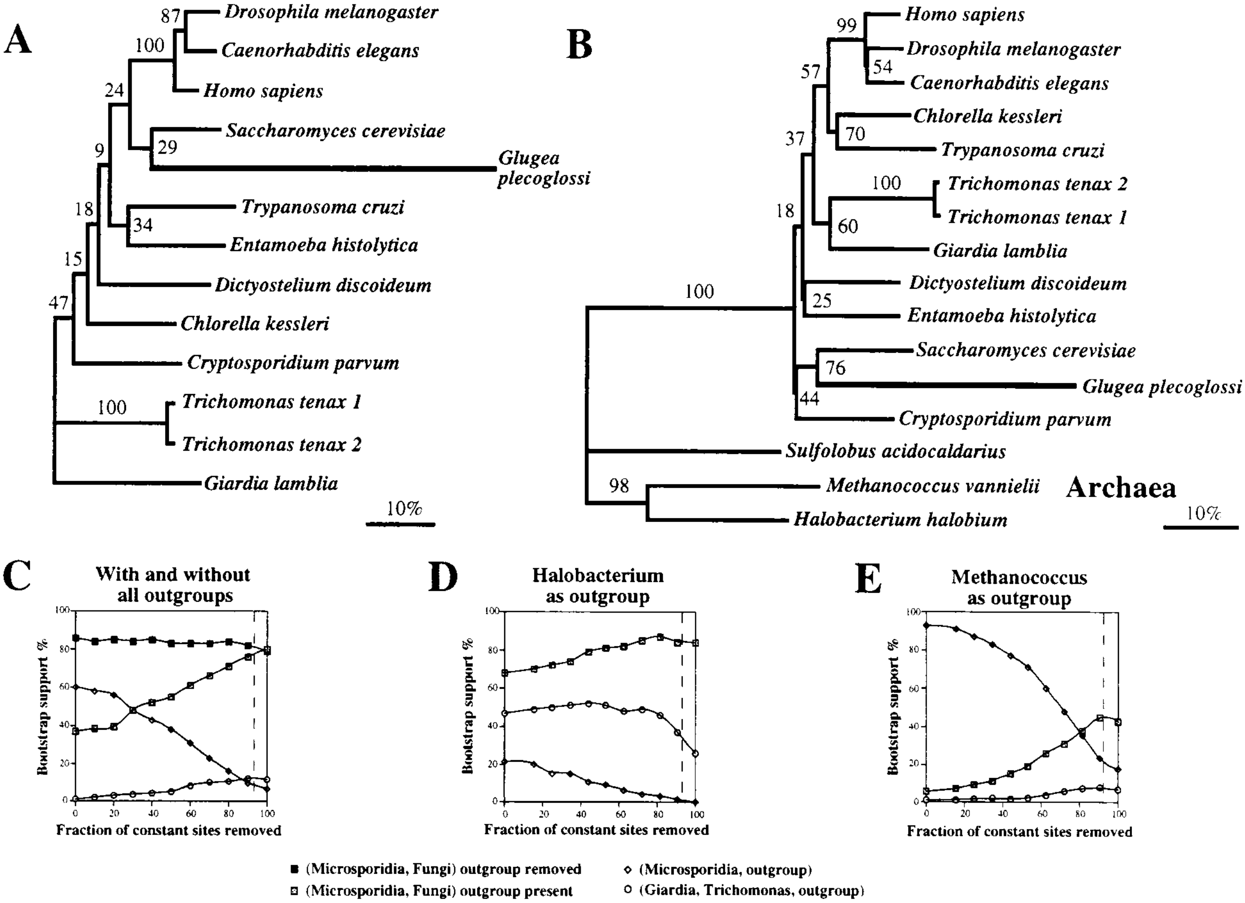

FIG. 2. Phylogenetic analyses of EF-2. (*A*) Protein ML tree with outgroups removed and FSR (72 sites). The scale bar represents 10% estimated sequence divergence under the JTT-F model. (*B–E*) Phylogenetic analyses of EF-2 DNA codon positions 1 + 2 using LogDet and investigating the influence of constant sites and choice of outgroup on bootstrap support. The DNA alignment was derived from a 542-aa alignment. (*B*) Bootstrap consensus tree topology for LogDet distances estimated from 812 variable sites (93% of constant sites removed as invariant) indicating support for M + F. The effect of incremental removal of observed constant sites on bootstrap support for different relationships in the presence or absence of all 3 outgroups (*C*), with *Halobacterium halobium* as outgroup (*D*), and with *Methanococcus vannielii* as outgroup (*E*). The dashed vertical lines represent the ML estimates of invariant sites: 93%, 92%, and 93% of constant sites for *C*, *D*, and *E*, respectively (see text for discussion).

Analyses of transitions versus transversions for EF-2 codon positions 1 or 2 (data not shown) suggest that Microsporidia sequences are potentially saturated. Furthermore, analyses of base compositions suggest that shared biases may potentially influence tree topology. For example, nucleotide frequency trees clustered *Glugea* with the outgroup Archaea *Sulfolobus* and *Methanococcus*, with which it shares a similar base composition (interestingly, *Giardia* and *Trichomonas* clustered with the other outgroup, *Halobacterium*). We therefore investigated the effect of reducing site rate heterogeneity by using progressive CSR and outgroup choice on support for M + F or Microsporidia-early from the EF-2 DNA dataset (Fig. 2 *C–E*).

ML estimates suggested that 93% of constant sites should be removed from the EF-2 dataset for LogDet analyses if the assumption of the method—that all sites can vary—is not to be violated (22, 23, 33). Under these conditions, the LogDet tree recovered *G. plecoglossi* as the sister group to *Saccharomyces* with 76% bootstrap support (Fig. 2*B*). In the absence of outgroups, the M + F relationship was strongly supported in all analyses (Fig. 2*C*). In the presence of outgroups, there were two conflicting positions for Microsporidia detected among bootstrap partitions: (*i*) Microsporidia-early when all or most constant sites were included and (*ii*) M + F, which increased in support as constant sites were removed (Fig. 2*C*). This suggests that EF-2 support for Microsporidia-early is mainly caused by a failure to adequately account for site-by-site rate variation coupled with an outgroup-attraction effect. This last phenomenon is clearly related to the base composition of the outgroup. When *Halobacterium* was used as outgroup, M + F was recovered even when all constant sites were included, and there was never any strong support for Microsporidia-early (Fig. 2*D*). The Microsporidia-early hypothesis was strongly supported only when (*i*) outgroup taxa, i.e., *Sulfolobus* (data not shown) or *Methanococcus* (Fig. 2*E*), shared the same base composition bias as *Glugea* and (*ii*) too many constant sites were included in the analysis.

Protein ML analyses of the EF-1α data set produced similar trees as previously published (4) with moderate (64%) bootstrap support for *Glugea* as the first branch. However, pairwise comparisons for EF-1α codon positions 1 and 3 (data not shown) indicate potential saturation with all comparisons involving *Glugea* strongly clustered and separated from other among-eukaryote comparisons. The pattern for changes at replacement position 2 is even more extreme where the clustered points for *Glugea* are outside the range for the other eukaryote and Archaea outgroup comparisons (data not shown). Because all changes at position 2 result in a change of amino acid, the use of the *Glugea* sequence for phylogenetic inference is compromised, and trees inferred from this data set must be considered unreliable. Consistent with the hypothesis that the *Glugea* EF-1α sequence is behaving in a manner that is different from the other eukaryote sequences are observations (44) that it contains many nonconservative amino acid substitutions at otherwise universally conserved positions. Moreover, some of these affect active-site residues, where change may not be compatible with enzyme function.

In summary, our analyses of the EF-2 and EF-1α data sets indicate that there are potentially severe problems for reconstructing some eukaryotic relationships from these proteins, results which cast serious doubt on their support for Microsporidia-early. The use of local bootstrap values has apparently inflated the perceived support for deep-branching relationships from EF-2; support for Microsporidia-early over M + F can be attributed to a failure to deal adequately with rate variation or compositional biases. Furthermore, our analyses indicate substitution saturation is a problem with the EF data sets, especially for EF-1α. Therefore, M + F cannot be significantly excluded by either EF-2 or EF-1α; in some analyses EF-2 actually supports M + F (Fig. 2). Clear support for M + F from EF-1α comes from the presence of an insertion in the *Glugea* EF-1α (5), which is in the same position in animals and Fungi (11, 44).

Do RPB1, EF-2, and EF-1α Sequences Support the Deep Divergence of *Giardia* and *Trichomonas*? We have also addressed whether these proteins provide support for the early divergence of the diplomonad, *Giardia* and the parabasalid, *Trichomonas*, the other amitochondriate protists in the Archezoa. In SSUrRNA trees, *Giardia* and *Trichomonas* consistently branch deep (3), and it has been strongly argued that *Giardia* in particular represents an ancient offshoot (6).

Our PROTML analyses of the RPB1 protein data sets suggest that strong bootstrap support for such deep relationships depends on the inclusion of invariant sites and is much reduced ($\leq$65%) in their absence (Fig. 1). The proportion of constant sites analyzed also affected support for *Giardia* plus the outgroup in LogDet distance analyses of the RPB1/RPA1 DNA sequences (data not shown). Our ML estimates indicate that 98% of constant sites should be excluded from the LogDet analysis; under these conditions, support for *Giardia* as the deepest branch was 41% from bootstrapping and support for a partition of *Giardia*, *Trypanosoma*, and *Trichomonas* from the other eukaryotes was only 26%. By calling into question the deep positions of *Giardia* and *Trichomonas* on the RPB1 tree, our results are entirely consistent with and complementary to the recent RPB1 analyses of Stiller *et al.* (37).

Published EF-2 protein trees (4) have recorded high ($\geq$83%) local bootstrap probabilities for *Giardia* and *Trichomonas* branching deep relative to outgroup Archaea. However, we found little support for this (BP $\leq$ 29%) from our own bootstrapping using resampled data sets. Analysis of EF-2 DNA sequences using LogDet with constant-site removal (Fig. 2 *B–E*) suggests that support for *Giardia* and *Trichomonas* branching deeper than other eukaryotes can be attributed almost entirely to violations of the method's assumption that all sites can vary or to outgroup attraction. For EF-1α, a recent comprehensive analysis of protein sequences (4) found only 46% local BP support for *Giardia* and *Trichomonas* branching deeper than other eukaryotes. Indeed, by removing invariant sites we found even lower support (BP = 16%) for this hypothesis.

Summary and Conclusions. The Archezoa hypothesis (2) posited that the amitochondriate Microsporidia, Metamonada and Parabasalia (*i*) lack mitochondria because they diverged from the rest of the eukaryotes before the acquisition of these organelles and thus (*ii*) compose the earliest lineages of eukaryotes. Even though evidence now suggests that none of these three protist groups are primitively amitochondriate (9, 10, 12–16), other data have often been taken as independent support for their early divergence. The deepest branches on eukaryotic SSUrRNA trees are consistently those leading to the amitochondriate Microsporidia, Metamonada and Parabasalia (3, 6, 17, 45). Several analyses of EF sequence data have also supported these early branchings (4, 5). The ultrastructural simplicity of these cells has certainly been interpreted to suggest primitivity. Other microsporidial features interpretable as primitive include (in addition to absence of mitochondria) fusion of 5.8S and 23S rRNA, possession of 70S ribosomes, and lack of peroxisomes and 9 + 2 microtubule structures (17, 46).

Phylogenetic trees are clearly central to our efforts to understand early eukaryote evolution, but phylogenetic reconstruction at such depth is difficult: phylogenetic signals are potentially weak, and no method is insensitive to noise. The deep position of Microsporidia (Microsporidia-early) was first challenged by tubulin data, which placed them with Fungi (7, 8). Tubulin trees are sometimes distrusted however, because of apparent long-branch effects (45) and because these proteins are eukaryote-specific (and thus cannot be properly outgroup-rooted) (7); the possibility of lateral transfer of tubulin genes has also been suggested (6). Thus the strong support for M + F given by the RPB1 data set (and the very strong rejection by this data of any early Microsporidial divergence) provide necessary and compelling support for the sisterhood of Microsporidia and Fungi. Consistent with this relationship are (*i*) the presence in Microsporidia of similar C-terminal heptapeptide repeats in RPB1, which also occur in crown RPB1 proteins but are absent from some protist RPB1s; (*ii*) phylogenetic analyses of Microsporidial Hsp70 sequences (9, 10); (*iii*) features of Microsporidial biochemistry and physiology (reviewed in refs. 7 and 10); and (*iv*) possession of spliceosomal components and introns (18, 41).

Can we explain why the RPB1 results for Microsporidia are at variance with published EF-1α and EF-2 trees? We have identified a number of potential sources of error (22, 23, 30–32, 35, 45) in such analyses. The RPB1 results do not appear strongly affected by such sources of error, but the EF-1α and EF-2 trees are. Our reanalyses of these data show, at most, that they cannot distinguish between deep divergence and a fungal origin for the Microsporidia, whereas the 11- to 12-aa insertion uniquely shared by animals, fungi, and Microsporidia EF-1α argues strongly for the latter.

Only SSUrRNA analyses appear to provide strong support for the deep divergence of Microsporidia (3, 45, 47). Because these analyses did not explore site-by-site rate variation, they may be subject to some of the problems deriving from constant or fast-evolving sites that we have observed with EF-2 and EF-1α. Kumar and Rzhetsky (48) included a site rate correction for SSUrRNA and concluded that the position of Microsporidia was difficult to resolve. Plots of transitions versus transversions (data not shown) show saturation for transitions for all comparisons involving Microsporidia. Furthermore, several papers (3, 45, 47) have commented on the problems presented by base composition and apparent rate inequalities at the base of the SSUrRNA tree. Very recently, a ML analysis of large subunit rRNA that included correction for site-rate variation indicated a crown placement of the microsporidian *Encephalitizoon*, although not specifically with Fungi (49).

Finally, neither RPB1, EF-2, or EF-1α provide strong support for the Archezoa *Giardia* and *Trichomonas* diverging before other eukaryotes, but (unlike the case for Microsporidia) no other position in the eukaryotic tree appears strongly favored for these taxa. *Giardia* and *Trichomonas* consistently branch deep in SSUrRNA trees (3), and the early branching position of *Giardia* has been highly touted (6). Yet, given the results of our analyses for Microsporidia and recognizing the enormous difficulties in inferring relationships from highly diverged sequences, we conclude that relationships of *Giardia* and *Trichomonas* to other eukaryotes are still unresolved and in need of further investigation.

We thank David Swofford for PAUP*, Andrew Roger and Naomi Fast for the *Nosema* genomic library, Hans-Peter Klenk for the *Giardia* RPB1 sequence, John Stiller for discussion, and James McInerney for help with ML bootstrapping. R.P.H. and B.H. are supported by a Wellcome Trust Career development fellowship to R.P.H. (045702/Z/95/Z). J.M.L. was supported for this work by a Sloan/National Science Foundation Postdoctoral Fellowship in Molecular Evolution, as well as a Medical Research Council (Canada) grant (MT4467) to W.F.D.

1. Canning, E. U. (1993) in *Parasitic Protozoa*, ed. Kreier, J. P. (Academic, New York), Vol. 6, pp. 299–370.
2. Cavalier-Smith, T. (1983) in *Endocytobiology II,* eds. Schwemmler, W. & Schenk, H. E. A. (de Gruyter, Berlin), pp. 1027–1034.
3. Leipe, D. D., Gunderson, J. H., Nerad, T. A. & Sogin, M. L. (1993) *Mol. Biochem. Parasitol.* **59,** 41–48.
4. Hashimoto, T., Nakamura, Y., Kamaishi, T. & Hasegawa, M. (1997) *Arch. Protistenkd.* **148,** 287–295.
5. Kamaishi, T., Hashimoto, T., Nakamura, Y., Masuda, Y., Nakamura, F., Okamoto, K., Shimizu, M. & Hasegawa, M. (1996) *J. Biochem. (Tokyo)* **120,** 1095–1103.
6. Sogin, M. L. (1997) *Curr. Opin. Genet. Dev.* **7,** 792–799.
7. Keeling, P. J. & Doolittle, W. F. (1996) *Mol. Biol. Evol.* **13,** 1297–1305.
8. Edlind, T. D., Li, J., Visversvara, G. S., Vodkin, M. H., McLaughlin, G. L. & Katiyar, S. K. (1996) *Mol. Phylogenet. Evol.* **5,** 359–367.
9. Hirt, R. P., Healy, B., Vossbrinck, C. R., Canning, E. U. & Embley, T. M. (1997) *Curr. Biol.* **7,** 995–998.
10. Germot, A., Philippe, H. & Le Guyader, H. (1997) *Mol. Biochem. Parasitol.* **87,** 159–168.
11. Baldauf, S. L. & Palmer, J. D. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11558–11562.
12. Horner, D. S., Hirt, R. P., Kilvington, S., Lloyd, D. & Embley, T. M. (1996) *Proc. R. Soc. London Ser. B* **263,** 1053–1059.
13. Bui, E. T. N., Bradley, P. J. & Johnson, P. J. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 9651–9656.
14. Germot, A., Philippe, H. & Le Guyader, H. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14614–14617.
15. Roger, A. J., Clark, C. G. & Doolittle, W. F. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14618–14622.
16. Roger, A. J., Svärd, S. G., Tovar, J., Clark, C. G., Smith, M. W., Gillin, F. D. & Sogin, M. L. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 229–234.
17. Cavalier-Smith, T. (1993) *Microbiol. Rev.* **57,** 953–994.
18. Fast, N. M., Roger, A. J., Richardson, C. A. & Doolittle, W. F. (1998) *Nucleic Acids Res.* **26,** 3202–3207.
19. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
20. Stiller, J. W. & Hall, B. D. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 4520–4525.
21. Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J. & Woese, C. R. (1996) *Nucleic Acids Res.* **24,** 82–85.
22. Lake, J. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1455–1459.
23. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) *Mol. Biol. Evol.* **11,** 605–612.
24. Lento, G. M., Hickson, R. E., Chambers, G. K. & Penny, D. (1995) *Mol. Biol. Evol.* **12,** 28–52.
25. Adachi, J. & Hasegawa, M. (1996) MOLPHY v2.3. Computer Science Monographs (Inst. Stat. Math., Tokyo), Vol. 28.
26. Charleston, M. (1998) *Bioinformatics* **14,** 98–99.
27. Strimmer, K. & von Haesler, A. (1996) *Mol. Biol. Evol.* **13,** 964–969.
28. Adachi, J. & Hasegawa, M. (1992) MOLPHY v2.2. Computer Science Monographs (Inst. Stat. Math., Tokyo), Vol. 27.
29. Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29,** 170–179.
30. Lockhart, P. J., Larkum, A. W. D., Steel, M. A., Waddel, P. J. & Penny, D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1930–1934.
31. Fitch, W. (1986) *Phil. Trans. R. Soc. London B.* **312,** 317–324.
32. Palumbi, S. R. (1989) *J. Mol. Evol.* **29,** 180–187.
33. Waddell, P. J. & Steel, M. A. (1997) *Mol. Phylogenet. Evol.* **8,** 398–414.
34. Philippe, H. & Adoutte, A. (1998) in *Evolutionary Relationships of Protozoa*, eds. Coombs, G. H., Vickermann, K., Sleigh, M. A. & Warren, A. (Chapman & Hall, London), pp. 25–56.
35. Yang, Z. (1996) *Trends Ecol. Evol.* **11,** 367–372.
36. Archambault, J. & Friesen, J. D. (1993) *Microbiol. Rev.* **57,** 703–724.
37. Stiller, J. W., Duffield, E. C. S. & Hall, B. D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 11769–11774.
38. Quon, D. V. K., Delgadillo, M. G. & Johnson, P. J. (1996) *J. Mol. Evol.* **43,** 253–262.
39. Stiller, J. W. & Hall, B. D. (1998) *J. Phycol.* **34,** 857–864.
40. Steinmetz, E. J. (1997) *Cell* **89,** 491–494.
41. Biderre, C., Méténier, G. & Vivarés, C. P. (1998) *Mol. Biochem. Parasitol.* **94,** 283–286.
42. Palmer, J. D. & Logsdon, J. M., Jr. (1991) *Curr. Opin. Genet. Dev.* **1,** 470–477.
43. Felsenstein, J. (1978) *Syst. Zool.* **27,** 401–410.
44. Baldauf, S. L. & Doolittle, W. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 12007–12012.
45. Philippe, H. & Adoutte, A. (1995) in *Protistological Actualities*, eds. Brugerolle, G. & Mignot, J.-P. (Blaise Pascal Univ. Press, Clermont-Ferrand, France), pp. 17–32.
46. Cavalier-Smith, T. (1987) *Nature (London)* **326,** 332–333.
47. Galtier, N. & Gouy, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11317–11321.
48. Kumar, S. & Rzhetsky, A. (1996) *J. Mol. Evol.* **42,** 183–193.
49. Peyretaillade, E., Biderre, C., Peyret, P., Duffieux, F., Méténier, G., Gouy, M., Michot, B. & Vivarès, C. P. (1998) *Nucleic Acids Res.* **26,** 3513–3520.