Rooting the Eukaryotic Tree with Mitochondrial and Bacterial Proteins

Romain Derelle*1 and B. Franz Lang2

1Bioinformatics and Genomics Program, Centre for Genomic Regulation and Universitat Pompeu Fabra, Barcelona, Spain
2Robert-Cedergren Centre for Bioinformatics and Genomics, Department of Biochemistry, Université de Montréal, Montreal, Quebec, Canada

*Corresponding author: E-mail: romain.derelle@crg.es.
Associate editor: Hervé Philippe

Abstract

By exploiting the large body of genome data and the considerable progress in phylogenetic methodology, recent phylogenomic studies have provided new insights into the relationships among major eukaryotic groups. However, confident placement of the eukaryotic root remains a major challenge. This is due to the large evolutionary distance separating eukaryotes from their closest relatives, the Archaea, implying a weak phylogenetic signal and strong long-branch attraction artifacts. Here, we propose a new approach to the rooting of the eukaryotic tree by using a subset of genomic information with more recent evolutionary origin—mitochondrial sequences, whose closest relatives are \( \alpha \)-Proteobacteria. For this, we identified and assembled a data set of 42 mitochondrial proteins (mainly encoded by the nuclear genome) and performed Bayesian and maximum likelihood analyses. Taxon sampling includes the recently sequenced \( T. \) trahens, a member of the phylogenetically elusive Apusozoa. This data set confirms the relationships of several eukaryotic supergroups seen before and places the eukaryotic root between the monophyletic “unikonts” and “bikonts.” We further show that \( T. \) trahens branches sister to Opisthokonta with significant statistical support and question the bikont/excavate affiliation of \( \text{Malawimonas} \) species. The mitochondrial data set developed here (to be expanded in the future) constitutes a unique alternative means in resolving deep eukaryotic relationships.

Key words: phylogenomic, eukaryotic origin, \( \alpha \)-Proteobacteria, unikonts, bikonts.

Introduction

The morphological and genetic simplicity of amitochondriate protists (such as diplomonads or microsporidians) were previously thought to resemble an ancestral premitochondrial phase of eukaryote evolution. Indeed, species were placed at the base of the eukaryotic tree in a crown-group phylogeny, known as the archezooan scenario (Sogin 1991). However, it is now evident that such lineages evolved from more complex mitochondrial ancestral forms by reductive evolution (Philippe et al. 2000; Brinkmann and Philippe 2007). The former placement of the amitochondriate protists at the base of the eukaryotic tree (Baldauf et al. 1996; Bapteste et al. 2002; Ciccarelli et al. 2006; Cox et al. 2008) is now attributed to long-branch attraction (LBA), a phylogenetic artifact that tends to erroneously group fast-evolving species, as well as attract them to a distant outgroup (Philippe and Gehrke 2000; Arisue et al. 2005; Brinkmann et al. 2005). The subsequent finding in these organisms of mitochondrial marker proteins, and even relict mitochondrial-derived organelles, confirms that these extant eukaryotic lineages descend from a premitochondrial archezooan (Van de Peer et al. 2000; Embley et al. 2003; Embley and Martin 2006). Thus, despite a wealth of molecular data that have become available in the past decade, the root of the phylogenetic tree of eukaryotes remains undefined (Baldauf 2003; Roger and Simpson 2004; Gribaldo et al. 2010; Koonin 2010).

Due to the lack of a closely related outgroup to Eukarya and the high potential for phylogenetic artifacts when rooting with Archaea, eukaryotic relationships are usually inferred without including an outgroup. In the past decade, this has led to a confident identification of several eukaryotic supergroups including Opisthokonta, Amoebozoa, Viridiplantae, Rhodophyta, Glaucomphyta, Haptophyta, Cryptophyta, SAR (stramenopiles plus Alveolata and Rhizaria), and Discoba (jakobids plus Euglenozoa and Heterolobosea, a subset of Excavata). For this, both phylogenomic data sets were used based on expressed sequence tag (EST) and genome data (with restricted taxon sampling), and taxon-rich data sets based on few highly conserved proteins that were for most part amplified by polymerase chain reaction (e.g., actin, tubulin) (Baldauf 2003; Simpson and Roger 2004; Keeling et al. 2005; Rodríguez-Ezepeleta et al. 2005; Burki et al. 2007, 2009; Rodríguez-Ezepeleta, Brinkmann, Burger, et al. 2007; Hampl et al. 2009; Parfrey et al. 2010). Yet, rooting of the resulting eukaryotic tree is essential for the correct assignment of supergroup relationships, and in the absence of an outgroup, these studies have relied on the validity of the proposed tree rooting.

Previous attempts at rooting the tree of Eukarya have used rare genomic changes (RGCs), that is, events that presumably retain information on phylogenetic relationships over very long periods of time. Yet, the interpretation of RGCs is often problematic since genomic changes are as...
prone to homoplasly as sequence characters (via convergence or reversion), and in the absence of evolutionary models for RGC changes, these are treated as parsimonious characters (Bapteste and Philippe 2002; Rodríguez-Ezpeleta, Brinkmann, Roure, et al. 2007). Other attempts at rooting the eukaryotic tree evaluate the phylogeny of paralogs, which arguably originated from gene duplications prior to the last universal common ancestor (e.g., Brinkmann and Philippe 1999). The weakness of this approach is reliance on only few sequence positions that carry useful information for the rooting, that is, statistical robustness is regularly lacking at this depth of inference.

The currently leading hypothesis places the eukaryotic root between unikonts (Opisthokonta plus Amoebozoa), which are arguably characterized by an ancestral single centriole and cilium, and bionts with two centrioles and kinetids. Support for this subdivision is further based on the presence of a gene fusion (Stechmann and Cavalier-Smith 2002) and the evolutionary history of myosin forms (Richards and Cavalier-Smith 2005). In fact, a bipartition of bikonts–unikonts is compatible with most phylogenomic analyses of eukaryotes that were performed without the use of an outgroup (Baldauf 2003; Simpson and Roger 2004; Keeling et al. 2005; Rodriguez-Ezepeleta et al. 2005; Burki et al. 2007, 2009; Rodriguez-Ezepeleta, Brinkmann, Burger, et al. 2007; Hampel et al. 2009; Parfrey et al. 2010). Alternatively, in a deviation from this unikont/bikont scenario, Cavalier-Smith (2010) recently proposed rooting the eukaryotic tree on Euglenozoa, based on several other molecular features. Finally, a study using rare replacements of highly conserved amino acid residues infers the root between Plantae and other eukaryotes (Rogozin et al. 2009; Koonin 2010). Evidently, there is little if any consensus, and although the various proposals may all seem reasonable, they lack support by rigorous phylogenetic inference based on evolutionary models and statistics. The main challenge with sequence-based rooting of the eukaryotic tree with Archaea is the excessively long evolutionary distance that separates them from eukaryotes (Brinkmann and Philippe 1999; Brinkmann et al. 2005; Gribaldo et al. 2010). Here, we propose to solve this issue by rooting the eukaryotic tree with genes that come from a relatively recent bacterial genome transfer: the acquisition of mitochondria in the last common ancestor of eukaryotes.

It is now widely accepted that mitochondria are derived from an endosymbiotic α-Proteobacterium close to the Rickettsiales (e.g., Gray 1998; Gray et al. 1999; Fitzpatrick et al. 2006), having been acquired in a most likely unique endosymbiotic event occurring in the common ancestor of all extant eukaryotes (Brinkmann and Philippe 2007; Koonin 2010). Most genes of the proto-mitochondron have been lost or transferred to the nuclear genome of the host, that is, vestigial mitochondrial genomes encode only 2–67 of the ~1,000 mitochondrial proteins (Lang et al. 1997, 1999; Andersson et al. 2003; Andreoli et al. 2004). The mitochondrial proteome further evolved by integrating proteins from the host and from various external genome sources, leading to a complex proteome with multiple origins (Karlberg et al. 2000; Kurland and Andersson 2000; Andersson et al. 2003; Szklarczyk and Huynen 2010). Furthermore, many parasites and anaerobic protists underwent mitochondrial reduction, including the loss of both the mitochondrial genome and the capacity to generate ATP through oxidative phosphorylation. Nevertheless, most of these amitochondriate species retain numerous mitochondrial proteins in their nuclear genomes, enough to allow their inclusion in mitochondrial protein–based phylogenies—as we will demonstrate in this paper. The reduced organelles in which these proteins are expressed are known as hydrogenosomes when they produce hydrogen or remnant mitochondria or mitosomes when their functions are cryptic (Embley et al. 2003; Hjort et al. 2010).

The use of mitochondrial proteins as phylogenetic markers is challenging for two reasons. First, from the ~1,000 or more mitochondrial proteins (e.g., Karlberg et al. 2000), only about half are classified paralogous specific, and an even smaller fraction may be traced with confidence to the α-proteobacterial ancestor of mitochondria (for a review, see Gray et al. 2001). According to comparative proteomic analyses, less than 200 of such candidate proteins exist in eukaryotes (Karlberg et al. 2000; Kurland and Andersson 2000; Andersson et al. 2003; Szklarczyk and Huynen 2010). For the subset that may be used in phylogenomic analyses, it is reasonable to demand that 1) they are known components of mitochondrial proteomes across eukaryotes, 2) are well conserved and of clear α-proteobacterial origin, and 3) that they are encoded, at least partially, by mitochondrial DNA (mtDNA) across eukaryotes (Lang et al. 1999). A second difficulty is that mitochondrial proteins usually evolve faster than their cytosolic counterparts, in particular mitoribosomal proteins. Whereas, cytosolic ribosomal proteins are well conserved, composing up to a half of classical phylogenomic data sets (Bapteste et al. 2002; Rodriguez-Ezepeleta et al. 2005; Burki et al. 2007; Hampel et al. 2009), it is sometimes even difficult to identify orthologs of mitoribosomal proteins due to their high level of divergence (Smits et al. 2007; Desmond et al. 2010). Despite these limitations, we were able to define a new multigene data set composed of 42 proteins that are of clear α-proteobacterial origin, with which we then conducted Bayesian analyses based on the CAT model (arguably, the most realistic evolutionary model that is least sensitive to phylogenetic artifacts; Lartillot and Philippe 2004; Lartillot and Philippe 2007; Lartillot and Philippe 2009; Lartillot and Philippe 2010) as well as maximum likelihood (ML) analyses. The robustness of the obtained tree was assessed by measuring the statistical support of the major branches in question (jackknife in Bayesian and bootstrap in ML analyses). In addition, for ML analyses, we evaluated likelihood ratio tests for alternative rooting of the eukaryotic tree.

Materials and Methods
Identification of Phylogenetic Markers
The predicted proteomes of Dictyostelium discoideum and Chlorella sp. were used as initial reference data sets for blasting locally a collection of 14 predicted bacterial
proteomes (four α-Proteobacteria, two β-Proteobacteria, two δ-Proteobacteria, two Firmicutes, two Actinobacteria, and two Bacteroidetes) and 10 Archaea (using blastp with a threshold value of 1 × 10⁻⁸). For proteins of potential α-proteobacterial origin, alignments were built with Muscle (Edgar 2004), together with the bacterial and archaeal species cited above and by further including Homo sapiens, Arabidopsis thaliana, Thalassiosira pseudonana, and Phytophthora infestans proteins (obtained by BLAST with the same threshold value). Alignments were automatically trimmed using trimAl (“automated” option; Capella-Gutiérrez et al. 2009), and ML analyses from these alignments were then conducted with RAxML version 7.2 (using the PROTGAMMALG model; Stamatakis 2006) to identify and select eukaryotic proteins of clear α-proteobacterial origin. A second round of phylogenetic analyses was then carried out to identify and remove proteins with complex evolutionary histories due to gene duplications or losses. Only those alignments were retained for which orthologous sequence relationships were assessed with confidence (i.e., one-to-one orthologous relationships, independently of any taxonomic assumption). Finally, a third round of phylogenetic analyses (with the same methodology as described above) was performed on selected phylogenetic markers, after adding sequences from a wide range of eukaryotic species. Again, we only retained proteins (a total of 32) displaying a simple history and where we were confident of orthologous relationships. In addition, we added to the selected data set ten well-conserved proteins that are encoded by mitochondrial genomes, using the GOBASE database (O’Brien et al. 2009) and in-house collections. The gene ontology of orthologous H. sapiens and A. thaliana proteins, from the Entrez Gene webpage in NCBI, revealed a cellular location of all 42 proteins within the mitochondrion (GO:0005739). Finally, this data set of 42 proteins was BLASTed on the predicted proteome of 15 reference mitochondrial genomes (including all main eukaryotic supergroups) to determine the genomic localization of these proteins among eukaryotes and on the 143 alignments used in Hampl et al. 2009 to estimate the intersection between our data set and published EST-based data sets.

Assembly of Sequences into the Phylogenetic Matrix

The phylogenetic matrix used to infer the eukaryotic root was restricted to eukaryotes, 12 α-Proteobacteria, and Magnetococcus sp. as their closest outgroup (according to Wu et al. 2009). To minimize the level of missing data, we created chimerical sequences in some cases by merging closely related species. The chimerical terminal taxa used in this study correspond to Blastocystis (B. hominis and Blastocystis sp.), Paramecium (P. tetraurelia and P. aurelia), Phytophthora (P. infestans, P. ramorum, and P. sojae), Rhizaria (B. globowella natans and Paracercomonas marina), and Malawimonas (M. jakobiformis and M. californiana; the most represented species are underlined). In cases where several sequences were present in the alignment, the slowest evolving one was selected (according to the branch lengths in RAxML trees). Finally, we removed from further analyses any eukaryotic sequence branching outside the “combined” group (Magnetococcus sp. + α-Proteobacteria + eukaryotes) in the RAxML tree, as these sequences were likely transferred laterally. Trimming of alignments was performed with Gblocks (Castresana 2000) underdefault/ stringent parameters, except for the maximum proportion of gaps per position (increased from 0% to 20%) and the maximum number of contiguous nonconserved positions (increased from 8 to 10). Finally, trimmed alignments were concatenated into a supermatrix using a custom-made script. This supermatrix has been deposited in the TreeBASE database (www.treebase.org).

Phylogenetic Analyses and Likelihood Ratio Tests

Bayesian inferences were performed with the CAT + Γ4 mixture model using the -dc option, by which constants sites are removed, implemented in the program PhyloBayes version 3.2 (Lartillot et al. 2009). We performed statistical comparisons of the CAT model with the empirical LG model by using cross-validation tests as described in Lartillot and Philippe (2008), based on the topology of figure 1. Ten replicates were performed: 9/10 for the learning set and 1/10 for the test set. Markov chain Monte Carlo (MCMC) chains were run for 3,000 cycles with a burn-in of 1,500 cycles for the CAT model and 1,500 cycles with a burn-in of 100 cycles for the LG model. The CAT model was found to have a much better statistical fit than did LG (a likelihood score of 1,648.6 ± 83.09 in favor of CAT). For the plain posterior estimation, two independent runs were performed with a total length of 24,000 cycles. Convergence between the two chains was ascertained by calculating the difference in frequency for all their bipartitions using “bpcomp” (maxdiff < 0.1). The first 10,000 points were discarded as burn-in, and the posterior consensus was computed by selecting one tree every ten over both chains. Because of computational efficiency reasons, jackknife was chosen instead of bootstrapping to estimate branch support. Hundred pseudoreplicates were generated using SEQBOOT (a sampling of 60% of the data set was chosen; Felsenstein 2001), and each of them was analyzed using PhyloBayes. Trees from all replicates were collected after the initial burn-in period, pooled, and a 50% majority-rule consensus tree was inferred with CONSENSE. Due to computational time constraints, we performed only 6,000 cycles for each replicate associated with a conservative burn-in of 3,000 cycles (verification of a few replicates indicated that the burn-in is generally <1,000). The recoding of amino acids into the six Dayhoff functional categories was performed using the “–recode” command implemented in PhyloBayes. We then performed Bayesian analyses using these recoded data sets under the CAT model: Two independent MCMC chains were run with a total
length of 12,000 cycles. Convergence of the two chains was checked with bpcomp, and the first 6,000 points were discarded as burn-in to compute the posterior consensus (by selecting one tree every ten over both chains). We performed the posterior predictive tests of compositional homogeneity using the “ppred –comp” command of PhyloBayes. For these tests, we ran for each data set one MCMC chain under the CAT model with a total length of 24,000 cycles (the first 10,000 points were discarded as burn-in).
ML analyses were performed using RAxML version 7.2, and searches for the best tree were conducted starting from ten random trees. Analyses were conducted with PROTGAMMALGF, which is the best-fitting model for the concatenated matrix according to the Akaike information criterion (computed with ProtTest 3.0; Darriba et al. 2011). We also defined the best-fitting model for each single-gene alignments in the same way to perform ML analyses under the separate model (for the list of models, see supplementary data S2, Supplementary Material online). ML bootstrap analyses shown in figures 1, 2B, and 3B were performed with the standard algorithm implemented in RAxML, but due to computational constraints, all other ML bootstrap analyses with the rapid BS algorithm. Site rates were calculated using CODEML (Yang 2007) under the LG + GAMMA model. As the phylogenetic topology used to estimate evolutionary rates might have a major impact on the results (e.g., Rodríguez-Ezepeleta, Brinkmann, Roure, et al. 2007), we successively calculated site rates on four different topologies including 8–12 species (topologies provided in supplementary data S2, Supplementary Material online). The underlying phylogenetic hypotheses are as follows: 1) no assumption on the eukaryotic root, that is, bacteria and eukaryotes were analyzed separately, 2) no more than two eukaryotic supergroups are present per topology (i.e., no assumption on eukaryotic supergroups relationships), and 3) species/lineages with ambiguous phylogenetic positions were discarded from these analyses (e.g., Thecamonas trahens, Malawimonas, Rhizaria, Haptophyta, Hartmannella vermiformis). Finally, using in-house scripts, a mean site rate was calculated for each position by combining the site rates calculated over these four analyses, and fast positions were removed from the data set in steps of 2% until 40%. Node supports for the unikont–bikont hypothesis and for the alternative positions of Malawimonas were calculated using the ETE package (Huerta-Cepas et al. 2010).

The likelihood-based approximately unbiased (AU) test of alternative eukaryotic root was performed with CONSEL (Shimodaira and Hasegawa 2001). For each alternative eukaryotic root topology, the best tree was inferred using the PROTGAMMALGF model, and an input tree file in which only the node corresponding to the eukaryotic root was constrained. Then ML branch lengths of alternative topologies were inferred using the “-f g” option of RAxML and the same model. Site-wise log-likelihood values were compared with CODEML (Yang 2007), and the P values of the different likelihood-based tests were calculated with CONSEL.

Results and Discussion

A New Large Multigene Data Set Based on Mitochondrial Proteins

Only ten mtDNA-encoded proteins that satisfy our selection criteria for phylogenetic analyses are encoded in the mitochondrial genome of the 15 eukaryotic species considered in our analysis (supplementary data S1, Supplementary Material online), which is most likely insufficient for rooting the eukaryotic tree. Therefore, using BLAST-based sequence similarity searches combined with phylogenetic analyses, we have developed a protocol to select additional bona fide mitochondrial proteins that may be either exclusively or partially nucleus encoded, depending on the species (table 1). The chosen proteins are present in a wide range of eukaryotes, are clearly orthologs, of α-proteobacterial origin, and as it turns out, all have a predicted mitochondrial localization. They function in various mitochondrial processes, namely the respiratory electron transport, the tricarboxylic acid cycle, ATP synthesis, protein folding, and translation (for more details, see table 1). Some of them have previously been shown to be of α-proteobacterial, that is, most likely mitochondrial origin (Karlb erg et al. 2000; Kurland and Andersson 2000; Szklarczyk and Huynen 2010). Among the 42 proteins, only 3 (mitochondrial Hsp 60 kDa, Hsp 70 kDa, and the alpha subunit of succinyl-CoA ligase) are in common with a popular phylogenomic data set based on 143 highly expressed proteins (the data set used in Hampl et al. 2009) that was originally introduced by Bapteste et al. 2002. It is composed of highly expressed proteins, essentially ribosomal proteins, proteasome subunits, and vacuolar ATPase subunits. Therefore, although most proteins of our mitochondrial data set are nucleus encoded, it has a negligible intersection with classical EST-based data sets, providing an independent view of eukaryotic evolution.

Trimmed alignments were concatenated into a multigene data set to yield a total of 11,500 amino acid positions, with a mean completeness of 80% amino acids per taxon (minimum of 31% for Sterkiella histriomuscorum and maximum of 100% for D. discoideum and T. pseudonana; supplementary data S1, Supplementary Material online). Interestingly, even the strictly anaerobic parasite B. hominis that has mitochondria-like organelles devoid of ATPase and cytochrome subunits (Nasirudeen and Tan 2004; Stechmann et al. 2008) has retained 21 of these 42 genes in its nuclear genome. Evidently, at least some species without mtDNA and with reduced mitochondria may be included in our data set. Species for which only EST data are available have a reduced coverage of only 15–50%, depending on the number of ESTs and the gene expression bias as well as on the availability of a separate mitochondrial genome sequence. Species with a high level of missing data have been removed from further analyses, and consequently, some key eukaryotic lineages are absent from this study (e.g., glaucophytes, centrohelids, cryptophytes). To allow for inclusion of Malawimonas, we reduced the fraction of missing data by combining M. californiana and M. jakobiiformis (18% and 44% of data coverage, respectively) into a single terminal unit “Malawimonas” (55% of data). Given that there is then only one representative of this group and the known difficulty for its positioning in the eukaryotic tree (Simpson et al. 2006; Rodríguez-Ezepeleta, Brinkmann, Burger, et al. 2007; Hampl et al. 2009), we
decided to conduct phylogenetic analyses both with and without *Malawimonas*.

**Phylogenetic Analyses Support the Unikont–Bikont Dichotomy**
Bayesian phylogenetic analysis was conducted under the CAT model as it has a better fit to the data than the empirical LG model (see Material and Methods). The Bayesian consensus LG model over both MCMC chains (fig. 1) shows a paraphyletic α-Proteobacteria, with Rickettsiales as a sister group to eukaryotes. Within eukaryotes, common supergroups and relationships are recovered (Burki et al. 2007, 2009; Hackett et al. 2007; Rodriguez-Ezpeleta, Brinkmann, Burger, et al. 2007; Hampl et al. 2009; Parfrey et al. 2010),

### Table 1. List of Phylogenetic Markers.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Gene Product</th>
<th>Human Protein</th>
<th>Refseq Accession Number</th>
<th>Biological Process</th>
<th>% of Alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ndhA</em></td>
<td>NADH dehydrogenase subunit 1</td>
<td>NP_004159</td>
<td>Respiratory electron transport chain</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td><em>ndhB</em></td>
<td>NADH dehydrogenase subunit 2</td>
<td>NP_002991</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadC</em></td>
<td>NADH dehydrogenase subunit 3</td>
<td>NP_00324033</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadD</em></td>
<td>NADH dehydrogenase subunit 4</td>
<td>NP_00324035</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadE</em></td>
<td>NADH dehydrogenase subunit 4L</td>
<td>NP_00324034</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadF</em></td>
<td>NADH dehydrogenase subunit 5</td>
<td>NP_00324036</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadG</em></td>
<td>NADH dehydrogenase subunit 6</td>
<td>NP_004541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadH</em></td>
<td>NADH dehydrogenase subunit 7</td>
<td>NP_002487</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadI</em></td>
<td>NADH dehydrogenase subunit 8</td>
<td>NP_004542</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadJ</em></td>
<td>NADH dehydrogenase subunit 9</td>
<td>NP_004543</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadK</em></td>
<td>NADH dehydrogenase subunit 10</td>
<td>NP_077718</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nadL</em></td>
<td>NADH dehydrogenase subunit 11</td>
<td>NP_004997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>ndufA</em></td>
<td>NADH dehydrogenase flavoprotein 1</td>
<td>NP_01159574</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>ndufB</em></td>
<td>NADH dehydrogenase flavoprotein 2</td>
<td>NP_066552</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>coxA</em></td>
<td>Apocytochrome b</td>
<td>NP_00324038</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>coxB</em></td>
<td>Apocytochrome c</td>
<td>NP_001907</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>coxC</em></td>
<td>Apocytochrome c</td>
<td>NP_061820</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>coxd</em></td>
<td>Apocytochrome c oxidase subunit 1</td>
<td>NP_00324037</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>coye</em></td>
<td>Apocytochrome c oxidase subunit 2</td>
<td>NP_00324029</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>cozf</em></td>
<td>Apocytochrome c oxidase subunit 3</td>
<td>NP_00324032</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>etfB</em></td>
<td>Electron transfer flavoprotein, beta subunit</td>
<td>NP_001976</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>atpA</em></td>
<td>ATP synthase F1 subunit alpha</td>
<td>NP_001001937</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>atpB</em></td>
<td>ATP synthase F1 subunit beta</td>
<td>NP_001001938</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>atpC</em></td>
<td>ATP synthase F1 subunit gamma</td>
<td>NP_001001939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>atpD</em></td>
<td>ATP synthase F0 subunit 6</td>
<td>NP_00324031</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>dldh</em></td>
<td>Dihydrolipoamide dehydrogenase</td>
<td>NP_000099</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>dlst</em></td>
<td>Dihydrolipoamide S-succinyltransferase</td>
<td>NP_001924</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>odpa</em></td>
<td>Pyruvate dehydrogenase</td>
<td>NP_005381</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>odpb</em></td>
<td>Pyruvate dehydrogenase</td>
<td>NP_005382</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>SUCLG1</em></td>
<td>SuccinylCoA ligase, alpha subunit</td>
<td>NP_003840</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>mHsp70</em></td>
<td>Mitochondrial Hsp 70 kDa</td>
<td>NP_004125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>mHsp70</em></td>
<td>Mitochondrial Hsp 70 kDa</td>
<td>NP_004125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>grpE</em></td>
<td>Mitochondrial cochaperone grpE</td>
<td>NP_079472</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Hsp78</em></td>
<td>Mitochondrial Hsp 78</td>
<td>NP_063444**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>HslVU</em></td>
<td>Heat shock protein HslVU</td>
<td>NP_063662**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>mlr3</em></td>
<td>Mitochondrial ribosomal protein L3</td>
<td>NP_009139</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>mtrf1</em></td>
<td>Peptide chain release factor</td>
<td>NP_061914</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>mtf1</em></td>
<td>Translation initiation factor</td>
<td>NP_061914</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>tufA</em></td>
<td>Translation elongation factor Tu</td>
<td>NP_003312</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>coxA1</em></td>
<td>Protein involved in heme biosynthesis</td>
<td>NP_003466</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>nfs1</em></td>
<td>Cysteine desulfurase, mitochondrial</td>
<td>NP_006923</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>sco1</em></td>
<td>Synthesis of cytochrome c oxidase factor</td>
<td>NP_004580</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.**—The first three columns indicate the name of human proteins, their NCBI RefSeq accession number, and their genomic localization (nuclear vs. mitochondrial genome). In two cases, human orthologs are absent and NCBI Refseq accession numbers correspond to orthologs found in the Dictyostelium discoideum proteome (indicated by **). A single asterisk (first column) indicates the presence of the protein in GOBASE. Markers are grouped according to their main gene ontology biological process (fourth column), and for each of the biological processes, its proportion in the concatenated multigene data set is indicated in the last column.
except for the position of Haptophyta as sister group to the SAR group and Rhizaria as sister group to Alveolata within the SAR group. The eukaryotic root lies between unikonts (including Opisthokonta, Amedoza, and Amoebozoa) and bikonts (all other eukaryotes). Jackknife analysis supports the monophyly of unikonts and bikonts at 85% and 81%, respectively. The best ML tree obtained under the PROTGAMMALGF model shows a weakly supported polyphyley of excavates, with fast-evolving excavates (Discicristata group; Naegleria gruberi, Euglena gracilis, Trypanosoma brucei, and Leishmania major) as a sister group of Amoebozoa and therefore no unikont–bikont root as in Bayesian analyses (supplementary data S2, Supplementary Material online). Other differences to the Bayesian topology are in relationships among some α-Proteobacteria and the grouping of Haptophyta and Rhodophyta. In contrast, ML analysis conducted under the separate model, which is expected to have a better fit to the data than the uniform LGF model, yields the same eukaryotic relationships as in the Bayesian analysis (supplementary data S2, Supplementary Material online), although recovering a unikont–bikont root with very low support values (51% and 47% for unikonts and bikonts, respectively, fig. 1). This absence of strong support of unikonts and bikonts is essentially due to the instability of the Excavata: In a large proportion of ML bootstrap samples (as well as in a few Bayesian jackknife replicates), the Discicristata group is a sister group of Amoeboboa or unikonts. Finally, in both Bayesian and ML analyses, the apusozoonan T. trahens branches as sister group to Opisthokonta with strong statistical support (fig. 1). This topology is in agreement with published studies that are, however, not well supported by statistics (Kim et al. 2006; Cavalier-Smith and Chao 2010).

Conceptually, it is problematic to assess the robustness of eukaryotic rooting with (bootstrap or jackknife) branch support values that indicate support of the monophyly of a single group and do not provide direct information about support of relationships between several lineages. We have therefore estimated the robustness of the unikont–bikont hypothesis as the proportion of trees in which both unikonts and bikonts are recovered (see Material and Method), a metric referred to in the following as “node support.” Node support for the unikont–bikont root is 80% and 47%, for the Bayesian jackknife and ML bootstrap replicates performed under the separate model, respectively (fig. 1). To explore the potential impact of fast-evolving (potentially saturated) sequence positions on the outcome of analyses and to overcome possible LBA artifacts (Brinkmann et al. 2005), we have applied a progressive site removal procedure. Evolutionary rates were determined for each position, with particular attention to avoiding assumptions on eukaryotic relationships (see Materials and Methods). Fast-evolving positions were deleted in 2% increments up to 40%, and a search for the best ML tree and 100 ML bootstraps under the uniform LGF model were performed on each data point. Note that along these 20 analyses (i.e., removing from 2% to 40% of sites in steps of two), the eukaryotic root is consistently between unikonts and bikonts, with support for the unikont–bikont root increasing from 52% to 95% until 22% of fast-evolving positions are removed. Support values fall toward 47% from this point on, likely because increasingly informative sequence positions are dropped (fig. 2A).

The initial increase of bootstrap values is due to stabilization of excavates (support for the monophyly of excavates reaches 100% at 18% site removal), whereas the subsequent decrease is due to instability of unikonts, with T. trahens tending to branch as sister group to all other eukaryotes. For further testing of the most robust data point (22% of sites removed, called the “filtered data set”), two additional analyses were performed: An ML analysis using the separate model and a Bayesian analysis under the CAT model. In contrast to the full data set, eukaryotic relationships are identical with the three types of analysis (a minor incongruence is within α-Proteobacteria; ML trees not shown). Most significantly, the unikont–bikont root is now supported by 90%, 95%, and 91% Bayesian jackknife and ML bootstrap (performed under the uniform and separate models) values, respectively (fig. 2B).

A final remaining question is whether the unikont–bikont root results from strong support by only few proteins or a summation of relatively weak phylogenetic signal across the data set. For this, we performed an ML search on all trimmed protein alignments using the best-fitting model according to ProtTest 3.0. In none of the 42 topologies was the eukaryotic root between unikonts and bikonts (trees are provided in supplementary data S3, Supplementary Material online), indicating that inference of the unikont–bikont root relies on a large data set that combines a weak phylogenetic signal across many proteins. In addition, we used for each gene the AU test (Shimodaira 2002) to compare the likelihoods of the best tree and a tree in which the unikont–bikont root was constrained (method and results are provided in supplementary data S2, Supplementary Material online). Only four genes reject the unikont–bikont root at the 5% confidence level (cox1, nad1, nad5, and nsf1), suggesting that the phylogenetic signal in support of the unikont–bikont root does not contradict and mask a conflicting signal.

Statistical Comparison of Alternative Eukaryotic Roots

To further test the stability of the eukaryotic root, we evaluated a combination of 26 alternative root topologies with the AU test. Tested topologies include variants observed in Bayesian jackknife and ML bootstrap replicates, and those proposed in the scientific literature. Analyses were conducted under the uniform LGF model with both the full and filtered data sets. With the full data set, the AU test qualifies the association of Discicristata with Amoebozoa (topology obtained in ML analysis under the LGF model) as most likely, and the unikont–bikont root (topology obtained in Bayesian analysis and ML analysis under the
Potential Phylogenetic Artifacts

Large variations in GC content between genomes can result in compositional amino acid biases (e.g., Singer and Hickey 2000) and if not modeled may cause phylogenetic artifacts. In our study, a compositional bias might be present due to Rickettsiales and mitochondrial genome sequences that have a low G + C content compared with other α-Proteobacteria and to the most nuclear genomes. We have therefore built and analyzed (ML search and bootstrap analysis) three alternative sub data sets: 1) without Rickettsiales, 2) without the ten proteins that are frequently encoded in mitochondrial genomes (i.e., genes present in all or all but one mtDNA of the taxonomic sampling used in supplementary data S1, Supplementary Material online), and 3) without Rickettsiales species and without these ten mtDNA-encoded proteins. The results shown in supplementary data S2 (Supplementary Material online) indicate that removal of Rickettsiales has no impact on phylogenetic analyses: The ML trees obtained under the LGF and separate models are identical to the ones obtained with the full data set under the corresponding models, and branch supports are very similar. Interestingly, the data set with the ten mtDNA-encoded proteins was removed, and the one in which Rickettsiales were removed in addition, recover the monophyly of unikonts and bikonts when analyzed under the LGF model. These results suggest that mtDNA-encoded markers likely contribute to a nonphylogenetic signal. An ML analysis of a matrix composed of these ten proteins shows lack of support for deep nodes and polyphyly of SAR, Excavata, and Amoebozoa groups (supplementary data S2, Supplementary Material online). This lack of phylogenetic signal may be caused by the strong A + T bias of mitochondrial genomes or by multiple independent changes in the mitochondrial genetic code along the evolution of eukaryotes (such process is known to occur at a higher rate in the mitochondrial genome than in the nuclear genome; e.g., Knight et al. 2001; Sengupta et al. 2007). From the combined experiments, we conclude that the unikont–bikont root hypothesis is not a result of a compositional artifact due to the presence in our data set of Rickettsiales species and proteins encoded by the mitochondrial genomes. Finally, in order to detect other possible compositional amino acid biases, we recoded the full and filtered data sets into the six Dayhoff functional categories and performed Bayesian analyses under the CAT model. In both cases, MCMC chains converge and both consensus trees infer the monophyly of unikonts and bikonts with posterior probabilities of 0.99 or 1 (supplementary data S2, Supplementary Material online).

Another potential source of phylogenetic artifact is due to the use of stationary evolutionary models (see, for instance, Cox et al. 2008), that is, the assumption of a substitution process constant in time. To test whether or not our data reject the hypothesis of stationarity, we performed posterior predictive tests of composition homogeneity as implemented in PhyloBayes (described in Blanquart and Lartillot 2008). The results of these tests (see supplementary data S2, Supplementary Material online) indicate that the stationary CAT model is only just rejected at a 5% confidence level when the full data set is analyzed, and the CAT model is accepted when the filtered data set is analyzed (Z scores of 2.3 and 1.4 for the full and filtered data sets, respectively). However, the power of the global posterior predictive test is likely limited, since for instance, 40 species of S3 (including T. trahens) are significant rejected for the filtered data set. Future studies are therefore needed to completely exclude the possibility that compositional bias affects our topology.

Phylogenetic Position of Malawimonas

With the full data set, Malawimonas is a sister group to Amoebozoa plus Discicristata in ML analysis under the LGF model, and it is a sister group to Amoebozoa in ML analysis under the separate model, although with negligible support in both analyses (supplementary data S2, Supplementary Material online). Using this data set, Malawimonas branches as a sister group of unikonts in Bayesian analyses under the CAT model (posterior probability equal to 0.99; data not shown), although the convergence of MCMC chains is less good than for other analyses (maxdiff = 0.216). Yet, when progressively removing fast-evolving positions, tree resolution improves significantly. In Bayesian analyses, Malawimonas branches initially with unikonts (0–18% positions removed) and then with T. trahens (from 22% to 40% of positions removed) (fig. 3A). When 24% of fast-evolving sites are removed, eukaryotic relationships are identical to the ML tree shown in figure 3B. In addition, MCMC chains converge, and the consensus tree shows
**FIG. 3.** Phylogenetic analyses including *Malawimonas*. (A) Evolution of support values for alternative positions of *Malawimonas* along the removal of fast-evolving sites from the "full" data set: Bayesian posterior probabilities (graphic on the left) and ML bootstrap values (using the uniform LGF model; graphic on the right). The gray vertical bar indicates on each graphic the data set analyzed in figure 3B. (B) Bayesian tree obtained from the data set with 24% of positions removed. The outgroup (z-Proteobacteria and *Magnetococcus* sp.) is not shown for design reasons (gain of space). Branch supports correspond from left to right to Bayesian posterior probabilities, Bayesian jackknife values, ML bootstrap values obtained with the LGF model, and ML bootstrap values obtained with the separate model. Branches with 100% support in all analyses are marked with a bullet. A dash indicates the absence of the corresponding node in the ML tree.
Malawimonas branching with *T. trahens*—although with only 61% jackknife support. However, *Malawimonas* groups with unikonts in 97% of jackknife replicate. In ML analyses under the LGF model, *Malawimonas* is either a sister group to Amoebozoa or to other unikonts (0–18% of positions removed) and at higher filtering values branches with *T. trahens* (fig. 3A). ML searches and bootstrap analyses under the LGF and separate models of the 24% data point confirm the Bayesian analysis: *Malawimonas* branches with *T. trahens* at low support, but the grouping of unikonts plus *Malawimonas* is relatively high (fig. 3B). Note that the grouping of Excavata plus *Malawimonas* identified in other investigations (Simpson et al. 2006; Rodríguez-Ezpeleta, Brinkmann, Burger, et al. 2007; Hampl et al. 2009) was never obtained in our analyses.

Conclusions

Here, we present a new large multigene data set of mitochondrial proteins with 11,500 amino acid positions, which includes a wide range of eukaryotic lineages. Based on this data set, Bayesian and ML analyses recover all eukaryotic supergroups and, after the removal of fast-evolving sites, converge to identical results: 1) the root of the eukaryotic supergroups and, after the removal of fast-evolving sites, the phylogenetic position of the eukaryotic tree is predicted to lie between unikonts and bikonts, 2) convergent to identical results: 1) the root of the eukaryotic supergroups and, after the removal of fast-evolving sites, the phylogenetic position of the eukaryotic tree is predicted to lie between unikonts and bikonts, 2) the phylogenetic position of *T. trahens* is confirmed as a sister group to Opisthokonta, although based on only a single representative (to be confirmed by adding more species), and 3) the enigmatic *Malawimonas* group branches as a sister group or even within unikonts, reminiscent of difficulties in placing them in a previous analysis (Rodríguez-Ezpeleta, Brinkmann, Burger, et al. 2007; Hampl et al. 2009). In the present study, *Malawimonas* does not associate with bikonts under any condition.

The unikont–bikont dichotomy is widely used as the default root hypothesis in almost all evolutionary studies of eukaryotes. However, to our knowledge, this is the first time that this topology has been assessed by phylogenetic analysis. In addition, this data set is an excellent complement to the classical EST-based data sets (Bapteste et al. 2002) to explore deep eukaryotic relationships. Note, however, that some extremely derived amitochondriate species, such as Metamonada or Microsporidia, cannot be studied through this data set, as the number of remaining nucleus-encoded mitochondrial proteins is too limited. For this, other approaches have to be found.

With the current mitochondrial data set, the branching position of the apusozoan *T. trahens* as sister group to Opisthokonta renders the proposed synapomorphy “unikont” inappropriate, as Apusozoa have two basal bodies per kinetid (Roger and Simpson 2009; Cavalier-Smith and Chao 2010). Note, however, that this topology is based on the inclusion of a single apusozoan species and that poor species sampling will increase the potential for phylogenetic error. The putative sister group relationship of *T. trahens* with *Malawimonas*, which also has two basal bodies per kinetid (O’Kelly and Nerad 1999), is in support of this new scenario that calls for abandoning “Unikonta” and “Bikonta” as taxonomic terms. Yet to confirm this topology beyond reasonable doubt, a range of *Malawimonas* and apusozoan genome projects are highly desirable. Finally, a significant fraction of the known eukaryotic diversity needs to be added to phylogenomic studies, including for instance the early emerging *Diphylloatea* (Brugerolle et al. 2002) and *Mantamonas* species (Glücksman et al. 2011).

Supplementary Material

Supplementary data S1–S3 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Fyodor Kondrashov for critical reading of the manuscript. Authors are also grateful to anonymous referees whose comments helped to improve significantly the quality of this study. This research was supported by the Plan Nacional grant BFI2009-09271 from the Spanish Ministry of Science and Innovation to R.D. and by the National Science and Engineering Research Council of Canada to B.F.L. (grant number 194560).

References


mitochondrial cristae that is phylogenetically distant from other flagellate groups. *Protist* 153:59–70.


