First publ in: Journal of Molecular Evolution 51 (2000), pp. 565–576

An Updated and Comprehensive rRNA Phylogeny of (Crown) Eukaryotes Based on Rate-Calibrated Evolutionary Distances

Yves Van de Peer,[1] Sandra L. Baldauf,[2,*] W. Ford Doolittle,[2] Axel Meyer[1]

[1] Department of Biology, University of Konstanz, D-78457 Konstanz, Germany
[2] Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4H7

Received: 23 May 2000 / Accepted: 3 August 2000

Abstract. Recent experience with molecular phylogeny has shown that all molecular markers have strengths and weaknesses. Nonetheless, despite several notable discrepancies with phylogenies obtained from protein data, the merits of the small subunit ribosomal RNA (SSU rRNA) as a molecular phylogenetic marker remain indisputable. Over the last 10 to 15 years a massive SSU rRNA database has been gathered, including more then 3000 complete sequences from eukaryotes. This creates a huge computational challenge, which is exacerbated by phenomena such as extensive rate variation among sites in the molecule. A few years ago, a fast phylogenetic method was developed that takes into account among-site rate variation in the estimation of evolutionary distances. This "substitution rate calibration" (SRC) method not only corrects for a major source of artifacts in phylogeny reconstruction but, because it is based on a distance approach, allows comprehensive trees including thousands of sequences to be constructed in a reasonable amount of time. In this study, a nucleotide variability map and a phylogenetic tree were constructed, using the SRC method, based on all available (January 2000) complete SSU rRNA sequences (2551) for species belonging to the so-called eukaryotic crown. The resulting phylogeny constitutes the most complete description of overall eukaryote diversity and relationships to date. Furthermore, branch lengths estimated with the SRC method better reflect the huge differences in evolutionary rates among and within eukaryotic lineages. The ribosomal RNA tree is compared with a recent protein phylogeny obtained from concatenated actin, α-tubulin, β-tubulin, and elongation factor 1-α amino acid sequences. A consensus phylogeny of the eukaryotic crown based on currently available molecular data is discussed, as well as specific problems encountered in analyzing sequences when large differences in substitution rate are present, either between different sequences (rate variation among lineages) or between different positions within the same sequence (among-site rate variation).

Key words: SSU rRNA database — Among-site rate variation — Eukaryotic diversity — Crown eukaryotes — Distance trees — Rate variation among lineages

Introduction

Due to its ubiquity, size, and generally slow rate of evolution, the small-subunit ribosomal RNA (SSU rRNA) has proven to be an invaluable tool in the field of molecular phylogeny (Olsen and Woese 1993, Doolittle and Brown 1994), and it is the most sequenced of all genes. Phylogenetic analyses based on this gene have contributed numerous novel and often-unexpected insights into the evolutionary history of life on this planet. For example, SSU rRNA studies completely altered our understanding of the overall organization of life by revealing the existence of an entire new organismal domain, the

Correspondence to: Yves Van de Peer; *e-mail:* (Yves.VandePeer@uni-konstanz.de)
* *Present address:* Department of Biology, University of York, Heslington, York YO10 5DD, UK

Archaea (Fox et al. 1980; Pace et al. 1986). Over 13,000 SSU rRNA sequences are now available from thousands of different species (Van de Peer et al. 2000a). Of these, about 3000 complete or nearly complete sequences exist for eukaryotes, of which about 2550 belong to the dense cluster of phyla known as the eukaryote crown (Knoll 1992).

The eukaryotic crown is formed by a huge grouping of what appear to be nearly simultaneously diverging taxa that constitute the vast majority of eukaryotic life, probably more than 30 million different species (Mayr 1998). These include the plants, fungi, and animals, as well as most algal and many protistan lineages such as rhodophytes, haptophytes, heterokont algae, ciliates, dinoflagellates, apicomplexans, oomycetes, hyphochytriomycetes, choanoflagellates, and many others. On the basis of ribosomal RNA data (but see further), this crown radiation is set off from a series of more basal lineages branching off successively from the base of the tree. These include the slime molds, especially the Myxogastridae (plasmodial slime molds), the Euglenozoa (*Euglena* and relatives plus kinetoplastids), and amitochondriate eukaryotes such as parabasalids, microsporidians, and diplomonads (Sogin 1987; Kumar and Rzhetsky 1996).

To study such ancient relationships dating back billions of years (Doolittle et al. 1996; Wang et al. 1999; Knoll 1999) on the basis of overall evolutionary distances, a reliable estimation of these distances among taxa, by applying an appropriate correction for multiple substitutions at the same site, is crucial. It is well known that unrealistic substitution models can cause serious artifacts in inferred tree topologies (Olsen 1987; Lockhart et al. 1994). In particular, differences in substitution rates among sites within a molecule, if unaccounted for, can cause inaccurate estimates of pairwise distances, and are thus an important cause of artifacts in phylogeny reconstruction (see Fig. 1) (Olsen 1987; Van de Peer et al. 1996a,b; Yang 1996; Swofford et al. 1996; Van de Peer and De Wachter 1997a; Philippe and Germot 2000). A few years ago, a method called "substitution rate calibration" (SRC), was developed for measuring the relative substitution rate of individual sites in a nucleotide sequence alignment on the basis of a distance approach by looking at the frequency at which sequence pairs differ at each site as a function of the distance between them (Van de Peer et al. 1996a). The obtained equation to compute evolutionary distances taking into account the distribution of substitution rates is similar to the general formula proposed by Rzhetsky and Nei (1994) to compute gamma distances (with $p = \frac{3}{4}\alpha$; see Fig. 1) but the novelty of the SRC method lies mainly in the estimation of individual substitution rates—using an approach based on genetic distances—which can be based on the comparison of thousands of sequences (as in this study) and which is independent of predefined tree topologies.

Here, we present an SSU rRNA nucleotide variability map and a rate-calibrated phylogenetic tree based on all known and compiled SSU rRNA sequences [2551 sequences; January 2000 (Van de Peer et al. 2000a)] from eukaryotes belonging to the crown radiation. By including all currently available sequences, this tree gives a clear idea about the species representation and sampling size of the different eukaryotic groups, within the SSU rRNA database and about the large differences in evolutionary rate in and among the different eukaryotic lineages. The resulting tree is the most comprehensive description by far of the overall organization and diversity of the eukaryotic crown published to date, previous studies being limited to at most 500 sequences or typically much fewer (Kumar and Rzhetsky 1996; Van de Peer and De Wachter 1997a; Lipscomb et al. 1998). Another large and comprehensive tree, which is basically a concatenation of several smaller maximum-likelihood trees, can be downloaded from the RDP server (Maidak et al. 2000). However, to our knowledge this tree has never been properly published or discussed. The comprehensive ribosomal RNA phylogeny presented in the current paper is compared with protein phylogenies discussed in the literature and with a recently published protein phylogeny constructed on the basis of four concatenated protein-coding gene sequences.

Materials and Methods

The European SSU rRNA database, established at the University of Antwerp (UIA) in 1984, is continuously updated by scanning all major molecular databases for new or corrected rRNA sequences. The UIA database includes all complete or nearly complete sequences stored in the form of a secondary structure-based alignment. Individual sequence files contain information on primary and secondary structure and sequence annotation including literature references, accession numbers, and detailed taxonomic information about the source organism (see http://rrna.uia.ac.be/). For this study, the complete SSU rRNA sequence alignment was used, except for the hypervariable region E-23 (Van de Peer et al. 2000a).

Individual substitution rates (Van de Peer et al. 1996a) were calculated based on an alignment of 2551 SSU rRNA sequences. Rates were estimated for each aligned position containing a nucleotide, rather than a gap, in at least 25% of the sequences, i.e., 1574 sites. Invariant positions were omitted from the phylogenetic analyses. Evolutionary distances were estimated according to the equation in Fig. 1 (with $p = 0.32$; see text), and evolutionary trees were constructed by the neighbor-joining method (Saitou and Nei 1987). Estimation of nucleotide substitution rates and evolutionary distances, phylogenetic tree construction, and bootstrap resampling were done with the software package TREECON for Windows (Van de Peer and De Wachter 1997b), which runs on IBM-compatible computers and is available from the authors upon request (URL http://www.evolutionsbiologie.uni-konstanz.de/). Application of the SRC method to the 2551 sequence data set used in this study and the computation of 100 bootstrap trees (Felsenstein 1985) takes about 4 weeks on a Pentium III computer running at 400 MHz and with 128 Mbytes of RAM. The nucleotide variability map was constructed with the software RNAViz (De Rijk and De Wachter 1997), a versatile program developed to draw secondary structures of molecules in a fast and user-friendly way.

Alignment and analysis of concatenated nucleotide and deduced amino acid sequences for actin, α-tubulin, β-tubulin, and EF-1α are

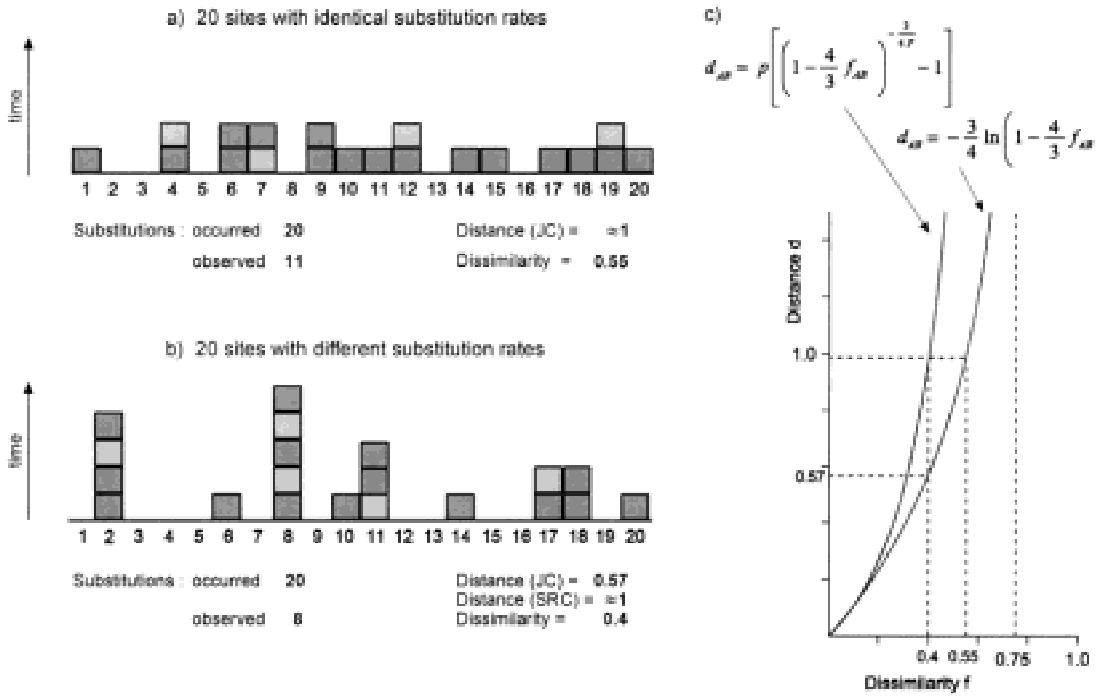

Fig. 1. **a** Hypothetical distribution of substitutions in a sequence of 20 nucleotides. It is assumed that the rate of substitution per site is the same for all sites in the sequence. In other words, substitutions (represented by squares) occur randomly. In this particular example, 11 substitutions are observed, although 20 have really occurred. *Dark grey squares* represent substitutions leading to a different nucleotide than the original one, while *light grey squares* represent substitutions resulting in the same nucleotide. Several sites have undergone multiple substitutions; e.g., site 4 has been mutated two times. When the dissimilarity (number of observed differences divided by the number of compared nucleotides, 11/20) is converted into evolutionary distance using the equation of Jukes and Cantor (1969), which assumes equal rates of substitution, a value of about 1 is obtained. This means that, on average, every site has been substituted once (20 substitutions in a sequence of 20 nucleotides). **b** Hypothetical distribution of substitutions in a sequence of 20 nucleotides but assuming different substitution rates among nucleotides. As a result, the majority of substitutions will take place at the same sites (i.e., sites 2, 8, and 11). Consequently, the number of observed substitutions will be smaller than in sequences where substitutions occur randomly (see a). In this particular example, the number of observed substitutions is only eight. If the distance is computed according to the Jukes and Cantor multiple hit correction (1969), the evolutionary distance is about 0.57. Therefore, the evolutionary distance is underestimated by about 40%, since the true evolutionary distance should be close to 1. **c** Graphic representation of the functions describing the relationship between dissimilarity (observed fraction of substitutions) and evolutionary distance (expected fraction of substitutions) when substitutions are assumed to occur randomly (*right curve*) and when substitution rates are assumed to differ among sites (*left curve*). Although the latter curve is hypothetical, its shape is similar to the function applicable to the eukaryotic SSU rRNA alignment (with $p = 0.32$).

described by Baldauf et al. (2000). In brief, after exclusion of several small gapped and ambiguously aligned regions and incomplete amino and carboxy termini, the final data set included 1528 amino acid positions, of which 423 were constant and 834 were parsimony informative (Baldauf et al. 2000). Amino acid sequences were analyzed by unweighted parsimony using PAUP* (Swofford 1998), while second codon-position nucleotide sequences were analyzed by maximum likelihood using DNAML as implemented in the PHYLIP package (Felsenstein 1993).

Results and Discussion

The Eukaryotic SSU rRNA Nucleotide Variability Map

After estimation of the nucleotide substitution rates, and dividing these into five variability subsets characterized by a different color, a color map superimposed on the secondary structure of the SSU rRNA can be constructed, as described previously for bacterial ribosomal RNA sequences (Van de Peer et al. 1996c). Such a color map can be interpreted in terms of higher-order structure, function, and evolution of the molecule. Figure 2 shows a nucleotide variability map superimposed on the secondary structure of the *Homo sapiens* SSU rRNA sequence and based on the comparison of all compiled eukaryotic SSU rRNA sequences belonging to the so called eukaryotic crown (see below). This map gives a much more detailed and quantitative description of positional variability than the crude distinction between variable and conserved areas that is often made by visual inspection of sequence alignments. The most conserved position has an estimated substitution rate of 0.00024, while the most variable position has a substitution rate of about 17. The varying amounts of colors, which indicate different levels of conservation, further indicate the non-

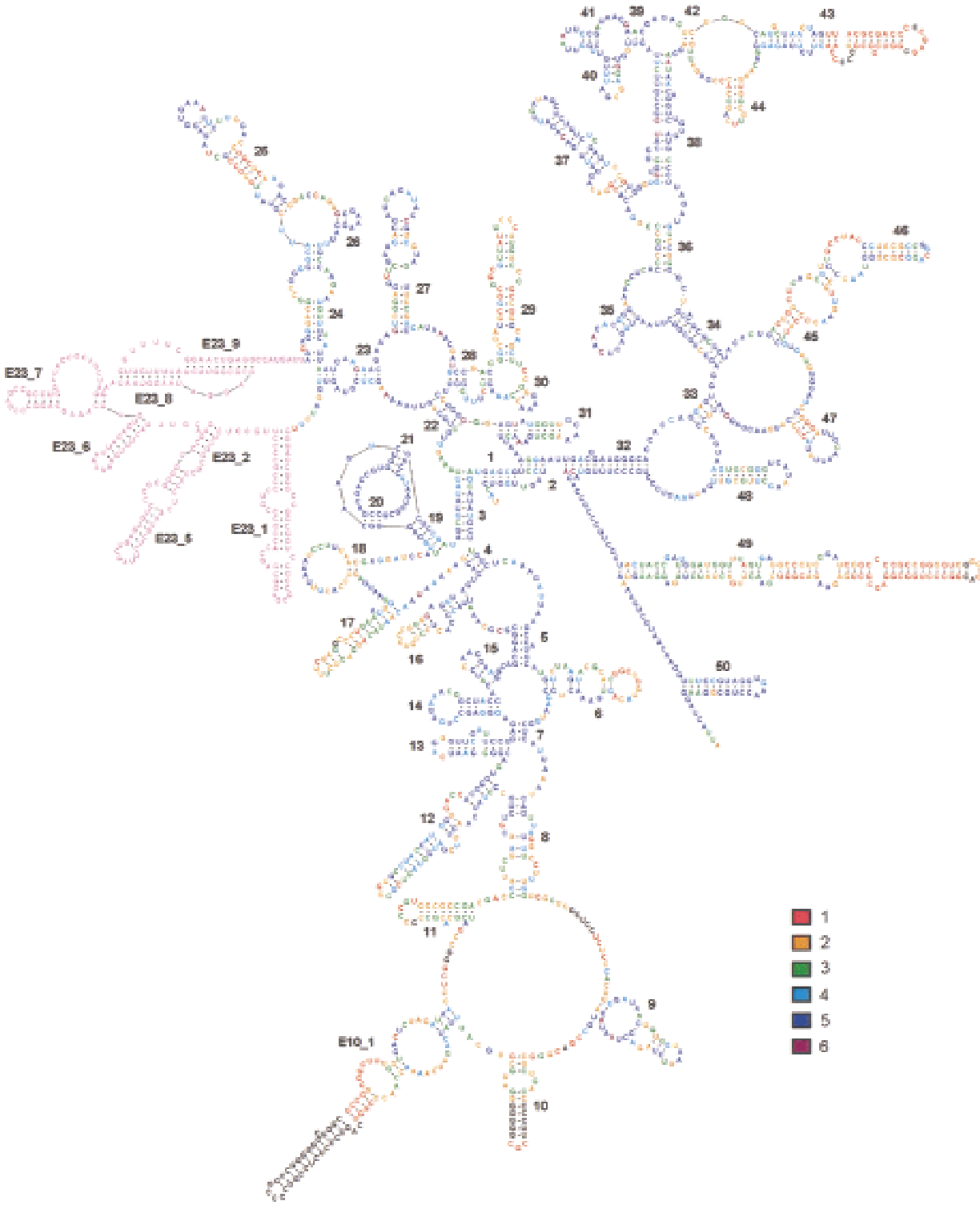

Fig. 2. Nucleotide variability map superimposed on the SSU rRNA secondary structure model of *Homo sapiens.* Nucleotides are divided into five groups of increasing variability with substitution rates measured relative to the average rate of the complete molecule. *Blue* indicates sites with estimated substitution rates smaller than 0.12, *cyan* indicates rates between 0.12 and 0.38, *green* indicates rates between 0.38 and 1.2, *orange* indicates rates between 1.2 and 3.8, and nucleotides with a substitution rate higher than 3.8 are shown in *red.* Absolutely conserved (invariant) positions (for which the substitution rate equals 0 i.e., 29 positions) are indicated in *purple.* Sites containing a nucleotide in *Homo sapiens* but vacant in >75% of the aligned sequences were not considered for the variability estimation and are indicated in *black.* The region in *pink* comprises the hypervariable region E23 (Van de Peer et al. 2000a) and was not taken into account in this study due to alignment problems.

rectangular nature of the rate distribution (Van de Peer et al. 1996c).

Effect of Taxon Sampling on Bootstrap Values

A comprehensive data set of all available SSU rRNA sequences for crown eukaryotes (Fig. 3) was analyzed using SRC-derived distances (Van de Peer et al. 1996a). The tree is arbitrarily rooted between the heterokont–alveolate cluster (see below) and all other taxa. Including all available sequences has the advantages of (1) giving a better idea of the sampling and representation of the different taxa studied and (2) showing the strikingly large differences in evolutionary rates in and among different lineages. Furthermore, it has been suggested that inferring large, comprehensive evolutionary trees may result in a higher accuracy. This is particularly true when long branches are included since these are then more likely to be subdivided in a larger data set (Hillis 1996; but see Poe and Swofford 1999).

Nonetheless, adding many long branches to a data set also tends to decrease the bootstrap support, particularly for clades containing these long branches, even when the true clade is recovered (Hillis and Bull 1993). Indeed, compared to a much smaller, 500-taxon tree published previously (Van de Peer and De Wachter 1997a), many nodes in our tree receive considerably lower bootstrap support (%BP) due to the inclusion of many long-branched taxa. Therefore we suggest that bootstrap values of >60% in such a large data set as the 2551-taxon tree probably indicate robust clades, since these trees still reconstruct similar overall topologies to smaller trees, for which these clades often show significantly higher bootstrap support (data not shown).

The detrimental effect of long branches on bootstrap values is further indicated by the observation that the removal of only a few long branches within specific clusters often significantly increases bootstrap numbers. For example, in the case of the Viridiplantae (green algae plus land plants), when the nucleomorph sequences (see further) are omitted from the analysis, the Viridiplantae form a highly supported monophyletic group with bootstrap support of 78%.

A Comprehensive Phylogeny of the Eukaryotic Crown

Overall, four major groups (or superphyla) are confidently reconstructed in the complete 2551-taxon tree. These are the Opisthokonta (animals, fungi, and choanoflagellates), the Plantae (green algae, land plants and red algae), the stramenopiles, and the alveolates.

The Opisthokonta. The opisthokont clade (Cavalier-Smith 1998) is now firmly established by SSU rRNA analyses [63% BP, Fig. 3; 94% BP in Van de Peer and De Wachter (1997a); see also Wainright et al. 1993], by its unique possession of a large insertion in protein synthesis elongation factor-1α (EF-1α), and by analyses of the conservative, taxonomically well-sampled proteins α-tubulin β-tubulin, EF-1α, and actin (Fig. 4) (Baldauf 1999).

Although SSU rRNA data seem to resolve fungal relationships consistently and with high bootstrap support for many major clades (e.g., Bruns et al. 1992; Van de Peer et al. 1995), the overall divergence order among animal taxa is not well resolved by these data (Adoutte and Philippe 1993; Abouheif et al. 1998). Nevertheless, some important conclusions about metazoan relationships can be drawn on the basis of SSU rRNA. These include recognition of the Ecdysozoa, which include molting invertebrates such as nematodes and arthropodes (Aguinaldo et al. 1997) (Fig. 3), as well as an emerging overall consensus regarding most relationships between and within major animal taxa (Adoutte et al. 1999, 2000). Likewise, protein sequence data also seem readily to resolve relationships among fungi but not among animals. The latter appears to reflect massive gene duplications very early in animal evolution, predating the diploblast–triploblast split (Iwabe et al. 1996). In addition, there often appear to be vast differences in evolutionary rates among different animal lineages for protein-encoding genes (Kim et al. 1999) (Fig. 3).

Within the animal–fungal clade, the choanoflagellates are of special interest, as they have long been considered candidates for being the sister group of multicellular animals due to their strong resemblance to the collar cells of sponges. Although SSU rRNA phylogenies place these taxa together with animals and fungi, it does not confidently resolve the branching order among these three clades (Fig. 3) (Van de Peer and De Wachter 1997a). Preliminary analysis of a single choanoflagellate EF-1α sequence gives a very similar result (Baldauf 1999).

The Plantae. Rate-calibrated phylogenies of SSU rRNA strongly support the monophyly of both the Rhodophyta and the Viridiplantae. Likewise, the monophyly of green algae and land plants is also reconstructed by most protein sequence phylogenies and strongly supported by trees of β-tubulin and actin. However, with the exception of actin, green algae are usually represented solely by members of the Volvocales (i.e., *Chlamydomonas* and/or *Volvox*). Rhodophyte algae, poorly represented for nonorganellar protein data, are depicted as monophyletic in phylogenies of actin, despite exhibiting highly accelerated evolutionary rates (Bhattacharya and Weber 1997), and in a recent phylogeny based on elongation factor 2 (EF-2) (Moreira et al. 2000).

Inferred relationships among the red algae, glaucocystophytes, and Viridiplantae address the long-standing debate over a single versus multiple independent origins of eukaryotic photosynthesis. Since the plastids of these or-

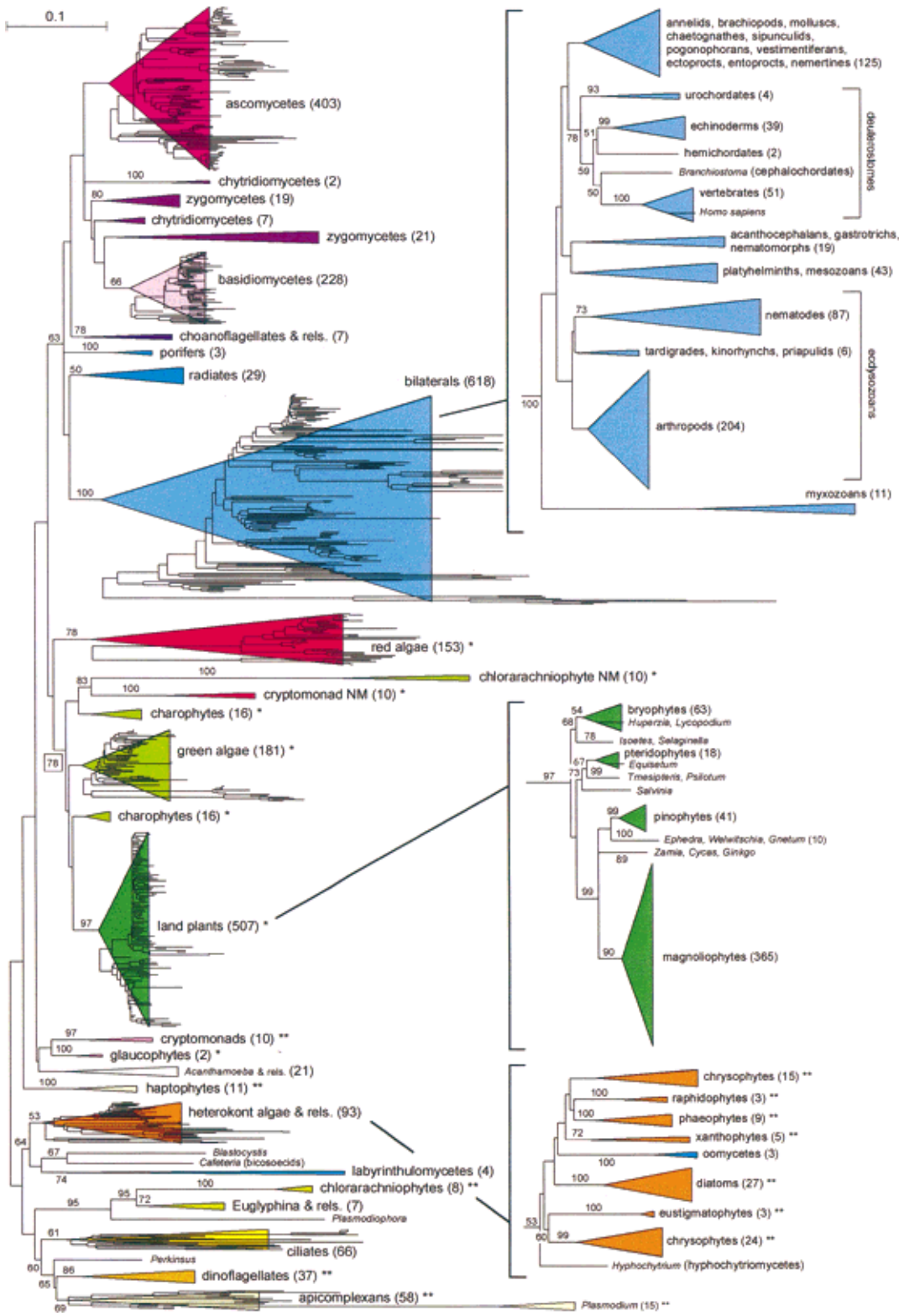

Fig. 3. Schematic representation of a phylogenetic tree reconstructed from all SSU rRNA sequences (2551) available in the European SSU rRNA database [January 2000 (Van de Peer et al. 2000a)] from organisms belonging to the so-called eukaryotic crown. The root was arbitrarily placed between the stramenopile–alveolate cluster and the other taxa. The *scale* on top measures evolutionary distance in substitutions per nucleotide, taking into account among-site rate variation. Clusters of organisms are represented as isosceles *triangles*, with a height equal to the average distance separating the terminal nodes from the deepest-branching point in the cluster and a base proportional to the number of sequences composing it. The number of sequences comprised in each cluster is shown in *parentheses* after the taxon name. Photosynthetic groups or organisms are indicated by an *asterisk*. Groups containing plastids of secondary origin are indicated by *two asterisks*. Bootstrap values above 50% are shown at the internodes. For the larger taxonomic groups, differences in branch lengths—estimated from SRC distances—are shown for a subset of species.

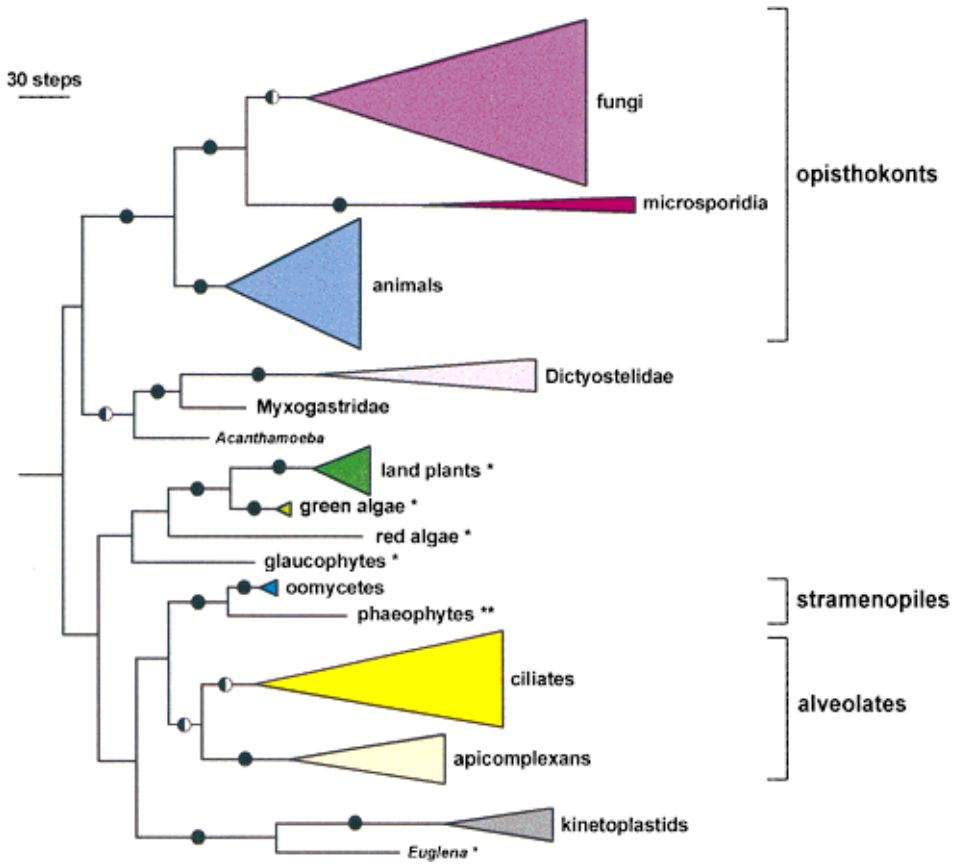

Fig. 4. Eukaryote phylogeny based on concatenated, deduced amino acid sequences of actin, α-tubulin, β-tubulin, and EF-1α. Clusters of organisms are represented as isosceles *triangles*, with a height equal to the average distance separating the terminal nodes from the deepest branching point in the cluster and a base proportional to the number of sequences composing it. Photosynthetic groups or organisms are indicated by an *asterisk*, while groups containing plastids of secondary origin are indicated by *two asterisks. Solid black* circles correspond to bootstrap support of more than 70% for both maximum-parsimony and maximum-likelihood analyses. *Half-black circles* correspond to bootstrap support of more than 70% for maximum-parsimony analysis only. The figure is adapted from Baldauf et al. (2000).

ganisms are surrounded by only two membranes, they are considered to be the result of a primary endosymbiotic event involving a eukaryote and a cyanobacterium. Thus, if the three groups of photosynthetic eukaryotes harboring these plastids form a monophyletic clade, then a single endosymbiotic event at the base of this clade could have given rise to all eukaryotic photosynthesis. In support of this scenario, the 2551-taxon phylogeny reconstructs a green–red clade (Fig. 3), although not statistically strongly supported. In fact, this relationship is often not reconstructed at all in SSU trees derived by other methods (see, e.g., Kumar and Rzhetsky 1996; Ragan and Gutell 1995).

Accumulating evidence from plastid (e.g., Turmel et al. 1999) and mitochondrial (Burger et al. 1999) genomes support a red–green algal clade including shared rearrangements in plastid genome operons (Douglas 1998). This grouping is also supported by trees based on glyceraldehyde-3-phosphate dehydrogenase (Ragan and Gutell 1995) and actin (Bhattacharya and Weber 1997) and particularly strongly by trees based on EF-2 (Moreira et al. 2000). Analyses of a four-protein data set including only a single incomplete red algal sequence also find this grouping but without significant statistical support (Fig. 4) (Baldauf et al. 2000), while a concatenated data set of 13 different and complete protein sequences supports this conclusion strongly (Moireira et al. 2000). Analyses of actin sequences (Bhattacharya and Weber 1997) and combined protein data sets also weakly support an all-inclusive red–green–glaucocystophyte clade (Fig. 4) (Baldauf et al. 2000; Moreira et al. 2000).

Two apparent members of this possible green–red clade warrant special mention. These are the chlorarachniophytes and the cryptomonads. Previous work has shown that the plastids of these organisms once belonged to free-living algae, green algae in the case of the chlorarachniophytes, red algae in the case of the cryptomonads (Van de Peer et al. 1996b). Both taxa retain a vestige of the plastid's original host nucleus, referred to as a nucleomorph, despite being the result of separate endosymbiotic events. Since these two nucleomorph genomes should be unrelated, the clustering of their sequences in the SSU tree (Fig. 3) is unexpected (Cavalier-Smith et al. 1994; Van de Peer et al. 1996b; Van de Peer

and De Wachter 1997a). This attraction appears to be artifactual since, when either the red or the green nucleomorph sequences are omitted from the analysis, the other group clusters as expected, i.e., the chlorarachniophyte nucleomorphs cluster with the green algae, and the cryptomonad nucleomorphs with the red algae (Van de Peer et al. 1996b; Van de Peer and De Wachter 1997a; data not shown). The strong artificial attraction between the nucleomorph SSU rRNAs may be explained by covariation (Pete Lockhart, personal communication) or convergences in their SSU rRNA sequences due to similar evolutionary pressure after becoming endosymbionts (Ishida et al. 1999). Alternatively, this may simply be an artifactual attraction of long branches from the same tree neighborhood. In fact, the latter would indirectly argue for a red algae–green plants clade (Moreira et al. 2000), since that would make the nucleomorph sequences each other's closest long-branch neighbor.

The Stramenopiles and Alveolates. The stramenopiles or heterokonts are united by the possession of two unequal flagella, one smooth and one tinsellated. This cluster [64% BP, Fig. 3; 84% BP in the 500-taxon tree (Van de Peer and De Wachter 1997a)] includes the bicosoecids (represented by *Cafeteria*), the enigmatic taxon *Blastocystis*, labyrinthulomycetes, oomycetes, hyphochytriomycetes, and the heterokont algae (Fig. 3). Within the heterokonts, the (chlorophyll a+c) algae, oomycetes and hyphochytriomycetes cluster together with low bootstrap support. The subdivisions of the heterokont algae are shown in more detail at the right in Fig. 3 and discussed in detail elsewhere (Leipe et al. 1994; Saunders et al. 1995). All protein phylogenies which sample these taxa reconstruct the group, including actin (Bhattacharya and Weber 1997). β-tubulin (Keeling and Doolittle 1996), and hsp70 (Budin and Phillippe 1998). However, only the oomycetes and heterokont algae are represented in these trees.

Finally, the alveolates comprise the ciliates, dinoflagellates, and apicomplexans and are characterized by the possession of alveoli: membrane-bound flattened vesicles or sacs underlying the plasma membrane. The alveolate clade is strongly supported in SRC-corrected and all other rRNA trees, as is a subclade of apicomplexans and dinoflagellates [65%BP, Fig. 3; 97%BP in Van de Peer and De Wachter (1997a)]. Although tubulin (Keeling and Doolittle 1996) and hsp70 (Budin and Philippe 1998) phylogenies support a monophyletic Alveolata, ciliate phylogeny is notoriously problematic with both EF-1α and actin due to highly accelerated evolutionary rates (Moreira et al. 1998, Baldauf 1999).

The tree in Fig. 3 gives little indication of the possible branching order among the superphyla. However, phylogenetic reconstructions based on SSU rRNA regularly suggest the heterokons and alveolates to be sister groups (e.g., Wolters 1991; Saunders et al. 1995), although bootstrap support for this grouping is usually low. However, this relationship is also seen in the four-protein tree with moderate bootstrap support (Fig. 4) (Baldauf et al. 2000) and with strong support in an analysis of combined cytosolic and endoplasmic reticulum forms of hsp70 (Germot and Philippe 1999). Furthermore, Cavalier-Smith (2000) has pointed out that such a grouping would significantly reduce the number of postulated secondary endosymbiotic events necessary to explain the distribution of plastids with greater than two membranes. In the tree in Fig. 3, chlorarachniophytes, Euglyphina and relatives, and *Plasmodiophora* also arise from within this heterokont and alveolate group.

Apart from the main clusters discussed, several smaller independent evolutionary lineages seem to exist, which show no specific relationship to any of the other major groups. Among these are the haptophytes, cryptomonads, glaucophytes, and *Acanthamoeba* and relatives (Fig. 3). These taxa may be important missing links among major clades, and the further elucidation of their biology and evolutionary history should be given special attention. For example, phylogenies based on actin (e.g., Bhattacharya and Weber 1997) and combined protein sequences strongly support *Acanthamoeba* in a clade together with the slime molds (Fig. 4) (Baldauf et al. 2000). This is consistent with traditional protistan taxonomy, which often placed these taxa together based on the similarity of their pseudopodia (Cavalier-Smith 1998).

Rates in the Eukaryotic Crown and the Effect of Among-Site Rate Correction

The SSU rRNA tree presented shows extreme differences in evolutionary rates between and within major taxonomic groups (Fig. 3). Furthermore, this tree shows important differences to trees where among-site rate variation is not considered. These particularly involve the apparently fast evolving lineages (the so-called long-branch sequences). Due to the serious underestimation of large evolutionary distances if site-to-site rate variation is not considered (see Fig. 1), distant species seem much closer to one another than they actually are. This often results in artificial clustering of long branches or to long branches being pulled closer to the base of the tree (Olsen 1987; Van de Peer et al. 1996a, 2000b; Philippe and Germot 2000; Philippe et al. 2000b), thereby seriously distorting the inferred tree topologies. When the branch lengths are more accurately estimated, making the long branches even longer, these artifactual attractions often disappear.

Of the extremely long branches in the tree, the bilateral animals and the apicomplexan *Plasmodium* warrant special note. Apart from showing huge differences in evolutionary rate between different bilateral taxa (Fig. 3), the bilateral animal SSU sequences have evolved much faster than those of the radiate animals and many

other crown taxa and are therefore likely to be seriously underestimated. This probably explains why many rRNA trees using simple rate corrections did not reconstruct a monophyletic animal clade (e.g., Field et al. 1988; Christen et al. 1991; Cavalier-Smith et al. 1994) and even did not show fungi and animals as sister groups, by now a widely accepted relationship. In distance trees based on SRC, the monophyly of animals is nearly always confirmed, although not necessarily supported by significantly high bootstrap values, and, importantly, animals are always strongly grouped with fungi (Fig. 3) (Van de Peer and De Wachter 1997a).

Apart from the fast-evolving SSU rRNA sequences encoded by the nucleomorphs of chlorarachniophytes and cryptomonads that are discussed above (and by Van de Peer et al. 1996b), another serious long-branch problem in SSU trees is the sequences of *Plasmodium.* In most studies that do not take into account the substitution rates of individual nucleotides, *Plasmodium* is clustered erroneously near the base of the trees. Even in SSU rRNA trees including only apicomplexans, dinoflagellates, and ciliates, *Plasmodium* clusters with the other apicomplexans but only with low statistical reliability and is separated from the other hematozoans by the coccidians (Escalante and Ayala 1995). However, in our rate-calibrated trees, the monophyly of the apicomplexa and the class Hematozoa, including *Plasmodium,* is well supported (Fig. 3) (Van de Peer et al. 1996a; Van de Peer and De Wachter 1997a).

It is interesting to note that many of the relative rate differences seen in SSU trees are absent or reversed in protein trees. In actin, and especially α- and β-tubulin phylogenies, it is the fungi that have remarkably long terminal branches relative to animals as well as most other eukaryotes [1.5 to 4.5-fold (Baldauf 1999); see also Fig. 4]. Likewise, while the ciliates have remarkably long branches in actin and EF-1α trees (Baldauf 1999), none of the broadly sampled protein data sets show any obvious signs of rate acceleration for *Plasmodium* or any other apicomplexan. Possibly, these phenomena are the result of different selection pressures on different genes in different lineages, a possibility that should be followed up with protein data sets of these groups where natural selection can be detected more clearly.

Is the SSU Eukaryote Crown Real?

As stated in the Introduction, the eukaryotic crown has been defined mainly on the basis of SSU rRNA trees. However, recent protein data disagree on both the composition and even the possible existence of the eukaryote crown radiation (see, e.g., Baldauf et al. 2000). Perhaps most notable are the microsporidia, which often appear strongly allied with the fungi (Keeling and Doolittle 1996), and the slime molds, which, unlike in SSU rRNA trees (but see below), are monophyletic and closely allied with animals and fungi with all examined proteins (Kuma et al. 1995; Keeling and Doolittle 1996; Baldauf and Doolittle 1997; Bhattacharya and Weber 1997) (Fig. 4). Thus it is possible that the increased evolutionary rates of the SSU rRNAs in these lineages may have blurred their true phylogenetic position in these trees, as suggested previously (e.g., Stiller and Hall 1999; Philippe et al. 2000). On the other hand protein data do generally agree with SSU rRNA on a basal position for the diplomonads, often represented solely by *Giardia* (Germot and Philippe 1999; Hirt et al. 1999; Roger et al. 1999). However, while some protein data also show the parabasalid *Trichomonas* arising on the same lineage or immediately after the diplomonads (Keeling and Doolittle 1996; Baldauf 1999), also in agreement with SSU rRNA phylogeny, other protein data suggest that many Trichomonad sequences exhibit very fast evolutionary rates (Germot and Philippe 1999). Thus protein sequences from this taxon often appear at highly incongruous positions in some protein trees, sometimes even being found within the animal–fungal clade (e.g., Germot and Philippe 1999).

It is important, albeit difficult, to test whether extreme-long branches are affecting some aspects of SSU rRNA based phylogenies. To this end, rate calibration should be applied to the sequences of microsporidians, parabasalids and diplomonads. However, how this should be done is not obvious. As can be seen in Fig. 1, the SRC function applicable to the eukaryotic SSU rRNA alignment lies considerably lower than the function applied to sequences for which substitutions occur randomly over the sequence alignment. As a result, estimation of evolutionary distances for highly divergent sequences (dissimilarity >0.4) becomes more error prone, due to the large variances for large distances, causing serious distortions in the inferred tree topologies.

Initial attempts to test the phylogenetic position of some of these more controversial taxa have yielded inconsistent results (YVdP, unpublished data; see also Philippe et al. 2000). For example, these analyses tend consistently to depict the slime molds as clustered together and diverging from within the eukaryotic crown (data not shown). However, while the parabasalids still diverge before the other eukaryotic taxa, so do the microsporidia. Therefore, we feel that it will be hard, if not impossible, to elucidate correctly the evolutionary position of the so-called early-branching protists on the basis of SSU rRNA. The large-subunit ribosomal RNA (LSU rRNA) may prove more promising in this respect. For example, we recently demonstrated that the SRC method applied to the LSU rRNA yields trees where microsporidia and fungi are consistently clustered together (Van de Peer et al. 2000b). Together with recent phylogenies constructed on the basis of protein data, and similarities such as the presence of chitin and trehalose and similar characteristics of the meiotic and mitotic cycles (Sprague

et al. 1992; Keeling and Doolittle 1996; Germot et al. 1997; Keeling, 1998), evidence for a close evolutionary relationship between fungi and microsporidia has now become quite convincing. This of course drastically changes our view about early eukaryotic evolution, and consequently the microsporidia example sheds doubt on the phylogenetic position of other early branching protists as well. Additionally, this even sheds doubt on the existence of the so-called eukaryotic crown since more and more taxa seem to diverge from within this crown (see, e.g., Embley and Hirt 1998; Keeling 1998; Stiller and Hall 1999; Philippe et al. 2000). In this respect, it is also noteworthy to mention that in preliminary studies based on SRC phylogenies of the LSU rRNA, the Euglenozoa (*Euglena* plus kinetoplastids; Fig. 4), which usually also form one of the early-branching lineages on the basis of SSU rRNA, seem to diverge from within the crown and even show a specific relationship with green algae and land plants (data not shown).

Conclusions

While it has been suggested that the crown structure of the SSU tree may itself be an artifact of many slowly evolving sequences clustering together to the exclusion of a rapidly evolving few (Palmer and Delwiche 1996; Stiller and Hall 1999; Philippe et al. 2000), many protein trees also show or suggest a difficult-to-root, largely unresolved clustering of most eukaryote phyla. This is most striking in α- and β-tubulin phylogenies (e.g., Keeling and Doolittle 1996; Keeling et al. 2000). Although most other proteins give a ladder-like progression of taxa sequentially branching off of a main line of ascent, the intermediate region of these trees is generally poorly resolved (Bhattacharya and Weber 1997; Roger et al. 1999). Therefore, individual protein trees are also consistent with a large, unresolved radiation of the majority of eukaryote phyla. Nevertheless, analyses of combined sequence data suggest that, given adequate information from multiple independent sources, many of these relationships can in fact be resolved (Baldauf et al. 2000; Moreira et al. 2000) and a consensus is beginning to emerge regarding the main structure, organization, and content of many of the crown taxa (Figs. 3 and 4).

Thus, despite considerable debate on the vices and virtues of various molecular phylogenetic markers, it appears that all have their strengths and weaknesses. Probably no single gene is completely immune to artifacts such as horizontal gene transfer, hidden paralogy, and changes in evolutionary mode or tempo among organismal lineages. Therefore, it is unlikely that any one gene will accurately reconstruct all eukaryote phylogeny, even at similar taxonomic levels (Baldauf et al. 2000). Nonetheless, the value of the SSU rRNA database as a coherent sequence database of all taxa makes it an incomparable tool for many areas of comparative biology. The extent of the rRNA database is still far and away unrivaled among proteins and will undoubtedly remain so for a long time, making SSU rRNA the foundation and reference point for other molecular phylogenetic studies. Furthermore, when analyzed carefully, particularly by taking into account the large differences in substitution rates among sites of the molecule, SSU rRNA remains an important tool for higher-order systematics.

Acknowledgments. S.L.B. and W.F.D. are supported by Grant 4467 from the Medical Research Council of Canada. Y.V.d.P. acknowledges the support of the Special Research Fund of the University of Antwerp and the University of Konstanz. Support from the NSF, USA, the German Science Foundation, University of Konstanz, and the FCI to A.M. is acknowledged. The authors thank Henner Brinkmann for critical comments and suggestions and Abdelghani Ben Ali for the construction of the nucleotide variability map. Y.V.d.P. is a Research Fellow of the National Fund for Scientific Research—Flanders (Belgium).

References

Abouheif E, Zardoya R, Meyer A (1998) Limitations of metazoan 18S rRNA sequence data: implications for reconstructing a phylogeny of the animal kingdom and inferring the reality of the Cambrian explosion. J Mol Evol 47:394–405

Adoutte A, Philippe H (1993) The major lines of metazoan evolution: Summary of traditional evidence and lessons from ribosomal RNA sequence analysis. In: Pichon Y (ed) Comparative molecular neurobiology. Birkhäuser Verlag, Basel, pp 1–30

Adoutte A, Balavoine G, Lartillot N, de Rosa R (1999) Animal evolution the end of the intermediate taxa? Trends Genet 15:104–108

Adoutte A, Balavoine G, Larttillot N, Lespinet O, Prud'homme B, de Rosa R (2000) The new animal phylogeny: reliability and implications. Proc Natl Acad Sci USA 97:4453–4456

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–493

Baldauf SL (1999) A search for the origins of animals and fungi, comparing and combining molecular data. Am Nat 154:S178–S188

Baldauf SL, Doolittle WF (1997) Origin and evolution of the slime molds. Proc Natl Acad Sci USA 94:12007–12012

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom level phylogeny of eukaryotes based on combined protein data. Science (in press)

Bhattacharya D, Weber K (1997) The actin gene of the glaucocystophyte *Cyanophora paradoxa:* Analysis of the coding region and introns, and an actin phylogeny of eukaryotes. Curr Genet 31:439–446

Bruns TD, Vilgalys R, Barns SM, Gonzalez D, Hibbett DS, Lane DJ, Simon L, Stickel S, Szaro TM, Weisburg WG, et al. (1992) Evolutionary relationships within the fungi: Analyses of nuclear small subunit rRNA sequences. Mol Phylogenet Evol 1:231–241

Budin K, Philippe H (1998) New insights into the phylogeny of eukaryotes based on ciliate hsp70 sequences. Mol Biol Evol 15:943–956

Burger G, Saint-Louis D, Gray MW, Lang BF (1999) Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*. Cyanobacterial introns and shared ancestry of red and green algae. Plant Cell 11:1675–1694

Cavalier Smith T (1998) A revised six kingdom system of life. Biol Rev Cambr Philos Soc 73:203–266

Cavalier-Smith T (2000) Membrane heredity and early chloroplast evolution. Trends Plant Sci 5:174–182

Cavalier-Smith T, Allsop MTEP, Chao EE (1994) Chimeric conundra: Are nucleomorphs and chromists monophyletic or polyphyletic? Proc Natl Acad Sci USA 91:11368–11372

Christen R, Ratto A, Baroin A, Perasso R, Grell KG, Adoutte A (1991) An analysis of the origin of metazoans, using comparisons of partial sequences of the 28S RNA, reveals an early emergence of triploblasts. EMBO J 10:499–503

De Rijk P, De Wachter R (1997) RnaViz, a program for the visualisation of RNA secondary structure. Nucleic Acids Res 25:4679–4684

Doolittle WF, Brown JR (1994) Tempo, mode, the progenote, and the universal root. Proc Natl Acad Sci USA 91:6721–6728

Doolittle RF, Feng DF, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Douglas SE (1998) Plastid evolution: Origins, diversity, trends. Curr Opin Genet Dev 8:655–661

Embley TM, Hirt RP (1998) Early branching eukaryotes. Curr Opin Genet Dev 8:624–629

Escalante AA, Ayala FJ (1995) Evolutionary origin of Plasmodium and other Apicomplexa based on rRNA genes. Proc Natl Acad Sci USA 92:5793–5797

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

Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5c. Department of Genetics, University of Washington, Seattle (distributed by the author)

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

Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, Wolfe RS, Balch WE, Tanner RS, Magrum LJ, Zablen LB, Blakemore R, Gupta R, Bonen L, Lewis BJ, Stahl DA, Luehrsen R, Chen KN, Woese CR (1980) The phylogeny of prokaryotes. Science 209:457–463

Germot A, Philippe H (1999) Critical analysis of eukaryotic phylogeny: a case study based on the hsp70 family. J Euk Microbiol 46:116–124

Hillis DM (1996) Inferring complex phylogenies. Nature 383:130–131

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 42:182–192

Hirt RP, Logsdon Jr JM, Healy B, Dorey MW, Doolittle WF, Embley TM (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

Ishida K, Green BR, Cavailier-Smith T (1999) Diversification of a chimeric algal group, the chlorarachniophytes: Phylogeny of nuclear and nuclear and nucleomorph small subunit rRNA genes. Mol Biol Evol 16:321–331

Iwabe N, Kuma KI, Miyata T (1996) Evolution of gene families and relationships with organismal evolution: Rapid divergence of tissue-specific genes in the early evolution of chordates. Mol Biol Evol 13:483–493

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HH (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–132

Keeling PJ (1998) A kingdoms progress: Archezoa and the origin of eukaryotes. BioEssays 20:87–95

Keeling PK, Doolittle WF (1996) Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol 13:1297–1305

Keeling PJ, Luker MA, Palmer JD (2000) Evidence from beta-tubulin phylogeny that Microsporidia evolved from within the fungi. Mol Biol Evol 17:23–31

Kim J, Kim W, Cunningham CW (1999) A new perspective on lower Metazoan relationships from 18S rDNA sequences. Mol Biol Evol 16:423–427

Knoll AH (1992) The early evolution of eukaryotes: A geological perspective. Science 256:622–627

Knoll AH (1999) A new molecular window on early life. Science 285:1025–1026

Kuma KI, Nikoh N, Iwabe N, Miyata T (1995) Phylogenetic position of Dictyostelium inferred from multiple protein data sets. J Mol Evol 41:238–246

Kumar S, Rzhetsky A (1996) Evolutionary relationships of eukaryotic kingdoms. J Mol Evol 42:183–193

Leipe DD, Wainright PO, Gunderson JH, Porter D, Patterson DJ, Valois F et al. (1994) The stramenopiles from a molecular perspective: 16S-like rRNA sequences from Labyrinthuloides minuta and Cafeteria roenbergensis. Phycologia 33:369–377

Lipscomb DL, Farris JS, Källersjö M, Tehler A (1998) Support, ribosomal sequences and the phylogeny of eukaryotes. Cladistics 14: 303–338

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Maidak BL, Cole JR, Lilburn TG, Parker CT Jr, Saxman PR, Stredwick JM, Garrity GM, Li B, Olsen GJ, Pramanik S, Schmidt TM, Tiedje JM (2000) The RDP (Ribosomal Database Project) continues. Nucleic Acids Res 28:173–174

Mayr E (1998) Two empires or three? Proc Natl Acad Sci USA 95: 9720–9723

Moreira D, Le Guyader H, Philippe H (1998) Unusually high evolutionary rate of the elongation factor 1a genes from the ciliophora and its impact on the phylogeny of eukaryotes. Mol Biol Evol 16:234–245

Moreira D, Le Guyader H, Philippe H (2000) The origin of red algae implications for the evolution of chloroplasts. Nature 405:69–72

Olsen GJ (1987) Earliest phylogenetic branchings: Comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol LII:825–837

Olsen GJ, Woese CR (1993) Ribosomal RNA: A key to phylogeny. FASEB J 7:113–123

Pace NR, Olsen GJ, Woese CR (1986) Ribosomal RNA phylogeny and the primary lines of evolutionary descent. Cell 9:325–326

Palmer JD, Delwiche CF (1996) Second-hand chloroplasts and the case of the disappearing nucleus. Proc Natl Acad Sci USA 93:7432–7435

Philippe H, Germot A (2000) Phylogeny of eukaryotes based on ribosomal RNA: Long branch attraction and models of sequence evolution. Mol Biol Evol 17:830–834

Philippe H, Lopez P, Brinkmann H, Budin K, Germot A, Laurent J, Moreira D, Müller M, Le Guyader H (2000) Early branching or fast evolving eukaryotes? An answer based on slowly evolving positions. Proc Roy Soc Ser B 267:1213–1222

Poe S, Swofford D (1999) Taxon sampling revisited. Nature 398:299–300

Ragan MA, Gutell RR (1995) Are red algae plants? Bot J Linn Soc 118:81–105

Roger AJ, Sandblom O, Doolittle WF, Philippe H (1999) An evaluation of elongation factor 1 alpha as a phylogenetic marker. Mol Biol Evol 16:218–233

Rzhetsky A, Nei M (1994) Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J Mol Evol 38:295–299

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Saunders GW, Potter D, Paskind MP, Andersen RA (1995) Cladistic analyses of combined traditional and molecular data sets reveal an algal lineage. Proc Natl Acad Sci USA 92:244–248

Sprague V, Becnel JJ, Hazard EI (1992) Taxonomy of phylum Microspora. Crit Rev Microbiol 18:285–395

Sogin ML (1989) Evolution of eukaryotic microorganisms and their small subunit RNAs. Am Zool 29:487–499

Stiller JW, Hall BD (1999) Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol 16:1270–1279

Swofford DL (1998) PAUP* Phylogenetic analysis using parsimony (*and other methods), version 4 Sinauer Associates Sunderland MA

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference In: Hillis DM, Moritz C, Mable BK (eds Molecular systematics. Sinauer, Sunderland MA, pp 407–514

Turmel M, Otis C, Lemieux C (1999) The complete chloroplast DNA sequence of the green alga Nephroselmis olivacea: Insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci USA 96:10248–10253

Van de Peer Y, De Wachter R (1995) Investigation of fungal phylogeny on the basis of small ribosomal subunit RNA sequences. In: Molecular microbial ecology manual 3.3.4. Kluwer Academic, Dordrecht, The Netherlands, pp 1–12

Van de Peer Y, De Wachter R (1997a) Evolutionary relationships among the eukaryotic crown taxa taking into account site to site rate variation in 18S rRNA. J Mol Evol 45:619–630

Van de Peer Y, De Wachter R (1997b) Construction of evolutionary distance trees with TREECON for Windows: Accounting for variation in nucleotide substitution rate among sites. Comput Applic Biosci 13:227–230

Van de Peer Y, Van der Auwera G, De Wachter R (1996a) The evolution of stramenopiles and alveolates as derived by "substitution rate calibration" of small ribosomal subunit RNA. J Mol Evol 42: 201–210

Van de Peer Y, Rensing S, Maier U-G, De Wachter R (1996b) Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae. Proc Natl Acad Sci USA 93:7732–7736

Van de Peer Y, Chapelle S, De Wachter R (1996c) A quantitative map of nucleotide substitution rates in bacterial ribosomal subunit RNA. Nucleic Acids Res 24:3381–3391

Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T, De Wachter R (2000a) The European small subunit ribosomal RNA database. Nucleic Acids Res 28:175–176

Van de Peer Y, Ben Ali A, Meyer A (2000b) Microsporidia: accumulating molecular evidence that a group of amitochondriate and suspectedly primitive eukaryotes are just curious fungi. Gene 246:1–8

Wainright PO, Hinkle G, Sogin ML, Stickel SK (1993) Monophyletic origins of the metazoa an evolutionary link with fungi. Science 260:340–342

Wang DY, Kumar S, Hedges SB (1999) Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc R Soc Lond B Biol Sci 266:163–171

Wolters J (1991) The troublesome parasites—Molecular and morphological evidence that Apicomplexa belong to the dinoflagellate ciliate clade. BioSystems 25:75–83

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372