Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions

Hervé Philippe^{1*}, Philippe Lopez¹, Henner Brinkmann¹, Karine Budin¹,
Agnès Germot¹, Jacqueline Laurent¹, David Moreira¹, Miklós Müller²
and Hervé Le Guyader¹

¹*Phylogénie et Evolution Moléculaires (UPRES-A 8080), Bât. 444, Université Paris-Sud, 91405 Orsay, France*

²*The Rockefeller University, New York, NY 10021, USA*

The current paradigm of eukaryotic evolution is based primarily on comparative analysis of ribosomal RNA sequences. It shows several early-emerging lineages, mostly amitochondriate, which might be living relics of a progressive assembly of the eukaryotic cell. However, the analysis of slow-evolving positions, carried out with the newly developed slow-fast method, reveals that these lineages are, in terms of nucleotide substitution, fast-evolving ones, misplaced at the base of the tree by a long branch attraction artefact. Since the fast-evolving groups are not always the same, depending on which macromolecule is used as a marker, this explains most of the observed incongruent phylogenies. The current paradigm of eukaryotic evolution thus has to be seriously re-examined as the eukaryotic phylogeny is presently best summarized by a multifurcation. This is consistent with the Big Bang hypothesis that all extant eukaryotic lineages are the result of multiple cladogeneses within a relatively brief period, although insufficiency of data is also a possible explanation for the lack of resolution. For further resolution, rare evolutionary events such as shared insertions and/or deletions or gene fusions might be helpful.

Keywords: eukaryotes; molecular phylogeny; long branch attraction artefact;
early-branching organisms; loss of function

1. INTRODUCTION

Due to its ubiquity and great evolutionary conservation, ribosomal RNA (rRNA) was amongst the first molecules used to infer the early evolution of life. The analysis of rRNA quickly led to recognition of the evolutionary distinctness of the Archaeobacteria (Woese & Fox 1977). This major achievement, together with the easy sequencing of rRNA, led to its extensive use in large-scale phylogenies. A phylogenetic framework of eukaryotic organisms emerged rapidly (Sogin *et al.* 1989) and was strengthened by further sequencing of several hundred species (Sogin & Silberman 1998). The biological validation of the rRNA tree was its ability to recover monophyletic groups supported by numerous morphological characters (e.g. red algae or ciliates), by only a few (e.g. Euglenozoa) or by single ones (e.g. stramenopiles or alveolates) at different taxonomic levels. Moreover, several taxonomically enigmatic protists have been classified through rRNA analysis, e.g. *Pneumocystis* as a fungus and *Blastocystis* as a stramenopile (Sogin & Silberman 1998). All these data seem to confirm the reliability of rRNA as a phylogenetic marker in studies of the evolution of eukaryotes.

Three amitochondriate lineages, diplomonads (*Giardia* being the best-known representative), microsporidia and trichomonads, are the first offshoots in the rRNA tree, with an unresolved relative branching order (Leipe *et al.* 1993). Several mitochondriate taxa (*Physarum*, *Dictyostelium*, Euglenozoa such as *Trypanosoma* and Percolozoa such as *Naegleria*) and amitochondriate ones (*Entamoeba* and *Mastigamoeba*) emerge stepwise after these three lineages

and before a vast multifurcation. This multifurcation, named the crown (Knoll 1992), is generally interpreted as evidence for a major evolutionary radiation during which fungi, animals, red and green algae, ciliates and many other groups of organisms appeared and rapidly diversified. This phylogenetic framework was viewed as reflecting a step-by-step construction of the eukaryotic cell (Cavalier-Smith 1987; Patterson & Sogin 1992), with some extant protists corresponding to lineages that derived from hypothetical intermediate forms. For example, diplomonads, microsporidia and trichomonads were considered relics of eukaryotes that never possessed mitochondria. This scenario attracted many adherents and has found its way into textbooks.

However, the heterogeneity of the G + C content of rRNA (35–70%) brought the first criticisms because it can lead to tree reconstruction artefacts (Hasegawa & Hashimoto 1993). For example, the early emergence of *Dictyostelium* was attributed to the low G + C content of its rRNA (Loomis & Smith 1990). Yet, the overall picture of the rRNA tree remained unchanged when only species with similar G + C contents were included (Leipe *et al.* 1993) or when methods of phylogenetic inference insensitive to this bias such as transversion parsimony (Woese *et al.* 1991) or log-Det distance (Lockhart *et al.* 1994) were applied. Indeed, such new methods seemed to confirm the earliest divergence of microsporidia (Galtier & Gouy 1995).

The phylogenetic position of this group turned out to be instructive. The tubulins, hsp70 and RNA polymerase phylogenies place microsporidia close to the fungi, in agreement with phenotypic characters such as the presence of chitin in the endospore wall and similarities of their cell cycles (see Hirt *et al.* (1999) for a recent

*Author for correspondence (herve.philippe@bc4.u-psud.fr).

Figure 1. Incongruities between eukaryotic phylogenies. The trees were obtained by the ML method on unambiguously aligned positions (875 nucleotides for SSU rRNA, 359 amino acids for actin and 398 amino acids for α -tubulin) and rooted on Archaeobacteria, centractin (CEN) and β -tubulin (β -TUB), respectively. In order to highlight incongruities between the markers, early-emerging taxa of the rRNA tree are in italics and marked with a solid square in all trees. Taxa from the crown are marked with an open triangle. The scale bar corresponds to a substitution rate of 5% per site.

analysis). Moreover, microsporidia are now believed to have once contained mitochondria, since they harbour bona fide mitochondrial genes in their nuclei. Similar findings were made also on diplomonads and trichomonads, whose lack of mitochondria has been attributed to secondary loss or conversion to hydrogenosome, respectively (Gray *et al.* 1999), though an acquisition by lateral gene transfer cannot be fully ruled out (Sogin 1997). Such findings have challenged the view that primitively amitochondriate eukaryotes ever existed, raising doubts as to the branching order at the base of the eukaryotic tree (Embley & Hirt 1998; Philippe & Adoutte 1998; Stiller & Hall 1999). Here we examine the phylogenetic properties of the small subunit (SSU) rRNA, actin, elongation factor EF-1 α and α -tubulin, which are commonly used to infer the branching order of eukaryotes.

2. MATERIAL AND METHODS

(a) Sequences

The SSU rRNA, EF-1 α , α -tubulin and actin sequences (retrieved from GenBank) were manually aligned using MUST (Philippe 1993). Only unambiguously aligned positions (875, 388, 359 and 398, respectively) were retained. Typically, two species representative of each eukaryotic phylum were selected for comparison of the rRNA, actin and tubulin trees, leading to 35, 32 and 31 taxa, respectively. For the slow-fast (S-F) analysis, taxonomic groups with at least four to five sequences were selected (when more were available, redundant and very divergent ones were discarded), giving 95, 32, 39 and 38 taxa, respectively. With regard to rRNA, only transversions were considered because they are less sensitive to the bias introduced by a heterogeneous G + C content (Woese *et al.* 1991), although the results were similar when both transitions and transversions were used (data not shown).

(b) Phylogenetic reconstruction

Maximum-likelihood (ML) analyses were run with the PROTML and NUCML v.2.3 programs (Adachi & Hasegawa 1996). Maximum-parsimony (MP) analyses were run with PAUP 3.1.1 software (Swofford 1993). The S-F method (Brinkmann & Philippe 1999) was applied according to the following protocol. The data set was split into monophyletic groups of equivalent size. An MP analysis was applied to each taxonomic group (four to five taxa) and the number of substitutions inferred for each position. This number was then summed over all the groups of the data set, giving an estimate of the variability of this position. Positions were then selected according to this estimated variability. For a threshold of i , the positions fell either into data set S_i (i.e. slow, an estimate equal to or lower than i changes) or data set F_i (i.e. fast, an estimate greater than i changes). This method does not pre-judge the relationships between the taxonomic groups and avoids the circularity of classical successive weighting (Farris 1969).

(c) Symmetry, among site rate variation and covarion model

A rooted dichotomic tree is perfectly symmetrical when, for each node, the two branches derived from this node contain the same number of species. In contrast, a tree is perfectly asymmetrical (forms a ladder) when, at each node, one branch contains a single species and the other all the remaining species (Heard 1992). The symmetry was estimated with a new index that is suitable for multifurcating trees (a detailed description can be found at <http://bufo.bc4.u-psud.fr/bigbang/main.htm>). The shape parameter of the gamma law (α) was estimated using PAMP from the PAML package (Yang 1997).

The relevance of the rate across-sites model was tested according to the method of Lopez *et al.* (1999). When considering the null hypothesis that the substitution rate of a given position is constant throughout time, its rejection means that a

Figure 2. Explaining the incongruities by the LBA artefact. When a distantly related outgroup is used, fast-evolving lineages are attracted by the long branch of the outgroup generating an artefactual asymmetrical base. The order of emergence will be correlated with the evolutionary rates. If the fast-evolving lineages are different for two markers (indicated by solid and dashed lines), the inferred trees will be markedly incongruent. Note that fast-evolving species will not necessarily display a long branch in the inferred tree (Philippe & Laurent 1998) rendering the detection of such artefacts difficult.

covariation model describes the data better. In a covariation model, the positions are allowed to switch between variable and invariable states (Fitch & Markowitz 1970; Tuffley & Steel 1998), which in turn allows the substitution rate of a given position to vary throughout time. Each individual position was tested as well as the complete data set.

The sequences alignments, all phylogenetic trees and supplementary information are available from <http://bufo.bc4.u-psud.fr/bigbang/main.htm>.

3. RESULTS AND DISCUSSION

(a) *Incongruities between phylogenies*

Many molecular markers are now available for most phyla. These markers include actin, α - and β -tubulins, hsp70, EF-1 α , cpn60, gapdh and RNA polymerase II. Yet their analysis has not provided a unifying picture as their inferred phylogenies are not congruent with the rRNA tree or with each other (for reviews, see Embley & Hirt 1998; Philippe & Adoutte 1998). This can be clearly seen when the rRNA phylogeny is compared with those of actin and α -tubulin (figure 1). Groups emerging first in the rRNA tree do not necessarily emerge first for actin and tubulin, as strikingly exemplified by the late branching of trichomonads, microsporidia and diplomonads in the α -tubulin tree. Conversely, the basal species for actin (ciliates and red algae) and α -tubulin (Mycetozoa) are not basal for rRNA.

Interestingly, despite these serious conflicts, the overall shapes of the three topologies are similar with a long branch leading to the outgroup, an asymmetrical base and a symmetrical top. This unexpected similarity is puzzling. We have previously proposed a model (Philippe & Adoutte 1998) in which the long branch attraction (LBA) artefact (Felsenstein 1978) generates an asymmetrical base in which the emergence order of the lineages is correlated with their evolutionary rates (figure 2). This model explains why similar tree shapes but highly incongruent phylogenies are obtained with different markers (figure 1). Computer simulations (H. Philippe, unpublished data) have shown that such an artefactual asymmetrical base is inferred using standard methods, assuming biologically meaningful differences of evolutionary rates (five- to 20-fold) (Ayala 1997; Pawlowski *et al.* 1997; Philippe 1997).

Rather than coping with the difficulty of finding an appropriate model of sequence evolution (Yang 1996; Sullivan & Swofford 1997; Goldman *et al.* 1998; but, see Philippe & Germot 2000), our analysis of the eukaryotic phylogeny was restricted to the most conserved positions, which are more likely to contain a reliable phylogenetic signal. If the model proposed in figure 2 is true, the main factors responsible for reconstruction artefacts are the fast-evolving positions, which carry a great deal of noise due to numerous multiple substitutions (mutational saturation). We therefore used the S–F method (Brinkmann & Philippe 1999), which increases the signal-to-noise ratio for internal branches by shortening the terminal branches. Starting from the complete data set, the method progressively removes the variable positions, eventually leading to a small matrix, which contains the most slowly evolving positions for which the LBA artefact will be reduced (figure 3). If the groups that branch early in the classical rRNA tree are correctly located, only the slow-evolving positions will have conserved the ancient phylogenetic signal. They should strengthen this early emergence while the fast-evolving positions should only weaken the signal. In contrast, if the model in figure 2 is correct, the misplaced early-emerging taxa should emerge much later as fast-evolving positions are removed, and subsequently they should have their branches lengthened (see figure 3 for a comparison of the two scenarios).

(b) *Re-examination of rRNA phylogeny*

The S–F method was applied to a large set of rRNA sequences. As shown in figure 4, the most conserved positions (matrices S_0 and S_1 ; zero or one ingroup substitution for the 19 groups) clearly demonstrate that (i) the outgroup (Archaeobacteria) is very distantly related, (ii) the early emergence of diplomonads, microsporidia and trichomonads is not supported at all, and (iii) these three groups evolve much faster than all other eukaryotes. Starting from the multifurcation, for matrix S_1 diplomonads and microsporidia have experienced *ca.* 20 substitutions and trichomonads *ca.* 12, whereas other eukaryotic groups such as ciliates, fungi, red algae and stramenopiles have undergone only zero or one substitution. When more variable characters are added (matrices S_2 – S_4), the pattern remains identical, but the trichomonads now emerge prior to the multifurcation, probably the first

Figure 3. LBA and the S-F method. The variability of each position in the data set can be estimated by the method described in Brinkmann & Philippe (1999). Then, positions that have a lower variability than a given threshold i are used in matrix S_i . When considering the topology inferred for the total data set, the S-F method allows distinction between two possible scenarios: (i) an actual early emergence of the bold taxa (left column), or (ii) an artefactual early emergence of the bold taxa, which are attracted towards the outgroup by a LBA artefact (right column). If scenario 1 is true, the slowest matrices should infer topologies like those described in the left column: the bold taxa should still emerge early, only with shorter branches. On the contrary, scenario 2 will result in bold taxa emerging later and displaying long branches for slow matrices like S_0 and they should emerge all the earlier as fast-evolving positions are added.

manifestation of LBA. An asymmetrical base (comprising the trichomonads, microsporidia and percolozoans) appears when positions exhibiting five changes are added. The usual asymmetrical base is recovered only when all positions are considered (Ts + Tv in figure 4).

The analysis of the slow-evolving positions of rRNA is thus fully congruent with scenario 2 in figure 3. Irrespective of their true position, fast-evolving lineages emerge at the base of the tree as a result of the LBA artefact, which is caused by the fast-evolving positions that blur the phylogenetic signal. Fast-evolving positions (more than three substitutions) account for as much as 90% of the total number of steps, which means that quantity overwhelms quality and, therefore, the reconstruction is prone to artefacts. While slow-evolving positions concentrate support for undisputed monophyletic groups (e.g. eukaryotes), they do not support the early emergence of any group. For example, enforcing the monophyly of fungi and microsporidia required 51 extra steps when all positions were considered, but none when only positions with less than two

substitutions were analysed. In conclusion, the general picture of the S trees is a wide range of evolutionary rates combined with a lack of resolution of interphyla relationships (see the vast multifurcations in figure 4).

(c) *Re-examination of protein phylogenies*

The model shown in figure 2 explains the asymmetrical base of the rRNA phylogeny and should also explain those of the phylogenies based on proteins such as EF-1 α , α -tubulin, RNA polymerase II or actin. The S-F method indeed gave the expected results when applied to the EF-1 α and α -tubulin sequences (figure 4). Analysis of the slow-evolving positions clearly reveals the high evolutionary rate of some lineages (diplomonads, euglenozoans and ciliates for EF-1 α and *Entamoeba*, *Dictyostelium*, diplomonads and microsporidia for α -tubulin). With the fast-evolving positions included, these phyla emerged at the base of the tree as a consequence of LBA artefacts leading, for example, to an artefactual paraphyly of ciliates in the EF-1 α tree (figure 4).

Figure 4. Impact of fast-evolving positions on the inferred trees. The trees were reconstructed using the MP method. The complete data sets contain 95, 32 and 39 sequences and 875, 388 and 411 positions for rRNA, EF-1 α and α -tubulin, respectively, and are indicated by Ts + Tv (both transitions and transversions) for rRNA and by C for proteins. Trees S_i were obtained only with positions undergoing less than i changes within the predefined groups (see Brinkmann & Philippe (1999) for a description of the S-F method). These trees are strict consensus trees of the often numerous, most parsimonious trees. Abbreviations are as in figure 1. The scale bars indicate the number of substitutions. For rRNA, trees S_2 and S_4 are virtually identical to tree S_3 (all trees are available at <http://bufo.bc4.u-psud.fr/bigbang/main.htm>). Taxa that emerge early in the topology inferred by the complete data set are indicated in bold.

However, with the exception of some mycetozoans, the α -tubulin data set exhibited much smaller differences in its evolutionary rates than the rRNA one. In congruence with our model, the asymmetrical base was also less pronounced. Interestingly, fungi, microsporidia and animals always grouped together except in the multifurcation obtained with the most slowly evolving positions (S_0). However, the statistical support for this clade always remained low. α -tubulin, although a good marker in terms of the small variations in its evolutionary rate, is of limited use because it contains few informative positions. Indeed, a marker is useful when it provides a great number of informative positions and behaves as much as possible like a molecular clock. In this respect, rRNA, with numerous informative positions, is a valuable phylogenetic marker, although its large rate variation, which violates a molecular clock behaviour, makes it unreliable for locating the fastest evolving species. rRNA remains useful for locating slow-evolving species and the topology of the crown of the rRNA tree can still be considered a backbone for eukaryotic phylogeny.

(d) *Limitations of molecular phylogenies*

At first glance it may appear shocking that probably none of the inferred phylogenies tell the true story. However, phylogenetic inference is a difficult task, even for more recent groups such as animals (Naylor & Brown 1998) or mammals (Philippe 1997). The difficulty in inferring eukaryotic phylogeny should be expected considering the long time-span involved, all the more so as the outgroup is usually very distant (figure 4). For example, in data sets containing positions with a single substitution (S_1), the numbers of substitutions in the branch leading to the outgroup were as high as 30 (EF-1 α), 40 (rRNA) or 65 (α -tubulin), whereas there were no substitutions (or few for tubulin) or in the internal branches that grouped separate eukaryotic phyla. Due to these huge distances, the outgroups behave like almost random sequences and their use is therefore fraught with hazards (Wheeler 1990). The misplacement of any lineage with an accelerated evolutionary rate is to be expected due to the LBA phenomenon. This artefact can be difficult to detect, because fast-evolving lineages do not always display very long branches when complete data sets are used (figure 4). In fact, for mutationally saturated sequences, fast- and slow-evolving species will exhibit similar distances to the outgroup, imitating a clock-like behaviour (Philippe & Laurent 1998). Only slow-evolving positions that are not saturated allow the identification of fast-evolving species when the outgroup is distant (figure 4).

For some markers, such as hsp70, the outgroup is closer (Budin & Philippe 1998). As expected, the asymmetrical base predicted by the model in figure 2 is reduced and the shape of the inferred eukaryotic phylogeny is more symmetrical (Germot & Philippe 1999). However, even with a close outgroup, the inference can be biased by an LBA artefact, as seen for mitochondrial rRNA for which many lineages exhibit independent accelerations. These fast-evolving species are attracted more by each other than by the outgroup and, thus, are grouped together (Philippe & Laurent 1998) in a meaningless way (Gray *et al.* 1999).

The limits of molecular phylogeny in inferring deep relationships are mainly due to (i) the existence of major differences in the evolutionary rates producing LBA artefacts, and (ii) the lack of an adequate model of sequence evolution (Philippe & Laurent 1998). Although very complex models have been implemented (Yang 1996; Sullivan & Swofford 1997; Goldman *et al.* 1998), they all assume that the evolutionary rate of a given position, though varying between positions, remains the same along the lineages (rates across-sites model). An alternative model states that only a fraction of the positions (the set of covarions) is free to vary at a given time and this fraction can change through time (Fitch & Markowitz 1970), rendering the evolutionary rate of a position variable over time. As shown for other markers (Miyamoto & Fitch 1995; Lockhart *et al.* 1998; Lopez *et al.* 1999), the covarion model describes the evolution of rRNA sequences better than rate across-sites models. Using the method of Lopez *et al.* (1999), 15% of the informative positions of the rRNA have a significantly heterogeneous evolutionary rate and the rate across-site model can be rejected at the 1% confidence level for the complete rRNA data set. The ML method was used in the hope of minimizing the effects of model violations because it is considered the most robust (Hasegawa & Fujiwara 1993). Yet, the trees in figure 1 are still plagued with LBA since ML is misled by the major violation of its assumptions (Lockhart *et al.* 1996; Cao *et al.* 1998; Philippe & Germot 2000). For example, taking into account among-site rate variation allowed locating microsporidia into the crown (though as a sister group to the alveolates instead of fungi), but left diplomonads and the nucleomorph at the base (Peyretailade *et al.* 1998). As the covarion model has so far not been implemented in any ML software, the S-F method, although less sophisticated, is probably more reliable because it considers the most reliable positions only. Indeed, trees inferred from slow-evolving matrices are not affected by LBA artefacts (figure 4).

The covarion model readily explains two puzzling observations. First, if we make the classical assumption that the evolutionary rates of the positions are distributed according to a gamma law, the shape parameter varies from 0.003 to 99 (data not shown) depending on the taxonomic groups. Second, the groups that exhibit many variable sites are not the most diverse, but are the fastest evolving ones. For example, microsporidia and diplomonads harbour more variable positions than such taxonomically diverse groups as fungi or stramenopiles (supplementary information is available at <http://bufo.bc4.u-psud.fr/bigbang/main.htm>). Evolutionary rates can increase through an augmentation of the substitution rate of the variable positions as generally thought, but also through an increase in the number of variable positions as demonstrated for hsp70 of *Giardia* and *Trichomonas* (Germot & Philippe 1999). This bias probably amplifies the LBA artefact and could explain the serious misplacements that occurred in the rRNA phylogeny.

(e) *Towards a resolution of the eukaryotic phylogeny*

The eukaryotic phylogenies inferred with diverse markers are mostly incongruent, which can be explained to a great extent by LBA artefacts. Interestingly, the only common feature of all the markers is their failure to resolve

the relationships between the eukaryotic phyla (Budin & Philippe 1998). Such a finding should not be unexpected. Efforts to develop a natural system of unicellular eukaryotes have encountered insurmountable difficulties. While an increasing number of independent lineages have been identified and easily delimited from each other on the basis of ultrastructural information, all attempts to establish their higher-order relationships have failed because no morphological synapomorphies were recognized in relating individual lineages (Patterson & Sogin 1992). The best summary of eukaryotic phylogeny is for now a soft polytomy including the major phyla. We have proposed the Big Bang hypothesis, i.e. a radiation event, to account for explaining the lack of resolution of both morphological and molecular data (Philippe & Adoutte 1998). Biologically speaking, this hypothesis assumes that the speciation events that gave birth to the phyla occurred in a short period of time, but 'short' has to be qualified to avoid misinterpretation. The resolving power of a phylogenetic marker mainly depends on the number of synapomorphies, which is very roughly proportional to the time elapsed between the speciation that gave birth to a group and the first speciation event within that group. Since the resolving power of molecular phylogeny is rather limited (Philippe *et al.* 1994), we assume that the duration of the eukaryotic radiation could be some tens of million years (Myr) up to 100 Myr, although many more data are required to clarify this point.

One way of increasing the resolving power is to use several genes simultaneously. For example, an analysis of a fusion of 13 nuclear markers (5171 amino acids) produced strong support for the sister-group relationship of red algae and green plants, a question that remained unresolved when single genes were used (Moreira *et al.* 2000). The rapid accumulation of sequence data should soon allow such analyses for locating many eukaryotic phyla. Yet, they can be misleading if the methods of tree reconstruction used are inconsistent, particularly because only limited species sampling will be used.

As another way of circumventing this limited resolving power, we suggest the use of rare events such as shared insertions and/or deletions or gene fusions. Such approaches have already proven useful, for instance the insertion of *ca.* 11 amino acids in the EF-1 α shared by fungi and animals (Baldauf & Palmer 1993) and also by microsporidia (Kamaishi *et al.* 1996). The fusion in the mitochondrial genome of *cox-1* and *cox-2*, which has only been observed in *Acanthamoeba* and *Dictyostelium* (Gray *et al.* 1999), constitutes a synapomorphy for the Mycetozoa, which has so far only been supported by phylogenies based on EF-1 α (Baldauf & Doolittle 1997) and actin (Philippe & Adoutte 1998). More interestingly, the fused dihydrofolate reductase and thymidylate synthase genes in alveolates, plants and Euglenozoa suggest that the fungi (+ microsporidia)/animals group could be among the first emerging lineages in the eukaryotic phylogeny, because these genes are also independent in prokaryotes.

Increases in evolutionary rates have been shown to obscure the phylogenetic signal when they occur after the emergence of a group (figure 4), but they can also enhance it when they occur between their emergence and diversification. In such a case, more synapomorphies can

accumulate and the internal branch will be longer, even if the time elapsed is rather short. This explains the strong support for the fungi (+ microsporidia)/animals group observed with α -tubulin and not with other markers (figure 4). Similarly, Bilateria display a relatively long branch in the rRNA tree (figure 4), particularly when compared with diploblasts. This acceleration probably explains why early rRNA phylogenies, which used fewer sequences and less efficient reconstruction methods did not recover the monophyly of animals (Field *et al.* 1988). Due to this acceleration, rRNA is the only known marker that robustly recovers the monophyly of the Bilateria (H. Philippe and P. Lopez, unpublished results). The hope of resolving the eukaryotic tree might lie in such accelerations, provided they have happened in a sufficient number of markers.

4. CONCLUSION

Our main finding is that all the early-emerging taxa of the rRNA and protein trees are misplaced by an LBA phenomenon and that all extant eukaryotic phyla become parts of the crown. Radiation can account for the lack of resolution, but the term 'radiation' is somewhat imprecise. Thus, the next necessary step is to determine the time-span for the diversification of eukaryotes in order to determine whether the polytomy observed reflects a Big Bang (Philippe & Adoutte 1998) or simply the limited amount of available data (Embley & Hirt 1998).

However attractive the idea might be that certain organisms could be regarded as living relics of a step-by-step assembly of the eukaryotic cell, it is unlikely that any contemporary eukaryote is similar in phenotype to the common ancestor of extant eukaryotes. This does not mean that the eukaryotic cell was miraculously assembled without any intermediate forms. The extinction of those intermediate forms so that a very long branch appears at the base of extant eukaryotes is a reasonable hypothesis, although hard to test. This phenomenon is common considering, for example, the well-documented cases of mammals, birds or angiosperms, for which intermediate forms are known mostly from the fossil record. Mammals diverged from other amniotes *ca.* 300 Myr ago and their first speciation leading to a still living group, the monotremes, occurred only *ca.* 120 Myr ago, yielding an unbroken branch spanning *ca.* 180 Myr, i.e. longer than the time in which extant mammals evolved. The lack of extant intermediate forms (mammalian reptiles in this example) renders the reconstruction of early evolutionary events very difficult. Considering the paucity of the early fossil record of eukaryotes before 1 Gyr (Knoll 1992), the question of their early evolution will be very difficult to answer.

Although the absence of some functions and structures (flagella, spliceosomal introns, peroxisomes, mitochondria, Golgi apparatus, etc.) in certain eukaryotic groups was often interpreted as evidence of a stepwise emergence of the complex eukaryotic cell, they probably more often resulted from secondary losses. This is not unexpected for eukaryotes, which have a parasitic way of life and, more importantly, this means that the last common ancestor of eukaryotes was more complex than previously thought. Loss of cellular functions as a major evolutionary

mechanism indeed has been repeatedly invoked, perhaps most eloquently by Lwoff (1943) more than half a century ago.

Though loss is an important evolutionary phenomenon, some events leading to increased complexity clearly occurred during the evolution of eukaryotes. To evaluate the level of complexity of the last common ancestor of extant eukaryotes, we will have to reassess all the data (palaeontological, biogeochemical, morphological, etc.) in the light of the revised phylogeny. The rapid progress of numerous genome projects will also provide a wealth of data, which should contribute significantly to this evaluation. Moreover, the phylogenetic analysis of the flood of new sequences together with improved analytical methods will certainly lead to major breakthroughs in the resolution of the phylogeny of eukaryotes.

We thank Nick Butterfield, Masami Hasegawa, Pete Lockhart, Geoff McFadden and Bill Martin for critical reading of the manuscript. We are grateful to Noëlle Narradon for technical help. This work was supported by the Centre National de la Recherche Scientifique, a grant from the Groupement de Recherches et d'Études sur les Génomes (décision d'aide no. 94/125) and by the ACC-SV7. M.M. was supported by United States National Institutes of Health grant AI 11942 and United States National Science Foundation grant DEB-9615659.

REFERENCES

- Adachi, J. & Hasegawa, M. 1996 MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr.* **28**, 1–150.
- Ayala, F. J. 1997 Vagaries of the molecular clock. *Proc. Natl Acad. Sci. USA* **94**, 7776–7783.
- Baldauf, S. L. & Doolittle, W. F. 1997 Origin and evolution of the slime molds (Mycetozoa). *Proc. Natl Acad. Sci. USA* **94**, 12 007–12 012.
- Baldauf, S. L. & Palmer, J. D. 1993 Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl Acad. Sci. USA* **90**, 11 558–11 562.
- Brinkmann, H. & Philippe, H. 1999 Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Mol. Biol. Evol.* **16**, 817–825.
- Budin, K. & Philippe, H. 1998 New insights into the phylogeny of eukaryotes based on ciliate hsp70 sequences. *Mol. Biol. Evol.* **15**, 943–956.
- Cao, Y., Waddell, P. J., Okada, N. & Hasegawa, M. 1998 The complete mitochondrial DNA sequence of the shark *Mustelus manazo*: evaluating rooting contradictions to living bony vertebrates. *Mol. Biol. Evol.* **15**, 1637–1646.
- Cavalier-Smith, T. 1987 The origin of eukaryotic and archaeobacterial cells. *Ann. NY Acad. Sci.* **503**, 17–54.
- Embley, T. M. & Hirt, R. P. 1998 Early branching eukaryotes? *Curr. Opin. Genet. Devel.* **8**, 624–629.
- Farris, J. 1969 A successive approximations approach to character weighting. *Syst. Zool.* **18**, 374–385.
- Felsenstein, J. 1978 Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.
- Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R. & Raff, R. A. 1988 Molecular phylogeny of the animal kingdom. *Science* **239**, 748–753.
- Fitch, W. M. & Markowitz, E. 1970 An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593.
- Galtier, N. & Gouy, M. 1995 Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl Acad. Sci. USA* **92**, 11 317–11 321.
- Germot, A. & Philippe, H. 1999 Critical analysis of eukaryotic phylogeny: a case study based on the hsp70 family. *J. Eukaryot. Microbiol.* **46**, 116–124.
- Goldman, N., Thorne, J. L. & Jones, D. T. 1998 Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458.
- Gray, M. W., Burger, G. & Lang, B. F. 1999 Mitochondrial evolution. *Science* **283**, 1476–1481.
- Hasegawa, M. & Fujiwara, M. 1993 Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.* **2**, 1–5.
- Hasegawa, M. & Hashimoto, T. 1993 Ribosomal RNA trees misleading? *Nature* **361**, 23.
- Heard, S. 1992 Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* **46**, 1818–1826.
- Hirt, R. P., Logsdon Jr, J. M., Healy, B., Dorey, M. W., Doolittle, W. F. & Embley, T. M. 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.
- Kamaishi, T., Hashimoto, T., Nakamura, Y., Masuda, Y., Nakamura, F., Okamoto, K., Shimizu, M. & Hasegawa, M. 1996 Complete nucleotide sequences of the genes encoding translation elongation factors 1 alpha and 2 from a microsporidian parasite, *Glugea plecoglossi*: implications for the deepest branching of eukaryotes. *J. Biochem. (Tokyo)* **120**, 1095–1103.
- Knoll, A. H. 1992 The early evolution of eukaryotes: a geological perspective. *Science* **256**, 622–627.
- Leipe, D. D., Gunderson, J. H., Nerad, T. A. & Sogin, M. L. 1993 Small subunit ribosomal RNA⁺ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. *Mol. Biochem. Parasitol.* **59**, 41–48.
- Lockhart, P., Steel, M., Hendy, M. & Penny, D. 1994 Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.* **11**, 605–612.
- Lockhart, P. J., Larkum, A. W., Steel, M., Waddell, P. J. & Penny, D. 1996 Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. *Proc. Natl Acad. Sci. USA* **93**, 1930–1934.
- Lockhart, P. J., Steel, M. A., Barbrook, A. C., Huson, D., Charleston, M. A. & Howe, C. J. 1998 A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol. Biol. Evol.* **15**, 1183–1188.
- Loomis, W. F. & Smith, D. W. 1990 Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc. Natl Acad. Sci. USA* **87**, 9093–9097.
- Lopez, P., Forterre, P. & Philippe, H. 1999 The root of the tree of life in the light of the covarion model. *J. Mol. Evol.* **49**, 496–508.
- Lwoff, A. 1943 *L'Évolution physiologique. Étude des pertes de fonctions chez les microorganismes*. Paris: Hermann et Cie.
- Miyamoto, M. M. & Fitch, W. M. 1995 Testing the covarion hypothesis of molecular evolution. *Mol. Biol. Evol.* **12**, 503–513.
- Moreira, D., Le Guyader, H. & Philippe, H. 2000 The origin of red algae: implications for the evolution of chloroplasts. *Nature*. (In the press.)
- Naylor, G. J. P. & Brown, W. M. 1998 Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47**, 61–76.
- Patterson, D. & Sogin, M. 1992 Eukaryote origins and protistan diversity. In *The origin and evolution of the cell* (ed. H. Hartman & K. Matsuno), pp. 13–46. Singapore: World Scientific.
- Pawlowski, J., Bolivar, I., Fahrni, J. F., De Vargas, C., Gouy, M. & Zaninetti, L. 1997 Extreme differences in rates of molecular evolution of foraminifera revealed by comparison of ribosomal

- DNA sequences and the fossil record. *Mol. Biol. Evol.* **14**, 498–505.
- Peyretailade, E., Biderre, C., Peyret, P., Duffieux, F., Metenier, G., Gouy, M., Michot, B. & Vivares, C. P. 1998 Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. *Nucl. Acids Res.* **26**, 3513–3520.
- Philippe, H. 1993 MUST, a computer package of management utilities for sequences and trees. *Nucl. Acids Res.* **21**, 5264–5272.
- Philippe, H. 1997 Rodent monophyly: pitfalls of molecular phylogenies. *J. Mol. Evol.* **45**, 712–715.
- Philippe, H. & Adoutte, A. 1998 The molecular phylogeny of Eukaryota: solid facts and uncertainties. In *Evolutionary relationships among Protozoa* (ed. G. Coombs, K. Vickerman, M. Sleigh & A. Warren), pp. 25–56. London: Chapman & Hall.
- 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. & Laurent, J. 1998 How good are deep phylogenetic trees? *Curr. Opin. Genet. Devel.* **8**, 616–623.
- Philippe, H., Chenuil, A. & Adoutte, A. 1994 Can the Cambrian explosion be inferred through molecular phylogeny? *Development* **120**(Suppl.), 15–25.
- Sogin, M. 1997 History assignment: when was the mitochondrion founded? *Curr. Opin. Genet. Devel.* **7**, 792–799.
- Sogin, M. L. & Silberman, J. D. 1998 Evolution of the protists and protistan parasites from the perspective of molecular systematics. *Int. J. Parasitol.* **28**, 11–20.
- Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. 1989 Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from *Giardia lamblia*. *Science* **243**, 75–77.
- Stiller, J. & Hall, B. 1999 Long-branch attraction and the rDNA model of early eukaryotic evolution. *Mol. Biol. Evol.* **16**, 1270–1279.
- Sullivan, J. & Swofford, D. L. 1997 Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mamm. Evol.* **4**, 77–86.
- Swofford, D. L. 1993 *PAUP: phylogenetic analysis using parsimony, version 3.1.1*. Champaign, IL: Illinois Natural History Survey.
- Tuffley, C. & Steel, M. 1998 Modeling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* **147**, 63–91.
- Wheeler, W. 1990 Nucleic acid sequence phylogeny and random outgroups. *Cladistics* **6**, 363–367.
- Woese, C. R. & Fox, G. E. 1977 Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl Acad. Sci. USA* **74**, 5088–5090.
- Woese, C., Achenbach, L., Rouviere, P. & Mandelco, L. 1991 Archaeal phylogeny: reexamination of the phylogenetic position of *Archaeoglobus fulgidus* in light of certain composition-induced artifacts. *Syst. Appl. Microbiol.* **14**, 364–371.
- Yang, Z. 1996 Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367–370.
- Yang, Z. 1997 *Phylogenetic analysis by maximum likelihood (PAML), version 1.3*. Berkeley, CA: Department of Integrative Biology, University of California at Berkeley.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.