New insights into the phylogenetic position of diplonemids: G+C content bias, differences of evolutionary rate and a new environmental sequence

David Moreira,1,2 Purificación López-García1,3 and Francisco Rodríguez-Valera1

Author for correspondence: David Moreira. Tel: +33 1 44 27 32 53. Fax: +33 1 44 27 34 45. e-mail: david.moreira@snv.jussieu.fr

The phylum Euglenozoa consists of three distinct groups: the euglenoids, diplonemids and kinetoplastids. The phylogenetic position of the diplonemids within this phylum remains unsettled, since both morphological and molecular data produce weak and contradictory results. It is shown here that taxonomic sampling, G+C content bias, mutational saturation and differences of evolutionary rate among lineages are major factors affecting the topology of the small-subunit rRNA euglenozoan tree. When these problems are minimized by using a larger diplonemid sampling (including a sequence of environmental origin) and correcting for G+C bias (by using both paralinear distances or an unbiased dataset), a diplonemids+euglenoids sisterhood is retrieved. Bootstrap support for this relationship is still moderate, but it is retrieved by all analysis methods, overcoming previously reported disagreements. In addition, the inclusion of a large number of euglenoid sequences in the analysis improves some phylogenetic relationships within this group. Some problematic taxa, such as the species Khawkinea quartana, are now placed with high bootstrap support and monophyly is found for two interesting groups (the photosynthetic genera Eutreptia and Eutreptiella and the loricate genera Strombomonas and Trachelomonas), although with weak statistical support.

**Keywords:** Diplonema, Euglenozoa, molecular phylogeny, environmental sequences, G+C content bias

INTRODUCTION

The diplonemids (containing the genera Diplonema and Rhynchopus) constitute one of the three main groups within the phylum Euglenozoa, together with the euglenoids and kinetoplastids (Cavalier-Smith, 1993; Simpson, 1997). Although the monophyly of the Euglenozoa is well established by both structural and molecular data, the relationships among these three groups are far from being resolved. All three groups are clearly distinct, but the diplonemids share a number of characters with both the euglenoids and kinetoplastids, making it very difficult to assess their phylogenetic position. Both their proximity to the euglenoids (Kivic & Walne, 1984; Willey *et al.*, 1988) and their status as an independent branch (Triemer & Farmer, 1991) have been posited on the basis of morphological data. However, morphological features appear to be insufficient to bear out either of these possibilities.

Recently, molecular data have been used to address this problem. Sequences for only two markers, the small-subunit (SSU) rRNA and cytochrome-c oxidase subunit I (COI), are available for the three euglenozoan groups. For both markers, maximum-likelihood (ML) analysis groups the diplonemids and kinetoplastids whereas, in contrast, maximum-par-
simony (MP) and distance (NJ) analyses favour the association of the diplonemids with the euglenoids (Maslov et al., 1999). In all cases, statistical support for the different trees was very weak [bootstrap proportions (BP) < 50% for the nodes concerning the relationships among the three lineages and no significant differences between the alternative tree topologies]. This uncertainty probably comes from a number of factors that affect the particular topology of the euglenozoan tree (see Fig. 1). Firstly, taxonomic sampling for the three lineages is unbalanced, especially for the SSU rRNA. Taxonomic sampling is a major factor affecting the reliability of molecular phylogenies (Lecointre et al., 1993). In the case of euglenozoans, the euglenoids are very well sampled, with a large number of sequences covering a wide diversity of this group. However, kinetoplastid sequences, although abundant, seem to be less diverse in terms of genetic distance, which results in a very long unbroken branch at the base of this group. Such a long unbroken basal branch also occurs in the diplonemids, for which, in addition, only two SSU rRNA sequences are available. Secondly, there are important differences of G+C content among the three lineages, notably higher in euglenoids than in diplonemids and kinetoplastids, which can induce the artificial grouping of clades with similar G+C contents (see below). Finally, euglenozoans, in particular the kinetoplastids and some euglenoids, are suspected to be rapidly evolving organisms (Stiller & Hall, 1999). These differences in sequence richness, diversity, G+C content and evolutionary rate among the three groups may induce phylogeny reconstruction artefacts, such as the long branch attraction (LBA) (Felsenstein, 1978), which could explain the instability of the euglenozoan tree.

It is well known that phylogenetic reconstruction can be improved by adding diverse sequences to alleviate problems due to LBA artefacts (Hendy & Penny, 1989; Moreira et al., 1999). In this work, we have re-examined the conflicting phylogenetic position of the diplonemids, improving both the analyses to minimize sources of error (in particular G+C content bias) and the diplonemid taxonomic sampling. To enlarge the taxonomic sampling, we propose a new strategy based on the use of ‘environmental sequences’ (i.e. sequences amplified directly from the environment). This strategy has been extremely useful in the analysis of the diversity and phylogeny of prokaryotes (Pace, 1997), but it has only very recently been applied to the study of eukaryotes, including protists (López-García et al., 2001; Moon-van der Staay et al., 2001). Diplonemids appear to be common inhabitants of deep-sea regions, in particular sediment ecosystems (Larsen & Patterson, 1990). Recently, we carried out a survey of the protist diversity existing at 3000 m depth in Antarctic waters using molecular methods (amplification and sequencing of SSU rRNA genes) and we found a sequence that branched close to the diplonemids (López-García et al., 2001). Phylogenetic analyses show that this sequence emerges before the two other available Diplonema sequences, therefore being useful to break the long diplonemid branch. Using this sequence and different strategies to minimize LBA and G+C content bias, we have obtained congruent results with all phylogenetic analysis methods (NJ, MP and ML), which yield trees where the diplonemids always appear as a sister-group of the euglenoids. This relationship was still supported by moderate bootstrap values (but higher than those reported by Maslov et al., 1999) and the preferred topologies were not significantly better than the alternatives (diplonemids + kinetoplastids or euglenoids + kinetoplastids as sister-groups). However, since all methods give similar results and our analyses suggest that artefacts such as G+C content bias or LBA are not responsible of the preferred topology, we favour the phylogenetic position of the diplonemids as a sister-group of the euglenoids.

**METHODS**

**Amplification of diplonemid SSU rRNA sequences from deep-sea samples.** In order to characterize the diversity of planktonic eukaryotes in the deep ocean, a volume of 20 l seawater from a depth of 3000 m was prefiltered through a nylon mesh and filtered through a filter (5 µm pore size) and the remaining plankton was collected in 0-2 µm Sterivex filters. After a proteinase K/SDS lysis step, nucleic acids were extracted from the 0-2 µm filter as described previously (Massana et al., 1997). SSU rRNA genes were amplified by PCR using the specific primers EK-1F (CTGGTTGATCGTGCCAC) and EK-1520R (CYGCGAGTCACCTAC) under conditions described previously (DeLong, 1992). rDNA clone libraries were constructed using the Topo TA Cloning system (Invitrogen). After plating, positive transformants were screened by PCR amplification of inserts using flanking vector primers. Amplicons of the expected size were subsequently purified using the QIAquick PCR purification system (Qiagen). Purified PCR products were partially sequenced directly in an ABI Prism 377 apparatus (Perkin Elmer ABI) using the ABI Prism dRhodamine terminator cycle sequencing ready reaction kit with primer EK-1F. All clones showed identical partial sequences and one of them (clone DH148-EKB1) was chosen for complete sequencing. The insert was sequenced twice using both the flanking vector primers and the two primers EK-555F (AGTCTGGTGCCAGCAGCCGC) and EK-1269R (AAGAACGGCCATGCACCAC), which were designed to complete and overlap the central insert sequence. The sequence produced was 2001 nucleotides long.

**Phylogenetic analyses.** Euglenozoan SSU rRNA sequences were retrieved from GenBank and the rRNA Database at the University of Antwerp (http://rRNA.uaa.be/). They were aligned together with the Antarctic DH148-EKB1 clone sequence using CLUSTAL W (Thompson et al., 1994) and the resulting multiple alignment was edited manually using the program ed from the MUST package (Philippe, 1993). Gaps and ambiguously aligned positions were excluded from the phylogenetic analyses and from the G+C content calculations. NJ, MP and ML trees were respectively constructed with the programs NJ from the MUST package (Philippe, 1993), PAUP 3.1 (Swofford, 1993) and NUCL from the MOLPHY 2.3 package (Adachi & Hasegawa, 1996). BP
Table 1. G+C contents (mol%) for the datasets studied in this work

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Diplonemids</th>
<th>Euglenoids</th>
<th>Kinetoplastids</th>
<th>Outgroup</th>
<th>Max. ΔG+C*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maslov et al. (1999)</td>
<td>48.95</td>
<td>52.47</td>
<td>48.36</td>
<td>48.45</td>
<td>4.11</td>
</tr>
<tr>
<td>Low-G+C outgroup (Fig. 1a, b)</td>
<td>48.95</td>
<td>52.47</td>
<td>48.36</td>
<td>43.29</td>
<td>9.18</td>
</tr>
<tr>
<td>High-G+C outgroup (Fig. 1c, d)</td>
<td>48.95</td>
<td>52.47</td>
<td>48.36</td>
<td>51.05</td>
<td>4.11</td>
</tr>
<tr>
<td>Homogeneous G+C (Fig. 3a)</td>
<td>48.17</td>
<td>49.75</td>
<td>48.64</td>
<td>49.12</td>
<td>1.58</td>
</tr>
</tbody>
</table>

* Maximum difference of G+C content among groups.

---

**Fig. 1.** ML and LogDet phylogenetic trees for euglenozoan SSU rRNA sequences constructed using low-G+C (a, b) or high-G+C (c, d) outgroup sequences. Numbers at nodes are bootstrap values. For the ML trees, ML (roman), NJ (italic) and MP (bold) BP are indicated for the node concerning the position of the diplonemids. A total of 1304 unambiguously aligned positions was used. Bars, 5 substitutions per 100 positions for a unit branch length.

---

were estimated by using 1000 replicates for the NJ and MP trees and by using the RELL method (Kishino et al., 1990) on the 2000 top-ranking trees for ML trees. NJ trees applying paralinear distances were constructed and bootstrapped using the DAMBE package (Xia, 2000). Saturation diagrams were constructed using the program COMP_MAT from the...
MUST package (Philippe, 1993). Alignments, trees and a list of species used are available upon request.

RESULTS
We have analysed two different classes of datasets to test the possible impact of G+C content bias and taxonomic sampling on the phylogenetic position of the diplonemids. Firstly, we have reanalysed the dataset used previously by Maslov et al. (1999) using outgroup sequences with low or high G+C content. Secondly, we have analysed a smaller dataset containing sequences selected to minimize the differences of G+C content among lineages. In addition, we have analysed a dataset with a larger taxonomic sampling of euglenoids to study some aspects of the internal phylogeny of this group.

G+C content bias and the position of the diplonemids
A potential source of phylogenetic error exists when different clades have sequences with significantly different G+C contents (Embly et al., 1992; Hasegawa & Hashimoto, 1993). Such a dependence of tree topology on the composition of the outgroup sequences has been reported for different groups (Tarrio et al., 2000). This may indeed be the case for euglenozoans since, for the unambiguously aligned regions in the dataset published by Maslov et al. (1999), values range from 47-71 (Trypanoplasma borrelii) to 53-60 (Euglena gracilis) mol% G+C; that is, a maximum difference of 6-43 mol%. Moreover, euglenoids show a mean value of 52-47 mol% G+C, kinetoplastids 48-36 mol% G+C and diplonemids 48-95 mol% G+C (i.e. a maximum mean difference of 4-11 mol% G+C between kinetoplastids and euglenoids; Table 1). The mean value for the complete set of euglenozoans was 49-92 mol% G+C. To test the possible influence of the observed G+C content bias on the reconstruction of the relationships between these three groups, we constructed two different datasets. In the first dataset, low-G+C-content outgroup sequences were used (mean value of 43-29 mol%). The resulting ML tree showed the diplonemids and euglenoids as sister-groups (Fig. 1a). In the second dataset, high-G+C-content outgroup sequences were used (mean value of 51-05 mol%). In this case, the resulting ML tree showed the kinetoplastids and diplonemids as sister-groups (Fig. 1c). In both cases, NJ and MP trees were very similar to the respective ML trees (not shown). This discrepancy between the two datasets seems to reflect the G+C contents of the different groups. Thus, when using a low-G+C outgroup, the ingroup clade with the lowest G+C content, the kinetoplastids, appears to be attracted by the outgroup. Conversely, when using a high-G+C outgroup, it is the ingroup clade with the highest G+C content, the euglenoids, that appears to be attracted by the outgroup. In previous analyses of euglenozoan phylogeny (Maslov et al., 1999), the simultaneous use of outgroup sequences with low (Saccharomyces cerevisiae, 44-88 mol%) and high (Physarum polycephalum, 52-03 mol%) G+C content may have induced biases that are very difficult to predict.

We analysed this problem using a double strategy. Firstly, we reconstructed phylogenetic trees for the different datasets by applying paralinear distances, which are designed to cope with unequal G+C contents among lineages (Lake, 1994) (Fig. 1b, d). This had especially significant effects on the high-G+C-outgroup dataset, since the diplonemids now appeared as a sister-group of the euglenoids (Fig. 1d). This suggested strongly that G+C content is indeed an important factor in determining the euglenozoan phylogeny. However, this distance method may be very sensitive to other reconstruction problems, in particular mutational saturation and differences of evolutionary rate among lineages producing LBA artefacts. In fact, the euglenozoan dataset exhibits a severe degree of mutational saturation (Fig. 2). The saturation diagram (Philippe et al., 1994) shows two clouds of points; a first cloud close to the diagonal, which represents the ideal case of sequences with no more than one substitution per sequence position, and a second cloud clearly distant from the diagonal. The first corresponds to intra-group sequence comparisons, while the second corresponds mostly to inter-group sequence comparisons. This second cloud reveals a strong saturation between the different groups, which is not surprising when the long distances separating them are taken into account. These distances are even longer than those separating groups as distant as metazoans and red algae (see Fig. 3). Mutational saturation combined with differences of evolutionary rate among lineages make phylogeny reconstruction methods prone to artefacts such as the LBA.

Fig. 2. Saturation diagram for the euglenozoan SSU rRNA sequences shown in Fig. 1. The number of observed differences between pairs of sequences is shown on the y-axis. The number of substitutions between the same pairs of sequences inferred by ML is shown on the x-axis. The diagonal represents the ideal case of no more than one substitution per sequence position.
In order to minimize the effects of both G + C content bias and LBA, we applied an alternative strategy for the study of the relationships among the three euglenozoan groups, based on analysis by ML (the method less sensitive to LBA) of datasets with a G + C content as homogeneous as possible and similar numbers of sequences for the different groups. We analysed several datasets containing combinations of sequences that exhibit similar G + C contents. These different datasets yielded similar results. Fig. 3(a) shows the ML tree derived from a dataset containing the three available diplonemid sequences, three kinetoplastids (Leishmania major, Rhyochobodo sp. and Trypanoplasma borelli) and three euglenoids (Peranema trichophorum, Eutreptiella gymnastica and Eutreptiella sp. CCMP389). Mean G + C contents for this dataset were 48.17 mol% for diplonemids, 48.64 mol% for kinetoplastids and 49.75 mol% for euglenoids (i.e. a maximum difference of 1.58 mol% G + C, less than half the difference observed for the previous, larger dataset; Table 1). The mean G + C content for the complete set of euglenozoans was 48.85 mol%. To minimize bias further, we selected an outgroup with a similar and homogeneous G + C content (49-12 mol%), resulting in a dataset containing 13 sequences (Fig. 3a).

All three analysis methods, NJ, MP and ML, retrieved identical relationships among the lineages and, as evidence that the G + C bias was reduced effectively for this dataset, no different tree topologies were retrieved by applying paralinear distances (not shown). In all cases, the diplonemids emerged as the sister-group of the euglenoids. This sisterhood was also retrieved in an exhaustive ML analysis, carried out imposing constraints (to the ingroup nodes that showed a BP > 95% for all analyses) to give a manageable number of possible trees (Fig. 3). This congruence was remarkable and was difficult to attribute to G + C content bias, since the two sister-groups were those with the most different values. Nevertheless, statistical support remained moderate (BP of 88% for NJ, 82% for MP and 62% for ML). In this sense, the difference of likelihood between the preferred topology (diplonemids + euglenoids) and the two alternatives was not significant under a Kishino–Hasegawa test (Kishino & Hasegawa, 1989). However, it is interesting to note that rejection was stronger for the topology diplonemids + kinetoplastids $[\Delta \ln L = 1.01 \text{ standard error (se)}]$ than for the topology euglenoids + kinetoplastids ($\Delta \ln L = 0.25 \text{ se}$), which is in disagreement with previous analyses (Maslov et al., 1999).

In order to test the effect of the addition of the new diplonemid sequence DH148-EKB1, a dataset excluding this sequence was constructed. All reconstruction methods, including an exhaustive ML search (Fig. 3b), once again retrieved the sisterhood of diplonemids and euglenoids. However, support for this node was, with the exception of the NJ analysis, weaker than in the previous analysis: BP of 93% for NJ, 78% for MP and 49% for ML. In addition, rejection of alternative topologies decreased under the Kishino–Hasegawa test; $\Delta \ln L$ was only 0.06 se for the kinetoplastids + euglenoids topology and 0.98 se for the kinetoplastids + diplonemids topology. All this suggests a stabilizing, although not definitive, positive effect of the addition of this sequence upon the euglenozoan phylogeny.

Relationships within the euglenoids

In addition to the study of the phylogenetic relationships between the three euglenozoan lineages, we have taken advantage of the number of euglenoid sequences determined recently to analyse the internal phylogeny of this group. Despite promising results on the use of new ultrastructural traits, such as pellicle patterns or flagellar structure (Leander & Farmer, 2000; Linton & Triemer, 2001), recent articles have discussed the limitations of morphological data for the determination of evolutionary relationships within this group (Linton et al., 1999, 2000). For instance, phylogenetic analyses of 20 euglenoid SSU rRNA sequences suggested strongly that the genera Phacus and Lepocinclis are polyphyletic and are intermixed with Euglena species (Linton et al., 2000). The phylogenetic positions of certain species, such as the osmotroph Khawkinea quartana, remained uncertain, while some
important genera, such as *Strombomonas* and *Trachelomonas* within the order Euglenales, were not represented.

An updated dataset of euglenoid SSU rRNA sequences includes 34 sequences (Fig. 4). When very similar sequences were found in databases (e.g. *Euglena gracilis* and *Euglena sp.* UTEX364, with 99% identity), only one representative was included in our study in order to accelerate the very time-consuming ML analyses. The topology of the ML tree obtained from these 34 sequences was basically congruent with that presented by Linton *et al.* (2000), except for some significant differences. Thus, the support found for the emergence of *Khawkeinae quartana* at the base of a group comprising *Euglena gracilis, Astasia longa* and *Euglena anabaena* was notably increased from a BP of 62% to a current BP of 93%. Also noticeable was the change in the position of *Euglena anabaena*, which formerly emerged in a more basal position, preceded only by *Eutreptiella* sp., *Peranema trichophorum* and *Petalomonas cantuscyngni* (Linton *et al.*, 2000). Using this larger taxonomic sampling, this species branches with a BP of 70% close to a group composed of *Phacus pyrum, Phacus megalopsis, Phacus splendens* and *Leptocinclis ovata*. Therefore, its basal position was most likely due to an LBA artefact, as proposed previously (Linton *et al.*, 2000). This LBA problem has been attenuated by the addition of new sequences. *Euglena anabaena* was proposed to belong to the Catilliferae, together with *Euglena gracilis* and *Euglena agilis* (Pringsheim, 1956). However, its relatively well-supported distance from the other members of the Catilliferae makes this grouping uncertain. Therefore, the structural characteristics that promote this clade, i.e. shield-shaped chloroplasts containing a double pyrenoid and lens-shaped paramylon caps, likely arose several times within the euglenoids or were ancestral characters lost in the remaining species. Given the topology of the euglenoid tree, this latter hypothesis requires a large number of independent losses, so we favour the first possibility.

The ML tree shows two remarkable clades for several species not included in previous studies. Firstly, the monophyly of *Eutreptia* and *Eutreptiella* species is weakly supported (BP of 29%). However, the topology of the ML tree suggests that the genus *Eutreptiella* may be paraphyletic. More importantly, the species of the genus *Distigma (Distigma curvata* and *Distigma proteus*), also proposed to be members of the Eutreptiales (Leedale, 1967), branch far from the *Eutreptia* + *Eutreptiella* clade, with strong statistical support. They instead form a group with *Gyropaigne lefevrei* (BP of 100%), a member of the order Rhabdomonadales. These data challenge the proposal for an order Eutreptiales including the genera *Distigma, Distignumopsis, Eutreptia* and *Eutreptiella* (Leedale, 1967). Nevertheless, both *Distigma* species show very long branches (i.e. extremely fast evolutionary rates), so the possibility that their emergence earlier than the other Eutreptiales could be due to an LBA artefact cannot be discarded. A second interesting group encompasses the genera *Strombomonas* and *Trachelomonas*, also with weak support (BP of 31%). Both genera possess characteristic lorica, which are mineralized with ferrous and manganic compounds (Conforti *et al.*, 1994; Kudo, 1966). Lorica may therefore be a valuable phenotypic character that unifies the two genera.

**DISCUSSION**

Different problems work against the resolution of the phylogeny of the euglenozoans, which remains problematic, especially for the relationships among their three main clades, the euglenoids, kinetoplastids and diplonemids. These problems are biological (differences in evolutionary rate among species, different G+C contents) and technical (unequal taxonomic sampling for the different groups). Our analyses show that these problems are significant in shaping the euglenozoan tree. We have applied a novel approach to try to minimize sampling bias: the search for sequences obtained directly from environments of interest without previous isolation of organisms. Thus, we have obtained a new diplonemid sequence that, indeed, helps to break the long branch of this group. In addition to taxonomic sampling, G+C content also seems to be a very important source of uncertainty,
determining the order of emergence of clades according to the G + C content of the outgroup sequences employed (see Fig. 1). When this problem is corrected by using a less biased species sampling, we found support for a diplonemids + euglenoids sisterhood. Although bootstrap support for this relationship was moderate (88% for NJ, 82% for MP and 62% for ML), the incongruities between the different methods reported in previous analyses, which supported an alternative diplonemids + kinetoplastids sisterhood (Maslov et al., 1999), were not observed. Therefore, although the phylogenetic position of the diplonemids should still be considered an open question, the congruence of our different analyses makes us favour the sisterhood of the diplonemids and euglenoids.

The sisterhood of the diplonemids and euglenoids is in agreement with previous morphological observations (Simpson, 1997; Willey et al., 1988). Nevertheless, the alternative diplonemids + kinetoplastids sisterhood also seems to agree with some biological features that unify these two groups, such as the presence of the unusual base β-D-glucosyl hydroxymethyluracil (also called base J) in their genomes (van Leeuwen et al., 1998). However, this base has recently also been found in the euglenoid Euglena gracilis, so it appears to be a universal character among euglenozoans (Dooijes et al., 2000). More interesting is the occurrence of characteristic 39-nt 5′ mini-exon genes involved in trans-splicing in both kinetoplastids and diplonemids (Campbell et al., 1997). Euglenoids also process mRNAs by trans-splicing, but have a distinctive 22-nt 5′ mini-exon gene. However, since trans-splicing occurs in all euglenozoan groups, it is not an adequate character to elucidate inter-group relationships. In fact, trans-splicing was very likely already present in the common ancestor of this group. This means that, independent of the type of 5′ mini-exon gene present in that ancestor, at least one size change (from 22 to 39 nt or vice versa) should have occurred during the diversification of the three lineages of euglenozoans. Therefore, if the diplonemids + euglenoids sisterhood is correct, it implies that the ancestor of euglenozoans had 39-nt 5′ mini-exon genes, which were reduced to 22-nt 5′ mini-exon genes in euglenoids. The same picture would be deduced from a putative kinetoplastids + euglenoids sisterhood but, in this case, nothing could be said about the precise nature of the ancestral 5′ mini-exon genes. Finally, phylogenetic analysis of mitochondrial COI sequences was also found to support a kinetoplastids + euglenoids sisterhood, although with weak statistical support (Maslov et al., 1999). However, only single sequences are available for both euglenoids and diplonemids. As in the case of SSU rRNA, a larger dataset is necessary in order to obtain a more confident phylogeny from this marker.

In the case of the intra-group phylogeny of euglenoids, the main problems appear to stem from taxonomic sampling and differences of evolutionary rate among lineages. Both can be alleviated by the addition of new sequences (Hendy & Penny, 1989). In this work, we have analysed a large number of euglenoid sequences available in databases. The increase in taxonomic sampling appeared to improve the euglenoid phylogeny. Nodes that previously lacked good statistical support (e.g. the one of Khawkinea quartana) or branches likely affected by LBA problems (e.g. that of Euglena anabaenoides) turn out to be better supported in the tree when a larger taxonomic sample was used (Fig. 4). Moreover, interesting relationships are found, such as the monophyly of the genera Eutreptia + Eutreptiella and Strombomonas + Trachelomonas.

Despite this improvement, potential artefacts could still confuse the euglenoid phylogeny. One example is the very low support for the relationships among the groups in the apical part of the tree (BP between 12 and 70%). This may be due to insufficient information in the SSU rRNA sequence. However, an alternative explanation could be a rapid diversification of these groups, since radiation processes are usually reflected in this lack of resolution (Philippe & Adoutte, 1998). If this is the case for euglenoids, it means that their very diverse structural and adaptive traits evolved in a relatively reduced time-span. This kind of rapid diversification phenomenon seems to be common in various eukaryotic groups, such as the alveolates (López-García et al., 2001). The species Euglena mutabilis represents another problem. Its position at the base of the apical part of the euglenoid tree is well supported (BP of 99%). However, it seems to be a very fast-evolving species and the possibility that its early emergence is due to LBA, as in the case of the Distigma species discussed above, should not be excluded.

LBA may also affect the basal region of the euglenoid phylogeny, which is highly asymmetrical, in contrast to the highly symmetrical apical region. Asymmetrical (i.e. ladder-shaped) trees are often the result of the LBA artefact (Moreira et al., 1999; Philippe & Laurent, 1998; Philippe et al., 2000). The very fast-evolving sequence of Euglena mutabilis may be a clear example, although it is likely that this phenomenon affects all taxa in this basal region, since it is populated by long branches. The addition of new sequences from basal taxa, such as the orders Eutreptiales, Heteronematales, Sphenomonadales and Rhabdomonadales, will be necessary in order to ascertain the reliability of this part of the euglenoid tree.

ACKNOWLEDGEMENTS

We thank Visitación Conforti for information about loricate euglenoids. This work was supported by the European Commission MIDAS project.

REFERENCES


