THE EVOLUTIONARY IMPLICATIONS OF DIPLONEMIDS AND THEIR SPLICEOSOMAL INTRONS

by

QING QIAN

B. Sc., Hangzhou University, 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Botany

We accept this thesis as conforming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 2000

© Qing Qian, 2000

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Botany

The University of British Columbia
Vancouver, Canada

Date September 18, 2000

Abstract

The phylum Euglenozoa consists of three main groups: euglenoids, kinetoplastids and diplonemids (Simpson 1997). This phylum is unique in having three types of introns: nuclear trans-spliced "introns", nuclear conventional and "aberrant" introns. In order to determine the evolutionary history of the introns in this phylum, it is very important to know the general distributions of intron types within the phylum, and the likely phylogeny of the phylum.

The nuclear genomes of euglenoids are known to contain all three types of introns, while only trans-spliced and conventional introns have been found in kinetoplastids. However, nothing is known about diplonemid introns, and the phylogenetic placement of diplonemids within the Euglenozoa is uncertain. Therefore, I looked for nuclear introns in diplonemids by sequencing four nuclear protein-coding genes (actin, alpha-tubulin, beta-tubulin, and GAPDH) from different diplonemids. I found 11 introns in nine of the twenty-nine newly obtained diplonemid nuclear protein-coding genes. They all have conventional 5'-GT-AG-3' splicing sites, but differ from well-studied eukaryotic conventional introns (mammalian introns) in several details.

I have added these nuclear encoded sequences from diplonemids to the tubulin, actin and GAPDH alignments and then made global phylogenetic trees based on these protein alignments. The discrepancy between the tubulin trees and actin tree is whether the diplonemids are closer to kinetoplastids (tubulin trees) or euglenoids (actin tree).

Taken together, I postulate that the GT-AG conventional introns were present in the euglenozoan ancestor and were largely lost in kinetoplastids and euglenoids. The "aberrant" intron is very likely a derived character restricted to euglenoids. The trans-spliced discontinuous "intron" is an ancestral character to this phylum and it is highly likely that it will be found in diplonemids as well.

The phylogenetic position of the four newly sequenced diplonemid GAPDH sequences turned out to be very interesting. None of the four diplonemid GAPDH sequences branch with those of other euglenozoa. Instead, three of the four diplonemid-sequences branch with the gap3 of cyanobacteria with 100% bootstrap support, indicating a lateral gene transfer from bacteria to eukaryotes, and one GAPDH sequence branches in an uncertain position with other eukaryotic GAPDH sequences.

TABLE OF CONTENTS

List of Tables

List of Figures

Acknowledgements

I would like to thank everyone who inspired and helped me to produce this thesis. On the inspiration side there stand, first, both of my supervisors, Dr. Tom Cavalier-Smith and Dr. Patrick Keeling. I am very grateful to Dr. Tom Cavalier Smith for his research funding that allowed me to pursue this project, for his high critical standards, and also for his insightful guidance about phylogeny and evolution. After Tom left for a faculty position at Oxford University in England, Dr. Patrick Keeling became my main supervisor. I enjoyed working in his lab and trying new techniques. I thank Dr. Patrick Keeling for his kindness in sharing his knowledge with me, especially for the very complicated phylogenetic analysis and the GAPDH phylogeny, and also for his generosity to provide me with almost all the protein alignments and degenerate primers utilized in this project.

My committee members, Dr. Carl Douglas and Dr. Martin Adamson, kindly gave me comments and suggestions during the progress of my project.

On the help side, I would first like to thank Dr. Naomi Fast, for her careful reading of the manuscripts of my thesis, and her help in the preparation for the presentation of my project. I would also like to thank Dr. Ken Ishida and Juan Saldarriaga who gave me valuable suggestions on my thesis. During the writing and re-writing process, I learned a lot in presenting my thoughts more strongly and precisely. I owe special thanks to Ema Chao for generously providing me with the genomic DNAs from several diplonemids, and Dr. Alexandra Marinets for showing me the basic molecular lab techniques in the beginning.

I couldn't possibly have finished my thesis without the support from my dear parents, and various help from my friends, especially my roommate Tanya Hooker. Finally, my thanks go to my friend Jens Happe, for his persistent encouragement over the last two years.

CHAPTER I:
Introduction

1.1 The Phylum Euglenozoa and its phylogeny

Cavalier-Smith (1981) first formally established the phylum Euglenozoa by grouping kinetoplastids and euglenoids together based on a list of shared characteristics, including: mitochondria with discoid cristae; paraxial rods; non-tubular mastigonemes (flagellar hairs) and closed mitosis with an endonuclear spindle (Cavalier-Smith 1981). The first electron microscopic observations of diplonemids (Triemer et al. 1990) suggested the addition of diplonemids to the phylum Euglenozoa. Simpson (1997) further proposed two potential synapomorphies uniting the phylum Euglenozoa: flagellar root pattern and paraxial rod substructure. The unique pattern of flagellar root organization of the Euglenozoa is the system of three microtubular roots: two roots closely associated with the outside of each basal body and one originating between the basal bodies. In addition, the paraxial rods of kinetoplastids, euglenoids and diplonemids share a distinctive substructure: the paraxial rod of the dorsal/anterior flagellum has a cylindrical cross-sectional appearance, while the structure in the ventral/recurrent flagellum is squarer in cross-section with a three-dimensional latticework substructure.

Another new addition to the phylum Euglenozoa is *Postgaardi mariagerensis.* It is a recently described organism that is covered by rod-shaped bacteria, and that has two thickened flagella inserting into an anterior pocket. A recent ultrastructural study of *Postgaardi mariagerensis* (Simpson et al. 1996/97) revealed a strong case for its inclusion within the Euglenozoa because it also shares the two major synapomorphies proposed by Simpson for the Euglenozoa (Simpson 1997).

While the Euglenozoa share several synapomorphies, each euglenozoan group also displays distinct features of their own. The euglenoids as a subgroup are identified by the presence of a pellicle- a system of strips of glycoprotein that appears under the plasma membrane and is supported by sub-pellicular microtubules (Triemer et al. 1991b). This group includes both photosynthetic euglenoids (e.g. *Euglena*) and non-photosynthetic euglenoids (e.g. *Entosiphon*). The photosynthetic euglenoids have attracted the attention of many researchers because of their intriguing chloroplasts, which are surrounded by three membranes, instead of two membranes. It is now clear that their chloroplasts are of secondary endosymbiotic origin, which means that a colourless euglenoid acquired its chloroplast by swallowing a green algal cell (Gibbs 1978).

The kinetoplastids are the euglenozoans that harbor one or more kinetoplasts (DNA-rich bodies) in their mitochondria (Lee et al. 1985; Opperdoes 1987). This group includes major disease-causing genera. For example, the genera *Trypanosoma* and *Leishmania* include serious human pathogens that cause African 'sleeping sickness', South American Chagas disease, as well as leishmaniasis in tropical and subtropical areas (Lee et al. 1985; Opperdoes 1987). In addition, this group also includes free-living flagellates such as *Bodo* (Lee et al. 1985).

Diplonemids are represented by only two genera, *Diplonema* and *Rhynchopus*, based on their very similar ultrastructural organizations (Schnepf et al. 1994). They have neither kinetoplasts nor a pellicle of glycoprotein strips, but do possess a distinctive feeding apparatus composed of vanes with fuzzy coats and giant, flat mitochondrial cristae (Triemer et al. 1990; Triemer et al. 1991a; Triemer et al. 1991b; Simpson, 1997). Diplonemids do not have chloroplasts and they are not human pathogens, and they live in either fresh-water or marine environments (Schnepf et al. 1994; Triemer et al. 1990).

Postgaardi mariagerensis lacks an euglenoid pellicle and possesses mitochondria without kinetoplast or cristae. So, although being part of the phylum Euglenozoa, it is neither an euglenoid nor a kinetoplastid. As far as its feeding apparatus is concerned, *P. mariagerensis* has no vanes or supporting rods, but only the MTR (a complex of reinforcing microtubules) to support its feeding apparatus. This distinction indicates that *Postgaardi mariagerensis* is not a diplonemid either (Simpson et al. 1996/97; Simpson 1997).

In short, there are many data based on light- and electron-microscopy to distinguish among the four groups of the Euglenozoa. Data are particularly abundant for euglenoids and kinetoplastids. Although these structural characters are helpful in revealing phenotypic similatities, they are not as helpful for inferring phylogeny. Molecular sequences, on the other hand, are much more suitable for the latter task. However, they are not available at all from *Postgaardi mariagerensis* and extremely limited from diplonemids compared to euglenoids and kinetoplastids. In fact, no nuclear protein-coding gene has been characterized from diplonemids so far. The only available molecular sequences at the onset of this study were the sequences of small subunit ribosomal RNA (SSU rRNA) genes from two diplonemids (*Diplonema papillatum* and *Diplonema sp.*) and a partial sequence of the mitochondrial gene for cytochrome c oxidase subunit I (Cox I protein) from one diplonemid (*Diplonema papillatum*) (Maslov et al. 1999). Maslove and Simpson (1999) performed a molecular phylogenetic study using these sequences in order to analyze the phylogenetic position of diplonemids within the phylum Euglenozoa. In their phylogenetic analyses, *Diplonema* was shown to be a sister-group of either kinetoplastids (in trees inferred with the maximum-likelihood method), or euglenoids (in trees inferred with the parsimony and distance methods). In either case, however, the affinity is not well supported by bootstrap analysis and the differences between the best tree and the alternative trees were not significant.

It remains unclear how diplonemids are related to euglenoids and kinetoplastids. In molecular trees, this may be due to two weaknesses in the phylogenetic analyses conducted by Maslove et al. (1999): 1) The very small sampling size. All the SSU gene trees were based on only seventeen taxa, including two diplonemid-sequences, and all the Cox I protein trees were based on only seven taxa, including one diplonemid-sequence. 2) The mitochondrial Cox I protein phylogeny can be unreliable due to the fast evolution of euglenozoan mitochondrial genes.

In this study, I have characterized diplonemid nuclear encoded genes for actin, alpha-tubulin, beta-tubulin and GAPDH, to construct novel phylogenetic trees to try to resolve the phylogenetic position of diplonemids within the Euglenozoa.

1.2 Intron types in the Euglenozoa

Three types of introns in euglenoids and/or kinetoplastids

Three types of introns occur in euglenoids and/or kinetoplastids: conventional 'GT-AG' spliceosomal introns, trans-spliced, or discontinuous 'introns' and "aberrant" introns. I will describe each of the three types and their distributions within the Euglenozoa in the following three sections.

GT-AG spliceosomal introns

GT-AG spliceosomal introns are abundant in higher eukaryotes. Genes in most eukaryotes are transcribed into pre-mRNAs that include introns. Only when all the introns in the pre-mRNAs are excised will the mature mRNAs be transported from the nucleus to the cytosol where translation takes place. The precise removal of the introns from the primary RNA transcripts is a critical step in gene expression in all eukaryotic cells. In general, it is a two-step catalytic process aided by a group of small nuclear ribonucleoprotein particles (snRNPs) together called the spliceosome. The spliceosome is mainly composed of five snRNPs (U1, U2, U5 and U4/U6), and assembles on the precursor messenger RNA through RNA-RNA, RNA-protein, and protein-protein interactions. The first step in cis-splicing is the cleavage of the 5' splice site by the formation of a 2'-5' phosphodiester bond between an adenosine within the intron and the guanosine residue at the 5' end of the intron. This generates a free 5' exon and an intermediate RNA in a lariat structure. The second step involves the cleavage of the 3' splice site, the ligation of the 5' exon and the 3' exon and the release of the intron in a lariat structure (Fig. 1).

Since the introns are removed before expression of the gene, most intron sequences accumulate mutations during evolution more rapidly than the flanking exons. The only highly conserved sequences within the intron are those required for intron removal or for recognition during formation of the spliceosome. In particular, the 'GT' at the 5' end and 'AG' at the 3' end of an intron are almost invariant. Mutational studies have shown that disrupting either the GT at the 5' splice site or the AG at the 3' splice site can block or reduce the rate of both steps during the cis-splicing (Sharp 1987).

In addition to the consensus 5' and 3' splice junction sequences, the next conserved sequence regions are the branchpoint region and the region between the branch point and the 3' splice site (Sharp 1987; Umen et al. 1995). The branchpoint site is where the lariat intermediate forms after the first step of the cis-splicing. During the first step of the splicing, the 2' hydroxyl of an adenosine in this branchpoint site attacks the phosphodiester bond between the guanosine at the 5' terminus of an intron and an ajacent exon nucleotide. This leads to the releasing of the 5' exon and the formation of a 2'-5' phosphodiester bond between the branchpoint adenosine and the guanosine at the 5' terminus of an intron. In

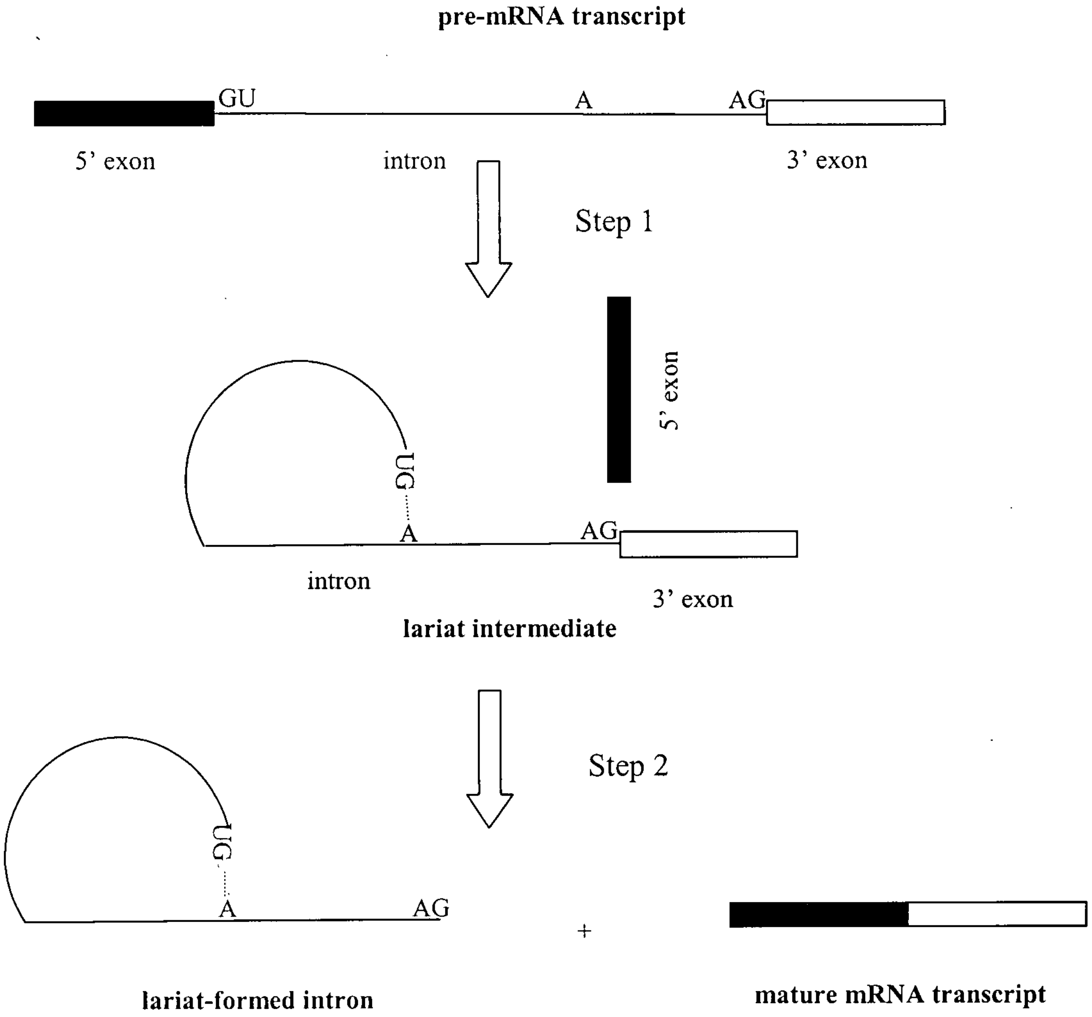


Fig. 1 The two-step cis-splicing. Filled square represents 5' exon and open square represents 3' exon. Intron is represented by black line. The consensus dinucleotides at either end of the intron are marked as GU and AG. The branchpoint adenosine is marked as A. The dashed line between G and A represents the 2'-5' phosphodiester bond formed after the first step of the splicing. See text for detailed description.

yeast, the branchpoint region is strictly maintained. It has the consensus sequence 5'-UACUA<u>A</u>CA-3' (the underlined adenosine is the adenosine participating in the formation of the 2'-5' phosphodiester bond). In mammals, branchpoint region is less conserved, but the region between the branchpoint and the 3' splice site is a conserved polypyrimidine tract, and is one of the essential recognition sites for the binding of splicing factors (Sharp 1987; Tazi et al. 1986; Gerker et al. 1986; Umen et al. 1995).

Conventional GT-AG spliceosomal introns have been found in both green and colourless euglenoids, although they are rare in both. In *Euglena*, so far, only three GT-AG introns have been found in the fibrillarin gene of *Euglena gracilis* (Breckenridge et al. 1999). In the colourless euglenoid, *Entosiphon sulcatum*, one spliceosomal intron has been found in a beta-tubulin gene (Ebel et al. 1999). In kinetoplastids, only two GT-AG cis-splicing introns have very recently been discovered in the poly (A) polymerase (PAP) genes from both *Trypanosoma brucei* and *Trypanosoma cruzi* (Mair et al. 2000).

Trans-spliced discontinuous introns

Trans-splicing is also a post-transcriptional RNA-splicing process. The distinguishing difference between cis- and trans-splicing is that, in trans-splicing, two exons flanking the discontinuous intron are on two different pieces of pre- messenger RNAs (Agabian 1990; Blumenthal et al. 1988; Nilsen 1995) (Fig. 2). However, trans-splicing is not an entirely novel RNA-splicing process, it is regarded as the splicing of a discontinuous GT-AG spliceosomal intron. Trans-splicing is similar to GT-AG spliceosomal cis-splicing in three fundamental ways. First, this discontinuous 'intron' also has consensus GT and AG dinucleotides sequences at either end. Second, the chemistry of trans-splicing involves two transesterification-reaction, with the discontinuous 'intron' forming a Y-branched intermediate that is structurally analogous to the cis-splicing lariat (Blumenthal et al. 1988;

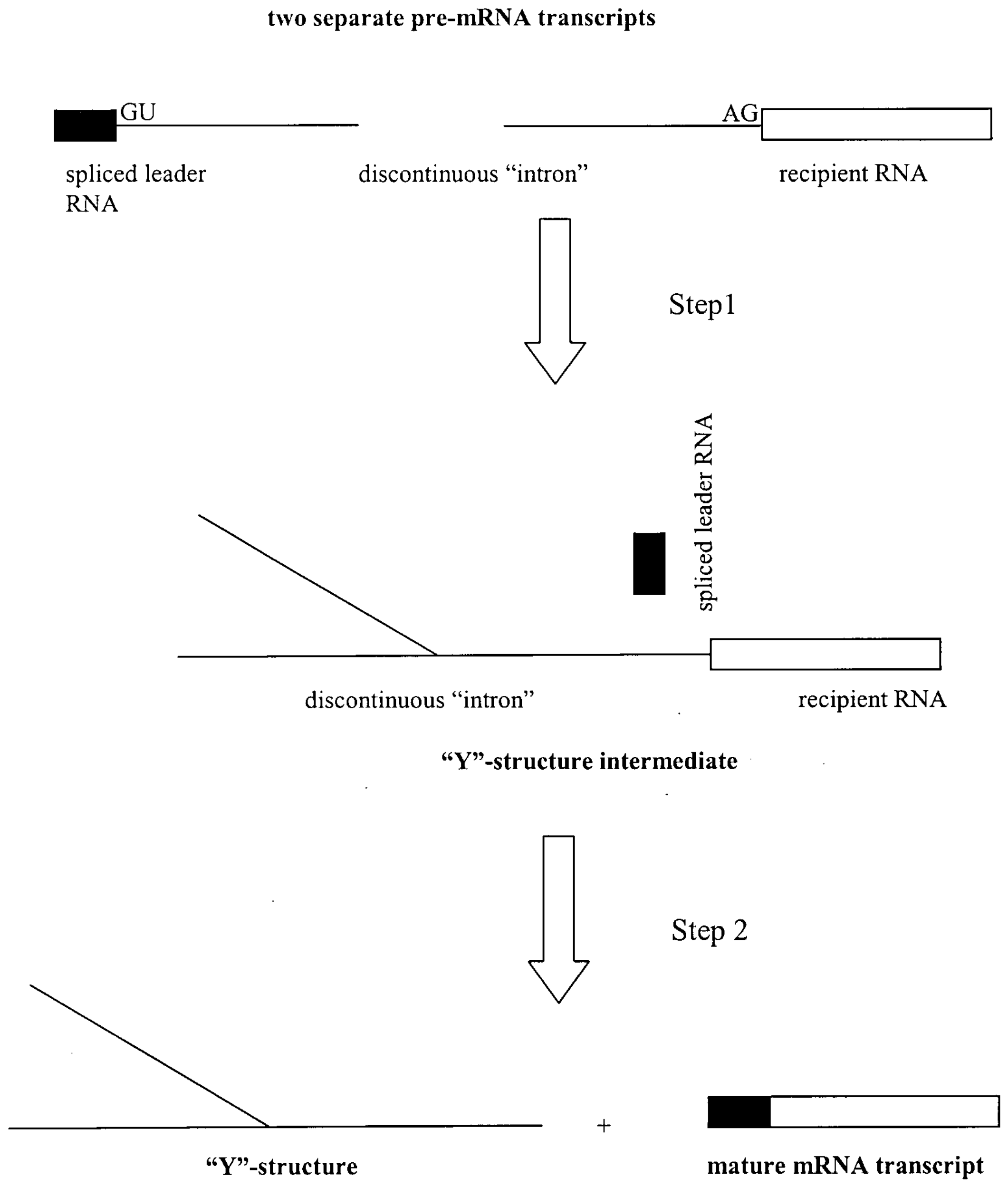


Fig. 2 The two-step trans-splicing. The filled square represents the spliced leader RNA and the open square represents the recipient exon RNA. The discontinuous "intron" is represented by black line. The consensus dinucleotides at either end of the discontinuous "intron" are marked as GU and AG. See text for detailed description.

Agabian 1990; Nilsen 1995) (Fig. 2). Third, cis- and trans-splicing share at least three small nuclear ribonucleoprotein particles- U2, U4 and U6 snRNPs (Agabian 1990; Nilsen 1995).

Around eighty percent of the mRNAs are trans-spliced in both the green euglenoid *Euglena gracilis* (Tessier et al. 1991) and the colourless euglenoid *Entosiphon sulcatum* (Ebel et al. 1999), whereas all known mRNAs are trans-spliced before they are translated into proteins in kinetoplastids (Agabian 1990; Nilsen 1994; Laird 1989).

In addition to kinetoplastids and euglenoids, trans-splicing has only been reported in the Metazoan worms, such as nematodes (e. g. *Caenorhabditis elegans*) (Blumenthal et al. 1988; Agabian 1990; Nilsen 1989; Nilsen 1994) and flatworms (e. g. trematodes) (Davis 1997; Nilsen 1995), but the process certainly evolved independently in these animals and euglenozoa.

"Aberrant" introns

A third type of intron, here simply called "aberrant" introns, has also been found in the genome of *Euglena gracilis*. In general, these introns have three distinctive features (Tessier et al. 1992; Muchhal et al. 1994; Henze et al. 1995). First, these introns do not have any consensus sequences at their borders. Second, they employ an unusual stable stem-loop secondary structure in the pre-mRNA (Fig. 3) (Tessier et al. 1992; Muchhal et al. 1994), and further secondary structures (stem-loop structures) are observed in the "aberrant" introns in the cytosolic GAPDH gene of *Euglena gracilis* (Henze et al. 1995). Third, they are usually flanked by short (2- to 4-bp) repeats (Tessier et al. 1992; Muchhal et al. 1994; Henze et al. 1995).

"Aberrant" introns have only been reported from *Euglena gracilis*. They are found in three nuclear encoded genes: 14 introns in the gene for the light harvesting chlorophyll a/b binding proteins of photosystem II (LHCPII) (Muchhal et al. 1994); 12 introns in the rbcS

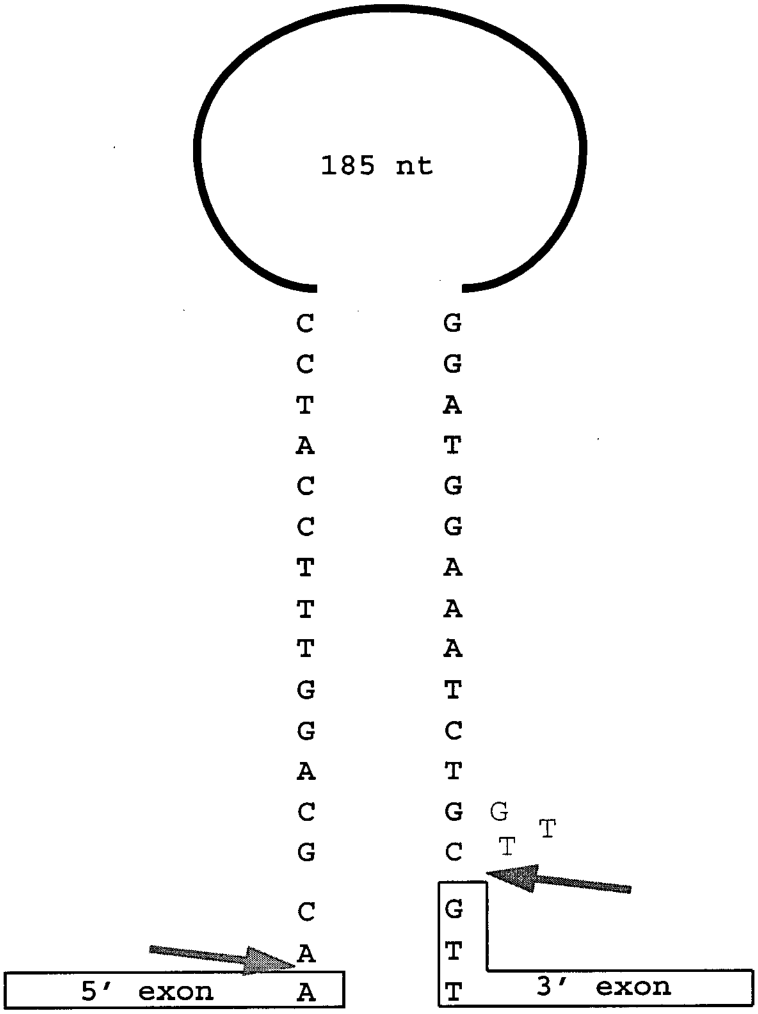


Fig. 3 The secondary stem-loop structure of an intron in the rbcS gene from *Euglena gracilis* (Tessier et al. 1992). The arrows point at the two cleavage sites of the intron. This intron does not have consensus dinucleotides (GT-AG) at either end. Two stretches of nucleotides at the 5' and 3' ends of the intron can base-pair to each other, usually with several nucleotides at the 3' end of the intron displaced by two adjacent nucleotides from the 3' exon.

genes, which encodes the small subunits of the ribulose 1,5 bisphosphate carboxylase oxygenase (Tessier et al. 1992), and four introns in the gene for cytosolic, glycolytic glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (Henze et al. 1995).

Introns in diplonemids?

Nothing is known about the intron distribution in diplonemids, the third major group of the Euglenozoa, because no nuclear protein-coding genes from any diplonemid have ever been sequenced, and their relationship to other Euglenozoa is unclear.

In order to be able to determine the origins of nuclear cis-splicing, trans-splicing, and the "abnormal" introns in the Euglenozoa, we need to know the distribution of these characters in all three main groups, and the internal phylogeny of the phylum. Diplonemids, as the third and most poorly studied major group of this phylum, are an important part of this puzzle, since the lack of known protein-coding gene sequences from diplonemids is a gap in our understanding of intron distribution, and hinders phylogenetic analyses to determine the branching order within the Euglenozoa.

The objective of the first part of this thesis, therefore, is to determine the evolutionary history of the intron-types in the Euglenozoa. In order to achieve this goal, I sequenced several nuclear protein-coding genes from diplonemids. By so doing, I sought to determine whether these nuclear encoded genes contain introns, and if they do, what kind of introns they possess. Also, I added my new diplonemid protein-sequences to protein alignments and constructed phylogenetic trees, hoping to solve the internal branching order of the three major groups of the phylum Euglenozoa. Then, based on the possible internal phylogeny, I attempt to infer the origins of the three intron types within the phylum Euglenozoa.

The nuclear encoded genes chosen for this study were: actin, alpha-tubulin and beta-tubulin. These proteins are the basic components for the cytoskeleton universally present in the eukaryotic cells. These genes are good candidates because they have been widely used as phylogenetic markers and sequences from many phylogenetically distinctive groups are available, including those from both euglenoids and kinetoplastids. In addition, these nuclear encoded genes often contain one or more introns in higher eukaryotes.

1.3 Diplonemid GAPDH

The second part of my thesis focuses on a phylogenetic analysis of glyceraldehyde-3-phosphate dehydrogenase (GAPDH). GAPDH was also chosen for analysis because it is well sampled and has a well-known intron distribution among extant eukaryotes.

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a central carbon metabolic enzyme. The phylogeny of GAPDH is complex, resulting from a complicated evolutionary history that includes gene duplications, endosymbiotic gene replacements and lateral gene transfers (Martin et al. 1993; Henze et al. 1995; Liaud et al. 1997). In global GAPDH trees constructed previously, GAPDH sequences can always be divided into two distinct clades, GapC and GapA/B. The GapC clade mainly represents the cytosolic GAPDH enzymes of eukaryotes. The GapC enzymes are generally involved in the glycolysis in the cytosol. The reaction catalised by the GapC enzyme is catabolic and this enzyme is NAD specific. However, the gap1 from proteobacteria and cyanobacteria are also included in the GapC clade. The reason for this is unclear so far. Conversely, the GapA/B clade mainly represents the GAPDH enzymes of bacteria. One exception to the eubacterial nature of the GapA/B clade is the inclusion of GAPDH from the eukaryotic phylum Parabasalia. Markos et al. (1993) and Viscogliosi et al. (1998) suggested that the close association of the GAPDH sequences from parabasalids with bacterial GAPDH sequences indicated a bacterial origin of the GAPDH genes in Parabasalia, most likely by a lateral gene transfer from a bacterium to the ancestor of this phylum. In addition, the GapA/B clade also includes the nuclear-encoded, chloroplast-targeted GAPDH sequences from photosynthetic eukatyotes, which are also bacterial due to the cyanobacterial origin of the chloroplast. Indeed, in the GapA/B clade of global GAPDH trees, the nuclear-encoded, plastid-targeted GAPDH genes of photosynthetic eukaryotes are always closely related to the gap2 of cyanobacteria (considered to be the free-living relatives of chloroplasts) (Martin et al. 1993; Henze et al. 1995; Liaud et al. 1997). The chloroplast GapA/B enzyme is involved in the Calvin cycle. The reaction catalised by this enzyme is anabolic and the substrate of this enzyme can be either NAD or NADP. Clermont et al. (1993) demonstrated that the amino acid at position 32 of a GAPDH gene plays an essential role in choosing the relative specificity of NAD or NADP as its substrate. In most catabolic NAD-specific cytosolic GAPDH enzyme this position is aspartic acid (D), whereas for the anabolic GAPDH enzyme that is both NAD- and NADP-specific, this position is occupied by a non-acidic amino acid, for instance alanine (A) in the chloroplast-targeted GAPDH of *Euglena gracilis*. This is because there is an electrostatic repulsion between the negatively charged carboxyl group of an acidic amino acid (Asp32) and the negatively charged 2'-phosphate of NADP.

The phylogeny of GAPDH in the phylum Euglenozoa is very complicated. It has been shown that there are two distantly related GAPDH genes in both euglenoids (chloroplast and cytosolic GAPDH genes) (Martin et al. 1993; Henze et al. 1995) and kinetoplastids (glycosomal and cytosolic GAPDH genes) (Michels et al. 1991; Michels et al. 1992).

The chloroplast GAPDH gene of *Euglena gracilis* is closely related to those of higher photosynthetic eukaryotes and the gap2 of the cyanobacteria that gave rise to chloroplast-targeted GAPDH genes in higher photosynthetic eukaryotes (Martin et al. 1993; Henze et al. 1995). The cytosolic GapC of *Euglena gracilis* has been shown to be closely related to the glycosomal GAPDH genes of kinetoplastids. Glycosomes are unique microbodies of kinetoplastids, which harbor most enzymes of the glycolytic pathway and are often thought to be of endosymbiotic origin (Opperdoes 1987; Borst et al. 1989; Opperdoes et al. 1989; Michels et al. 1994). This *Euglena* cytosol/kinetoplastids glycosomes clade branches basally to the GapC clade (Henze et al. 1995; Liaud et al. 1997).

Most GAPDH in kinetoplastids is found in glycosomes (Opperdoes 1987; Borst et al. 1989; Opperdoes et al. 1989; Michels et al. 1991; Michels et al. 1992). However, *Trypanosoma brucei* and *Leishmania mexicana* possess a second, distinct cytosolic GAPDH enzyme in addition to the glycosomal form (Michels et al. 1991; Michels et al. 1992). These two cytosolic GAPDH enzymes are extraordinarily closely related to *E.coli* gap1 (=gapA) (Michels et al. 1991, Henze et al. 1995). Michels et al. (1992) have further proved that more distantly related kinetoplastids, such as a bodonid *Trypanoplasma borelli,* only have the typical glycosomal GAPDH enzyme. This strongly supports the speculation that the *Euglena* cytosol/kinetoplastid glycosome clade represents the original GAPDH form to the phylum Euglenozoa, and a horizontal gene transfer, perhaps from a γ-purple bacterium related to *E. coli*, resulted in the cytosolic GAPDH in *Trypanosoma* and *Leishmania* after their ancestor diverged from the Bodonids (Michels et al. 1992; Michels et al. 1994; Henze et al. 1995; Liaud et al. 1997).

How the GAPDH sequences of diplonemids fit into this picture is entirely unknown. Outstanding questions include: how many types of GAPDH genes are there in diplonemids? Where, in the global GAPDH tree, are they going to branch? Since it is generally thought that the *Euglena* cytosol/kinetoplastid glycosome clade represents the original GAPDH form to the phylum Euglenozoa (Michels et al. 1992; Michels et al. 1994; Henze et al. 1995; Liaud et al. 1997), the positions of the GAPDH sequences from diplonemids in the GAPDH tree may either confirm this speculation or possibly reveal new relationships between diplonemid GAPDH sequences and those of eukaryotes or prokaryotes.

CHAPTER II:

Materials and Methods

2.1 Strains and culture conditions

Axenic cultures of *Diplonema ambulator* (ATCC 50223), *Diplonema papillatum* (ATCC 50162), *Diplonema sp.* 3 (new strain) (ATCC 50225), *Diplonema sp.* 4 (ATCC 50232) and *Rhynchopus sp.* 3 (ATCC 50231) were obtained from the ATCC (American Type Culture Collection). Cultures were maintained in four 150x15 mm sterilized, disposable plastic petri dishes (FISHER) in ATCC Culture medium1728, enriched *Isonema* medium (ATCC 1405 HESNW Medium) with 10% heat-inactivated horse serum (Sigma Cat. # H1270) added aseptically just before use. (Detailed recipes are given at the end of this section). Cultures were incubated at room temperature. After significant growth was observed by light microscopy, cells were harvested by centrifugation at 2000xg, 4^0C, for 10 minutes.

Genomic DNA of *Diplonema sp.* 2 (ATCC 50224), *Diplonema sp.* 3 (ATCC 50231), *Diplonema sp.* 4 (ATCC 50232), *Rhynchopus sp.* 1 (ATCC 50226), and *Rhynchopus sp.* 2 (ATCC 50229) were provided by Ema Chao.

ATCC Medium 1405:

Natural seawater	1.0 L
Enrichment Solution (see below)	10.0 ml
Vitamin Solution (see below)	1.0 ml

Two-month-old seawater was filter-sterilized and all components were combined aseptically.

Enrichment Solution:

EDTA · $2H_2O$	0.553 g
$NaNO_3$	4.667 g
$Na_2SiO_3 \cdot 9H_2O$	3.000 g
Sodium glycerophosphate	0.667 g
H_3BO_3	0.380 g
$Fe(NH_4)_2(SO_4)_2 \cdot 6H_2O$	0.234 g

$FeCl_3 \cdot 6H_2O$	0.016 g
$MnSO_4 \cdot 4H_2O$	0.054 g
$ZnSO_4 \cdot 7H_2O$	7.3 mg
$CoSO_4 \cdot 7H_2O$	1.6 mg
Distilled water	1.0 L

Na_2SiO_3 was neutralized with 1 N HCl. All ingredients were combined in the order listed. This solution was filter-sterilized.

Vitamin Solution:

Thiamine	0.1 g
Vitamin B_{12}	2.0 mg
Biotin	1.0 mg
Distilled water	1.0 L

This solution was filter-sterilized.

2.2 DNA extraction procedures

Cell pellets of diplonemids were resuspended in a 1.5 ml CTAB Solution (4% (w/v) CTAB (Hexadecyltrimethylammonium bromide, SIGMA H-5882), 100mM MES (SIGMA M-8250), 1.4M NaCl and 1% 2-Mercaptoethanol) pre-heated to 65^{0}C. The mixture was incubated at 65^{0}C for 30 minutes to allow for digestion and lysis. DNA was then gently extracted (to avoid extensive shearing) from the mixture with an equal volume of chloroform/isoamyl alcohol (24:1). DNA was precipitated from the aqueous phase by adding 2/3 volume isopropyl alcohol and incubating overnight at 4^{0}C. DNA was collected the following day by successive centrifugation of 1.5 ml aliquot portions in the same 1.5 ml tube at maximum speed (usually 12500 rpm) for 2.5 minutes. The DNA pellet was washed twice with 95% ethanol and twice with 70% ethanol to remove salt before being air-dried pellet and resuspended in TE (10/1 Tris/EDTA, pH 8.0).

2.3 PCR conditions

Degenerate PCR primers for alpha-tubulin, beta-tubulin, actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) were designed by Dr. Patrick Keeling based on the conserved amino acid sequences at the extreme Amino- and Carboxyl- termini of the corresponding protein.

The conserved regions at the N- and C- termini of the four proteins used for primer designing are given below:

Alpha-tubulin gene,

N- terminus	5'-QVGNAGWE-3'
C- terminus	5'-WYVGEGM-3'.

Beta-tubulin gene,

N- terminus	5'-GQCGNQ-3'
C- terminus	5'-MDEMEFT-3'.

Actin gene,

N- terminus	5'-EKMTQIMFE-3'
C- terminus	5'-VHRKCF-3'.

GAPDH gene,

N- terminus	5'-KVGINGFG-3'
C- terminus	5'-WYDNEWGYS-3' .

The degenerate primer pairs used for these four nuclear encoded genes were:

Alpha-tubulin gene,

TUBA1	5'-TCC GAA TTC ARG TNG GNA AYG CNG GYT GGG A-3'
TUBA2	5'-CGC GCC ATN CCY TCN CCN ACR TAC CA-3'.

Beta-tubulin gene,

TUBB1	5'-GCC TGC AGG NCA RTG YGG NAA YCA-3'
TUBB2	5'-TCC TCG AGT RAA YTC CAT YTC RTC CAT-3'

Actin gene,

actF2	5’-GAG AAG ATG CAN CAR ATH ATG TTY GA-3’
actR1	5’-GGC CTG GAA RCA YTT NCG RTG NAC-3’

GAPDH gene,

gap1F	5’-CCA AGG TCG GNA THA AYG GNT TYG G-3’
gap1R	5’-CGA GTA GCC CCA YTC RTT RTC RTA CCA-3’

Amplification of DNA was carried out using standard methods. Typically, 250ng of diplonemid genomic DNA was used as a template in 50 µl reactions, with each primer at 10 µM, 0.25 units of Taq polymerase, 2.5mM concentration of dNTPs and reaction buffer (Gibco BRL). Cycle parameters were 94^{0}C/ 30 Sec & pause 2.15min (1x); 94^{0}C /30sec, 50^{0}C /30sec, 72^{0}C/ 2min (30x); and 72^{0}C /5min(1x).

2.4 Cloning of amplified fragments

An aliquot of 5µl of each 50µl PCR reaction was run on an agarose gel (0.7-0.8% agarose), together with the DNA molecular weight marker (1 Kb DNA ladder), to check the size of product. If the product has the expected molecular weight, the remaining portion of the reaction was run on another agarose gel (0.7-0.8% agarose), and the fragment of interest was isolated from the gel using either the Prep-a-gene kit (BIO-RAD) or GeneClean II kit (BIO 101 BIO/CAN SCIENTIFIC).

Isolated fragments were ligated into the pCR 2.1-TOPO T-tailed vectors as specified by the manufacturer’s protocol (Invitrogen). After 5 minutes of incubation at room temperature, ‘One Shot Competent TOPO 10’ *Escherichia coli* cells were transformed with the ligated plasmids following the manufacturer’s protocol (Invitrogen). The cells were plated on selective LB medium containing 50 µg/ml ampicillin and 40 µl of 40 mg/ml X-gal and incubated overnight at 37^{0}C. The presence of X-gal allowed for ‘blue-white screening’ where colonies containing vectors with inserts appear white, whereas colonies containing 'empty vectors' appear blue. In order to determine the sizes of the inserts within the plasmids, 6-10 or more white colonies per cloning reaction were chosen and either a restriction analysis (digest with *Eco*R I) or a screening reaction (PCR) by amplifying these inserts using M13 Forward (-20) and M13 Reverse primers were performed on them. On average, six white colonies, each of which contained plasmid with the insert of expected size, per cloning reaction were cultured overnight in individual tubes containing liquid LB medium with 50 μg/ml ampicillin. Plasmid DNA with the expected size of PCR insert was isolated using either the standard alkaline lysis (miniprep) method (Sambrook et al. 1989) or the Perfect prep Plasmid DNA Kit following the manufacturer's protocol (Eppendorf).

2.5 DNA sequencing

Automated sequencing using the dideoxy method was employed to obtain the sequences of cloned PCR products. The full-length sequences of genes were obtained using a primer-walking strategy for alpha- and beta- tubulin genes. (Walking primer sequences are given below.) Regions sequenced only on a single strand were confirmed by two independent sequences.

The forward strands of the alpha-tubulin genes of *Diplonema sp.* 2 and *Diplonema sp.* 3 were sequenced using the following oligonucleotide primer:

<u>TUAA32</u> GCG GCG AAC AAC TAC GC.

The reverse strands of the alpha-tubulin genes of *Diplonema sp.* 3, *Diplonema sp.* 4 , *Rhynchopus sp.*1 and *Rhynchopus sp.* 2 were sequenced using the following primer:

<u>TUAA41</u> GG CAG CAC GCC ATG TAC.

The forward strand of the beta-tubulin gene of *Rhynchopus sp.* 1 was sequenced using the following oligonucleotide primer:

<u>TUBB32</u> GG TGC GGG GAA CAA CTG.

The reverse strands of the beta-tubulin genes of *Diplonema sp.* 3, *Rhynchopus sp.* 1 and *Rhynchopus sp.* 2 were sequenced using the following oligonucleotide primer:

TUBB42 GAC TTG ATG TTG TTC GGG.

The forward strands of the beta-tubulin genes of *Diplonema sp.* 2, *Diplonema sp.* 3, *Diplonema sp.* 4 and *Rhynchopus sp.* 1 were sequenced using the following oligonucleotide primer:

TUBB3 GGA GCT GGT AAC AAC TGG.

The reverse strands of the beta-tubulin genes of *Diplonema sp.* 2, *Diplonema sp.* 3, *Diplonema sp.* 4 and *Rhynchopus sp.* 1 were sequenced using the following oligonucleotide primer:

TUBB4 C TTG ATG TTG TTT GGA ATC.

The forward strand of the beta-tubulin gene of *Rhynchopus sp.* 2 was sequenced using the following oligonucleotide primer:

TUBB33 GG TGC GGG CAA CAA CTG.

2.6 Sequence alignment and phylogenetic analyses

The nature of the obtained sequences was confirmed by BLAST searches against the GenBank database. Introns were tentatively identified by insertions in genes that couldn't be aligned to the amino acid sequences of the same gene from other organisms. The sequences were then imported into the Sequencher 3.1.1 software package, where contigs were assembled. Once the contigs were complete, they were translated into amino acid sequence using the DNA Strider 1.2 program. By comparing the inferred amino acid sequences with the corresponding protein alignments, introns were positively identified by the presence of canonical GT-AG boundaries. Introns were removed, then the nucleotide sequences were translated into amino acid sequences. All the inferred amino acid sequences were added to the corresponding protein alignments.

Amino acid sequence alignments of alpha-tubulin (436 amino acids), beta- tubulin (428 amino acids) and actin (373 amino acids) that included broad samplings of eukaryotes, and GAPDH (290 amino acids) from a wide range of both eukaryotes and prokaryotes were provided by Dr. Patrick Keeling and Dr. Naomi Fast. Amino acid sequences from diplonemids were added to these four alignments. Regions in the alignments that did not appear optimal were subsequently adjusted manually using a text editor. Phylogenetic analyses were performed on the aligned protein datasets using a distance method.

PUZZLE version 4.0.2 (Strimmer and von Haeseler 1996) was used to calculate maximum likelihood distances between pairs of sequences. The distance matrices were corrected by the JTT substitution frequency matrix with amino acid usage estimated from the data, site-to-site rate variation modeled on a gamma distribution with eight rate categories (except for the GAPDH alignment with 100 taxa). The Gamma distribution parameter alpha was estimated from each dataset. Trees were constructed from the distance matrices with the neighbor-joining (NJ) algorithm using the BioNJ program (Gascuel 1997). A hundred bootstrap resamplings of the data were generated by the SEQBOOT program implemented in the Phylip 3.572 package (Felsenstein 1993). One hundred distance matrices were inferred from the 100 resampled alignments by PUZZLE version 4.0.2 (using the settings described above but not with the gamma-distribution), using the shell script puzzleboot (by M. Holder & A. Roger). A hundred trees were generated by analyzing the 100 distance matrices with the neighbor-joining (NJ) method using BioNJ. The bootstrap majority-rule consensus tree was constructed using the CONSENSE program from the Phylip 3.752 package.

Alternative internal topologies of the phylum Euglenozoa in the alpha- and beta-tubulin and actin trees were tested statistically using the Kishino-Hasegawa (K-H) method (Kishino and Hasegawa 1989). This method evaluates the standard error of the difference in ln likelihood between alternative topologies, that is, it allows one to test whether a tree topology with higher likelihood is significantly preferred over others with lower likelihood. All K-H tests were performed using PUZZLE version 4.0.2 with gamma-distributed rates and user-defined trees (the parameters used were based on the first input tree). For the present studies, differences of log likelihood greater than 1.96 standard errors (corresponding to a 95% confidence interval) were considered significant (Kishino and Hasegawa 1989).

CHAPTER III:
Results

3.1 Sequences for nuclear encoded genes from diplonemids

I sequenced genes for alpha- and beta-tubulin, actin and GAPDH from nine different diplonemids in this study. I obtained twenty-nine seqences in total: ten alpha-tubulin sequences from eight different diplonemids, thirteen beta-tubulin sequences from eight different diplonemids, two actin sequences from two different diplonemids and four GAPDH sequences from three different diplonemids. (See Table 1 for a summary). The predicted amino acid sequences inferred from the nucleotide sequences for these twenty-nine nuclear encoded genes are given in Fig. 4-Fig. 7. Lengths of these sequences (with neither intron nor PCR primer sequences included) are approximately: 733 nt for actin, 1153 nt for alpha-tubulin, 1162 nt for beta-tubulin, and 904-949 nt for GAPDH.

Actin gene sequences

A band close to the expected size (777 nt) was obtained from *Diplonema ambulator*. A band larger than the expected size was obtained from *Diplonema sp.* 3. These two amplified products were cloned and seqeuenced, and blast searches against the GenBank database confirmed both encoded actin genes. Blast searches also revealed that both sequences contain introns: two in the actin gene of *Diplonema sp.* 3 (80 nt and 176 nt in length) and one in the actin gene of *Diplonema ambulator* (40 nt in length). The length of each of the two predicted protein sequences is 244 amino acids, representing about two-thirds of a complete actin sequence.

Table 1. Twenty-nine nuclear encoded genes from nine diplonemids. Numbers indicate copy/copies sequenced from one specific diplonemid.

diplonemid	**α-tubulin**	**β-tubulin**	**Actin**	**GAPDH**
Diplonema sp. **2**	1	1		2
Diplonema sp. **3**	1	1	1	1
Diplonema sp. **3 new**	2	2		
Diplonema sp. **4**	1	3		
Diplonema ambulator		2	1	
Diplonema papillatum	1			
Rhynchopus sp. **1**	1	1		
Rhynchopus sp. **2**	1	1		
Rhynchopus sp. **3**	2	2		1

```
1 D.sp.3       --------------- --------------- --------------- --------------- --------------- ---------------      0
2 D.ambulator  --------------- --------------- --------------- --------------- --------------- ---------------      0
3 E.gracilis   MAEEIEQQALVCDNG SGMVKAGFAGDDAPR CVFPSIVGRRKNDSA MMGTAKKDAYIGDDA QAKRGILFIKYPIEH GIVTNWDDMEKIWHH     90
4 T.cruzi      -MSDEEQSAIVCDNG SGMVKAGFSGDDALC HVFPSIVGRPKNEQA MMGSASKKLFVGDEA QAKRGVLSLKYPIEH GIVTNWDDMEKIWHH     89

1 D.sp.3       --------------- --------------- -------TFNSPAMY VGIQAVLSLYSSGRT TGIVLDSGDGVTHTV PIYEGYSLPHAVLRI     53
2 D.ambulator  --------------- --------------- -------TFNSPAMY VGIQAVLSLYSSGRT TGIVLDSGDGVTHTV PIYEGYSLPHAVLRI     53
3 E.gracilis   TFFNELRVAPEDHPV LLTEAPMNPKSNREK MTQIMFETFNVPALY VSIQAVLSLYSSGRT TGIVLDCGDGVSHTV PIYEGYSLPHAVLRI    180
4 T.cruzi      TFYNDVRVNPESHSV LLTEAPMNPKQNREK MTQIMFETFGVPAMY VGIQAVLSLYSSGRT TGIVLDAGDGVTHTV PIYEGYSLPHAIRRM    179

1 D.sp.3       DMAGRDMTDWMIKLL TERGNSFVTSAEREI VRDIKEKLAYVALDF DEEMSLASSSSSLEK DYELPDGQVITVGSE RFRCPEALFRPAFIG    143
2 D.ambulator  DMAGRDMTDWMIKLL TERGNSFVTSAEREI VRDIKEKLAYVALDF DEEMSLAASSSSLEK DYELPDGQVITVGSE RFRCPEALFKPAFIG    143
3 E.gracilis   DMAGRDLTDYMMKLL TERGLSFTTSAEREI VRDVKEKLCYVALDF DEEMSLATSSSSVEK EYELPDGNIIQVGSE RFRCPEVLMKPSMIG    270
4 T.cruzi      DMAGRDLTEYLMKLL MESGMTFTTSAEKEI VRNVKEQLCYVALDF DEEVTNS-AKTVNEE PFELPDGTIMQVGNQ RFRCPEALFKPMLIG    268

1 D.sp.3       LEAN-GIHETVYNSI MKCDIDVRKDLYANI VLSGGTTMYEGLADR LSKEVTNLAPNSMKI KVVAPPERKYSVWIG GSILSSLSTFASMWV    232
2 D.ambulator  LEAN-GIHETVYNSI MKCDIDVRKDLYANI VLSGGTTMYEGLADR LSKEVTNLAPNSMKI KVVAPPERKYSVWIG GSILSSLSTFATMWV    232
3 E.gracilis   LEAC-GVHETTFNSI NKCDIDVRKDLYSNI VLSGGTTMYEGLPER MSKEITNLAPNSMKI KVVAPPERKYSVWIG GSILASLSTFQSMWI    359
4 T.cruzi      LDEAPGFHEMTFQSI NKCDIDVRRDLYGNI VLSGGTTMFKNLPER LGKEISNLAPSSIKP KVVAPPERKYSVWIG GSILSSLTTFQTMWI    358

1 D.sp.3       KKEEYDESGPGI--- ---     244
2 D.ambulator  KKEEYDESGPGI--- ---     244
3 E.gracilis   KKEEYDEAGPGIVHR KCF     377
4 T.cruzi      KKSEYDEAGPSIVHN KCF     376
```

Fig. 4 An alignment of four amino-acid sequences of actin. Amino acid sequences were aligned to obtain maximal similarity. Dashes indicate the absence of amino acids at the corresponding positions. The amino acid sequences are as follows: *Diplonema sp.* 3 (from this study), *Diplonema ambulator* (from this study), *Euglena gracilis*, and *Trypanosoma cruzi*. Bold characters in *Diplonema sp.* 3 and *Diplonema ambulator* indicate the positions of introns found in the corresponding genes.

The deduced amino acid sequences of *Diplonema sp.*3 and *Diplonema ambulator* were aligned with the homologous region of the actin sequences from 63 other eukaryotes. Figure 4 shows a representative alignment including actin sequences from *Diplonema sp.*3, *Diplonema ambulator*, *Euglena gracilis* and *Trypanosoma cruzi*. The two diplonemid sequences were very similar to each other, with only 3 amino acid differences over the total length of 244 amino acids (sequence differences were calculated by PAUP version 4.0). In addition, they were similar to the sequences of other euglenozoa: when the two diplonemid actin sequences were compared to that of *Euglena gracilis,* only 37 of the 244 amino acids were different, whereas sequence differences between diplonemids and kinetoplastids were higher (64-71 of the 244 amino acids).

Alpha-tubulin gene sequences

PCR products of the expected size (1197 nt) were obtained from *Diplonema sp.* 2, *Diplonema sp.* 3, *Diplonema sp.* 3 (new strain), *Diplonema papillatum, Rhynchopus sp.* 1, *Rhynchopus sp.* 3. PCR products of larger than the intronless sizes were obtained from *Diplonema sp.* 4 and *Rhynchopus sp.* 2. The above PCR products were cloned, and both strands of two independent clones from each source were sequenced. It was confirmed that all of them were true alpha-tubulin sequences by BLAST searches against the GenBank database. The results of the BLAST searches also revealed that there was one intron in each of the two alpha-tubulin clones from *Diplonema sp.* 4 (109 nt) and *Rhynchopus sp.* 2 (126 nt). In addition, by comparison, I found that the two alpha-tubulin clones from both *Diplonema sp.* 3 (new strain) and *Rhynchopus sp.* 3 were slightly different from each other. The two sequences from *Diplonema sp.* 3 (new strain) vary at several nucleotides and two amino acids and those of *Rhynchopus sp.* 3 vary at several nucleotides but not at any amino acid. These differences indicate that two different alpha-tubulin genes were sequenced from

```
 1 T.cruzi       MR-EAICIHIGQAGC QVGNACWELFCLEHG IQPDGAMPSDKTIGV EDDAFNTFFSETGAG KHVPRAVFLDLEPTV VDEIRTGTYRQLFHP    89
 2 E.gracilis    MR-EIISIHLGQGGI QIGNACWELYCLEHG IQPDGSMPSDKAIGV EDDAFNTFFSETGAG KHVPRAVFLDLEPSV VDEVRTGTYRQLFHP    89
 3 D.papillatum  --------------- --------LYCLEHG IQPDGAMPSDKTIGI EDDAYNTFFSETGAG KHVPRAVFLDLEPTV IDEVRTGTYRQLFHP    67
 4 D.sp.3        --------------- --------LYCLEHG IQPDGALPSDKTIGI EDDAFSTFFSETGSG KHVPRAVLLDLEPTV IDEVRTGTYRQLFHP    67
 5 D.sp.3new C1  --------------- --------LYCLEHG IQPDGALPSDKTIGI EDDAFNTFFSETGSG KHVPRAVLLDLEPTV IDEVRTGTYRQLFHP    67
 6 D.sp.3new C2  --------------- --------LYCLEHG IQPDGALPSDKTIGI EDDAFNTFFSETGSG KHVPRAVLLDLEPTV IDEVRTGTYRQLFHP    67
 7 D.sp.2        --------------- --------LYCLEHG IQPDGAMPSDKTIGI EDDAFNTFFSETGSG KHVPRAVFLDLEPTV IDEVRTGTYRQLFHP    67
 8 D.sp.4        --------------- --------LYCLEHG IQPDGAMPSDKTIGI EDDAFNTFFSETGAG KHVPRAVFLDLEPTV IDEVRTGTYRQLFHP    67
 9 R.sp.3 C1     --------------- --------LYCLEHG IQPDGAMPSDKTIGI EDDAYNTFFSETGAG KHVPRAVMLDLEPTV IDEVRTGTYRQLFHP    67
10 R.sp.3 C2     --------------- --------LYCLEHG IQPDGAMPSDKTIGI EDDAYNTFFSETGAG KHVPRAVMLDLEPTV IDEVRTGTYRQLFHP    67
11 R.sp.1        --------------- --------LYCLEHG IQPDGAMPSDKTCGV EDDAFNTFFSETGAG KHVPRAVLLDLEPTV IDEVRTGTYRQLFHP    67
12 R.sp.2        --------------- --------LYCLEHG IQPDGAMPSDKTCGV EDDAFNTFFSETGAG KHVPRAVMLDLEPTV IDEVRTGTYRQLFHP    67

 1 T.cruzi       EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLGHNCTGLQG FLVYHAVGAGTGSGL GALLLERLSVDYGKK SKLGYTVYPSPQVST   179
 2 E.gracilis    EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLAFHAVGGGTGSGL GRLLLERLSVDYGKK SKLGFTIYPSPQIST   179
 3 D.papillatum  EQLISGKEDAANNYA RGHYTIGKEMVDLCL DRIRKLADNCTGLZG FLVFNSVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
 4 D.sp.3        EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNSVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
 5 D.sp.3new C1  EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNSVGGGTGSGL GALLLERLSVDCGRK SKLGFTVYPSPQVST   157
 6 D.sp.3new C2  EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNSVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
 7 D.sp.2        EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNSVGGGTGSGL GALLLERLSVDYGRK SKLGFTIYPSPQVST   157
 8 D.sp.4        EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNSVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
 9 R.sp.3 C1     EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNAVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
10 R.sp.3 C2     EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNAVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
11 R.sp.1        EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNAVGGGTGSGL GALLLERLSVDYGRK SKLGFTVYPSPQVST   157
12 R.sp.2        EQLISGKEDAANNYA RGHYTIGKEIVDLCL DRIRKLADNCTGLQG FLVFNSVGGGTGSGL GALLLERLSVDYGRK SKLGLRCTRRRRCRR   157

 1 T.cruzi       AVVEPYNSVLSTHSL LEHTDVAAMLDNEAI YDLTRRNLDIERPTY TNLNRLIGQVVSALT ASLRFDGALNVDLTE FQTNLVPYPRIHFVL   269
 2 E.gracilis    AVVEPYNSVLSTHSL LEHTDVAVMLDNEAI YDICRRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDITE FQTNLVPYPRIHFVL   269
 3 D.papillatum  AVVEPYNSVLSTHSL LEHTDVAVMLDNEAI YDICRRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
 4 D.sp.3        AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
 5 D.sp.3new C1  AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
 6 D.sp.3new C2  AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
 7 D.sp.2        AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDICRRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
 8 D.sp.4        AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
 9 R.sp.3 C1     AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLISQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
10 R.sp.3 C2     AVVEPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLISQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
11 R.sp.1        AVVGPYNSVLSTHSL LEHTDVACMLDNEAI YDIARRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   247
12 R.sp.2        PWLSRTAPGGD???? ??????????????? YDIARRNLDIERPTY TNLNRLIAQVISSLT ASLRFDGALNVDVTE FQTNLVPYPRIHFML   228
```

```
 1 T.cruzi      TSYAPVISAEKAYHE QLSVSEISNAVFEPA SMMTKCDPRHGKYMA CCLMYRGDVVPKDVN AAVATIKTKRTIQFV DWSPTGFKCGINYQP   359
 2 E.gracilis   SSYAPIISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWCPTGFKCGINYQP   359
 3 D.papillatum SSFAPVISAEKAYHE QLSVAEITNSVFEPA AMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
 4 D.sp.3       SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPHHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
 5 D.sp.3new C1 SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
 6 D.sp.3new C2 SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGRYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
 7 D.sp.2       SSYAPVISAEKAYHE QLFVAEITNAAFEPA SSMVKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
 8 D.sp.4       SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
 9 R.sp.3 C1    SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
10 R.sp.3 C2    SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
11 R.sp.1       SSYAPVISAEKAYHE QLPVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   337
12 R.sp.2       SSYAPVISAEKAYHE QLSVAEITNAAFEPA SMMAKCDPRHGKYMA CCLMYRGDVVPKDVN ASVATIKTKRTIQFV DWSPTGFKCGINYQP   318

 1 T.cruzi      PTVVPGGDLAKVQRA VSMIANSTAIAEVFA RIDHKFDLMYSKRAF VHWYVGEGMEEGEFS EAREEY--------- ---------------   425
 2 E.gracilis   PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VHWYVGEGMEEGEFS EAREDLAALEKDYEE VGAESADVEGEEDVE   449
 3 D.papillatum PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
 4 D.sp.3       PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
 5 D.sp.3new C1 PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
 6 D.sp.3new C2 PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
 7 D.sp.2       PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
 8 D.sp.4       PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
 9 R.sp.3 C1    PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
10 R.sp.3 C2    PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAF VH------------- --------------- ---------------   384
11 R.sp.1       PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAL VH------------- --------------- ---------------   384
12 R.sp.2       PTVVPGGDLAKVQRA VCMISNSTAIAEVFA RIDHKFDLMYSKRAL VH------------- --------------- ---------------   365

 1 T.cruzi       --    425
 2 E.gracilis    EY    451
 3 D.papillatum  --    384
 4 D.sp.3        --    384
 5 D.sp.3new C1  --    384
 6 D.sp.3new C2  --    384
 7 D.sp.2        --    384
 8 D.sp.4        --    384
 9 R.sp.3 C1     --    384
10 R.sp.3 C2     --    384
11 R.sp.1        --    384
12 R.sp.2        --    365
```

Fig. 5 An alignment of twelve amino-acid sequences of alpha-tubulin. Amino acid sequences were aligned to obtain maximal similarity. Dashes indicate the absence of amino acids at the corresponding positions. '?' in Rhynchopus sp.2 indicates unavailable amino acid sequences at the corresponding positions. The ten amino acid sequences of diplonemids were obtained from this study. The amino acid sequences are as follows: *Trypanosoma cruzi*, *Euglena gracilis*, *Diplonema papillatum*, *Diplonema sp.* 3, , *Diplonema sp.* 3 (new, copy 1), *Diplonema sp.* 3 (new, copy 2), *Diplonema sp.* 2, *Diplonema sp.* 4, *Rhynchopus sp.* 3 (copy 1), *Rhynchopus sp.* 3 (copy 2), *Rhynchopus sp.*1, *Rhynchopus sp.* 2. Bold characters in *Diplonema sp.* 4 and *Rhynchopus sp.* 2 indicate the positions of introns found in this study.

these two diplonemids, but only one gene was sequenced from each of the remaining six diplonemids (see Table 1). The sequence of the PCR product was 1153 nt in length (excluding primer and intron sequences) for each of the nine alpha-tubulin sequences, recovering more than 80% of a complete intronless alpha-tubulin sequence. However, for the alpha-tubulin sequence of *Rhynchopus sp.* 2, I was unable to sequence about 120 nt (see Fig. 5).

The inferred translation of 384 amino acids for each of the ten alpha-tubulins from diplonemids (with no primer sequences) was aligned with those from a sampling of 54 other eukaryotic taxa. Figure 5 shows a small sampling of this alignment, including the ten diplonemid alpha-tubulin sequences as well as those of *Euglena gracilis* and *Trypanosoma cruzi*. The sequence differences among the ten diplonemid sequences were slight (only 0-15 amino acid differences over the total length of 384 amino acids). The sequence differences between diplonemids and *Euglena gracilis* were 18 to 39 amino acids. Similarly, the sequence differences between diplonemids and kinetoplastids were 23 to 44 amino acids.

Beta-tubulin gene sequences

PCR products of the expected size (1200 nt) were obtained from *Diplonema sp.* 2, *Diplonema sp.* 3, *Diplonema sp.* 3 (new strain), *Diplonema ambulator, Rhynchopus sp.* 1, *Rhynchopus sp.* 2 and *Rhynchopus sp.* 3. PCR products of larger than the expected size were obtained from *Diplonema sp.* 4 and *Diplonema sp.* 2. All the above PCR products were cloned and both strands of several clones from each were sequenced. Again, BLAST searches of these sequences confirmed that each of them was beta-tubulin, and corresponded to about 86% (387-388 amino acids) of a full-length beta-tubulin gene. In addition, the sequences from three independent clones from *Diplonema sp.* 4 showed that each were slightly different copies of beta-tubulin (1-2 amino acid differences). The sequences from

```
 1 R.sp.3 C1        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 2 R.sp.3 C2        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 3 R.sp.2           --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 4 D.sp.3 C1        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 5 D.sp.3 C2        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 6 D.sp.3           --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 7 D.sp.4 C1        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 8 D.sp.4 C2        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
 9 D.sp.4 C3        --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
10 D.ambulator C1   --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
11 R.sp.1           --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYLNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
12 D.ambulator C2   --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
13 D.sp.2           --------------- IGSKFWEVVSDEHGV DPTGTYQGDSDLQLE RINVYFNEATGGRYV PRAVLMDLEPGTMDS VRAGPYGQIFRPDNF    75
14 T.brucei         MREIVCVQAGQCGNQ IGSKFWEVISDEHGV DPTGTYQGDSDLQLE RINVYFDEATGGRYV PRSVLIDLEPGTMDS VRAGPYGQIFRPDNF    90
15 E.gracilis       --------------- --------------- --------------- --------------- -------LEPGTMDS VRAGPYGQIFRPDNF    23

 1 R.sp.3 C1        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRIMM TFSVIPSPKVSDTVV   165
 2 R.sp.3 C2        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRIMM TFSVIPSPKVSDTVV   165
 3 R.sp.2           VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRIMM TFSVIPSPKVSDTVV   165
 4 D.sp.3 C1        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLRGSSC RTPWVGGTGSGMGTL LISKLREEYPDRMMI TFSVIPSPKVSDTVV   165
 5 D.sp.3 C2        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLRGSSC RTPWVGGTGSGMGTL LISKLREEYPDRMMI TFSVIPSPKVSDTVV   165
 6 D.sp.3           VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRMMI TFSVIPSPKVSDTVV   165
 7 D.sp.4 C1        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRIMV TFSVIPSPKVSDTVV   165
 8 D.sp.4 C2        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRIMV TFSVIPSPKVSDTVV   165
 9 D.sp.4 C3        VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRIMV TFSVIPSPKVSDTVV   165
10 D.ambulator C1   VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL AHSLGGGTGSGMGTL LISKLREEYPDRIMM TFSVIPSPKVSDTVV   165
11 R.sp.1           VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL AHSLGGGTGSGMGTL LISKLREEYPDRVMM TFSVIPSPKVSDTVV   165
12 D.ambulator C2   VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL AHSLGGGTGSGMGTL LISKLREEYPDRIMM TFSVIPSPKVSDTVV   165
13 D.sp.2           VFGQTGAGNNWAKGH YTEGAELIDSVLDVC RKEAESCDCLQGFQL SHSLGGGTGSGMGTL LISKLREEYPDRMML TFSVIPSPKVSDTVV   165
14 T.brucei         IFGQSGAGNNWAKGH YTEGAELIDSVLDVC CKEAESCDCLQGFQI CHSLGGGTGSGMGTL LISKLREQYPDRIMM TFSIIPSPKVSDTVV   180
15 E.gracilis       VFGQTGAGNNWAKGH YTEGPELIDSVLDVV RKEAESCDCLQGFQI AHSLGGGTGSGMGTL LISKIREEYPDRMMM TFSVIPSPKVSDTVV   113

 1 R.sp.3 C1        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 2 R.sp.3 C2        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 3 R.sp.2           EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 4 D.sp.3 C1        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 5 D.sp.3 C2        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 6 D.sp.3           EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 7 D.sp.4 C1        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 8 D.sp.4 C2        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
 9 D.sp.4 C3        EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
10 D.ambulator C1   EPYNATLSIHQLVEN ADESVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
11 R.sp.1           EPYDATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
12 D.ambulator C2   EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNADLRKLAV NLIPFPRLHFFLVGF   255
13 D.sp.2           EPYNATLSIHQLVEN ADECVMIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVMSGVTCCL RFPGQLNSDLRKLAV NLIPFPRLHFFLVGF   255
14 T.brucei         EPYNTTLSVHQLVEN SDESMCIDNEALYDI CFRTLKLTTPTFGDL NHLVSAVVSGVTCCL RFPGQLNSDLRKLAV NLVPFPRLHFFMMGF   270
15 E.gracilis       EPYNTTLSVHQLVEN ADEVMCIGNEALYDI CLPTLKLTTPTFG-H ETLVSAVMSGVTCCL RFPGQLNSDLRKLAV NLIPFPRLHFFLVGF   202
```

```
 1 R.sp.3 C1       APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ LFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 2 R.sp.3 C2       APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ LFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 3 R.sp.2          APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ LFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 4 D.sp.3 C1       APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 5 D.sp.3 C2       APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 6 D.sp.3          APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 7 D.sp.4 C1       APLTSRGSQQYRALT VPELTQQSFDAKDMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
 8 D.sp.4 C2       APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSISDIPPKG   345
 9 D.sp.4 C3       APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
10 D.ambulator C1  APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSAICDIPPKG   345
11 R.sp.1          APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVZNKNSSYFVEWI PNNIKSSICDIPPKG   345
12 D.ambulator C2  APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTACQ MFRGRISTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSICDIPPKG   345
13 D.sp.2          APLTSRGSQQYRALT VPELTQQSFDAKNMM CASDPRHGRYLTASQ LFRGRISTKEVDEQM LNVQNKSSSYFVEWI PNNIKSSVCDIPPKG   345
14 T.brucei        APLTSRGSQQYRGLS VPELTQQMFDAKNMM QAADPRHGRYLTASA LFRGRMSTKEVDEQM LNVQNKNSSYFIEWI PNNIKSSVCDIPPKG   360
15 E.gracilis      APLTSRGSQQYRALT VPELTQQMFDAKNMM AASDPAHGRYLTASA MFRGRMSTKEVDEQM LNVQNKNSSYFVEWI PNNIKSSVCDIPPKG   292

 1 R.sp.3 C1       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 2 R.sp.3 C2       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 3 R.sp.2          LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 4 D.sp.3 C1       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 5 D.sp.3 C2       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 6 D.sp.3          LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 7 D.sp.4 C1       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 8 D.sp.4 C2       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
 9 D.sp.4 C3       LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
10 D.ambulator C1  LKMSTCFIGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
11 R.sp.1          LKMSTCFIGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
12 D.ambulator C2  LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEG--- --------------- --------------- ---------------   387
13 D.sp.2          LKMSTCFVGNNTCIQ EMFKRVSEQFTAMFR RKAFLHWYTGEGM-- --------------- --------------- ---------------   388
14 T.brucei        LKMAVTFIGNNTCIQ EMFRRVGEQFTLMFR RKAFLHWYTGEGMDE MEFTEAESNMNDLVS EYQQYQDATIEEEGE FDEEEQY--------   442
15 E.gracilis      LKMSATFIGNNTAIQ EMFKRVSEQFTAMFR RKAFLHWYTGEGMDE MEFTEAESNMNDLVS EYQQYQDATVEEEGE FDEEEDVEQY-----   377
```

Fig. 6 An alignment of fifteen amino-acid sequences of beta-tubulin. Amino acid sequences were aligned to obtain maximal similarity. Dashes indicate the absence of amino acids at the corresponding positions. The thirteen amino acid sequences of diplonemids were obtained from this study. The amino acid sequences are as follows: *Rhynchopus sp.* 3 (copy 1), *Rhynchopus sp.* 3 (copy 2), *Rhynchopus sp.* 2, *Diplonema sp.* 3 (new, copy 1), *Diplonema sp.* 3 (new, copy 2), *Diplonema sp.* 3, *Diplonema sp.* 4 (copy 1), *Diplonema sp.* 4 (copy 2), *Diplonema sp.* 4 (copy 3), *Diplonema ambulator* (copy 1), *Rhynchopus sp.* 1, *Diplonema ambulator* (copy 2), *Diplonema sp.* 2, *Trypanosoma brucei, Euglena gracilis.* Bold characters in *Diplonema sp.* 2 and *Diplonema sp.* 4 indicate the positions where introns were found.

two independent clones from *Diplonema ambulator* also represented two different beta-tubulin sequences (3 amino acid differences). The sequences from two independent clones from *Diplonema sp.* 3 (new strain), and *Rhynchopus sp.* 3, respectively, revealed two different beta-tubulin gene sequences, however the differences were only detectable at the nucleotide level, not at the amino acid level. From each of the rest four diplonemids, the two independent sequences from two independent clones are identical (Table 1). One intron was found in each of the three different beta-tubulin sequences from *Diplonema sp.* 4 (140 nt, 126 nt and 149 nt in length, respectively) and in the beta-tubulin gene of *Diplonema sp.* 2 (71 nt).

The inferred amino acids for each of the thirteen beta-tubulins from diplonemids (with no primer sequences) were aligned with those from 46 other eukaryotic taxa. Figure 6 shows a small part of this alignment, which includes not only the thirteen new diplonemid beta-tubulin sequences, but also the sequences from *Euglena gracilis* and *Trypanosoma brucei*. All of the thirteen diplonemid-sequences were very similar to each other (0-15 amino acid differences over a total of 387 residues). The sequence difference between diplonemids and *Euglena gracilis* (23-44 amino acids) was similar to the sequence difference between diplonemids and kinetoplastids (39-54 amino acids).

GAPDH gene sequences

PCR products of the expected size (around 1000 nt) were obtained from *Diplonema sp.* 3 and *Rhynchopus sp.* 3. PCR products of both expected size and larger than expected size were obtained from *Diplonema sp.* 2. All these PCR products were cloned and both strands of several clones were sequenced from each. BLAST searches of these sequences confirmed that they all encoded GAPDH, recovering over 90% of a full-length GAPDH gene. In addition, the sequence of the larger PCR product from *Diplonema sp.* 2 contained two intron sequences (77 nt and 120 nt in length).

The deduced amino-acid sequences of the four diplonemid GAPDH genes (with no PCR primer sequences) were aligned with those from 96 other taxa, including both eukaryotes and prokaryotes. During the aligning process, I found it was difficult to align the four newly obtained GAPDH sequences from diplonemids with those of any euglenozoa. However, it was comparatively easier to align the three intron-lacking GAPDH sequences (from *Diplonema sp.* 3, *Rhynchopus sp.* 3 and *Diplonema sp.* 2) with *Anabaena variabilis* gap3 (I will refer to these three diplonemid sequences as "gap1"), and the second *Diplonema sp.* 2 GAPDH with the GapC of cryptomonads (I will refer to this diplonemid sequence as "gap2"). Figure 7 is a representation of this alignment. Comparison among the four diplonemid GAPDH sequences reveals that *Diplonema sp.* 3, *Rhynchopus sp.* 3 and *Diplonema sp.* 2 (gap1) are far more similar to each other than to *Diplonema sp.* 2 (gap2). The sequence differences between *Diplonema sp.* 3, *Rhynchopus sp.* 3 and *Diplonema sp.* 2 (gap1) are 81-93 amino acids over a total length of 301-316 amino acids, whereas the sequence differences between these three and *Diplonema sp.* 2 gap2 are 187-196 amino acids. Further pairwise sequence comparisons showed that the sequence differences between *Diplonema sp.*3, *Rhynchopus sp.* 3, *Diplonema sp.*2 (gap1) and *Anabaena variabilis* gap3 are only 148-162 amino acids, whereas those between these three diplonemid GAPDH sequences and those of *E. gracilis* GapC, *T. bruci* GapC, and *L. mexicana* GapC were 195-207 amino acids. Meanwhile, the sequence differences between *Diplonema sp.* 2 (gap2) and the GapC of two cryptomonads (*Pyrenomonas salina* GapC and the *Guillardia theta* GapC) were only 115-118 amino acids, whereas those between *Diplonema sp.*2 (gap2) and the *E. gracilis* GapC, *T. bruci* GapC, and *L. mexicana* GapC were 129-152 amino acids. In

```
                                                     ↓
1 1R.sp.3       -----------RMGR LSFRLAWDMP--ELE IVHVNEIIGGSVHAA YLVKFDSVHGTWEKE CEDG--EDGKYFTVA GKRVGFSEEKDFRAV   75
2 1D.sp.2       --------------- ----LLGICL--RSK VVHINEIIGGCEHAA YLVKFDSVHGTWEKE CEPG--NDKKHIKVE GKRVNFTDKKDFTQV   67
3 1D.sp.3       -----------RMGR LACRMAWDMD--EVE IVHVNEIIGGSAHAA YLLQFDSVHGTWGHE CVDADSEEDPHFTVD GKKITFSDEKDFTKV   79
4 3A.variabilis MKIRVGINGFGRMGR LALRAAWGWP--ELE FVHINEIKGGAVAAA HLLKFDSVHGRWTPE VEAE—GERVLID-- STPLSFSEYGKPEDV   84
5 G.theta       MPTKIGINGFGRIGR LVTRAAFSNKKADCQ VVAVNDPFIDLDYMV YMFKYDSTHGIWKGS CEKK—GDKLVID-- GCEITCFTERDPTAI   86
6 2D.sp.2       -----------RIGR LVLRAALQN--PQAT VVAVNDPFISVDYMV YMFKYDSVHGRYNGK VEAR—NGKFIVD-- GVEITVFTQKDPSSI   75

1 1R.sp.3        DWKGAGVDIVLDCTG KFLKTEVLQDYITKC GVKKVVVSAPVKEPS VVNIVMGCNDDKVTR DHHIVTAASCTTNCI GPIIKVIHENLGIET   165
2 1D.sp.2        DWKGSGCDIVVDSSG KFTKTDKLKPYLTQC GMKKVVVSAPVKEPS VLNIVVGVNDQKLTA SHDIITAASCTTNCL APVIKVMHENLGIES   157
3 1D.sp.3        DWKGKGIDIVLECSG KFTTTKKLQPYLSEC GVKKVVVSAPVKEPE VLNIVVGVNDDKLEA KHDIVTAASCTTNCL APIVKVIHENLTIES   169
4 3A.variabilis  PWEDFGVDLVLECSG KFRTPATLDPYFKRG -VQKVIVAAPVKE-E ALNIVMGVNDYLYEPEKHHLLTAASCTTNCL APVVKVIHEGLGIKH   173
5 G.theta        KWAAAGAIYVVESTG VFTTIDKAQAHLSGG -AKKVVISAPSA--D APMFVMGVNNKAYDG TPTIVSNASCTTNCL APLAKVVERNFGIVE   173
6 2D.sp.2        PWGEAGAEVVVESTG VFTKIESASAHLTGG -AKRVVISAPSP--D APMFVVGVNEEKFNP SMKVISNASCTTNCL APLAKVLNDNFGIVE   162

1 1R.sp.3        GMITTVHDVTGTQSL VDMVNTKKNDLRRSR SGMLNLAPTSTGSAT AICEVFPELKGKLNG HAIRVPLLNASITDI VLNVGKDTSAEEVNA   255
2 1D.sp.2        GMITTVHNITATQSL VDLVQTKKNDMRRSR SGMLNLAPTSTGSAT AIAEVFPELKGKLNG HAIRVPLLNGSITDI VLNVKKETSVAEVNK   247
3 1D.sp.3        GMITTVHNITGTQSL VDMVNTKKNDLRRSR SGMLNLCPTSTGSAT AIAEVFPELKGKLDG LAIRVPLLNSSITDM VFNVSRGTTKAEVNA   259
4 3A.variabilis  GIITTIHDNTNTQTL VDAP-HK--DLRRAR ATSLSLIPTTTGSAK AIALIYPELKGKLNG IAVRVPLLNASLTDC VFEVNRPTTVEEINA   260
5 G.theta        GLMSTVHATTATQKT VDGPSGK--DWRGGR GAAQNIIPSATGAAK AVGKVLPELNGKLTG MAFRVPTPDVSVVDL TVKLAKPASYQQICD   261
6 2D.sp.2        GLMTTVHAATATQKT VDAPSKK--DWRGGR GILGNIIPSSTGAAK AVGKVIPELNGKLTG MAFRVPTADVSVVDL TVRLRKGASKKDIDA   250

1 1R.sp.3        LLQDAAAEGPLAATS EHGAILGFETRPLVA TDYTNDKRSTIVDAP STMVVAKRMVKVYA- --------------- ---------   314
2 1D.sp.2        LLEAASSSGPLSANS EHGSILGYETRPLVS TDYTNDKRSSIIDAP STMVVDKKMVKLY-- --------------- ---------   305
3 1D.sp.3        LLEKAAQEGPLAGTA EHGSILGYEARPLVS TDFTNDERSSIVDAP STLVVGDKMVKIYA- --------------- ---------   318
4 3A.variabilis  LLKAASEQAPL---- --QGILGYEERPLVS IDYKDDPRSSIIDAL STMVVDETQVKILAW YDNEWGYVNRMVELA RKVALSLK-   337
5 G.theta        AIKAATTDPEY---- --CGVIAYTDDEVVS TDFLGNSYSSIFDAK AGIALNDTFVKLVSW YDNEWGYSNRVVDLI AHMATVDKF   339
6 2D.sp.2        AVLKASQSGKM---- --AGVIGFTNEDVVS TDFIGDTRSSIYDSK ASICLNDNFVKLVP- --------------- ---------   303
```

Fig. 7 An alignment of six amino-acid sequences of GAPDH. Amino acid sequences were aligned to obtain maximal similarity. Dashes indicate the absence of amino acids at corresponding positions. The four amino acid sequences of diplonemids were obtained from this study. The amino acid sequences are as follows: *Rhynchopus sp.* 3 gap1, *Diplonema sp.* 2 gap1, *Diplonema sp.* 3 gap1, *Anabaena variabilis* gap3, *Guillardia theta* gapC, *Diplonema sp.* 2 gap2. Bold characters in *Diplonema sp.*2 gap2 indicate the positions of introns found in this study. The five unusual insersions in the GAPDH sequences of *Rhynchopus sp.* 3 gap1, *Diplonema sp.*2 gap1, *Diplonema sp.*3 gap1 were shaded in gray. The arrow points at the position 32 of the GAPDH sequences, according to the numbering of the amino acid sequence of GAPDH by Biesecker et al. (1977). This position will be discussed later.

addition, *Diplonema sp.* 3, *Rhynchopus sp.* 3 and *Diplonema sp.* 2 (gap1) share five unusual insertions (see Fig. 7), which are not found in either the gap2 of *Diplonema sp.* 2, or in the GAPDH from any other members of the Euglenozoa, *Anabaena variabilis* gap3, GapC of cryptomonads, or in the GAPDH genes from most other prokaryotes and eukaryotes.

In summary, all the diplonemid actin, alpha- and beta-tubulin sequences are very similar to each other. Diplonemid actin sequences show more resemblance to euglenoid sequences than to kinetoplastid sequences, and diplonemid alpha- and beta- tubulin sequences are nearly as similar to those from euglenoids as to those from kinetoplastids. On the other hand, among the four newly obtained diplonemid-GAPDH sequences, three are very similar to each other but very different to the remaining one sequence. In addition, none of the four is particularly similar to any of the euglenozoa GAPDH sequence, instead, three are more similar to the gap3 of cyanobacteria and one is more similar to the GapC of cryptomonads.

3.2 Diplonemid introns

Position of diplonemid introns

As mentioned previously, introns were found in several of the PCR products. They were tentatively identified by insertions in genes that couldn't be aligned to the amino acid sequences of the same gene from other organisms by BLAST searching. After the contigs were complete, introns were confirmed by the presence of the canonical GT-AG cleavage-sites (see Materials and Methods). They were found in all three phases (if an intron is found between two codons, then it is termed as a phase 0 intron; if an intron is found between the first nucleotide and the second nucleotide of one codon, then it is termed as a phase 1 intron; if an intron is found between the second nucleotide and the third nucleotide of one codon, then it is termed as a phase 2 intron).

I have characterized 11 introns in nine of the 29 nuclear-encoded genes from diplonemids. The amino acids, where the corresponding introns occur, are highlighted in bold in the alignments (Fig. 4-Fig. 7). I found two introns in the actin gene from *Diplonema sp.* 3 and one intron from *Diplonema ambulator*. These three introns are all phase 0 introns (Fig. 4). There is one intron in each of the alpha-tubulin genes from *Diplonema sp.* 4 and *Rhynchopus sp.* 2, respectively. In *Diplonema sp.* 4, the intron is a phase 1 intron whereas in *Rhynchopus sp.* 2, the intron is a phase 0 intron in a different position (Fig. 5). I found one intron in the beta-tubulin gene of *Diplonema sp.* 2. It is a phase one intron. In *Diplonema sp.* 4, there are three different introns in three different copies of the gene. All the three introns are in the same position and same phase (phase 2) (Fig. 6). I found two introns in one copy of the GAPDH gene from *Diplonema sp.* 2 (gap2). They are both phase zero introns.

Diplonemid intron characterization

The 5' and 3' ends of the 11 introns were aligned (Fig. 8). All have the consensus GT and AG boundaries as expected of canonical spliceosomal introns. The highly conserved six nucleotides at the 5' splice sites of the diplonemid introns are GTRTGY, which also closely correspond to the six conserved nucleotides at the 5' splice site of a classical GT-AG mammalian intron: the only difference lies in the fourth position, where a T-residue is present in diplonemid-introns rather than an A-residue as in mammalian introns. All eleven diplonemid introns end with CAG consensus nucleotides, just as classical spliceosomal introns do. Interestingly, however, the twelve nucleotides preceding the final CAG are mostly C or A in all eleven diplonemid introns (see Fig. 8). In ten of the eleven-diplonemid introns (except for the intron in the alpha-tubulin gene of *Rhynchopus sp.* 2), A-residues are present at least five times each and T-residues are

```
Actin-D3-1  gtatgtggcggggc......( 80nt ).....acagcaccaacacag
Actin-D3-2  gtgtgtggcccccg......(176nt ).....cacaccaccacacag
Actin-Da    gtatgccattttact......( 40nt ).....tacgaaaccacctag
gap2-d2-1   gtatgtagttattga......( 77nt ).....tccaacaaatcacag
gap2-d2-2   gtgtgtttactttt......(120nt ).....aataacaacaaacag
Alpha-d4    gtatgctacactaac......(109nt ).....gtaaaacacacacag
Alpha-rh2   gtgtgtttgttggtc......(126nt ).....ttttcacacccacag
Beta-d2     gtcatgaattgattt......( 71nt ).....caaacaaccaaacag
Beta-d4-C1  gtatgttaacttctt......(140nt ).....aacccacacacacag
Beta-d4n-C2 gtatgttgacttttt......(126nt ).....cacacacacacacag
Beta-d4n-C3 gtatgttgacttttt......(149nt ).....cacacacacacacag

CONSERVED   GTATGT...............................ACACACACACACAG
              G C
CLASSICAL   GTAAGT..............................TTTTTTTTTTTNCAG
              G C                               CCCCCCCCCCC
```

Fig. 8 Alignment of eleven-diplonemid introns from 5'-end to 3'-end. They are: intron1 (closest to the 5' end of the gene) in the actin gene of *Diplonema sp.* 3; intron 2 (second close to the 5' end of the gene) in the same actin gene of *Diplonema sp.* 3; one intron from the actin gene of *Diplonema ambulator*; intron 1 in the gap2 of *Diplonema sp.* 2; intron 2 in the gap2 of *Diplonema sp.* 2; one intron in the alpha-tubulin gene of *Diplonema sp.* 4; one intron in the alpha-tubulin gene of *Rhynchopus sp.* 2; one intron in the beta-tubulin gene of *Diplonema sp.* 2; one intron in a first copy of beta-tubulin gene of *Diplonema sp.* 4; one intron in a second copy of beta-tubulin gene of *Diplonema sp.* 4; one intron in a third copy of beta-tubulin gene of *Diplonema sp.* 4. CONSERVED shows the majority consensus sequences of the eleven introns. CLASSICAL shows the strictly conserved sequences of the classical 'GT-AG' spliceosomal introns from higher eukaryotes (mammals, in particular). Numbers in the brackets indicate the entire lengths of the eleven introns.

either completely absent or present only once or twice each. The C-residues are also present around five times each on average. This contrasts with the classical GT-AG mammalian introns, which contain a polypyrimidine tract in this region. Moreover, in the introns of the alpha- and beta-tubulin genes from *Diplonema sp.* 4, continuous 'CA' repeats were observed.

The branchpoint region in a classical GT-AG intron is usually closer to the 3' splice site than to the 5' splice site. More specifically, this region generally appears 15-40 nt upstream of the 3' splice site (Umen et al. 1995). In yeast, this branchpoint consensus sequence is strictly conserved, that is 5'-TACTAACA-3' (Umen et al. 1995). In contrast, this branchpoint region is loosely conserved in the introns of mammals: 5'-YNYTRACN-3' (Umen et al. 1995). The branchpoint consensus sequence from yeast introns was not observed in any of the eleven diplonemid-introns, but the branchpoint consensus sequence from mammalian introns was observed six times in five of the eleven diplonemid-introns. However, it was observed four times in four different introns at either position +4 (referring to the 5' cleavage site) to +11 (three introns in three different copies of *Diplonema sp.* 4 beta-tubulin genes) or, at position +9 to +16 (one intron in the alpha-tubulin gene of *Diplonema sp.* 4). Both regions are highly unlikely to be real branchpoint sites since they are too close to or even overlapping with the 5' consensus splice sites of the introns. This branchpoint sequence of mammalian introns was also observed between position -23 (referring to the 3' cleavage site) to -16 (TGTTGACT) in the intron from *Diplonema sp.* 4 alpha-tubulin gene, and between position -36 and -29 (TCCTGACC) in the first intron (closest to the 5' end of the gene) in the *Diplonema sp.* 2 gap2.

3.3 Phylogeny of the Euglenozoa

In addition to looking for introns in diplonemids, I constructed three protein phylogenetic trees with the newly obtained diplonemid sequences in an attempt to determine the phylogenetic position of diplonemids within the phylum Euglenozoa.

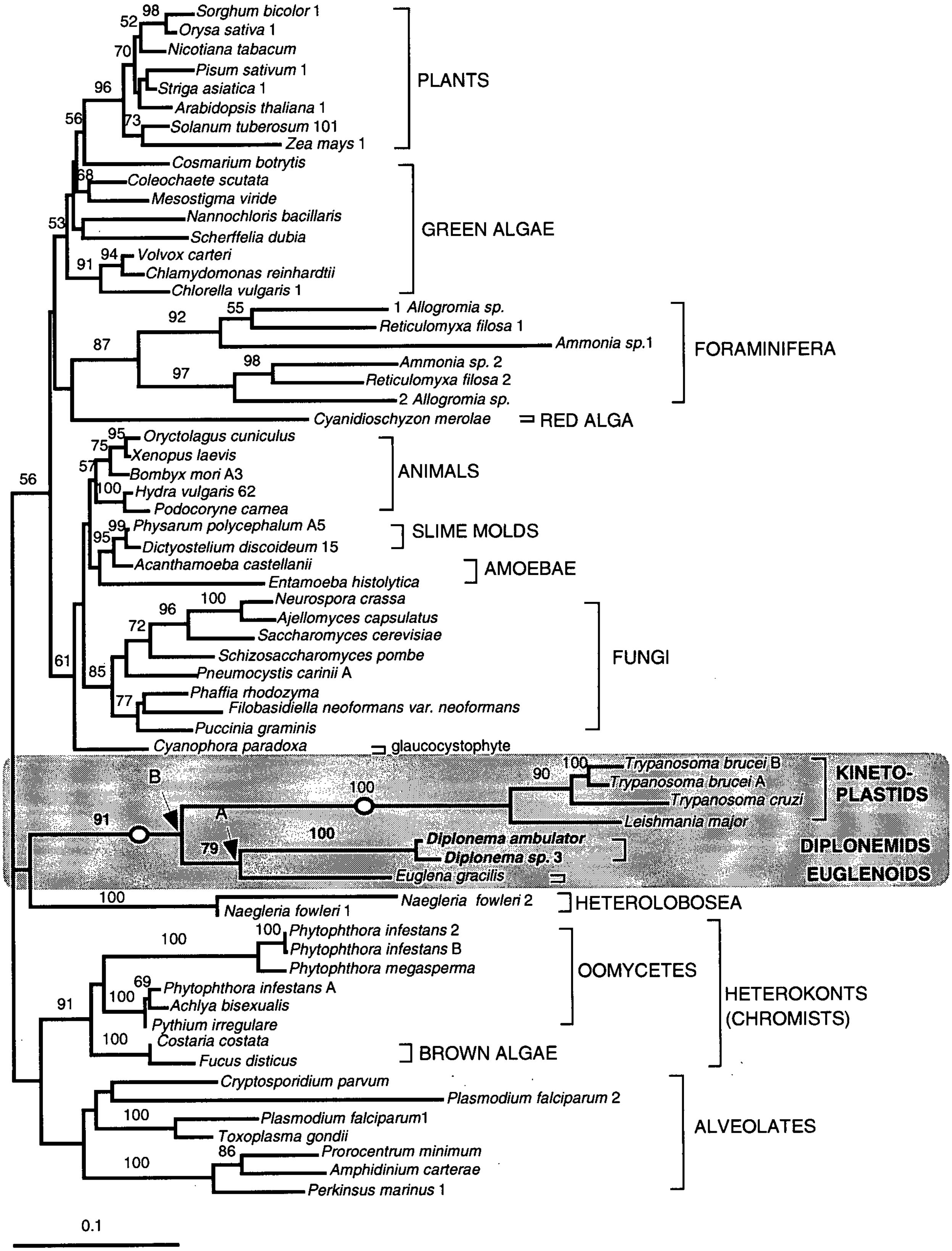
Actin phylogeny

An actin phylogeny was constructed from 373 alignable characters from a total of 65 eukaryotic taxa using distance and neighbor joining methods (Figure 9).

Most of the phylogenetically distinct eukaryotic groups including land plants, green algae, animals, fungi, heterokonts, and alvolates are recovered in the actin tree (Fig. 9). The two new diplonemid sequences are closely related to each other and form a clade with 100% bootstrap support. In fact, the whole phylum Euglenozoa (shaded in Fig. 9), consisting of three major groups (diplonemids, euglenoids and kinetoplastids), is well supported by this tree (91% bootstrap value). Furthermore, this actin tree also strongly suggests that the two diplonemid sequences are more closely related to the euglenoid sequences than to the kinetoplastid sequences. The node uniting diplonemids with euglenoids (at the exclusion of kinetoplastids, node A in Fig. 9) is well supported by bootstrap (79%).

In an effort to test the likelihood of alternative positions for diplonemids within the phylum Euglenozoa, Kishino-Hasegawa tests were carried out on the actin data. In this case, I tested two alternative positions for the diplonemids. In one alternative, the diplonemids branch with the kinetoplastids, so the internal topology of the phylum became ((diplonemids, kinetoplastids), euglenoids) or ((D, K), E). The other possible position of diplonemids is at

Fig. 9 Neighbor-joining tree based on actin protein sequences of various eukaryotes, including two new sequences of diplonemids (in bold). This unrooted BioNJ tree was constructed by calculating maximum-likelihood (ML) distances between pairs of sequences. Values on selected branches indicate neighbor-joining bootstrap support greater than 50% and the bootstrap values of particular interest are in bold. Scale bar indicates amino acid substitutions per site. Node A is the last common ancestor of diplonemid and euglenoid sequences and node B is the last common ancestor of the phylum Euglenozoa (shaded). Alternative positions for the diplonemids were assessed with Kishino-Hasegawa tests at the nodes marked with open circles. The two alternatives were not rejected at 5% levels.


98
52
Sorghum bicolor 1
Orysa sativa 1
70
Nicotiana tabacum
Pisum sativum 1
Striga asiatica 1
96
Arabidopsis thaliana 1
56
73
Solanum tuberosum 101
Zea mays 1
PLANTS
Cosmarium botrytis
68
Coleochaete scutata
Mesostigma viride
Nannochloris bacillaris
53
Scherffelia dubia
GREEN ALGAE
94
Volvox carteri
91
Chlamydomonas reinhardtii
Chlorella vulgaris 1
55
1 Allogromia sp.
92
Reticulomyxa filosa 1
87
Ammonia sp.1
FORAMINIFERA
98
Ammonia sp. 2
97
Reticulomyxa filosa 2
2 Allogromia sp.
Cyanidioschyzon merolae
RED ALGA
95
Oryctolagus cuniculus
75
Xenopus laevis
57
Bombyx mori A3
ANIMALS
56
100
Hydra vulgaris 62
Podocoryne carnea
99
Physarum polycephalum A5
SLIME MOLDS
95
Dictyostelium discoideum 15
Acanthamoeba castellanii
AMOEBAE
Entamoeba histolytica
100
Neurospora crassa
96
Ajellomyces capsulatus
72
Saccharomyces cerevisiae
Schizosaccharomyces pombe
FUNGI
61
85
Pneumocystis carinii A
Phaffia rhodozyma
77
Filobasidiella neoformans var. neoformans
Puccinia graminis
Cyanophora paradoxa
glaucocystophyte
100
Trypanosoma brucei B
90
Trypanosoma brucei A
KINETO-
PLASTIDS
B
100
Trypanosoma cruzi
Leishmania major
91
A
100
Diplonema ambulator
79
Diplonema sp. 3
DIPLONEMIDS
Euglena gracilis
EUGLENOIDS
100
Naegleria fowleri 1
Naegleria fowleri 2
HETEROLOBOSEA
100
Phytophthora infestans 2
100
Phytophthora infestans B
Phytophthora megasperma
OOMYCETES
69
Phytophthora infestans A
HETEROKONTS
(CHROMISTS)
91
100
Achlya bisexualis
Pythium irregulare
100
Costaria costata
Fucus disticus
BROWN ALGAE
Cryptosporidium parvum
Plasmodium falciparum 2
100
Plasmodium falciparum1
ALVEOLATES
Toxoplasma gondii
86
Prorocentrum minimum
100
Amphidinium carterae
Perkinsus marinus 1
0.1

the base of this phylum, so the internal topology became (D, (K, E)). These two alternative positions of diplonemids are indicated by open circles in Fig. 9. The K-H tests found that these two alternative topologies of Euglenozoa were not significantly worse than the original topology ((D, E), K), at a confidence level of 5% (Table 2).

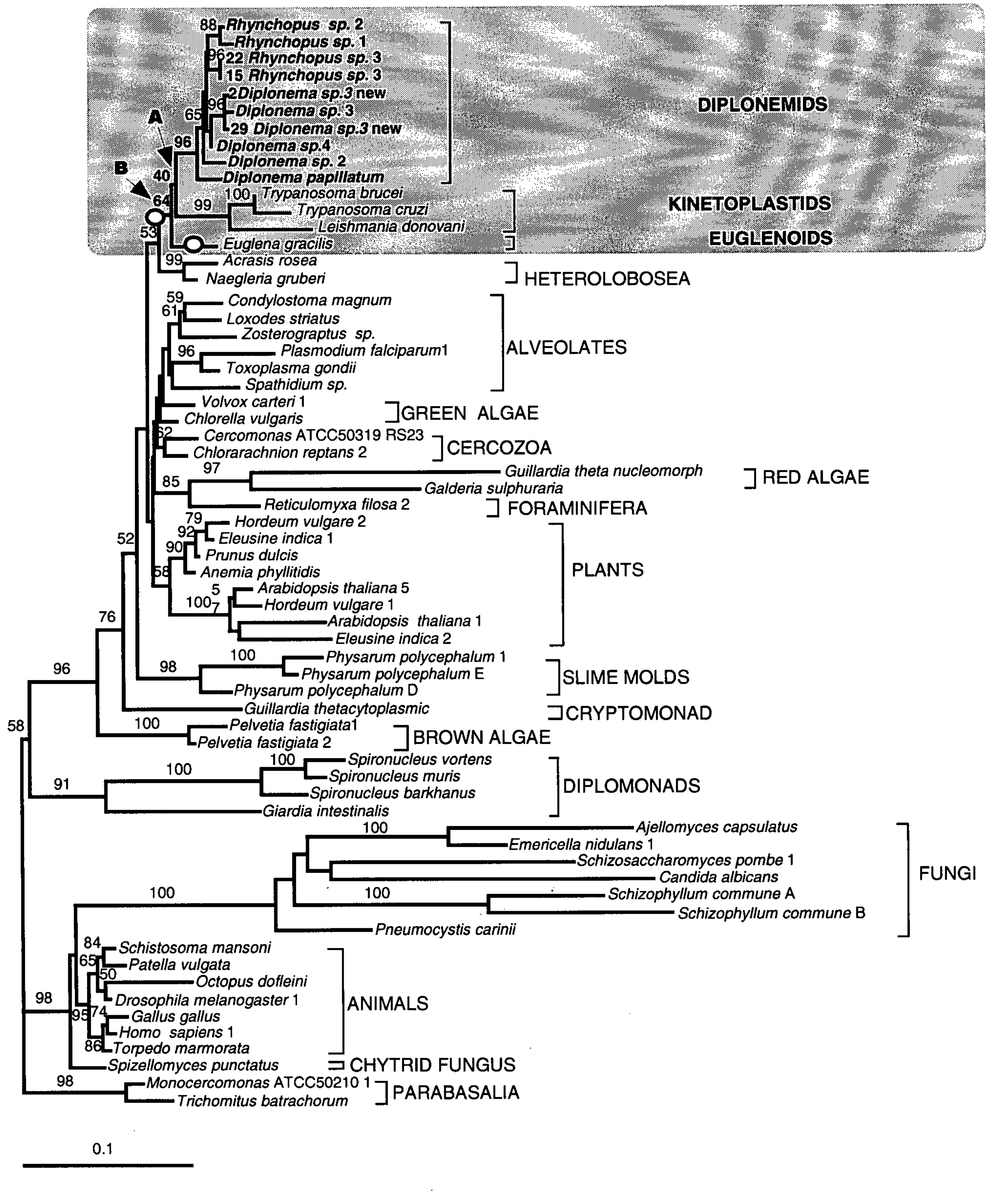
In conclusion, actin tree strongly supports the inclusion of diplonemids within the phylum Euglenozoa, but the close association of diplonemids to euglenoids at the exclusion of kinetoplasteds is only supported by bootstrap, and not by the K-H tests.

Alpha-tubulin phylogeny

An alpha-tubulin phylogeny was constructed from 436 alignable characters of a total of 64 eukaryotic taxa using distance and neighbor joining methods. The resulting alpha-tubulin tree (Fig. 10) supports most of the major eukaryotic groups, including alveolates, green algae, red algae, land plants, diplomonads, fungi, animals and parabasalia. The 10 diplonemid sequences from this study form a single group with a very high bootstrap value (96%). The phylum Euglenozoa (euglenoids, kinetoplastids and diplonemids) also forms a single clade, which is supported at 64% by bootstrap.

When the internal phylogeny of the Euglenozoa is considered, the alpha-tubulin tree tells a different story than the actin tree. The alpha-tubulin phylogeny favors diplonemids being closer to kinetoplastids than to euglenoids. However, the node uniting diplonemids and kinetoplastids (node A in Fig. 10) is poorly supported (40% bootstrap). To test the strength of this position for the diplonemids, I did K-H tests on two alternative positions for diplonemids (marked by two open circles in Fig. 10). The results showed that the alternative topologies for the phylum Euglenozoa- (D, (E, K)) and (K, (D, E))- were not significantly worse than (E, (D, K)) at confidence levels of 5%, as suggested by the low bootstrap values in the original alpha-tubulin tree (Table 3).

Fig. 10 Neighbor-joining tree based on alpha-tubulin protein sequences of various eukaryotes, including ten new sequences of diplonemids (in bold). This unrooted BioNJ tree was constructed by calculating maximum-likelihood (ML) distances between pairs of sequences. Values on selected branches indicate neighbor-joining bootstrap support greater than 50% and the bootstrap values of particular interest are in bold. Scale bar indicates amino acid substitutions per site. Node A is the last common ancestor of diplonemids and kinetoplastids and node B is the last common ancestor of the phylum Euglenozoa (shaded). Alternative positions for diplonemids were assessed with Kishino-Hasegawa tests at nodes marked with open circles. The alternatives were not rejected at 5% levels.


88 Rhynchopus sp. 2
Rhynchopus sp. 1
96 22 Rhynchopus sp. 3
15 Rhynchopus sp. 3
2Diplonema sp.3 new
96 Diplonema sp. 3
65
29 Diplonema sp.3 new
A
96 Diplonema sp.4
Diplonema sp. 2
B
40
Diplonema papillatum
DIPLONEMIDS
64
100 Trypanosoma brucei
99 Trypanosoma cruzi
Leishmania donovani
KINETOPLASTIDS
53
Euglena gracilis
EUGLENOIDS
99 Acrasis rosea
Naegleria gruberi
HETEROLOBOSEA
59 Condylostoma magnum
61 Loxodes striatus
Zosterograptus sp.
96 Plasmodium falciparum1
Toxoplasma gondii
Spathidium sp.
ALVEOLATES
Volvox carteri 1
Chlorella vulgaris
GREEN ALGAE
62 Cercomonas ATCC50319 RS23
Chlorarachnion reptans 2
CERCOZOA
97 Guillardia theta nucleomorph
85 Galderia sulphuraria
RED ALGAE
Reticulomyxa filosa 2
FORAMINIFERA
79 Hordeum vulgare 2
92 Eleusine indica 1
52
90 Prunus dulcis
Anemia phyllitidis
58
PLANTS
5 Arabidopsis thaliana 5
100 7 Hordeum vulgare 1
76
Arabidopsis thaliana 1
Eleusine indica 2
100 Physarum polycephalum 1
96
98 Physarum polycephalum E
Physarum polycephalum D
SLIME MOLDS
Guillardia thetacytoplasmic
CRYPTOMONAD
58
100 Pelvetia fastigiata1
Pelvetia fastigiata 2
BROWN ALGAE
100 Spironucleus vortens
100 Spironucleus muris
91 Spironucleus barkhanus
Giardia intestinalis
DIPLOMONADS
100 Ajellomyces capsulatus
Emericella nidulans 1
Schizosaccharomyces pombe 1
Candida albicans
FUNGI
100
100 Schizophyllum commune A
Schizophyllum commune B
Pneumocystis carinii
84 Schistosoma mansoni
65 Patella vulgata
50 Octopus dofleini
Drosophila melanogaster 1
ANIMALS
98
95 74 Gallus gallus
Homo sapiens 1
86 Torpedo marmorata
Spizellomyces punctatus
CHYTRID FUNGUS
98 Monocercomonas ATCC50210 1
Trichomitus batrachorum
PARABASALIA
0.1

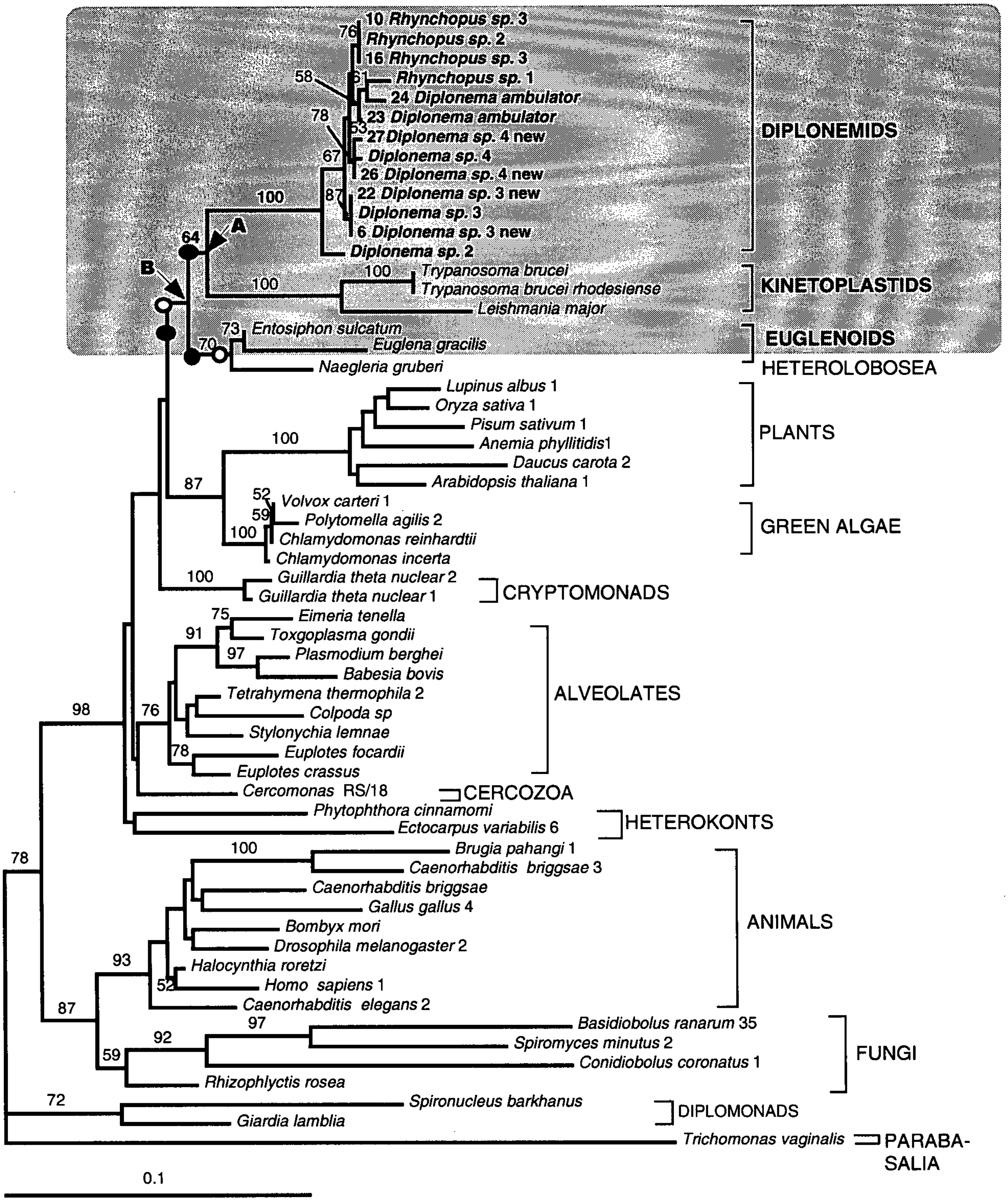
In conclusion, although in alpha-tubulin tree diplonemids are placed closer to kinetoplastids than to euglenoids, this placement is not supported by either bootstrap or K-H tests.

Beta-tubulin phylogeny

A beta-tubulin phylogeny was constructed from 428 alignable characters from a total of 59 eukaryotic taxa using distance and neighbor joining methods. The beta-tubulin tree (Fig. 11) also supports most of the common eukaryotic groups: land plants, green algae, alveolates, heterokonts, animals, fungi, and diplomonads. The thirteen new diplonemid sequences obtained from this study also branch together with 100% bootstrap support. However, in this tree, a heterolobosean sequence (*Naegleria gruberi*) branches within the Euglenozoa, specifically with euglenoids. The close association of the beta-tubulin sequences from *Naegleria gruberi* and euglenoids was also indicated by the previously constructed global beta-tubulin tree (Keeling et al. 1996). In both actin and alpha-tubulin trees, the heterolobosea form a separate phylogenetic group closest to the phylum Euglenozoa. The inclusion of a member of a different phylogenetic group may cause the low support (less than 50%) for the whole group (the Euglenozoa and *Naegleria gruberi*, node B in Fig. 11). Furthermore, the suspiciously close association of *Naegleria gruberi* and euglenoids may affect the real phylogenetic relationship of diplonemids to kinetoplastids and euglenoids.

As for the phylogenetic placement of the diplonemids within the phylum Euglenozoa, this beta-tubulin tree agrees with the alpha-tubulin tree in placing the diplonemids with the kinetoplastids, in contrast to the actin tree. The bootstrap value for the

Fig. 11 Neighbor-joining tree based on beta-tubulin protein sequences of various eukaryotes, including thirteen new sequences of diplonemids (in bold). This unrooted BioNJ tree was constructed by calculating maximum-likelihood (ML) distances between pairs of sequences. Values on selected branches indicate neighbor-joining bootstrap support greater than 50% and the bootstrap values of particular interest are in bold. Scale bar indicates amino acid substitutions per site. Node A is the last common ancestor of diplonemids and kinetoplastids and node B is the last common ancestor of the phyla Euglenozoa (shaded) and Heterolobosea (represented by *Naegleria gruberi* in this tree). Alternative positions for diplonemids were assessed with the Kishino-Hasegawa test at nodes with open circles (when *Naegleria gruberi* branches together with euglenoids) and filled circles (when *Naegleria gruberi* was moved out of the phylum Euglenozoa, suggested by a dashed line). The five alternatives were not rejected at confidence levels of 95%.


10 Rhynchopus sp. 3
76
Rhynchopus sp. 2
16 Rhynchopus sp. 3
58
61
Rhynchopus sp. 1
24 Diplonema ambulator
78
23 Diplonema ambulator
53
27 Diplonema sp. 4 new
67
Diplonema sp. 4
26 Diplonema sp. 4 new
100
87
22 Diplonema sp. 3 new
Diplonema sp. 3
6 Diplonema sp. 3 new
64
A
Diplonema sp. 2
DIPLONEMIDS
B
100
Trypanosoma brucei
100
Trypanosoma brucei rhodesiense
Leishmania major
KINETOPLASTIDS
73
Entosiphon sulcatum
70
Euglena gracilis
EUGLENOIDS
Naegleria gruberi
HETEROLOBOSEA
Lupinus albus 1
Oryza sativa 1
Pisum sativum 1
100
Anemia phyllitidis1
Daucus carota 2
87
Arabidopsis thaliana 1
PLANTS
52
Volvox carteri 1
59
Polytomella agilis 2
100
Chlamydomonas reinhardtii
Chlamydomonas incerta
GREEN ALGAE
100
Guillardia theta nuclear 2
Guillardia theta nuclear 1
CRYPTOMONADS
75
Eimeria tenella
91
Toxgoplasma gondii
97
Plasmodium berghei
Babesia bovis
Tetrahymena thermophila 2
98
76
Colpoda sp
Stylonychia lemnae
78
Euplotes focardii
Euplotes crassus
ALVEOLATES
Cercomonas RS/18
CERCOZOA
Phytophthora cinnamomi
Ectocarpus variabilis 6
HETEROKONTS
100
Brugia pahangi 1
78
Caenorhabditis briggsae 3
Caenorhabditis briggsae
Gallus gallus 4
Bombyx mori
Drosophila melanogaster 2
93
Halocynthia roretzi
52
Homo sapiens 1
87
Caenorhabditis elegans 2
ANIMALS
97
Basidiobolus ranarum 35
92
Spiromyces minutus 2
59
Conidiobolus coronatus 1
Rhizophlyctis rosea
FUNGI
72
Spironucleus barkhanus
Giardia lamblia
DIPLOMONADS
Trichomonas vaginalis
PARABA-
SALIA
0.1

Tree	log L	difference	S.E.	Significantly worse
1 (K,(D,E))	-8698.30	3.22	5.84	no
2 (E,(D,K))	-8700.06	4.99	5.10	no
3 (D,(E,K))	-8695.08	0.00		best tree

Table 2. Kishino-Hasegawa test of the positions of diplonemids within Euglenozoa in the actin tree. D-diplonemids, E-euglenoids, K-kinetoplastids.

Tree	log L	difference	S.E.	Significantly worse
1 (E,(D,K))	-8439.05	5.04	4.84	no
2 (D,(E,K))	-8434.01	0.00		best tree
3 (K,(D,E))	-8437.58	3.57	5.57	no

Table 3. Kishino-Hasegawa test of the positions of diplonemids within Euglenozoa in the alpha-tubulin tree. D-diplonemids, E-euglenoids, K-kinetoplastids.

Tree	log L	difference	S.E.	Significantly worse
1 (E,(D,K))	-6424.70	0.00		best tree
2 (D,(E,K))	-6436.15	11.45	8.51	no
3 (K,(D,E))	-6434.70	10.00	8.90	no
4 (E,(D,K))	-6425.10	0.41	10.26	no
5 (D,(E,K))	-6436.54	11.84	13.89	no
6 (K,(D,E))	-6438.95	14.26	13.46	no

Table 4. Kishino-Hasegawa test of the positions of diplonemids within Euglenozoa in the beta-tubulin tree. The topologies of Euglenozoa of Tree 1-Tree 3, with *Naegleria gruberi* branching with euglenoids. Tree 4-Tree 6 exclude *Naegleria gruberi* from the Euglenozoa. D-diplonemids, E-euglenoids, K-kinetoplastids.

diplonemids node uniting diplonemids and kinetoplastids (node A in Fig. 11) is 64%, which is higher than the bootstrap value indicated by the alpha-tubulin tree, but is still relatively low. Considering the phylogenetic position of *Naegleria gruberi* within the Euglenozoa, and that its close association with euglenoids might affect the phylogenetic placement of within this phylum, I did K-H tests on alternative positions for the group diplonemids with *Naegleria gruberi* branching with euglenoids within the phylum Euglenozoa ((D, (E, K)) and ((D, E), K), marked by open circles), and on three alternative positions for the group diplonemids with *Naegleria gruberi* constrained outside the phylum Euglenozoa ((E, (D, K)), (D, (E, K) and ((D, E), K), marked by closed circles). None of these alternative topologies were rejected at the 5% level (Table 4).

In conclusion, beta-tubulin also supports a closer relationship between diplonemids and kinetoplastids than between diplonemids and euglenoids. However, this relationship is not supported by either bootstrap or K-H tests. Moreover, the validity of the phylogenetic position of diplonemids within the Euglenozoa suggested by the beta-tubulin tree is questioned by the inclusion of a member from a separate phylogenetic group into the Euglenozoa.

To summarize, among the three protein phylogenetic tree constructed in this study, the actin tree shows the strongest bootstrap support, not only for the phylum Euglenozoa, but also for the phylogenetic placement of diplonemids within this phylum. The alpha-tubulin tree indicates low ability to resolve the internal phylogeny of the phylum Euglenozoa and the beta-tubulin tree does not support the Euglenozoa as a monophyletic phylum.

3.4 Lateral gene transfer indicated by GAPDH phylogeny

On the basis of the comparison of the 290 alignable amino-acid sequences for GAPDH from 100 taxa, a BioNJ tree was constructed (Fig. 12) using a distance and neighbor-joining analysis. This global tree includes not only diverse eukaryotic groups but also diverse prokaryotic groups. The resulting tree revealed a very complex picture of GAPDH gene evolution (Fig. 12) and recovered the basic relationships of the two separate classes of GAPDH sequences, GapC and GapA/B (divided by a dashed line in Fig. 9), typical of GAPDH phylogeny (Michels et al. 1991; Martin et al. 1993; Henze et al. 1995; Liaud et al. 1997). The GapC clade (above the dashed line) includes the cytosolic GAPDH from most eukaryotes. In published global GAPDH trees, the GapC of most eukaryotes form a sub-clade, with unresolved relationships. This is also indicated in my global GAPDH tree by the lack of bootstrap support for the backbone of the GapC sub-clade. Moreover, my global GAPDH tree also shows that the gap1 sequences from a group of proteobacteria (including the gapA (=gap1) from *E.coli*) and a group of cyanobacteria are basal to this eukaryotic crown sub-clade, but separated from GapA/B by the GapC sequences from another eukaryotic phylum, Heterolobosea. The GapA/B clade (below the dashed line) includes the GAPDH genes from most bacteria and the plastid targeted GAPDH genes from photosynthetic eukaryotes. The eukaryotic plastid-targeted GapA/B sequences form a sub-clade that is closely related to the gap2 sequences from cyanobacteria, in keeping with the cyanobacterial origin of chloroplasts.

In order to do careful phylogenetic analysis that would be impossible on 100 taxa (I did not perform the gamma-distribution correction on the distance matrices inferred from the GAPDH sequences alignment with 100 taxa), I constructed two smaller BioNJ GAPDH trees based on two sub-alignments. Both alignments retained the 290 alignable characters. One includes all 39 taxa in the GapA/B clade (below the dashed line) of the larger GAPDH tree (Fig. 13) and the other includes all 61 taxa in the GapC clade (above the dashed line) of

Fig. 12 Phylogeny of diverse eukaryotes and prokaryotes based on GAPDH protein sequences, including the four new sequences of diplonemids (in bold). This unrooted BioNJ tree was constructed by calculating maximum-likelihood (ML) distances between pairs of sequences. Values on selected branches indicate neighbor-joining bootstrap support greater than 50% and the bootstrap values of particular interest are in bold. Dashed line divides the two classes of GAPDH: GapC (above the dashed line) and GapA/B (below the dashed line). Scale bar indicates amino acid substitutions per site. The five shaded regions include all the members of the phylum Euglenozoa in this tree. Node A unites diplonemid and cyanobacterial sequences, and node B unites a second copy GAPDH of *Diplonema sp.* 2 and two sequences of cryptomonads.

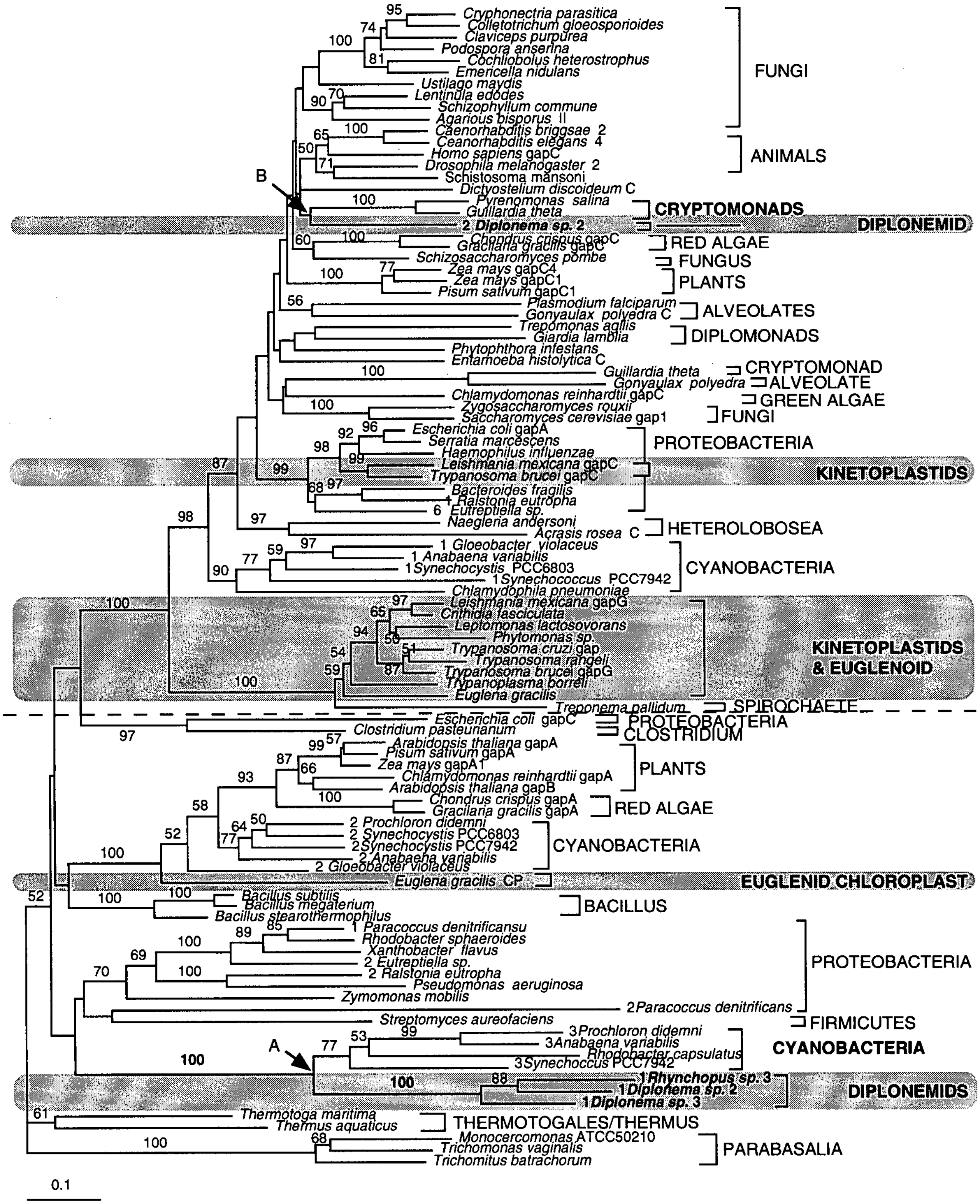

Cryphonectria parasitica
Colletotrichum gloeosporioides
Claviceps purpurea
Podospora anserina
Cochliobolus heterostrophus
Emericella nidulans
Ustilago maydis
Lentinula edodes
Schizophyllum commune
Agarious bisporus II
FUNGI
Caenorhabditis briggsae 2
Ceanorhabditis elegans 4
Homo sapiens gapC
Drosophila melanogaster 2
Schistosoma mansoni
ANIMALS
Dictyostelium discoideum C
B
Pyrenomonas salina
Guillardia theta
CRYPTOMONADS
2 Diplonema sp. 2
DIPLONEMID
Chondrus crispus gapC
Gracilaria gracilis gapC
RED ALGAE
Schizosaccharomyces pombe
FUNGUS
Zea mays gapC4
Zea mays gapC1
Pisum sativum gapC1
PLANTS
Plasmodium falciparum
Gonyaulax polyedra C
ALVEOLATES
Trepomonas agilis
Giardia lamblia
DIPLOMONADS
Phytophthora infestans
Entamoeba histolytica C
Guillardia theta
CRYPTOMONAD
Gonyaulax polyedra
ALVEOLATE
Chlamydomonas reinhardtii gapC
GREEN ALGAE
Zygosaccharomyces rouxii
Saccharomyces cerevisiae gap1
FUNGI
Escherichia coli gapA
Serratia marcescens
Haemophilus influenzae
PROTEOBACTERIA
Leishmania mexicana gapC
Trypanosoma brucei gapC
KINETOPLASTIDS
Bacteroides fragilis
1 Ralstonia eutropha
6 Eutreptiella sp.
Naegleria andersoni
Acrasis rosea C
HETEROLOBOSEA
1 Gloeobacter violaceus
1 Anabaena variabilis
1 Synechocystis PCC6803
1 Synechococcus PCC7942
CYANOBACTERIA
Chlamydophila pneumoniae
Leishmania mexicana gapG
Crithidia fasciculata
Leptomonas lactosovorans
Phytomonas sp.
Trypanosoma cruzi gap
Trypanosoma rangeli
Trypanosoma brucei gapG
Trypanoplasma borreli
Euglena gracilis
KINETOPLASTIDS & EUGLENOID
Treponema pallidum
SPIROCHAETE
Escherichia coli gapC
PROTEOBACTERIA
Clostridium pasteurianum
CLOSTRIDIUM
Arabidopsis thaliana gapA
Pisum sativum gapA
Zea mays gapA1
Chlamydomonas reinhardtii gapA
Arabidopsis thaliana gapB
PLANTS
Chondrus crispus gapA
Gracilaria gracilis gapA
RED ALGAE
2 Prochloron didemni
2 Synechocystis PCC6803
2 Synechocystis PCC7942
2 Anabaena variabilis
2 Gloeobacter violaceus
CYANOBACTERIA
Euglena gracilis CP
EUGLENID CHLOROPLAST
Bacillus subtilis
Bacillus megaterium
Bacillus stearothermophilus
BACILLUS
1 Paracoccus denitrificansu
Rhodobacter sphaeroides
Xanthobacter flavus
2 Eutreptiella sp.
2 Ralstonia eutropha
Pseudomonas aeruginosa
Zymomonas mobilis
2 Paracoccus denitrificans
PROTEOBACTERIA
Streptomyces aureofaciens
FIRMICUTES
A
3 Prochloron didemni
3 Anabaena variabilis
Rhodobacter capsulatus
3 Synechoccus PCC7942
CYANOBACTERIA
1 Rhynchopus sp. 3
1 Diplonema sp. 2
1 Diplonema sp. 3
DIPLONEMIDS
Thermotoga maritima
Thermus aquaticus
THERMOTOGALES/THERMUS
Monocercomonas ATCC50210
Trichomonas vaginalis
Trichomitus batrachorum
PARABASALIA
0.1

Fig. 13 GAPDH phylogeny of protein sequences of prokaryotes and some eukaryotes, including three new sequences of diplonemids (in bold). This unrooted BioNJ tree was constructed by calculating maximum-likelihood (ML) distances between pairs of sequences. Values on selected branches indicate neighbor-joining bootstrap support greater than 50% and the bootstrap values of particular interest are in bold. For technical details see Material and Methods. Scale bar indicates amino acid substitutions per site. The number in front of the species indicates a particular copy of the GAPDH from that species. The two shaded regions include all the members of the phylum Euglenozoa in this tree. Node A unites diplonemid sequences (*Rhynchopus sp.* 3, *Diplonema sp.* 2 gap1, *Diplonema sp.* 3), cyanobacterial gap3 sequences and one proteobacterial GAPDH sequence.

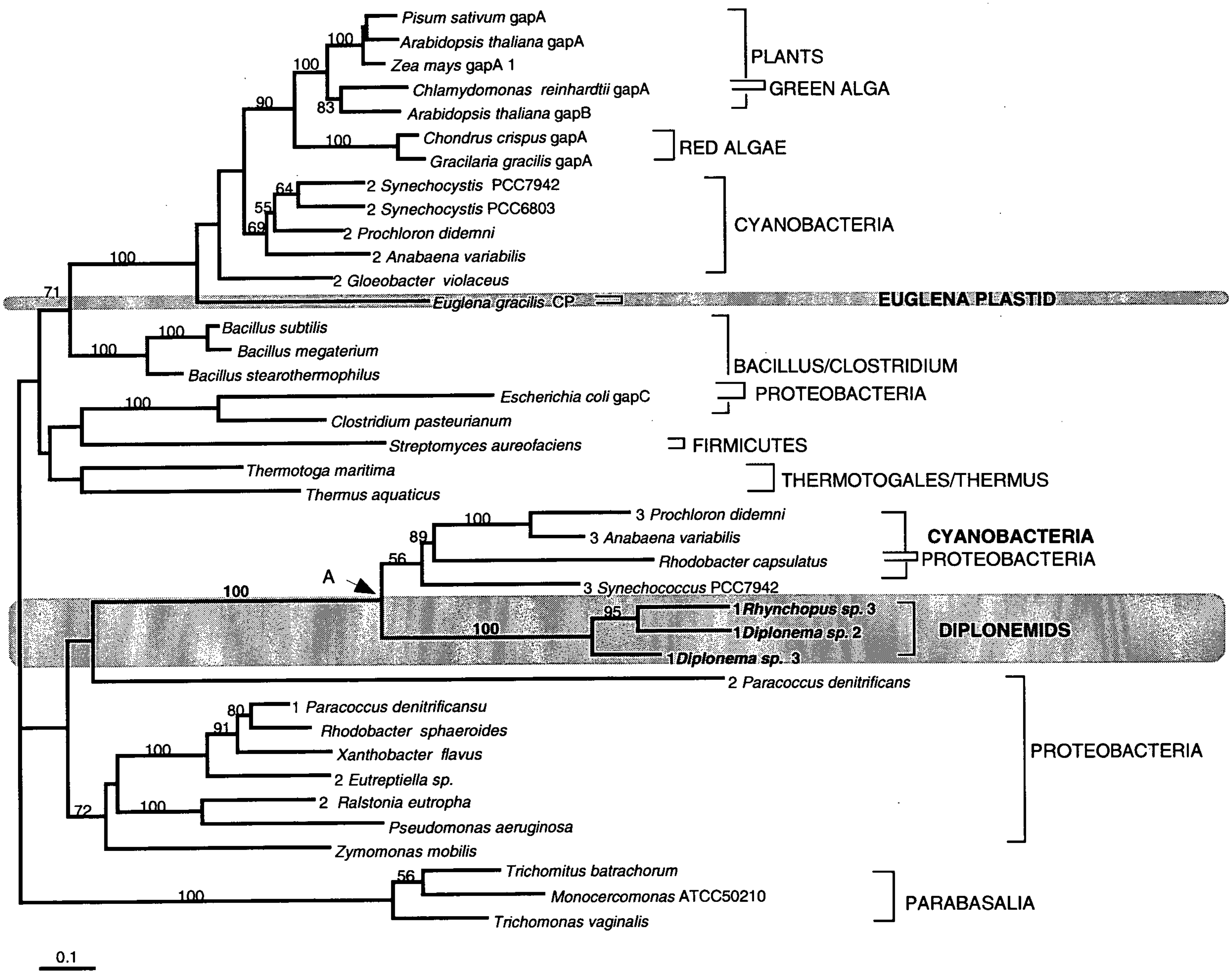
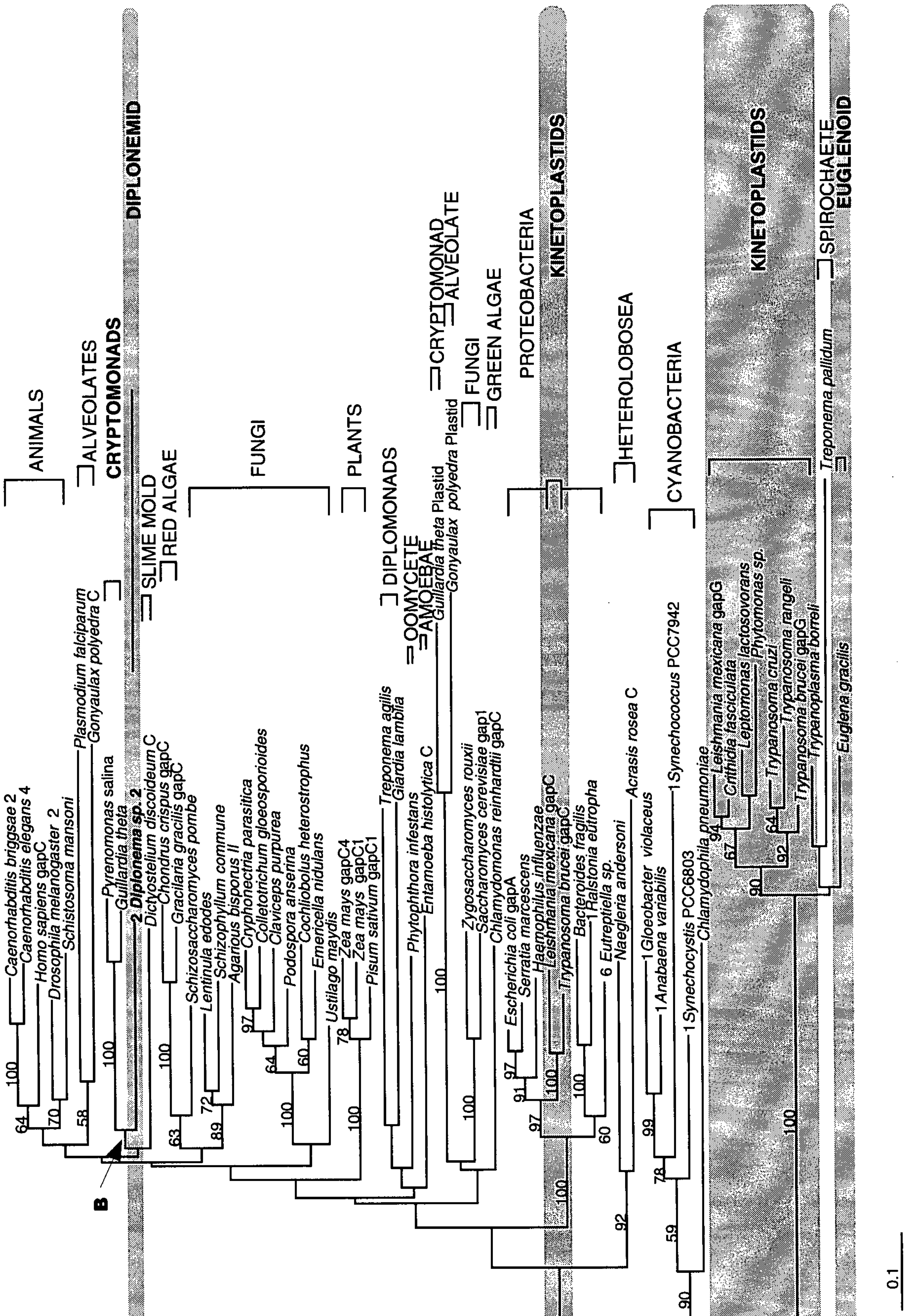

Pisum sativum gapA
Arabidopsis thaliana gapA
Zea mays gapA 1
Chlamydomonas reinhardtii gapA
Arabidopsis thaliana gapB
Chondrus crispus gapA
Gracilaria gracilis gapA
2 Synechocystis PCC7942
2 Synechocystis PCC6803
2 Prochloron didemni
2 Anabaena variabilis
2 Gloeobacter violaceus
Euglena gracilis CP
Bacillus subtilis
Bacillus megaterium
Bacillus stearothermophilus
Escherichia coli gapC
Clostridium pasteurianum
Streptomyces aureofaciens
Thermotoga maritima
Thermus aquaticus
3 Prochloron didemni
3 Anabaena variabilis
Rhodobacter capsulatus
3 Synechococcus PCC7942
1 Rhynchopus sp. 3
1 Diplonema sp. 2
1 Diplonema sp. 3
2 Paracoccus denitrificans
1 Paracoccus denitrificansu
Rhodobacter sphaeroides
Xanthobacter flavus
2 Eutreptiella sp.
2 Ralstonia eutropha
Pseudomonas aeruginosa
Zymomonas mobilis
Trichomitus batrachorum
Monocercomonas ATCC50210
Trichomonas vaginalis
PLANTS
GREEN ALGA
RED ALGAE
CYANOBACTERIA
EUGLENA PLASTID
BACILLUS/CLOSTRIDIUM
PROTEOBACTERIA
FIRMICUTES
THERMOTOGALES/THERMUS
CYANOBACTERIA
PROTEOBACTERIA
DIPLONEMIDS
PROTEOBACTERIA
PARABASALIA
A
100
100
83
90
100
64
55
69
100
71
100
100
100
100
89
56
100
95
100
80
91
100
100
72
56
100
0.1

Fig. 14 GAPDH phylogeny of protein sequences of eukaryotes and some bacteria, including one new sequences of diplonemids (in bold). This unrooted BioNJ tree was constructed by calculating maximum-likelihood (ML) distances between pairs of sequences. Values on selected branches indicate neighbor-joining bootstrap support greater than 50%. For technical details see Material and Methods. Scale bar indicates amino acid substitutions per site. The number in front of the species indicates a particular copy of the GAPDH from that species. The four shaded regions include all the members of the phylum Euglenozoa in this tree. Node B unites *Diplonema sp.* 2 gap2 and the GapC of cryptomonads.


B
Caenorhabditis briggsae 2
Caenorhabditis elegans 4
Homo sapiens gapC
Drosophila melanogaster 2
Schistosoma mansoni
Plasmodium falciparum
Gonyaulax polyedra C
Pyrenomonas salina
Guillardia theta
2 Diplonema sp. 2
Dictyostelium discoideum C
Chondrus crispus gapC
Gracilaria gracilis gapC
Schizosaccharomyces pombe
Lentinula edodes
Schizophyllum commune
Agaricus bisporus II
Cryphonectria parasitica
Colletotrichum gloeosporioides
Claviceps purpurea
Podospora anserina
Cochliobolus heterostrophus
Emericella nidulans
Ustilago maydis
Zea mays gapC4
Zea mays gapC1
Pisum sativum gapC1
Treponema agilis
Giardia lambia
Phytophthora infestans
Entamoeba histolytica C
Guillardia theta Plastid
Gonyaulax polyedra Plastid
Zygosaccharomyces rouxii
Saccharomyces cerevisiae gap1
Chlamydomonas reinhardtii gapC
Escherichia coli gapA
Serratia marcescens
Haemophilus influenzae
Leishmania mexicana gapC
Trypanosoma brucei gapC
Bacteroides fragilis
1 Ralstonia eutropha
6 Eutreptiella sp.
Naegleria andersoni
Acrasis rosea C
1 Gloeobacter violaceus
1 Anabaena variabilis
1 Synechococcus PCC7942
1 Synechocystis PCC6803
Chlamydophila pneumoniae
Leishmania mexicana gapG
Crithidia fasciculata
Leptomonas lactosovorans
Phytomonas sp.
Trypanosoma cruzi
Trypanosoma rangeli
Trypanosoma brucei gapG
Trypanoplasma borreli
Treponema pallidum
Euglena gracilis
100
64
70
58
100
63
89
72
100
97
64
60
78
100
100
100
91
97
97
100
100
60
6
99
78
92
59
90
100
100
94
67
64
92
90
ANIMALS
ALVEOLATES
CRYPTOMONADS
SLIME MOLD
RED ALGAE
FUNGI
PLANTS
DIPLOMONADS
OOMYCETE
AMOEBAE
CRYPTOMONAD
ALVEOLATE
FUNGI
GREEN ALGAE
PROTEOBACTERIA
HETEROLOBOSEA
CYANOBACTERIA
SPIROCHAETE
DIPLONEMID
KINETOPLASTIDS
KINETOPLASTIDS
EUGLENOID
0.1

the larger GAPDH tree (Fig. 14). The two smaller GAPDH trees have essentially the same branching order as the global GAPDH tree.

The phylogeny of the phylum Euglenozoa based on GAPDH sequences is a lot more complicated than those based on actin, alpha-tubulin or beta-tubulin sequences. The various Euglenozoan GAPDH genes do not branch together as they do in these other trees: instead, euglenozoan sequences are scattered all over the GAPDH tree (see shaded regions of Fig. 12). As with previous analyses (e.g. Michels et al. 1991; Martin et al. 1993; Henze et al. 1995; Liaud et al. 1997), the *Euglena* cytosolic/kinetoplastid glycosomal clade is basal to GapC, and the gapA(=gap1) from proteobacteria and the gap1 from cyanobacteria. The cytosolic GAPDH genes of *Leishmania mexicana* and *Trypanosoma brucei* are extraordinarily close to *Escherichia coli* gapA (=gap1). The chloroplast GapA of *Euglena gracilis* branches at the base of the clade comprised of plastid-targeted GapA/B sequences of photosynthetic eukaryotes, and the gap2 of cyanobacteria.

The phylogenetic positions of the four diplonemid GAPDH sequences from this study are intriguing. None of the four is closely related to any of the GAPDH sequences from other euglenozoa. Three of the four diplonemid-GAPDH sequences (*Rhynchopus sp.* 3, *Diplonema sp.* 3, and one copy of GAPDH s*equence* from *Diplonema sp.* 2) form a group (gap1) with 100% bootstrap support. This group, surprisingly, branches with the gap3 from cyanobacteria and one proteobacterial gap (*Rhodobacterium*). The union of *Diplonema sp.* 2, *Diplonema sp.* 3, and *Rhynchopus sp.* 3 gap1 with these bacterial GAPDHs is robust (supported by a 100% bootstrap value). The most reasonable explanation for this unusual association between prokaryotic and eukaryotic genes is interkingdom lateral gene transfer, much as that suggested to explain the extraordinary affinity between the *L. mexicana* and *T. brucei* GapC genes and the *E. coli* gapA (Michels et al. 1991).

The second copy of GAPDH from *Diplonema sp.* 2 (gap2) is not closely related to the other three diplonemid sequences, nor is it closely related to the other euglenozoan GAPDH sequences. Instead, it weakly branches with cytosolic GAPDH sequences from cryptomonads, which branch with animals and fungi in Fig. 12 and Fig. 14, with a very low bootstrap support (31%). The low bootstrap support makes the phylogenetic placement of the second copy of GAPDH from *Diplonema sp.* 2 very tentative and questionable, but it is certain that it is not related to the diplonemid gap1 genes.

CHAPTER IV:
Discussion

In an effort to address questions about the evolutionary history of introns in the phylum Euglenozoa, I have sequenced twenty-nine nuclear encoded genes from nine different diplonemids. I discovered eleven introns in nine of the twenty-nine genes. I have also inferred phylogenetic trees from these protein genes (actin, alpha- and beta-tubulins), including the new diplonemid-sequences to attempt to reconstruct the evolutionary history of the Euglenozoan introns.

In order to gain a better understanding of the GAPDH phylogeny of the phylum Euglenozoa, I also constructed a global GAPDH tree, including four sequences from diplonemids. The resulting phylogenetic positions of diplonemid-sequences were unexpected, which makes the GAPDH phylogeny of the Euglenozoa even more intriguing.

4.1 Phylogeny of the Euglenozoa

The actin, alpha- and beta- tubulin trees constructed in this study confirm that diplonemids represent a third group in the phylum Euglenozoa, along with euglenoids and kinetoplastids (there are no molecular data on *Postgaardi*), as previously proposed based on the morphological (Triemer et al. 1990; Triemer et al. 1991b; Simpson 1997) and molecular phylogenetic evidence (Maslov et al. 1999). While some phylogenetic relationship between diplonemids and the other two euglenozoan groups seems certain, the phylogenetic relationships among the three groups has never been clear. There are three possible topologies for a tree of three lineages (Fig. 15). The topology of the actin tree I constructed with the two-diplonemid sequences obtained in this study suggested that diplonemids are more closely related to euglenoids than to kinetoplastids. On the other hand, phylogenetic analyses of alpha- and beta- tubulin, including many new diplonemid sequences, weakly support a different topology, where diplonemids are more closely related to the kinetoplastids than to euglenoids. The third topology, in which diplonemids are at the base of the phylum, and euglenoids and kinetoplastids are closer to each other, has not been supported by any phylogenetic tree constructed in this study or in previous studies (Maslov et al. 1999).

In order to assess the reliability of different topologies suggested by different protein trees, I did bootstrap analysis (for the details, see results). Bootstrap analysis for the actin tree gives 79% support for the union of diplonemids and euglenoids, while neither alpha- nor beta-tubulin tree give strong bootstrap support for the node uniting the diplonemids and kinetoplastids (only 40% and 64%, respectively). Moreover, the reliability of the beta-tubulin tree is questionable because the phylum Euglenozoa is not holophyletic: the last common ancestor for the three major euglenozoan groups (diplonemids, kinetoplastids and euglenoids) is also an ancestor of *Naegleria gruberi*. It is well known that *Naegleria gruberi* belongs to a related but phylogenetically distinct group, the Heterolobosea. The separation of the Heterolobosea and the Euglenozoa is supported by nearly all known molecular phylogenies, including the actin and alpha-tubulin trees I constructed. Taken together, the topology of the actin tree, in which diplonemids are closer to euglenoids, is probably more reliable than that of tubulin trees.

This conclusion agrees with the topologies suggested by distance and parsimony trees based on the sequences of the SSU rRNA gene and the Cox I (cytochrome c oxidase subunit I) protein, but differs from maximum likelihood trees based on the same molecules (Maslov et al. 1999). Maximum likelihood has been proven to be a powerful method of phylogenetic reconstruction. However, as it is a complicated and computationally heavy process, the sampling size in maximum likelihood methods is very limited, especially when the data are composed of protein sequences. The limited sampling size raises doubts as to whether the phylogenetic position of diplonemids in the maximum likelihood trees of Maslove et al. (1999) are reliable. Moreover, none of the phylogenetic positions of diplonemids suggested by the maximum likelihood trees (Maslov et al. 1999) are well-supported by bootstrap analysis: in the maximum likelihood Cox I protein tree, the bootstrap support for the diplonemid/kinetoplastid branch is very low (56%) and in maximum likelihood SSU rRNA tree, it is even lower, less than 50%.

Compared with the phylogenetic analysis conducted by Maslov et al. (1999), the phylogenetic analysis I performed has three advantages. First, the sampling size is comparatively large: the number of diplonemid-species is significantly increased, especially in both alpha- and beta-tubulin trees (10 and 13 diplonemid-species, respectively). Second, the number of outgroups is also greatly increased: all of the three protein trees, actin, alpha-tubulin and beta-tubulin, contain a variety of phylogenetically distant groups. Thus, I may have avoided the bias caused by the choice of outgroups. Different choices of outgroup sequences may lead to variable support for the ingroup phylogeny. This has indeed been noticed by Maslov et al. (1999) in the phylogenetic analysis performed on the Euglenozoa: in the maximum likelihood analysis of SSU rRNA, they found that the choice of *Giardia lamblia* and *Vairimorpha necatrix* as outgroups greatly increased the bootstrap support for the association of diplonemids and kinetoplastids, compared to the bootstrap support obtained when choosing *Physarum polycephalum* and *Saccharomyces cerevisiae* as outgroups. However, they pointed out that this seemingly high bootstrap support might be caused by the biased nucleotide composition or fast substitution rates in *Giardia lamblia* and *Vairimorpha necatrix*. Third, my analyses were performed at the amino acid level, rather than alignments of DNA sequences as in the SSU rRNA analyses conducted by Maslov et al. (1999). This is also an advantage because a substitution of an amino acid may be more evolutionary informative than a substitution of a nucleotide, especially when the change of a nucleotide is synonymous.

On the other hand, although the phylogenetic analysis conducted in this study probably favours a closer association of diplonemids with euglenoids, in each case the difference between the best tree and the alternative trees was not significant at a 95% confidence level, as inferred by K-H test (see results). In order to further assess the reliability of the union of diplonemids and euglenoids, it may be helpful to perform a combined analysis, in which actin, alpha-tubulin and beta-tubulin sequences are combined into a single alignment, and a phylogenetic tree constructed from this alignment. In addition, it might be helpful to try to use different combinations of outgroups chosen from the actin, alpha- and beta- tubulin trees, in maximum likelihood analyses with the newly obtained protein sequences of diplonemids.

4.2 Possible origins of the intron-types in the Euglenozoa

As mentioned before, three types of introns (conventional GT-AG spliceosomal intron, trans-spliced discontinuous intron, and "aberrant" intron) have been reported in the phylum Euglenozoa. Since conventional spliceosomal introns are seemingly rare in the Euglenozoa, trans-spliced discontinuous introns are rarely found out of this phylum, and the "aberrant" introns are unique to photosynthetic euglenoids, it would be interesting to determine the distribution of intron types in the third major lineage of the Euglenozoa, diplonemids. In this section, I am going to discuss the possible origins of the three types of intron based on the avalaible information on the distribution of the intron types in the Euglenozoa, and the internal phylogeny of this phylum discussed in the preceding section.

GT-AG spliceosomal introns

Among the twenty-nine newly sequenced nuclear encoded genes (actin, alpha-tubulin, beta-tubulin and GAPDH), eleven GT-AG introns were found in nine genes. Thus, GT-AG introns seem to be frequently present in the actin, alpha-tubulin, beta-tubulin and GAPDH genes of diplonemids. As mentioned in the Introduction, conventional GT-AG introns are very rare in euglenoids and altogether absent from the actin and tubulin genes of *Euglena gracilis*. The apparent rarity of GT-AG introns in euglenoids could be due to limited sequence sampling. When more nuclear genes from different euglenoids are examined, more GT-AG introns could be discovered. In fact, three GT-AG spliceosomal introns have been recently reported in the fibrillarin gene of *Euglena gracilis* (Breckenridge et al. 1999), in addition to one GT-AG spliceosomal intron in a beta-tubulin gene of *Entosiphon sulcatum* (Ebel et al. 1999). Because GT-AG spliceosomal introns have been detected in the nuclear genes of most eukaryotes, including the closest relatives of the Euglenozoa, the phylum Heterolobosea (Remillard et al. 1995), it is reasonable to think that GT-AG spliceosomal introns already exited in the ancestor of the Euglenozoa. Thus, the reason that GT-AG spliceosomal introns are very rare in kinetoplastids and euglenoids is likely due to a high frequency of intron loss.

Trans-spliced discontinuous introns

Trans-splicing occurs abundantly in the post-transcriptional process of pre-mRNA in both kinetoplastids and euglenoids (see introduction), but no information available is available on whether this process is present in diplonemids or not. However, by combining the internal phylogeny and the known distribution of trans-splicing within the phylum Euglenozoa, it is possible to make predictions regarding the origin of this unusual process.

There are three possible topologies to describe the phylum Euglenozoa (Fig. 15). In my phylogenetic analysis based on actin, alpha- and beta- tubulin sequences, only two of the three topologies were ever recovered (Fig. 15 A and Fig. 15 B), while the third possible topology, favouring diplonemids at the base of the Euglenozoa (Fig. 15 C), was supported neither by my protein trees (actin, alpha- and beta-tubulin trees) nor by any other phylogenetic analysis conducted so far (Maslov et al. 1999).

Either of the two plausible euglenozoan phylogenies, that which unites diplonemids with euglenoids or, alternatively, with kinetoplastids, implies that trans-splicing arose in the common ancestor of all Euglenozoa. If this is true, then all three major groups of Euglenozoa, including diplonemids, should contain trans-spliced, discontinuous introns.

On the other hand, if one accepts the third topology of this phylum (with diplonemids basal) and then considers the known distribution of trans-splicing within Euglenozoa, there could be two possible origins of trans-splicing (Fig. 15 C): it either originated in the common ancestor of this phylum or, alternatively, in the common ancestor of euglenoids and kinetoplastids after the separation of the diplonemid-lineage.

Since the third topology of the Euglenozoa is not supported in any of the phylogenetic analysis, trans-splicing is highly likely to be an ancestral character of the phylum Euglenozoa, and therefore, it will also be found in diplonemids.

"Aberrant" introns

In nine of the twenty-nine nuclear encoded genes sequenced from diplonemids, I discovered eleven GT-AG introns. None of them resemble the "aberrant" introns unique to *Euglena gracilis*. In kinetoplastids, over 4000 protein sequences are available in Genbank at present, and none of them contains any such "aberrant" intron either. Therefore, it is tempting to speculate that "aberrant" introns are a derived character unique to euglenoids.

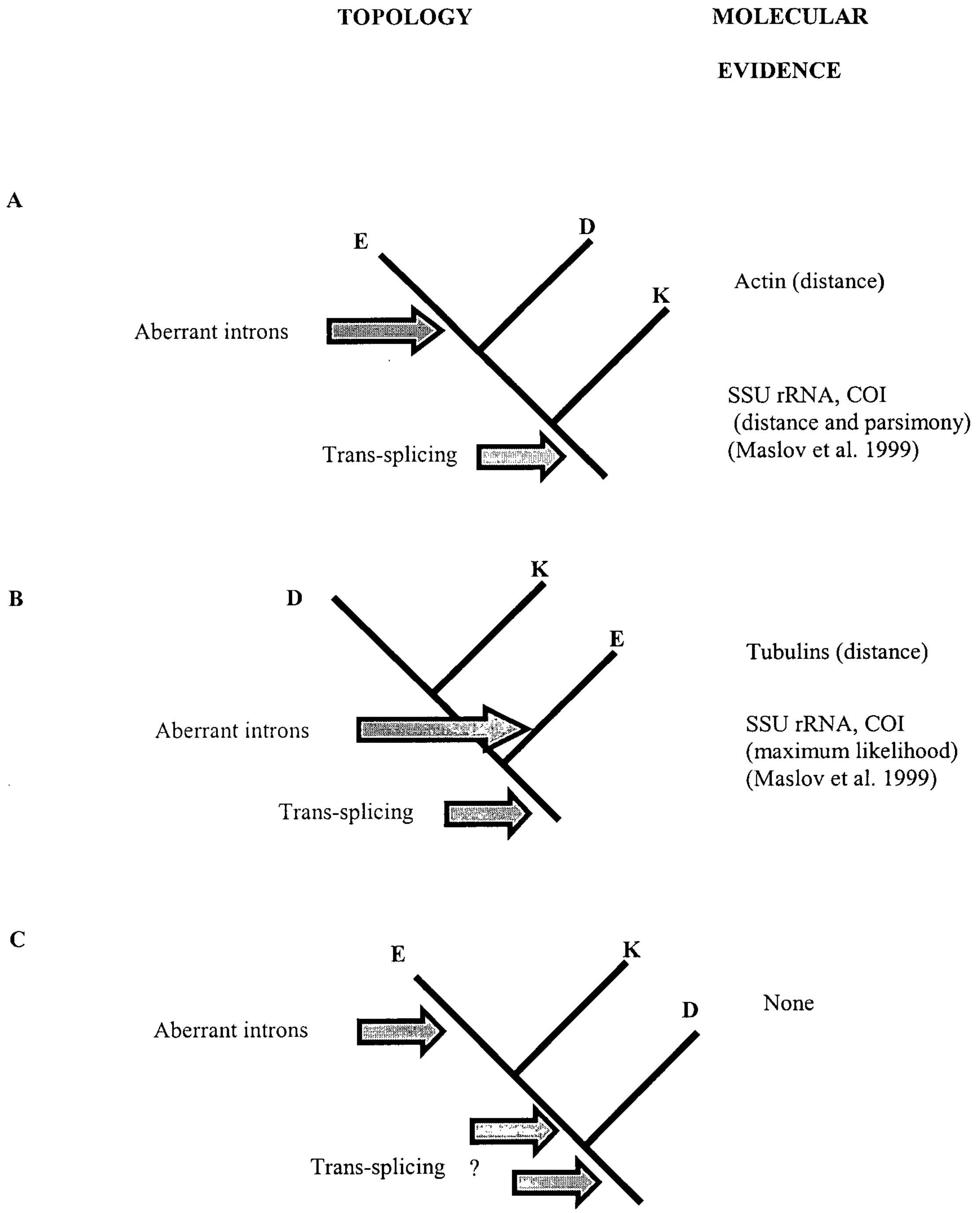


Fig. 15 Three possible topologies (A, B, C) for the internal phylogeny of the Euglenozoa. E-euglenoids; D-diplonemids; K-kinetoplastids. Arrows point at the most likely origin of either "aberrant" introns or trans-splicing. "?" indicates the uncertainty of the most likely origin of trans-splicing between the two sites: either before or after the divergence of the diplonemid-lineage from other euglenozoons.

They are perhaps unique to photosynthetic euglenoids, or they may even be *Euglena gracilis* specific, since all the thirty "aberrant" introns reported so far are from *Euglena gracilis*: 26 from two different nuclear-encoded chloroplast-targeted genes and four from the nuclear encoded, cytosolic GAPDH gene (Tessier et al. 1992; Muchhal et al. 1994; Henze et al. 1995).

4.3 Features of diplonemid introns

Although there were no "aberrant" introns in any of the 29 newly sequenced diplonemid genes, the eleven GT-AG introns from diplonemids do share four unusual features when compared with conventional GT-AG introns, especially those of mammals, which are the best studied.

First, although the lengths of these introns are not uncommon among other protist-introns, they are relatively short when compared with those GT-AG spliceosomal introns in mammals. They range in size from 40 to 149 nt whereas the sizes of the GT-AG spliceosomal introns in mammals generally range from 80 to 10000 nucleotides or more.

Second, the 5' splice consensus sequence of a typical diplonemid-intron is one nucleotide different from that of a typical mammalian-intron. As observed by previous researchers, the consensus sequences of a mammalian GT-AG spliceosomal intron are G/GURAGY at the 5' splice site and CAG/ at the 3' splice site (a slash marks the cleavage site; R represents purine; Y represents pyrimidine; N can be any nucleotide: Umen et al. 1995). In diplonemid-introns, the 5' splice site consensus is G/GURUGY while the 3' splice site consensus is the same as that of mammalian introns. These consensus sequences at the 5' splice sites and 3' splice sites of the eleven diplonemid-introns are thus very similar to those of the introns in mammals, the only difference being the fourth position at the 5' splice site. It is a U in the diplonemid-intron whereas an A in the animal-intron. This is consistent with a previous finding in euglenoids: the conserved sequences at the 5' splice sites of all the euglenoid GU-AG conventional introns so far (three introns in the fibrillarin gene of *Euglena gracilis* and one intron in the beta-tubulin of *Entosiphon sulcatum*) also have this single nucleotide substitution (Breckenridge et al. 1999; Ebel et al. 1999). In animal introns, the consensus region at the 5' splice site is recognised through complementary base pairing by U1 snRNA (Sharp 1987). It has been shown that in euglenoids, the highly conserved 5' extremity of U1 sequences contain one complimentary substitution (U to A) at the fourth position (Ebel et al. 1999; Breckenridge et al. 1999). Therefore, one would expect an analogous compensatory change at the 5' extremity in the U1 snRNA of diplonemids.

A third unusual feature is that no conventional branchpoint site can be clearly identified in these diplonemid introns. I have searched the branchpoint consensus sequence of both mammalian introns (5'-YNYURACN-3') and yeast introns (5'-TACTAACA-3') in the 11 diplonemid-introns (the branchpoint adenosine is underlined; Umen et al. 1995). As mentioned in the Results, the branchpoint consensus sequence of yeast introns was not present in any of the 11 diplonemid-introns. On the other hand, the branchpoint consensus sequence of mammalian intron was observed six times in five of the 11 diplonemid-introns (twice in one intron). But, in four of the six times, this branchpoint consensus sequence was observed either between position +4 and +11 or between position +9 and +16 (referring to the 5' cleavage site) of the introns. These can hardly be true branchpoint sites, since they are too close to the 5' splice sites of the introns. We know that the branchpoint site is closer to the 3' splice site than to the 5' splice site of an intron, usually 15-40 nucleotides upstream of the 3' splice site of an intron (Umen et al. 1995). It is highly unlikely that the branchpoint site would be present so close to the 5' splice site or even overlapping with the 5' splice site consensus sequence. In two introns, this branchpoint consensus sequence was observed once between position -23 and -16; once between position -36 and -29 in a different diplonemid gene. These two sites are also unlikely to be true branchpoint sites for two reasons. First of all, if they represent real branchpoint sites in diplonemid-introns, then they should be observed in the other nine diplonemid-introns as well. Second, the branchpoint consensus sequence is relatively redundant in a mammalian intron. Among the eight nucleotides YNYTRACN, only three nucleotides are specific. So, the chance to find such eight continuous nucleotides within a piece of nucleotide-sequence is comparatively high. In short, the branchpoint consensus sequence in a diplonemid-intron may be different from that of a mammalian intron or a yeast intron.

By analysis of the 14 alignable nucleotides at the 3' splice sites of the eleven diplonemid-introns, another unusual feature of diplonemid-introns becomes apparent. The 11-nucleotide regions preceding ACAG/ in diplonemid-introns are generally CA-rich (see Results). We know that in a conventional GT-AG intron, especially in mammals, there is a polypyrimidine tract between the branchpoint region and the 3' splice site (Umen et al. 1995). Previous experiments have demonstrated that the polypyrimidine tract in mammalian introns provides recognition sites for a splicing factor (PSF) and a negative regulatory factor, pyrimidine tract binding protein (PTB) (Gerke 1986; Tazi 1986; Singh et al. 1995). The binding of PSF to the polypyrimidine tract is essential for both splicing steps (Gerke 1986; Tazi 1986; Singh et al. 1995; Umen et al. 1995). It has also been demonstrated that PSF has strong RNA-sequence preferences (Singh et al. 1995). PTB acts as a negative regulator of splicing by binding to the pyrimidine tract and thus preventing the binding of PSF to the pyrimidine tract (Singh et al. 1995). The 'CA' rich regions adjacent to the consensus CAG/ at 3' splice site raises the possibility that the role of this region in diplonemid-introns is different from other introns.

This 'CA'-rich region is absent at the 3' splice site of the GT-AG intron in the beta-tubulin gene from the colourless euglenoid *Entosiphon sulcatum,* where, instead, a typical polypyrimidine tract is present (Ebel et al. 1999). Among the three GU-AG introns in the fibrillarin gene from *Euglena gracilis* (Breckenridge et al. 1999), the introns A and C have CT-rich tracts rather than CA-rich tracts at their 3' splice sites while intron B seems to have a weakly CA-rich tract: four A, four C, two T and two G residues, preceding the CAG/. If introns with both CA-rich and polpyrimidine tracts can exist in the same pre-mRNA transcript, then it is possible that a splicing factor in euglenoids could recognise both the CA-rich tract and the CT-rich tract at the 3' splice selection site. It is also possible that the splicing factor in diplonemids has the same dual functions, since the 11-nucleotide region preceding ACAG/ in one of the 11 diplonemid-introns is clearly CT-rich (Alpha-rh2 in Fig. 8).

In summary, introns seem to be more common in nuclear encoded genes of diplonemids than those of either euglenoids or kinetoplastids. The eleven diplonemid-introns discovered in this study are all GT-AG introns. However, they distinguish themselves from classical GT-AG spliceosomal introns in four ways: 1) They are short. 2) Nearly all the diplonemid-introns possess a T residue at the fourth position at their 5' splice sites. 3) They don't have branchpoint consensus sequence of either mammalian introns or yeast introns. 4) There is a CA-rich region comprised of 12 nucleotides preceding CAG/ at the 3' splice site of a typical diplonemid-intron. All these differences suggest that the spliceosomes in diplonemids might be slightly different from comparatively well-studied spliceosomes of other eukaryotes.

4.4 Evolutionary origin of diplonemid GAPDH

The phylogenetic positions of the four diplonemid GAPDH sequences obtained in this study (see Fig. 12) are unexpected, as none of the diplonemid sequences branch with three other known groups of euglenozoan sequences (the kinetoplastid glycosome/*Euglena* cytosol clade, the chloroplast GapA of *Euglena gracilis*, or the *Leishmania mexicana* and *Trypanosoma brucei* cytosolic clade). Instead, the gap1 sequences of three diplonemids (*Rhynchopus sp.* 3, *Diplonema sp.* 3 and *Diplonema sp.* 2) branch within the GapA/B clade, specifically with the gap3 genes of cyanobacteria while the *Diplonema sp.* 2 gap2 sequence branches with the cytosolic GAPDH genes of eukaryotes, specifically with those from cryptomonads.

The three gap1 sequences are much more similar to each other than to the gap2 sequence from *Diplonema sp.* 2 (for details, see Results). In addition, they share five insertions that are neither in the gap2 sequence from *Diplonema sp.* 2 or in any other GAPDH sequences examined (see Fig. 7 and Results). I suggest that the sequence differences between the three gap1 sequences and the gap2 sequence were largely caused by their different evolutionary histories, rather than considered to be the evolutionary consequences of different localizations or different functions within the cell. This is because that it seems likely that both types of GAPDH in diplonemids are NAD-specific, possibly playing roles in catabolic glycolysis in the cytosol. As mentioned in the Introduction, the amino acid at position 32 of a GAPDH gene is considered as an important indicator of the relative specificity of GAPDH for NAD or NADP as a substrate. The amino acid at position 32 of *Diplonema* gap2 is aspartic acid (D), the same as nearly all other NAD-specific cytosolic GAPDH enzymes found in different prokaryotes and eukaryotes, while in the diplonemid gap1 sequences, the same positions are occupied by glutamic acid (E), which is also shared by gap3 in *A. varibilis* (see Fig. 7). Comparative studies of the substrate-binding properties of various mutants have suggested that replacing Asp32 (D32) by Glutamic acid (E) will not compromise activity with NAD, but both prevent activity with NADP (Clermont et al. 1993). This means that both types of GAPDH in diplonemids are likely NAD-specific. Since we know that cytosolic GAPDH is NAD-specific and chloroplast GAPDH is both NAD- and NADP-specific, it is likely that both diplonemid gap1 and gap2 perform the same role (catabolic) in the same location (cytosol).

In the global GAPDH tree (Fig. 12), the three gap1 sequences of three different diplonemids (*Diplonema sp.* 2, *Diplonema sp.* 3 and *Rhynchopus sp.* 3) unite themselves robustly with the cyanobacterial gap3 clade at a 100% bootstrap level. It is unlikely that the diplonemid gap1 genes come from a cyanobacterial contaminant since these three gap1 sequences were isolated from three independently grown axenic cultures (see Materials and Methods). So, three related but different cyanobacteria would have to have contaminated the three diplonemid cultures, which is highly unlikely. Since it isn't contamination, lateral gene transfer is the only way to explain why diplonemids have eubacterial genes, and the close and strongly supported relationship with cyanobacterial gap3 suggests that diplonemids aquired their gap1 from a cyanobacterium through horizontal gene transfer. Lateral gene transfer has been cited previously to explain unusual association observed in GAPDH phylogeny. In addition to the GAPDH genes of parabasalids mentioned in the Introduction, another case is the extraordinarily close relationship among the cytosolic GapC sequences of *T. brucei* and *L. mexicana* with *E.coli* gap1 sequence. Michels et al. (1991) and Martin et al. (1993) have postulated that the ancestor of the trypanosome-lineage received this gene by a prokaryote-to-eukaryote lateral gene transfer from an *E. coli*-like ancestor relatively recently in evolution. Those kinetoplastids that separated early in evolution from the trypanosome-lineage (such as the bodonid *Trypanoplasma borelli*), possess only the glycosomal GAPDH gene, and lack the cytosolic genes found in "higher" kinetoplastids (Michels et al. 1992). Therefore, Henze et al. (1995) concluded that the genes for cytosolic GAPDH in kinetoplastids provides evidence for an evolutionarily recent gene transfer.

The second copy of GAPDH from *Diplonema sp.* 2 (gap2) differs considerably from the other three diplonemid-gap1 genes in sequence and appears from the phylogeny to be unrelated to the gap1 genes (see Results). In the GAPDH tree (Fig. 12), this gap2 from *Diplonema sp.* 2 falls into the eukaryotic crown taxa (the GapC clade), and branches specifically with the cytosolic GAPDH genes from cryptomonads. These are in turn closely related to animals and fungi but very distant from other diplonemid GAPDH sequences and GAPDH sequences from either kinetoplastids or euglenoids. The association of gap2 and the cryptomonad-GAPDH sequences is very weak (31% bootstrap support), however, it is clear that the diplonemid gap2 sequence is not related to any other euglenozoan GAPDH sequence. The origin of this GAPDH and its evolutionary relationships to other Euglenozoan GAPDH sequences are difficult to predict since the phylogenetic position of *Diplonema sp.*2 gap2 is so tentative. However, the presence of a *Diplonema* GAPDH sequence within the eukaryotic crown taxa (GapC sub-clade) raises tantalising question: could this be descended from the original GapC of the Euglenozoa, which is now lost in both euglenoids and kinetoplastids?

In summary, what can be inferred from this GAPDH phylogeny of the phylum Eulgenozoa are the following three points: 1) The GAPDH phylogeny of the phylum Euglenozoa is complex. There are two distinct types of GAPDH enzymes in each of the three major groups: euglenoids, kinetoplastids and diplonemids. Except for the cytosolic GAPDH in *Euglena* and the glycosomal GAPDH in kinetoplastids, the remaining four types of GAPDH sequences (chloroplast GapA in *Euglena gracilis,* cytosolic GAPDH in trypanosomes, cyanobacteria-related GAPDH in diplonemids, and the *Diplonema sp.* 2 gap2) are scattered all over the global GAPDH tree. 2)The extraordinarily close association of *Diplonema sp.* 2, *Diplonema sp.* 3 and *Rhynchopus sp.* 3 GAPDH sequences with the gap3 sequences of the cyanobacteria suggested a inter-domain horizontal gene transfer from a prokaryotic to a eukaryotic genome. 3) A different copy of diplonemid GAPDH (*Diplonema sp.* 2 gap2) branches with the GapC sequences of other eukaryotes, and may represent the ancestral euglenozoan GAPDH.

References

Agabian, N. 1990. Trans splicing of nuclear pre-mRNAs. Cell **61**:1157-60.

Biesecker, G., J. I. Harris, J. C. Thierry, J. E. Walker, and A. J. Wonacott 1977. Sequence and structure of D-glyceraldehyde 3-phosphate dehydrogenase from *Bacillus stearothermophilus*. Nature **266**:328-33.

Blumenthal, T., and J. Thomas 1988. Cis and trans mRNA splicing in *C. elegans*. TIG **4**:305-8.

Borst, P., and B. W. Swinkels 1989. The evolutionary origin of glycosomes: how glycolysis moved from cytosol to organelle. In Evolutionary tinkering in gene expression (Grunberg-Manago, M., Clark, B. F. Zachau, H.G., eds.) pp. 163-74, Plenum Publishing Corporation, New York.

Breckenridge, D. G., Y. Watanabe, S. J. Greenwood, M. W. Gray, and M. N. Schnare 1999. U1 small nuclear RNA and spliceosomal introns in *Euglena gracilis*. Proc Natl Acad Sci U S A **96**:852-6.

Cavalier-Smith, T. 1981. Eukaryote kingdoms: seven or nine? Biosystems **14**:461-81.

Clermont, S., C. Corbier, Y. Mely, D. Gerard, A. Wonacott, and G. Branlant 1993. Determinants of coenzyme specificity in glyceraldehyde-3-phosphate dehydrogenase: role of the acidic residue in the fingerprint region of the nucleotide binding fold. Biochemistry **32**:10178-84.

Davis, R. E. 1997. Surprising diversity and distribution of spliced leader RNAs in flatworms. Mol Biochem Parasitol **87**:29-48.

Ebel, C., C. Frantz, F. Paulus, and P. Imbault 1999. Trans-splicing and cis-splicing in the colourless Euglenoid, *Entosiphon sulcatum*. Curr Genet **35**:542-50.

Farmer, M. A., and R. E. Triemer 1988. Flagellar systems in the euglenoid flagellates. Biosystems **21**:283-91.

Felsenstein, J. 1993. PHYLIP (phylogeny inference package). Distributed by the author, Department of Genetics, University of Washington, Seattle, Version 3.57c.

Gascuel, O. 1997. BioNJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol **14:** 685-95.

Gerke, V., and J. A. Steitz 1986. A protein associated with small nuclear ribonucleoprotein particles recognizes the 3' splice site of premessenger RNA. Cell **47**:973-84.

Gibbs, S. P. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. Can J Bot **56**:2883-9.

Henze, K., A. Badr, M. Wettern, R. Cerff, and W. Martin 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. Proc Natl Acad Sci U S A **92**:9122-6.

Keeling, P. J., and W. F. Doolittle 1996. Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family. Mol Biol Evol **13**:1297-305.

Kishino, H., and M. Hasegawa 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol. **29:** 170-9.

Laird, P. W. 1989. Trans splicing in trypanosomes--archaism or adaptation? Trends Genet **5**:204-8.

Lee, J. J., Hutner, S. H., and E. C. Bovee 1985. Oder 11. Kinetoplastida, pp. 141-55, in An illustrated guide to the protozoa. Allen Press, Lawrence, USA.

Liaud, M. F., U. Brandt, M. Scherzinger, and R. Cerff 1997. Evolutionary origin of cryptomonad microalgae: two novel chloroplast/cytosol-specific GAPDH genes as potential markers of ancestral endosymbiont and host cell components. J Mol Evol **44 Suppl 1**:S28-37.

Mair, G., H. Shi, H. Li, A. Djikeng, H. O. Aviles, J. R. Bishop, F. H. Falcone, C. Gavrilescu, J. L. Montgomery, M. I. Santori, L. S. Stern, Z. Wang, E. Ullu, and C. Tschudi 2000. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. RNA **6**:163-9.

Markos, A., A. Miretsky, and M. Muller 1993. A glyceraldehyde-3-phosphate dehydrogenase with eubacterial features in the amitochondriate eukaryote, Trichomonas vaginalis. J Mol Evol **37**:631-43.

Martin, W., H. Brinkmann, C. Savonna, and R. Cerff 1993. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci U S A **90**:8692-6.

Maslov, D. A., S. Yasuhira, and L. Simpson 1999. Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences. Protist **150**:33-42.

Michels, P. A., and V. Hannaert 1994. The evolution of kinetoplastid glycosomes. J Bioenerg Biomembr **26**:213-9.

Michels, P. A. M., F. R. Opperdoes, V. Hannaert, E. A. C. Wiemer, S. Allert, and N. Chevalier 1992. Phylogenetic analysis based on glycolytic enzymes. Belg Journ Bot **125**:164-73.

Michels, P. A., M. Marchand, L. Kohl, S. Allert, R. K. Wierenga, and F. R. Opperdoes 1991. The cytosolic and glycosomal isoenzymes of glyceraldehyde-3-phosphate dehydrogenase in *Trypanosoma brucei* have a distant evolutionary relationship. Eur J Biochem **198**:421-8.

Muchhal, U. S., and S. D. Schwartzbach 1994. Characterization of the unique intron-exon junctions of *Euglena* gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II. Nucleic Acids Res **22**:5737-44.

Nilsen, T. W. 1995. trans-splicing: an update. Mol Biochem Parasitol **73**:1-6.

Nilsen, T. W. 1994. Unusual strategies of gene expression and control in parasites. Science **264**:1868-9.

Nilsen, T. W. 1989. Trans-splicing in nematodes. Exp Parasitol **69**:413-6.

Opperdoes, F. R. 1987. Compartmentation of carbohydrate metabolism in trypanosomes. Annu Rev Microbiol **41**:127-51.

Opperdoes, F. R., and P. A. M. Michels 1989. Biogenesis and evolutionary origin of peroxisomes. In Organelles in eukaryotic cells:molecular structure and interactions (Tager, J. M., Azzi, A., Papa, S. and Guerrieri, F., eds) pp. 187-95, Plenum Publishing Corporation, New York.

Remillard, S. P., E. Y. Lai, Y. Y. Levy, and C. Fulton 1995. A calcineurin-B-encoding gene expressed during differentiation of the amoeboflagellate *Naegleria gruberi* contains two introns. Gene **154**:39-45.

Sambrook, J., E. F. Fritsch, and T. Maniatis 1989. Small-scale preparations of plasmid DNA. In Molecular cloning (a laboratory manual, second edition) pp. 1.25-1.32, Cold Spring Harbor Laboratory Press, U S A.

Schnepf, E. 1994. Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a colorless, phagotrophic Euglenozoon with concealed flagella. Arch Protistenkd **144**:63-74.

Sharp, P. A. 1987. Splicing of messenger RNA precursors. Science **235**:766-71.

Simpson, A. G. B. 1997. The identity and composition of the Euglenozoa. Arch Protistenkd **148**:318-28.

Simpson, A. G. B., D. H. J. Van, C. Bernard, H. R. Burton, and D. J. Patterson 1996/97. The ultrastructure and systematic position of the euglenozoon *Postgaardi mariagerensis*. Fenchel et al. Arch. Protistenkd. **147:** 213-25.

Singh, R., J. Valcarcel, and M. R. Green 1995. Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science **268**:1173-6.

Strimmer, K., and A. von Haeseler 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13:** 964-9.

Tazi, J., C. Alibert, J. Temsamani, I. Reveillaud, G. Cathala, C. Brunel, and P. Jeanteur 1986. A protein that specifically recognizes the 3' splice site of mammalian pre-mRNA introns is associated with a small nuclear ribonucleoprotein. Cell **47**:755-66.

Tessier, L. H., R. L. Chan, M. Keller, J. H. Weil, and P. Imbault 1992. The *Euglena gracilis* rbcS gene contains introns with unusual borders. FEBS Lett **304**:252-5.

Tessier, L. H., M. Keller, R. L. Chan, R. Fournier, J. H. Weil, and P. Imbault 1991. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. EMBO J **10**:2621-5.

Triemer, R. E., and M. A. Farmer 1991a. An ultrastuctural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. Protoplasma **164**:91-104.

Triemer, R. E., and M. A. Farmer 1991b. The ultrastuctural organization of the heterotrophic euglenoids and its evolutionary implications, pp. 183-204. In Patterson, D. J., and J. Larsen (ed.), The biology of free-living heterotrophic flagellates. Clarendon Press, Oxford.

Triemer, R. E., and D. W. Ott 1990. Ultrastructure of *Diplonema ambulator* Larsen & Patterson (Euglenozoa) and its relationship to *Isonema*. Europ J Protistol **25**:316-20.

Umen, J. G., and C. Guthrie 1995. The second catalytic step of pre-mRNA splicing. RNA **1**:869-85.

Viscogliosi, E., and M. Muller 1998. Phylogenetic relationships of the glycolytic enzyme, glyceraldehyde-3-phosphate dehydrogenase, from parabasalid flagellates. J Mol Evol **47**:190-9.

Addendum

Spliced leader sequences have recently been isolated from *Diplonema papillatum* and *Diplonema sp.* by D. A. Campbell, University of California at Los Angeles (unpublished data, personal communication with Dr. Patrick Keeling). This discovery strongly supports my prediction that trans-splicing is an ancestral characteristic to the phylum Euglenozoa.