Leslie E. Harrell. IS GLAUCOCYSTOPHYTA THE PROTISTAN ANCESTOR OF GREEN PLANTS? (Under the direction of Dr. John Stiller). Department of Biology, May 2005.

Three diverse eukaryotic groups are thought to be related by primary endosymbiosis and have been placed together in the Kingdom Plantae: these are the glaucocystophytes, rhodophytes and green plants. Certain physical characteristics of their plastids have indicated that glaucocystophytes could be the ancestor of this group. The purpose of this research is to find evidence that glaucocystophytes are the protistan ancestors of green plants. The sequence of the largest subunit of RNA polymerase II (RPB1) is examined because it is evolutionarily conserved and, therefore, makes a suitable gene for comparative analyses of early eukaryote evolution. Three complementary approaches are used to place Glaucocystophyta on the evolutionary tree; phylogenetic analyses of RPB1 sequences, presence of a conserved C-terminal domain, and intron positions in RPB1. Data from all three characters are consistent with a close relationship between green plants and glacucocystophytes, but do not provide compelling support that they are immediate sister groups.

IS GLAUCOCYSTOPHYTA THE PROTISTAN ANCESTOR OF GREEN PLANTS?

A Thesis

Presented to

The Faculty of the Department of Biology

East Carolina University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Biology

by

Leslie Elizabeth Harrell

May 2005

IS GLAUCOCYSTOPHYTA THE PROTISTAN

ANCESTOR OF GREEN PLANTS?

By

Leslie E. Harrell

APPROVED BY:

DIRECTOR OF THESIS _____

John W. Stiller, Ph.D.

COMMITTEE MEMBER _____

Charles E. Bland, Ph.D.

COMMITTEE MEMBER _____

Jason E. Bond, Ph.D.

COMMITTEE MEMBER _____

James P. Coleman, Ph.D.

CHAIR OF THE DEPARTMENT OF BIOLOGY

_____

Ronald J. Newton, Ph.D.

INTERIM DEAN OF THE GRADUATE SCHOOL

_____

Paul D. Tschetter, Ph.D.

ACKNOWLEDGEMENTS

I want to thank my advisor, Dr. John Stiller for all he has done to help me complete my thesis and for his understanding and supportive nature. I also want to thank my committee members, Dr. Charles Bland, Dr. Jason Bond and Dr. James Coleman for their guidance and support during the course of my graduate work.

I could not have done all the lab work without the help of everyone in Dr. Stiller's lab, including Matthew Cook, Puja Gatiwala, Zhenhua Guo, Caitlin Mattos, Ana McClanahan and Patrick Ford who helped with *Glaucocystis*. I also want to thank my husband Nick who always encouraged me to finish school, and his family for helping with our daughter Zoe.

This research was funded by the Natural Science Foundation, grant number 0133295.

TABLE OF CONTENTS

List of Figures..... iv

List of Tables..... v

Introduction..... 1

Plastid Origin and Analysis..... 2

Nuclear Genome Analysis..... 5

3 Complementary Approaches..... 7

Phylogenetic Analyses of RPB1..... 8

CTD Analysis..... 8

Intron Analysis..... 10

Materials and Methods..... 11

Obtaining Cultures..... 11

Preparation of Samples..... 11

Glaucocystis..... 11

Cyanophora..... 11

RNA Extraction and cDNA Amplification..... 12

DNA Amplification..... 13

Cloning..... 14

Sequencing..... 14

DNA Isolation..... 15

Phylogenetic Analyses..... 16

Intron Analysis..... 18

Results and Discussion..... 21

Isolation and Analysis of Primary Sequences..... 21

CTD Analysis..... 23

Phylogenetic Analyses..... 24

Intron Analysis..... 27

Conclusions.....31

References.....63

Appendix. Alignment of all RPB1 Intron Positions..... 68

LIST OF FIGURES

1. Schematic Diagram of RPB1 and Primers.....34

2. GeneRacer Oligo Linkers.....36

3. CTD Comparison.....38

4. Bayesian Tree of 46 Sequences.....40

5. Bayesian Tree of 28 Sequences.....42

6. Tree from PROTPARS Analysis.....44

7. Distance tree from Neighbor-Joining Analysis.....46

8. Intron Alignment for Major Eukaryotic Groups.....48

9. Dollo Parsimony Tree based on Intron Analysis.....52

10. Dollo Parsimony Tree Created with Combined Introns.....54

LIST OF TABLES

1. Degenerate Primer Sequences for RPB1..... 56

2. KH and SH Test Results..... 58

3. Intron Positions in Eukaryotes..... 60

INTRODUCTION

The evolutionary origin of green plants and their relationship to other eukaryotic organisms is still widely debated. Green plants are characterized primarily by the presence of chloroplasts, also known as plastids. Plastids are the light harvesting organelles of photosynthetic eukaryotes and are the product of an endosymbiotic event or events between a eukaryote and cyanobacterium. A number of characteristics of green plants, including the nature of their chloroplasts, indicate that green plants are descended from more developmentally simple green algae. The question that remains unanswered is what are the closest eukaryotic relatives of this "green" lineage. Based on some molecular analyses, it has been proposed that green plants are related to red algae and a small and enigmatic group of protists known as the Glaucocystophyta; it has also been proposed that all three groups are descended from a single common photosynthetic ancestral organism (Delwiche 1999, McFadden 2001, Palmer 2003). Glaucocystophytes currently are thought to have diverged first with green plants and rhodophytes evolving later. This is because the chloroplast of glaucocystophytes, which is called a cyanelle, has retained its cyanobacterial peptidoglycan wall, while the wall has been lost in the plastids of rhodophytes and green plants (Delwiche 1999).

The Glaucocystophyta consists of three taxa, the two most prominent being *Cyanophora paradoxa* and *Glaucocystis nostichinearum*. Both *Cyanophora* and *Glaucocystis* are single-celled, flagellated organisms that are found in freshwater or in a soil sample. One difference between the two is that *Cyanophora* is a naked flagellate while *Glaucocystis* has a cellulosic cell wall. In addition to retention of their

peptidoglycan cell wall, the cyanelles found in glaucocystophytes are characterized by the presence of chlorophyll *a*, phycobilins (photosynthetic pigments also found in cyanobactteria and red algae), unstacked thylakoids and carboxysomes (sites where carbon dioxide can be concentrated) (Bhattacharya and Schmidt, 1997). These characteristics suggest that Glaucocystophytes may represent an intermediate or "missing-link" in the evolution of plastids.

Plastid Origin and Analysis

The origin of plastids is one of the most significant events in the evolution of eukaryotes because it gave rise to all phototrophic eukaryotes (Archibald and Keeling 2002). Many studies have been performed in an effort to understand the evolution of plastids, and in hopes of uncovering the origin of green plants. The number of endosymbiotic events, as well as the relationships among different plastid lineage, is still under debate (Stiller et al. 2003, Palmer 2003); however, the dominant view emerging is that all eukaryotic plastids are descended from a single "primary" endosymbtiotic association between a phagotrophic protist and cyanobacterium (Palmer 2003). Many groups of algae have obtained their plastids via secondary endosymbiosis involving the capture of a photosynthetic eukaryotic endosymbiont, which, unlike the "primary" organelles of green plants, red algae and glaucocystophytes, resulted in more than two membranes surrounding the plastid (Delwiche 1999). Most molecular evidence suggests that plastids of green algae and plants, red algae and glaucocystophytes are all derived from one primary endosymbiotic event (Delwiche & Palmer 1997, Martin et al. 1998). Several of these analyses have placed the glaucocystophyte plastid as the outgroup to the

other primary plastids (Delwiche 1999, Martin et al. 1998, Helmchen et al. 1995). If this interpretation is correct and plastids are monophyletic, this places the Glaucocystophyta in an important evolutionary position, diverging before a single loss of the ancestral plastid cell wall.

Molecular research on glaucocystophytes has focused primarily on the cyanelle genome because of its possible key position in plastid evolution; *Cyanophora* has been the most extensively studied of the Glaucocystophyta. The entire cyanelle genome has been sequenced from *Cyanophora* and used to infer the evolutionary position of the glaucocystophytes through its plastid lineage. One of the strongest indications for a common origin of all plastid types is the cluster of the -RPS20 gene- and three transcriptional units (rpoB, C1, C2, rps2-tsf, and atpI,H,F,G,D,A,C). This cluster is shared by all known plastids, but has been found in no extant cyanobacterium to date. This appears to be most easily explained by a fusion event after a single endosymbiotic event joining a cyanobacterium with a host eukaryote, from which all plastids are descended (Loffelhardt et al. 1997).

Small subunit or 16S ribosomal plastid RNA sequences from several members of the Glaucocystophyta were analyzed using maximum-likelihood, maximum parsimony, and neighbor-joining phylogenetic methods (Helmchen et al. 1995). A global search using maximum-likelihood was performed, suggesting that cyanelles were the first to diverge among all plastids. This places Glaucocystophytes as the outgroup to rhodophytes and green plants, a result consistent with a single loss of the cyanobacterial peptidoglycan wall. However, the other two methods indicated some discrepancies and

did not reveal as well supported a position as did the maximum-likelihood method. Martin et al. (1998) examined known protein-coding genes from nine chloroplast genomes and used ML to conclude that glaucocystophyte plastids are the basal branch to chlorophyte and rhodophyte plastids. This data set also created different topologies depending upon the method of analysis, as occurred in the small-subunit rRNA analyses. Because phlyogenetic trees of plastid genes must be rooted using relatively distant sequences from cyanobacteria, the point of origin of plastid trees remains suspect (Moreira and Philippe 2001).

Another method for comparing plastid genomes has been examining similarities in gene content. It has been demonstrated that plastids have lost many genes throughout their evolution but probably seldom if ever gain genes (Martin et al. 2002). Nozaki and his colleagues (2003) used a different approach and created a phylogeny based on plastid gene loss. Cladistic analysis of 274 genes resolved two major monophyletic groups with robust support, the red lineage and a large monophyletic group consisting of green lineage and the basal glaucocystophyte plastid (Nozaki et al. 2003). These phlyogenetic relationships suggest a specific sister relationship of plastids in the green lineage and the Glaucocystophyta.

Although data from plastid analyses generally suggest a single plastid origin, and often well-supported relationships among different groups of plastids, this does not necessarily indicate a close relationship among their host organisms. Real conflicts could exist due to plastids evolving independently from their hosts through gene loss, substitution, or horizontal transfer of plastids (Delwiche 1999). Therefore, examining the

nuclear genome of the host cell provides more appropriate data for analyzing host cell evolution.

Nuclear Genome Analysis

The results from phylogenetic analyses comparing chloroplast and nuclear sequences are often contradictory. This could be due to methodological differences in the types of programs used in the analyses, or to unusual or unequal evolutionary rates among genomes. It also may reflect real differences in the evolutionary histories of plastids and their host organisms. Most studies have used genes from the cyanelle genome rather than the nuclear genome in determining the evolutionary positions of glaucocystophytes relative to green plants. Many more nuclear genes need to be examined before firm conclusions can be drawn. Complex eukaryotes, including animals, fungi, plants, and most algae, appear to have emerged as a broad radiation known as the eukaryotic "crown" (Knoll 1992). Below the crown, the more distinct and structurally simpler organisms generally diverged in a ladder-like pattern (Philippe et al. 2000).

Nuclear encoded small-subunit ribosomal rRNA (SSU rRNA) has been used extensively as an evolutionary marker because of its relatively large size and its highly conserved function (Bhattacharya et al. 1995). SSU rDNA analyses provide strong evidence for an origin of the Glaucocystophyte host cell within the eukaryote crown group radiation. Analyses of this gene further suggest that glaucocystophytes and a group known as the cryptophytes share a common ancestry. Bootstrap analyses with

these data sets also support the monophyly of the Gluacocystophyta itself (Bhattacharya and Schmidt 1997).

Actin is a highly conserved protein found in eukaryotes that is used in phylogenetic analyses to corroborate results from sequence analyses of other genes. When actin is found in single copies, problems associated with multi-gene families may be avoided, resulting in a more useful phylogenetic marker (Bhattacharya and Schmidt 1997). The *Cyanophora* genome was found to contain a single copy actin gene sequence. Maximum-likelihood and maximum parsimony analyses of actin sequences also place *Cyanophora* within the crown group radiation (Bhattacharya and Weber 1997). Chlorophyte host cells are closely related to glaucocystophytes in the actin trees and this is consistent with a monophyletic origin of the plastids in these groups, with glaucocystophytes emerging as the outgroup to green algae and plants. Therefore, actin trees agree with phylogenetic results obtained using plastid genes.

A combination of genes can also be used in phylogenetic analyses to provide a more diverse array of sequences to be studied. Moreira et al. (2000) did an ML search using a multigene fusion which included six genes (actin, α-tubulin, β-tubulin, EF-1-α Hsp 70 and ATPase-vatB) including representatives from each primary plastid group. The best tree showed that red algae, green plants and the Glaucocystophyta make up a weakly supported monophyletic group. However, the association was not significant according to the Kishino-Hasegawa test (Moreira et al. 2000). A similar study was done by Baldauf and Roger (2000) using another combination of protein sequences. *Porphyra*, a red alga, grouped with green plants as some other nuclear studies have shown, but the

βtub-2 sequence proved to be from a contaminant, so the position of reds was not actually supported (see note added at proof stage). After the correct sequence of βtub-1 was obtained, the position of *Porphyra* on the tree moved, leaving only greens and glaucocystophytes as a monophyletic group. Even with several genes used in different combinations, multi-gene data still do not agree upon the exact position of glaucocystophytes, red algae and green algae. Moreira and colleagues (2000) suggested that more gene sequences are needed to determine how these groups are related.

Three Complementary Approaches

The complexity of evolution at broader time scales, along with the history of conflicts between results from different sequences and methods of analyses, demonstrate that no single character or methodology can reliably define ancient evolutionary relationships. Although large molecular data sets have provided strong statistical support for a certain hypothesis, it is important to show that the tree-building signal clustering plastid and nuclear genes in phylogenetic analyses comes from shared history and not data biases (Stiller 2004). As discussed above, differences between phylogenetic analyses of an organism's position in evolution can result from real differences in the histories of characters, or to methodological factors. The aim of this study is to use three distinct and complementary approaches to examine the gene encoding the largest subunit of RNA polymerase II (RPB1). This provides multiple lines of evidence to test the hypothesis that Glaucocystophytes are the direct protistan ancestor of green algae and plants. The three approaches are as follows: first, the primary sequence of *RPB1* is determined and used in phylogenetic analyses to place glaucocystophytes on the

evolutionary tree; second, the presence or absence and overall structure of the C-terminal domain is investigated to see if glaucocystophytes have a CTD as is predicted for the ancestor of green plants; finally, evidence of shared intron position is investigated for indications of a relationship between the two groups.

Phylogenetic Analyses of RPB1. The RPB1 gene encodes the largest sub-unit from RNA Polymerase II (RNAP II), the enzyme responsible for synthesis of messenger RNA (mRNA) in the eukaryotic nucleus. This makes much of the sequence of the gene highly conserved in eukaryotic organisms (Stiller and Hall 1998). All RPB1 homologs share conserved domains, designated regions A through H, which can be aligned among all eukaryotic, eubacterial, and archaebacterial polymerase largest subunit sequences (Jokerst et al. 1989). In addition, the RPB1 gene has been sequenced from many different organisms, so it can be used in a fairly comprehensive investigation of eukaryotic relationships. It already has been the focus of a number of phylogenetic studies of ancient evolutionary events (Sidow and Thomas 1994, Klenk et al. 1995, Puhler et al. 1989, Hirt et al. 1999).

CTD Analysis. Unique among RNA polymerases, RNAP II has an extended C-terminal domain, or CTD, of RPB1 consisting of tandemly repeated heptapeptides. This CTD is present in all RNAP II largest subunits from animals, green plants and fungi examined so far (Stiller and Hall 2002). The protistan ancestors of animals and fungi also have a CTD, and these groups, along with plants, fall into a monophyletic group. Bayesian analysis of RPB1 produces this monophyletic group, termed the "CTD-clade." However, a growing number of genes have been sequenced from diverse protists and algae that do

not encode the C-terminal repeats (Stiller and Hall, 2002). The invariable presence of the CTD in all members of the CTD-clade, predicts that it also should be conserved in the specific ancestor of green plants. The presence of a CTD, and specifically, one that consists of the canonical heptads (YSPTSPS), would place Glaucocystophytes in the CTD-clade and provide further support that it could be related to green plants. Organisms that do not have this "standard" CTD, fall outside of the clade when Bayesian inference is used to analyze the data.

The CTD plays an important role in transcription by binding essential proteins that help regulate gene expression, promote efficient elongation, and couple transcription to pre-mRNA processing (Howe 2002, Stiller and Hall 2002). It also is involved in binding a group of spliceosomal factors that are involved in intron splicing in different tissues and at different developmental stages (Bourquin et al. 1997, Tanner et al. 1997). The CTD is essential for gene regulation and co-transcriptional pre-mRNA processing. If the CTD is not present to bind various transcription and processing factors, then a given organism must have alternative mechanisms for controlling the RNAP II transcription cycle (Stiller and Hall 2002). Presumably, major changes in the mechanisms of transcription are uncommon evolutionary processes, so two organisms are not expected to be closely related if their RNAP II largest subunits do not both contain the conserved CTD. Therefore, the absence of the CTD from Glaucocystophyta RPB1 would suggest that the group does not share core transcription and mRNA processing functions common to green plants, animals, and fungi.

Intron Analyses. The presence and position of introns can also be an important indication of phylogenetic relationships. It is thought that spliceosomal introns were not present in prokaryotes, but originated sometime after the origin of the stem eukaryote (Lynch and Richardson, 2002). There are two competing theories for the explanation of most current intron positions. The introns-late hypothesis states that most introns were inserted relatively recently into "proto-splice sites" of genes (Federov et al. 2002, Rogozin et al. 2003). Therefore, most shared intron positions among organisms could represent shared-derived evolutionary characters. The alternative hypothesis states that most intron positions are ancestral, meaning they existed at least before the animal, fungal and plant kingdoms diverged. Fedorov and colleagues (2002) demonstrated that for the genes with the most introns, 60% of fungal introns have positions common to either animal or plant introns, and 39% of fungal introns share common positions with both plant and animal introns. This large percentage of shared introns supports the second hypothesis and indicates that many introns are likely to be ancestral in these three groups (Fedorov et al. 2002). Introns that do not appear to be ancestral were probably gained more recently, so if an intron is shared uniquely by two lineages, it could be a shared-derived position (Rogozin et al. 2003). A reasonable percentage of introns appear to be derived, rather than ancestral (about 40% based on the numbers previously stated). These derived introns could provide a phylogenetic signal for grouping glaucocystophytes with plants.

MATERIALS AND METHODS

Obtaining Cultures

Most available cultures of *Cyanophora*, including the culture examined originally, contain bacteria and an unidentified eukaryotic protist. This protist poses problems because degenerate PCR primers can potentially amplify *RPB*1 from any eukaryotic organism, not just a glaucocystophyte. In order to make certain that the gene recovered was from *Cyanophora,* an axenic culture (containing only one organism) was needed so that *Cyanophora paradoxa* was guaranteed to be the only eukaryote present. This axenic culture was obtained from the CCAC, Culture Collection of Algae at the University of Cologne in Germany. *Cyanophora* was grown in 10 ml and 250 ml bubbling cultures of soil water medium with barley seeds, under constant fluorescent light. Subsequent cultures were ordered from Carolina Biological, Burlington, NC, as were cultures of *Glaucocystis nostichineraum*. *Glaucocystis* was grown in 250 ml flasks of Alga Gro media (Carolina Biological).

Preparation of Samples

Glaucocystis. Samples were collected and spun down in 10 ml conical tubes so that the media could be poured off. Then the sample was placed in a chilled mortar and flash frozen by pouring in liquid nitrogen. The algal specimen was pulverized with a pestle and kept frozen with liquid nitrogen until it became a fine powder. The ground sample was used either in DNA or RNA extraction.

Cyanophora. Since *Cyanophora* does not have a cell wall, it did not need to be pulverized under liquid nitrogen. For RNA extractions, the sample was spun down in a

10 ml conical tube, and the medium poured off. Lysis buffer was immediately added and no other steps were necessary to break open the cells. For DNA extraction, the same thing was done, only CTAB extraction buffer was added instead (Doyle and Doyle 1990).

RNA Extraction and cDNA Amplification

The Promega (Madison, WI) SV Total RNA Isolation System was used to isolate total RNA from a culture of *Cyanophora* or *Glaucocystis*. Beginning with the addition of SV RNA Lysis Buffer, RNA purification was done by column centrifugation under manufacturer's directions. The total RNA was then eluted from the silica surface membrane by the addition of nuclease-free water. The RNA was then quantified and stored at -80°C.

The Invitrogen (Carlsbad, CA) GeneRacer kit was used to make complementary DNA (cDNA) from the extracted RNA. This results in a copy of the coding region of the gene with no introns. The 5' end was obtained using rapid amplification of cDNA ends (RACE) method. The total mRNA was dephosphorylated, the 5' cap structure removed, and then the GeneRacer RNA oligo was ligated to the 5' end of the mRNA. Because mRNA molecules lacking a cap were dephosphorylated prior to this ligation step, only messages complete on the 5' end bound the oligo. This oligo included a linker with two nested priming sites. The final step is an AMV reverse transcription (RT) reaction, and a gene-specific reverse primer was used from near the 5' end of the RPB1 coding region (Figure 1b). In other reactions, where the 3' end was the target, the GeneRacer Oligo dT primer was used, so that the cDNA recovered contained the entire 3' end of the gene (Figure 1c). The GeneRacer Oligo dT primer (like the GeneRacer 5' RNA Oligo)

contains a linker with two nested priming sites along with the poly-T reverse transcription primer to facilitate subsequent recovery of the gene by the polymerase chain reaction (PCR) (Figure 2).

DNA Amplification

PCR was used to amplify segments of *RPB*1 from both cDNA and total genomic DNA. Eppendorf Taq polymerase was used in 30 µl reactions for shorter pieces of DNA, less than 1200 base pairs. Standard *RPB*1 degenerate primers (Table 1) were used in nested pairs to recover a significant amount of amplified DNA. Since degenerate primers were involved, a "touchdown" PCR program was used with the steps as follows:

Step	Profile
1	94° for 1 minute
2	94° for 30 seconds
3	58° for 1 minute minus 1° /cycle
4	72° for 2 minutes
5	15 times to step 2
6	94° for 30 seconds
7	55° for 30 seconds
8	72° for 2 minutes plus 5 seconds/cycle
9	25 times to step 6
10	72° for 10 minutes

PCR products were electrophoresed on a 1% agarose gel and stained in ethidium bromide so that the amplified-fragment size could be compared to bands in the Promega 1 kb ladder. If the band appeared to be the correct size, the remaining PCR product was purified by low-melt agarose gel electrophoresis in TAE (Tris/acetic acid/EDTA) buffer. Bands were cut out of the gel, cleaned using a QIA miniprep column (Qiagen, Carlsbad, CA), and then cloned directly. Only PCR reactions that had little product or had

additional artifacts that would interfere with easy direct cloning were gel purified; clear, single bands were cloned directly.

Cloning

The Invitrogen Topo-TA Cloning kit for sequencing was used to clone and sequence each overlapping segment of the gene. The One Shot Chemical Transformation protocol was followed per manufacturer's instructions. In the Topo cloning reaction, the purified PCR product is mixed with the plasmid vector PCR 4-Topo vector. This reaction directly inserts a PCR product into a plasmid vector for sequencing. The vector was then transformed into Top10 competent *E. coli* cells and plated on LB (Luria Broth) + ampicillin (75 μg/ml) plates that were treated with X-Gal (final concentration 10 g/ml) so that transformed colonies containing inserts could be selected. X-Gal was added to the plates to permit blue/white selection; colonies that contained an insert remained white because the insert disrupts the reading frame of β-galactosidase, while those lacking an insert turn blue. White colonies were screened using a PCR-stab technique described by Palumbi and Baker (1994). The standard PCR screen was done using vector primers, M13 forward and reverse, to ensure that the colonies contained the right size insertion of DNA. Colonies with the correct size insert were cultured overnight in 10 ml of LB + kanamycin (50 μg/ml media). QIAprep Spin Miniprep kit from Qiagen was used to isolate the plasmid DNA for sequencing.

Sequencing

Big-Dye Terminator (ABI version 3.0, dideoxy terminator chemistry) was used with M13 forward and reverse primers to sequence the plasmid DNA using a standard

Big-Dye cycle sequencing program. After the cycle-sequencing program, the DNA was cleaned using a G-50 fine grade Sephadex column. Applied Biosystems Inc. (ABI, Foster City, CA) Automated Sequencer was used to visualize the sequenced fragments on a 5% long-range polyacrylamide gel. The Sequencher 4.0 program (Gene Codes Corporation, Ann Arbor, MI) was used to examine the sequence, assemble contigs and compare them to sequences of the other clones to obtain the most accurate sequence possible.

DNA Isolation. Once *RPB*1 cDNA was sequenced, a genomic DNA sequence was needed to determine the intron positions. After the sample was prepared as described previously, the pellet of cells was resuspended in an equal volume of CTAB extraction buffer (2.5% CTAB w/v, 1.4 M NaCl, 100mM Tris pH 8, 10 mM EDTA, 1% PVP and 0.2% -mercaptoethanol). The resuspended pellet then was heated to 70°C for 5 minutes. Then 1 μl of proteinase K (20 mg/ml) was added, and the temperature was reduced to 50°C for 15 minutes. An equal volume of 24:1 chloroform:isoamyl-alcohol was added, the solution was shaken to emulsify, then spun at 13,000g for 10 minutes to separate the organic and aqueous phases. The aqueous phase containing genomic DNA was removed, one μl of RNase A (10 mg/ml) was added and the mixture was allowed to incubate at 42°C for 15 minutes. A Qiagen DNA clean-up kit was used to column purify the DNA following the manufacturer's directions using a standard centrifugation protocol. Gene-specific primers designed from cDNA were used to PCR amplify *RPB*1 genomic sequences. The products were cloned, sequenced and analyzed as described above.

After an initial segment from *Cyanophora* and *Glaucocystis RPB*1 was sequenced, sequence specific primers were designed for each organism and used in conjunction with other downstream degenerate primers matching the most highly conserved RPB1 motifs.

Phylogenetic Analyses

The RPB1 inferred amino acid sequences of *Glaucocystis* and *Cyanophora* were aligned with a data set of other RPB1 genes from organisms present in GenBank and various genome-sequencing databases. First the amino acid sequences for universally conserved regions A-H were aligned using CLUSTAL W (Thompson et al. 1994). Areas of the sequences that could not be aligned with confidence were excluded from phylogenetic analyses.

Bayesian inference was performed using MRBAYES (Hulsenbeck and Ronquist 2001) to assess the strength of support for the hypothesis that glaucocystophytes should group with green plants. Two separate data sets were analyzed. In one, incomplete sequences were removed from the alignment, while maintaining at least three or four organisms per major eukaryotic group where possible. Another set of analyses was performed using all RPB1 data available, including a partial sequence from *Cyanophora* (A-G). A model of invariant sites plus four discrete categories estimating a Γ (gamma) distribution was used to model rate variation among sites, along with a JTT (Jones et al. 1992) matrix for substitution probabilities among amino acids. Four chains, one heated, were run for one million generations and trees were sampled every ten generations. Trees saved before the likelihoods converged on a stable mean value, the so-called "burn-

in," were eliminated and the remaining trees were combined into a 50% majority rule consensus tree using PAUP 4.0 (Swofford 1998) to determine whether glaucocystophytes and green plants grouped together.

Branch lengths for both Bayesian trees were calculated using TREE-PUZZLE 5.0 (Strimmer and von Haeseler 1996). A consensus tree was created in PAUP from the MRBAYES trees, and it was imported into TREE-PUZZLE to estimate maximum likelihood branch lengths and parameters. Tree reconstruction was performed using the user defined tree search procedure. The JTT model of amino acid substitutions was used along with the same parameters for rate variation among sites as in MRBAYES.

To determine the significance of differences between the consensus Bayesian tree and alternative tree topologies, both Shimodaira-Hasegawa (SH) (Shimodaira and Hasegawa 1999) and the Kishino-Hasegawa (KH) (Kishino and Hasegawa 1989) tests were performed on the smaller but more complete data set of *RPB*1 sequences. In order to test several *a priori* hypotheses against the Bayesian consensus tree, several constrained trees were constructed using McClade 3.06 (Maddison and Maddison 1992). First, to test whether the glaucocystophytes could be rejected as the specific protistan ancestor of chlorophytes, *Glaucocystis* was constrained in a monophyletic relationship with green plants; second, to test whether the proposed kingdom "Plantae" (Gray 1993, Gray and Spencer 1996, Delwiche and Palmer 1997) could be rejected, *Glaucocystis*, green plants and rhodophytes were constrained as a monophyletic group; and third, *Acanthamoeba* was constrained as sister to the Opisthokonts (animals + fungi) because this relationship has been supported strongly in other molecular phylogenetic analyses

(Bapteste et al. 2002). In all cases, the minimum required tree rearrangements were made by hand in MacClade, to avoid complicating tests with additional introduced variables. The KH tests were performed using TREEPUZZLE, and SH tests using PHYLIP (version 3.61).

Parsimony and neighbor-joining bootstrap analyses also were performed using PHYLIP version 3.6. For parsimony analyses, one hundred data sets were produced in SEQBOOT and used as the infile in the PROTPARS (protein parsimony) program, each jumbled 10 times for random sequence addition. The CONSENSE program then produced a majority rule consensus tree from the one hundred analyses along with bootstrap values. Distance matrices were constructed in the PROTDIST program using the same one hundred data sets, with JTT weighting for probability of change among amino acids with a four-category estimate for a Γ distribution of rate variation among sites. The parameter α was first estimated in TREEPUZZLE using the same model of sequence evolution. The resulting one hundred distance matrices were then imported into NEIGHBOR for neighbor-joining analyses to produce trees, then CONSENSE was used to recover a consensus tree and bootstrap values.

Intron Analysis

Intron positions were determined in *Glaucocystis* and *Cyanophora* by comparing the DNA sequence with the cDNA sequence directly; this is the most accurate method to determine precise intron insertion positions. Many gene sequences in Genbank are annotated, including intron positions; however, in most cases it is done using bioinformatics tools rather than by direct sequence analyses. A computer algorithm is

used to comparatively align sequences to look for putative intron splice junction sites and estimate intron positions (Federov et al. 2002). In practice, this method appears to misestimate many intron positions (see results section). Therefore, each sequence used in this investigation was examined individually to estimate intron positions that were most consistent with conservation of the protein sequence. Annotated *RPB*1 sequences, as well as raw sequences retrieved from genome sequencing projects, were compared to a global *RPB*1 alignment to determine whether conserved protein regions remained intact after putative introns were removed. The standard splice junction sites consist of GT at the 5' end and AG at the 3' end of an intron, but other combinations are possible. In all sequences for which introns had not been determined experimentally, intron positions were estimated by choosing putative splice junctions that restored the most conserved amino acid alignment possible. Splice junction sites were also examined carefully and preference was given to GT-AT intron boarders, unless they significantly disrupted conserved domains of the protein.

An amino acid alignment was created in CLUSTAL X of corrected *RPB*1 sequences from those organisms containing introns in the gene (note: intron positions were corrected before assembling the data matrix for phylogenetic analyses described above). Then each intron position and its insertion frame were noted; that is, whether the intron fell between codons, or interrupted the reading frame after the first or second nucleotide of a given codon. A table of intron positions as discrete characters was created using zero for absence and one for presence. This data set underwent parsimony analyses in PAUP using shared intron positions as characters. Under the assumption that

introns are far more likely to be lost than inserted convergently in the same locations (Rogozin et al. 2003), Dollo parsimony was used with 200 random sequence repetitions in each heuristic search. An initial analysis was performed on all gene sequences in the global *RPB*1 alignment that contained any introns.

Based on the results, and clear evidence of massive intron loss in various taxa, an additional intron character set was constructed by pooling all intron positions present in members of each major eukaryotic groups represented in this study (plants, animals, fungi, glaucocystophytes). This was done to explore parsimony analyses under the assumption that an intron present in any member of one of these groups was present in its common ancestor and, therefore, could be combined into one hypothetical ancestral data set for each lineage. Organisms from groups without multiple representatives were treated as individuals. Dollo parsimony again was used to create the most parsimonious tree based on shared intron positions.

RESULTS AND DISCUSSION

Isolation and Analysis of Primary Sequences

Cyanophora paradoxa is the most widely studied glaucocystophyte and is the one most commonly used in phylogenetic analyses; therefore, obtaining its RPB1 gene was attempted first. As discussed in the Materials and Methods section, it was necessary to obtain an axenic culture because of the presence of an additional eukaryote in the original culture that interfered with the amplification of *Cyanophora*'s RPB1 gene. The axenic culture ensured that only the RPB1 gene from *Cyanophora* would be amplified using the degenerate primers. Isolation of cDNA was attempted first since there are no introns to interfere with the annealing of the primers. Degenerate *RPB*1 primers were used to amplify overlapping segments between conserved regions A and G (Figure 1a). *Cyanophora*-specific primers were then designed to permit complete sequencing of each segment. In *Cyanophora*, this yielded a total sequence of 3114 base pairs from region A to region G. Isolation of sequence from region G to the end of the gene was not successful because of a pernicious PCR artifact from the reverse oligo dT linker primers used in RACE. This artifact was preferentially amplified under all conditions, so that the actual *Cyanophora* sequence could not be obtained from cDNA. No universally conserved domains adequate for PCR primer design are present downstream of the G region (Stiller and Hall 1997).

To obtain the intron positions in the sequenced regions, specific primers were made from the cDNA to amplify the genomic DNA sequence. Regions A through D were successfully sequenced, which yielded 1293 base pairs. The positions of introns

were determined by comparing the cDNA to DNA sequences. Ten introns were found between regions A and D. In the middle of sequencing the genomic DNA, new *Cyanophora paradoxa* cultures were ordered from Carolina Biological for extraction of additional DNA to permit isolation of genomic regions D through G. Because *Cyanophora*-specific primers were now available, it seemed unnecessary to obtain further axenic material from Europe, which is extremely slow-growing and expensive and would have delayed the sequencing effort. The D region sequence from the new DNA differed from the cDNA obtained from the axenic culture from CCAC, but only in some degenerate third codon positions. This slight difference in the nucleotide sequence suggests that there are multiple strains of *Cyanophora* present in culture collections.

There could also be two recently duplicated copies of *RPB*1 in *Cyanophora*, which would account for the amino acid sequence remaining the same but the nucleotide sequence differing slightly. However, whatever the cause, the specific primers created from the cDNA would not amplify any RPB1 DNA from regions D through G from the *Cyanophora* culture from Carolina Biological. Given the number of difficulties encountered with *Cyanophora*, it appeared worthwhile to attempt to isolate a complete and unequivocal *RPB*1 sequence from a second glaucocystophyte. In particular, to determine whether glaucocystophytes do indeed have a CTD, which could not be isolated from *Cyanophora*, the cDNA and DNA sequences from *Glaucocystis* were obtained.

A small portion of the *Glaucocystis* gene (regions D through F) had previously been isolated by Dr. John Stiller, indicating some shared intron positions with *Arabidopsis* (which prompted this study). The remaining 3' portion of the gene was

isolated as cDNA by RT-PCR, cloned and sequenced. The full sequence of 9544 bp was determined by walking across the clone using specific primers created from the cDNA, as well as primers from the oligo dT linker. The RPB1 gene was then PCR amplified from genomic DNA and sequenced fully to determine intron positions through the H region. Beyond conserved domains of the H region, RPB1 sequences cannot be aligned reliably, meaning any introns positions present could not be aligned with those of other eukaryotes; therefore, only cDNA was sequenced from region H through the polyA tail (Figure 1c).

Although not useful for broad-scale phylogenetic analyses, the 5' end was also needed since several organisms do have introns in the area leading up to region A. Primers included on the GeneRacer RNA oligo linker and *Glaucocystis*-specific A reverse primers were used in RACE to amplify the untranslated region beginning at the initiation site prior to the start codon, through the A region. The DNA was also sequenced in this region so that the intron positions could be located. In all, the entire RPB1 gene was sequenced from genomic and/or cDNA, a total of 9544 base pairs from the start to stop codon, plus both 5' and 3' untranslated sequence. Eleven introns were found between the start codon and region H. Intron positions are difficult to determine in the CTD region and they are not usually annotated in GenBank, so these were not determined for *Glaucocystis* or included in the comparative study.

CTD Analysis

In green plants, as in animals and fungi, the C-terminal domain of RPB1 plays an essential role in RNAP II transcription, and the region is strongly conserved in plant

genes. Because the CTD also is strongly conserved in animals and fungi, and is present in the protistan ancestors of both groups (Stiller and Hall 2002), I hypothesized that it also should be present in Glaucocystophytes if they are ancestral to green plants. *Glaucocystis* was found to have a well-conserved CTD with 35 canonical heptads and a more degenerate tail. This is very similar to the length and overall structure of the CTD from *Arabidopsis* and other green plants and algae (Figure 3), consistent with the hypothesis that glaucocystophytes could be the ancestor to green plants. While the comparable length and structure of the CTDs in *Glaucocystis* and *Arabidopsis* could indicate a shared history, it also could be coincidental. Both animals and fungi have C-terminal domains of varying lengths (Corden 1990) and it would be surprising if the CTD length is found to be strongly conserved across the entire plant kingdom. It will be interesting to see whether the plant CTD remains so conserved in its primary sequence features with the addition of new organisms in comparative sequence analyses.

Phylogenetic Analyses

Bayesian analyses of both the larger and smaller data sets resulted in similar trees. The analyses of the larger data set (Figure 4) of 46 *RPB*1 sequences, many of them partial, show that 100% of the time *Acanthamoeba* and *Dictyostelium* group with glaucocystophytes, and that green plants are the sister group to that monophyletic group. These analyses of the large data set also maintain the integrity of the CTD clade with a 100% Bayesian support value. Both parsimony and distance analyses of this larger data set produced clearly nonsensical trees in which green plants were not recovered as a monophyletic group. Maximum-likelihood analysis revealed that *Chlamydomonas* has a

much longer branch length than other plants. Combined with the effects of large regions of missing data, the highly divergent nature of the *Chlamydomonas* sequence may explain why it does not group with the other plants in parsimony or neighbor-joining analyses.

The smaller data set (which included a sub-sample of each major eukaryotic lineage) also groups *Acanthamoeba* with Glaucocystis 94% of the time in Bayesian analyses, and this monophyletic group remains the sister group to green plants (Figure 5). The CTD clade also is conserved with 95% Bayesian support. Maximum-likelihood branch lengths for the Bayesian tree were also calculated, as were parsimony and distance bootstrap values for commonly supported nodes. The results from both data sets indicate that Amoebozoans (*Acanthamoeba* and *Dictyostelium*) form a monophyletic group with the glaucocystophytes, with green plants as the sister group. Interestingly, these taxa have proven problematic in previous phylogenetic analysis, such as those using actin genes (Bhattacharya and Weber 1997), which did not recover relationships expected from investigations using ribosomal RNA and other sequences. Actin sequences from *Acanthamoeba, Dictyostelium*, and *Cyanophora* had to be excluded in order to obtain bootstrap support for the expected monophyly of the fungi and animals (Bhattacharya and Weber 1997). Likewise, sequences of *Acanthamoeba* and *Dictyostelium* are drawn to *Glaucocystis* and *Cyanophora* in *RPB*1 analysis, which is unexpected based on prior phylogenetic results (Bapteste et al. 2002). The sequences from these organisms seem to create discrepancies between phylogenetic analyses based on different gene sequences. This could be due to phylogenetic artifacts produced by sequences of these taxa, or

simply that their defined evolutionary positions are not yet accurately defined. The most probable explanation for these discrepancies, according to Bhattacharya and colleagues, is the high variation in rates of sequence divergence within and among these lineages (Bhattacharya et al. 1991, Bhattacharya and Weber 1997).

Because *Acanthamoeba* groups with the glaucocystophytes in all phylogenetic analyses of *RPB*1 sequences, and red algae are not found to be monophyletic with green plants and glaucocystophytes, I determined whether several *a priori* hypotheses developed from other phylogenetic investigations could be rejected based on *RPB*1 data. Using the smaller and more complete data set, constrained trees were created and KH and SH tests were run to see whether they were significantly worse than the Bayesian consensus tree. Both KH and SH tests gave similar results (Table 2). The Bayesian tree was found to have the highest likelihood of all trees tested. Neither the tree in which *Glaucocystis* was constrained in a monophyletic relationship with green plants, nor the tree in which *Acanthamoeba* was constrained as sister to the Opisthokonts (animals + fungi) could be rejected. The tree that contrained *Glaucocystis*, green plants and rhodophytes as a monophyletic group, however, was rejected with a p value of 0.001 in the SH test and 0.002 in the KH test. These results are in conflict with the proposition that glaucocystophytes, green plants and red algae form a monophyletic kingdom "Plantae," which has been proposed primarily based on the theory of a single plastid origin (Martin et al. 1998, Moreira et al. 2000). Phylogenetic analyses of many plastid genes and some nuclear genes give some support to this theory, although none provide

statistically significant evidence thus far (Bhattacharya and Weber 1997, Nozaki et al. 2003).

Finally, trees produced through parsimony and neighbor-joining analyses do not recover some known relationships (Figures 6 and 7, respectively). In particular, the green alga *Chlamydomonas* appears to produce long-branch artifacts and is not recovered in a monophyletic group with other green plants. Therefore, these analyses do not appear to be reliable using this data set, particularly since the focus of the study is a putative relationship of plants to glaucocystophytes.

Intron Analysis

Most unicellular protists do not contain many introns. Glaucocystophytes, which are also unicellular, are different in that both *Glaucocystis* and *Cyanophora* contain a large number of introns in their RPB1 genes. The number of introns in eukaryotic species differ widely, which is likely due to cases of intron insertion, loss, and high intron turnover rate (Rogozin et al. 2003). Results of broad scale comparative investigations suggest the possibility that the common ancestor of many eukaryotic groups had an intron-rich genome; the majority of ancestral introns appeared to have survived in plants and vertebrates but have been lost in yeasts, nematodes and arthropods (Rogozin et al. 2003). For many protist groups with uncertain phylogenetic affinities, the situation is less clear; some probably have lost ancestral introns, as have yeasts and some animals, while others may never have contained introns at all (Rogozin et al. 2003)

Like plants and some vertebrates, however, glaucocystophytes have retained a large number of introns. As stated earlier, *Cyanophora* has ten introns from region A to

region D alone, while *Glaucocystis* has eleven introns between the start codon and the end of region G. *Cyanophora* and *Glaucocystis* share one intron position that is not found in any other organism, and each have three entirely unique introns. One or both of the glaucocystophytes share a total of nine introns with plants. Of these shared intron positions, seven are also shared with animals. This indicates that at least these seven introns are most likely ancestral and not shared-derived characters linking glaucocystophytes with plants (Figure 8 and Table 3).

A more rigorous phylogenetic analysis was undertaken to see if any relationships became apparent based on shared intron positions. I made a table of individual intron insertion sites as characters for use in Dollo parsimony analyses. An initial tree found by heuristic search including all organisms that contain any *RPB*1 introns (Figure 9) was created. This tree recovered *Chlamydomonas* and mouse as a monophyletic group, with *Glaucocystis* as sister to those two. The remaining green plants formed a clade, which grouped with the former three organisms with a bootstrap value of 67. *Chlamydomonas* and mouse each contain many RPB1 introns, by far the most of all organisms in this study, a number of which are shared. Most of these are probably ancestral introns that have been lost in many of the other organisms investigated. The tree also recovered fungi as a clade, but the remainder of the tree was unresolved, including the other animals, due to convergent intron loss and the presence of many unique introns positions in different organisms.

To try to compensate for the large amount of parallel intron loss that appears to have occurred, a second matrix of intron insertion sites was created. Introns were

combined from organisms in each well-defined eukaryotic group (animals, green plants, fungi, glaucocystophytes), including those found in the protistan ancestral organism (if known). This assumes that any intron found in a given group is most likely to be ancestral, and reduces the complication of parallel intron losses in different lineages. The tree recovered with these combined introns revealed that plants and animals group together, with glaucocystophytes sister to that clade (Figure 10). This tree corresponds to the tree made using all individual organisms. Plants and animals most likely grouped together because of the large number of shared positions in mouse and *Chlamydomonas*. Changes in character states were examined in MacClade to determine the number of changes supporting each node, and whether they were losses or gains. Plants and fungi each were supported as a group by 20 changes and animals by 22. Most of these changes are inferred losses. Unlike yeasts, basidiomycete fungi have a large number of introns, but most are unique to the group and are not shared with either animals or plants. Given that animals and fungi are believed to be closely related (Baldauf and Roger 2000), this suggests that fungi lost most of the introns that were present in their common ancestor, and then gained new introns in different positions later. Alternatively, RPB1 intron positions could be taken as evidence that the animal-fungi sister relationship recovered in phylogenetic trees is inaccurate.

Chlamydomonas is the green plant with the most introns, several of which occur after region G; this also is true in mouse and some of these positions are shared. Introns are not found in this region in the other plants. Based on their presence or absence in *Chlamydomonas*, lack of introns in this region in other plants is most likely due to intron

loss. Interestingly, *Glaucocystis* also contains no introns after the G region, which could be coincidental or it could reflect some history of loss shared with green plants. Unfortunately, for the reasons discussed earlier, I was unable to isolate this region from *Cyanophora*, *Cyanophora* has many more introns than *Glaucocystis* and it would be interesting to see if introns found after region D shared positions with *Chlamydomonas*, and whether the *Cyanophora* RPB1 gene contains introns after the G region. The overall analysis of shared intron positions is consistent with the hypothesis that glaucophytes and plants are closely related, since they do share several insertion positions that could be interpreted as shared-derived characters. Intron evidence also is consistent with the hypothesis that the common ancestor of many, if not most eukaryotes, had an intron-rich genome and that ancestral introns have survived in glaucocystophytes, green plants and vertebrates, but were lost in yeasts and many other simple eukaryotes. This scenario is further complicated by the need to explain additional impressive gains of new introns in basidiomycete fungi and several protists, after they first lost virtually all ancestral intron positions. Overall, there appears to be no simple explanation for the distribution of *RPB*1 introns, regardless of whether they are interpreted as ancestral or derived features.

CONCLUSIONS

The aim of this study was to use three distinct and complementary approaches to examine the gene encoding the largest subunit of RNA polymerase II (*RPB*1). This combination of data provides multiple lines of evidence to test the hypothesis that glaucocystophytes are the direct protistan ancestor of green algae and land plants. First, the primary sequence of *RPB1* was determined. Bayesian analyses of *Cyanophora* and *Glaucocystis RPB1* sequence placed glaucocystophytes as sister group to green plants on the evolutionary tree, but only with the assumption that mycetezoans are a non-photosynthetic descendent of their common ancestor. Trees recovered from parsimony and distance analyses also tend to place these groups together; however, other recognized relationships are not supported by these trees, apparently due to phylogenetic artifacts. In no case do phylogenetic analyses reject the hypothesis that green plants and glaucocystophytes form a monophyletic group.

A conserved RNAP II C-terminal domain was predicted to be present in glaucocystophytes if they are derived from a common ancestor with green plants. *Glaucocystis* does have repeated heptapeptides in its CTD of the same conserved sequence present in green plants and other CTD-containing organisms. The presence of a canonical CTD in the glaucocystophytes suggests that they share similar RNAP II transcriptional functions with green plants and other developmentally complex eukaryotes. Indeed, the established ancestors of animals and fungi both contain a well-conserved RNAP II CTD. Therefore, glaucocystophytes remain a reasonable candidate

for the protistan ancestor of green plants, if green plants follow the same pattern of CTD conservation as other major eukaryotic groups.

Finally, evidence from shared intron positions suggest a reasonably recent relationship between glaucocystophytes and green plants. There is still some debate over the origin and history of introns in eukaryotic evolution, but it is believed that unique introns (not ancestral) are unlikely to have inserted in exactly the same location independently. Therefore, the presence of introns in the same location in different lineages suggests that the introns are shared-derived characters. The overall analysis of shared intron positions is consistent with the hypothesis that glaucocystophytes and plants are closely related, since they do share several insertion positions that could be interpreted as unique, shared-derived characters. A complete sequence of the *Cyanophora* RPB1 gene would probably reveal even more shared intron positions with *Chlamydomonas*, because both of these organisms have the most intron-rich genes for members of their respective groups.

The three aspects of the RPB1 gene examined in this study are consistent with the sister relationship between green plants and glaucocystophytes, but do not provide compelling support. Therefore, glaucocystophytes may or may not represent the direct protistan ancestors to green plants. Overall, the data do provide strong support for the hypothesis that green plants and glaucocystophytes are more closely related to each other and to other eukaryotes, than either is to red algae. This is inconsistent with the hypothesis of an inclusive kingdom Plantae, comprising all photosynthetic groups with primary plastids.

Figure 1. **A.** Schematic of the RPB1 gene showing relative locations of conserved domains A through H. Degenerate primer sites are shown within the conserved regions. **B.** Strategy for isolating the 5' region of *RPB1,* including start codon, untranslated region to the initiation site, ligation point for GeneRacer RNA Oligo linkers. The GeneRacer primers and the specific A reverse primer used to obtain the 5' end also are shown. **C.** The *RPB*1 cDNA 3'end, including the CTD encoding region, poly-A tail and 3' GeneRacer Oligo primers. The gene-specific primer sites are shown in the conserved regions.

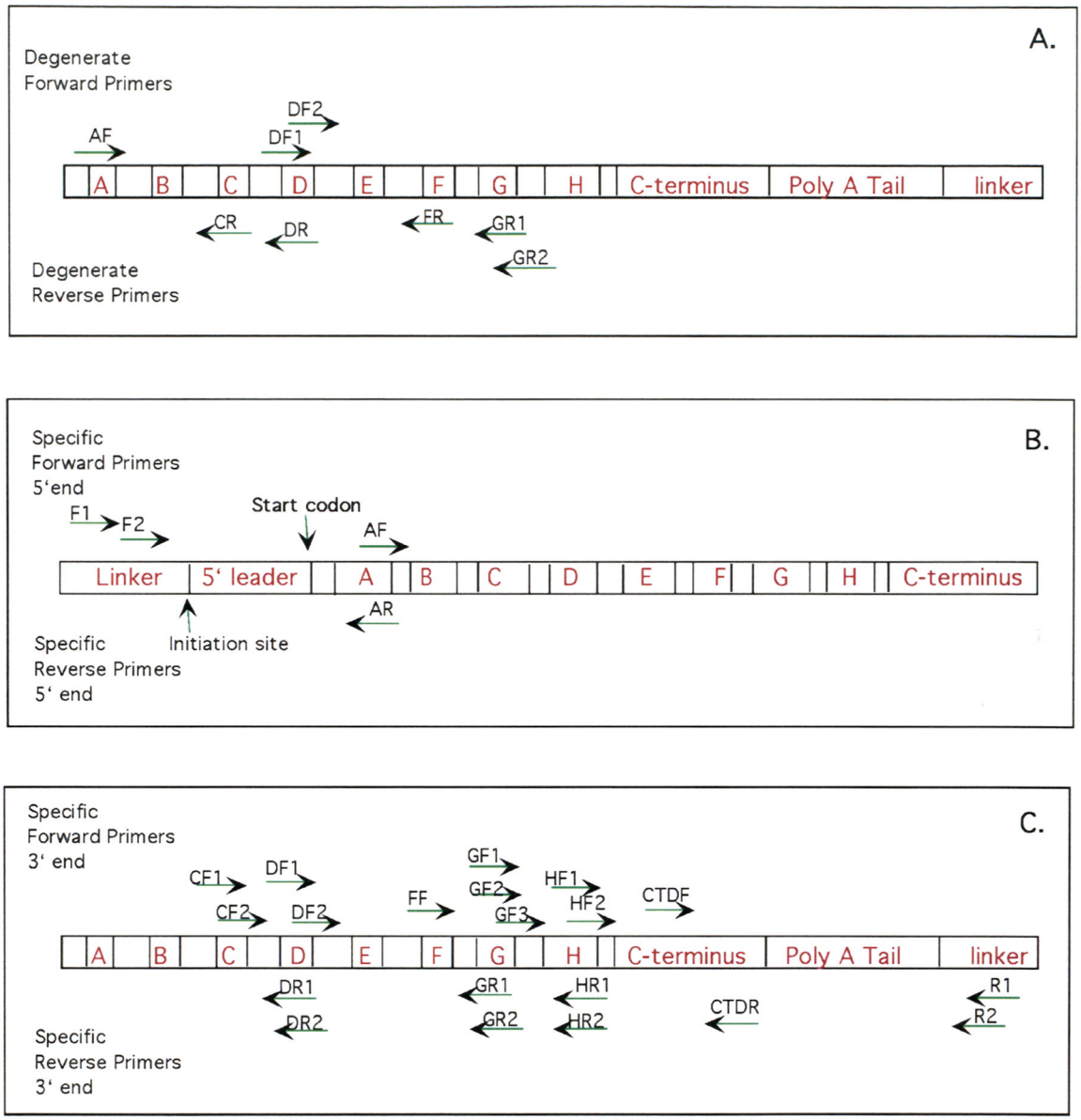


Figure 1. The RPB1 gene with conserved regions A-H, degenerate and specific primer locations.

Figure 2. Oligo linkers, with primer sequences, used to obtain the complete 3' and 5' ends of *RPB*1. The GeneRacer RNA Oligo 43 bp linker is ligated to the 5' end of complete mRNA after dephosphorylation of incomplete messages and removal of the cap structure from complete messages. The GeneRacer Oligo dT linker is a 54 bp oligonucleotide that contains a poly-dT tail of 18 nucleotides to prime first strand cDNA synthesis through reverse transcription. The sequence also contains priming sites for the GeneRacer 3' and the GeneRacer 3' nested primers.

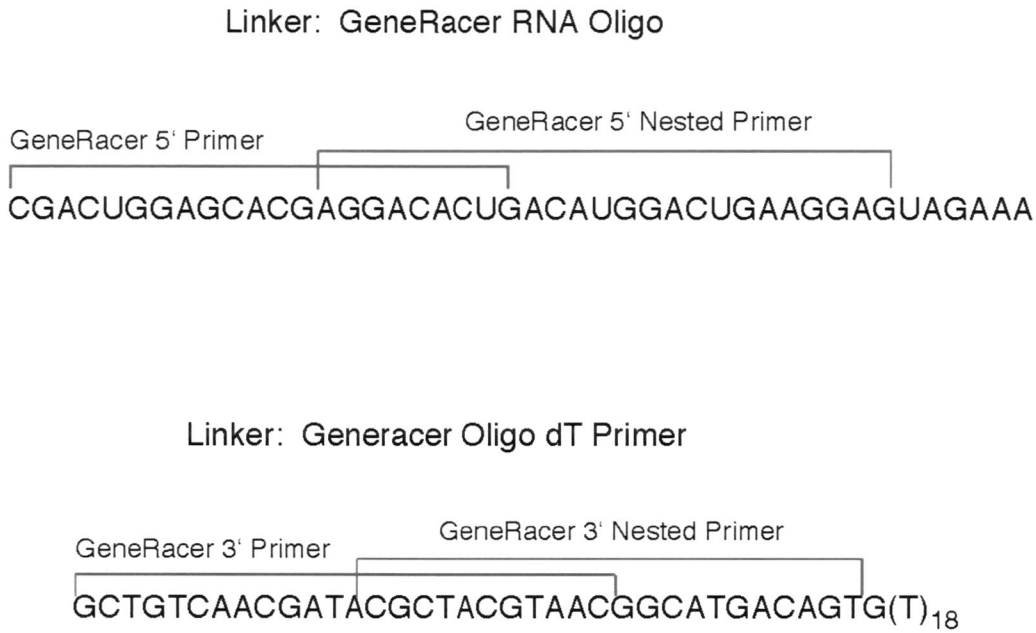


Figure 2. Oligo linker used in RACE isolation of RPB1 5' and 3' ends.

Figure 3. Comparison of C-terminal domains from *Glaucocystis* and *Arabidopsis*. Both CTDs are of comparable length and maintain the tandemly repeated heptapeptide (YSPTSPS) throughout most of the sequence. Each has a leader sequence and a tail end that becomes more degenerate.

Glaucosystis	***Arabidopsis***
YAESPFVDGSFSPHRG	SPNMRLSPMSDAQFSP
YSPSSPP	YVGGMAF
YSPRSPGF	SPSSSPG
DYAASPS	YSPSSPG
YSPASPS	YSPTSPG
YSAASPR	YSPTSPG
YSPTSPG	YSPTSPG
YSASSPS	YSPTSPT
YSPTSPS	YSPSSPG
YSPTSPS	YSPTSPA
YSPTSPS	YSPTSPS
YSPTSPA	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPG	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPR	YSPTSPS
YSPTSPS	YSPTSPA
YSPTSPS	YSPTSPA
YSPTSPG	YSPTSPA
YSPTSPS	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPS	YSPTSPA
YSPTSPS	YSPTSPG
CSPTSPS	YSPTSPS
YSPTSPS	YSPTSPS
YSPTSPS	YGPTSPS
YSPSSPS	YNPQSAK
CSPASPT	YSPSIA
YSPSSPS	YSPSNAR
YSPSSPS	LSPASP
YSPSSPS	YSPTSPN
YSPSSPS	YSPTSPS
YSPSSPQ	YSPTSPS
YTPASAN	YSPSSPT
YSPPSPL	YSPSSP
HSSPA	YSSGASPD
YSPTSQ	YSPSAG
YSPTSPA	YSPTLPG
YGASSPA	YSPSSTGQ
YTPQSNLSPASLPSPALSPVSQKSPPQDR	YTPHEGDKKDKTGKKDASKDDKGNP

Figure 3. CTD comparison

Figure 4. Phylogenetic tree recovered through Bayesian inference on an alignment of inferred amino acid sequences of RPB1 regions A to H from 46 organisms. Bayesian support values are shown for the clades relevant to the relationship between glaucocystophytes and green plants. The analysis applied a discrete estimate of 1 invariable + 4 γ-distributed rates among sites and a JTT model for probabilities of changes among amino acids. The tree was rooted with *Giardia* species.

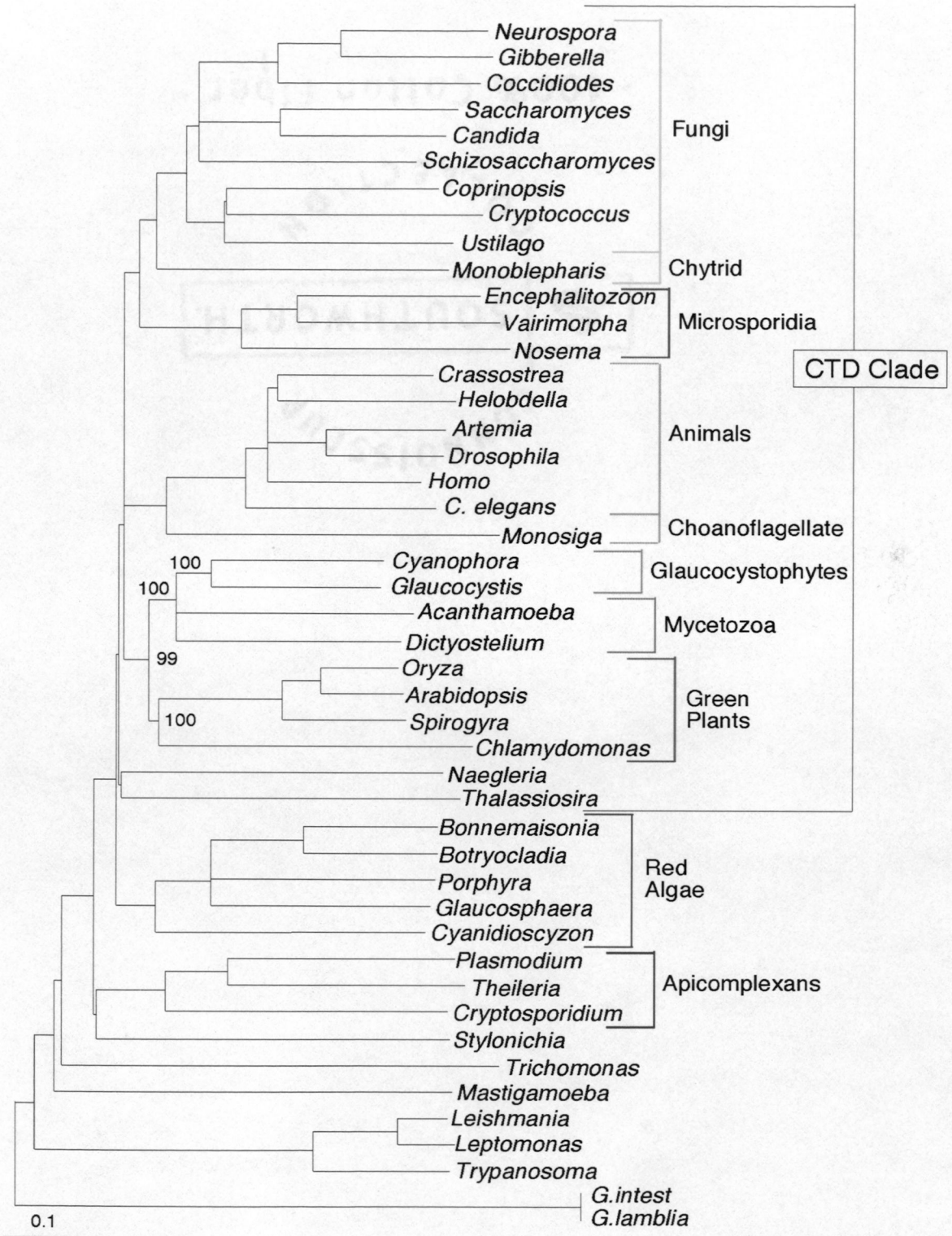


Figure 4. Tree recovered by Bayesian Inference of 46 sequences.

Figure 5. Phylogenetic tree recovered through Bayesian inference on an alignment of inferred amino acid sequences of RPB1 from 28 organisms in which the complete sequence from regions A through H is known. The analysis applied a discrete estimate of 1 invariable + 4 γ-distributed rates among sites and a JTT model for probabilities of changes among amino acids. The first number shown represents the Bayesian values. The second two numbers indicate bootstrap values from distance and parsimony analyses respectively. The bootstrap values were inferred from100 replications. The distance bootstrap values were inferred using the PROTDIST method with a neighbor-joining tree reconstruction. The parsimony values were inferred using the PROTPARS method. The stars indicate that all three analyses give 100% support to the clade. Only bootstrap values >50% are shown.

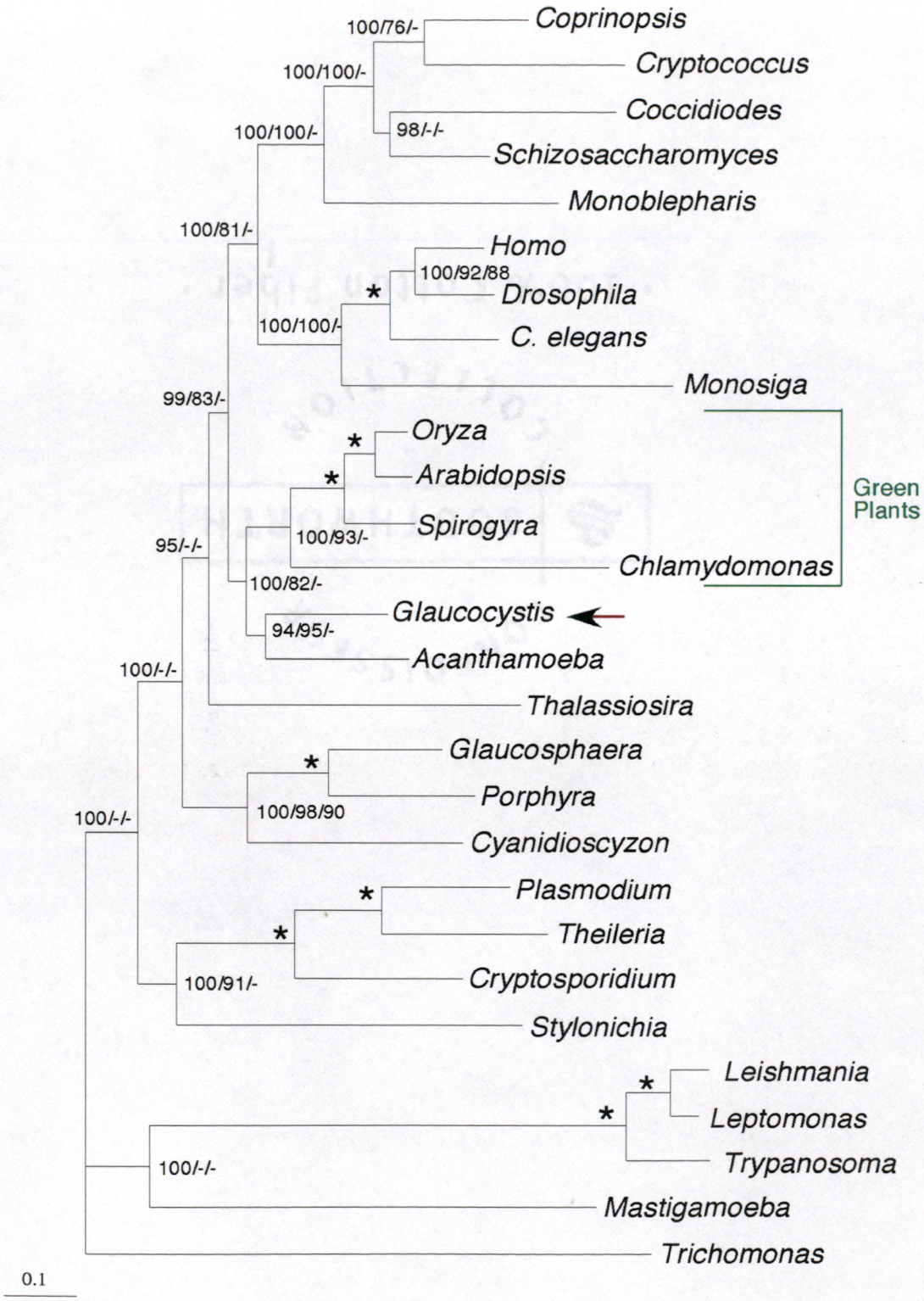


Figure 5. Tree recovered by Bayesian inference of 28 sequences.

Figure 6. Single most parsimonious tree based on RPB1 sequences recovered through PROTPARS analysis in PHYLIP. Arrow points out *Chlamydomonas*, which is not recovered with green plants in this analysis.

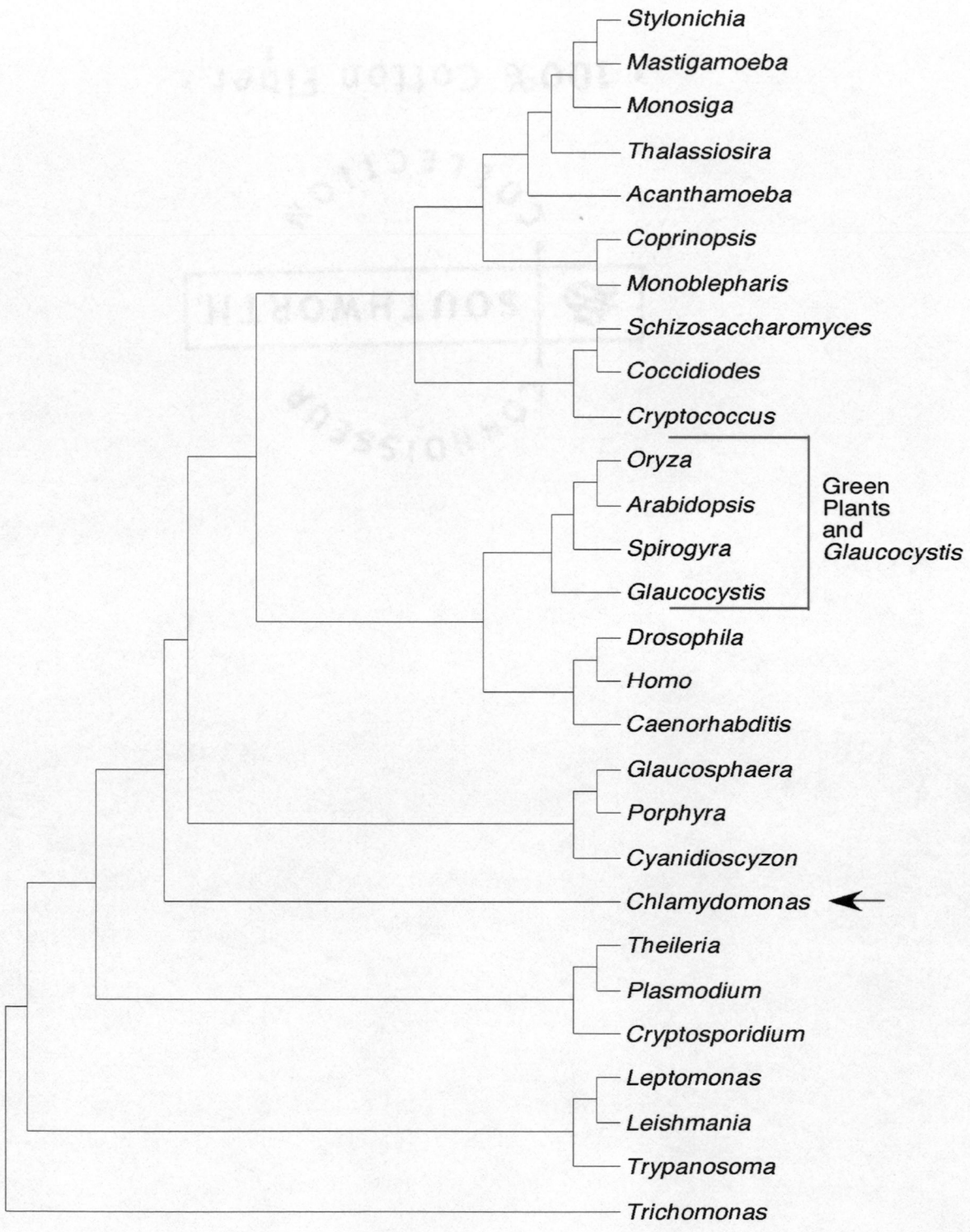


Figure 6. Most parsimonious tree from PROTPARS analysis.

Figure 7. PROTDIST was used to compute a distance measure for the RPB1 sequences, using a JTT matrix model with the gamma-distributed rates of change among sites. The distance maxtrix was used in neighbor-joining analysis. Branch lengths are shown and the tree is rooted with *Trichomonas*.

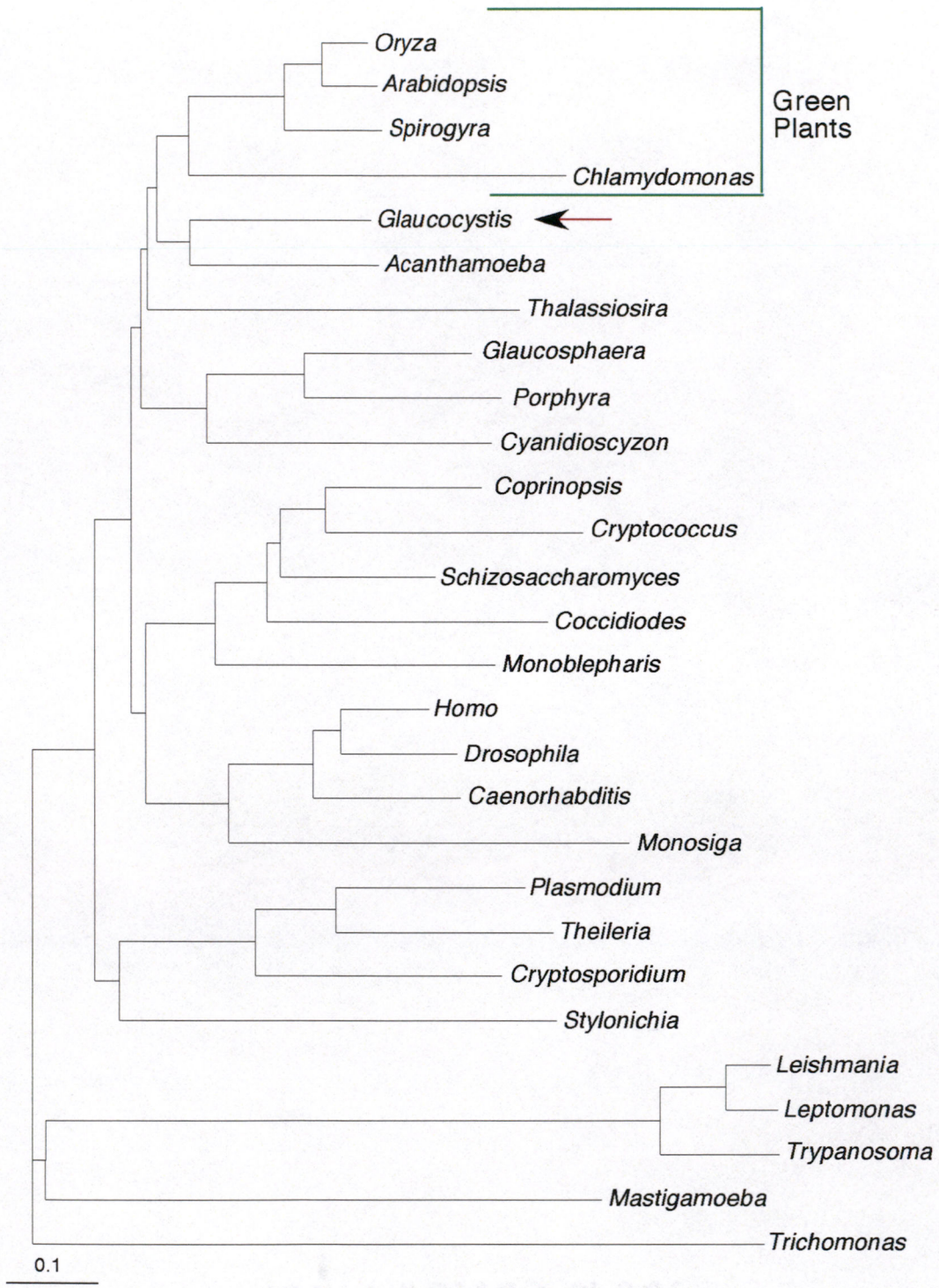
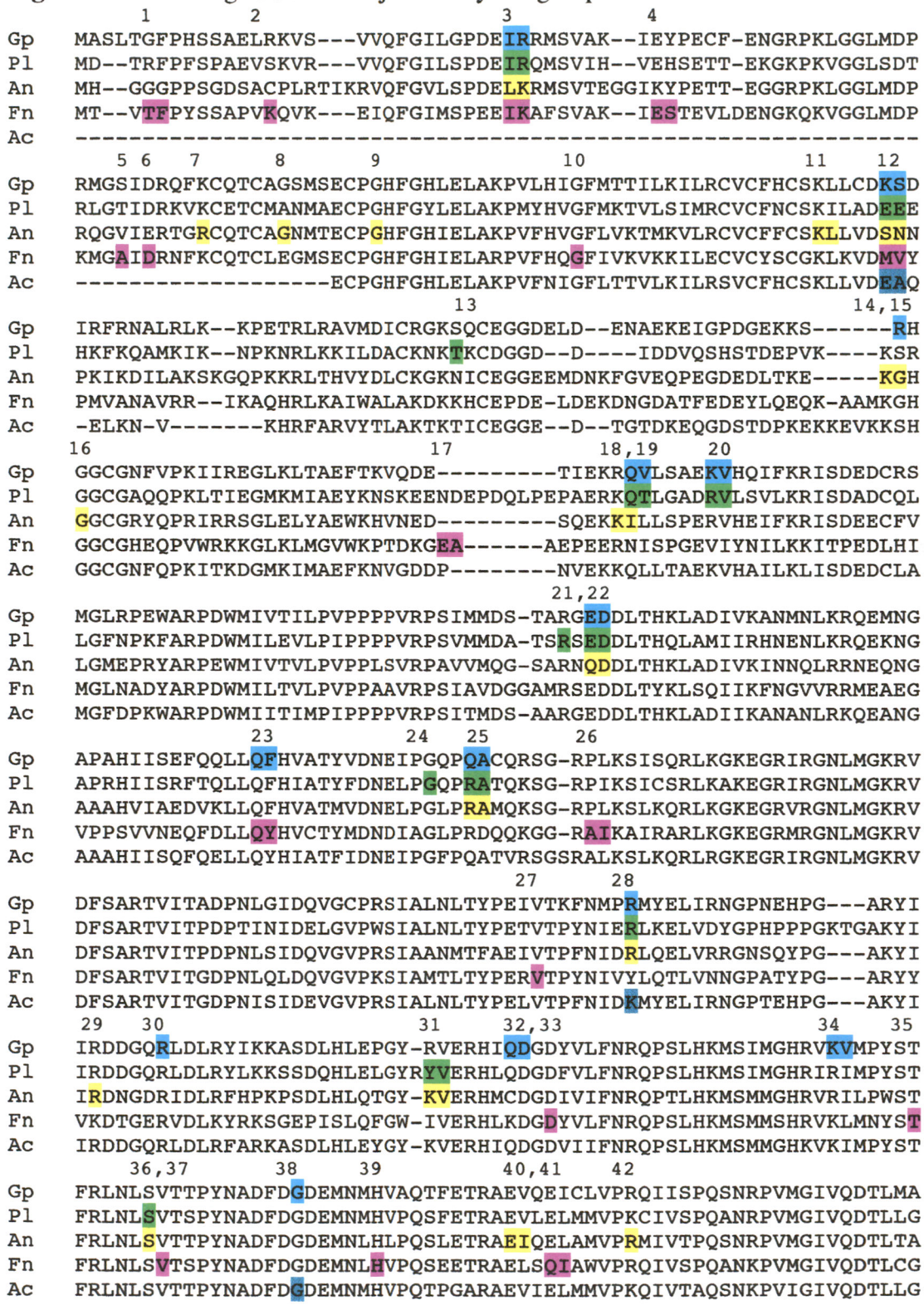
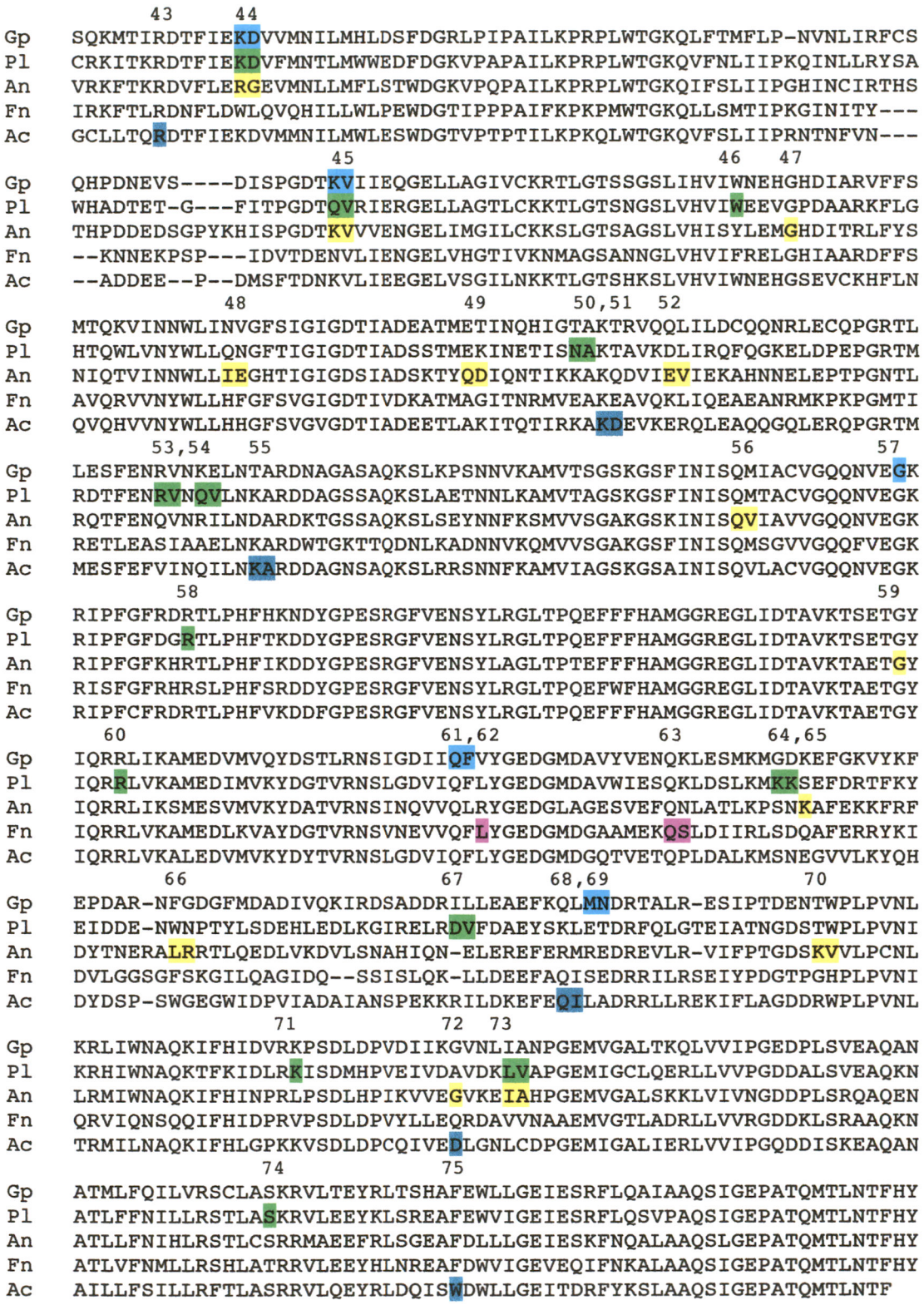


Figure 7. Distance tree from neighbor-joining.

Figure 8. Introns combined from each of the major eukaryotic groups represented in this investigation. If the intron comes between two codons, both residues are highlighted. If the intron interrupts the codon, only that codon is highlighted. The numbers above the amino acids correspond with the numbers on Table 3, which lists introns from all organisms individually.
Gp = glaucocystophytes – (*Cyanophora* and *Glaucocystis*)
Pl = plants (*Oryza, Arabidopsis, Spirogrya,* and *Chlamydomonas*)
An = animals (*Drosophila, Mus musculus, C. elegans, and Monosiga)*
Fn = fungi *(Schizosaccharomyces, Coprinopsis, Coccidioides, Cryptococcus*, and *Monoblepharis)*
Ac = *Acanthamoeba*

Figure 8. Intron alignment for major eukaryotic groups.




43 44
Gp SQKMTIRDTFIEKDVVMNILMHLDSFDGRLPIPAILKPRPLWTGKQLFTMFLP-NVNLIRFCS
Pl CRKITKRDTFIEKDVFMNTLMWWEDFDGKVPAPAILKPRPLWTGKQVFNLIIPKQINLLRYSA
An VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHS
Fn IRKFTLRDNFLDWLQVQHILLWLPEWDGTIPPPAIFKPKPMWTGKQLLSMTIPKGINITY---
Ac GCLLTQRDTFIEKDVMMNILMWLESWDGTVPTPTILKPKQLWTGKQVFSLIIPRNTNFVN---
45 46 47
Gp QHPDNEVS----DISPGDTKVIIEQGELLAGIVCKRTLGTSSGSLIHVIWNEHGHDIARVFFS
Pl WHADTET-G---FITPGDTQVRIERGELLAGTLCKKTLGTSNGSLVHVIWEEVGPDAARKFLG
An THPDDEDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYS
Fn --KNNEKPSP---IDVTDENVLIENGELVHGTIVKNMAGSANNGLVHVIFRELGHIAARDFFS
Ac --ADDEE--P--DMSFTDNKVLIEEGELVSGILNKKTLGTSHKSLVHVIWNEHGSEVCKHFLN
48 49 50,51 52
Gp MTQKVINNWLINVGFSIGIGDTIADEATMETINQHIGTAKTRVQQLILDCQQNRLECQPGRTL
Pl HTQWLVNYWLLQNGFTIGIGDTIADSSTMEKINETISNAKTAVKDLIRQFQGKELDPEPGRTM
An NIQTVINNWLLIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTL
Fn AVQRVVNYWLLHFGFSVGIGDTIVDKATMAGITNRMVEAKEAVQKLIQEAEANRMKPKPGMTI
Ac QVQHVVNYWLLHHGFSVGVGDTIADEETLAKITQTIRKAKDEVKERQLEAQQGQLERQPGRTM
53,54 55 56 57
Gp LESFENRVNKELNTARDNAGASAQKSLKPSNNVKAMVTSGSKGSFINISQMIACVGQQNVEGK
Pl RDTFENRVNQVLNKARDDAGSSAQKSLAETNNLKAMVTAGSKGSFINISQMTACVGQQNVEGK
An RQTFENQVNRILNDARDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGK
Fn RETLEASIAAELNKARDWTGKTTQDNLKADNNVKQMVVSGAKGSFINISQMSGVVGQQFVEGK
Ac MESFEFVINQILNKARDDAGNSAQKSLRRSNNFKAMVIAGSKGSAINISQVLACVGQQNVEGK
58 59
Gp RIPFGFRDRTLPHFHKNDYGPESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTSETGY
Pl RIPFGFDGRTLPHFTKDDYGPESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTSETGY
An RIPFGFKHRTLPHFIKDDYGPESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGY
Fn RISFGFRHRSLPHFSRDDYGPESRGFVENSYLRGLTPQEFWFHAMGGREGLIDTAVKTAETGY
Ac RIPFCFRDRTLPHFVKDDFGPESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGY
60 61,62 63 64,65
Gp IQRRLIKAMEDVMVQYDSTLRNSIGDIIQFVYGEDGMDAVYVENQKLESMKMGDKEFGKVYKF
Pl IQRRLVKAMEDIMVKYDGTVRNSLGDVIQFLYGEDGMDAVWIESQKLDSLKMKKSEFDRTFKY
An IQRRLIKSMESVMVKYDATVRNSINQVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRF
Fn IQRRLVKAMEDLKVAYDGTVRNSVNEVVQFLYGEDGMDGAAMEKQSLDIIRLSDQAFERRYKI
Ac IQRRLVKALEDVMVKYDYTVRNSLGDVIQFLYGEDGMDGQTVETQPLDALKMSNEGVVLKYQH
66 67 68,69 70
Gp EPDAR-NFGDGFMDADIVQKIRDSADDRILLEAEFKQLMNDRTALR-ESIPTDENTWPLPVNL
Pl EIDDE-NWNPTYLSDEHLEDLKGIRELRDVFDAEYSKLETDRFQLGTEIATNGDSTWPLPVNI
An DYTNERALRRTLQEDLVKDVLSNAHIQN-ELEREFERMREDREVLR-VIFPTGDSKVVLPCNL
Fn DVLGGSGFSKGILQAGIDQ--SSISLQK-LLDEEFAQISEDRRILRSEIYPDGTPGHPLPVNI
Ac DYDSP-SWGEGWIDPVIADAIANSPEKKRILDKEFEQILADRRLLREKIFLAGDDRWPLPVNL
71 72 73
Gp KRLIWNAQKIFHIDVRKPSDLDPVDIIKGVNLIANPGEMVGALTKQLVVIPGEDPLSVEAQAN
Pl KRHIWNAQKTFKIDLRKISDMHPVEIVDAVDKLVAPGEMIGCLQERLLVVPGDDALSVEAQKN
An LRMIWNAQKIFHINPRLPSDLHPIKVVEGVKEIAHPGEMVGALSKKLVIVNGDDPLSRQAQEN
Fn QRVIQNSQQIFHIDPRVPSDLDPVYLLEQRDAVVNAAEMVGTLADRLLVVRGDDKLSRAAQKN
Ac TRMILNAQKIFHLGPKKVSDLDPCQIVEDLGNLCDPGEMIGALIERLVVIPGQDDISKEAQAN
74 75
Gp ATMLFQILVRSCLASKRVLTEYRLTSHAFEWLLGEIESRFLQAIAAQSIGEPATQMTLNTFHY
Pl ATLFFNILLRSTLASKRVLEEYKLSREAFEWVIGEIESRFLQSVPAQSIGEPATQMTLNTFHY
An ATLLFNIHLRSTLCSRRMAEEFRLSGEAFDLLLGEIESKFNQALAAQSLGEPATQMTLNTFHY
Fn ATLVFNMLLRSHLATRRVLEEYHLNREAFDWVIGEVEQIFNKALAAQSIGEPATQMTLNTFHY
Ac AILLFSILLRFTLASRRVLQEYRLDQISWDWLLGEITDRFYKSLAAQSIGEPATQMTLNTF

```
                                                    76      77              78
Gp   AGVSSKNVTLGVPRLKEIINVAKKVKTPSLVVYLKEYCRRDSEKAKAVQCQLEFTTLRHVTAA
Pl   AGVSAKNVTLGVPRLREIINVAKRIKTPSLSVYLTPEASKSKEGAKTVQCALEYTTLRSVTQA
An   AGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKDILCRLEHTTLRKVTAN
Fn   AGVASKSVTGGVPRLKEIINVAVNIRTPALNVYLEPEYSKTEEDAHQIMRKLTYTRLRDITAT
Ac   AGVSAKNVTLGVPRLKELINIAKTIKTPSLTVYLEPHCSRDHEAAKNVQCSLQHTTLRDVTAA
                                    79     80
Gp   TEIYYDPDLQNTVIEEDREFVQCYYEIPDIPQSDFARMSTWLLRIELNREMMTDKKLTMKDID
Pl   TEVWYDPDPMSTIIEEDFEFVRSYYEMPDEDVS-PDKISPWLLRIELNREMMVDKKLSMADIA
An   TAIYYDPNPQSTVVAEDQEWVNVYYEMPDD----VARISPWLLRVELDRKHMTDRKLTMEQIA
Fn   VEIFYDPKLDSTDIEEDKDFVDAFFAIPDEDIR-LELHSPWLLRLELDRAKVLEGGYEMSQIV
Ac   TEIFYDPDPVNTVITEDQDFVRAYFLMPEEEIN-TSNLSPWLLRIELNREKMTDTKLSMQEIA
          81                          82          83
Gp   ERIAVGLWAGDLHVIYSDDNAAKLSLHIRIKNQEEEAKNQEEEASG--DEHEFLRRIEGDMLT
Pl   EKINLE-FDDDLTCIFNDDNAQKLILRIRIMNDEGPKGELQDESAE--D-DVFLKKIESNMLT
An   EKINAG-FGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNMLT
Fn   DAIAET-VGKDVFVIHSEDNAPKLVIRIRVVAEKED----EE-LLG--DEDMFLKRIEGTLLD
Ac   ERIHAD-FGGDLNCIYNDDNADKLILRIRINNDEENKAPDQEGSVG--DDDVFLKQIESNMLT
                84
Gp   EMALRGIAQIRKVFMR-----EAKKITFDADHKYMHEKEWVLDTEGCNLLEVMSCQDVDHTRT
Pl   EMALRGIPDINKVFIK-----QVRKSRFDEEGGFKTSEEWMLDTEGVNLLAVMCHEDVDPKRT
An   DMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRT
Fn   QVELGGITGITRVFIS-----EGKQVVVSQNGEYDQEKEWFLETDGINLKEVMAVDGVNAFRT
Ac   EMDLKGIEGIKKVFIR----EDKNKVAIDARGEYTKANELVLDTEGTNLLAVMSHPDVDHTRT
              85    86      87
Gp   TSNDIVEIIQVLGIEAVRQAALKELRDVISFDGSYVNYRHLATLADVMTYRGHLMSITRHGIN
Pl   TSNHLIEIIEVLGIEAVRRALLDELRVVISFDGSYVNYRHLAILCDTMTYRGHLMAITRHGIN
An   TSNDIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVN
Fn   YSNNCYEVYETLGIEAARNALYKELNGVIEMGGSYVNYRHLALLCDLMCSKGALMSITRHGIN
Ac   TSNNIIETIEVLGIEAVRNALLRELRNVISFDGAYVNYRHLAILADVMTYRGHLMAITRHGIN
                  88    89
Gp   RVPTGAMMRCSFEETVEILMDAAVYAETDHCRSVLKNIMLGQLCQLGTG
Pl   RNDTGPLMRCSFEETVDILLDAAAYAETDCLRGVTENIMLGQLAPIGTG
An   RQDTGPLMKCSFEETVDVLMEAAAHGESYPMKGVSENIMLGQLAPAGTG
Fn   RTDAGALSRAAFEETVEILLEAAAVGDVDDCKGVAENVLLGQMAPMGTG
Ac   RVETGCLMRCSFEETADILLEAATFSELDPLSGVSENILLGQLPPLGTG
```

Figure 9. Tree from Dollo parsimony of 19 organisms with known introns in RPB1. A data matrix was created from intron absence or presence (encoded as 0 = absent or 1 = present). A heuristic search was done with random stepwise addition of 200 repetitions in PAUP. Tree is a 50% majority rule of two equally parsimonious trees.

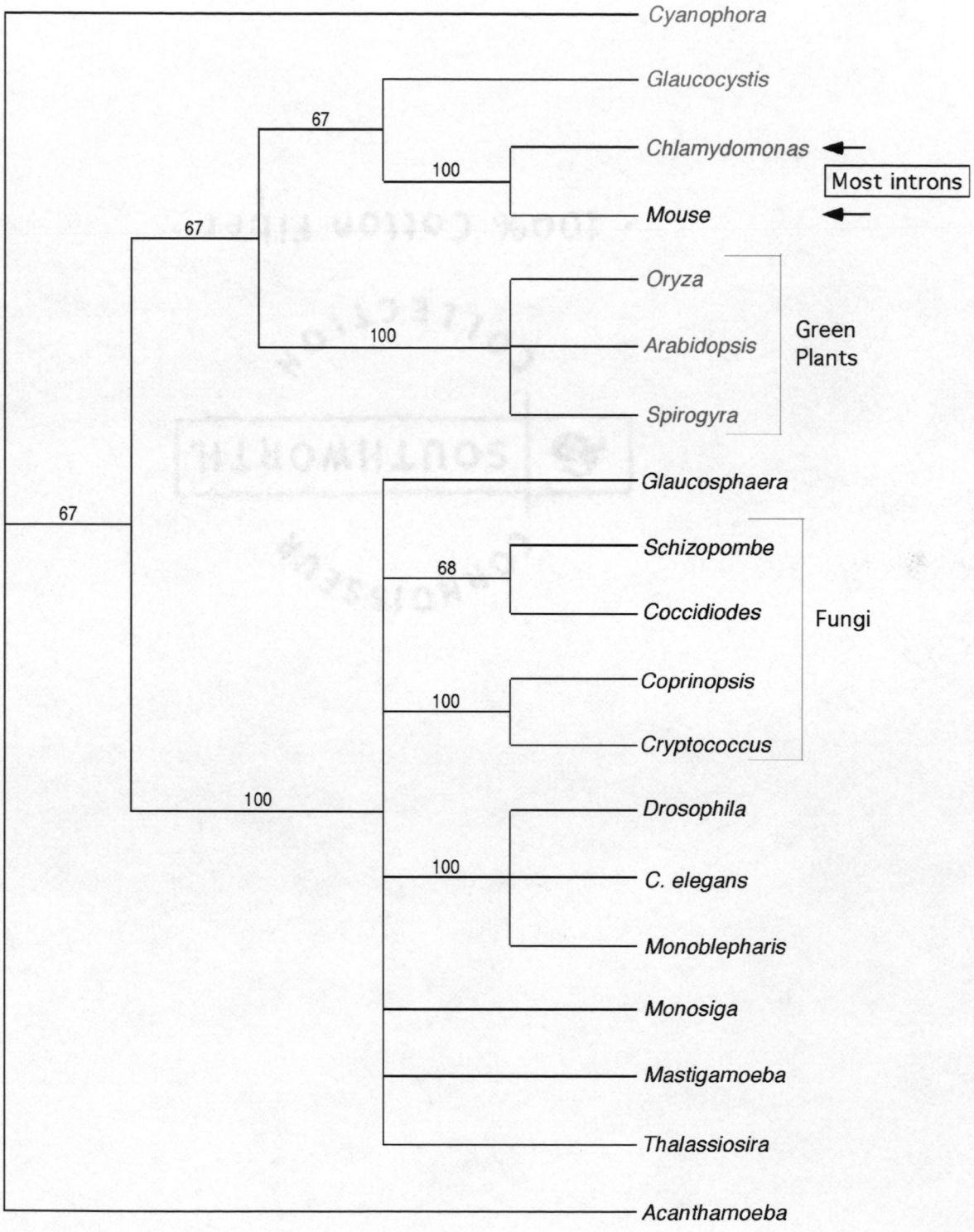


Figure 9. Tree created by Dollo parsimony.

Figure 10. Phylogenetic tree created using Dollo parsimony. Another intron character set was constructed by pooling all intron positions present in members of each major eukaryotic group represented in this study (plants, animals, fungi, glaucocystophytes). This data matrix then was used in phylogenetic analysis in PAUP. A heuristic search was done with random stepwise addition of 200 repetitions. Branch lengths indicate number of changes.

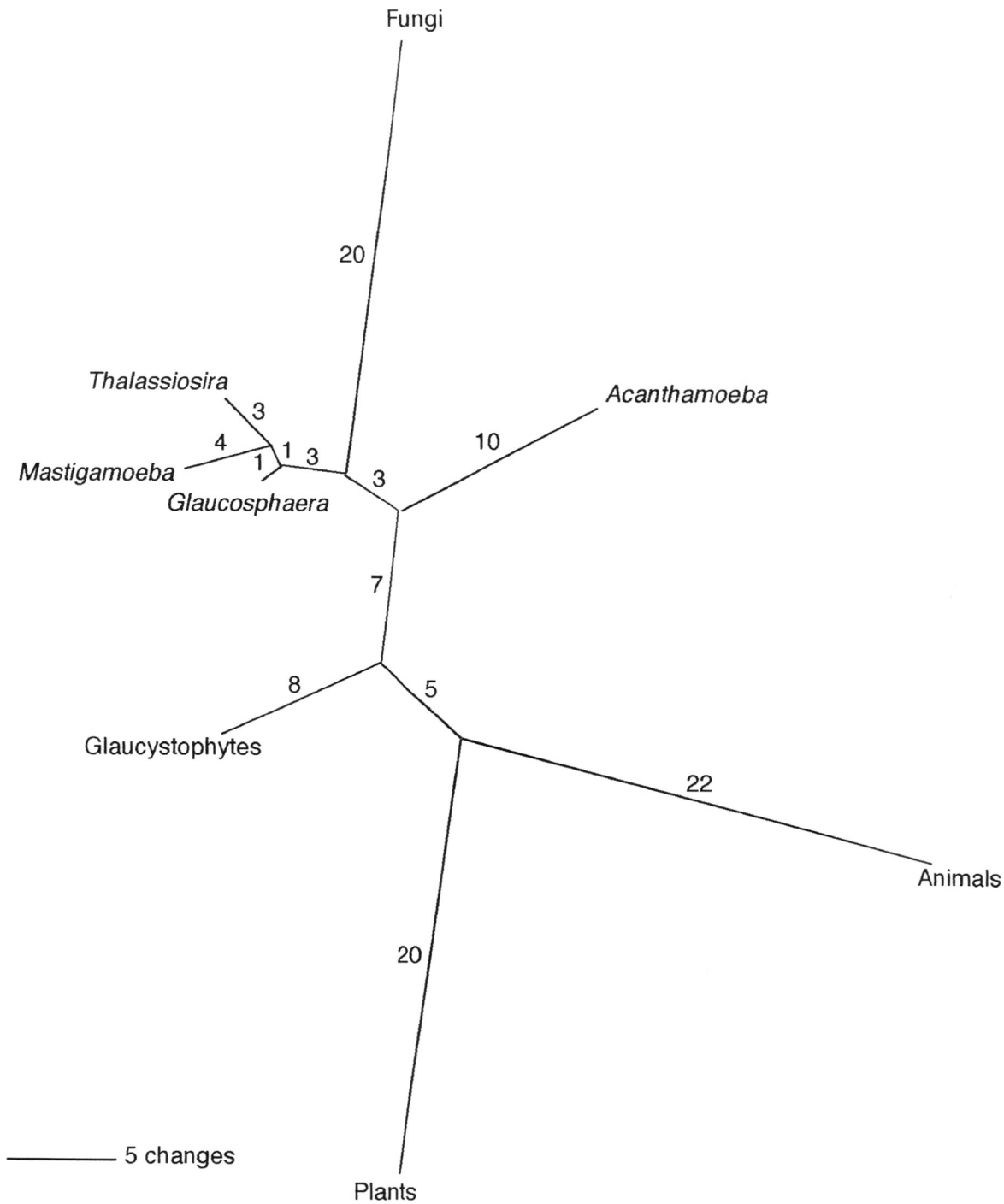


Figure 10. Tree of combined introns created by Dollo parsimony.

Table 1. Degenerate primer sequences used to amplify *RPB1*.

```
Aforward (1997)
GAK TGT CCK GGW CAT TTT GG
 E   C   P   G   H   F   G

Aforward (revised)
GAI TGY CCI GGI/C CAY TTY GG
 E   C   P   G     H   F   G

Dforward
TAY AAY GCI/C GAY TTY GAY GG
 Y   N   A     D   F   D   G

Dforward2
GAY TTY GAY GGI/C GAY GAR ATG AA
 D   F   D   G     D   E   M   N

Fforward
CAY GCI ATG GGI GGI MGI/C GAR GG
 H   A   M   G   G   R     E   G

Dreverse
TT CAT YTC RTC ICC RTC RAA RTC
N   M   E   D   G   D   F   D

Freverse
CC YTC ICK ICC ICC CAT C/IGC RTG
G   E   R   G   G   M     A   H

Greverse1
TG RAA IGT RTT IAG C/IGT CAT YTG
H   F   T   N   L     T   M   Q

Greverse2
GT CAT YTG IGT IGC IGG YTC C/ICC
T   M   Q   T   A   P   E     G
```

Table 1. Degenerate primer sequences for *RPB1*.

Table 2. Kishino-Hasegawa and Shimodaira-Hasegawa tests were performed to determine the significance of differences between the consensus Bayesian tree (of 28 sequences) and alternative tree topologies. The KH tests were performed using TREEPUZZLE, and SH tests using PHYLIP (version 3.61). The trees involved in the tests were constrained based on results from the original MrBayes analysis. P&G tree had green plants and glaucocystophytes constrained; O&A had all opisthokonts and *Acanthamoeba* constrained; R,P&G had red algae, green plants and glaucocystophytes constrained.

	Shimodiara-Hasegawa test				Kishino-Hasegawa test			
Tree	logL	Δ	P value	Sig?	logL	Δ	P value	Sig?
Best tree	-44498.6	-	-	-	-44272.7	-	-	-
P&G	-44508.9	-10.3	0.351	No	-44284.1	11.38	0.107	No
O&A	-44520.9	-22.3	0.086	No	-44295.1	22.43	0.054	No
R,P&G	-44555.5	-56.9	0.001	Yes	-44331.4	58.69	0.002	Yes

Table 2. SH and KH test results.

Table 3. Introns found in all organisms are represented in the table, which corresponds with the alignment in Figure 8.
The complete alignment with all introns can be found in the Appendix.

[1] Species abbreviations: Cp-*Cyanophora paradoxa*, Gn–*Glaucocystis nostochinearum*, Os–*Oryza sativa*, At-*Arabidopsis thaliana*, S. sp?–*Spirogrya*, Cr-*Chlamydomonas reinhardtii*, Dm-*Drosophila melangaster*, Mu.m-*Mus musculus*, Ce-*Caenorhabditis elegans*, Mb-*Monosiga brevicollis*, Sp-*Schizosaccharomyces pombe*, Cc-*Coprinopsis cinerea*, Ci-*Coccidioides immitis*, Cn-*Cryptococcus neoformans*, Mm-*Monoblepharis macrandra*, Ac-*Acanthamoeba castellanii*, Mi-*Mastigamoeba invertens*, Tp-*Thalassiosira pseudonana*, Gv-*Glaucosphaera vacuolata*

[2] These three organisms were not included in Figure 8. The introns of these organisms are unique, and therefore do not provide any additional information with respect to shared intron positions between glaucocystophytes and other eukaryotes.

Table 3. Intron positions in Eukaryotes. Numbers correspond to positions in Figure 8.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Cp[1]												●			●			
Gn			●									●						
Os			●									●						
At			●									●						
S.sp?			●									●						
Cr			●									●	●					
Dm			●															
Mu.m			●					●				●		●				
Ce			●				●									●		
Mb									●									●
Sp	●	●		●		●						●						
Cc										●		●					●	
Ci										●		●						
Cn			●		●												●	
Mm										●	●						●	
Ac												●						
Mi																		
Tp																		
Gv			●															

	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Cp		●		●	●					●		●		●		●		
Gn	●						●			●				●				
Os		●		●			●			●			●					
At		●		●			●			●			●					
S.sp?		●		●			●			●			●					
Cr	●		●			●				●			●					●
Dm																		
Mu.m				●			●			●			●					●
Ce											●							
Mb																		
Sp																		
Cc					●				●						●		●	
Ci																		
Cn									●						●		●	
Mm								●										
Ac										●								
Mi																		
Tp																		
Gv																		

	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
Cp		●																
Gn								●	●									
Os								●		●								●
At								●		●								●
S.sp?								●		●								●
Cr								●	●					●			●	
Dm																		
Mu.m								●	●		●					●		
Ce						●												
Mb				●														
Sp																		
Cc	●		●		●													
Ci																		
Cn	●		●		●													
Mm												●	●					
Ac		●					●								●			
Mi																		
Tp																		
Gv																		

	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Cp																		
Gn			●				●								●			
Os																		
At																		
S.sp?																		
Cr				●		●				●			●				●	
Dm		●																
Mu.m		●			●						●					●		●
Ce		●										●						
Mb																		
Sp																		
Cc																		
Ci																		
Cn									●									
Mm		●						●										
Ac	●													●				●
Mi																		
Tp																		
Gv																		

	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89
Cp																	
Gn																	
Os																	
At																	
S.sp?																	
Cr	●		●	●			●			●		●		●			●
Dm																	
Mu.m			●		●				●		●	●	●			●	
Ce						●											
Mb																	
Sp																	
Cc																	
Ci																	
Cn																	
Mm																	
Ac		●						●							●		
Mi																	
Tp																	
Gv																	

Unshared introns from miscellaneous taxa [2]								
Mi				●	●	●		●
Tp	●	●	●					
Gv							●	

REFERENCES

Andersson, J.O. and Roger, A.J. 2002. A cyanobacterial gene in non-photosynthetic protists – An early chloroplast acquisition in eukaryotes? *Curr. Biol.* 12:115-119.

Archibald, J.M. and Keeling, P.J. 2002. Recycled plastids: a 'green movement' in eukaryotic evolution. *Trends in Genetics* 18: 577-584.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M., and Philippe, H. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium, Entamoeba*, and *Mastigamoeba. Proc. Natl. Acad. Sci. USA* 3:1414-1419.

Bauldauf, S. and Roger, A.J. 2000. A Kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* 290: 972-977.

Bhattacharya, D., Helmchen, T., Bibeau, C., and Melkonian, M. 1995. Comparisons of nuclear-encoded small-subunit ribosomal RNAs reveal the evolutionary position of Glaucocystophyta. *Mol. Biol. Evol.* 12: 415-420.

Bhattacharya, D., and Schmidt, H.A. 1997. Division *Glaucocystophyta. Plant Sys. Evol.* 11: 139-148.

Bhattacharya, D. and Weber, K. 1997. The actin gene of the Glaucocystophyte *Cyanophora paradoxa*: analysis of the coding region and introns, and an actin phylogeny of eukaryotes. *Curr. Genet.* 31: 439-446.

Bhattacharya, D., Yoon, H.S., and Hackett, J.D. 2003. Photosynthetic eukaryotes unite: endosymbiosis connects the dots. *BioEssays* 26: 50-60.

Cavalier-Smith, T. 1985. Selfish DNA and the origin of introns. *Nature* 315: 283-284.

Cavalier-Smith, T. 2002. Chloroplast evolution: secondary symbiogenesis and multiple losses. *Curr. Biol.* 12: R62-R64.

Corden, J.L. 1990. Tails of polymerase II. *TIBS* 15: 383-387.

Delwiche, C.F. and Palmer, J.D. 1997. The origin of plastids and their spread via secondary symbiosis. *Plant Syst. Evol.* Suppl 11: 53-86.

Delwiche, C.F. 1999. Tracing the thread of plastid diversity through the tapestry of life. *Am. Nat.* 154: S164-177.

Douglas, S.E. 1998. Plastid evolution: origins, diversity, trends. *Curr. Opin. Genet. Dev.* 8: 655-661.

Doyle, J.J. and Doyle, J.L. 1990. Isolation of plant DNA from fresh tissue. *Focus* 12:13-15.

Federov, A., Merican, A.F., and Gilbert, W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci. USA* 99: 16128-16133.

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

Gray, M.W. 1993. Origin and evolution of organelle genomes. *Curr. Opin. Genet. Dev.* 3: 153-162.

Gray, M.W. and Spencer, D.F. 1996. Organellar evolution. In *Evolution of Microbial Life*, vol 54. Roberts, D.M., Sharp, P., Alderson, G., and Collins, M.A. Eds. Cambridge: Cambridge University Press. pp. 109-126.

Hall, W.T., and Claus, G. 1963. Ultrastructural studies on the blue-green algal symbiont in *Cyanophora paradoxa. Cell. Biol.* 19: 551-563.

Helmchen, T.A., Bhattacharya D., and Melkonian, M. 1995. Analyses of ribosomal-RNA sequences from Glaucocystophyte cyanelles provide new insights into the evolutionary relationships of plastids. *J. Mol. Evol.* 41: 203-210.

Hirt, R.P., Logsdon,J.M., Healy, B., Dorey,w., Doolittle, W.F., and Embley, T.M. 1999. Microsporidia are related to fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. USA 96: 580-585.

Howe, K.J. 2002. RNA polymerase II conducts a symphony of pre-mRNA processing activities. *Biochimica et Biophysica Acta.* 1577: 308-324.

Huelsenbeck, J.P. and Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17: 754-755.

Jokerst, R.S., Weeks, J.R., Zehring, W.A., and Greenleaf, A.L. 1989. analysis of the gene encoding the largest subunit of RNA polymerase II in Drosophila. *Mol. Gen. Genet.* 215: 266-275.

Jones, D. T., Taylor, W. R. and Thornton, J. M. 1992. The rapid generation of mutation data matricies from protein sequences. *Comput. Appl. Biosci.* 8: 275-282.

Kies, L., and Kremer, B.P. 1986. Typification of the Glaucocystophyta. *Taxon* 35: 128-133.

Kim, Y.J., Bjorklund, S., Li, Y., Sayre, M.H., and Kornberg, R.D. 1994. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* 77: 599-608.

Kishino, H. and Hasegawa, M. 1989. Evaluation of the maximum-likelihood estimate of the evolutionary tree topologies from DNA - sequenced data and the branching order in hominoidea. *J. Mol. Evol.* 29: 170-179.

Klenk, H.-P., Zillig, W., Lanzendorfer, M., Brampp,B., and Palm, P. 1995. Location of protist lineages in a phylogenetic tree inferred from sequences of DNA-dependent RNA polymerases. *Archiv. Protisten. Kunde* 145: 221-230.

Knoll, A.H. 1992. The early evolution of eukaryotes: a geological perspective. *Science* 256: 622-627.

Kumar, S. and Rzhetsky, A. 1996. Evolutionary relationships of eukaryotic kingdoms. *J. Mol. Evol.* 42: 183-193.

Loffelhardt, W., Bohnert, H.J., and Bryant, D.A. 1997. The complete sequence of the *Cyanophora paradoxa* cyanelle genome. *Plant Syst. Evol.* 11: 149-162.

Lynch, M. and Richardson, A.O. 2002. The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* 12: 701-710.

Maddison, W.P., and Maddison, D.R. 1992. MacClade, version 3.0. Sinauer, Sunderland, Mass.

Martin, W., Stoebe, B., Goremykin, V., Hansmann, S., Hasegawa, M., and Kowallik, K.V. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393: 162-165.

McFadden, G.I. 2001. Primary and secondary endosymbiosis and the origin of plastids. *J. Phycol.* 37: 951-959.

McFadden, G.I., and van Dooren, G.G. 2004. Evolution: red algal genome affirms a common origin of all plastids. *Curr. Bio.* 14: R514-R516.

Moreira, D., LeGuyader,H., and Phillippe, H. 2000. The origin of red algae and the evolution of chloroplasts. *Nature* 405: 69-72.

Moreira, D., and Philippe, H. 2001. Sure facts and open questions about the origin and

evolution of photosynthetic plastids. *Res. Microbiol.* 152: 771-780.

Nawrath, C., Schell, J., and Konez, C. 1990. Homologous domains of the largest subunit of eukaryotic RNA polymerase II are conserved in plants. *Mol. Gen. Genet.* 223: 65-75.

Nozaki, H., Ohta, N., Matsuzaki, M., Misumi, O., and Kuroiwa, T. 2003. Phylogeny of plastids based on cladistic analysis of gene loss inferred from complete genome sequences. *J. Mol. Evol.* 57: 377-382.

Nozaki, H., Matsuzaki, M., Takahara, M., Misumi, O., Kuroiwa, H., Hasegawa, M., Shin-I, T., Kohara, Y., Ogasawara, N., and Kuroiwa,T. 2003. The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. *J. Mol. Evol.* 56: 485-497.

Palmer, J.D. 2003. The symbiotic birth and spread of plastids: How many times and whodunit? *J. Phcol.* 39: 4-11.

Palumbi, S. and Baker, C. 1994. Contrasting population structure from nuclear intron sequences and mtDNA of humpback whales. *Mol. Biol. Evol.* 11: 426-435.

Philippe, H., Lopez, P., Brinkmann, H., Budin, K., Germot, A., Laurent, J., Moreira, D., Muller, M., and Le Guyader, H. 2000. Early branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. *Proc. R. Soc. Lond. B.* 267: 1213-1221.

Puhler, G., Leffer, H., Gropp, F., Palm, P., Klenk, H-P., Lottspeich, F., Garret, R.A., and Zillig, W. 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. *Proc. Natl. Acad. Sci. USA* 86: 4569-4573.

Rogozin, I.B., Wolf, Y.I., Sorokin, A.F., Mirkin, B.G., and Koonin, E.V. 2003. Remarkable Interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Bio.* 13: 1512-1517.

Shimodaira, H. and Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* 16: 1114–1116.

Sidow, A. and Thomas, W.K. 1994. A molecular evolutionary framework for eukaryotic model organisms. *Curr. Biol.* 4: 596-603.

Stiller, J.W., and Hall, B.D. 1997. The origin of red algae: implications for plasmid

evolution. *Proc. Natl. Acad. Sci. USA* 94: 4520–4525.

Stiller, J.W. and Hall, B.D. 1998. Sequences of the largest subunit of RNA polymerase II from two red algae and their implications for rhodophyte evolution. *J. Phycol.* 34: 857-864.

Stiller, J.W. and Hall, B.D. 2002. Evolution of the RNA polymerase II C-terminal domain. *Proc. Natl. Acad. Sci. USA* 99: 6091-6096.

Stiller, J.W., Reel, D.C. and Johnson, J.C. 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. *J. Phycol.* 39: 95-105.

Stiller, J.W. 2004. Eukaryotic phylogeny in the age of genomics: evolutionary implications of functional differences. In: *Organelles, Genomes and Eukaryotic Phylogeny: An Evolutionary Synthesis in the Age of Genomics*. Hirt, R.P and Horner, D.S. Eds. CRC Press London. pp. 171-188.

Strimmer, K. and von Haeseler, A. 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* 13: 964-969.

Swofford, D.L. 1998. PAUP. Phylogenetic analysis using parsimony (and other methods). Version 4. Sinauer Associates, Sunderland, MA.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. *Nucleic Acids Res.* 22: 4673-4680.

APPENDIX

ALIGNMENT OF ALL RPB1 INTRON POSITIONS

Complete Alignment of unedited RPB1 sequences with intron positions. Includes entire *Glaucocystis* sequence to region H and the *Cyanophora* sequence regions A-G. Introns lie between residues in bold, red indicates the intron is inserted after first nucleotide of the codon and green indicates it is after the second nucleotide.

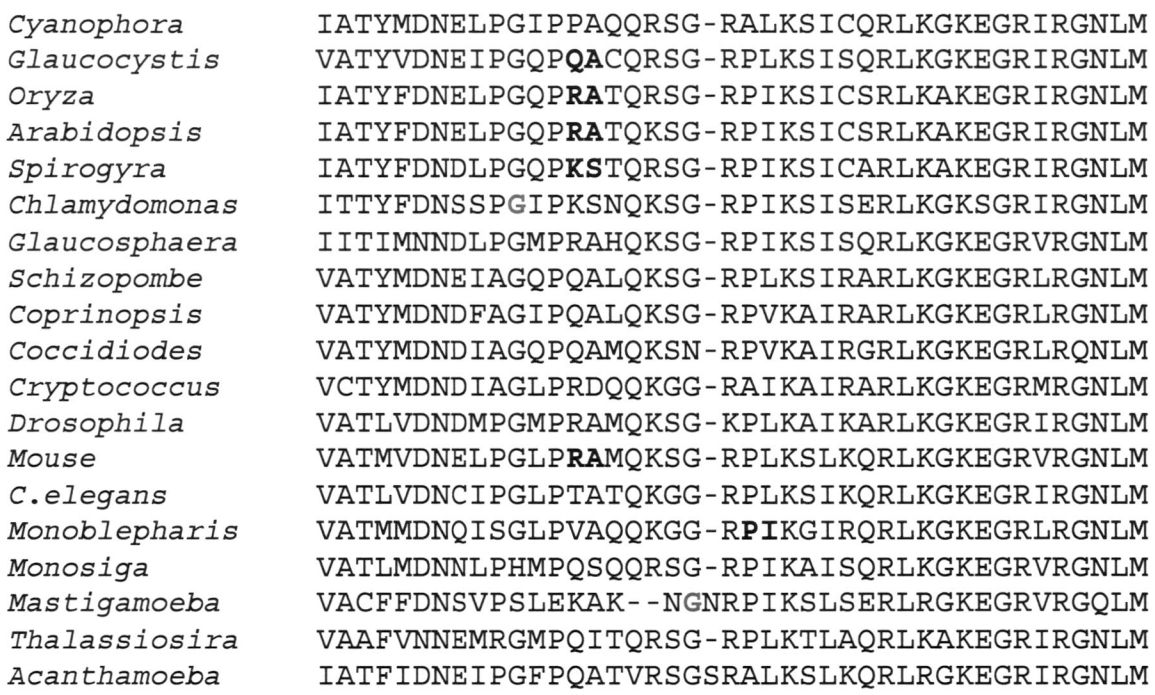
```
Cyanophora      ----------------------------------------
Glaucocystis    -MASLTGFPHSSAELRKVSVVQFGILGPDEIRRMSVAK--IEY
Oryza           --MDAARFPYSPAEVAKVEAVQFGVLSPDEIRQMSVVH--IEH
Arabidopsis     ---MDTRFPFSPAEVSKVRVVQFGILSPDEIRQMSVIH--VEH
Spirogyra       ---MDQRFPFSPAQLSKVKRVQFGILSPDEIRQMSVAT--IES
Chlamydomonas   -MASVDRFPFSSAPVRQVKAIQFSVWDPAEIKAYSVAE--IVT
Glaucosphaera   -MSFVKDFLYSSAEVKKVKAVQFGVLSPDEIRRMSVCE--IKF
Schizopombe     --MSGIQFSPSSVPLRRVEEVQFGILSPEEIRSMSVAK--IEF
Coprinopsis     ----------------------------------------
Coccidiodes     -MASVK-FPFSKAPLRTIKEIQFGLLSPEEIKRMSVVH--VEY
Cryptococcus    ---MTVTFPYSSAPVKQVKEIQFGIMSPEEIKAFSVAK--IES
Drosophila      ----MSTPTDSKAPLRQVKRVQFGILSPDEIRRMSVTEGGVQF
Mouse           MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKY
C.elegans       -MALVG-VDFQ-APLRIVSRVQFGILGPEEIKRMSVAH--VEF
Monoblepharis   ----------------------------------------
Monosiga        ----------------------------------------
Mastigamoeba    ----------------------------------------
Thalassiosira   -----TNFGHSSARLRRIKKLQFGIINPEELRQYSVTQ----A
Acanthamoeba    ----------------------------------------

Cyanophora      ----------------------------------------
Glaucocystis    PECFEN--GRPKLGGLMDPRMGSIDRQFKCQTCAGSMSECPGH
Oryza           AETMEK--GKPKPGGLSDPRMGTIDRKIKCETCMAGMAECPGH
Arabidopsis     SETTEK--GKPKVGGLSDTRLGTIDRKVKCETCMANMAECPGH
Spirogyra       SETYDR--GKPKLGGLSDPRLGTIDRAVKCETCGGNMADCPGH
Chlamydomonas   SETYEK--GKPKLGGLSDPRMGTMDRGGICTTDGANANDSPGY
Glaucosphaera   ETAFEM--GKPKTEGLMDPRLGAIGRDLPCKTCHCDEKNCPGH
Schizopombe     PETMDESGQRPRVGGLLDPRLGTIDRQFKCQTCGETMADCPGH
Coprinopsis     ----------------------------------------ECPGH
Coccidiodes     PETMDEARQRPREKGLNDPRLGTIDRNWRCATCEEGINDCPGH
Cryptococcus    TEVLDE-NGKQKVGGLMDPKMGAIDRNFKCQTCLEGMSECPGH
Drosophila      AETMEG--GRPKLGGLMDPRQGVIDRTSRCQTCAGNMTECPGH
Mouse           PETTEG--GRPKLGGLMDPRQGVIERTGRCQTCAGNMTECPGH
C.elegans       PEVYEN--GKPKLGGLMDPRQGVIDRRGRCMTCAGNLTDCPGH
Monoblepharis   ----------------------------------------ECPGH
Monosiga        ----------------------------------------CPGH
Mastigamoeba    ----------------------------------------CPGH
Thalassiosira   ITTY-MS-GQPVYGGVNDPRLGDLHDKS-----------DPGY
Acanthamoeba    ----------------------------------------ECPGH
```

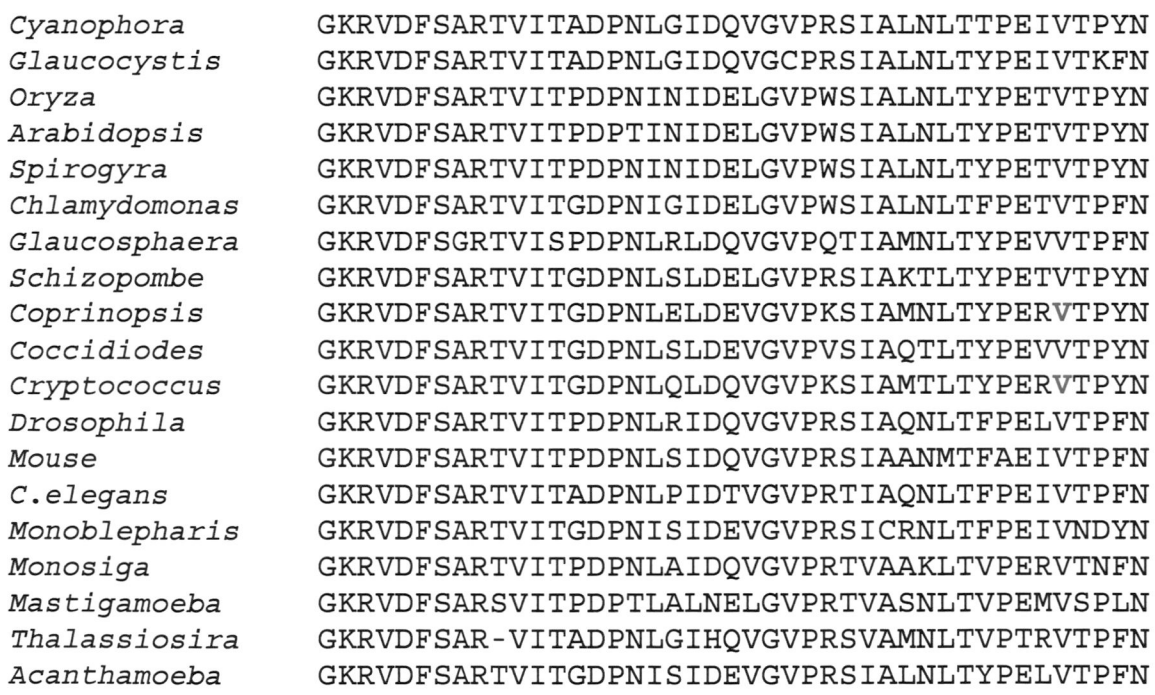
```
Cyanophora          ------------IGFLNTILKVLRCVCFNCSKILCDKNDVRFK
Glaucocystis        FGHLELAKPVLHIGFMTTILKILRCVCFHCSKLLCDKSDIRFR
Oryza               FGHLELAKPMFHIGFIKTVLSIMRCVCFNCSKILADEDDIKFK
Arabidopsis         FGYLELAKPMYHVGFMKTVLSIMRCVCFNCSKILADEEEHKFK
Spirogyra           FGHLELAKPMFHIGFLKTVLSVLRCVCHNCSRILADKNSHQFK
Chlamydomonas       FGHIALARPVYHIGFIKTVVRVLRCVSYHTSRLLVDKDDPKYA
Glaucosphaera       FGHIELVKPMFNIGFFGMTLRVLRCVCFFCSKLLIDPRDPKAD
Schizopombe         FGHIELAKPVFHIGFLSKIKKILECVCWNCGKLKIDSSNPKFN
Coprinopsis         FGHIELARPVFHPGFFNKVKKILECICVNCGKLKADITDPAFA
Coccidiodes         FGHIELSTPVFHIGFLTKIKKLLETVCHNCGKIKANTSDQKYL
Cryptococcus        FGHIELARPVFHQGFIVKVKKILECVCYSCGKLKVDMVYPMVA
Drosophila          FGHIDLAKPVFHIGFITKTIKILRCVCFYCSKMLVSPHNPKIK
Mouse               FGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIK
C.elegans           FGHLELAKPVFHIGFLTKTLKILRCVCFYCGRLLIDKSAPRVL
Monoblepharis       FGHIELAKPVFHEGFVTTIKKLLESVCLFCGKLKSDDSDPDFF
Monosiga            FGHIVLAKPVFHAGFLPKIISVLRCVCLYCSKLRVEHSDQQLQ
Mastigamoeba        FGHVTLAKAVYHLAYIRTVCRVLQCVCMNCSRLLVDRKIS---
Thalassiosira       FGHIELARPVYHQGFIVT-LKALRCVCFHCSRITMEDTEYKFQ
Acanthamoeba        FGHLELAKPVFNIGFLTTVLKILRSVCFHCSKLLVDEAQELKN

Cyanophora          AAQ----RIKR--PEAKLRAVMDICKSKSM-CEGGEELDMNVS
Glaucocystis        NAL----RLKK--PETRLRAVMDICRGKSQ-CEGGDELDENA-
Oryza               QAL----KIRN--PKNKLKRIYDACKNRKI-CAGGDNLDVQE-
Arabidopsis         QAM----KIKN--PKNRLKKILDACKNKTK-CDGGDDIDDVQ-
Spirogyra           LAS----RIRI--PKQRLRRMLDCCKSKTV-CEGSSNKEE--N
Chlamydomonas       YGL----RISD--SEKRLRYFVHLCQVRGPGSNGCMHTCR---
Glaucosphaera       FEVKKVLRIKN--RINRLKKVQNLC-SKVKICKHCSNE-----
Schizopombe         DTQ----RYRD--PKNRLNAVWNVCKTKMV-CDTGLSAGSDN-
Coprinopsis         EKIR---HIRN--PKQRMAVVWDHCKTKTI-CEADEPREEGAE
Coccidiodes         DAL----RFRD--PKRRFDAIWRLSKDVLI-CEADPPPDD---
Cryptococcus        NAVR---RIKA-Q--HRLKAIWALAKDKKH-CEPDELDEKDNG
Drosophila          EIV----MKSRGQPRKRLAYVYDLCKGKTI-CEGGEDMDLTKE
Mouse               DIL----AKSKGQPKKRLTHVYDLCKGKNI-CEGGEEMDNKFG
C.elegans           EILK---KTGTN-SKKRLTMIYDLCKAKSV-CEGAAEKEEG-M
Monoblepharis       RKS----RLRD--RRARFTGVWSLCKTKMV-CQGSEEDADEEQ
Monosiga            AILS---QTKDN-PKHRLQAVSRLCSNKRQ-CHRDEPPDTFNP
Mastigamoeba        AST----RLRN--GAVRLNEVSKMCLQYKT-CGECAVKTE---
Thalassiosira       RAR----QIRN--RKRRLDAMHALIRPKKK-CD----------
Acanthamoeba        VKH----RFAR---------VYTLAKTKTI-CEGGEDTGTDKE
```

```
Cyanophora      LDPAAAD------KAGVPKRVGCGNPQPKIMKDG-----LKLT
Glaucocystis    -EKEIGPDG----EK-KSRHGGCGNFVPKIIREG-----LKLT
Oryza           -QQGT---D----DPVKK-RGGCGAQQPNITVDG-----MKMV
Arabidopsis     ----SHSTD----EPVKKSRGGCGAQQPKLTIEG-----MKMI
Spirogyra       LDEEEGKEN---------PNEGCGAEQPRFSIDG-----IRIV
Chlamydomonas   -AAATMQGK--------KMDEATGNPQPSYRLDG-----LKIM
Glaucosphaera   --------------------------QPKFAKEG-----LQFR
Schizopombe     FDLSNPSAN--------MGHGGCGAAQPTIRKDG-----LRLW
Coprinopsis     GDGDEPKK----------GHGGCGHQQPQIRKEG-----LKMF
Coccidiodes     -DDPFGKESS---KIVK-GHGGCGNAQPQIRKEG-----ISLV
Cryptococcus    DATFEDEYLQ-EQKAAMKGHGGCGHEQPVWRKKG-----LKLM
Drosophila      NQQPDPNKK--------PGHGGCGHYQPSIRRTG-----LDLT
Mouse           VEQPEGDED---LTKE-KGHGGCGRYQPRIRRSG-----LELY
C.elegans       PDDPDDPMNDG-----KKVAGGCGRYQPSYRRVG-----IDIN
Monoblepharis   QQQQQQGAN--LLGRRKGAHGGCGRRQPKIRKEG-----LTLL
Monosiga        EGQPEVALS----------AGGCGKMQPRVRRSQEEGRTLQLL
Mastigamoeba    -NAQVTR-------------TGCGARQPLWRVSK------LSI
Thalassiosira   ---------------------HCNGYQPKYTKVG-----LHVE
Acanthamoeba    QGDSTDPKE---KKEVKKSHGGCGNFQPKITKDG-----MKIM

Cyanophora      AEFKKTLD----------ENQE-KKIILMPEQVYN-IFKRISD
Glaucocystis    AEFTKVQD----------ETIE-KRQVLSAEKVHQ-IFKRISD
Oryza           AEYKAPKKK-NDDQEQLPEPVE-RKQILSAERVLN-VLKRISD
Arabidopsis     AEYKNSKEE-NDEPDQLPEPAE-RKQTLGADRVLS-VLKRISD
Spirogyra       AEFKLKRK-KNEDMEQTLEPTE-RKQHLGADKILN-ILKKISD
Chlamydomonas   AEFP-KLK---GDEEQEQDNVE-RKQELTAARALE-IMKRIPA
Glaucosphaera   AEFKG----------SADDTME-KKQIVSAEKVLS-IFKHISD
Schizopombe     GSWK------RGKD----ESDLPEKRLLSPLEVHT-IFTHISS
Coprinopsis     VQYKK-----SKDD-DDMKSLQPDKRLITPSEILT-VFKKISD
Coccidiodes     GTWKPNKMRDMMDDTDIQQP-E-KKQ-ITPQMALN-IFRNVSE
Cryptococcus    GVWKP-------TDKGEA--AEPEERNISPGEVIYNILKKITP
Drosophila      AEWKH----------QNEDSQE-KKIVVSAERVWE-ILKHITD
Mouse           AEWKHVNE----------DSQE-KKILLSPERVHE-IFKRISD
C.elegans       AEWKKN---------VNEDTQE-RKIMLTAERDLE-VFQQITD
Monoblepharis   AHY---------AKETEDADLAP--RPLLPDQVQK-ILLQISD
Monosiga        GEWEE----------VNDDEVDRKRPITAAEVLK-CFRGISD
Mastigamoeba    TRVPRD----SSGG--E-LPEVQ----DSEEVVR--ILKNISD
Thalassiosira   IEY-------ADEME-RVGSSGDKKQFMSAQKAVD-IFKKMRD
Acanthamoeba    AEFKNVG---------DDPNVE-KKQLLTAEKVHA-ILKLISD
```

```
Cyanophora        EDCRHMGLDPRWARPDWFCITHLPVPPAAVRPGIAMNSV---Q
Glaucocystis      EDCRSMGLRPEWARPDWMIVTILPVPPPPVRPSIMMDST---A
Oryza             EDCLLLGLNPKFARPDWMILQVLPIPPPPVRPSVMMDTS---S
Arabidopsis       ADCQLLGFNPKFARPDWMILEVLPIPPPPVRPSVMMDAT---S
Spirogyra         EDCETLGLNPKFARPDWMILQVLPIPPPPVRPSVMMDSS---S
Chlamydomonas     EDCRALGFDCRFTRPDWMIIQNMPVPPPPVRPSVMMDSS---S
Glaucosphaera     DDIEIMGLSPQHSRPEWFILTLFPVPPPHIRPSVMMDAS---M
Schizopombe       EDLAHLGLNEQYARPDWMIITVLPVPPPSVRPSISVDGT---S
Coprinopsis       HDLHLLGLSDEYARPEWMILTVLPVPPPPVRPSIAVDGG--TM
Coccidiodes       EDVRILGLSNDYARPEWMVITVLPVPPPPVRPSVLVGGSSGGQ
Cryptococcus      EDLHIMGLNADYARPDWMILTVLPVPPAAVRPSIAVDGG--AM
Drosophila        EECFILGMDPKYARPDWMIVTVLPVPPLAVRPAVVMFG---AA
Mouse             EECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGS---A
C.elegans         EDILVIGMDPQFARPEWMICTVLPVPPLAVRPAVVTFGS---A
Monoblepharis     ADCVSLGLNPEWSRPEWMIITVLPVPPPQVRPSVVGDGT--GL
Monosiga          EDCRIIGFDPVYTRPDFMIIEVLPVPPLAVRPSVELGES---G
Mastigamoeba      EDCVLIGLNPRFARPEWMMITLLPVPPMTVRPSIAFGSV---G
Thalassiosira     DEVKALGLDVTWARPEWMCVSVMPVPPLHVRPSVVMGGG--AQ
Acanthamoeba      EDCLAMGFDPKWARPDWMIITIMPIPPPPVRPSITMDSA---A
```

```
Cyanophora        RAEDDLTSKLMDIVRANAQLRKNEQNAAPAHHINELVTQLQYH
Glaucocystis      RGEDDLTHKLADIVKANMNLKRQEMNGAPAHIISEFQQLLQFH
Oryza             RSEDDLTHQLAMIIRHNENLRRQERNGAPAHIITEFAQLLQFH
Arabidopsis       RSEDDLTHQLAMIIRHNENLKRQEKNGAPRHIISRFTQLLQFH
Spirogyra         RSEDDLTYQLSMIIRHNNNLKKQEQNGTLSHVINEFVQLLQFH
Chlamydomonas     RCEDDLTHKLAEILRTNNAIKKQDATGTPQHVIAEQIMALQYH
Glaucosphaera     RGEDDLTYKLGDIVRNNNALREMERTGAPAHRLNEQIHVLQAH
Schizopombe       RGEDDLTHKLSDIIKANANVRRCEQEGAPAHIVSEYEQLLQFH
Coprinopsis       RSEDDLTYKLGDIIKASSNVKRCEEEGVPAHIISEFEQLLQFH
Coccidiodes       RGEDDLTYKLAEIIRANQNVTRCEQEGSPEHVVREFESLLQYH
Cryptococcus      RSEDDLTYKLSQIIKFNGVVRRMEAEGVPPSVVNEQFDLLQYH
Drosophila        KNQDDLTHKLSDIIKANNELRKNEASGAAAHVIQENIKMLQFH
Mouse             RNQDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFH
C.elegans         KNQDDLTHKLSDIIKTNQQLQRNEANGAAAHVLTDDVRLLQFH
Monoblepharis     RSQDDLTYKLADILKANASLKRHEQDGGPTHVVNEFWDLLQYH
Monosiga          HSADDLTYQLGDIIKTNISLQENIANGATPHMIEECVEYIQFK
Mastigamoeba      RAEDDLTKQLSTIVKNNAQIRKLKLDGAAPINIQDTLDVLQIN
Thalassiosira     SSEDDLTHQLVNIVSN-IALKTAIQNGEPNIIVEQFEQALQHN
Acanthamoeba      RGEDDLTHKLADIIKANANLRKQEANGAAAHIISQFQELLQYH
```


Cyanophora IATYMDNELPGIPPAQQRSG-RALKSICQRLKGKEGRIRGNLM
Glaucocystis VATYVDNEIPGQPQACQRSG-RPLKSISQRLKGKEGRIRGNLM
Oryza IATYFDNELPGQPRATQRSG-RPIKSICSRLKAKEGRIRGNLM
Arabidopsis IATYFDNELPGQPRATQKSG-RPIKSICSRLKAKEGRIRGNLM
Spirogyra IATYFDNDLPGQPKSTQRSG-RPIKSICARLKAKEGRIRGNLM
Chlamydomonas ITTYFDNSSPGIPKSNQKSG-RPIKSISERLKGKSGRIRGNLM
Glaucosphaera IITIMNNDLPGMPRAHQKSG-RPIKSISQRLKGKEGRVRGNLM
Schizopombe VATYMDNEIAGQPQALQKSG-RPLKSIRARLKGKEGRLRGNLM
Coprinopsis VATYMDNDFAGIPQALQKSG-RPVKAIRARLKGKEGRLRGNLM
Coccidiodes VATYMDNDIAGQPQAMQKSN-RPVKAIRGRLKGKEGRLRQNLM
Cryptococcus VCTYMDNDIAGLPRDQQKGG-RAIKAIRARLKGKEGRMRGNLM
Drosophila VATLVDNDMPGMPRAMQKSG-KPLKAIKARLKGKEGRIRGNLM
Mouse VATMVDNELPGLPRAMQKSG-RPLKSLKQRLKGKEGRVRGNLM
C.elegans VATLVDNCIPGLPTATQKGG-RPLKSIKQRLKGKEGRIRGNLM
Monoblepharis VATMMDNQISGLPVAQQKGG-RPIKGIRQRLKGKEGRLRGNLM
Monosiga VATLMDNNLPHMPQSQQRSG-RPIKAISQRLKGKEGRVRGNLM
Mastigamoeba VACFFDNSVPSLEKAK--NGNRPIKSLSERLRGKEGRVRGQLM
Thalassiosira VAAFVNNEMRGMPQITQRSG-RPLKTLAQRLKAKEGRIRGNLM
Acanthamoeba IATFIDNEIPGFPQATVRSGSRALKSLKQRLRGKEGRIRGNLM


Cyanophora GKRVDFSARTVITADPNLGIDQVGVPRSIALNLTTPEIVTPYN
Glaucocystis GKRVDFSARTVITADPNLGIDQVGCPRSIALNLTYPEIVTKFN
Oryza GKRVDFSARTVITPDPNININDELGVPWSIALNLTYPETVTPYN
Arabidopsis GKRVDFSARTVITPDPTININDELGVPWSIALNLTYPETVTPYN
Spirogyra GKRVDFSARTVITPDPNININDELGVPWSIALNLTYPETVTPYN
Chlamydomonas GKRVDFSARTVITGDPNIGIDELGVPWSIALNLTFPETVTPFN
Glaucosphaera GKRVDFSGRTVISPDPNLRLDQVGVPQTIAMNLTYPEVVTPFN
Schizopombe GKRVDFSARTVITGDPNLSLDELGVPRSIAKTLTYPETVTPYN
Coprinopsis GKRVDFSARTVITGDPNLELDEVGVPKSIAMNLTYPERVTPYN
Coccidiodes GKRVDFSARTVITGDPNLSLDEVGVPVSIAQTLTYPEVVTPYN
Cryptococcus GKRVDFSARTVITGDPNLQLDQVGVPKSIAMTLTYPERVTPYN
Drosophila GKRVDFSARTVITPDPNLRIDQVGVPRSIAQNLTFPELVTPFN
Mouse GKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFN
C.elegans GKRVDFSARTVITADPNLPIDTVGVPRTIAQNLTFPEIVTPFN
Monoblepharis GKRVDFSARTVITGDPNISIDEVGVPRSICRNLTFPEIVNDYN
Monosiga GKRVDFSARTVITPDPNLAIDQVGVPRTVAAKLTVPERVTNFN
Mastigamoeba GKRVDFSARSVITPDPTLALNELGVPRTVASNLTVPEMVSPLN
Thalassiosira GKRVDFSAR-VITADPNLGIHQVGVPRSVAMNLTVPTRVTPFN
Acanthamoeba GKRVDFSARTVITGDPNISIDEVGVPRSIALNLTYPELVTPFN

```
Cyanophora      MDRMYQLIRAGPTEYPG---ARYIIRSDGTRFDLRYVPKAS--
Glaucocystis    MPRMYELIRNGPNEHPG---ARYIIRDDGQRLDLRYIKKAS--
Oryza           IERLKELVEYGPHPPPGKTGAKYIIREDGQRLDLRYVKKSS--
Arabidopsis     IERLKELVDYGPHPPPGKTGAKYIIRDDGQRLDLRYLKKSS--
Spirogyra       LEKLKELVENGPHPPPGKTGAKYIIRDDGQRLDLRFLKKNS--
Chlamydomonas   IEKLQKLVDNGPNPPPGETGAKHIIREDGRRVSLAVVKGDA--
Glaucosphaera   VEKMRRLVMNGPNEYPG---AKYIERLDGSKVNLAFVKNRS--
Schizopombe     IYQLQELVRNGPDEHPG---AKYIIRDTGERIDLRYHKRAG--
Coprinopsis     IAYLQELVRNGPTTYPG---ARYVVRDTGERIDLRYNKRA---
Coccidiodes     INKLGQLVDNGPDVHPG---ARYVIRSSGERIDLRHHKGGGGR
Cryptococcus    IVYLQTLVNNGPATYPG---ARYYVKDTGERVDLKYRKSGE--
Drosophila      IDRMQELVRRGNSQYPG---AKYIVRDNGERIDLRFHPKSS--
Mouse           IDRLQELVRRGNSQYPG---AKYIIRDNGDRIDLRFHPKPS--
C.elegans       VDKLQELVNRGDTQYPG---AKYIIRENGARVDLRYHPRAA--
Monoblepharis   RDHLQKLVANGTNEYPG---ANRVIRANGDIVSLKHTAGSG--
Monosiga        LNKMNQYVANGPTQHPG---ARYIIQEDGTRIDLHIARSHS--
Mastigamoeba    IHRLAALVRNGPGTYPG---AKYVIRDDGARIAL-TSRGDM--
Thalassiosira   IQELSALVANGPTEHPG---AKHIIRSDGLRIDLRYVKNKS--
Acanthamoeba    IDKMYELIRNGPTEHPG---AKYIIRDDGQRLDLRFARKAS--
```

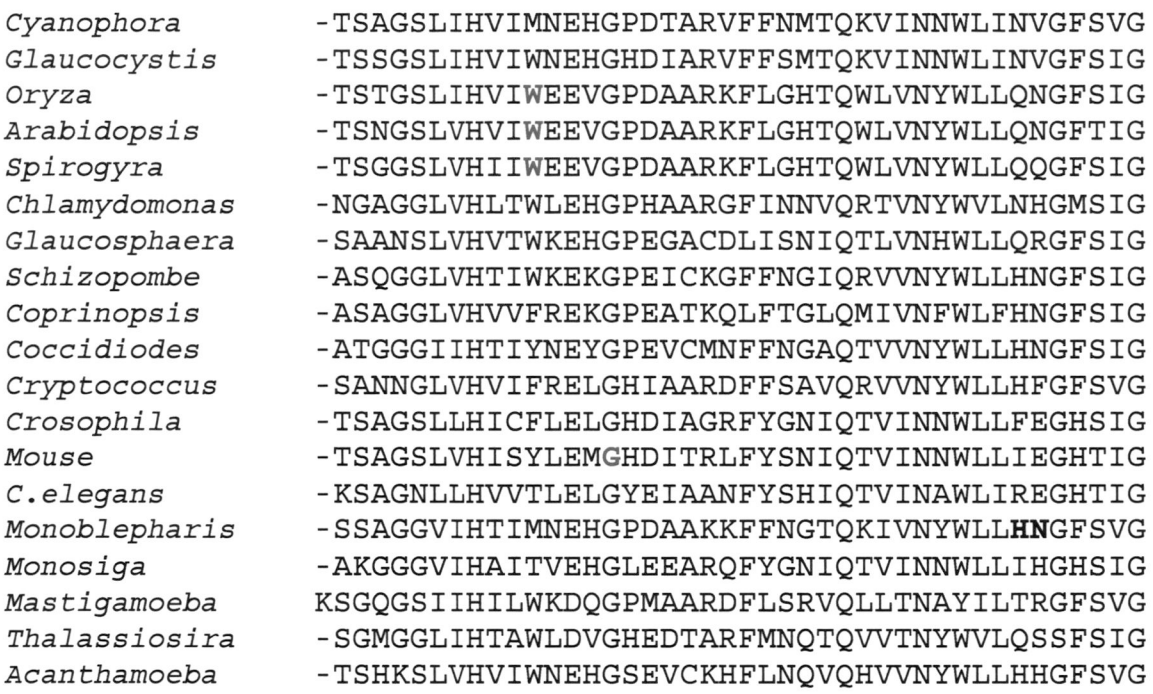
```
Cyanophora      DLHLEPGY-RVERHIQDGDYVLFNRQPSLHKMSIMAHRIKVMP
Glaucocystis    DLHLEPGY-RVERHIQDGDYVLFNRQPSLHKMSIMGHRVKVMP
Oryza           DQHLELGY-KVERHLNDGDFVLFNRQPSLHKMSIMGHRIKIMP
Arabidopsis     DQHLELGYRYVERHLQDGDFVLFNRQPSLHKMSIMGHRIRIMP
Spirogyra       DRHLELGY-KVERHLVDGDFVLFNRQPSLHKMSIMGHRIRIMP
Chlamydomonas   DRRLEPGD-KVERHLINGDLVLFNRQPSLHKMSMMGHRVRILP
Glaucosphaera   DIHLNNGE-KVIRHLLDGDYVIFNRQPSLHKMSIMGHRVKVMR
Schizopombe     DIPLRYGW-RVERHIRDGDVVIFNRQPSLHKMSMMGHRIRVMP
Coprinopsis     DAFLQLGW-IVERHLKDGDYVLFNRQPSLHKMSMMSHRVRLMP
Coccidiodes     N-FLQWGW-RVERHLMDGDVILFNRQPSLHKESMMAHRVRVMP
Cryptococcus    PISLQFGW-IVERHLKDGDYVLFNRQPSLHKMSMMSHRVKLMN
Drosophila      DLHLQCGY-KVERHLRDDDLVIFNRQPTLHKMSMMGHRVKVLP
Mouse           DLHLQTGY-KVERHMCDGDIVIFNRQPTLHKMSMMGHRVRILP
C.elegans       DLHLQPGY-RVERHMKDGDIIVFNRQPTLHKMSMMGHRVKILP
Monoblepharis   ETLLNIGD-RVERHLVDGDYIIFNRQPSLHKMSMMGHKVRVMP
Monosiga        DRTLQPGF-IVERHLQDNDVIIFNRQPTLHKMSMMGHRVKVLP
Mastigamoeba    --PIVPGY-IIERHLMDGDHVVFNRQPSLHKMSMMGHQVRVLP
Thalassiosira   DLLLANGW--VERHLRDGDIVLFNRQPSLHKMSIMGHMAKVLD
Acanthamoeba    DLHLEYGY-KVERHIQDGDVIIFNRQPSLHKMSMMGHKVKIMP
```

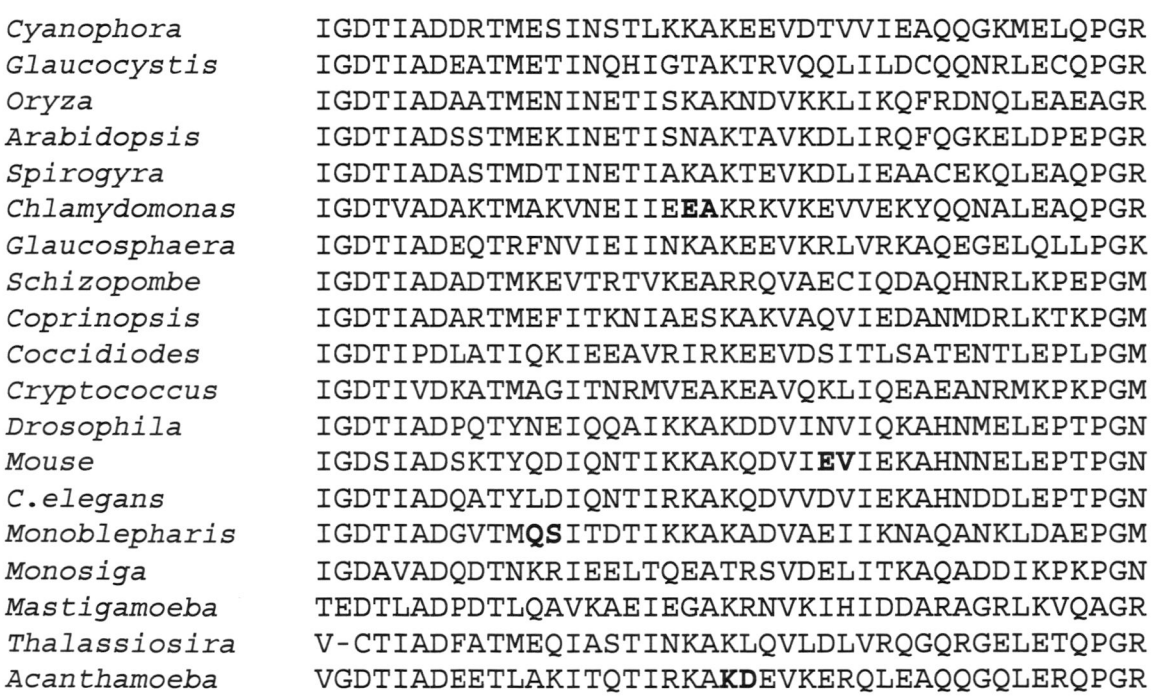
Cyanophora YSTFRLNLSVTSPYNADFDGDEMNLHVPQSFETRAEASEIIAV
Glaucocystis YSTFRLNLSVTTPYNADFDGDEMNMHVAQTFETRAEVQEICLV
Oryza YSTFRLNLSVTSPYNADFDGDEMNMHVPQSFETRAEVLELMMV
Arabidopsis YSTFRLNLSVTSPYNADFDGDEMNMHVPQSFETRAEVLELMMV
Spirogyra YSTFRLNLSVTSPYNADFDGDEMNMHVCQTFETRAETMELMMV
Chlamydomonas YSTFRLNLSVTTPYNADFDGDEMNMHVAQTHETRAEMANLMMV
Glaucosphaera YSSFRLNLSCTSPYNADFDGDEMNLHVPQSPQARAEVMQLMMV
Schizopombe YSTFRLNLSVTSPYNADFDGDEMNMHVPQSEETRAEIQEITMV
Coprinopsis YSTFRLNLSVTPPYNADFDGDEMNMHVPQSEETRAELS**QI**AWV
Coccidiodes YSTFRLNLSVTTPYNADFDGDEMNLHVPQSEEARAELNQLCLV
Cryptococcus YSTFRLNLSVTSPYNADFDGDEMNLHVPQSEETRAELS**QI**AWV
Drosophila WSTFRMNLSCTSPYNADFDGDEMNLHVPQSMETRAEVENIHIT
Mouse WSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMV
C.elegans WSTFRMNLSVTSPYNADFDGDEMNLHLPQSLETRAEIEEIAMV
Monoblepharis FSTFRLNLSVTSPYNADFDGDEMNLHAPQSYETRAEIMEIMRV
Monosiga WSTFRMNLSVTSPYNADFDGDEMNLHVPQSLGSRA**EI**EEIMMV
Mastigamoeba YSSFRLNLSVTTPYNADFDGDEMNMHVPNSLEAISEVKNLMAV
Thalassiosira WSTFRLNLSCTSPYNADFDGDEMNLHVPQSLPARAEA-ELMMH
Acanthamoeba YSTFRLNLSVTTPYNADFDGDEMNMHVPQTPGARAEVIELMMV

Cyanophora PRQIVSPQSNRPVMGIVQDTLMASQKMTLRDTFIEKDVIMNIV
Glaucocystis PRQIISPQSNRPVMGIVQDTLMASQKMTIRDTFIE**KD**VVMNIL
Oryza PKCIVSPQSNRPVMGIVQDTLLGCRKITKRDTLIE**KD**VFMNIL
Arabidopsis PKCIVSPQANRPVMGIVQDTLLGCRKITKRDTFIE**KD**VFMNTL
Spirogyra PKCVVSPQSNRPVMGIVQDTLLGCRKVTKRDTFIE**KD**VFMNIL
Chlamydomonas PRNIVSPQANKPVMGIVQDALLGTRMMTKRDIFIE**KD**NFMNCV
Glaucosphaera PRCIVSPQGNKPVMGIVQDTLVGTMLFTQRDTFMEKDLVMNLL
Schizopombe PKQIVSPQSNKPVMGIVQDTLAGVRKFSLRDNFLTRNAVMNIM
Coprinopsis PRQIVSPQANKPVMGIVQDTLCGIRKFTLRDTFLDWNHVQNIL
Coccidiodes PLNIVSPQRNGPLMGIVQDTLCGIYKICRRDVFLTKEQVMNMM
Cryptococcus PRQIVSPQANKPVMGIVQDTLCGIRKFTLRDNFLDWLQVQHIL
Drosophila PRQIITPQANKPVMGIVQDTLTAVRKMTKRDVFITREQVMNLL
Mouse PRMIVTPQSNRPVMGIVQDTLTAVRKFTKRDVFLE**RG**EVMNLL
C.elegans PRQLITPQANKPVMGIVQDTLCAVRMMTKRDVFIDWPFMMDLL
Monoblepharis PRQIISPKHNQPVMGIVQDTLCGIRKFTIRDNFLSRDMVMNIL
Monosiga HKNILTPQSNRPVMGIIQDTLTVARRMTFRDCFIEREQLMHLL
Mastigamoeba PFQIVTPQKNSPCMGVVQDSLLGCSLISRRDTFLTEDVMMNLA
Thalassiosira SPRVVSGQSNRPVMGIVQDSLLAVQKMTKRDVFVKKDLMMNIL
Acanthamoeba PKQIVTAQSNKPVIGIVQDTLLGGCLLTQRDTFIEKDVMMNIL

```
Cyanophora        MHLFDGFHTLPIPAIVKPKPRWTGKQLFSTFLP-NVN---VVR
Glaucocystis      MHLDSFDGRLPIPAILKPRPLWTGKQLFTMFLP-NVN---LIR
Oryza             MWWEDFDGKVPAPAILKPRPIWTGKQVFNLIIPKQIN---LIR
Arabidopsis       MWWEDFDGKVPAPAILKPRPLWTGKQVFNLIIPKQIN---LLR
Spirogyra         MWWEDFEGKIPSPTILKPRPLWTGKQVFSLIIPRAVN---LER
Chlamydomonas     MGIEDWDGTVPMPAVLKPRPLWTGKQIFSMFVP-DVN---LKN
Glaucosphaera     LHFRGWDGVIPPPAILKPRQLWTGKQLFSMILP-DVN---LVR
Schizopombe       LWVPDWDGILPPPVILKPKVLWTGKQILSLIIPKGIN---LIR
Coprinopsis       LWVPEWDGVVPTPAILKPKPLWTGKQILSLTIPRGIN---IQR
Coccidiodes       LWVPEWDGVLPQPAILKPRPRWTGKQMISMVLPSGL----NLL
Cryptococcus      LWLPEWDGTIPPPAIFKPKPMWTGKQLLSMTIPKGIN---ITY
Crosophila        MFLPTWDAKMPQPCILKPRPLWTGKQIFSLIIPGNVN---MIR
Mouse             MFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHIN---CIR
C.elegans         MYLPTWDGKVPQPAILKPKPLWTGKQVFSLIIPGNVN---VLR
Monoblepharis     MWVEDWDGVIPLPAIVKPVPLWTGKQMMSLAIP-DIN---LVG
Monosiga          MWMPNWDGKIPVPAIVAPKELWTGKQLVSMIIPRRIN---LEG
Mastigamoeba      LVISYDPDKIPQPAILRPKPLWTGKQAFSMLLPKDLSYYFNAE
Thalassiosira      MWVEDWDGRIPPPAIYRPEELWTGKQIMSMILPK-IN---LTG
Acanthamoeba      MWLESWDGTVPTPTILKPKQLWTGKQVFSLIIPRNTN---FVN
```

```
Cyanophora        FHSTHPDGEST-----DISPGDTQVRIENGELLCGIVCKRTLG
Glaucocystis      FCSQHPDNEVS-----DISPGDT**KV**IIEQGELLAGIVCKRTLG
Oryza             FSGWHSEAETR-----FITPGDTMVRIEKGELLSGTLCKKTLG
Arabidopsis       YSAWHADTETG-----FITPGDTQVRIERGELLAGTLCKKTLG
Spirogyra         YSAWHPDSERG-----DFSPGDTQVRVEKGELLAGILCKKSLG
Chlamydomonas     KSAWYKDADV-----PDMSVDDA**QV**LIRQGELLTGALCKKTVG
Glaucosphaera     FSNTHDDDD-------IENAGDTKVLISRGELLSGTVDKRTVG
Schizopombe       DDDKQSLS----------NPTDSGMLIENGEIIYGVVDKKTVG
Coprinopsis       ASENKSSN----------PVFDDGMLIENGEFIYGIVDKKIVG
Coccidiodes       R-VDRDKVPLAE---KFSPLTDTGLLVHGGELMYGMFSKKTVG
Cryptococcus      KNNEKPSP---------IDVTDENVLIENGELVHGTIVKNMAG
Drosophila        THSTHPDEEDEG-PYKWISPGDTKVMVEHGELIMGILCKKSLG
Mouse             THSTHPDDEDSG-PYKHISPGDT**KV**VVENGELIMGILCKKSLG
C.elegans         THSTHPDSEDSG-PYKWISPGDTKVIIEHGELLSGIVCSKTVG
Monoblepharis     FSATHPDDEK-----GDMSVGDTKVIIEGGTILSGTMCKKTVG
Monosiga          KHSTHEDKVDRATPYL--TVHDTKVIISKGVLISGILCKSMLG
Mastigamoeba      QSVSREVAAAQEE--YMNSLLDSVVRIK**KG**QLHTGVITNKAVG
Thalassiosira     KNNGGPGPNT-------FNAYDNLVRVMEGELIEGTIDKKTIG
Acanthamoeba      -----ADDEE-----PDMSFTDNKVLIEEGELVSGILNKKTLG
```


Cyanophora -TSAGSLIHVIMNEHGPDTARVFFNMTQKVINNWLINVGFSVG
Glaucocystis -TSSGSLIHVIWNEHGHDIARVFFSMTQKVINNWLINVGFSIG
Oryza -TSTGSLIHVIWEEVGPDAARKFLGHTQWLVNYWLLQNGFSIG
Arabidopsis -TSNGSLVHVIWEEVGPDAARKFLGHTQWLVNYWLLQNGFTIG
Spirogyra -TSGGSLVHIIWEEVGPDAARKFLGHTQWLVNYWLLQQGFSIG
Chlamydomonas -NGAGGLVHLTWLEHGPHAARGFINNVQRTVNYWVLNHGMSIG
Glaucosphaera -SAANSLVHVTWKEHGPEGACDLISNIQTLVNHWLLQRGFSIG
Schizopombe -ASQGGLVHTIWKEKGPEICKGFFNGIQRVVNYWLLHNGFSIG
Coprinopsis -ASAGGLVHVVFREKGPEATKQLFTGLQMIVNFWLFHNGFSIG
Coccidiodes -ATGGGIIHTIYNEYGPEVCMNFFNGAQTVVNYWLLHNGFSIG
Cryptococcus -SANNGLVHVIFRELGHIAARDFFSAVQRVVNYWLLHFGFSVG
Crosophila -TSAGSLLHICFLELGHDIAGRFYGNIQTVINNWLLFEGHSIG
Mouse -TSAGSLVHISYLEMGHDITRLFYSNIQTVINNWLLIEGHTIG
C.elegans -KSAGNLLHVVTLELGYEIAANFYSHIQTVINAWLIREGHTIG
Monoblepharis -SSAGGVIHTIMNEHGPDAAKKFFNGTQKIVNYWLLHNGFSVG
Monosiga -AKGGGVIHAITVEHGLEEARQFYGNIQTVINNWLLIHGHSIG
Mastigamoeba KSGQGSIIHILWKDQGPMAARDFLSRVQLLTNAYILTRGFSVG
Thalassiosira -SGMGGLIHTAWLDVGHEDTARFMNQTQVVTNYWVLQSSFSIG
Acanthamoeba -TSHKSLVHVIWNEHGSEVCKHFLNQVQHVVNYWLLHHGFSVG


Cyanophora IGDTIADDRTMESINSTLKKAKEEVDTVVIEAQQGKMELQPGR
Glaucocystis IGDTIADEATMETINQHIGTAKTRVQQLILDCQQNRLECQPGR
Oryza IGDTIADAATMENINETISKAKNDVKKLIKQFRDNQLEAEAGR
Arabidopsis IGDTIADSSTMEKINETISNAKTAVKDLIRQFQGKELDPEPGR
Spirogyra IGDTIADASTMDTINETIAKAKTEVKDLIEAACEKQLEAQPGR
Chlamydomonas IGDTVADAKTMAKVNEIIEEAKRKVKEVVEKYQQNALEAQPGR
Glaucosphaera IGDTIADEQTRFNVIEIINKAKEEVKRLVRKAQEGELQLLPGK
Schizopombe IGDTIADADTMKEVTRTVKEARRQVAECIQDAQHNRLKPEPGM
Coprinopsis IGDTIADARTMEFITKNIAESKAKVAQVIEDANMDRLKTKPGM
Coccidiodes IGDTIPDLATIQKIEEAVRIRKEEVDSITLSATENTLEPLPGM
Cryptococcus IGDTIVDKATMAGITNRMVEAKEAVQKLIQEAEANRMKPKPGM
Drosophila IGDTIADPQTYNEIQQAIKKAKDDVINVIQKAHNMELEPTPGN
Mouse IGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGN
C.elegans IGDTIADQATYLDIQNTIRKAKQDVVDVIEKAHNDDLEPTPGN
Monoblepharis IGDTIADGVTMQSITDTIKKAKADVAEIIKNAQANKLDAEPGM
Monosiga IGDAVADQDTNKRIEELTQEATRSVDELITKAQADDIKPKPGN
Mastigamoeba TEDTLADPDTLQAVKAEIEGAKRNVKIHIDDARAGRLKVQAGR
Thalassiosira V-CTIADFATMEQIASTINKAKLQVLDLVRQGQRGELETQPGR
Acanthamoeba VGDTIADEETLAKITQTIRKAKDEVKERQLEAQQGQLERQPGR

Cyanophora SFIESFESKVNRILNQARDAAGKTAQGSLQRSNNVKTMLTSGS
Glaucocystis TLLESFENRVNKELNTARDNAGASAQKSLKPSNNVKAMVTSGS
Oryza TTMESFENRVN**EV**LNKARDVAGSSAEKSLSESNNLKAMATAGS
Arabidopsis TMRDTFENRVN**QV**LNKARDDAGSSAQKSLAETNNLKAMVTAGS
Spirogyra TLMESFENRVN**QV**LNKARDDAGRAAQSSLSESNNVKAMVTAGS
Chlamydomonas TIHESFEA**QV**NQILNKCRDDAGKAAYMSVDLSNNIIKMVTAGS
Glaucosphaera TMMESFETGVNKVLNGARDKSGSDAQKSLLKTNNIKRMVNAGS
Schizopombe TLRESFEAKVSRILNQARDNAGRSAEHSLKDSNNVKQMVAAGS
Coprinopsis TLRESFESEVEGILNKARDSSGQYAQKHLKEDNNVKQMVVAGS
Coccidiodes SVRETFESKVSRALNNARDEAGTETEKSLKDLNNAVQMARSGS
Cryptococcus TIRETLEASIAAELNKARDWTGKTTQDNLKADNNVKQMVVSGA
Drosophila TLRQTFENKVNRILNDARDKTGGSAKKSLTEYNNLKAMVVSGS
Mouse TLRQTFENQVNRILNDARDKTGSSAQKSLSEYNNFKSMVVSGA
C.elegans TLRQTFENKVNQILNDARDRTGSSAQKSLSEFNNFKSMVVSGS
Monoblepharis TVRETFESKVNKVLNKARDDAGKKAEKSLKIYNNVKQMVIAGS
Monosiga TVRETFEVEVNTVLNDMVNKAGKSAQNSLSIHNNFKAMSTAGS
Mastigamoeba SLVESFEAKTNKSLQDALSNAGKKSLSSLKYDNNFKLMIESGS
Thalassiosira TMIESFEQFVNKVLNTARDHAGKSAQASLDETNSVKAMV-AGS
Acanthamoeba TMMESFEFVINQILN**KA**RDDAGNSAQKSLRRSNNFKAMVIAGS

Cyanophora KGSFINISQMFACVGQQNVEGKRIPFGFRNRTLPHFTKEEYGP
Glaucocystis KGSFINISQMIACVGQQNVEGKRIPFGFRDRTLPHFHKNDYGP
Oryza KGTFINISQMTACVGQQNVEGKRIPFGFTNRTLPHFTKNDYGP
Arabidopsis KGSFINISQMTACVGQQNVEGKRIPFGFDGRTLPHFTKDDYGP
Spirogyra KGSFINISQMIACVGQQNVEGKRIPFGFVDRTLPHFTKDDYGP
Chlamydomonas KGSNINISQMMGCVGQQNVEGKRIPFGFQNRTLPHFNKDDLGP
Glaucosphaera KGSFINISQICACVGQQNVEGKRISYGFRRRTLPHFVLDDLGP
Schizopombe KGSFINISQMSACVGQQIVEGKRIPFGFKYRTLPHFPKDDDSP
Coprinopsis KGSFINISQMSVCVGQQSVEGRRIPFGFRHRTLPHFTKDDFSP
Coccidiodes KGSTINISQMTAVVGQQSVEGKRIPFGFKYRTLPHFTKDDYSP
Cryptococcus KGSFINISQMSGVVGQQFVEGKRISFGFRHRSLPHFSRDDYGP
Drosophila KGSNINIS**QV**IACVGQQNVEGKRIPYGFRKRTLPHFIKDDYGP
Mouse KGSKINIS**QV**IAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYGP
C.elegans KGSKINIS**QV**IACVGQQNVEGKRIPFGFRHRTLPHFIKDDYGP
Monoblepharis KGSFINIS**QM**TACVGQQNVEGKRIPYGFRYRTLPHFTKDDHSP
Monosiga KGGPINISQIIACVGQQNVEGKRIPFGFKYRTLPHFVKDDYGP
Mastigamoeba KGSEMNICQITGMVGQQNVEGKRIAFGFQRRTLPHFRKDDYEP
Thalassiosira KGSFINISQIIACVGQQNVEGNRIPYGFKRRTLPHFSKDDLGP
Acanthamoeba KGSAINISQVLACVGQQNVEGKRIPFCFRDRTLPHFVKDDFGP

```
Cyanophora      ESRGFVENSYLRGLTPQEFFFHAMGREGGLIDTAVKTSETGYI
Glaucocystis    ESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTSETGYI
Oryza           ESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTSETGYI
Arabidopsis     ESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTSETGYI
Spirogyra       ESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTSETGYI
Chlamydomonas   EARGFVGNSYLRGLTPQEFFFHAMGGREGLIDTAVKTASTGYI
Glaucosphaera   ESRGFVENSYLRGLTATEFFFHAMGGREGLIDTAVKTSETGYI
Schizopombe     ESRGFIENSYLRGLTPQEFFFHAMAGREGLIDTAVKTAETGYI
Coprinopsis     EARGFVENSYLRGLTPQEFFFHAMAGREGLIDTAVKTAETGYI
Coccidiodes     ESRGFVENSYLRGLTPTEFFFHAMAGREGLIDTAVKTAETGYI
Cryptococcus    ESRGFVENSYLRGLTPQEFWFHAMGGREGLIDTAVKTAETGYI
Drosophila      ESRGFVENSYLAGLTPSEFYFHAMGGREGLIDTAVKTAETGYI
Mouse           ESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYI
C.elegans       ESKGFVENSYLAGLTPSEFFFHAMGGREGLIDTAVKTAETGYI
Monoblepharis   ESRGFVENSYLRGLSPQEFFFHAMGGREGLIDTAVKTAETGYI
Monosiga        DSRGFVFNSYFKGLSAQEMFFHAMGGREGLIDTAVKTAETGYI
Mastigamoeba    EARGFVENSYVSGLSPQEFFFHMMGGREGLIDTAVKTSESGYI
Thalassiosira   ESRGFVENSYLSGLTPQEFFFHAMGGREGLIDT-CKTAETGYI
Acanthamoeba    ESRGFVENSYLRGLTPQEFFFHAMGGREGLIDTAVKTAETGYI

Cyanophora      QRRLIKAMEDVMVQYDNTTRNSMGEIIQFLYGEDGMAGEFIEG
Glaucocystis    QRRLIKAMEDVMVQYDSTLRNSIGDIIQFVYGEDGMDAVYVEN
Oryza           QRRLVKAMEDIMVKYDGTVRNSLGDVIQFLYGEDGMDAIWIES
Arabidopsis     QRRLVKAMEDIMVKYDGTVRNSLGDVIQFLYGEDGMDAVWIES
Spirogyra       QRRLVKAMEDVMVKYDGTVRNSLGDVIQFLYGEDGMDAVWIES
Chlamydomonas   QRRLVKAMEDMVVRYGGTVRNSQGDVVQFLYGEDGMDAVRIEG
Glaucosphaera   QRRLIKALEDVMVKYDGTVRNSTGLIVQFLYGEDGMDGAMVED
Schizopombe     QRRLVKAMEDVMVRYDGTVRNAMGDIIQFAYGEDGLDATLVEY
Coprinopsis     QRRLVKALEDVMVCYDGTVRNSLGDLIQFVYGEDGMDAAFIEK
Coccidiodes     QRKLVKALEEVMVKYDGTVRNSLGDVIQFLYGEDGLDGAFIEN
Cryptococcus    QRRLVKAMEDLKVAYDGTVRNSVNEVVQFLYGEDGMDGAAMEK
Drosophila      QRRLIKAMESVMVNYDGTVRNSVGQLIQLRYGEDGLCGELVEF
Mouse           QRRLIKSMESVMVKYDATVRNSINQVVQLRYGEDGLAGESVEF
C.elegans       QRRLIKAMESVMVNYDGTVRNSLAQMVQLRYGEDGLDGMWVEN
Monoblepharis   QRRLVKALEDVMVKYDGTVRNSLGHVVQFVYGEDGMDATGVER
Monosiga        QRRLVKSMEGLRLNYDGSIRNSNGDLIQLCYGEDGMDGACLEK
Mastigamoeba    QRKLMKSMEDISVKYDTTVRNAAGMILQFAYGEDGIDGTAHEN
Thalassiosira   QRRLVKAMETVMARYDGTLRTSGGQIVQFLYGEDGMDAVWIER
Acanthamoeba    QRRLVKALEDVMVKYDYTVRNSLGDVIQFLYGEDGMDGQTVET
```

```
Cyanophora      QRLESLELDDKEMEKKFKMNLDD-----RNFGEGFLDPDVAAE
Glaucocystis    QKLESMKMGDKEFGKVYKFEPDA-----RNFGDGFMDADIVQK
Oryza           QKLDSLKMKKAEFDNVFRYELDD-----ENWKPNYLSTQHAED
Arabidopsis     QKLDSLKMKKSEFDRTFKYEIDD-----ENWNPTYLSDEHLED
Spirogyra       QSLPSMKMNKSTFDATYRYEIDQE-----DWSPDYMDPQFAKD
Chlamydomonas   QMFEYLKWDPAKLDKAYRIDTTR------DMPPDWLSAEEYEA
Glaucosphaera   QNLRTLSMNDAKLEQTFHLDPFDLDFGIGENRRSYLAQEVIDD
Schizopombe     QVFDSLRLSTKQFEKKYRIDLME---------DRSLSLYMENS
Coprinopsis     QNIETFALSDKEFDHRYRVDVMDP-------AGGFTENALQVG
Coccidiodes     QRVDVIRCSDELFRNRFRIDLMDPEN---SLSPELLEQAT--E
Cryptococcus    QSLDIIRLSDQAFERRYKIDVLGGSG----FSKGILQAGID--
Drosophila      QNMPTVKLSNKSFEKRFKFDWSNERL-----MKKVFTDDVIKE
Mouse           QNLATLKPSNKAFEKKFRFDYTN-----ERALRRTLQEDLVKD
C.elegans       QNMPTMKPNNAVFERDFRVSVAQN----AIKLMDLTDNKFLRK
Monoblepharis   QQLETIRMTEEEFKRRYKLDLMTPETR--KWLPTVVTNDVRHE
Monosiga        QGLPTLTPSDQGF-ERDFCNTGNKLAS----LRQYLVGNIVDD
Mastigamoeba    QDIPSLNADDAKMRKLYDWNTADWDQ----YRPDFAPETL-DE
Thalassiosira   QSFDSLTLNKREFDERYLLHSDDPDFG-DDQNIPFLEAEVIED
Acanthamoeba    QPLDALKMSNEGVVLKYQHDYDSP-----SWGEGWIDPVIADA


Cyanophora      IRGSA--VIRHQLDKE---------YEQLLKDREQLRSEIVQS
Glaucocystis    IRDSA--DDRILLEAE---------FKQLMNDRTALRESIPTD
Oryza           LKTIS--EIRNVFEAE---------VQKLEADRFQLGTEIATT
Arabidopsis     LKGIR--ELRDVFDAE---------YSKLETDRFQLGTEIATN
Spirogyra       AKIVA--EFRQVMDAE---------VLQLEQDRRTLGLEIAPT
Chlamydomonas   LRTDP--AVEQAMRDE---------MAQIKEDLRVLREEVLTN
Glaucosphaera   VRND--PQLIKALEDE---------FAQIKLDRDRLQREIIRS
Schizopombe     IENDS--SVQDLLDEE---------YTQLVADRELLCKFIFPK
Coprinopsis     IDDSS-LELQSKLDEE---------YARLVEDRRLLRNFIFPR
Coccidiodes     ITGD--MEVQRYLDEE---------WEQLQKDRAFLR-SVAKE
Cryptococcus    --QSS-ISLQKLLDEE---------FAQISEDRRILRSEIYPD
Drosophila      MTDSS--EAIQELEAE---------WDRLVSDRDSLR-QIFPN
Mouse           VLSNA--HIQNELERE---------FERMREDREVLR-VIFPT
C.elegans       NYSED--VVREIQESEDGISLVESEWSQLEEDRRLLR-KDFPR
Monoblepharis   LLGDGMDENAKVLEME---------YTQLVEDRNVMRNFIFRN
Monosiga        INADD--EMLQELEDE---------WQRLQDDRQFLR-SVFPK
Mastigamoeba    IRRDA--RVAEVVERE---------FAQLMADRATVR-AIKHA
Thalassiosira   CRHNP--EVQQMLDRE---------MEVLKEDQAMLR-IIMAN
Acanthamoeba    IANSP--EKKRILDKE---------FEQILADRRLLREKIFLA
```

```
Cyanophora      G------EGAVHLPVNLKRLIWNAQKIYKVDTSKP--SDLRPD
Glaucocystis    -------ENTWPLPVNLKRLIWNAQKIFHIDVRKP--SDLDPV
Oryza           G------DNTWPMPVNLKRLIWNAQKTFKIDLRRP--SDMHPM
Arabidopsis     G------DSTWPLPVNIKRHIWNAQKTFKIDLRKI--SDMHPV
Spirogyra       G--DSS----WPLPVNIKRLIWNAQKIFKIDLRKP--SDMNPM
Chlamydomonas   G--DE----KVNIPLNLARLIWNAQTKFNCKPHRPGWTGLQVK
Glaucosphaera   RES------KWPLAVNIDRLIWNVKTLFNIRQDSV--SDLNPR
Schizopombe     GDA------RWPLPVNVQRIIQNALQIFHLEAKKP--TDLLPS
Coprinopsis     RSAN----VDFYLPVNLQRLVQNASQIFHIDRRKP--SDLDPA
Coccidiodes     D---EE---MMQLPINVQRILESAKTTFRIREGTI--SDLHPA
Cryptococcus    GTP------GHPLPVNIQRVIQNSQQIFHIDPRVP--SDLDPV
Drosophila      GES------KVVLPCNLQRMIWNVQKIFHINKRLP--TDLSPI
Mouse           G------DSKVVLPCNLLRMIWNAQKIFHINPRLP--SDLHPI
C.elegans       G--DAK----IVLPCNLLRLIWNAQKIFKVDLRNA--VNLSPL
Monoblepharis   GDS------NWPLPSNIRRLIWNARNMAKISHNTL--SNLRPE
Monosiga        GEAD------VVLPVQLRRLIISAQKEFNIDPRRP--TDLSPA
Mastigamoeba    T--GAR-DDNVFLPVHVGRIIAKAKQHYHISETTR--SDLSPA
Thalassiosira   REAGRESDVNSYAPGNVKRVIQNALRQFQIDKGPS---DLHPK
Acanthamoeba    G------DDRWPLPVNLTRMILNAQKIFHLGPKK--VSDLDPC

Cyanophora      IILQGVELLSTQLVVIA----------GED----------
Glaucocystis    DIIKGVNLLTKQLVVIP----------GED----------
Oryza           EIVDAIDKLQERLKVVP----------GDD----------
Arabidopsis     EIVDAVDKLQERLLVVP----------GDD----------
Spirogyra       DVVDGMDKLQERLKVVV----------GDD----------
Chlamydomonas   EVITKVRELCERLVVVI----------GSD----------
Glaucosphaera   DILKGVQILLSRCNVVALPPKSIKLLDEGDGDQVAEVDTNSVS
Schizopombe     DIINGLNELIAKLTIFR----------GSD----------
Coprinopsis     YIVDSIHELGKRLVIVR----------GDD----------
Coccidiodes     DVIPQVQGLLDRLVVVR----------GDD----------
Cryptococcus    YLLEQRDALADRLLVVR----------GDD----------
Drosophila      RVIKGVKTLLERCVIVT----------GND----------
Mouse           KVVEGVKELSKKLVIVN----------GDD----------
C.elegans       HVISGVRELSKKLIIVS----------GND----------
Monoblepharis   EIVRDVNDLLSRCIVIT----------GED----------
Monosiga        HVAQAIRTFTDRLVVIP----------GDD----------
Mastigamoeba    DIYTTLDLLITRELNITG-----AAYAEGRR----------A
Thalassiosira   DVIEKIEAMLRRLVVVV----------GDD----------
Acanthamoeba    QIVEDLGNLIERLVVIP----------GQD----------
```

```
Cyanophora      -TLSREAQENATRLFNILRATFLATKRVLSEYRLSAAALEWLL
Glaucocystis    -PLSVEAQANATMLFQILVRSCLASKRVLTEYRLTSHAFEWLL
Oryza           -DISIEAQKNATLFFNILLRSTFASKRVLKEYRLTKEAFEWVI
Arabidopsis     -ALSVEAQKNATLFFNILLRSTLASKRVLEEYKLSREAFEWVI
Spirogyra       -HISREAQKNATLFFNCLLRSTMSSKQVLKDYRLNMEAFQWVI
Chlamydomonas   -GLSVEAQRNATIMFHSLVRMHLASKRVMSEYKLNNKALDWLL
Glaucosphaera   RQLAVEAQENATLLWSIHIRAFLASKVILKEYRLDKKAFMHLL
Schizopombe     -RITRDVQNNATLLFQILLRSKFAVKRVIMEYRLNKVAFEWIM
Coprinopsis     -PLSKEAQDNSTLLFRMHLRATFAARKVIEQMRLTREAFEWIL
Coccidiodes     -VISKEAQENATLLFKAQLRSRLAFKRLVVEYSLNKLAFQHVL
Cryptococcus    -KLSRAAQKNATLVFNMLLRSHLATRRVLEEYHLNREAFDWVI
Drosophila      -RISKQANENATLLFQCLIRSTLCTKYVSEEFRLSTEAFEWLV
Mouse           -PLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDLLL
C.elegans       -EISKQAQYNATLLMNILLRSTLCTKNMCTKSKLNSEAFDWLL
Monoblepharis   -GLSTMAQHNATLLFNILVRHLLATKRVYAEHRLNRNAFRWLI
Monosiga        -PLSKVAQRNATLLFNCNLRSVLSSRQCIVRHRLNTKAFDWIL
Mastigamoeba    DPLGEEQRSTALTMFAIMLRAQLASKVMILKHHITLDCWGYII
Thalassiosira   -LLSVEAQNNATTLYRILIRSYLASKRVLREYRLSEAALIWVL
Acanthamoeba    -DISKEAQANAILLFSILLRFTLASRRVLQEYRLDQISWDWLL

Cyanophora      GVIESRFNQAIAHPGEMVGAIAAQSIGEPATQMT---------
Glaucocystis    GEIESRFLQAIANPGEMVGAIAAQSIGEPATQMTLNTFHYAGV
Oryza           GEIESRFLQSLVAPGEMIGCVAAQSIGEPATQMTLNTFHYAGV
Arabidopsis     GEIESRFLQSLVAPGEMIGCVPAQSIGEPATQMTLNTFHYAGV
Spirogyra       GEIEARFLQSQVAPGEMIGCVAAQSIGEPATQMTLNTFHFAGV
Chlamydomonas   GELETRFFLAQAHAGECVGTVAAQSLGEPTTQMTLNTFHFAGV
Glaucosphaera   GEIENRFLHAKCPPGEMIGAVAPQSIGEPATQMTLNTFHFAGV
Schizopombe     GEVEARFQQAVVSPGEMVGTLAAQSIGEPATQMTLNTFHYAGV
Coprinopsis     GEIETKFNQSIAHPGEMCGTLAAQSIGEPATQMTLNTFHYAGV
Coccidiodes     GAIESRFAKAAANPGEMVGVLAAQSIGEPATQMTLNTFHFAGV
Cryptococcus    GEVEQIFNKAVVNAAEMVGTLAAQSIGEPATQMTLNTFHYAGV
Drosophila      GEIETRFQQAQANPGEMVGALAAQSLGEPATQMTLNTFHFAGV
Mouse           GEIESKFNQAIAHPGEMVGALAAQSLGEPATQMTLNTFHYAGV
C.elegans       GEIESRFQQAIAQPGEMVGALAAQSLGEPATQMTLNTFHYAGV
Monoblepharis   GEIESRFKAAKVNPAEMVGTIAAQSIGEPATQMTLNTFHFAGV
Monosiga        GEVEKRFMQAQAQPGEMVGALAAQSLGEPATQMTLNTFHLAGV
Mastigamoeba    GEVREKFGRALVAPGEMVGAIAAQSFGEPTTQMTLNTFHSAGV
Thalassiosira   GEIEARFHH-AVSPGEMAGVLAAQSIGEPATQMTLNTFHYAGV
Acanthamoeba    GEITDRFYKSLCDPGEMIGALAAQSIGEPATQMTLNTFHFAGV
```

```
Cyanophora      ----------------------------------------
Glaucocystis    SSKNVTLGVPRLKEIINVAKKVKTPSLVVYLKE---YCRRDSE
Oryza           SAKNVTLGVPRLREIINVAKNIKTPSLSVHLKP---EVNKKKE
Arabidopsis     SAKNVTLGVPRLREIINVAKRIKTPSLSVYLTP---EASKSKE
Spirogyra       SAKNVTLGVPRLREIINVAKKAKTPSLTVYLHK---RIDQNKE
Chlamydomonas   SAKNVTLGVPRLTELINLAKNIKTPTLTVHLLP---GLRENKD
Glaucosphaera   SAKNVTLGVPRLKEIINVAKHCKTPSLTVYLRG---EAARDAE
Schizopombe     SSKNVTLGVPRLKEILNVAKNIKTPSLTIYLMP---WIAANMD
Coprinopsis     SSKNVTLGVPRLKEIINVATNIKTPSLTIHLQP---EIAVAPE
Coccidiodes     SSKNVTLGVPRLKEILNVATHIKTPSMTVYQDP---ARAMDKE
Cryptococcus    ASKSVTGGVPRLKEIINVAVNIRTPALNVYLEP---EYSKTEE
Drosophila      SSKNVTLGVPRLKEIINISKKPKAPSLTVFLTG---GAARDAE
Mouse           SAKNVTLGVPRLKELINISKKPKTPSLTVFLLG---QSARDAE
C.elegans       SAKNVTLGVPRLKEIINVSKTLKTPSLTVFLTG---AAAKDPE
Monoblepharis   GSKNVTLGVPRLKEIINVATNLRTPSLLVCLKANKDGRPLDLP
Monosiga        SAKNVTLGVPRLKEIINVSKKPKTPSMVLFIGE---DKYQDRN
Mastigamoeba    GTKNVTLGVPRLKELINVNMALRTPNISIFVQD---PFCNSEE
Thalassiosira   SAKNVTLGVPRLKEIINVAKTVKTPGLTIYLQN---HVSGDAD
Acanthamoeba    SAKNVTLGVPRLKELINIAKTIKTPSLTVYLEP---HCSRDHE
```

```
Cyanophora      ----------------------------------------
Glaucocystis    KAKAVQCQLEFTTLRHVTAATEIYYDP-------------DLQ
Oryza           LAKNVQCALEYTTLRSVTHATEIWYDP-------------DPL
Arabidopsis     GAKTVQCALEYTTLRSVTQATEVWYDP-------------DPM
Spirogyra       AAKSVQVELEYTTLHSVTEATEIWYDP-------------DPL
Chlamydomonas   AAKEIQSKLEYTCFKNVVKTTEIWYDP-------------VDP
Glaucosphaera   RAKQVQAELQHTTLNHVTQSTEIYYDP-------------NPQ
Schizopombe     LAKNVQTQIEHTTLSTVTSATEIHYDP-------------DPQ
Coprinopsis     LAKNVQQELAFTSLKTVTSAVEIWYDP-------------NPT
Coccidiodes     SAKQLRSVVEHTNLRSVTEATEIYYDP-------------DIQ
Cryptococcus    DAHQIMRKLTYTRLRDITATVEIFYDP-------------KLD
Drosophila      KAKNVLCRLEHTTLRKVTANTAIYYDP-------------DPQ
Mouse           RAKDILCRLEHTTLRKVTANTAIYYDP-------------NPQ
C.elegans       KAKDVLCKLEHT---TVTCNTAIYYDP-------------DPK
Monoblepharis   QAKKIQAKIEYTTLGKLAVQSEVYYDP-------------DLR
Monosiga        TAMKVLFSIEHSTLKTIIHSSSIYYDP-------------DEM
Mastigamoeba    NATKVPLHIKHTTVRDLVSRAEIIYDPMYVVDQMTGEAYFDCT
Thalassiosira   VAKLVHS-IEYTVLTDLTKLTEIYYDP-------------DPV
Acanthamoeba    AAKNVQCSLQHTTLRDVTAATEIFYDP-------------DPV
```

```
Cyanophora      ------------------------------------------
Glaucocystis    NTVIEEDREFVQCYYEIP-DIPQSDFARMSTWLLRIELNREMM
Oryza           GTIIEEDAEFVQSYYEMP--DEDIDPDKISPWLLRIELNREMM
Arabidopsis     STIIEEDFEFVRSYYEMP--DEDVSPDKISPWLLRIELNREMM
Spirogyra       STVIEEDLETVKSYYEMP--DEDVNPELLSPWLLRIELNREMM
Chlamydomonas   ATVIGEDEEMLSSYYDLA--GDDVLPR-LSHWLLRFELNRGAM
Glaucosphaera   DTVIEADQELVRSYYELPD-DENSSAN-LSPWLLRLNLSKEMM
Schizopombe     DTVIEEDKDFVEAFFAIPDEEVEENLYKQSPWLLRLELDRAKM
Coprinopsis     TTIIEEDEVFVESFFAIPDEEIERNLHKQSPWLLRLVLDRSKM
Coccidiodes     STVIEADRDMVESYFIIP-EDVMDDLSRQSKWLLRIILSRPKL
Cryptococcus    STDIEEDKDFVDAFFAI--PDEDIRLELHSPWLLRLELDRAKV
Drosophila      RTVISEDQEFVNVYYEMP----DFDPTRISPWLLRIELDRKRM
Mouse           STVVAEDQEWVNVYYEMP---DDVARI--SPWLLRVELDRKHM
C.elegans       NTVIAEDEEWVSIFYEMP----DHDLSRTSPWLLRIELDRKRM
Monoblepharis   NSVVEADRDWLDAYFEI--EDADALQARLSSWALRIEIDRGRR
Monosiga        NSVIPEDREFLEN-YDIG----DFDMNKWHRWVLRWELSRREI
Mastigamoeba    RTVVPEDEFFVQDNYMQA-DRLNYQTHSYSSFVLRVMLDTKVM
Thalassiosira   NTIVAKDREFVKEYYEMG-EETEEDLRRLSPWVLRIELNQVVM
Acanthamoeba    NTVITEDQDFVRAYFLMP--EEEINTSNLSPWLLRIELNREKM

Cyanophora      ------------------------------------------
Glaucocystis    TDKKLTMKDIDERIAVGLWAG-------DLHVIYSDDNAAK-L
Oryza           VDKKLSMADIAEKIN-HEFDD-------DLSCIFNDDNADK-L
Arabidopsis     VDKKLSMADIAEKIN-LEFDD-------DLTCIFNDDNAQK-L
Spirogyra       VDKKLTMEMINERIN-SEFSQ-------ELSCINSDDNAPK-L
Chlamydomonas   LDKGLTLGAVQKALQ-DEWED-------FINVLVNDDNAEK-L
Glaucosphaera   TDRKLSMNHVKNKIH-HDLGG-------DVNVMASEDNDAN-L
Schizopombe     LDKKLSMSDVAGKIA-ESFER-------DLFTIWSEDNADK-L
Coprinopsis     LDRKLTMHYVASRIA-ECFKS-------DLFVIWSEDNSET-L
Coccidiodes     LDKGLTVQDVAMKIK-ESYPQ-------DIAVIFSDNNADE-Q
Cryptococcus    LEGGYEMSQIVDAIA-ETVGK-------DVFVIHSEDNAPK-L
Drosophila      TDKKLTMEQIAEKIN-VGFGE-------DLNCIFNDDNADK-L
Mouse           TDRKLTMEQIAEKIN-AGFGD-------DLNCIFNDDNAEK-L
C.elegans       VDKKLTMEMIADRIH-GGFG-------NDVHTIYTDDNAEK-L
Monoblepharis   IDKDISMSEIARKIS-SQYQD-------VLHVMHSDDNAPS-L
Monosiga        NTRRILPENVAKLIK-RAMGAVQDTERDTVMVHFQQGNIEDHT
Mastigamoeba    QQMELKEESVVSIVK-AEFED-------AVEIISTVNDLNK-P
Thalassiosira   VDKKIKMNEIAAEI--AEYGS-------DLNVLVSDDNADD-L
Acanthamoeba    TDTKLSMQEIAERIH-ADFGG-------DLNCIYNDDNADK-L
```

```
Cyanophora      ----------------------------------------
Glaucocystis    SLHIRIKNQE----------EEAKNQE----EEASG--DEHE
Oryza           ILRVRITNDE----------AQKGE-I----QDEYG--EDDV
Arabidopsis     ILRIRIMNDE----------GPKGE-L----QDES--AEDDV
Spirogyra       ILRIRIRNDNAP----------KGEGK----DS-VE-GEDDV
Chlamydomonas   IMRIRKDES--------------ADP-------------TED
Glaucosphaera   VLRIRIAAQKEPEK---------MAEGEEHAQEEEED---ECL
Schizopombe     IIRCRIIRDDD----------RKAEDD----DNMIE---EDV
Coprinopsis     VIRCRVLSSGD-------------KDDP----EEVG--LEEDI
Coccidiodes     VIRIRQIQD-------------PKQDEE----DDDDTE--YDV
Cryptococcus    VIRIRVVAEK--------------EDE----ELLGD---EDM
Drosophila      VLRIRIMNNEENK----------FQDE----DEAVDKMEDDM
Mouse           VLRIRIMNSD----------ENKMQEE----EEVVDKMDDDV
C.elegans       VFRLRIAGEDK-------------GEAQ----EEQVDKMEDDV
Monoblepharis   ALVIRINQDE----------AMQESLK----ESAQT---EDV
Monosiga        VMRIRLPRDK-------------SEAD----DSLG----PDA
Mastigamoeba    RMRIRLVRER-----------AEAEAA----VDEDRG--EDD
Thalassiosira   VARIRIVNDLPMVQGMDEDGNPIMTDDD----VELGQE--DDV
Acanthamoeba    ILRIRINNDEE----------NKAPDQ----EGSVG--DDDV
```

```
Cyanophora      ----------------------------------------
Glaucocystis    FLRRIEGDMLTEMALRGIAQIRKVFMRE-------AKKITFDA
Oryza           FLKKIESNMLTEMALRGIPGINKVFIKE------GNVNKFEDN
Arabidopsis     FLKKIESNMLTEMALRGIPDINKVFIKQ-------VRKSRFDE
Spirogyra       FLKQILGNLLNEMALRGIPDIRKVFIR------ELKKQKFDEI
Chlamydomonas   TMLKLSGAIMD-VRLHGVPNIRKVFIRAE------NQHVYDKE
Glaucosphaera   FLRKVEGSLLSEMNLGGVPDIKKVYMREA------DRIVPDPR
Schizopombe     FLKTIEGHMLESISLRGVPNITR--VYMME-----HKIVRQIE
Coprinopsis     FLRQLENTMLNSVTLRGVRGINRVFLTASD-------RVTLAP
Coccidiodes     TLKKLEAHLLDTLTLRGVAGVERAFINE-------KSRVRTLE
Cryptococcus    FLKRIEGTLLDQVELGGITGITR--VFISE----GKQVVVS-Q
Drosophila      FLRCIEANMLSDMTLQGIEAIGKVYMHLPQTD--SKKRIVITE
Mouse           FLRCIESNMLTDMTLQGIEQISKVYMHLPQT--DNKKKIIITE
C.elegans       FLRCIEANMLSDLTLQGIPAISKVYMNQPNT--DDKKRIIITP
Monoblepharis   ILRDVESQLMASLALGGITNIRR--VYISE-----QKVAVQNT
Monosiga        YLRLLEDRILHSITLQGIPDITRGYLVTGDSKQPGKQRIFINE
Mastigamoeba    RLRDMEEIIMG-LHVHGIRGMTKVMVDDS------RTEKWVDE
Thalassiosira   FLKRLEKNMLHTLKLRGVDDVKK--VMRGG----AKKTVWDDE
Acanthamoeba    FLKQIESNMLTEMDLKGIEGIKKVFIRE------DKNKVAIDA
```

```
Cyanophora      ------------------------------------------
Glaucocystis    DHK-------YMHEKEWVLDTEGCNLLEVMSCQDVDHTRTTSN
Oryza           DG--------FKTEKGWMLDTEGVNLLAVMCHEDVDATRTTSN
Arabidopsis     EGG-------FKTSEEWMLDTEGVNLLAVMCHEDVDPKRTTSN
Spirogyra       EG--------FKYEKEWILDTEGVNLLRVMCHEAVDSTRTTSN
Chlamydomonas   KG-------AFRKYEEWILDTEGINLEQVLAFEGVDSRRTMSN
Glaucosphaera   TG-------GLSVEKEWVLDTDGTNLLAVMSHKDVDYTRTVSN
Schizopombe     DGT-------FERADEWVLETDGINLTEAMTVEGVDATRTYSN
Coprinopsis     DGS-----IKTEKGAEWILETDGTNLKGVMCLDGVDSTRTYSN
Coccidiodes     DGSLYKN-SEDPQCKEWVLETSGSALGDVLAIPGVDATRTYSN
Cryptococcus    NG-------EYDQEKEWFLETDGINLKEVMAVDGVNAFRTYSN
Drosophila      TGE-------FKAIGEWLLETDGTSMMKVLSERDVDPIRTSSN
Mouse           DGE-------FKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
C.elegans       EGG-------FKSVADWILETDGTALLRVLSERQIDPVRTTSN
Monoblepharis   ATG------AIEQAQEHILETDGINLREVMAVEEVDFTRTYSN
Monosiga        EGE-------FKPQDQFLVETDGSSIQSVFAQRYVDQERSTSN
Mastigamoeba    RGALQTR-RFWPLDGEVKKDTAPPLL-DVMVHPAVDPYRTTTN
Thalassiosira   KG--------FGITDEWVLETDGTNLMSVLGIDYVDATRTISN
Acanthamoeba    RG-------EYTKANELVLDTEGTNLLAVMSHPDVDHTRTTSN
```

```
Cyanophora      ------------------------------------------
Glaucocystis    DIVEIIQVLGIEAVRQAALKELRDVISFDGSYVNYRHLATLAD
Oryza           HLIEVIEVLGIEAVRRALLDELRVVISFDGSYVNYRHLAVLCD
Arabidopsis     HLIEIIEVLGIEAVRRALLDELRVVISFDGSYVNYRHLAILCD
Spirogyra       HLTEVMEVLGIEAVRQSLLKELRDVISFDGSYVNYRHLAILCD
Chlamydomonas   SIIEVLEVLGI**EA**ARAALFKEVRNVIQFDGSYVNYRHLACLVD
Glaucosphaera   DVVEMFVTLGIEGVRAALLNEIRGVISFDGAYVNYRHLAILAD
Schizopombe     SFVEILQILGIEATRSALLKELRNVIEFDGSYVNYRHLALLCD
Coprinopsis     NCVEIFTVLGVEAARAALMKEIRNVIEFDGSYVNYRHLGLLCD
Coccidiodes     QFVEILEVFGIEATRTALLRELTQVLAFDGSYVNHRHLALLCD
Cryptococcus    NCYEVYETLGIEAARNALYKELNGVIEMGGSYVNYRHLALLCD
Drosophila      DICEIFQVLGIEAVRKSVEKEMNAVLQFYGLYVNYRHLALLCD
Mouse           DIVEIF**TV**LGIEAVRKALERELYHVISFDGSYVNYRHLALLCD
C.elegans       DICEIFEVLGIEAVRKAIEREMDNVISFDGSYVNYRHLALLCD
Monoblepharis   APLEIYAVLGIEAGRAALLKEIRKVIEFDGSYVNYRHLALLVD
Monosiga        DIVEIFSTFGIEGARKAIEAEMYNVISFGGSYVNARHLSLLCD
Mastigamoeba    SITVINDVLGVEAARAALLKELRDVLGFDGSYINIRHLLMLVD
Thalassiosira   DIVEVFMVLGIEGVRAAILSELRNVISFDGS-VNYRHLACLVD
Acanthamoeba    NIIETIEVLGIEAVRNALL**RE**LRNVISFDGAYVNYRHLAILAD
```

```
Cyanophora      ------------------------------------------
Glaucocystis    VMTYRGHLMSITRHGINRVPTGAMMRCSFEETVEILMDAAVYA
Oryza           TMTYRGHLMAITRHGINRNDTGPLMRCSFEETVDILLDAAVYA
Arabidopsis     TMTYRGHLMAITRHGINRNDTGPLMRCSFEETVDILLDAAAYA
Spirogyra       IMTYRGHLMSITRHGINRNDTGPMMRCSFEETVDILLDASVYA
Chlamydomonas   IMTCKGCLMAITRHGINRNGNGPMTQCSFEETVDILYRAAMFG
Glaucosphaera   IMTCRGHIMSITRHGINRVETGALMRCSFEETADILLEAAMYG
Schizopombe     VMTSRGHLMAITRHGINRAETGALMRCSFEETVEILMDAAASG
Coprinopsis     VMTHRGTLMPITRHGINRADTGALMRCSFEETVEILMEAAAVG
Coccidiodes     VMTSRGFLMAVTRHGINRADTGALMRCSFEETVEILLDAAAFA
Cryptococcus    LMCSKGALMSITRHGINRTDAGALSRAAFEETVEILLEAAAVG
Drosophila      VMTAKGHLMAITRHGINRQDTGALMRCSFEETVDVLMDAAAHA
Mouse           TMTCRGHLMAITRHGVNRQDTGPLMKCSFEETVDVLMEAAAHG
C.elegans       VMTAKGHLMAYSRHGINRQEVGALMRCSFEETVDILMEAAVHA
Monoblepharis   VMCQKGGLMSITRHGINRTENGALARSSFEETVEILMDAAGSG
Monosiga        IMTTKGYLMAITRHGVNRQNTGVLMKASFEETVDILLEAAAHG
Mastigamoeba    VMTFRGYLMSITRFGMNRTDNGILMRASFEETVDVLQDAAQFA
Thalassiosira   VMTMHGHLMAVDRHGINRVESGPLLRCSFEETVDMLNDAACFA
Acanthamoeba    VMTYRGHLMAITRHGINRVETGCLMRCSFEETADILLEAATFS
```

```
Cyanophora      ------------------------
Glaucocystis    ETDHCRSVLKNIMLGQLCQLGTG
Oryza           ESDPLRGVSENIMLGQLAPIGTG
Arabidopsis     ETDCLRGVTENIMLGQLAPIGTG
Spirogyra       ETDYLRGVTENIILGQLAPIGTG
Chlamydomonas   ERDKLMGVSDNIMMGNMCPLGTG
Glaucosphaera   EVDDMKGVAENIMLGQMAPVGTG
Schizopombe     EKDDCKGISENIMLGQLAPMGTG
Coprinopsis     EKDDCHGIAENVMFGQMAPMGTG
Coccidiodes     ELDDCRGVSENLILGQMAPAGTG
Cryptococcus    DVDDCKGVAENVLLGQMAPMGTG
Drosophila      ETDPMRGVSENIIMGQLPKMGTG
Mouse           ESYPMKGVSENIMLGQLAPAGTG
C.elegans       EEDPVKGVSENIMLGQLARCGTG
Monoblepharis   ETDICRGVSENIILGQLAPLGTG
Monosiga        ETDYLKGVSENIMLGQVIPGGTG
Mastigamoeba    GHDELRGVSDNIMLGQVAPVGTG
Thalassiosira   EEEVLRGVTENIMMGQLARVGTG
Acanthamoeba    ELDPLSGVSENILLGQLPPLGTG
```