The Chloroplast Genome Sequence of *Chara vulgaris* Sheds New Light into the Closest Green Algal Relatives of Land Plants

Monique Turmel, Christian Otis, and Claude Lemieux

Département de Biochimie et de Microbiologie, Université Laval, Québec, Canada

The phylum Streptophyta comprises all land plants and six monophyletic groups of charophycean green algae (Mesostigmatales, Chlorokybales, Klebsormidiales, Zygnematales, Coleochaetales, and Charales). Phylogenetic analyses of four genes encoded in three cellular compartments suggest that the Charales are sister to land plants and that charophycean green algae evolved progressively toward an increasing cellular complexity. To validate this phylogenetic hypothesis and to understand how and when the highly conservative pattern displayed by land plant chloroplast DNAs (cpDNAs) originated in the Streptophyta, we have determined the complete chloroplast genome sequence (184,933 bp) of a representative of the Charales, *Chara vulgaris*, and compared this genome to those of *Mesostigma* (Mesostigmatales), *Chlorokybus* (Chlorokybales), *Staurastrum* and *Zygnema* (Zygnematales), *Chaetosphaeridium* (Coleochaetales), and selected land plants. The phylogenies we inferred from 76 cpDNA-encoded proteins and genes using various methods favor the hypothesis that the Charales diverged before the Coleochaetales and Zygnematales. The Zygnematales were identified as sister to land plants in the best tree topology (T1), whereas *Chaetosphaeridium* (T2) or a clade uniting the Zygnematales and *Chaetosphaeridium* (T3) occupied this position in alternative topologies. *Chara* remained at the same basal position in trees including more land plant taxa and inferred from 56 proteins/genes. Phylogenetic inference from gene order data yielded two most parsimonious trees displaying the T1 and T3 topologies. Analyses of additional structural cpDNA features (gene order, gene content, intron content, and indels in coding regions) provided better support for T1 than for the topology of the above-mentioned four-gene tree. Our structural analyses also revealed that many of the features conserved in land plant cpDNAs were inherited from their green algal ancestors. The intron content data predicted that at least 15 of the 21 land plant group II introns were gained early during the evolution of streptophytes and that a single intron was acquired during the transition from charophycean green algae to land plants. Analyses of genome rearrangements based on inversions predicted no alteration in gene order during the transition from charophycean green algae to land plants.

Introduction

More than 470 MYA, green algae belonging to the class Charophyceae emerged from their aquatic habitat to colonize the land (Kenrick and Crane 1997; Graham, Cook, and Busse 2000; Lewis and McCourt 2004; Sanderson et al. 2004), giving rise to more than 500,000 land plant species currently found on our planet. In contrast to the large diversity of land plants, only a few thousand charophycean species are living today. These algae exhibit great variability in cellular organization and reproduction (Lewis and McCourt 2004), and together with the land plants, they form the green plant lineage Streptophyta (Bremer et al. 1987). Most, if not all, of the other extant green algae (more than 10,000 species) belong to the sister lineage Chlorophyta (Lewis and McCourt 2004). Six monophyletic groups of charophycean green algae are currently recognized: the Mesostigmatales (Karol et al. 2001), Chlorokybales, Klebsormidiales, Zygnematales, Coleochaetales, and Charales (Mattox and Stewart 1984), given here in order of increasing cellular complexity. Whether the unicellular flagellate *Mesostigma viride* (Mesostigmatales) belongs to the Streptophyta remains controversial: some phylogenetic analyses placed this alga at the base of the Streptophyta (Bhattacharya et al. 1998; Marin and Melkonian 1999; Karol et al. 2001; Martin et al. 2002) and others before the divergence of the Chlorophyta and Streptophyta (Lemieux, Otis, and Turmel 2000; Turmel et al. 2002; Turmel, Otis, and Lemieux 2002*a*; Martin et al. 2005).

On the basis of morphological characters, the Charales or the Coleochaetales have been proposed to share a sister relationship with land plants (Qiu and Palmer 1999; Chapman and Waters 2002). Recent analyses of the combined sequences of four genes from the nucleus (18S rRNA gene), chloroplast (*atpB* and *rbcL*), and mitochondria (*nad5*) of 25 charophycean green algae, eight land plants, and five chlorophytes revealed that the Charales and land plants form a highly supported clade; however, moderate bootstrap support was observed for the positions of the other charophycean groups (Karol et al. 2001; McCourt, Delwiche, and Karol 2004). The best trees inferred by Bayesian inference and maximum likelihood (ML) in this four-gene analysis support an evolutionary trend toward increasing cellular complexity (McCourt, Delwiche, and Karol 2004). All previously reported phylogenies of charophycean green algae were reconstructed from a smaller number of genes, showed poor resolution, and yielded conflicting topologies, thus providing no conclusive information regarding the branching order of charophycean lineages and their specific relationships with land plants (Qiu and Palmer 1999; Chapman and Waters 2002; McCourt, Delwiche, and Karol 2004).

To unravel the phylogenetic relationships among charophycean lineages and to understand how and when the highly conservative pattern displayed by land plant chloroplast DNAs (cpDNAs) originated in the Streptophyta, we have undertaken the sequencing of the chloroplast genome from representatives of all charophycean lineages. We have reported thus far the cpDNA sequences of *Mesostigma* (Mesostigmatales) (Lemieux, Otis, and Turmel 2000), *Chaetosphaeridium globosum* (Coleochaetales) (Turmel, Otis, and Lemieux 2002*b*), *Staurastrum punctulatum*, and *Zygnema circumcarinatum* (Zygnematales) (Turmel, Otis, and Lemieux 2005). Comparative analyses of *Mesostigma* cpDNA

Key words: Streptophyta, charophycean green algae, phylogenomics, chloroplast genome evolution, genome rearrangements, introns.

E-mail: monique.turmel@rsvs.ulaval.ca.

Mol. Biol. Evol. 23(6):1324–1338. 2006
doi:10.1093/molbev/msk018
Advance Access publication April 12, 2006

(137 genes, no introns) with its land plant counterparts (110–120 genes, about 20 introns) revealed that the chloroplast genome underwent substantial changes in its architecture during the evolution of streptophytes (namely gene losses, intron insertions, and scrambling in gene order). At the levels of gene content (125 genes), intron composition (18 introns), and gene order, *Chaetosphaeridium* cpDNA is highly similar to land plant cpDNAs. *Mesostigma*, *Chaetosphaeridium*, and most land plant cpDNAs exhibit a quadripartite structure that is characterized by the presence of two copies of an rRNA-containing inverted repeat (IR) separated by large single-copy (LSC) and small single-copy (SSC) regions. All the genes they have in common, with a few exceptions, reside in corresponding genomic regions. Although the chloroplast genomes of the zygnematalean algae *Staurastrum* and *Zygnema* closely resemble their *Chaetosphaeridium* and bryophyte counterparts at the gene content level, they feature substantial differences in overall structure, gene order, and intron content (8 and 13 introns in *Staurastrum* and *Zygnema*, respectively). Like the partially characterized cpDNA of *Spirogyra maxima* (Manhart, Hoshaw, and Palmer 1990), they lack a large IR, suggesting that loss of one copy of the IR sequence occurred very early during the evolution of this group of charophycean green algae. From these observations, we inferred that the chloroplast genome of the last common ancestor of *Staurastrum* and *Zygnema* carried at least 16 of the approximately 20 introns found in land plant cpDNAs and bore more similarity to its land counterparts than to either zygnematalean cpDNA at the gene order level.

In the present study, we report the complete cpDNA sequence of *Chara vulgaris* (Charales). We have compared this genome sequence with that of *Chlorokybus atmophyticus* (Chlorokybales) and with those previously reported for the above-mentioned charophycean green algae and selected land plants. Our phylogenetic analyses of shared proteins and genes as well as our analyses of independent sets of structural genomic data provide robust support for the notion that the Charales occupy a basal position relative to both the Coleochaetales and Zygnematales, thus challenging our current view about the phylogeny of charophycean green algae. Our comparative analyses of structural genomic data also indicate that the chloroplast genome remained largely unchanged at the gene order, gene content, and intron composition levels during the transition from charophycean green algae to land plants.

Materials and Methods

DNA Cloning, Sequencing, and Sequence Analysis

Chara vulgaris was collected from a pond located in Quebec City (Quebec, Canada); a voucher (no. QFA468020) of this plant material is held at the herbarium Louis-Marie of Laval University. A random clone library was prepared from a fraction containing both cpDNA and mitochondrial DNA (Turmel, Otis, and Lemieux 2003). DNA templates were obtained with the QIAprep 96 Miniprep kit (Qiagen Inc., Mississauga, Canada) and sequenced as described previously (Turmel, Otis, and Lemieux 2005). Sequences were assembled using SEQUENCHER 4.1.1 (Gene Codes Corporation, Ann Arbor, Mich.). Genes were identified using a custom-built suite of bioinformatic tools allowing the automated execution of the three following steps: (1) open reading frames (ORFs) were found using GETORF in EMBOSS (Rice, Longden, and Bleasby 2000), (2) their translated products were identified by BlastP (Altschul et al. 1990) searches against a local database of cpDNA-encoded proteins or the nonredundant database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Blast/), and (3) consecutive 100-bp segments of the genome sequence were analyzed with BlastN and BlastX (Altschul et al. 1990) to identify genes. tRNA genes were localized using tRNAscan-SE (Lowe and Eddy 1997). Intron boundaries were determined by modeling intron secondary structures (Michel, Umesono, and Ozeki 1989; Michel and Westhof 1990) and by comparing intron-containing genes with intronless homologs using FRAMEALIGN of the Genetics Computer Group (Madison, Wis.) software (version 10.3) package. Repeated sequence elements were searched using GenAlyzer (Choudhuri et al. 2004) and Vmatch (http://www.vmatch.de/).

Phylogenetic Reconstructions from cpDNA Sequences

A data set of 76 concatenated protein sequences derived from the chloroplast genomes of *Chara* (this study), *Mesostigma* (Lemieux, Otis, and Turmel 2000), *Chlorokybus* (M. Turmel, C. Otis, and C. Lemieux, unpublished data), *Staurastrum* (Turmel, Otis, and Lemieux 2005), *Zygnema* (Turmel, Otis, and Lemieux 2005), *Chaetosphaeridium* (Turmel, Otis, and Lemieux 2002*b*), *Marchantia polymorpha* (Ohyama et al. 1986), *Anthoceros formosae* (Kugita et al. 2003), and *Physcomitrella patens* (Sugiura et al. 2003) was prepared as described previously (Turmel, Otis, and Lemieux 2003). In addition, two nucleotide data sets containing the genes for these proteins and differing only by the presence/absence of third codon positions were prepared. To obtain the data set with all three codon positions, the multiple sequence alignment of each protein was converted into a codon alignment, the poorly aligned and divergent regions in each codon alignment were excluded using GBLOCKS 0.91b (Castresana 2000) with the $-t = c$ option, and the individual codon alignments were concatenated. Third codon positions were excluded from the latter data set with PAUP* 4.0b10 (Swofford 2002). Using the same methods, we also generated amino acid and nucleotide data sets with a broader representation of land plants by including the 56 protein-coding genes shared by the nine above-mentioned streptophytes, the lycophyte *Huperzia lucidula* (Wolf et al. 2005), and the 23 other tracheophyte taxa examined by Leebens-Mack et al. (2005).

The amino acid data sets were used to reconstruct phylogenies with ML, maximum parsimony (MP), and distance methods. ML trees were computed with PHYML 2.4.4 (Guindon and Gascuel 2003) under the cpREV45 + Γ + I model (Adachi et al. 2000). Bootstrap support for each node was calculated using 100 replicates. The confidence limits of the alternative tree topologies recovered in analyses of the 76-protein data set were evaluated under the cpREV45 + Γ + I model using the Shimodaira-Hasegawa test as implemented in CODEML 3.14 (Yang 1997).

Support of the individual proteins for alternative topologies was estimated with CODEML 3.14 using the MGENE = 1 option. MP trees and ML-distance trees were inferred using PROTPARS and NEIGHBOR, respectively, in PHYLIP 3.63 (Felsenstein 1995). The ML distances were computed with PUZZLEBOOT 1.03 and Tree-Puzzle 5.2 (Strimmer and von Haeseler 1996) under the cpREV45 + Γ + I model. Robustness of MP and distance trees was assessed by bootstrap percentages after 100 replications.

ML, MP, and ML-distance trees were inferred from each nucleotide data set using PAUP* 4.0b10. Trees were searched with the full heuristic option, and optimization was performed by branch swapping using tree bisection and reconnection. For ML analyses, starting trees were obtained by the stepwise addition of sequences. For MP and ML-distance analyses, starting trees were obtained by the stepwise random addition of sequences with one tree held per addition; 10 replications of the addition procedure were performed. ML and ML-distance trees were constructed under the GTR + Γ + I model, a model that was selected by Modeltest 3.6 (Posada and Crandall 1998) as the one best fitting our nucleotide data. Confidence of branch points was estimated by 1,000 bootstrap replications in all analyses. Shimodaira-Hasegawa tests of alternative tree topologies were carried out using PAUP* 4.0b10.

LogDet-distance trees were computed using PAUP* 4.0b10 with the neighbor-joining search setting. The LogDet distances were calculated with LDDist (Thollesson 2004), and the proportion of invariant sites was estimated using the capture-recapture method of Steel, Huson, and Lockhart (2000). Confidence of branch points was estimated by 1,000 bootstrap replications.

Relative rate tests were performed with RRTree (Robinson-Rechavi and Huchon 2000) using the 76-gene data set containing all three codon positions and the topology of the best ML tree. The three land plant sequences and the two zygnematalean sequences were treated as single lineages, and the *Chlorokybus* sequence served as an outgroup. The number of synonymous substitutions per synonymous site (K_s) and the number of nonsynonymous substitutions per nonsynonymous site (K_a) were compared.

Analyses of Gene Order Data

The GRIMM Web server (Tesler 2002) was used to infer the number of gene permutations by inversions in pairwise comparisons of green algal cpDNAs. For comparisons involving two IR-containing genomes, genes within one of the two copies of the IR were excluded from the data set, and the SSC and LSC + IR regions were considered as two separate chromosomes. The SSC and LSC regions were assumed to be independent from one another because the conserved gene-partitioning pattern displayed by streptophyte genomes is not consistent with the occurrence of inversions spanning the SSC and LSC regions. For comparisons involving an IR-containing genome and an IR-lacking genome, genes within one copy of the IR were excluded, and no constraints were imposed on the inversion endpoints. Phylogenies based on inversion medians were inferred with GRAPPA 2.0 (http://www.cs.unm.edu/~moret/GRAPPA/) using the algorithm of Caprara (2003) and a data set of 113 gene positions.

Analyses of Other Structural Genomic Data

MacClade 4.06 (D. Maddison and W. Maddison 2000) was used to generate matrices of gene and intron contents, to trace the encoded characters on tree topologies, and to calculate tree lengths under the Dollo parsimony method (Farris 1977). The presence of a gene, the presence of a pseudogene, and the absence of a gene were coded with values of 2, 1, and 0, respectively. The presence/absence of an intron was coded as binary characters. Gene-coding regions carrying unambiguous insertions/deletions were identified by visual inspection of the nucleotide and amino acid alignments, and indels were coded as unordered, binary characters. The resulting binary data matrix was subjected to parsimony analysis using PAUP* 4.0b10.

Results

Features of *Chara* cpDNA

Chara cpDNA consists of a circular molecule of 184,933 bp with the typical quadripartite structure found in streptophyte cpDNAs (fig. 1). Table 1 compares the main features of this genome with those of its counterparts in *Mesostigma*, other charophycean green algae, and a representative of land plants, the bryophyte *Marchantia* (a liverwort). Most land plant chloroplast genomes that have been completely sequenced to date closely resemble *Marchantia* cpDNA in gene content, intron content, and gene order. Among all streptophyte chloroplast genome sequences determined thus far, that of *Chara* has the largest size, the longest single-copy regions, and the highest A + T content. Its increased size is mainly accounted for by expanded intergenic spacers and introns representing 38.8% and 13.4% of the total genome, respectively. The average size of intergenic spacers in *Chara* cpDNA is 521 bp versus 223 bp and 178 bp in *Chaetosphaeridium* and *Marchantia* cpDNAs, respectively, whereas the average intron size is 1,372 bp versus 686 bp and 650 bp in *Chaetosphaeridium* and *Marchantia*, respectively. Probing of *Chara* cpDNA with Vmatch and GenAlyzer revealed that dispersed repeats $\geq$25 bp are virtually absent. Both intergenic spacers and introns from *Chara* are richer in A + T (82.9% and 78.5%, respectively) than their homologs in *Chaetosphaeridium* (79.8% and 77.6%) and *Marchantia* (80.6% and 76.8%).

At both the levels of gene content and intron composition, *Chara* cpDNA bears most similarity with *Chaetosphaeridium* cpDNA (table 1). With 127 genes, its gene repertoire is slightly larger than those of *Chaetosphaeridium*, *Staurastrum*, *Zygnema*, and *Marchantia* cpDNAs. These four charophycean genomes differ for the presence/absence of 13 genes (Supplementary Table S1, Supplementary Material online), and interestingly, *Chara* cpDNA features four genes (*rpl12*, *trnL*(gag), *rpl19*, and *ycf20*) that are entirely missing from the three other charophycean cpDNAs and land plant cpDNAs. Considering that *rpl12* is also absent from *Mesostigma* cpDNA, *Chara* is the first streptophyte in which this chloroplast gene is

FIG. 1.—Gene map of *Chara* cpDNA. Genes (filled boxes) shown on the outside of the map are transcribed in a clockwise direction, whereas those on the inside of the map are transcribed counterclockwise. Genes absent from *Marchantia* cpDNA are represented in beige. Gene clusters shared with *Marchantia* cpDNA are shown as alternating series of green and red boxes. Genes present in *Marchantia* cpDNA but located outside conserved clusters are shown in gray. Genes containing introns (open boxes) are denoted by asterisks. The intron sequences bordering the *rps12* exons (*rps12a* and *rps12b*) are spliced in *trans* at the RNA level. tRNA genes are indicated by the one-letter amino acid code (Me, elongator methionine and Mf, initiator methionine) followed by the anticodon in parentheses.

reported. With regards to *tufA*, our results are consistent with previous reports indicating that this chloroplast gene is functional in Charales (Baldauf, Manhart, and Palmer 1990) but not in Coleochaetales (Baldauf, Manhart, and Palmer 1990; Turmel, Otis, and Lemieux 2002*b*). In contrast to its *Chaetosphaeridium* and Coleochaete counterparts, the *tufA* sequence in *Chara* cpDNA shows substantial conservation along its entire length with its *Mesostigma* and chlorophyte counterparts (data not shown). As in *Chaetosphaeridium* cpDNA, a total of 18 introns reside in *Chara* cpDNA; however, they are not all inserted at common sites in the two genomes (fig. 2). *Chara* cpDNA lacks three group II introns (in *petD*, *rpoC1*, and *ycf66*) found in *Chaetosphaeridium*, *Zygnema*, and land plant cpDNAs and exhibits one group II intron (in *ycf3* at site 124) missing in its *Chaetosphaeridium* counterpart but present in zygnematalean and land plant cpDNAs. Moreover, in contrast to *Chaetosphaeridium* cpDNA whose introns all have homologs in land plant cpDNAs, two of the *Chara* chloroplast introns, the group II intron in *atpF* at site 138 and the group I intron in *rrl* at site 2500, have not been previously identified in streptophyte cpDNAs. The latter *rrl* intron, however, has an homolog at the same position within the large subunit rRNA gene of *Chara* mitochondrial DNA

Table 1
General Features of Streptophyte cpDNAs

Feature	*Mesostigma*	*Chlorokybus*	*Chara*	*Chaetosphaeridium*	*Zygnema*	*Staurastrum*	*Marchantia*
Size[a] (bp)							
IR	6,057	7,640	10,919	12,431	—	—	10,058
SSC	22,619	27,876	27,280	17,639	—	—	19,813
LSC	83,627	109,098	135,815	88,682	—	—	81,095
Total genome	118,360	152,254	184,933	131,183	165,372	157,089	121,024
A + T content (%)	69.9	61.8	73.8	70.4	68.9	67.5	71.2
Gene content[b]	137	138	127	125	125	121	120
Introns							
Group I	0	1	2	1	1	1	1
Group II							
cis-Spliced	0	0	15	16	11	6	18
trans-Spliced	0	0	1	1	1	1	1
Gene order relative to *Marchantia*							
No. of gene clusters	22	19	14	12	20	22	—
No. of genes in clusters	95	87	113	116	82	101	—
No. of inversions	38	47	17	12	56	34	—

[a] Because *Staurastrum* and *Zygnema* cpDNAs lack an IR, only the genome size is given for each of these cpDNAs.
[b] Unique ORFs, intron ORFs, and pseudogenes were not taken into account. Note that *Chaetosphaeridium tufA* was considered to be a functional gene.

(Turmel, Otis, and Lemieux 2003) as well as in the cpDNAs and mitochondrial DNAs of several chlorophyte green algae (Côté et al. 1993; Turmel et al. 1993; Turmel, Mercier, and Côté 1993; Turmel et al. 1999).

At the level of gene organization, *Chara* cpDNA differs substantially from its land plant and *Chaetosphaeridium* counterparts even though it is much less scrambled than zygnematalean cpDNAs. Relative to *Marchantia* cpDNA, it is more rearranged than is *Chaetosphaeridium* cpDNA. Using GRIMM, we estimated that 17 and 12 inversions would be required to convert the gene orders of *Chara* and *Chaetosphaeridium* cpDNAs, respectively, into that of *Marchantia* cpDNA (table 1). As 23 inversions would be necessary to interconvert the gene orders of the *Chara* and *Chaetosphaeridium* cpDNAs, each charophycean genome more closely resembles its *Marchantia* counterpart. By searching for the presence of all possible gene pairs with conserved polarities in *Marchantia* and charophycean genomes, we found that *Chara* and *Marchantia* share only three gene pairs (3′*ndhF*-5′*trnN*(guu), 3′*psbA*-5′*trnH*(gug), and 3′*trnY*(gua)-5′*trnD*(guc)) that are not present in all other charophyceans, that *Chaetosphaeridium* and *Chara* shares specifically six gene pairs (3′*atpA*-3′*trnR*(ucu), 3′*cemA*-5′*petA*, 3′*chlB*-5′*trnK*(uuu), 3′*psaI*-5′*ycf4*, 3′*psbZ*-5′*trnG*(gcc), and 3′*ycf3*-5′*psaA*), and that *Marchantia* and *Straurastrum/Zygnema* shares specifically three gene pairs (3′*trnV*(uac)-5′*ndhC*, 5′*trnS*(gga)-5′*ycf3*, and 3′*trnH*(gug)-5′*ftsH*).

Phylogenetic Analyses of Chloroplast Sequences

Various phylogenetic inference methods were used to analyze an amino acid data set (16,024 sites) and two nucleotide data sets (codons excluding third positions, 34,370 sites; all three codon positions, 51,555 sites) that were derived from the 76 protein-coding genes common to the cpDNAs of *Mesostigma*, *Chlorokybus*, *Zygnema*, *Staurastrum*, *Chaetosphaeridium*, and the bryophytes *Marchantia* (liverwort), *Anthoceros* (hornwort), and *Physcomitrella* (moss). In all analyses, the *Mesostigma* sequences served as an outgroup to root the trees. As shown in figure 3, all three best ML trees share the same branching order for the charophycean green algae. It can be seen that the clade formed by the two zygnematalean species occupies a sister position

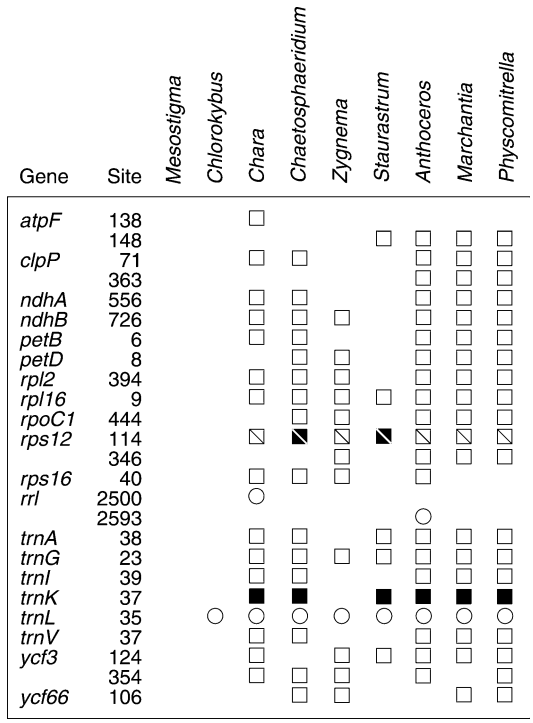

FIG. 2.—Intron distribution in *Chara* cpDNA and other streptophyte cpDNAs. Circles denote the presence of group I introns, and squares denote the presence of group II introns. Divided squares represent *trans*-spliced group II introns. Open symbols denote the absence of intron ORFs, whereas filled symbols denote their presence. Intron insertion sites in protein-coding and tRNA genes are given relative to the corresponding genes in *Mesostigma* cpDNA; insertion sites in *rrl* are given relative to the *Escherichia coli* 23S rRNA. For each insertion site, the position corresponding to the nucleotide immediately preceding the intron is reported. Note that *rps16* is lacking in *Marchantia* and *Physcomitrella* cpDNAs.

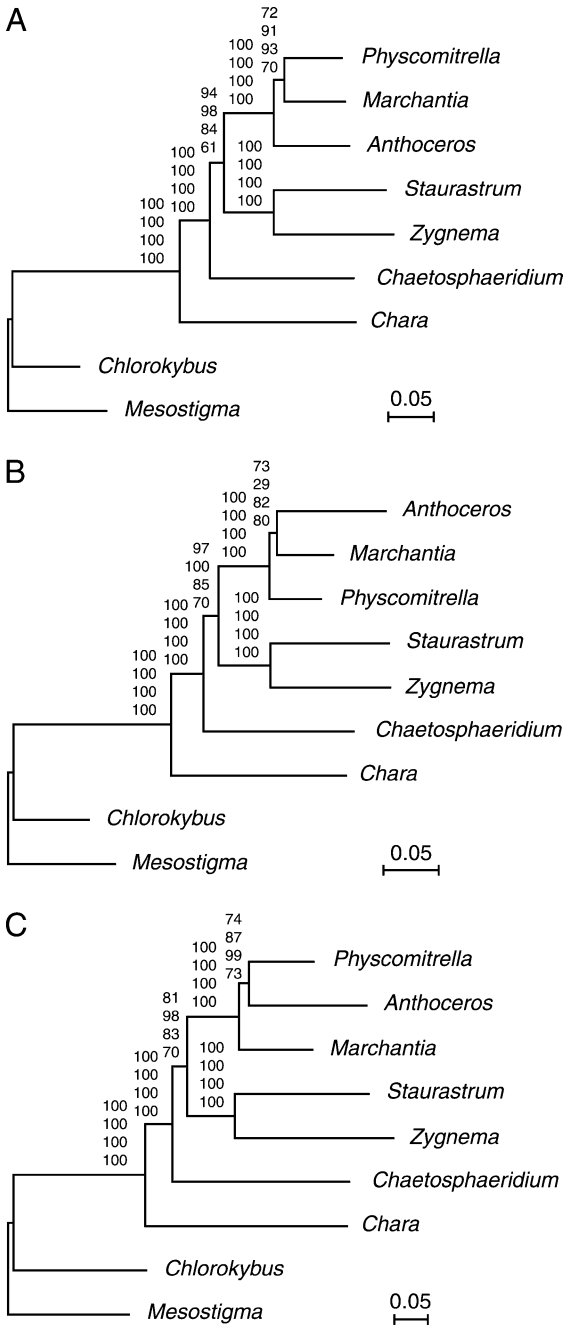

FIG. 3.—Phylogenies inferred from 76 chloroplast protein-coding genes and their deduced amino acid sequences. (*A*) Best ML tree inferred from proteins. (*B*) Best ML tree inferred from first and second codon positions. (*C*) Best ML tree inferred from all three codon positions. Bootstrap values obtained in ML, MP, ML-distance, and LogDet-distance analyses are listed in this order from top to bottom on the corresponding nodes. Branch lengths are drawn to scale. The 76 genes and proteins used for these phylogenetic reconstructions are listed in table S2 (Supplementary Material online).

relative to the land plants, that the representative of the Coleochaetales emerges just before the zygnematalean lineage, that *Chara* is basal relative to the zygnematalean and coleochaetalean lineages, and that *Chlorokybus* is the first lineage emerging after the outgroup. A bootstrap support value of 100% is found at all nodes describing these relationships, except that identifying the Zygnematales as sister to land plants where bootstrap values vary from 81% to 97%. Among the alternative topologies recovered from the analysis of the amino acid data set, the sister group to land plants is either *Chaetosphaeridium* or a clade consisting of *Chaetosphaeridium* and the two zygnematalean green algae. In the single alternative topology recovered from the analysis of the nucleotide data sets, *Chaetosphaeridium* is sister to land plants.

The distance and MP methods yielded protein and gene trees highly congruent with ML trees and also revealed that the alternative topologies for the charophycean green algae differ only in the relative positions of *Chaetosphaeridium* and the zygnematalean green algae (fig. 3). The most notable difference between the trees inferred from the three data sets concern the relative positions of the bryophytes. The earliest diverging bryophyte is *Anthoceros* in the protein trees, either *Physcomitrella* or *Anthoceros* in the gene trees inferred from the first two codon positions, and *Marchantia* in the gene trees inferred from all three codon positions. This instability is not surprising considering that the relative branching order of the moss, hornwort, and liverwort lineages is currently a problematic issue in phylogenetic studies of land plants (Nishiyama et al. 2004; Shaw and Renzaglia 2004; Goremykin and Hellwig 2005). The addition of 32 RNA-coding genes (the three rRNA and 29 tRNA genes) to the data set containing all three codon positions improved slightly the resolution of charophyte lineages but had no significant effect on the resolution of bryophyte lineages (data not shown).

The two alternative topologies differing in the branching order of charophycean green algae were evaluated using the likelihood-based statistical test of Shimodaira-Hasegawa (Shimodaira and Hasegawa 1999). When this test is applied to the amino acid data set and to the nucleotide data set featuring the first two codon positions, it rejects the hypothesis that the Coleochaetales are sister to land plants (T2) with *P* values of 0.043 and 0.049, respectively; however, it does not reject it when the nucleotide data set containing all three codon positions is used. The alternative hypothesis that a clade formed by the Coleochaetales and Zygnematales is sister to land plants (T3) is rejected by analyses of the two nucleotide data sets with *P* values of 0.003 (first two codon positions) and 0.011 (all three codon positions) but not by the analysis of the amino acid data set. We identified no conflict among the proteins in the amino acid data set with respect to their phylogenetic signals; 68 of these proteins favor a single topology among the three detected but are unable to reject either one or both of the other topologies (Supplementary Table S2, Supplementary Material online). A total of 27 individual proteins support the Zygnematales as being the sister to land plants (best topology or T1), 18 proteins support T2, and 23 proteins support T3. The remaining eight proteins equally support the three topologies.

Because most conventional methods of phylogeny reconstructions yield the true topology only when the sequences evolved under a stationary Markov process (Gu and Li 1996), heterogeneity in base and amino acid composition may lead to incorrect topologies (Steel, Lockhart, and Penny 1993; Lockhart et al. 1994). To overcome this potential problem and to also minimize potential problems

of phylogeny reconstruction due to heterotachy (Lopez, Casane, and Philippe 2002) or among-lineage heterogeneity (Kolaczkowski and Thornton 2004; Spencer, Susko, and Roger 2005), we used LogDet distances (Lake 1994; Lockhart et al. 1994; Steel 1994; Gu and Li 1996). LogDet-distance analysis of each of the three data sets analyzed identified the same branching order for the charophycean green algae as did ML, MP, or ML-distance analyses (fig. 3).

The long-branch attraction phenomenon can also lead to erroneous phylogenies (Felsenstein 1978). This reconstruction artifact tends to cluster long branched but otherwise unrelated taxa in the inferred phylogeny. Although there is no reliable way to determine whether a phylogenetic analysis is plagued with this problem, two sets of results provide no indication that the long-branch attraction phenomenon might be a cause of errors in our phylogeny reconstructions. First, relative rate tests using the nucleotide data set containing all three codon positions and *Chlorokybus* as an outgroup detected no significant difference ($P < 0.05$) in substitution rates among streptophyte lineages (data not shown). K_s was found to be saturated in all six pairwise comparisons, and there was no significant variation in K_a in all comparisons. Second, removal of the nine fastest evolving proteins from our amino acid data set (i.e., those with a specific rate of evolution >3.50 relative to AtpA, see Supplementary Table S2, Supplementary Material online) had no effect on the best tree topology recovered in analyses with the ML, MP, ML-distance, and LogDet-distance methods (data not shown). The use of slowly evolving sequences is one of the three approaches proposed by Philippe, Lartillot, and Brinkmann (2005) to reduce the impact of potential long-branch attraction artifacts.

To explore whether the restricted representation of land plants in our phylogenies might have lead to misleading topologies, we analyzed using ML and LogDet-distance methods an amino acid data set (11,025 sites) and a nucleotide data set (22,712 sites, first two codon positions) that were supplemented with 24 tracheophyte taxa (fig. 4). In all analyses, the positions of *Chlorokybus* and *Chara* remained identical to those shown in figure 3 and received robust support. The ML trees provided support for the Zygnematales being sister to all land plants (fig. 4*A* and *C*), whereas the LogDet-distance analyses favored a clade uniting *Chaetosphaeridium* and the Zygnematales as sister to all land plants (fig. 4*B* and *D*). In agreement with previously reported chloroplast phylogenies (Nishiyama et al. 2004; Goremykin and Hellwig 2005), the branching order of the three bryophyte lineages varied depending upon the method of analysis and the data set used. Bryophytes were paraphyletic in the ML gene tree (fig. 4*C*) and monophyletic in the ML protein tree (fig. 4*A*) and the two LogDet trees (fig. 4*B* and *D*). Regarding the relationships among tracheophytes, the ML trees were consistent with recently published chloroplast phylogenies (Goremykin et al. 2005; Leebens-Mack et al. 2005; Wolf et al. 2005).

Testing Phylogenetic Hypotheses Using Structural Genomic Features

The availability of complete chloroplast genome sequences offers the opportunity to test phylogenetic hypotheses using structural genomic features. We analyzed four independent data sets of structural genomic features (gene order, gene content, intron content, and insertions/deletions in gene-coding regions) to evaluate the branching orders of charophycean green algae inferred from the 76 chloroplast proteins and genes analyzed in the present study and from the four genes encoded by the nuclear, chloroplast, and mitochondrial genomes in the study of Karol et al. (2001).

We used a gene order data set featuring all chloroplast genes common to *Marchantia*, *Mesostigma*, and the charophycean algae to infer a phylogeny based on inversion medians using GRAPPA and the algorithm of Caprara (2003). A single land plant taxon was selected because this type of analysis is computationally intensive and feasible with only a limited number of taxa and also because the three bryophytes used in our phylogenetic analyses of cpDNA sequences exhibit very similar gene orders. Although GRAPPA can use breakpoint medians to compute trees, we chose inversion medians because this method has been shown to greatly outperform breakpoint medians in phylogeny reconstruction (Moret et al. 2002). We recovered two best trees with 144 inversion medians that display the T1 and T3 topologies (fig. 5). A separate GRAPPA analysis constrained to the topology of the four-gene tree (T4) yielded a tree featuring nine extra inversions (fig. 5). The difference in length between the latter inversion tree and that carrying the T1 topology is more pronounced when we consider only the portions featuring the crown taxa (i.e., *Marchantia*, the zygnematalean green algae, *Chaetosphaeridium*, and *Chara*); 106 inversions are then scored for the tree compatible with T1 versus 125 inversions for that compatible with T4.

Thirty-six of the 144 chloroplast genes predicted to have been present in the common ancestor of *Mesostigma* and streptophytes sustained losses during the evolution of streptophytes. Mapping of these genes on the T1 and T4 topologies revealed that 27 were lost only once in T1 (fig. 6) and that 23 sustained single losses in T4, whereas all remaining genes were lost on two or more occasions. To determine which of these topologies is the most parsimonious in term of predicted gene losses, we analyzed a gene content data set in which we coded the presence/absence of a gene or pseudogene as three-state characters. The predicted scenario of gene losses inferred from T1 was found to be associated with 109 steps (fig. 6), whereas that inferred from T4 involves 19 additional steps. Five genes (*tufA*, *rpl19*, *ycf20*, *trnL*(gag), and *rpl12*) account for these extra steps (fig. 7). The *trnL*(gag), *rpl19*, and *ycf20* genes are each lost once in T1 (two steps/gene) as compared to three times in T4 (six steps/gene), whereas *rpl12* suffers three independent losses in T1 (six steps) and five losses in T4 (10 steps). Note that the highly divergent *tufA* sequence present in the Coleochaetales was considered to be a pseudogene in this analysis. In T1, conversion of the functional *tufA* gene present in *Chara* into a pseudogene (one step) supports the basal position of the Charales relative to the Coleochaetales, whereas the subsequent disappearance of this pseudogene before the emergence of the Zygnematales (one step) supports the latter lineage as being sister to land plants. On the other hand, in T4, *tufA* is associated with two losses (four steps) and is independently converted into a pseudogene (one step).

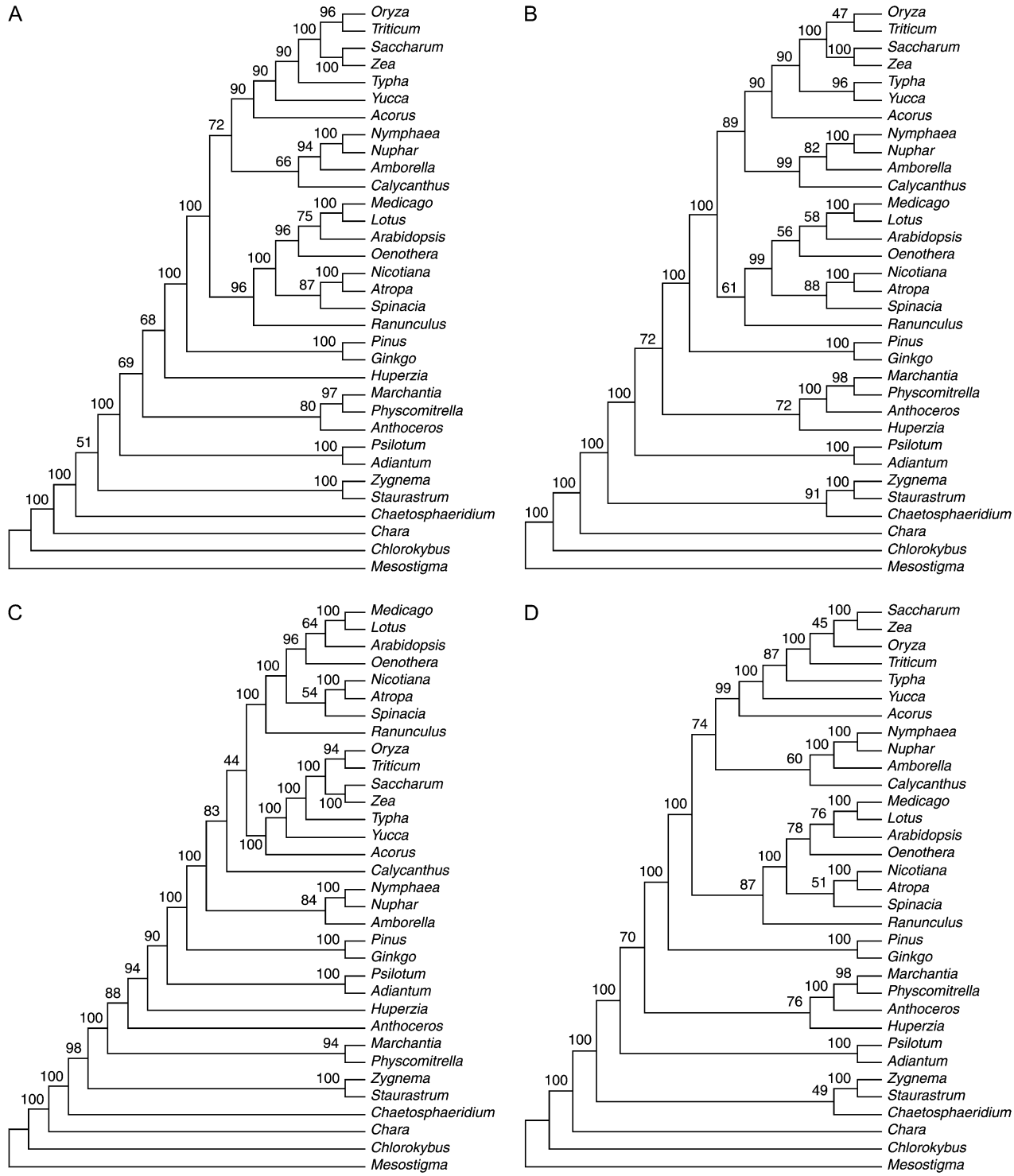

FIG. 4.—Phylogenies inferred from 56 chloroplast protein-coding genes (first two codon positions) and their deduced amino acid sequences. (*A*) Best ML tree inferred from proteins. (*B*) LogDet tree inferred from proteins. (*C*) Best ML tree inferred from genes. (*D*) LogDet tree inferred from genes. Bootstrap values are shown on the corresponding nodes. The following genes/proteins were used for these phylogenetic reconstructions: *atpA*,*B*,*E*,*F*,*H*,*I*, *cemA*, *clpP*, *petA*,*B*,*D*,*G*,*L*,*N*, *psaA*,*B*,*C*,*I*,*J*, *psbA*,*B*,*C*,*D*,*E*,*F*,*H*,*I*,*J*,*K*,*L*,*M*,*N*,*T*,*Z*, *rbcL*, *rpl2*,*14*,*16*,*20*,*33*,*36*, *rpoB*,*C1*,*C2*, *rps2*,*3*,*4*,*7*,*8*,*11*,*12*,*14*,*18*,*19*, *ycf3*,*4*.

Figure 8 shows the scenarios of gains/losses of chloroplast group II introns predicted by T1 and T4. By coding the presence/absence of each group II intron as binary characters, we found that the pattern of intron gains/losses based on T1 is associated with 41 steps, whereas that based on T4 comprises two additional steps. Compared to T4, T1 predicts that group II introns were gained more progressively by the chloroplast genome during the evolution of

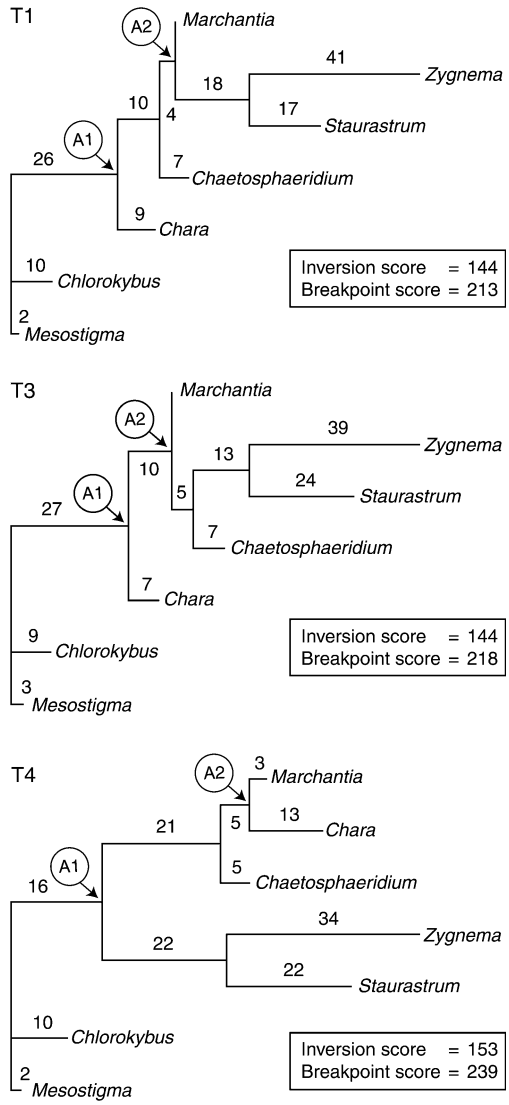

FIG. 5.—Scenarios of genome rearrangements by inversions predicted by the T1, T3, and T4 topologies. Inversion phylogenies were reconstructed using a data set of 113 gene positions in which the two distinct positions of *rps12* were coded as distinct genes. The values shown in the figure indicate the number of inversion medians. A1 and A2 refer to ancestral genomes discussed in the text.

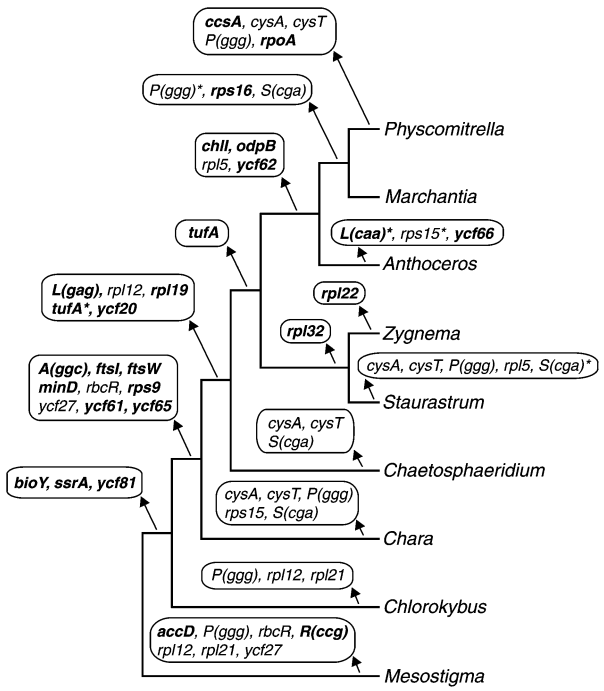

FIG. 6.—Scenario of chloroplast gene losses along the best ML tree inferred from 76 chloroplast proteins. Genes that were lost independently in two or more lineages are indicated in bold characters. Asterisks denote genes that were converted into pseudogenes.

streptophytes and that fewer introns underwent more than one loss event. In each scenario, 15 of the 20 introns common to charophycean green algae and land plants (with five introns specific to each scenario) took their origin before the emergence of the Charales, Coleochaetales, and Zygnematales. T1 also predicts that following their birth, chloroplast group II introns remained generally stable in all streptophyte lineages, with the exception of the zygnematalean lineages where losses of 16 introns were mapped in the common ancestor of *Staurastrum* and *Zygnema* and in the separate lineages leading to these green algae.

Figure 9 presents our phylogenetic analysis of insertions/deletions in protein- and RNA-coding genes. We identified a total of 131 loci carrying unambiguous insertions/deletions in our alignments of proteins and genes. MP analysis of these indels, 52 of which are phylogenetically informative, yielded a majority-rule consensus tree that is congruent with the T3 topology (fig. 9). The basal divergence of the *Chara* lineage relative to those occupied by *Chaetosphaeridium*, the Zygnematales, and the bryophytes is supported by 90% bootstrap value.

Discussion

Phylogenetic Relationships Among Charophycean Green Algae and Their Interrelationships with Land Plants

Although previous studies have provided robust support for the notion that the Charales, Coleochaetales, and Zygnematales diverged after the Klebsormidiales (Karol et al. 2001; Turmel et al. 2002), unequivocal resolution of the branching order of the former three charophycean lineages has been problematic. In the phylogenetic study reported here, which is based on 76 chloroplast genes/proteins from five charophycean green algae and three bryophytes and on 56 chloroplast genes/proteins from a broader taxon sampling including 24 tracheophytes, all methods of analysis yielded a best tree in which *Chara* is basal with respect to *Chaetosphaeridium* and the zygnematalean algae *Staurastrum* and *Zygnema* (figs. 3 and 4). Regardless of the method used, the zygnematalean lineage was sister to land plants (T1) in the best trees inferred from the 76-gene/protein data sets as well as in ML analyses of the 56-gene/protein data sets, whereas a clade consisting of *Chaetosphaeridium* and the two zygnematalean green algae showed a sister relationship with land plants (T3) in LogDet analyses of the latter data sets. In agreement with these results, best trees consistent with the T1 and T3 topologies were recovered in analyses of the chloroplast large and

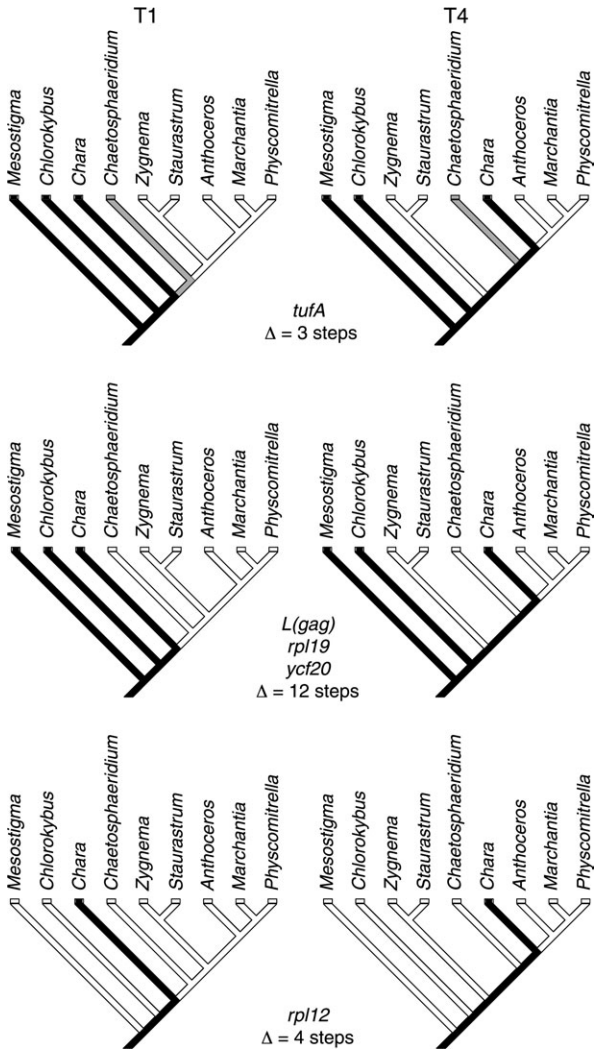

FIG. 7.—Differences between the scenarios of chloroplast gene losses predicted by the best ML tree inferred from 76 chloroplast proteins (T1) and the best ML tree inferred from four genes encoded by three different cellular compartments (T4). Five genes account for the 19 extra steps in the scenario predicted by T4. Different shades denote the presence of a functional gene (black), the presence of a pseudogene (gray), and the absence of a gene (white).

small subunit rRNA genes from 14 charophycean green algae and five bryophytes (Turmel et al. 2002); however, four alternative topologies not significantly different from the best tree, three of which showed the Charales as sister to land plants, were also recovered in these analyses. A phylogenetic study of GAPDH proteins also favors the notion that *Chara* is basal relative to the Coleochaetales and land plants, but no representative of the Zygnematales was included in this analysis (Petersen, Brinkmann, and Cerff 2003).

The multigene chloroplast phylogenies reported here are not congruent with the four-gene phylogeny inferred by Karol et al. (2001). Violations of the models used in these phylogenetic analyses may be the underlying cause of the conflicting results. The processes of sequence evolution assumed by current phylogeny inference programs may deviate substantially from those occurring in nature;

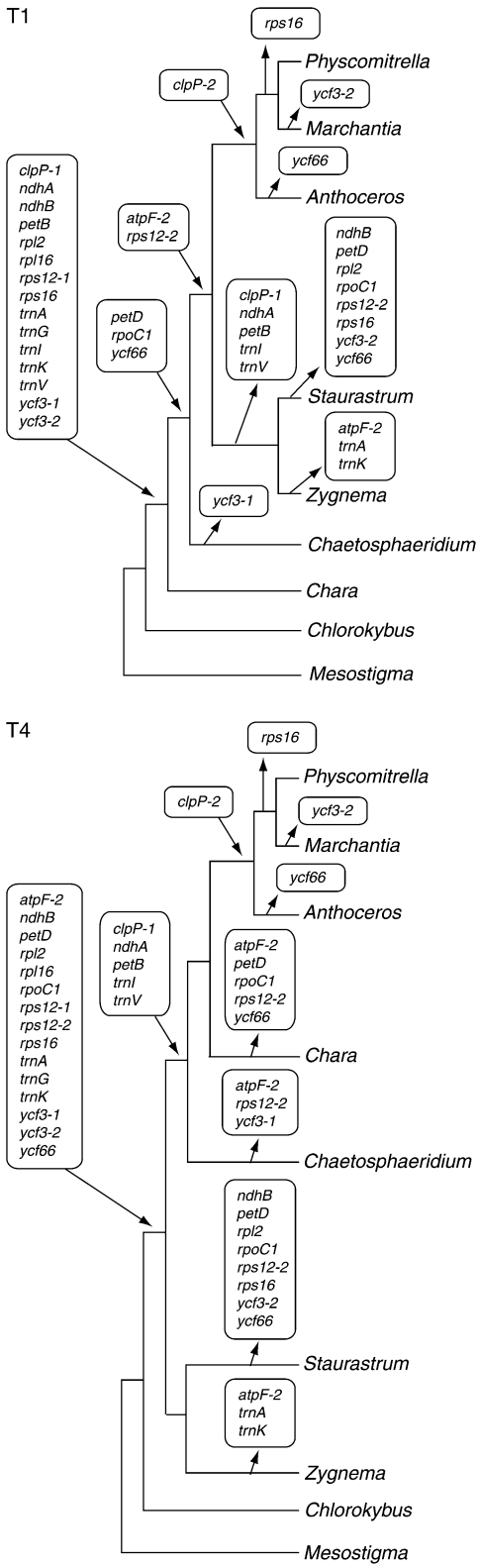

FIG. 8.—Gains/losses of chloroplast group II introns predicted by the best ML tree inferred from 76 chloroplast proteins (T1) and the best ML tree inferred from four genes encoded by three different cellular compartments (T4). Intron gains are denoted by arrows pointing toward the tree, whereas intron losses are denoted by arrows pointing away from the tree. The *Chara atpF* intron was not considered here because it has no homolog at the same insertion site in any other streptophyte cpDNAs.

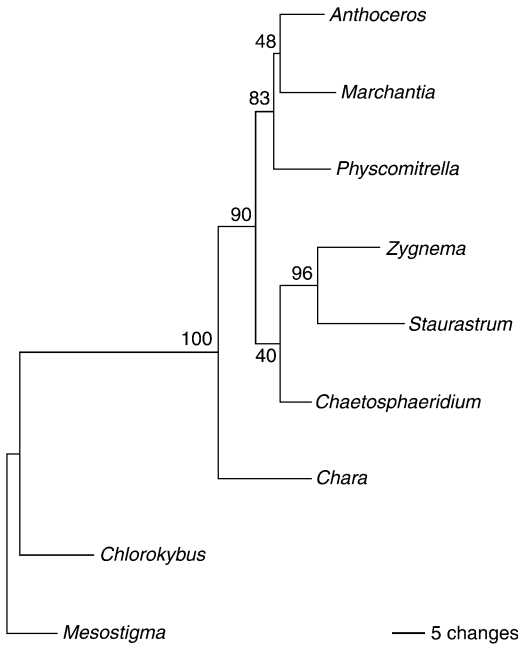

FIG. 9.—MP analysis of 131 indels mapping to coding regions of chloroplast genes. The majority-rule consensus tree is shown, with bootstrap values indicated at the corresponding nodes. Bootstrap support was calculated using 1,000 replicates. Branch lengths are drawn to scale.

these deviations include lineage-specific departures from an assumed common distribution of across-site rate variation (covarion evolution) and lineage-specific departures from an assumed symmetric substitution model (compositional heterogeneity) (Delsuc, Brinkmann, and Philippe 2005; Martin et al. 2005). Considering that only a few chloroplast genome sequences are currently available for charophycean green algae, artifacts in phylogeny reconstructions due to limited taxon sampling of these algae may also explain the incongruence between the multigene chloroplast phylogenies and the four-gene tree. Genome-scale phylogenetic studies with sparse taxon sampling are particularly susceptible to long-branch attraction (Soltis et al. 2004; Leebens-Mack et al. 2005; Philippe, Lartillot, and Brinkmann 2005). We found no evidence, however, that our 76-gene chloroplast phylogeny is artifactual. Methods that are insensitive to differences in base and protein composition (LogDet analyses) had no effect on the topology of the best tree recovered (fig. 3), no significant difference in the rate of nucleotide substitution was identified, and removal of rapidly evolving proteins from the amino acid data set in an attempt to attenuate potential problems of long-branch attraction revealed no change in the best topology. Selection of an outgroup more closely related to the taxa examined is another strategy that has been recommended to attenuate potential long-branch attraction problems (Malek et al. 1996; Philippe, Lartillot, and Brinkmann 2005), but unfortunately a taxon that meets this criterion is not available. Considering that *Mesostigma* may represent a lineage that predates the split of the Streptophyta and Chlorophyta (Lemieux, Otis, and Turmel 2000; Turmel, Otis, and Lemieux 2002*a*; Martin et al. 2005), we tried to use all currently available chlorophyte chloroplast genome sequences as outgroup and found that *Chara* still remains basal relative to *Chaetosphaerium* and the two zygnematalean green algae (data not shown). If we assume that sparse taxon sampling of charophycean green algae is the main cause of the conflict between our multigene chloroplast phylogenies and the four-gene tree of Karol et al. (2001), analysis of the four genes examined by these authors using a reduced taxon sampling comparable to that of the 76-gene analysis would be expected to position *Chara* before the divergence of the Coleochaetales and Zygnematales; however, the sister relationship of *Chara* and land plants was still observed (Supplementary Fig. S1, Supplementary Material online) in this analysis. It will be necessary to expand taxon sampling by including sequence data derived from additional charophycean chloroplast genomes to identify without any ambiguity the divergence order of the Charales, Coleochaetales, and Zygnematales.

In addition to the multigene chloroplast phylogenies reported here, our analyses of three independent sets of structural genomic data (gene order, gene content, and indels) provide robust support for the early emergence of *Chara*. Scenarios of genome rearrangements and gene losses inferred from T1 were found to be more parsimonious than those inferred from the topology of the four-gene tree (T4) (figs. 5–7). Phylogenetic inference from gene order data yielded two best trees with topologies corresponding to T1 and T3. Similarly, the gene content data did not allow us to distinguish unequivocally between the T1 and T3 hypotheses as the gene-loss scenarios inferred from these hypotheses are almost identical. In the gene-loss scenario compatible with T1, the *tufA* pseudogene is lost only once (in the common ancestor of land plants and the Zygnematales), whereas in the scenario compatible with T3, it is lost independently in the lineages leading to the Zygnematales and to the land plants. As observed for the gene content and gene order data, phylogenetic analysis of indels in coding regions favored the notion that *Chara* is basal relative to the Zygnematales and *Chaetosphaeridium* but failed to identify the precise branching orders of the Coleochaetales and Zygnematales relative to the Charales (fig. 9). In contrast, the intron content data provided weak support for the basal position of *Chara* (fig. 8) as the scenarios of intron gains/losses inferred from T1 and T4 are associated with about the same number of steps and invoke multiple intron losses due to intron instability in the Zygnematales.

The basal placement of *Chara* in our multigene chloroplast phylogenies is in apparent conflict with the mitochondrial genomic data currently available for *Chara* (Turmel, Otis, and Lemieux 2003) and *Chaetosphaeridium* (Turmel, Otis, and Lemieux 2002*b*). *Chara* mitochondrial DNA (mtDNA) more closely resembles its land plant counterparts at the levels of gene content, gene order, and intron composition than does *Chaetosphaeridium* mtDNA, and consistent with this structural comparison and the four-gene phylogeny, *Chara* affiliates robustly with land plants in phylogenetic analyses based on 23 mitochondrial genes and proteins (Turmel, Otis, and Lemieux 2003). Obviously, because no members of the Zygnematales were included in these analyses, it will be necessary to examine a representative of this lineage before drawing any firm conclusions

about the branching order of charophycean lineages in the mitochondrial tree. Assuming that the divergence order of the Charales and Coleochaetales in this phylogeny was not affected by sparse taxon sampling, we propose the following explanation for the incongruence between our chloroplast phylogenies and both the mitochondrial and four-gene trees. Given that several cases of horizontal gene transfers have been recently documented for the mitochondria of land plants (Bergthorsson et al. 2003, 2004), we speculate that the difference in topology between our chloroplast and mitochondrial phylogenies might reflect the occurrence of such events in the charalean lineage. Similarly, if the mitochondrial *nad5* gene experienced horizontal transfer, conflicting phylogenetic signals in the concatenated data set of chloroplast and mitochondrial gene sequences analyzed by Karol et al. (2001) might have generated an incorrect tree. Our ML analyses of two subsets of the four genes analyzed by these authors (i.e., *nad5* alone and the *atpB* and *rbcL* gene pairs) indicate that the mitochondrial *nad5* gene is mainly responsible for the high level of support observed for the clade uniting the Charales and land plants (data not shown). We found 99% bootstrap support for this clade when we used the data set containing *nad5* alone but only 24% support with the data set containing the two chloroplast genes. It should be mentioned that, in agreement with the ML analyses of *atpB* and *rbcL* previously conducted by Delwiche et al. (2002) and by Cimino and Delwiche (2002), the analysis of the two-gene data set recovered a best tree whose topology is identical to that of the four-gene tree. In all three analyses, the sister relationship of the Charales and land plants received less than 50% bootstrap support.

Alternatively, the multigene chloroplast phylogenies and both the mitochondrial and four-gene phylogenies can be reconciled by assuming that chloroplast genes from an early-diverging charophycean green alga were transferred horizontally to the chloroplast genome of a charalean alga. This hypothesis, however, seems less plausible than that invoking the horizontal transfer of mitochondrial genes because it implies the almost complete replacement of the recipient genome via transfers of large segments of the donor chloroplast genome. Such massive gene transfers must be postulated to account for the numerous ancestral characters observed at the level of gene order throughout the *Chara* genome. Because no cases of horizontal transfers of chloroplast genes have been reported in green plants, it is hard to envision that the Charales acquired most or all their chloroplast genome from a different charophycean lineage through this process. Perhaps, replacement of the chloroplast genome occurred in the Charales by organellar introgression rather than by horizontal gene transfer per se. Considering that epiphytic and endophytic charophycean algae frequently colonize the Charales in nature (Cimino and Delwiche 2002), one can contemplate the idea that interactions of this sort might have lead to the integration of foreign chloroplasts into the cytoplasm of an ancestral charalean alga and ultimately to the total replacement of host chloroplasts. Note that we have not considered the hypothesis that chloroplast genes from an ancestral land plant were transferred horizontally to the chloroplast genome of a zygnematalean alga because this hypothesis fails to explain the strong support for the early emergence of the Charales.

An accurate phylogeny encompassing streptophyte algae and basal land plants will be essential to understand the suite of molecular events that allowed green plants to adapt and colonize the land. If the position of *Chara* in the multigene chloroplast phylogenies reported here proves to be correct, it will alter significantly our current view of streptophyte evolution. The four-gene phylogeny reported by Karol et al. (2001) supports the notion that the evolution of charophycean green algae was accompanied by a gradual increase in cellular complexity (McCourt, Delwiche, and Karol 2004). Given the little similarity in cellular organization between the Zygnematales, Coleochaetales, and land plants, we were very surprised to find that the zygnematalean lineage or a clade affiliating the Zygnematales and Coleochaetales has a sister relationship with land plants in the best trees supported by our chloroplast genomic data. In placing the charophycean group exhibiting the most complex cellular organization (the Charales) at a basal position relative to the Coleochaetales and Zygnematales, our comparative analyses of chloroplast genomes can be interpreted as indicating that a number of cellular features displayed by land plants were acquired from charalean green algae earlier than anticipated and that some of these features were lost from extant members of the coleochaetalean and zygnematalean lineages. In other words, the common ancestor of the Charales, Coleochaetales, and Zygnematales might have displayed a relatively complex cellular organization that became reduced in more derived green algal lineages. Perhaps, such a reductive evolution became necessary because the ancestral green algae bearing a complex organization were unable to compete with land plants for survival in the same habitat. Note here that the idea that charophycean green algae became secondarily aquatic has been previously proposed (Stebbins and Hill 1980). Alternatively, cellular features currently considered to be shared by the Charales and land plants might have emerged independently in these lineages.

Chloroplast Genome Evolution in the Streptophyta

One of the most important observations that emerged from our comparative analysis of structural cpDNA features is that the chloroplast genome has remained largely unchanged in terms of gene content, gene order, and intron composition during the transition from charophycean green algae to land plants. Our results indicate that during this evolutionary period, the chloroplast genome has lost only four genes, has gained a single intron, and has become resistant to gene rearrangements. As shown by the scenario of genome rearrangements inferred from the T1 topology, *Marchantia* has retained a chloroplast gene order identical to that of its common ancestor with the zygnematalean green algae (ancestor A2 in fig. 5), whereas the IR-lacking cpDNAs of *Staurastrum* and *Zygnema* have diverged considerably from this ancestral gene order ($\geq$35 inversions). Like its *Chaetosphaeridium* counterpart, *Marchantia* cpDNA displays about 15 inversions relative to the genome of the common ancestor of the Charales, Coleochaetales, and Zygnematales (ancestor A1 in fig. 5).

Chara cpDNA has also retained a high degree of ancestral features, this genome being the most similar to the

A1 ancestral genome in terms of primary sequence, gene content, and gene order. The most distinctive features of *Chara* cpDNA relative to other IR-containing streptophyte cpDNAs are its loosely packed genes and expanded introns. These features, accounting for the larger size of *Chara* cpDNA compared to its streptophyte counterparts, are likely to be shared by the chloroplast genomes of other charalean species, as *Nitella translucens* cpDNA has been estimated to be 400 kb in size (Palmer 1991).

Our analysis of intron content data revealed that 15 of the 21 group II introns found in land plant genomes took their origin early during the evolution of streptophyte algae, that is, during the evolutionary interval separating the Chlorokybales and Charales. Considering that the Coleochaetales and Charales are each represented by a single member, that occasional loss of introns may have occurred in these lineages, and that chloroplast introns sharing identical insertion sites may have been lost independently in the *Staurastrum* and *Zygnema* lineages, this predicted number of introns in the common ancestor of these algae must be considered as a minimal estimate. Because of their significant similarity at the sequence level, introns inserted at identical insertion sites in charophycean and land plant chloroplast genomes most likely share a common vertical ancestry. As none of these introns has been identified in the same gene context outside of the Streptophyta, it appears that intragenomic proliferation mainly accounts for the numerous group II introns in streptophyte cpDNAs. These introns have remained generally stable in all streptophyte lineages examined, except in the zygnematalean green algae *Staurastrum* and *Zygnema*.

Supplementary Material

All data sets used in phylogenetic analyses and Supplementary Tables S1 and S2 and Fig. S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/). The *Chara* and *Chlorokybus* chloroplast genome sequences have been deposited in the GenBank database under the accession numbers DQ229107 and DQ422812, respectively.

Acknowledgments

We thank Marc-André Bureau and Jean-François Rochette for their help in determining the *Chara* cpDNA sequence and Patrick Charlebois and Jules Gagnon for their assistance with the bioinformatic analyses. This work was supported by the Natural Sciences and Engineering Research Council of Canada (to C.L. and M.T.).

Literature Cited

Adachi, J., P. J. Waddell, W. Martin, and M. Hasegawa. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. **50**:348–358.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. **215**: 403–410.

Baldauf, S. L., J. R. Manhart, and J. D. Palmer. 1990. Different fates of the chloroplast *tufA* gene following its transfer to the nucleus in green algae. Proc. Natl. Acad. Sci. USA **87**:5317–5321.

Bergthorsson, U., K. L. Adams, B. Thomason, and J. D. Palmer. 2003. Widespread horizontal transfer of mitochondrial genes in flowering plants. Nature **424**:197–201.

Bergthorsson, U., A. O. Richardson, G. J. Young, L. R. Goertzen, and J. D. Palmer. 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. Proc. Natl. Acad. Sci. USA **101**: 17747–17752.

Bhattacharya, D., K. Weber, S. S. An, and W. Berning-Koch. 1998. Actin phylogeny identifies *Mesostigma viride* as a flagellate ancestor of the land plants. J. Mol. Evol. **47**:544–550.

Bremer, K., C. J. Humphries, B. D. Mishler, and S. P. Churchill. 1987. On cladistic relationships in green plants. Taxon **36**:339–349.

Caprara, A. 2003. The reversal median problem. INFORMS J. Comput. **15**:93–113.

Castresana, J. 2000. Selection of conserved bloks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**:540–542.

Chapman, R. L., and D. A. Waters. 2002. Green algae and land plants—an answer at last? J. Phycol. **38**:237–240.

Choudhuri, J. V., C. Schleiermacher, S. Kurtz, and R. Giegerich. 2004. GenAlyzer: interactive visualization of sequence similarities between entire genomes. Bioinformatics **20**:1964–1965.

Cimino, M. T., and C. F. Delwiche. 2002. Molecular and morphological data identify a cryptic species complex in endophytic members of the genus *Coleochaete* Bréb. (Charophyta: Coleochaetaceae). J. Phycol. **38**:1213–1221.

Côté, V., J.-P. Mercier, C. Lemieux, and M. Turmel. 1993. The single group-I intron in the chloroplast *rrnL* gene of *Chlamydomonas humicola* encodes a site-specific endonuclease (*I-ChuI*). Gene **129**:69–76.

Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. **6**:361–375.

Delwiche, C. F., K. G. Karol, M. T. Cimino, and K. J. Sytsma. 2002. Phylogeny of the genus *Coleochaete* (Coleochaetales, Charophyta) and related taxa inferred by analysis of the chloroplast gene *rbcL*. J. Phycol. **38**:394–403.

Farris, J. S. 1977. Phylogenetic analysis under Dollo's Law. Syst. Zool. **26**:77–88.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27**: 401–410.

———. 1995. PHYLIP (phylogeny inference package). Version 3.5. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Goremykin, V. V., and F. H. Hellwig. 2005. Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. Plant Syst. Evol. **254**:93–103.

Goremykin, V. V., B. Holland, K. I. Hirsch-Ernst, and F. H. Hellwig. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol. **22**: 1813–1822.

Graham, L. E., M. E. Cook, and J. S. Busse. 2000. The origin of plants: body plan changes contributing to a major evolutionary radiation. Proc. Natl. Acad. Sci. USA **97**:4535–4540.

Gu, X., and W. H. Li. 1996. Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. Mol. Biol. Evol. **13**: 1375–1383.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

Karol, K. G., R. M. McCourt, M. T. Cimino, and C. F. Delwiche. 2001. The closest living relatives of land plants. Science **294**: 2351–2353.

Kenrick, P., and P. R. Crane. 1997. The origin and early evolution of plants on land. Nature **389**:33–39.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980–984.

Kugita, M., A. Kaneko, Y. Yamamoto, Y. Takeya, T. Matsumoto, and K. Yoshinaga. 2003. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. Nucleic Acids Res. **31**:716–721.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.

Leebens-Mack, J., L. A. Raubeson, L. Cui, J. V. Kuehl, M. H. Fourcade, T. W. Chumley, J. L. Boore, R. K. Jansen, and C. W. Depamphilis. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol. Biol. Evol. **22**:1948–1963.

Lemieux, C., C. Otis, and M. Turmel. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature **403**:649–652.

Lewis, L. A., and R. M. McCourt. 2004. Green algae and the origin of land plants. Am. J. Bot. **91**:1535–1556.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. **11**:605–612.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. **19**:1–7.

Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25**:955–964.

Maddison, D., and W. Maddison. 2000. MacClade 4: analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Mass.

Malek, O., K. Lattig, R. Hiesel, A. Brennicke, and V. Knoop. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. EMBO J. **15**:1403–1411.

Manhart, J. R., R. W. Hoshaw, and J. D. Palmer. 1990. Unique chloroplast genome in *Spirogyra maxima* (Chlorophyta) revealed by physical and gene mapping. J. Phycol. **26**:490–494.

Marin, B., and M. Melkonian. 1999. Mesostigmatophyceae, a new class of streptophyte green algae revealed by SSU rRNA sequence comparisons. Protist **150**:399–417.

Martin, W., O. Deusch, N. Stawski, N. Grunheit, and V. Goremykin. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. **10**:203–209.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc. Natl. Acad. Sci. USA **99**:12246–12251.

Mattox, K. R., and K. D. Stewart. 1984. Classification of the green algae: a concept based on comparative cytology. Pp. 29–72 *in* D. E. G. Irvine, and D. M. John, eds. The systematics of the green algae. Academic Press, London.

McCourt, R. M., C. F. Delwiche, and K. G. Karol. 2004. Charophyte algae and land plant origins. Trends Ecol. Evol. **19**:661–666.

Michel, F., K. Umesono, and H. Ozeki. 1989. Comparative and functional anatomy of group II catalytic introns—a review. Gene **82**:5–30.

Michel, F., and E. Westhof. 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J. Mol. Biol. **216**:585–610.

Moret, B. M. E., A. C. Siepel, J. Tang, and T. Liu. 2002. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. Pp. 521–536. Proceedings of the Second International Workshop on Algorithms in Bioinformatics (WABI 2002), Lecture Notes in Computer Science. Springer-Verlag, New York.

Nishiyama, T., P. G. Wolf, M. Kugita et al. (12 co-authors). 2004. Chloroplast phylogeny indicates that bryophytes are monophyletic. Mol. Biol. Evol. **21**:1813–1819.

Ohyama, K., H. Fukuzawa, T. Kohchi et al. (13 co-authors). 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature **322**:572–574.

Palmer, J. D. 1991. Plastid chromosomes: structure and evolution. Pp. 5–53 *in* L. Bogorad, and K. Vasil, eds. The molecular biology of plastids. Academic Press, San Diego, Calif.

Petersen, J., H. Brinkmann, and R. Cerff. 2003. Origin, evolution, and metabolic role of a novel glycolytic GAPDH enzyme recruited by land plant plastids. J. Mol. Evol. **57**:16–26.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. **22**:1246–1253.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817–818.

Qiu, Y. L., and J. D. Palmer. 1999. Phylogeny of early land plants: insights from genes and genomes. Trends Plant Sci. **4**:26–30.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16**:276–277.

Robinson-Rechavi, M., and D. Huchon. 2000. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics **16**:296–297.

Sanderson, M. J., J. L. Thorne, N. Wikstrom, and K. Bremer. 2004. Molecular evidence on plant divergence times. Am. J. Bot. **91**:1656–1665.

Shaw, J., and K. Renzaglia. 2004. Phylogeny and diversification of bryophytes. Am. J. Bot. **91**:1557–1581.

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

Soltis, D. E., V. A. Albert, V. Savolainen et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. **9**:477–483.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. **22**:1161–1164.

Stebbins, G. L., and G. J. C. Hill. 1980. Did multicellular plants invade the land? Am. Nat. **115**:342–353.

Steel, M. 1994. Recovering a tree from the leaf colorations it generates under a Markov model. Appl. Math. Lett. **7**:19–23.

Steel, M., D. Huson, and P. J. Lockhart. 2000. Invariable sites models and their use in phylogeny reconstruction. Syst. Biol. **49**:225–232.

Steel, M. A., P. J. Lockhart, and D. Penny. 1993. Confidence in evolutionary trees from biological sequence data. Nature **364**:440–442.

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

Sugiura, C., Y. Kobayashi, S. Aoki, C. Sugita, and M. Sugita. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA*

from the chloroplast to the nucleus. Nucleic Acids Res. **31**:5324–5331.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0 for MacIntosh. Sinauer Associates, Sunderland, Mass.

Tesler, G. 2002. GRIMM: genome rearrangements web server. Bioinformatics **18**:492–493.

Thollesson, M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics **20**:416–418.

Turmel, M., M. Ehara, C. Otis, and C. Lemieux. 2002. Phylogenetic relationships among streptophytes as inferred from chloroplast small and large subunit rRNA gene sequences. J. Phycol. **38**:364–375.

Turmel, M., R. R. Gutell, J. P. Mercier, C. Otis, and C. Lemieux. 1993. Analysis of the chloroplast large subunit ribosomal RNA gene from 17 *Chlamydomonas* taxa. Three internal transcribed spacers and 12 group I intron insertion sites. J. Mol. Biol. **232**:446–467.

Turmel, M., C. Lemieux, G. Burger, B. F. Lang, C. Otis, I. Plante, and M. W. Gray. 1999. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. Plant Cell **11**:1717–1729.

Turmel, M., J.-P. Mercier, and M.-J. Côté. 1993. Group I introns interrupt the chloroplast *psaB* and *psbC* and the mitochondrial *rrnL* gene in *Chlamydomonas*. Nucleic Acids Res. **21**:5242–5250.

Turmel, M., C. Otis, and C. Lemieux. 2002*a*. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol. Biol. Evol. **19**:24–38.

———. 2002*b*. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc. Natl. Acad. Sci. USA **99**:11275–11280.

———. 2003. The mitochondrial genome of *Chara vulgaris*: insights into the mitochondrial DNA architecture of the last common ancestor of green algae and land plants. Plant Cell **15**:1888–1903.

———. 2005. The complete chloroplast DNA sequences of the charophycean green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biol. **3**:22.

Wolf, P. G., K. G. Karol, D. F. Mandoli, J. Kuehl, K. Arumuganathan, M. W. Ellis, B. D. Mishler, D. G. Kelch, R. G. Olmstead, and J. L. Boore. 2005. The first complete chloroplast genome sequence of a lycophyte, *Huperzia lucidula* (Lycopodiaceae). Gene **350**:117–128.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555–556.

Charles Delwiche, Associate Editor

Accepted April 10, 2006