The Chloroplast Genome Sequence of the Green Alga *Pseudendoclonium akinetum* (Ulvophyceae) Reveals Unusual Structural Features and New Insights into the Branching Order of Chlorophyte Lineages

Jean-François Pombert, Christian Otis, Claude Lemieux, and Monique Turmel

Département de biochimie et de microbiologie, Université Laval, Québec G1K 7P4, Canada

One major lineage of green plants, the Chlorophyta, is represented by the green algal classes Prasinophyceae, Ulvophyceae, Trebouxiophyceae, and Chlorophyceae. The Prasinophyceae occupies the most basal position in the Chlorophyta, but the branching order of the Ulvophyceae, Trebouxiophyceae, and Chlorophyceae remains unresolved. The chloroplast genome sequences currently available for representatives of three chlorophyte classes have revealed that this genome is highly plastic, with *Chlamydomonas* (Chlorophyceae) and *Chlorella* (Trebouxiophyceae) showing fewer ancestral features than *Nephroselmis* (Prasinophyceae). We report the 195,867-bp chloroplast DNA (cpDNA) sequence of *Pseudendoclonium akinetum* (Ulvophyceae), a member of the class that has not been previously examined for detailed cpDNA analysis. This genome shares common evolutionary trends with its *Chlorella* and *Chlamydomonas* homologs. The gene content, number of ancestral gene clusters, and abundance of short dispersed repeats in *Pseudendoclonium* cpDNA are intermediate between those observed for *Chlorella* and *Chlamydomonas* cpDNAs. Although *Pseudendoclonium* cpDNA features a large inverted repeat, its quadripartite structure is unusual in displaying an rRNA operon transcribed toward the large single-copy (LSC) region and a small single-copy region containing 14 genes that are normally found in the LSC region. Twenty-seven group I introns lie in nine genes and fall within four subgroups (IA1, IA2, IA3, and IB); 19 encode putative homing endonucleases, and 7 have homologs at identical insertion sites in other chlorophyte or streptophyte organelle genomes. The high similarity observed among the 14 IA1 and 7 IA2 introns and their encoded endonucleases suggests that many introns arose from intragenomic proliferation of a few founding introns in the lineage leading to *Pseudendoclonium*. Interestingly, one intron (in *atpA*) and some of the dispersed repeats also reside in *Pseudendoclonium* mitochondria, providing strong evidence for interorganellar lateral transfer of these genetic elements. Phylogenetic analyses of 58 cpDNA-encoded proteins and genes support the hypothesis that the Ulvophyceae is sister to the Trebouxiophyceae but cannot eliminate the hypothesis that the Ulvophyceae is sister to the Chlorophyceae. We favor the latter hypothesis because it is strongly supported by phylogenetic analyses of gene order data and by independent structural evidence based on shared gene losses and rearrangement break points within ancestrally conserved gene clusters.

Introduction

The green algae are divided into the phyla Streptophyta and Chlorophyta. The Streptophyta (Bremer 1985) contains all land plants and the green algae belonging to the class Charophyceae (Graham, Cook, and Busse 2000), whereas the Chlorophyta (Sluiman 1985) contains virtually all of the other green algae, i.e., the members of the classes Prasinophyceae, Ulvophyceae, Trebouxiophyceae, and Chlorophyceae (Lewis and McCourt 2004). The basal position of the Prasinophyceae in the Chlorophyta is generally well established, but the branching order of the Ulvophyceae, Trebouxiophyceae, and Chlorophyceae (UTC) remains unresolved (Friedl and O'Kelly 2002; Pombert et al. 2004). A third lineage at the base of the Streptophyta and Chlorophyta is possibly represented by *Mesostigma viride*; however, some studies suggest that this green alga traditionally classified within the Prasinophyceae represents a basal divergence within the Streptophyta (Bhattacharya et al. 1998; Marin and Melkonian 1999; Karol et al. 2001).

To improve our understanding of phylogenetic relationships among green algae and also to better understand how the chloroplast genome has evolved in this algal group, we have undertaken the complete sequencing of chloroplast DNA (cpDNA) from representatives of various lineages in the Streptophyta and Chlorophyta. Five green algal chloroplast genome sequences are currently available in public databases. The genomes of the prasinophyte *Mesostigma* (Lemieux, Otis, and Turmel 2000) and of the charophyte *Chaetosphaeridium globosum* (Turmel, Otis, and Lemieux 2002) most closely resemble their land plant counterparts in terms of overall structure and gene organization. Like most land plant cpDNAs, they feature a quadripartite structure that is characterized by the presence of two copies of an rRNA-encoding inverted repeat (IR) sequence separating a small single-copy (SSC) and a large single-copy (LSC) region. In the Streptophyta, each of these genomic regions shows a highly conserved gene content, and the rRNA operon within the IR is transcribed toward the SSC region.

In contrast, the complete cpDNA sequences of the prasinophyte *Nephroselmis olivacea* (Turmel, Otis, and Lemieux 1999), the trebouxiophyte *Chlorella vulgaris* (Wakasugi et al. 1997), and the chlorophycean alga *Chlamydomonas reinhardtii* (Maul et al. 2002) indicate that the architecture of the chloroplast genome is very fluid in the Chlorophyta. *Nephroselmis* cpDNA displays ancestral gene clusters and the typical quadripartite structure observed in the streptophytes, whereas *Chlorella* cpDNA lacks the IR as well as several genes and has gained three group I introns. *Chlamydomonas* cpDNA is even more scrambled in its structure and gene organization. Although it features an IR, the two single-copy regions are about equal in size and each includes genes usually present in the SSC and LSC regions. *Chlamydomonas* cpDNA carries fewer genes but more introns than its *Chlorella* homolog; moreover, it harbors fragmented ancestral operons, and the coding regions of some genes are greatly expanded or broken relative

Key words: green algae, Ulvophyceae, *Pseudendoclonium akinetum*, chloroplast genome evolution, group I introns, repeated sequences.

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

Mol. Biol. Evol. 22(9):1903–1918. 2005
doi:10.1093/molbev/msi182
Advance Access publication June 1, 2005

© The Author 2005. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please e-mail: journals.permissions@oupjournals.org

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

to the corresponding genes in other completely sequenced cpDNAs. Many repeated sequence elements (Maul et al. 2002) are dispersed throughout both the *Chlamydomonas* and *Chlorella* genomes; however, such repeats have not been identified in *Nephroselmis*, *Mesostigma*, and *Chaetosphaeridium* cpDNAs.

Here we report the chloroplast genome sequence of *Pseudendoclonium akinetum*, a unicellular green alga thought to belong to a deep-branching lineage within the Ulvophyceae (Floyd and O'Kelly 1990). This genome has a unique architecture, although it shares some features with *Chlorella* cpDNA and others with *Chlamydomonas* cpDNA. Based on our phylogenetic inferences from chloroplast sequences and gene order data as well as on independent structural evidence, we strongly favor the hypothesis that the Ulvophyceae and Chlorophyceae are sister groups.

Materials and Methods

DNA Isolation and Sequencing

Pseudendoclonium akinetum (Tupa 1974) was obtained from the University of Texas Algal Culture Collection (UTEX 1912) and grown in modified Volvox medium (McCracken, Nadakavukaren, and Cain 1980) under 12 h light-dark cycles. An A+T–rich fraction containing cpDNA and mitochondrial DNA (mtDNA) was isolated and sequenced as described previously (Pombert et al. 2004). The nuclear-encoded 18S rRNA gene was amplified by polymerase chain reaction (PCR) from total cellular DNA using primers NS1 (White et al. 1990) and 18L (Hamby et al. 1988) as described in Helms et al. (2001). Sequences were edited and assembled with SEQUENCHER 4.2.1 (Gene Codes, Ann Arbor, Mich.).

Sequence Analyses

Genes were identified by Blast homology searches (Altschul et al. 1990) against the nonredundant database of the National Center for Biotechnology Information (NCBI) server (http://www.ncbi.nlm.nih.gov/BLAST/). Positions of open reading frames (ORFs) and protein-coding genes were determined using ORFFINDER at NCBI and various programs of the GCG Wisconsin package (version 10.2) (Accelrys, Burlington, Mass.), whereas genes coding for tRNAs were localized with tRNAscan-SE 1.23 (Lowe and Eddy 1997). Patterns of codon usage for protein-coding genes and ORFs were compared using the CORRESPOND and CODONPREFERENCE programs of the GCG package and the CAI program of the EMBOSS package (version 2.6.0) (http://emboss.sourceforge.net). Repeated sequences were identified with PipMaker (Schwartz et al. 2000) and REPuter 2.74 (Kurtz et al. 2001). Repeats were sorted with REPEATFINDER (Volfovsky, Haas, and Salzberg 2001), and the retrieved classification was refined manually. Putative stem-loop structures and degenerated repeats were identified with PALINDROME and ETANDEM, respectively; these two programs are part of the EMBOSS package. Homologous introns were identified by BlastN searches (Altschul et al. 1990) against the nonredundant database of NCBI using an *E* value threshold of 1×10^{-6}. Homologous introns inserted at identical positions within the same gene were identified by manual screening of the GOBASE database (O'Brien et al. 2003).

Phylogenetic Analyses

The 18S rDNA sequence alignment of Friedl and O'Kelly (2002) was retrieved from TreeBASE (http://www.treebase.org/treebase/), modified to include the *P. akinetum* sequence, and analyzed as described by these authors. To obtain the chloroplast amino acid data set, sequences of individual chloroplast proteins were aligned using T-COFFEE 1.37 (Notredame, Higgins, and Heringa 2000), the ambiguously aligned regions of these alignments were removed with GBLOCKS 0.91b (Castresana 2000), and the filtered sequences were concatenated. The nucleotide data set was generated by aligning the sequences of individual protein-coding genes on the basis of corresponding protein alignments, filtering the alignments with GBLOCKS 0.91b, concatenating the filtered sequences, and subsequently removing third codon positions with PAUP* 4.0b10 (Swofford 2002). Maximum likelihood (ML) trees inferred from amino acid sequences were computed with PHYML 2.4.4 (Guindon and Gascuel 2003) under the cpREV45 + Γ model (Adachi et al. 2000), whereas ML trees inferred from nucleotide sequences were computed with PAUP* 4.0b10 (Swofford 2002) under the general time reversible + Γ + I model. Modeltest 3.6 (Posada and Crandall 1998) identified the latter model as the one best fitting our nucleotide data. Bootstrap support for each node was calculated using 100 replicates. The confidence limits of alternative tree topologies were evaluated using the Shimodaira-Hasegawa test as implemented in CODEML 3.14 (Yang 1997) and PAUP* 4.0b10 (Swofford 2002). Support for the T1, T2, and T3 topologies by individual proteins in the amino acid data set was estimated with CODEML 3.14 using the MGENE = 1 option.

Intron sequences were aligned manually on the basis of secondary structure predictions, and regions judged to be ambiguously aligned were removed. Neighbor-Joining analyses of the resulting data sets were carried out with PAUP* 4.0b10 (Swofford 2002) using Hasegawa-Kishino-Yano 85 distances with 1,000 bootstrap replicates.

Phylogenetic reconstructions from gene order data were performed with GRAPPA 2.0 (Moret et al. 2001) using the default parameters and the −m (tighter circular lower bound) option. Because this program requires that all genes analyzed be shared between the taxa examined, the genes present within copy B of the IR were excluded from the data set to accommodate the lack of the IR in *Chlorella* cpDNA.

Results

Phylogenetic Affiliation of *P. akinetum*

Because it has not been previously demonstrated that *P. akinetum* (UTEX 1912) belongs to the Ulvophyceae, we have ascertained this affiliation by conducting phylogenetic analyses of nuclear-encoded 18S rDNA sequences with ML, maximum parsimony, and distance methods (Supplementary Fig. S1, Supplementary Material online). *Pseudendoclonium akinetum* was found to clearly affiliate

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

Table 1
Compared Features of *Pseudendoclonium* and Other Green Algal cpDNAs

Feature	*Mesostigma*	*Nephroselmis*	*Chlorella*	*Pseudendoclonium*	*Chlamydomonas*
Size (bp)					
Total	118,360	200,799	150,613	195,867	203,827
IR	6,057	46,137	—[a]	6,039	22,211
LSC	83,627	92,126	—[a]	140,914	81,307
SSC	22,619	16,399	—[a]	42,875	78,088
A+T(%)	69.9	57.9	68.4	68.5	65.5
Coding sequences (%)[b]	74.5	68.7	60.9	62.3	50.1
Genes (no.)[c]	137	128	112	105	94
Introns (no.)					
Group I	0	0	3	27	5
Group II	0	0	0	0	2

[a] *Chlorella* cpDNA has no IR.
[b] Conserved genes, unique ORFs, introns, and intron ORFs were considered as coding sequences.
[c] Genes present in the IR were counted only once. Unique ORFs and intron ORFs were not taken into account.

with other members of the Ulvophyceae, being part of a clade including *Pseudendoclonium basiliense* and five other members of the Ulotrichales. The *P. akinetum* and *P. basiliense* 18S rDNA sequences differ at only seven positions in the data set analyzed.

Genome Structure and Gene Partitioning

Pseudendoclonium cpDNA (195,867 bp) contains two copies of an IR sequence that are separated from one another by LSC and SSC regions (table 1 and fig. 1). The *Pseudendoclonium* IR encodes only the rRNA operon, in contrast to the IRs of all previously sequenced green plant cpDNAs that include additional genes. Surprisingly, the *Pseudendoclonium* rRNA operon is transcribed toward the LSC region rather than toward the SSC region. Another unusual feature of *Pseudendoclonium* cpDNA concerns the pattern of gene partitioning. The SSC region of this genome features 14 genes that map to the LSC region in *Mesostigma*, *Nephroselmis*, and land plant cpDNAs (fig. 1). However, the *Pseudendoclonium* LSC region contains no genes that are usually found in the SSC region.

Gene Content and Gene Density

Table 2 compares the gene content of *Pseudendoclonium* cpDNA with those of other green algal cpDNAs. It can be seen that a large number of genes found in both *Mesostigma* and *Nephroselmis* (including all *ndh* genes) have been lost before the emergence of the UTC lineages. With its 105 genes, the chloroplast gene repertoire of *Pseudendoclonium* is intermediate in size between those of *Chlorella* and *Chlamydomonas* (see table 1). Eight of the genes identified in *Chlorella* (*chlB*, *chlL*, *chlN*, *cysA*, *cysT*, *trnL*(gag), *trnS*(gga), and *trnT*(ggu)) are absent from *Pseudendoclonium*, whereas all of the genes found in *Pseudendoclonium* except *trnR*(ccu) are present in *Chlorella*. Fourteen of the genes identified in *Pseudendoclonium* (*accD*, *chlI*, *infA*, *minD*, *psaI*, *psaM*, *rpl12*, *rpl19*, *rpl32*, *ycf20*, *ycf62*, *trnL*(caa), *trnR*(ccg), and *trnR*(ccu)) are absent from *Chlamydomonas*, whereas all of the genes found in *Chlamydomonas* except *chlB*, *chlL*, and *chlN* are present in *Pseudendoclonium*.

A total of 19 of the 27 introns displayed by *Pseudendoclonium* cpDNA carry an internal ORF encoding a putative homing endonuclease (see table 4). In addition to these intron ORFs, 12 free-standing ORFs larger than 100 codons are present in *Pseudendoclonium* cpDNA; only 2 (*orf286* and *orf521*) show a codon usage not significantly different from that observed for the protein-coding genes.

Genes are less tightly packed in *Pseudendoclonium* cpDNA than in *Mesostigma* and *Nephroselmis* cpDNAs (table 1). The intergenic spacers in this ulvophyte genome are up to 3,215 bp in size, with an average of 600 bp. The higher proportion of coding sequences found in *Pseudendoclonium* cpDNA relative to *Chlorella* cpDNA is explained by the greater number of introns in the former genome (table 1).

Expansion of Coding Regions

Eight protein-coding genes in *Pseudendoclonium* cpDNA are nearly two or three times larger than their *Mesostigma* homologs (table 3). For all of these genes, with the exception of *cemA*, expansion of coding regions appears to be the main cause of their increased sizes. Because the sizes of the corresponding *Mesostigma* genes match those of their counterparts in streptophytes and the cyanobacterium *Synechocystis* sp. PCC 6803, we eliminated the alternative hypothesis that shrinkage of the *Mesostigma* coding regions is responsible for the observed size variations. The *ftsH*, *rpoB*, *rpoC1*, and *rpoC2* genes represent the most notable cases of gene expansion; the levels of expansion noted for these genes are comparable or greater than those for the corresponding genes in *Chlamydomonas* (table 3). The extra sequences accounting for the expansion of the *Pseudendoclonium ftsH* and *rpo* genes map mainly to internal coding regions. In *Chlamydomonas*, *rpoB, rpoC1*, as well as *rps2* each consist of two separate ORFs, and even though these ORFs are adjacent, no intron has been detected in the noncoding sequence separating them. In *Pseudendoclonium*, each of these three genes occurs as a single ORF.

Introns

The 27 introns in *Pseudendoclonium* cpDNA account for 14.8% of the genome size and are preferentially located in genes coding for photosynthetic proteins (table 4). They all belong to the group I family and fall within subgroups IA1, IA2, IA3, and IB according to the classification system

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

FIG. 1.—Gene map of *Pseudendoclonium* cpDNA. Genes outside the map are transcribed clockwise. The transcription direction of the rRNA operons is indicated by arrows. The genes shown in yellow, blue, and red map to the IR, LSC, and SSC regions in *Mesostigma* cpDNA, respectively. Genes and ORFs absent from *Mesostigma* cpDNA are shown in gray. Gene clusters shared specifically with *Chlorella* are denoted by thick brackets, whereas the clusters shared with *Chlorella* and one or more additional chlorophytes are represented by thin brackets. Only the free-standing ORFs larger than 100 codons are shown. tRNA genes are indicated by the one-letter amino acid code followed by the anticodon in parentheses (Me, elongator methionine; Mf, initiator methionine).

of Michel and Westhof (1990). The subgroup IA1 is well represented by the *Pseudendoclonium* introns, with 14 introns assigned to this category.

Structural and phylogenetic analyses of the IA1 and IA2 introns revealed close relationships among members of the same subgroup. These introns are well conserved in both primary sequences and secondary structures (fig. 2*A*, *B*, *C*, and *D*); moreover, those sharing similar ORFs tend to cluster together in phylogenetic trees inferred from core sequences (compare fig. 2*E* with fig. 2*C* and *D*). The seven IA1 intron-encoded H-N-H endonucleases are all specified by an ORF in L5 (table 4) and can be assigned to three categories based on sequence similarity within the region containing the H-N-H motif (fig. 2*E*). Pa.*psbA*.1 differs from the other IA1 introns in carrying an ORF in L9 that potentially codes for an endonuclease of the LAGLIDADG family (fig. 2*E*). Five IA2 introns exhibit an ORF in L6 that codes for a putative GIY-YIG endonuclease; these IA2 intron-encoded proteins can be divided into two categories based on sequence similarity (fig. 2*D* and *E*).

The IA3 and IB introns show more sequence divergence than their IA1 and IA2 counterparts. As shown in Supplementary Figure S2 (Supplementary Material online) and table 4, the two IA3 introns (Pa.*psbA*.3 and Pa.*rrl*.1) differ greatly with respect to their core structures and encoded endonucleases. The P3.1 and P3.2 pairings are present in Pa.*psbA*.3 but not in Pa.*rrl*.1, and Pa.*psbA*.3 encodes a putative GIY-YIG endonuclease, whereas Pa.*rrl*.1 encodes a putative endonuclease with a single LAGLIDADG motif (table 4). Each of the four IB introns contains an ORF in L8 that codes

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

Table 2
Gene Content in *Pseudendoclonium* and Other Green Algal cpDNAs

Gene[a]	*Mesostigma*	*Nephroselmis*	*Chlorella*	*Pseudendoclonium*	*Chlamydomonas*
accD	○	●	●	●	○
bioY	●	○	○	○	○
chlB	●	●	●	○	●
chlI	●	●	●	●	○
chlL	●	●	●	○	●
chlN	●	●	●	○	●
cysA	●	●	●	○	○
cysT	●	●	●	○	○
ftsI	●	●	○	○	○
ftsW	●	●	○	○	○
infA	●	●	●	●	○
minD	●	●	●	●	○
ndhA	●	●	○	○	○
ndhB	●	●	○	○	○
ndhC	●	●	○	○	○
ndhD	●	●	○	○	○
ndhE	●	●	○	○	○
ndhF	●	●	○	○	○
ndhG	●	●	○	○	○
ndhH	●	●	○	○	○
ndhI	●	●	○	○	○
ndhJ	●	○	○	○	○
ndhK	●	●	○	○	○
odpB	●	○	○	○	○
petN	●	●	○	○	○
psaI	●	●	●	●	○
psaM	●	○	●	●	○
rne	○	●	○	○	○
rnpB	○	●	○	○	○
rpl12	○	●	●	●	○
rpl19	●	●	●	●	○
rpl22	●	○	○	○	○
rpl32	●	●	●	●	○
rpl33	●	○	○	○	○
rps15	●	○	○	○	○
rps16	●	○	○	○	○
ssrA	●	○	○	○	○
ycf20	●	○	●	●	○
ycf47	○	●	○	○	○
ycf61	●	○	○	○	○
ycf62	●	●	●	●	○
ycf65	●	○	○	○	○
ycf66	●	○	○	○	○
ycf81	●	●	○	○	○
trnA(ggc)	●	○	○	○	○
trnL(caa)	●	●	●	●	○
trnL(gag)	●	●	●	○	○
trnR(ccg)	○	●	●	●	○
trnR(ccu)	○	○	○	●	○
trnS(cga)	●	●	○	○	○
trnS(gga)	●	●	●	○	○
trnT(ggu)	●	●	●	○	○
trnV(gac)	●	○	○	○	○

[a] A filled/open circle denotes the presence/absence of a gene. Only the chloroplast genes that are missing in one or more cpDNAs are indicated. A total of 91 genes are shared by all compared cpDNAs: *atpA*, *B*, *E*, *F*, *H*, *I*, *ccsA*, *cemA*, *clpP*, *ftsH*, *petA*, *B*, *D*, *G*, *L*, *psaA*, *B*, *C*, *J*, *psbA*, *B*, *C*, *D*, *E*, *F*, *H*, *I*, *J*, *K*, *L*, *M*, *N*, *T*, *Z*, *rbcL*, *rpl2*, *5*, *14*, *16*, *20*, *23*, *36*, *rpoA*, *B*, *C1*, *C2*, *rps2*, *3*, *4*, *7*, *8*, *9*, *11*, *12*, *14*, *18*, *19*, *rrf*, *rrl*, *rrs*, *tufA*, *ycf1*, *3*, *4*, *12*, *trnA*(ugc), *C*(gca), *D*(guc), *E*(uuc), *F*(gaa), *G*(gcc), *G*(ucc), *H*(gug), *I*(cau), *I*(gau), *K*(uuu), *L*(uaa), *L*(uag), *Me*(cau), *Mf*(cau), *N*(guu), *P*(ugg), *Q*(uug), *R*(acg), *R*(ucu), *S*(gcu), *S*(uga), *T*(ugu), *V*(uac), *W*(cca), *Y*(gua).

for a putative LAGLIDADG endonuclease with one or two copies of this motif (table 4); however, these ORFs are distantly related in sequence.

Our BlastN searches of the GenBank database for homologs of the *Pseudendoclonium* chloroplast introns identified only introns from green algae and land plants (*E* value threshold of 1×10^{-6}). The most surprising result of this analysis was the finding of homologous introns (Pa.*atpA*.1 and Pa.*atp1*.1) inserted at the same gene position in the chloroplast and mitochondrial genomes of *Pseudendoclonium* (fig. 3). In addition to highly similar primary sequences and secondary structures, these introns feature in L8 very similar ORFs encoding putative endonucleases with a double LAGLIDADG motif.

Table 5 reports the homologous group I introns of chloroplast origin that proved to be inserted at the same

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

Table 3
Compared Sizes of Expanded Genes in *Pseudendoclonium*, *Chlorella*, and *Chlamydomonas* cpDNAs

	Chlorella		*Pseudendoclonium*		*Chlamydomonas*	
Gene	Size (bp)	Expansion[a]	Size (bp)	Expansion[a]	Size (bp)	Expansion[a]
cemA	801	1.6	909	1.8	1,503	3.0
ftsH	5,163	1.9	7,791	2.9	8,916	3.3
rpoA	837	0.9	1,734	1.8	2,213[b]	2.3
rpoB	3,906	1.2	6,537	2.0	4,967[c]	1.5
rpoC1	2,511	1.3	4,737	2.4	5,739[c]	2.9
rpoC2	4,689	1.3	10,389	2.8	9,423	2.6
ycf1	2,460	2.0	2,505	2.0	5,988	4.9
ycf62	1,515	1.6	1,803	1.9	—[d]	—

[a] Size relative to the corresponding gene in *Mesostigma*. Each value was obtained by dividing the size of the green algal gene by the size of the *Mesostigma* gene.

[b] The indicated size is derived from our unpublished sequence data. We found that a portion of the *rpoA*-coding sequence is missing in accession BK000554 as a result of a sequencing error introducing a frameshift mutation.

[c] The indicated size includes the intergenic spacer separating the ORFs corresponding to the 5′ and 3′ portions of the gene.

[d] *ycf62* is absent from *Chlamydomonas* cpDNA.

sites as their *Pseudendoclonium* counterparts. Five *Pseudendoclonium* introns have known homologs in chlorophycean green algae, whereas homologs of Pa.*rrl*.1 have been observed in prasinophytes, trebouxiophytes, chlorophycean green algae, and the hornwort *Anthoceros punctatus*. The three *psbA* and *rrl* introns in *C. reinhardtii* cpDNA exhibit a high degree of primary sequence and secondary structure conservation with their *Pseudendoclonium* counterparts (Supplementary Fig. S2, Supplementary Material online). Even the unusual P3.1 and P3.2 pairings found in Pa.*psbA*.3 are present in Cr.*psbA*.2.

It is intriguing that L9 of Pa.*psbA*.1 codes for a putative single LAGLIDADG endonuclease and that no endonuclease motif has been assigned to the ORF of 102 codons found within the same loop in the homologous Cr.*psbA*.1 IA1 intron (Holloway, Deshpande, and Herrin 1999). Our analysis of the published *Chlamydomonas* cpDNA sequence mapping at this locus revealed the presence of an ORF of 46 codons encoding a LAGLIDADG motif 29 nt upstream of *orf102*; interestingly, the predicted protein sequence of this ORF exhibits high similarity with the N-terminal domain of the Pa.*psbA*.1-encoded endonuclease (fig. 2*E*). To test the possibility that there is a frameshift in the sequence originally published by Holloway, Deshpande, and Herrin (1999), we sequenced a PCR product containing the Cr.*psbA*.1 intron; however, we found no nucleotide difference. We conclude that a nonsense mutation led to loss of the LAGLIDADG motif in the Cr.*psbA*.1-encoded protein.

Repeated Elements

The *Pseudendoclonium* chloroplast genome contains a large number of repeated elements (fig. 4). Two types of repeated sequences can be distinguished: short tandem repeats and short dispersed repeats (SDRs). The short tandem repeats are found in the vicinity of *trnMe*(cau) in a region spanning about 2 kb. The units forming this repeat region are 10–20 bp in size, rich in A + T, and degenerated in sequence.

Table 4
Group I Introns in *Pseudendoclonium* cpDNA

			ORF		
Designation	Subgroup[a]	Size (bp)	Location[b]	Type[c]	Size (codons)
Pa.*atpA*.1	IB	1,634	L8	LAGLIDADG (2)	302
Pa.*atpA*.2	IA1	344	—	—	—
Pa.*psaA*.1	IB	1,520	L8	LAGLIDADG (2)	301
Pa.*psaB*.1	IB	1,368	L8	LAGLIDADG (2)	234
Pa.*psaB*.2	IB	1,488	L8	LAGLIDADG (1)	253
Pa.*psaB*.3	IA1	1,273	—	—	—
Pa.*psbA*.1	IA1	1,068	L9	LAGLIDADG (1)	156
Pa.*psbA*.2	IA1	1,120	L5	H-N-H	241
Pa.*psbA*.3	IA3	1,452	L3.2	GIY-YIG	294
Pa.*psbA*.4	IA1	1,092	L5	H-N-H	217
Pa.*psbA*.5	IA2	1,216	L6	GIY-YIG	231
Pa.*psbA*.6	IA1	978	L5	H-N-H	222
Pa.*psbA*.7	IA1	1,073	L5	H-N-H	254
Pa.*psbA*.8	IA1	333	—	—	—
Pa.*psbA*.9	IA1	1,045	—	—	—
Pa.*psbA*.10	IA1	314	—	—	—
Pa.*psbB*.1	IA2	966	L6	GIY-YIG	223
Pa.*psbB*.2	IA1	1,282	L5	H-N-H	105
Pa.*psbB*.3	IA1	1,230	L5	H-N-H	105
Pa.*psbC*.1	IA2	887	—	—	—
Pa.*psbC*.2	IA1	907	—	—	—
Pa.*psbC*.3	IA2	960	L6	GIY-YIG	212
Pa.*psbC*.4	IA2	1,051	—	—	—
Pa.*psbD*.1	IA2	1,118	L6	GIY-YIG	214
Pa.*psbD*.2	IA2	921	L6	GIY-YIG	143
Pa.*rbcL*.1	IA1	1,682	L5	H-N-H	246
Pa.*rrl*.1	IA3	816	L6	LAGLIDADG (1)	168

[a] Introns were classified according to Michel and Westhof (1990). The subcategories of IB introns could not be identified unambiguously.

[b] L followed by a number refers to the loop extending the base-paired region identified by the number. A dash denotes the absence of an ORF.

[c] The conserved motif in the predicted endonuclease is given, with the number of copies of the LAGLIDADG motif indicated in parentheses.

The SDRs map to intergenic spacers and/or introns and can be classified into four groups of repeat units (A, B, C, and D) on the basis of their primary sequences (table 6). They occur frequently as palindromic sequences separated by 7–8 bp, and the stem-loop structures often contain more than one SDR unit. As deletion of palindromic structures and other repeated elements is known to occur during cloning in *Escherichia coli*, we confirmed by direct sequencing of PCR products the sizes and sequences of the *Pseudendoclonium* SDRs that we identified by analysis of cloned fragments. To our surprise, we found that numerous SDRs in *Pseudendoclonium* cpDNA bear identical or close sequence similarity to those present in the mitochondrial genome of this alga (Pombert et al. 2004). These homologous mitochondrial SDRs fall within three of the four above-mentioned groups of repeat units (A, B, and C) (table 6) and also occur as stem-loop structures composed of one or more repeat units. To our knowledge, the presence of closely related SDRs in different organelles of the same eukaryotic cell has not been previously reported.

Genome Organization

Our pairwise comparisons of *Pseudendoclonium* cpDNA with previously sequenced green plant cpDNAs indicate that its gene organization is most similar to that of

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

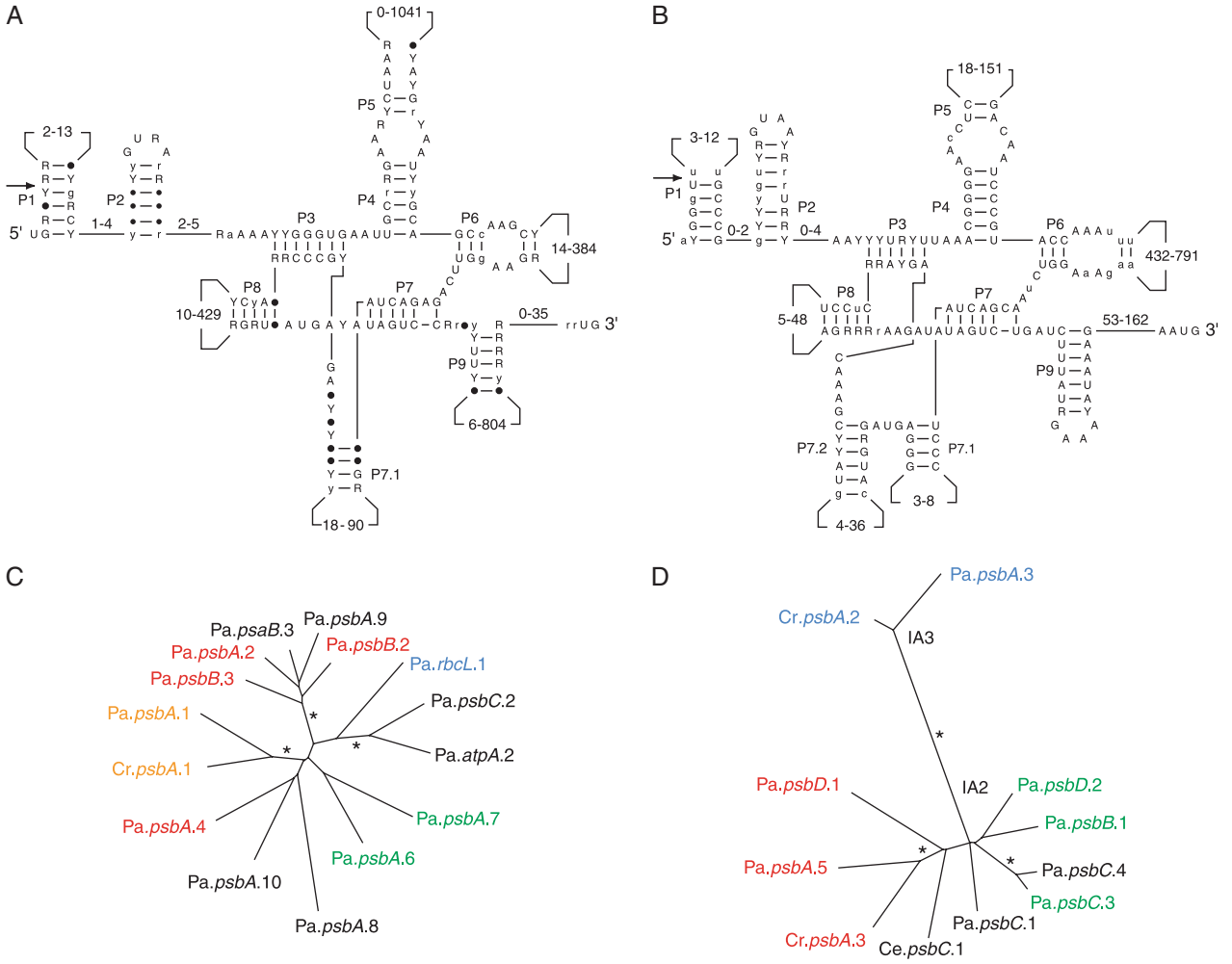

FIG. 2.—Comparative analyses of group I introns and intron ORFs in *Pseudendoclonium* cpDNA. (*A* and *B*) Consensus secondary structure models of introns displayed according to Burke et al. (1987): (*A*) IA1 introns and (*B*) IA2 introns. Arrows denote the 5′ splice sites between exon and intron residues. Highly conserved residues (in 11 IA1 and 5 IA2 introns) and slightly less conserved residues (in nine IA1 and four IA2 introns) are shown in uppercase and lowercase characters, respectively; the other residues are represented by filled circles. Conserved base pairings (in 11 IA1 and 6 IA2 introns) are denoted by bars. Numbers inside the variable loops indicate the size variations of these loops in the compared introns. Note that the P2 pairing is missing in one IA1 intron (Pa.*psbA*.7). (*C* and *D*) Neighbor-Joining analyses of intron core sequences: (*C*) IA1 introns and (*D*) IA2/IA3 introns. The *Chlamydomonas reinhardtii* (Cr) and *Chlamydomonas eugametos* (Ce) introns homologous to the *Pseudendoclonium* (Pa) introns (see table 5) were included in these analyses. Nodes that were identified in at least 80% of the bootstrap replicates are labeled with asterisks. The introns denoted in green, red, blue, and orange encode putative homing endonucleases; in each panel, those denoted by the same color specify closely related proteins. (*E*) Alignment of the endonuclease regions containing the H-N-H, LAGLIDADG, or GIY-YIG motif. The amino acid residues making up each motif are underlined.

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

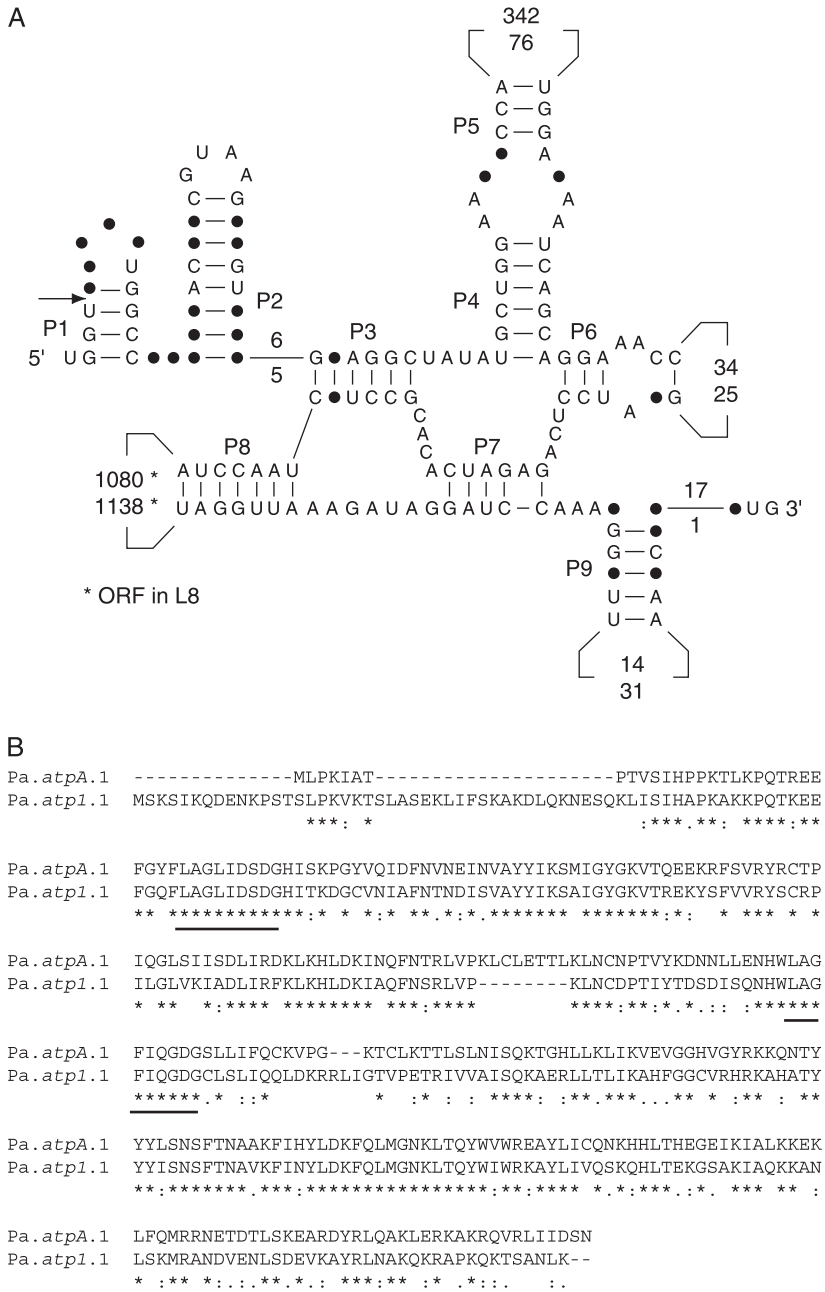

FIG. 3.—Comparative analyses of the chloroplast Pa.*atpA*.1 and mitochondrial Pa.*atp1*.1 introns. (*A*) Consensus secondary structure models of the two introns displayed according to Burke et al. (1987). The arrow denotes the 5′ splice site between the exon and intron residues. Identical residues are shown in uppercase characters, whereas different residues are represented by filled circles. Conserved base pairings are denoted by bars. Upper and lower numbers inside the variable loops indicate size variations in the Pa.*atpA*.1 and Pa.*atp1*.1 introns, respectively. (*B*) ClustalW alignment of the putative LAGLIDADG endonucleases encoded by the two introns. The LAGLIDADG motifs are underlined.

Chlorella. These two green algal genomes share 16 distinct gene clusters, 3 of which (*rps18-rpl20*, *psbK-ycf12-psaM*, and *trnE*(uuc)-*trnMf*(cau)) have not been identified in any other sequenced chloroplast genomes (fig. 1). A single gene cluster (*trnC*(gca)-*trnT*(ugu)) is shared specifically between *Pseudendoclonium* and *Chlamydomonas* cpDNAs.

Twelve of the gene clusters that are conserved between *Mesostigma* and *Nephroselmis* cpDNAs have been broken in *Pseudendoclonium* cpDNA (fig. 5). As revealed by our analysis of the rearrangement break points within these ancestrally conserved clusters, the *Pseudendoclonium* genome shares 9 of the 10 break points found in *Chlorella* cpDNA and features 13 additional break points (fig. 5). All of these nine break points are also observed in *Chlamydomonas* cpDNA (fig. 5). Of the 33 break points exhibited by this chlamydomonad genome, 11 are specifically shared with *Pseudendoclonium*, whereas a single one is specifically shared with *Chlorella*. The *psbB-psbT-psbN-psbH* cluster displays only two break points that are unique to *Pseudendoclonium*. Although the *psbN* gene of this

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

Table 5
Group I Introns at Identical Gene Locations in *Pseudendoclonium* cpDNA, Other Green Algal cpDNAs, and Land Plant cpDNAs

	Homologous Intron	
Pseudendoclonium Intron	Green Plant[a]/Intron Number[b]	Accession Number
Pa.*psbA*.1	*Chlamydomonas reinhardtii* (C) i1	BK000554
Pa.*psbA*.3	*C. reinhardtii* (C) i2	BK000554
Pa.*psbA*.5	*C. reinhardtii* (C) i3	BK000554
Pa.*psbC*.1	*Chlamydomonas eugametos* (C) i1	M90639
Pa.*rbcL*.1	*Chlorogonium capillatum* (C) i1	AB010236
	Chlorogonium euchlororum (C) i1	AB010224
Pa.*rrl*.1	*Chlamydomonas agloeformis* (C)	L43351
	Chlamydomonas callosa (C)	L43501
	Chlamydomonas iyengarii (C)	L43354
	Chlamydomonas mexicana (C)	L43538
	Chlamydomonas nivalis (C)	L42990
	Chlamydomonas peterfii (C)	L43538
	C. reinhardtii (C)	BK000554
	Carteria lunzensis (C)	L42986
	Carteria olivieri (C)	L43500
	Haematococcus lacustris (C)	L49151
	Pediastrum biradiatum (C)	L49156
	Neochloris aquatica (C)	L49155
	Scenedesmus obliquus (C)	L43360
	Trichosarcina mucosa (U)	AY008341
	Chlorella vulgaris (T)	NC_001865
	Monomastix species M722 (P)	L44124
	Monomastix species OKE-1 (P)	L49154
	Scherffelia dubia (P)	L44126
	Anthoceros punctatus (E)	AF393576

[a] The letter in parentheses indicates the specific chlorophyte/streptophyte lineage comprising the green algal/plant indicated. P, Prasinophyceae; U, Ulvophyceae; T, Trebouxiophyceae; C, Chlorophyceae; and E, Embryophyta.

[b] The intron number is given only when more than one intron is present.

ulvophyte has retained its location between *psbT* and *psbH*, it has been relocated to the DNA strand encoding the other *psb* genes (fig. 1).

Phylogenetic Analyses

The amino acid and nucleotide sequences derived from the 58 protein-coding genes (see Supplementary Table S1, Supplementary Material online) that are shared between the cpDNAs of *Pseudendoclonium*, *Chaetosphaeridium*, *Chlamydomonas*, *Chlorella*, *Marchantia*, *Nephroselmis*, and *Nicotiana* were concatenated and analyzed with ML inference methods using the homologous sequences of *Mesostigma* as the outgroup. As expected, analyses of both the amino acid (11,225 positions) and nucleotide (25,318 positions) data sets yielded a best tree (T1) in which the chlorophytes and streptophytes form two distinct lineages (fig. 6). In the chlorophyte lineage, the prasinophyte *Nephroselmis* occupies the most basal position, whereas the trebouxiophyte *Chlorella*, the ulvophyte *Pseudendoclonium*, and chlorophycean alga *Chlamydomonas* form a clade in which *Pseudendoclonium* is sister to *Chlorella*. Topology T1 accounted for 91% and 80% of the bootstrap replicates in the analyses of the amino acid and nucleotide data sets, respectively, and was also found to be the most highly supported topology in distance (ML and LogDet distances) and maximum parsimony analyses (data not shown). In all analyses, instability in the branching order of taxa was observed only for the UTC clade. The alternative topology showing *Pseudendoclonium* as sister to *Chlamydomonas* (T2) was recovered in 9% and 20% of the bootstrap replicates in the ML analyses of the amino acid and nucleotide data sets, respectively, whereas the topology placing *Pseudendoclonium* at a basal position in the UTC clade (T3) was not detected in any of the 100 bootstrap replicates. The confidence levels of these alternative topologies were assessed using the statistical test of Shimodaira-Hasegawa. T3, but not T2, proved to be significantly worse than T1 at $P < 0.05$ in the analyses of both the amino acid and nucleotide data sets (fig. 7).

Separate ML analyses of the amino acid data set with CODEML revealed no strong disagreement among the phylogenetic signals provided by the individual proteins (Supplementary Table S1, Supplementary Material online). Topologies T1, T2, and T3 were found to be supported by 22, 25, and 11 proteins, respectively, with RELL bootstrap values ranging from 39% to 93%. However, all 58 proteins, with the exception of *rpoB*, failed to provide a signal of sufficient strength to reject one or both of the alternative topologies at $P < 0.05$ in the Shimodaira-Hasegawa test. *rpoB* supported T1 and rejected T3 (but not T2) at $P = 0.02$. Thus, it appears that general homoplasy throughout the data set is the most likely explanation for the lack of resolution between T1 and T2.

Phylogenetic relationships were also inferred from gene order data using the relative positions of the 80 chloroplast genes shared by *Pseudendoclonium* and the seven green plants mentioned above. GRAPPA 2.0, the program used for these analyses, reconstructs phylogenies by assuming that gene rearrangements occur by inversions. The best tree was found to display the T2 topology and to feature a total length of 240 inversion steps (fig. 8). User-tree analyses constrained to the T1 and T3 topologies yielded trees with 6 and 10 extra steps, respectively.

Mapping of shared gene losses and rearrangement break points located within ancestrally conserved gene clusters on topologies T1, T2, and T3 also revealed that these two sets of structural characters independently support topology T2 as the most parsimonious scenario (fig. 7). A total of 36 rearrangement break points and 41 gene losses were mapped on T2 compared to 47 break points and 46 gene losses on T1 and 46 break points and 45 gene losses on T3. In contrast to T1 and T3, T2 shows several rearrangement events and gene losses that are specifically shared between sister taxa in the UTC lineage.

Discussion

Unusual Features of *Pseudendoclonium* cpDNA

Like the chloroplast genomes of the chlorophycean green algae belonging to the genus *Chlamydomonas*, *Pseudendoclonium* cpDNA is unusual with respect to its quadripartite structure (fig. 1). Fourteen of the genes found in the SSC region of this ulvophyte genome map to a different genomic region (most often the LSC region) in all streptophytes featuring an IR as well as in the prasinophytes *Nephroselmis* and *Mesostigma*. Moreover, the *Pseudendoclonium* IR features an rRNA operon that is transcribed toward the LSC region. During the evolution of ulvophytes,

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

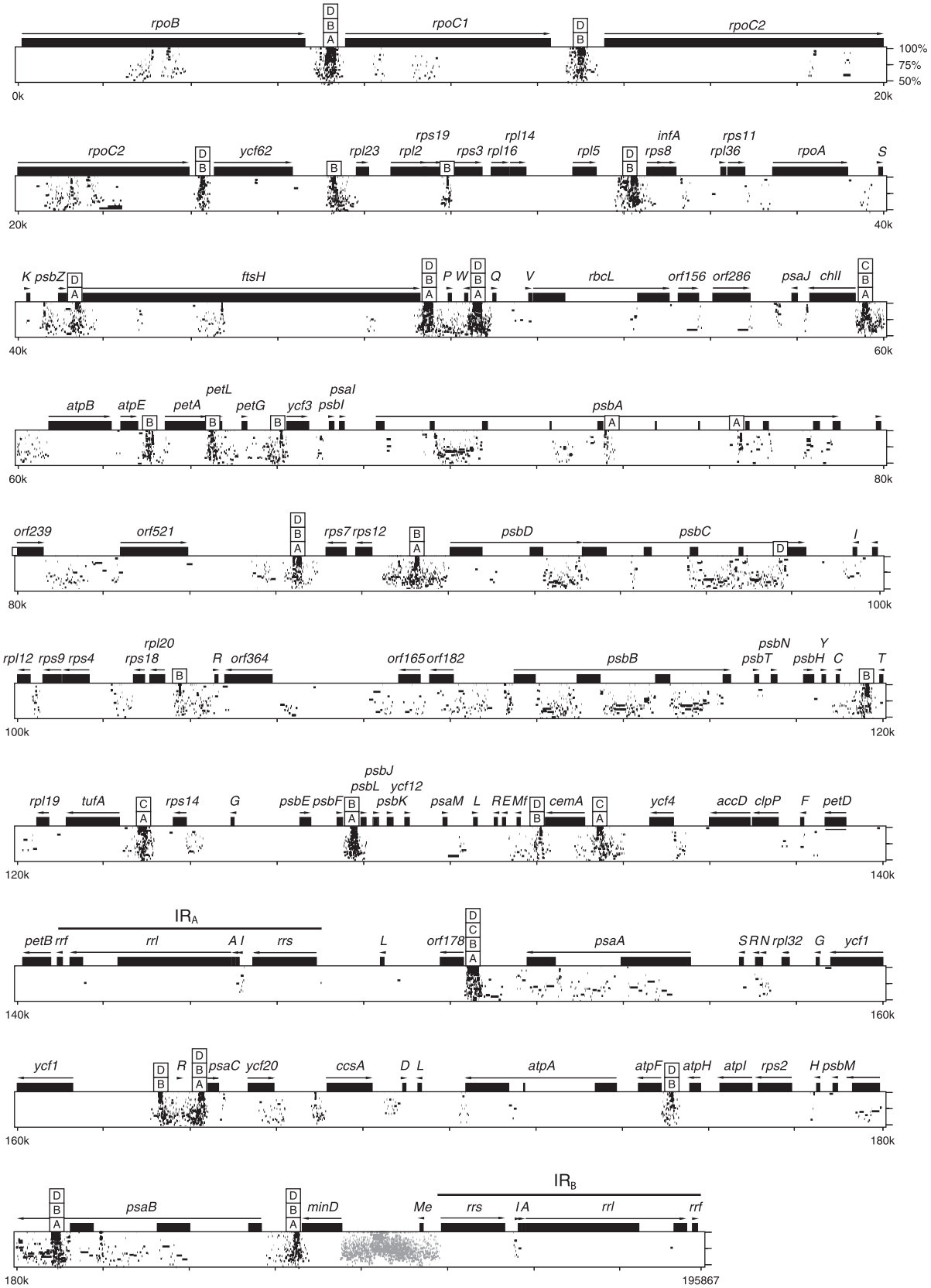

FIG. 4.—PipMaker analysis of *Pseudendoclonium* cpDNA. This genome sequence was aligned against itself. At the top of the alignment, genes and their polarities are denoted by horizontal arrows, and coding sequences are represented by filled boxes. The SDR repeat units described in table 6 are denoted by boxed letters. The short tandem repeats near *trnMe*(cau) are shown as gray dots. Similarities between aligned regions are shown as average percent identity between 50% and 100% identity.

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

Table 6
SDR Repeat Units in *Pseudendoclonium* cpDNA and mtDNA

Designation	Sequence	Number[a] cpDNA	Number[a] mtDNA
A	TTTTGCCCCCGCAGTCTTTAAGCCTGCGAAAAGA	15 (22)	22 (35)[b]
B	AAGCAATTTGTTCGTAAACTCACAAACTTC	38 (45)	0 (31)
C	TTTTGCCCGCAGCAAAAGCAAGGGGCAG	4 (4)	8 (8)
D	CTCCGGACTTTTGCCC	17 (23)	0 (0)

[a] Numbers of repeat units with 100% or 90% (in parentheses) sequence identity are given.
[b] Only the underlined sequence is present in mtDNA.

the transfer of genes from the LSC to the SSC region may have been closely associated with the inversion of the IR. In this context, it may be envisioned that an ancestral genome with a conventional quadripartite structure has undergone two consecutive inversions, each involving a segment encompassing the IR and part of the LSC or, alternatively, one inversion involving a segment encompassing the entire IR sequence and the other the second IR sequence. Whatever the mechanism, additional ulvophyte chloroplast genomes will need to be examined to determine whether the atypical quadripartite structure reported here for *Pseudendoclonium* cpDNA is commonly found in the Ulvophyceae. Prior to our study, the only ulvophyte chloroplast genome that had been investigated by physical and gene mapping is that of *Codium fragile*, a green alga thought to belong to a later-diverging lineage than the *Pseudendoclonium* lineage (Manhart et al. 1989). Considering that *Codium* cpDNA is highly reduced in size (89 kb), lacks an IR, and has a broken rRNA operon, the chloroplast genome appears to have evolved under relaxed constraints in the Ulvophyceae.

Another unusual feature of *Pseudendoclonium* cpDNA concerns its intron composition (table 4). Of all the chloroplast genome sequences reported so far, this chlorophyte genome is the richest in group I introns. The striking similarity among the 14 IA1 and 7 IA2 introns and the putative homing endonucleases encoded by these introns (fig. 2) suggests that many of the 27 group I introns found in *Pseudendoclonium* cpDNA arose from intragenomic proliferation of a few founding introns in the lineage leading to *Pseudendoclonium*. The possibility that numerous *Pseudendoclonium* introns took their origin from independent invasions of closely related introns by lateral transfer rather than by intragenomic proliferation seems less likely but cannot be excluded. Interestingly, the intron composition of *Pseudendoclonium* cpDNA resembles that of its mitochondrial counterpart (Pombert et al. 2004) in featuring exclusively group I introns, a substantial fraction of which specifies potential homing endonucleases.

The highly similar group IB introns sharing the same insertion site in the *Pseudendoclonium* chloroplast *atpA* and mitochondrial *atp1* genes (fig. 3) as well as the closely related SDRs in both organelle DNAs of this green alga (table 6) provide compelling evidence for interorganellar, lateral transfer of these genetic elements. The lateral transfer of the *atpA*/*atp1* introns appears to have occurred specifically in the ulvophyte lineage, considering that these introns represent the first examples of intron insertion at their cognate site in *atpA*/*atp1*. Cases of interorganellar transfers of group I introns have been previously documented for green algae (Turmel, Mercier, and Côté 1993; Turmel et al. 1995, 1999); however, the rDNA introns associated with these transfers exhibit a sporadic distribution pattern and occur in more than one chlorophyte lineage, making it difficult to estimate the timing of the lateral transfer event(s).

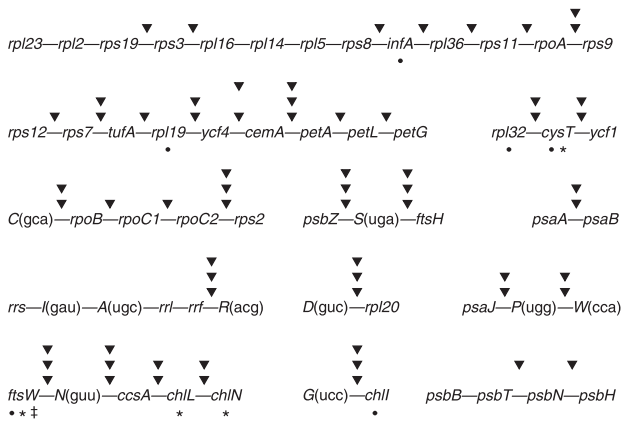

FIG. 5.—Fragmented ancestral gene clusters in *Pseudendoclonium*, *Chlorella*, and *Chlamydomonas* cpDNAs. The indicated clusters are found in both *Mesostigma* and *Nephroselmis* cpDNAs. Note that *rpl22* has not been represented in the large ribosomal protein gene cluster; this gene is found in *Mesostigma* between *rps19* and *rps3* but is missing from *Nephroselmis* and the three other chlorophytes. Sites of fragmentation are denoted by arrowheads above the clusters, with the arrowheads at the lower, middle, and upper positions pointing to sites in *Chlamydomonas*, *Pseudendoclonium*, and *Chlorella*, respectively. Genes missing from *Chlamydomonas*, *Pseudendoclonium*, and *Chlorella* are denoted by circles, asterisks, and double dagger, respectively, below the clusters. Gene polarities are not shown.

The cpDNAs of *Pseudendoclonium* and Other Advanced Chlorophytes Share Common Evolutionary Trends

A number of shared derived features between the cpDNAs of *Pseudendoclonium* and other advanced chlorophytes highlight common evolutionary trends among the members of the UTC clade. Like its *Pseudendoclonium* homolog, the completely sequenced cpDNA of the chlorophycean green alga *C. reinhardtii* as well as other chlamydomonad cpDNAs that have been analyzed by physical and gene mapping display an atypical quadripartite structure in which gene partitioning among the two single-copy regions differs from the conserved pattern observed in all other green plant cpDNAs containing an IR (Boudreau, Otis, and Turmel 1994; Boudreau and Turmel 1995, 1996; Maul et al. 2002). The single-copy regions of these genomes are about equal in size, and their genes have been so extensively

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

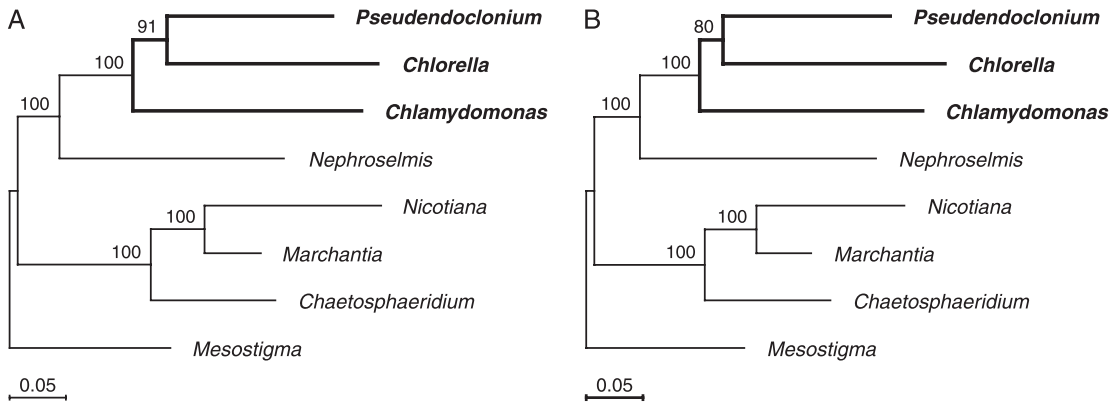

FIG. 6.—Phylogenetic position of *Pseudendoclonium* as inferred by ML analyses of 58 cpDNA-encoded proteins and genes. (*A*) Best protein tree. (*B*) Best gene tree. Bootstrap values are indicated on the corresponding nodes, and branch lengths are drawn to scale.

shuffled that it is impossible to trace the events that led to their relocations and to determine if the rRNA operons have been altered in their orientation. On the other hand, studies of *C. vulgaris* and *Chlorella ellipsoidea* cpDNAs clearly indicate that the IR was the subject of rearrangement or deletion during the evolution of trebouxiophytes (Yamada 1991; Wakasugi et al. 1997). Given the absence of the IR in *C. vulgaris* and the fragmentary information currently available for the IR-containing cpDNA of *C. ellipsoidea*, it remains unknown if genes were relocated from one single-copy region to the other. Unlike its *C. vulgaris* homolog which is intact, the *C. ellipsoidea* rRNA operon is fragmented into two pieces, one containing the *rrs* and *trnI*(gau) genes and the other containing the *trnA*(ugc), *rrl*, and *rrf* genes (Yamada and Shimaji 1987). The orientation of these genes is such that a single inversion of the latter fragment would recreate an intact rRNA operon that is transcribed toward the LSC region. From these data alone, one cannot conclude about the timing of the event(s) that led to the change in orientation of the *Pseudendoclonium* and *C. ellipsoidea* rRNA operons. It is possible that the chloroplast genome of the last common ancestor of the UTC algae had already acquired an IR with an rRNA operon transcribed toward the LSC region and an SSC region with some genes usually present in the opposite single-copy region.

UTC algae also exhibit common features at the levels of gene content, gene order, repeated sequences, and intron content. *Pseudendoclonium*, *Chlorella,* and *Chlamydomonas* cpDNAs have each lost many genes that are present in *Nephroselmis* cpDNA, with 18 missing genes shared between all three chlorophyte genomes (table 2). The three chlorophyte genomes share nine rearrangement break points within seven gene clusters that are common to *Mesostigma* and *Nephroselmis* (fig. 5), and they all possess SDRs (this study; Maul et al. 2002). In this context, it is interesting to note that intramolecular recombination events between SDRs have been proposed to cause the fragmentation of ancestral operons (Palmer 1991). Altogether, these observations support the notion that the chloroplast genome of the last common ancestor of UTC algae was relatively gene-poor, punctuated with SDRs, and endowed with broken ancestral clusters. Although multiple introns are present in *Pseudendoclonium*, *Chlorella,* and *Chlamydomonas* cpDNAs, there is no evidence that some of these introns took their origin just before the emergence of the UTC clade. Considering that five of the six introns that *Pseudendoclonium* shares with other green plants are structurally and positionally homologous to chloroplast introns that have been documented only in chlorophycean green algae (table 5), the origin of these introns might be attributed to vertical inheritance from the last common ancestor of ulvophytes and chlorophycean algae. However, given that all five introns, except one (Pa.*psbC*.1), encode homing endonucleases (table 4) and are thus most probably mobile, we cannot exclude the possibility that horizontal transfer accounts for their presence in both ulvophytes and chlorophycean green algae. Pa.*rrl*.1, the remaining intron that *Pseudendoclonium* shares with other green plants, encodes a LAGLIDADG homing endonuclease; interestingly, of all the introns identified so far in green plant cpDNAs, this intron shows the broadest phylogenetic distribution, being found in representatives of the four chlorophyte classes as well as in streptophytes (Lucas et al. 2001; Turmel et al. 2002).

Pseudendoclonium cpDNA features an intermediary level of ancestral characters relative to its *Chlorella* and *Chlamydomonas* counterparts. Its gene complement (105 genes) displays more genes than *Chlamydomonas* cpDNA (94 genes), but fewer than *Chlorella* cpDNA (112 genes) (table 2). Moreover, ancestral gene clusters have been less extensively rearranged in *Pseudendoclonium* than in *Chlamydomonas*, but more in *Pseudendoclonium* than in *Chlorella* cpDNAs (fig. 5). Assuming that ancestral characters were lost gradually during evolution, these data would support the idea that the Trebouxiophyceae appeared before the emergence of the Ulvophyceae and Chlorophyceae. However, as loss of ancestral characters is known to occur at variable rates, independent evidence is necessary to validate this hypothesis.

Phylogenetic Inferences from Sequences and Structural Features of the Chloroplast Genome Favor the Hypothesis that the Ulvophyceae Is Sister to the Chlorophyceae

Although our phylogenetic analyses of chloroplast protein and gene sequences have failed to identify unambiguously the divergence order of the UTC lineages, they reject with high confidence the hypothesis that the Ulvophyceae

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

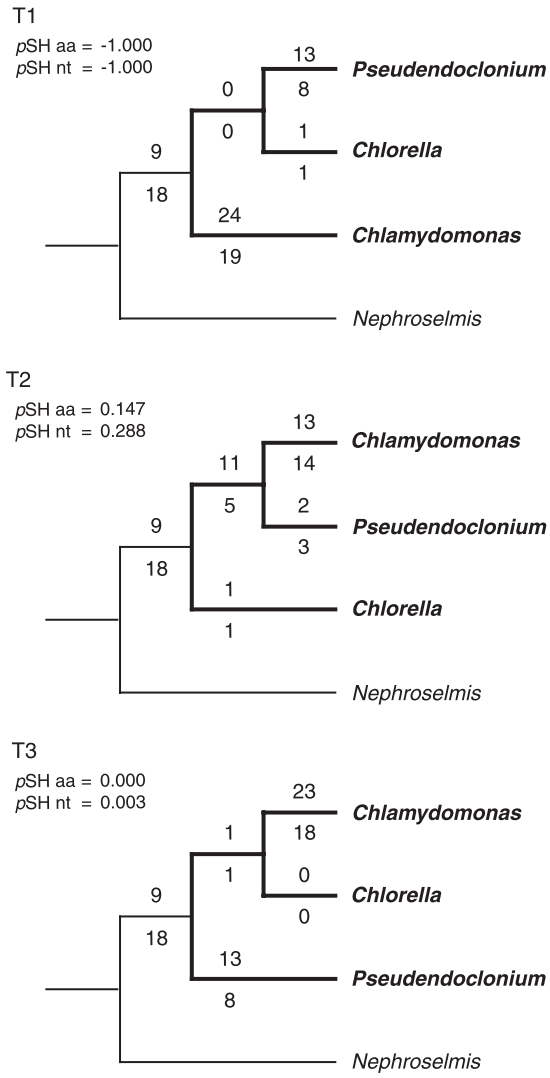

FIG. 7.—Evaluation of the three phylogenetic hypotheses for the branching order of UTC taxa using the Shimodaira-Hasegawa test and structural evidence based on shared gene losses and rearrangement break points within ancestrally conserved gene clusters. Values above and below the nodes indicate the number of break points and gene losses, respectively. The *P* values obtained in the Shimodaira-Hasegawa test are indicated for the ML analyses of both the amino acid (aa) and nucleotide (nt) data sets.

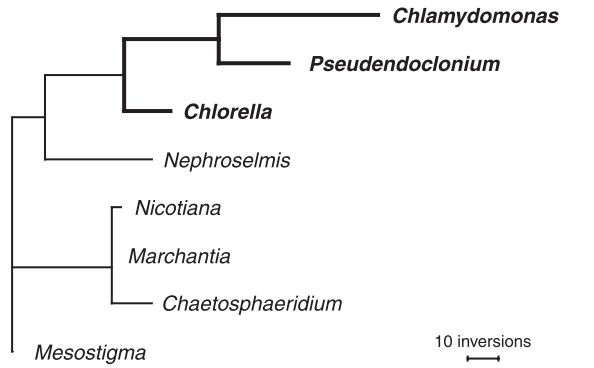

FIG. 8.—Phylogenetic position of *Pseudendoclonium* as inferred by gene order analysis of 80 chloroplast genes. The best tree is shown with branch lengths drawn to scale.

occupies a basal position relative to the Trebouxiophyceae and Chlorophyceae (T3 in fig. 7). About 20 years ago, Mattox and Stewart (1984) proposed this hypothesis based on comparisons of flagellar structure, characteristics of cell division, and nature of cell covering. Our phylogenetic analyses of 58 chloroplast proteins and genes rather favor the idea that the Chlorophyceae diverged first (T1 in fig. 7), but they cannot eliminate the possibility that the Ulvophyceae is sister to the Chlorophyceae (T2 in fig. 7).

Of the latter two hypotheses, we strongly favor that identifying the Ulvophyceae and Chlorophyceae as sister groups because it is robustly supported by three independent sets of data: (1) phylogenetic analysis of gene order data (fig. 8), (2) structural evidence based on derived characters such as shared gene losses and rearrangement break points within ancestrally conserved gene clusters (fig. 7), and (3) phylogenetic analyses of multiple mtDNA-encoded proteins (Pombert et al. 2004). Among the derived cpDNA characters supporting the notion that the Ulvophyceae and Chlorophyceae diverged after the Trebouxiophyceae, we find not only gene losses and rearrangement break points specifically shared between *Pseudendoclonium* and *Chlamydomonas* but also shared patterns of expansion for the *rpo* genes (table 3) and closely related group I introns occupying identical positions. In this context, the occurrence of *rpoB* and *rpoC1* as two distinct ORFs in *Chlamydomonas* but as single ORFs in *Pseudendoclonium* suggests that the capture of new sequences by the coding regions of these genes preceded their fragmentation.

Published phylogenies of chlorophytes inferred from nuclear genes (Friedl and O'Kelly 2002) are not in disagreement with the two hypotheses supported by our phylogenetic analyses of cpDNA-encoded sequences. Because interclass relationships are poorly resolved in these nuclear phylogenies based mostly on a single gene (18S rRNA gene), the branching order of the UTC lineages still remains ambiguous in spite of a broad representation of taxa (Friedl and O'Kelly 2002). Although the sequences of green algal organelle genomes offers great potential toward the resolution of interclass relationships, the current data set of whole-genome sequences suffers from very limited taxon sampling. Clearly, the chloroplast and mitochondrial genome sequences of additional chlorophytes will be required to provide unambiguous support for a sister-group relationship between the Ulvophyceae and Chlorophyceae.

Repeated Elements as an Evolutionary Force

From the four chlorophyte chloroplast genome sequences that are currently available, we find that there exists a correlation between the abundance of SDRs and the extent of gene rearrangements. The chloroplast genome of the prasinophyte *Nephroselmis* contains virtually no SDRs and displays the most ancestral gene organization among the chlorophytes (figs. 5 and 9). Although the cpDNAs of the UTC algae feature SDRs, the abundance of these elements is variable (fig. 9). *Chlamydomonas* cpDNA displays the greatest density of SDRs and is the most scrambled chlorophyte genome in gene order, whereas *Chlorella* cpDNA exhibits the lowest density of

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

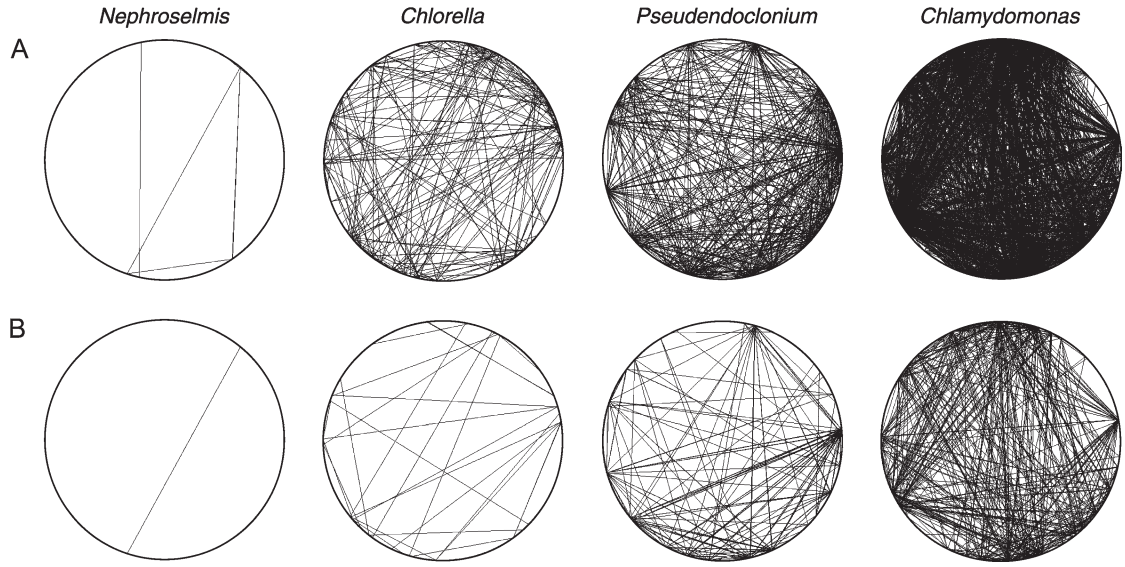

FIG. 9.—Densities of repeated sequences in chlorophyte cpDNAs as revealed by REPuter. (*A*) Repeats of 30 nt or more. (*B*) Repeats of 45 nt or more. For these analyses, one copy of the IR sequence was deleted from the *Nephroselmis*, *Pseudendoclonium*, and *Chlamydomonas* genomes. Repeated elements with identical sequences are connected by lines.

SDRs and is the least rearranged. Again, *Pseudendoclonium* cpDNA occupies an intermediate position among the UTC algae with respect to the density of SDRs and the extent of gene rearrangements. The proliferation of SDRs in the chloroplast genome thus appears to be a major cause of genome reorganization in the UTC clade. More chloroplast genome sequences from representatives of each chlorophyte class will be needed to understand the dynamics of SDR evolution.

SDRs in green plant cpDNAs could serve as hot spots for nonhomologous recombinational events and lead to inversions when they are in an inverted orientation (Palmer 1991). As indicated by comparative analyses of cpDNAs from numerous land plants (Palmer 1991) and from closely related pairs of chlamydomonads (Boudreau and Turmel 1995, 1996), inversion is most likely the predominant mode of chloroplast gene reorganization. However, for the highly rearranged cpDNAs of the angiosperm family Campanulaceae (Cosner, Raubeson, and Jansen 2004) and subclover (Milligan, Hampton, and Palmer 1989), SDRs have also been proposed to promote transposition events. No sequence elements with characteristics of transposons have been identified in *Pseudendoclonium* cpDNA; the Wendy element harbored by *Chlamydomonas* is the only transposon-like sequence that has been found among green plant cpDNAs (Fan, Woelfle, and Mosig 1995).

Supplementary Material

The *Pseudendoclonium* cpDNA and 18S rDNA sequences reported in this study have been deposited in the GenBank database (accession numbers AY835431 and DQ011230, respectively). All data sets used in phylogenetic analyses are available as supplementary data files. Supplementary Figures S1 and S2 and Supplementary Table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Patrick Charlebois for his help with the analysis of conserved gene clusters and gene order data. We also thank Charles F. Delwiche and the two anonymous reviewers for their valuable comments and suggestions. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (to M.T. and C.L.). J.-F.P. gratefully acknowledges a scholarship from CREFSIP (Centre de Recherche sur la Fonction, la Structure et l'Ingénierie des Protéines).

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.

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.

Boudreau, E., C. Otis, and M. Turmel. 1994. Conserved gene clusters in the highly rearranged chloroplast genomes of *Chlamydomonas moewusii* and *Chlamydomonas reinhardtii*. Plant Mol. Biol. **24**:585–602.

Boudreau, E., and M. Turmel. 1995. Gene rearrangements in *Chlamydomonas* chloroplast DNAs are accounted for by inversions and by the expansion/contraction of the inverted repeat. Plant Mol. Biol. **27**:351–364.

———. 1996. Extensive gene rearrangements in the chloroplast DNAs of *Chlamydomonas* species featuring multiple dispersed repeats. Mol. Biol. Evol. **13**:233–243.

Bremer, K. 1985. Summary of green plant phylogeny and classification. Cladistics **1**:369–385.

Burke, J. M., M. Belfort, T. R. Cech, R. W. Davies, R. J. Schweyen, D. A. Shub, J. W. Szostak, and H. F. Tabak.

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

1987. Structural conventions for group I introns. Nucleic Acids Res. **15**:7217–7221.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. **17**:540–552.

Cosner, M. E., L. A. Raubeson, and R. K. Jansen. 2004. Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol. Biol. **4**:27.

Fan, W. H., M. A. Woelfle, and G. Mosig. 1995. Two copies of a DNA element, 'Wendy', in the chloroplast chromosome of *Chlamydomonas reinhardtii* between rearranged gene clusters. Plant Mol. Biol. **29**:63–80.

Floyd, G. L., and C. J. O'Kelly. 1990. Phylum Chlorophyta. Class Ulvophyceae. Pp. 617–635 *in* L. Margulis, J. O. Corliss, M. Melkonian, and D. J. Chapman, eds. Handbook of Protoctista. Jones and Bartlett Publishers, Boston.

Friedl, T., and C. J. O'Kelly. 2002. Phylogenetic relationships of green algae assigned to the genus *Planophila* (Chlorophyta): evidence from 18S rDNA sequence data and ultrastructure. Eur. J. Phycol. **37**:373–384.

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.

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

Hamby, R. K., L. Sims, L. Issel, and E. Zimmer. 1988. Direct ribosomal RNA sequencing: optimization of extraction and sequencing methods for work with higher plants. Plant Mol. Biol. Rep. **6**:175–192.

Helms, G., T. Friedl, G. Rambold, and H. Mayrhofer. 2001. Identification of photobionts from the lichen family *Physciaceae* using algal-specific ITS rDNA sequencing. Lichenologist **33**:73–86.

Holloway, S. P., N. N. Deshpande, and D. L. Herrin. 1999. The catalytic group-I introns of the *psbA* gene of *Chlamydomonas reinhardtii*: core structures, ORFs and evolutionary implications. Curr. Genet. **36**:69–78.

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.

Kurtz, S., J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. **29**:4633–4642.

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.

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.

Lucas, P., C. Otis, J.-P. Mercier, M. Turmel, and C. Lemieux. 2001. Rapid evolution of the DNA-binding site in LAGLIDADG homing endonucleases. Nucleic Acids Res. **29**:960–969.

Manhart, J. R., K. Kelly, B. S. Dudock, and J. D. Palmer. 1989. Unusual characteristics of *Codium fragile* chloroplast DNA revealed by physical and gene mapping. Mol. Gen. Genet. **216**:417–421.

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

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.

Maul, J. E., J. W. Lilly, L. Cui, C. W. dePamphilis, W. Miller, E. H. Harris, and D. B. Stern. 2002. The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. Plant Cell **14**:2659–2679.

McCracken, D. A., M. J. Nadakavukaren, and J. R. Cain. 1980. A biochemical and ultrastructural evaluation of the taxonomic position of *Glaucosphaera vacuolata* Korsch. New Phytol. **86**:39–44.

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.

Milligan, B. G., J. N. Hampton, and J. D. Palmer. 1989. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol. Biol. Evol. **6**:355–368.

Moret, B. M., L. S. Wang, T. Warnow, and S. K. Wyman. 2001. New approaches for reconstructing phylogenies from gene order data. Bioinformatics **17**(Suppl. 1):S165–S173.

Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. **302**:205–217.

O'Brien, E. A., E. Badidi, A. Barbasiewicz, C. deSousa, B. F. Lang, and G. Burger. 2003. GOBASE—a database of mitochondrial and chloroplast information. Nucleic Acids Res. **31**:176–178.

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

Pombert, J. F., C. Otis, C. Lemieux, and M. Turmel. 2004. The complete mitochondrial DNA sequence of the green alga *Pseudendoclonium akinetum* (Ulvophyceae) highlights distinctive evolutionary trends in the Chlorophyta and suggests a sister-group relationship between the Ulvophyceae and Chlorophyceae. Mol. Biol. Evol. **21**:922–935.

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

Schwartz, S., Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. 2000. PipMaker—a web server for aligning two genomic DNA sequences. Genome Res. **10**:577–586.

Sluiman, H. J. 1985. A cladistic evaluation of the lower and higher green plants (*Viridiplantae*). Plant Syst. Evol. **149**:217–232.

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

Tupa, D. D. 1974. An investigation of certain chaetophoralean algae. Beih. zur Nova Hedwigia **46**(Suppl.):64–67.

Turmel, M., V. Côté, C. Otis, J.-P. Mercier, M. W. Gray, K. M. Lonergan, and C. Lemieux. 1995. Evolutionary transfer of ORF-containing group I introns between different subcellular compartments (chloroplast and mitochondrion). Mol. Biol. Evol. **12**:533–545.

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., 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. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis*

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024

olivacea: insights into the architecture of ancestral chloroplast genomes. Proc. Natl. Acad. Sci. USA **96**:10248–10253.

———. 2002. 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.

Volfovsky, N., B. J. Haas, and S. L. Salzberg. 2001. A clustering method for repeat analysis in DNA sequences. Genome Biol. **2**:Research0027.

Wakasugi, T., T. Nagai, M. Kapoor et al. (15 co-authors). 1997. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: the existence of genes possibly involved in chloroplast division. Proc. Natl. Acad. Sci. USA **94**:5967–5972.

White, T. J., T. Bruns, S. Lee, and J. Taylor. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Pp. 315–322 *in* M. A. Innis, D. H. Gelfand, J. J. Sninsky, and T. J. White, eds. PCR protocols: a guide to methods and applications. Academic Press, New York.

Yamada, T. 1991. Repetitive sequence-mediated rearrangements in *Chlorella ellipsoidea* chloroplast DNA: completion of nucleotide sequence of the large inverted repeat. Curr. Genet. **19**:139–147.

Yamada, T., and M. Shimaji. 1987. Splitting of the ribosomal RNA operon on chloroplast DNA from *Chlorella ellipsoidea*. Mol. Gen. Genet. **208**:377–383.

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

Charles Delwiche, Associate Editor

Accepted May 25, 2005

Downloaded from https://academic.oup.com/mbe/article/22/9/1903/982132 by guest on 08 March 2024