Nucleic Acids Research, 1995, Vol. 23, No. 23 4745–4752

Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus *Euglena*

Michael D. Thompson, Donald W. Copertino, Eric Thompson, Mitchell R. Favreau and Richard B. Hallick*

Department of Biochemistry, University of Arizona, Tucson, AZ 85721, USA

Received October 25, 1995; Accepted October 25, 1995

GenBank accession nos U21004–U21010 (incl.)

ABSTRACT

The origin of present day introns is a subject of spirited debate. Any intron evolution theory must account for not only nuclear spliceosomal introns but also their antecedents. The evolution of group II introns is fundamental to this debate, since group II introns are the proposed progenitors of nuclear spliceosomal introns and are found in ancient genes from modern organisms. We have studied the evolution of chloroplast introns and twintrons (introns within introns) in the genus *Euglena*. Our hypothesis is that *Euglena* chloroplast introns arose late in the evolution of this lineage and that twintrons were formed by the insertion of one or more introns into existing introns. In the present study we find that 22 out of 26 introns surveyed in six different photosynthesis-related genes from the plastid DNA of *Euglena gracilis* are not present in one or more basally branching *Euglena* spp. These results are supportive of a late origin for *Euglena* chloroplast group II introns. The *psbT* gene in *Euglena viridis*, a basally branching *Euglena* species, contains a single intron in the identical position to a *psbT* twintron from *E.gracilis*, a derived species. The *E.viridis* intron, when compared with 99 other *Euglena* group II introns, is most similar to the external intron of the *E.gracilis psbT* twintron. Based on these data, the addition of introns to the ancestral *psbT* intron in the common ancester of *E.viridis* and *E.gracilis* gave rise to the *psbT* twintron in *E.gracilis*.

INTRODUCTION

Are introns ancient or are they evolutionarily recent introductions to genetic systems? What is the evolutionary history of introns? These questions have stimulated lively debate for 20 years. Two theories have predominated. According to the introns early view introns have existed since the origin of life and were selectively lost in bacterial genomes and to a lesser extent in eukaryotic genomes (1,2). The alternative introns late view is that ancient genes lacked introns and that contemporary introns were inserted into genes late in eukaryotic evolution (3–5). Organellar group I and group II introns have received little attention in this debate compared with nuclear spliceosomal introns. The timing of group II intron origin is fundamental to this debate, since group II introns are found in contemporary organisms within genes for ancient photosynthetic reaction center polypeptides. Evidence for photosynthetic organisms resembling modern cyanobacteria is present in 3.5 billion-year-old fossil deposits (6). The positions of introns in chloroplast-encoded, photosynthesis-related genes has previously been cited in support of a 'limited universe of exons' in the prebiotic world (7). Furthermore, group II introns have been proposed as the progenitors of nuclear spliceosomal introns (8–18). Therefore, understanding the evolution of group II introns in organisms that exemplify the vast diversity of early eukaryotic evolution is central to the question of an early or late origin for all introns. By tracing the evolutionary history of chloroplast introns we can test some predictions of the introns early versus introns late hypotheses.

Since *Euglena gracilis* chloroplast DNA is the richest known source of introns, the genus *Euglena* is an ideal system for studying the evolution and proliferation of group II and group III introns. The 143 kbp *E.gracilis* plastid genome contains 155 group II and group III introns (Fig. 1), nearly 10 times the number in any other known plastid DNA. The group III introns appear to be streamlined versions of group II introns (19), sharing a common evolutionary ancestor with a group II intron. The origin of group III introns from group II introns may parallel that of nuclear spliceosomal introns from a group II progenitor (20). Among the *E.gracilis* introns are 15 twintrons and complex twintrons, which are introns within introns (19,21–25). Since all known internal introns disrupt a functional domain of the external intron, twintrons are spliced from inside out. Excision of internal introns occurs first. Copertino (25) proposed that the occurrence of twintrons is most consistent with the introns late view and that twintrons may have been formed by the insertion of one or more mobile introns into another intron.

We have exploited the unique structure of twintrons, the availability of cultures of different *Euglena* spp. and information on a large number of individual introns in *Euglena* to test the hypotheses that introns were added to intronless progenitors during the evolution of *Euglena* plastid DNA and that twintrons were formed by the insertion of one or more introns into an existing intron. The rationale is that genes from species of *Euglena* with ancestral forms of the plastid genome should resemble progenitors

* To whom correspondence should be addressed at present address: Department of Neurobiology, The Scripps Research Institute, La Jolla, CA 92037, USA

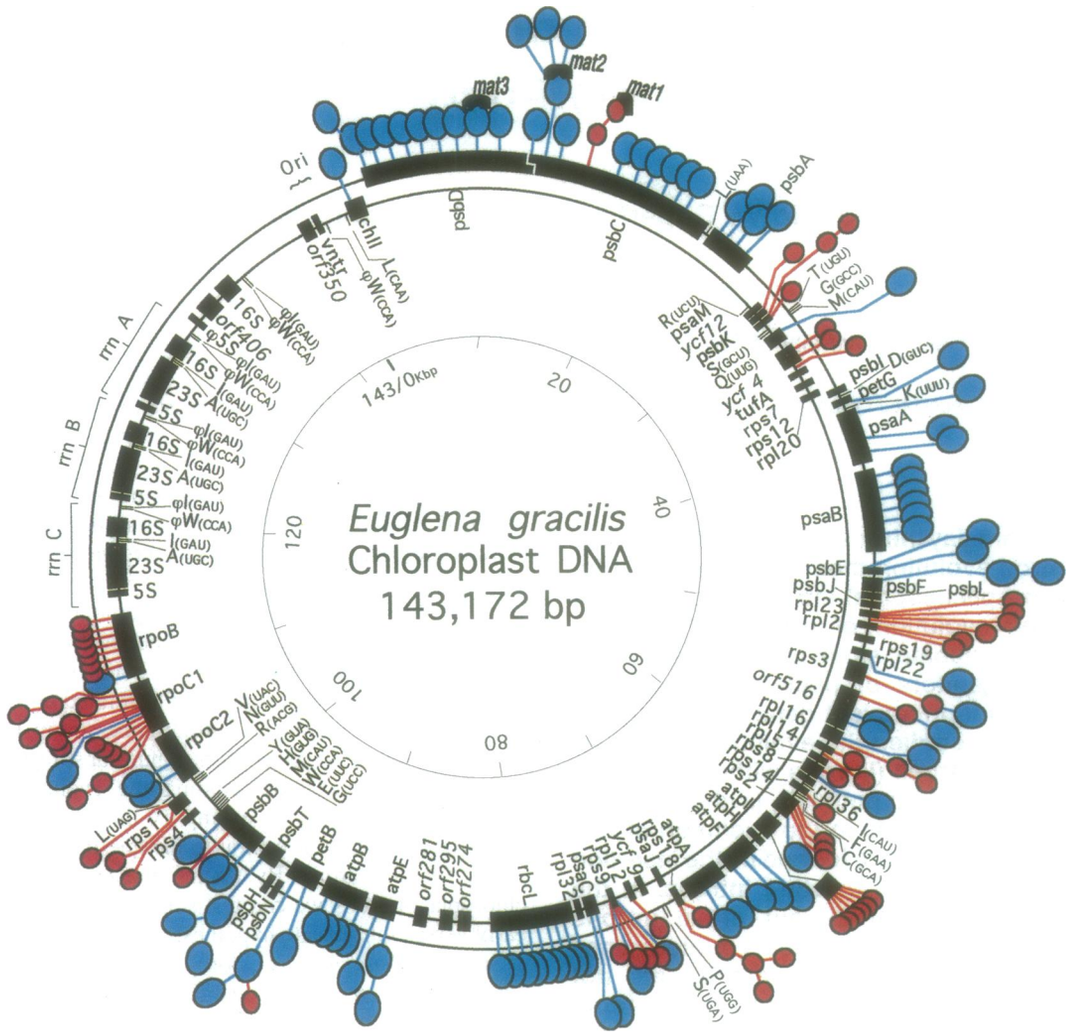

Figure 1. Circular map of *E.gracilis* chloroplast DNA (adapted from 44). The locations of introns are shown. Blue lollipops represent group II introns. Red lollipops represent group III introns. One lollipop inserted into another represents a twintron.

to those genes from derived forms. By tracing the evolutionary history of specific introns and twintrons through a portion of the *Euglena* plastid lineage we can infer whether introns are ancestral or derived characters. According to the introns early hypothesis progenitor genes should contain more introns than derived genes. In contrast, our hypothesis is that introns and twintrons are a derived trait and that twintron precursors (consisting of fewer internal introns or an external intron only) may occur in genes from basally branching *Euglena* spp. Two genes, *rbcL* and *psbT*, were chosen for detailed analysis. In *E.gracilis rbcL* is interrupted by nine group II introns (3) and *psbT* contains a complex twintron, with two group II introns internal to a third group II intron (20).

MATERIALS AND METHODS

Euglena cultures

The following cultures of Euglenophycea were obtained from the culture collection of algae at the University of Texas at Austin (UTEX): *Euglena anabaena* (UTEX 373), *Euglena geniculata* var. terricola (UTEX 366), *Euglena mutabilis* (UTEX 364), *Euglena myxocylindracea* (UTEX 1989), *Euglena pisciformis* var. *typica* (UTEX 1604), *Euglena stellata* (UTEX 372), *Euglena viridis* (UTEX 85), *Euglena gracilis* var. *Z* strain (UTEX 753), *Cryptoglena pigra* (UTEX LB 571) and *Eutreptia* spp. (UTEX LB 2003).

Nucleic acid isolation

Total nucleic acid was extracted from each species. Cells contained in one loop (for cultures maintained on solid media) or 1.5 ml (for cultures maintained in liquid media) were concentrated by centrifugation and resuspended in 300 µl NTES, pH 7.5 (0.1 M NaCl, 0.01 M Tris–HCl, 1 mM Na_2EDTA, 1% SDS). Resuspended cells were extracted twice with phenol/chloroform/isoamyl alcohol (25:24:1). DNA was ethanol precipitated and resuspended in 500 µl water.

cDNA synthesis, DNA amplification and sequencing

The synthetic oligonucleotide primer C1 (5′-CCAAACTTAAC-AAGCGGCAGC-3′) was used for cDNA synthesis of mature RNA encoding ribulose bisphosphate carboxylase/oxygenase large subunit (*rbcL*). cDNA reactions were carried out using 5 µl

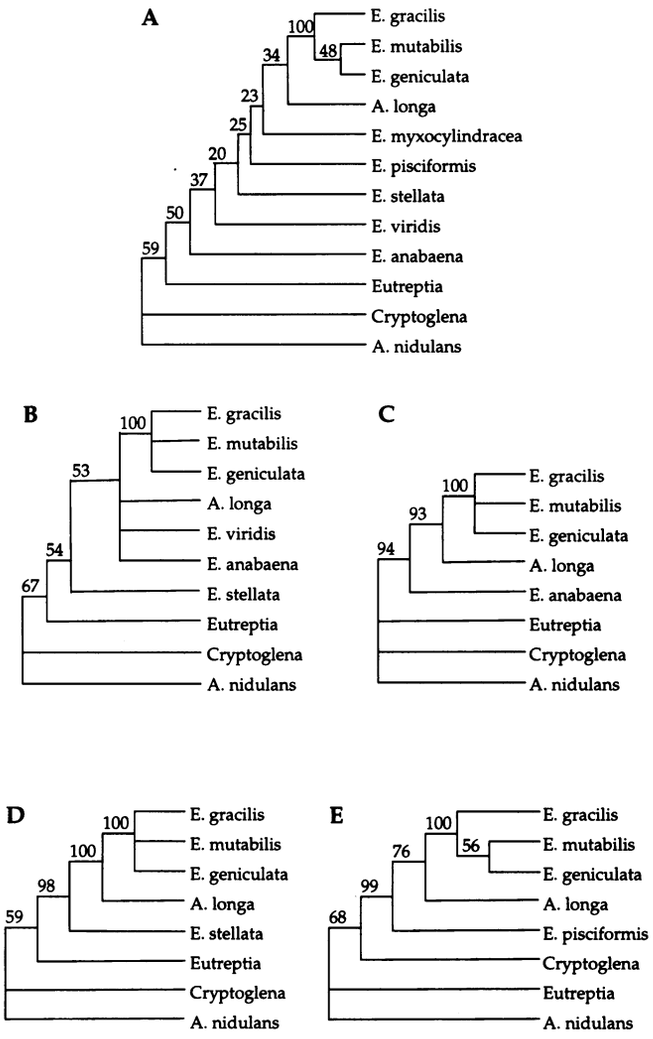

Figure 2. *Euglena* phylogeny. *Euglena* phylogenys based on aligned ribulose bisphosphate carboxylase/oxygenase large subunit coding sequences are shown. Bootstrap analysis using the branch and bound algorithm was performed on all (**A**) or subsets (**B–E**) of the taxa in this study. Results of bootstrap analyses are shown at each node.

total nucleic acid extract as described by Coppertino and Hallick (21). cDNAs were amplified by the polymerase chain reaction (PCR) using the primers C1 and P1 (5′-AAAACTGGA(G/A)-CTG(G/A)ATTTAAAGC-3′). PCR amplification of gene segments was performed according to Copertino and Hallick (21) using 1 µl total nucleic acid extract.

PCR products containing *rbcL* coding sequence were cloned into the plasmid vector pBluescript KS+ (Stratagene) as described by Holton and Graham (26). *rbcL* clones were sequenced on both strands by the chain termination method described by Sanger *et al.* (27) using a series of synthetic oligonucleotide primers: C1, C2 (5′-CAACGTAAGCATCACGC-3′), C2A (5′-GGCATTGT-CCCACCCATACC-3′), C3 (5′-CTCTTCGCAAGTACCTGC-3′), C3A (5′-GCATTTAGATAATGTCC-3′), C4 (5′-GCGCAAATC-TTCTAAACG-3′), C4A (5′-CCAAAAAGTTTTGATATAA-GC-3′), P1, P2 (5′-CTTTTTGAAGAAGGTTCGG-3′), P3 (5′-GCGTTGGAGAGATCGTTTC-3′) and P4 (5′-CGTATGTC-TGGTGGTGATC-3′).

The *Euglena rbcL* cDNA sequences have been submitted to the GenBank DNA sequnce databank. The following accession nos have been assigned: *E.anabaena*, U21004; *E.geniculata*, U21005; *E.mutabilis*, U21006; *E.myxocylindracea*, U21007; *E.pisciformis*, U21008; *E.stellata*, U21009; *E.viridis*, U21010. Sequences for *Astasia longa* (GenBank accession no. X16004) and *Anacystis nidulans* (GenBank accession no. J01536) were obtained from the GenBank nucleotide sequence database.

Phylogeny determination

Phylogenetic trees were produced using the Phylogenetic Analysis Using Parsimony program (PAUP version 3.1.1; 28) and the distance matrix-based program NJ boot. Of the *rbcL* coding region 1296 nt were aligned using Pileup (29). Of the 1296 nt aligned 357 were phylogenetically informative. Bootstrap analysis was based on 100 replicates using the branch and bound algorithm. *Anacystis nidulans* was used as an outgroup.

RESULTS

rbcL cDNA cloning and sequencing

Synthetic oligonucleotide primers were used to synthesize cDNA copies of mature mRNA encoding ribulose bisphosphate carboxylase/oxygenase large subunit (*rbcL*) from seven different *Euglena* spp. and two additional Euglenophycea. cDNAs were amplified by PCR. The amplified cDNAs from *E.anabaena*, *E.geniculata* var. terricola, *E.mutabilis*, *E.myxocylindracea*, *E.pisciformis* var. *typica*, *E.stellata* and *E.viridis* extend from P1 to C1 (Fig. 3). Two amplified cDNAs from *Cryptoglena pigra* extended from P1 to C4 and P4 to C1. *rbcL* coding sequence from *Eutreptia* spp. was derived from an amplified cDNA extending from P1 to C4. Each amplified PCR product was cloned and sequenced on both strands.

The *rbcL* nucleotide and predicted amino acid sequences for all genera in this study are co-linear, completely alignable without insertions or deletions. The amino acid sequence is very highly conserved, ranging from 97 to 99% identical among the *Euglena* spp. The percent identical nucleotides between the *Euglena* species ranges from 80 to 99.8%. *rbcL* sequences from *E.gracilis*, *E.geniculata* and *E.mutabilis* are nearly identical, with at least 99.6% of the coding positions sharing identical nucleotides. *rbcL* sequences from these species are between 82.9 and 86.3% identical to the remaining *Euglena* spp. and ~67% identical to *A.nidulans rbcL*. These relationships are consistent with the parsimony-based *Euglena* phylogeny described below.

Euglena phylogeny

To establish an evolutionary context for inferring the state (ancestral versus derived) of *Euglena* chloroplast introns a *Euglena* chloroplast phylogeny was produced. The phylogeny was based on 1296 nt of *rbcL* coding sequence. *rbcL* coding sequences were aligned using Pileup (29). This alignment formed the basis for inferring phylogenetic trees. The consensus (majority rule) of three equally and most parsimonious trees (length 1044) resulting from a branch and bound search by the Phylogenetic Analysis Using Parsimony program (PAUP version 3.1.1; 28) is shown in Figure 2A. The tree produced by the distance matrix method using the neighbor joining algorithm is generally consistent with the parsimony-based tree. The same topology for *E.gracilis*, *E.geniculata*, *E.mutabilis* and *A.longa* is

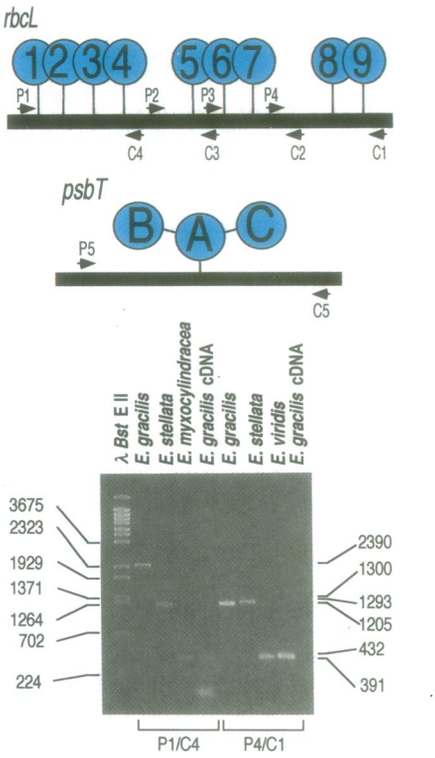

Figure 3. *rbcL* and *psbT* primer locations and representative PCR data. (**A**) Schematic diagrams of the *E.gracilis rbcL* and *psbT* genes are shown. Black boxes represent exons. White and hatched boxes represent introns. Locations of PCR primers (P1–P5) and cDNA primers (C1–C5) are indicated. (**B**) Representative data. The oligonucleotide combinations P1/C4 and P4/C1 were used to amplify two gene segments from *E.gracilis*, *E.stellata*, *E.myxocylindracae*, *E.viridis* and from a cloned *E.gracilis rbcL* cDNA. PCR reaction products were separated in agarose gels and visualized by ethidium bromide staining.

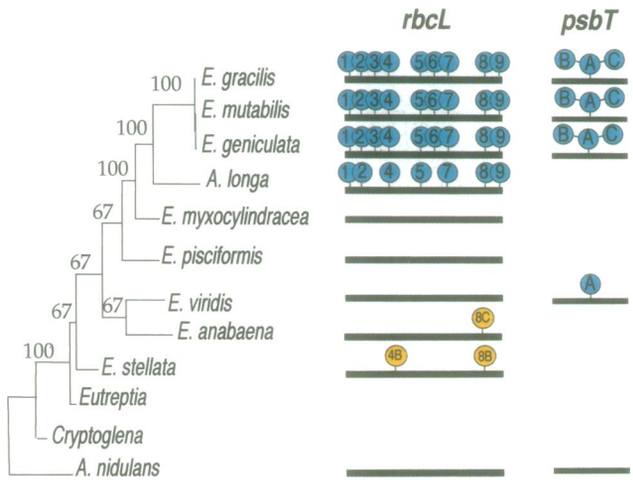

Figure 4. *rbcL* and *psbT* intron evolution in *Euglena*. The tree is the consensus (majority rule) of three equally and most parsimonious trees resulting from a branch and bound search by the Phylogenetic Analysis Using Parsimony program (PAUP version 3.1.1; 28). Branch frequencies among the three most parsimonious trees are shown at each node. The intron contents of *rbcL* and *psbT* are shown on the right. Black boxes represent coding region. Introns are indicated by numbered or lettered lollipops.

produced by both methods. In the neighbor joining tree *E.pisciformis* and *E.myxocylindracea* are grouped and *E.anabaena* is the earliest branching species, followed by *E.stellata*, *E.viridis* and the *E.myxocylindracea–E.pisciformis* group. According to the parsimony-based tree the *E.stellata* branch is the most basal, followed by the branch containing *E.viridis* and *E.anabaena* and then the *E.pisciformis* and *E.myxocylindracea* branches. Topological differences for *E.pisciformis*, *E.myxocylindracea*, *E.stellata*, *E.anabaena* and *E.viridis* do not affect the general conclusions on the evolution of introns.

A high degree of homoplasy in the data results in relatively weak bootstrap support for several nodes of the tree shown in Figure 2A. In order to support the overall structure of the phylogeny bootstrap analysis was performed on subsets of the genera (Fig. 2B–E). Improved overall support, by bootstrap analysis, was achieved by removing *E.pisciformis* and *E.myxocylindracea* from the analysis (Fig. 2B). Additional support for the basal positions of the *E.stellata*, *E.anabaena* and *E.pisciformis* branches and the intermediate position of the *A.longa* branch relative to the *E.gracilis* branch was achieved by the bootstrap analyses shown in Figure 2C–E.

Based on the phylogenies shown in Figure 2, *E.gracilis*, *E.geniculata* and *E.mutabilis* are very closely related and represent the most derived branch. *Astasia longa* is most closely related to this group. The branches containing *E.pisciformis*, *E.myxocylindracea*, *E.anabaena*, *E.viridis* and *E.stellata* are basal relative to the *A.longa* and *E.gracilis* branches and contain terminal taxa that are as divergent from one another as each is to *E.gracilis*. These evolutionary relationships are also supported by parsimony analysis of an intron-encoded open reading frame, *mat1*, from *E.gracilis*, *A.longa*, *E.myxocylindacea* and *E.viridis* (Thompson, Doetsch and Hallick, unpublished observation) and a cyanobacterium *Anabaena* spp. (GenBank accession no. U13767).

rbcL intron content

To investigate intron evolution introns of the *rbcL* gene were studied. *Euglena gracilis* has nine *rbcL* introns (3). The *rbcL* gene of *A.longa* (a closely related non-photosynthetic euglenoid) contains seven introns, corresponding exactly in location to introns 1, 2, 4, 5 and 7–9 of *E.gracilis* (30).

The distribution of *rbcL* introns in seven species of *Euglena* was determined. The presence of introns was initially inferred based on the size of PCR-amplified gene segments. The completely sequenced *E.gracilis* genes served as (+) intron controls. Cloned cDNAs from each species served as (–) intron controls. A schematic diagram of the *E.gracilis rbcL* gene and the locations of oligonucleotide primers is shown in Figure 3. Representative data from these experiments are also shown in Figure 3. In *E.gracilis* the P1/C4-amplified gene segment contains introns 1–4 and the P4/C1-amplified gene segment contains introns 8 and 9. In *E.myxocylindracae* and *E.viridis* these amplified gene segments co-migrated with those derived from the cloned *E.gracilis* cDNA. Based on these data, *rbcL* introns 1–4, 8 and 9 are absent in *E.myxocylindracae* and *E.viridis*. This conclusion was confirmed by sequencing across the sites of intron insertion. In *E.stellata* the P1/C4-amplified gene segment is intermediate in size between *E.gracilis* and *E.myxocylindracae*. The absence of introns 1–4 and the presence of a new intron 4B, located between

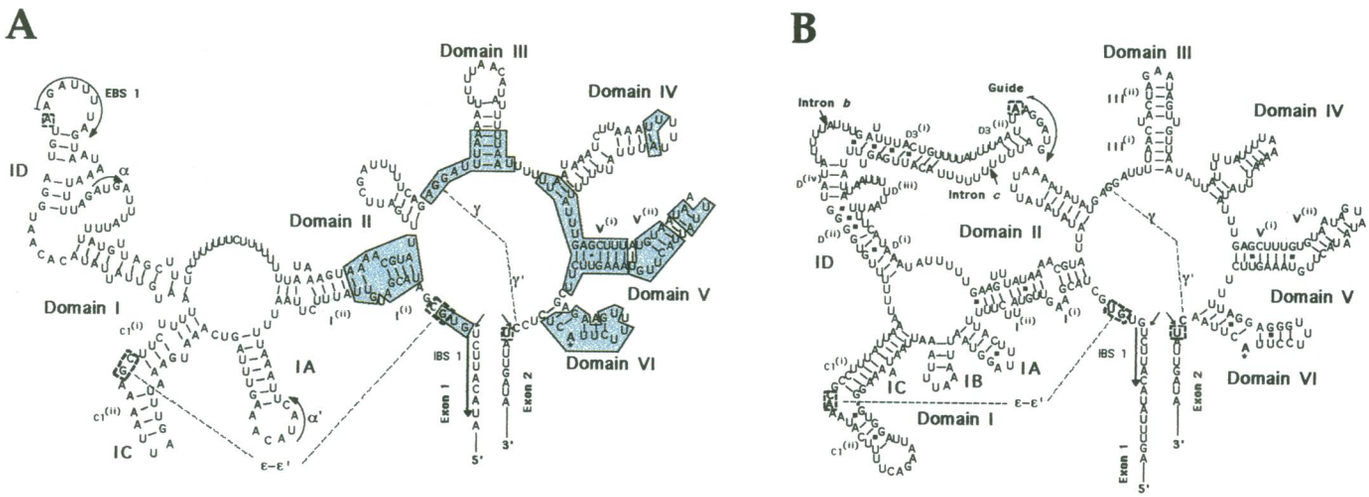

Figure 5. Secondary structure models for *psbT* introns. (**A**) *Euglena viridis psbT* intron. (**B**) *Euglena gracilis psbT* intron A. Conserved nucleotides are boxed in (A). Dashed lines and dashed boxes indicate tertiary interactions (γ–γ′, ε–ε′ and the guided pair). The branch site A is marked with an asterisk (*). Nucleotides involved in the EBS1–IBS1 and the α–α′ interactions are marked with arrows.

codons 84 and 85 in *E.stellata*, was confirmed by DNA sequence analysis. Similarly, introns 8 and 9 are missing and a new intron 8B is present within *E.stellata* codon 364.

Euglena gracilis, *E.mutabilis* and *E.geniculata* contain the same number of introns in the same positions within the *rbcL* gene. *Euglena myxocylindracea*, *E.pisciformis* and *E.viridis* completely lack *rbcL* introns. *Euglena stellata* and *E.anabaena* contain one (*E.anabaena*) or two (*E.stellata*) *rbcL* introns that do not correspond in position to one another or to any of the *rbcL* introns in *E.gracilis*.

psbT cloning, sequencing and intron content

To investigate twintron evolution the *psbT* gene (formerly called *ycf8*; 20) was studied. *psbT* encodes a photosystem II protein (31). In *E.gracilis* the *psbT* gene contains a complex twintron consisting of two group II introns inserted into a third group II intron (20).

A segment of the *psbT* gene from three different *Euglena* spp. was amplified by PCR. The amplified gene segments extending from P5 to C5 (Fig. 3) were cloned and completely sequenced. The cloned *psbT* products from *E.gracilis, E.mutabilis* and *E.geniculata* were identical in size, being 1433 nt long. This 1433 nt long segment is composed of 23 nt of *psbT* exon 1, a 1352 nt twintron and 58 nt of *psbT* exon 2. The twintron includes a 359 nt external group II intron and two internal group II introns of 601 and 392 nt. The DNA sequence of the *psbT* gene segment is 99% identical among these three species. In *E.viridis* the PCR-amplified *psbT* gene segment is 402 nt long. This 402 nt gene segment is made up of 23 and 58 nt of *psbT* exons 1 and 2, respectively, and a 321 nt intron. The *E.viridis* intron and the *E.gracilis* twintron interrupt their respective *psbT* coding regions in the same position, between the first and second nucleotides of the ninth codon.

Comparison of intron secondary structures from *E.viridis* and *E.gracilis*

Since the single group II intron of *E.viridis psbT* is located in the identical position to the external group II intron of the *E.gracilis psbT* twintron, the possibility that both of these introns may have evolved from a common ancestor was further explored by comparative secondary structure analysis. Figure 5 shows a comparison of the secondary structural models for the *psbT* intron A from *E.gracilis* and the corresponding intron from *E.viridis*. The overall lengths of these introns are similar, with *E.gracilis* being 359 nt and *E.viridis* being 321 nt. All of the major secondary structural domains are conserved. The primary sequence of much of the core and the base of each stem, except for domain II, are nearly identical. The percent identities between the *psbT* intron domains from *E.viridis* and *E.gracilis* are shown in Table 1. Domains I–III are 44–53% identical, slightly lower than that for the entire length of the intron (63%). The domains IV are 68% identical. Domain IV is not typically well conserved. The high degree of conservation here is most likely due to the high AU content of this domain. The domains V are 91% identical.

DISCUSSION

The *E.gracilis rbcL* gene evolved from an intronless ancestor

The most parsimonious explanation for the distribution of *rbcL* introns in the genus *Euglena* is that these introns were absent in the ancestral genes and were added during the evolution of this lineage (Fig. 4). *rbcL* introns 1, 2, 4, 5 and 7–9 were acquired after the divergence of *A.longa* or a closely related ancestor from the main *Euglena* line. Introns 3 and 6 of *rbcL* were subsequently acquired after divergence of the ancestor of the *E.gracilis*, *E.mutabilis*, *E.geniculata* clade from a common ancestor with *A.longa*. All of the basally branching species in our study lack this

set of *rbcL* introns. Three additional events, giving rise to the three different *rbcL* introns in *E.stellata* and *E.anabaena*, most likely occurred after the divergence of these species from the main *Euglena* line. Based on this 'introns late' scenario, 12 intron gain events are required to account for this distribution of *rbcL* introns.

Table 1. Percent identities between *E.viridis psbT* intron A and *E.gracilis* introns

Intron	Identity (%)
Intron A	63
Core	70
Domain I	53
Domain II	44
Domain III	52
Domain IV	68
Domain V	91
Domain VI	65
Average of 96 domains V from *E.gracilis*	56

The intron domains used for comparisons correspond to those illustrated in Figure 5. In each comparison the *E.viridis psbT* intron or intron domain was compared with an *E.gracilis* intron or intron domain.

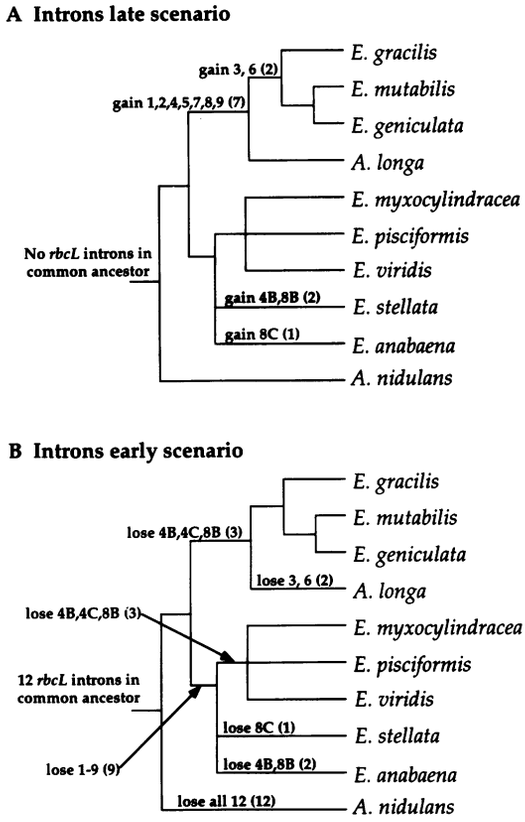

Figure 6. Gain versus loss of *rbcL* introns. A possible phylogeny postulating the least number of intron loss events for *rbcL* introns is shown. The observed distribution of *rbcL* introns among these species is shown in Figure 4. Proposed intron gain events, based on the introns late scenario, are indicated in (**A**). Proposed intron loss events, based on the introns early scenario, are shown in (**B**). The total number of gain (A) or loss (B) events for each branch are shown in parentheses.

The alternative hypothesis is that the *Euglena rbcL* introns are primitive and were retained in only a few species. While possible, this view seems improbable, since more than 12 intron loss events would be necessary to give rise to the present distribution of *rbcL* introns in *Euglena*. Based on the *rbcL* gene phylogeny shown in Figure 2A and an introns early senario, 74 intron loss events would be required to account for the observed *rbcL* intron distribution. Since several branches in that tree are largely unsupported, an alternative phylogeny, postulating the fewest number of intron losses, is shown in Figure 6. Here only 32 loss events (assuming introns early) must occur, compared with 12 gain events (assuming introns late).

Astasia longa shared a common photosynthetic ancestor with *Euglena*

The position of *A.longa* within the *Euglena rbcL* (and *ycf13*) phylogeny may offer new insight into the evolution of *Astasia*. The 73 kbp plastid genome of *A.longa* has been extensively characterized by Hachtel and colleagues (30,32–35). The rRNA and ribosomal protein operons of the plastid genome of this colorless, non-photosynthetic protist are similar in organization to rRNA and ribosomal protein operons of *E.gracilis*, but the *Astasia* plastid genome appears to lack all photosynthesis-related genes except *rbcL*. Based on the phylogeny shown in Figure 2, *A.longa* could be re-classified within genus *Euglena*. *Astasia* and the *Euglena* spp. in this study appear to have shared a common, photosynthetic ancestor. A secondary loss of photosynthetic ability and photosynthesis-related genes has occurred specifically in the *Astasia* lineage.

The first *Euglena* chloroplast introns may have carried genes for maturase-related proteins

The pattern of fewer introns in basal compared with derived branches of the *Euglena* lineage is not restricted to the *rbcL* and *psbT* genes. Figure 7 shows a comparison of the intron content of six different genes from *E.gracilis* and *E.viridis*. This intron content summary includes results based on PCR analysis, similar to that described for *rbcL*. Twenty six of the 40 introns that are distributed among the *rbcL*, *psbF*, *psbT*, *petB*, *psbC* and *psbD* genes in *E.gracilis* are shown. Only four of these 26 introns are present in *E.viridis*, the terminal taxon of a basal branch in the *Euglena* lineage. Two of these introns form a group III twintron that encodes a maturase-like protein, *mat1*, in the internal intron (36). *ycf13* is located within the internal intron of this twintron (23). Maturases are required *in vivo* for splicing in yeast mitochondria (11). The group III twintron in *psbC* and the *mat1* locus are present in five other *Euglena* spp., including *E.geniculata*, *E.mutabtlis*, *E.myxocylindracae*, *E.pisciformis* and *E.viridis* (Thompson, Doetsch and Copertino, unpublished data), representing basal and derived branches of the *Euglena* lineage. A free-standing ORF that is 55.5% identical to *E.gracilis mat1* is

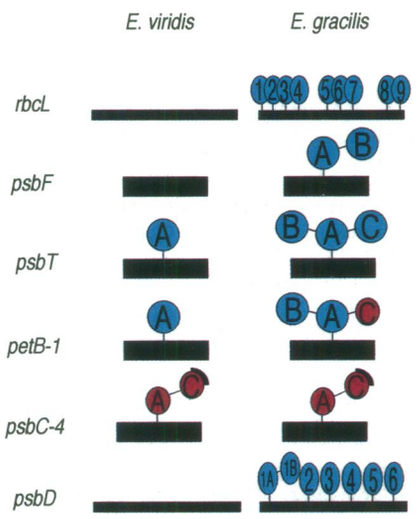

Figure 7. Occurrence of introns in six different genes from *E.viridis* and *E.gracilis*. Black boxes indicate coding regions. Group II introns are shown as blue lollipops and group III introns are shown as red lollipops. The black box on the perimeter of *psbC* intron C represents an intron-encoded open reading frame, *ycf 13*.

present in the plastid DNA of *A.longa* (35). *Astasia longa* also contains group III introns in its plastid genome. The phylogenetic distribution of *mat1* and group III introns is consistent with the hypothesis that the encoded protein may be important for group III intron excision or intron mobility (23). An intriguing possibility is that one or more intron-encoded genes were harbored by early intron invaders of the *Euglena* plastid lineage. These intron-encoded proteins may have then played key roles in the proliferation of introns in this lineage by promoting intron mobility. To see if there is a correlation between intron-encoded ORFs and an increased number of plastid introns, we are currently examining other intron-encoded ORFs and expanding our study to a broader range of euglenoids.

The *E.viridis psbT* intron: an evolutionary building block to twintron formation

The *psbT* intron A is absent in *A.nidulans* and all other known *psbT* genes, but present in both basal and derived branches of the *Euglena* lineage. Because the single *psbT* group II intron in *E.viridis* is present in the same position in the *psbT* coding sequence as the external *psbT* intron (intron A) of *E.gracilis*, most likely it is derived from a common ancestor with the external intron of the *E.gracilis psbT* twintron. The acquisition of internal introns, forming the *psbT* twintron, must have occurred in a common ancestor of *E.gracilis*, *E.geniculata* and *E.mutabilis*, since no internal *psbT* introns are present in *E.viridis*, whose immediate ancestor is basal to this group. This conclusion is supported by direct comparison of the secondary structures and primary sequences of this intron from *E.viridis* and *E.gracilis*.

Group II intron domain V: an indicator of direct ancestry among *Euglena* chloroplast introns

The secondary structural model for group II introns proposed by Michel (16), described recently as a central wheel from which radiate six spokes that define six major ribozyme domains (37), was originally based on a comparative analysis and sequence alignment. The secondary structures of group II introns are evolutionarily conserved across all mitochondrial and chloroplast taxa. These structures vary considerably in evolutionary rate. Learn *et al.* (38) studied the rate of evolution of each domain of the group II intron in the gene encoding $tRNA^{Val}$ (UAC) from seven plant taxa. Domain II is evolving at the highest rate, exceeding the rate of evolution of protein-encoding chloroplast genes. Domain V is evolving at ~1/10 the rate of domain II. This is slower than protein-encoding plastid genes. Domain V is the most slowly evolving domain, followed by domain VI. Domains V and VI are also essential for intron splicing (39,40).

Evolutionary rates similar to those found in the plant $tRNA^{Val}$ (UAC) intron are seen in the *Euglena psbT* intron A (see Table 1). The most stiking conservation (91% identical) is seen between the *psbT* intron A domains V from *E.gracilis* and *E.viridis*. As expected, this is substantially higher than the identity between any of the other domains. The average identity between the *E.viridis psbT* intron domain V and the domains V from 99 different *E.gracilis* group II introns is 56%. The exceptionally high degree of conservation between the *psbT* domains V from *E.viridis* and *E.gracilis*, as well as the strong similarity between their cores, is strongly supportive of these introns arising from a common ancestor. The overall conservation of secondary structure and the relatively high degree of conservation at the nucleotide level of the central core and domains V and VI and the closures of domains I and III (Fig. 5) between these homologous introns from *E.viridis* and *E.gracilis* lend credibility to secondary structure models that are consistent with the structures of group II introns from other organisms. Further comparative analysis is required to decipher conserved features in distal segments of intron domain I.

The late acquisition of chloroplast introns is not limited to genus *Euglena*

The introns late view is supported by an analysis of the distribution of chloroplast introns from divergent organisms (reviewed in 41). The genes from land plant and green algal chloroplasts and from cyanobacteria, three major lineages, contain highly variable numbers of introns. Only one group II intron has been found in cyanobacteria (42). Two additional prokaryotic group II introns occur in *Azotobacter vinelandii* (42) and *Escherichia coli* (43). The extreme variation in intron content among different chloroplast lineages is illustrated by *Cyanophora paradoxa* and *E.gracilis*. In the nearly complete sequence of the *C.paradoxa* cyanelle genome only a single intron has been found. In contrast, the *E.gracilis* chloroplast genome contains 155 introns (82 are group II) (44). Only one chloroplast intron, a group I intron in the gene encoding $tRNA^{Leu}$, is shared between the land plant, green algal and *C.paradoxa* lineages. This single intron is also present in cyanobacteria and was most likely present in the progenitor of all of these chloroplast lineages. The remaining chloroplast introns were most likely acquired after divergence of these lineages.

Since protists are ancient eukaryotes, protist intron evolution may be highly relevant to our understanding of intron origins. In the present study we find that introns in six different photosynthesis-related genes appear to be a derived characteristic, absent from the same gene in one or more basally branching *Euglena* spp.

REFERENCES

1 Gilbert,W., Marchionni,M. and McKnight,G. (1986) *Cell*, **46**, 151–154.
2 Darnell,J.E. and Doolittle,W. F. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 1271–1275.
3 Gingrich,J.C. and Hallick,R.B. (1985) *J. Biol. Chem.*, **260**, 16156–16161.
4 Cavalier-Smith,T. (1991) *Trends Genet.*, **7**, 145–148.
5 Palmer,J. D. and Logsdon,J.M.,Jr (1991) *Curr. Opin. Genet. Dev.*, **1**, 470–477.
6 Schopf,J.W. (1993) *Science*, **260**, 640–646.
7 Dorit,R.L. and Gilbert,W. (1991) *Curr. Opin. Genet. Dev.*, **1**, 464–469.
8 Cech,T.R. (1986) *Cell*, **44**, 207–210.
9 Sharp,P.A. (1985) *Cell*, **42**, 397–400.
10 Roger,J.A. and Doolittle,W.F. (1993) *Nature*, **364**, 289–290.
11 Saldanha,R., Mohr,G., Belfort,M. and Lambowitz,A.M. (1993) *FASEB J.*, **7**, 15–24.
12 Sontheimer,E.J. and Steitz,J.A. (1993) *Science*, **262**, 1989–1996.
13 Michel,F. and Dujon,B. (1983) *EMBO J.*, **2**, 33–38.
14 Jacquier,A. and Michel,F. (1990) *J. Mol. Biol.*, **213**, 437–447.
15 Jarrell,K.A., Dietrich,R.C. and Perlman,P.S. (1988) *Mol. Cell. Biol.*, **8**, 2361–2366.
16 Michel,F., Umesono,K. and Ozeki,H. (1989) *Gene*, **82**, 5–30.
17 Madhani,H.D. and Guthrie,C. (1992) *Cell*, **71**, 803–817.
18 Wise,J.A. (1993) *Science*, **262**, 1978–79.
19 Copertino,D.W. and Hallick,R.B. (1993) *Trends Biochem. Sci.*, **18**, 467–471.
20 Hong,L. and Hallick,R.B. (1994) *Genes Dev.*, **8**, 1589–1599.
21 Copertino,D.W. and Hallick,R.B. (1991) *EMBO J.*, **10**, 433–42.
22 Copertino,D.W., Christopher,D.A. and Hallick,R.B. (1991) *Nucleic Acids Res.*, **19**, 6491–6497.
23 Copertino,D.W., Van Hook,F.W., Hall,E.T., Jenkins,K.P. and Hallick,R.B. (1994) *Nucleic Acids Res.*, **22**, 1029–1036.
24 Copertino,D.W., Shigeoka,S. and Hallick,R.B. (1992) *EMBO J.*, **11**, 5041–5050.
25 Copertino,D.W. (1992) PhD dissertation, University of Arizona.
26 Holton,T.A. and Graham,M.W. (1991) *Nucleic Acids Res.*, **19**, 1156.
27 Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
28 Swofford,D.L. (1993) *PAUP: Phylogenetic Analysis Using Parsimony.* Illinois Natural History Survey, Champaign, IL.
29 Genetics Computer Group (1994) *Program Manual for the Wisconsin Package, Version 8.* University of Wisconsin, Madison, WI.
30 Siemeister,G. and Hachtel,W. (1990) *Plant Mol. Biol.*, **14**, 825–833.
31 Monod,C., Takahashi,T., Goldschmidt-Clermont,M. and Rochaix,J.D. (1994) *EMBO J.*, **13**, 2747–2754.
32 Gockel,G., Hachtel,W., Baier,S., Fliss,C. and Henke,M. (1994) *Curr. Genet.*, **26**, 256–262.
33 Siemeister,G., Buchholz,C. and Hachtel,W. (1990) *Curr. Genet.*, **18**, 457–64.
34 Siemeister,G. and Hachtel,W. (1990) *Curr. Genet.*, **17**, 433–8.
35 Siemeister,G., Buchholz,C. and Hachtel,W. (1990) *Mol. Gen. Genet.*, **220**, 425–432.
36 Mohr,G., Perlman,P.S. and Lambowitz,A.M. (1993) *Nucleic Acids Res.*, **21**, 4991–4997.
37 Michel,F. and Jean-Luc,F. (1995) *Annu. Rev. Biochem.*, **64**, 435–461.
38 Learn,G.H.J., Shore,J.S., Furnier,G.R., Zurawski,G. and Clegg,M.T. (1992) *Mol. Biol. Evol.*, **9**, 856–871.
39 Padgett,R.A., Podar,M., Boulanger,S.C. and Perlman,P.S. (1994) *Science*, **266**, 1685–1688.
40 Chanfreau,G. and Jacquier,A. (1994) *Science*, **266**, 1383–1387.
41 Palmer,J.D. (1991) In Bogorad,L. and Vasil,I.K. (eds), *Molecular Biology of Plastids.* Academic Press, San Diego, CA, pp. 5–53.
42 Ferat,J.-L. and Michel,F. (1993) *Nature*, **364**, 358–361.
43 Knoop,V. and Brennicke,A. (1994) *Nucleic Acids Res.*, **22**, 1167–1171.
44 Hallick,R.B., Hong,L., Drager,R.G., Favreau,M.R., Monfort,A., Orsat,B., Spielmann,A. and Stutz,E. (1993) *Nucleic Acids Res.*, **21**, 3537–3544.