A Possible Role for Short Introns in the Acquisition of Stroma-Targeting Peptides in the Flagellate *Euglena gracilis*

**Matej Vesteg**¹, **Rostislav Vacula**¹, **Jürgen M. Steiner**², **Bianka Matešíková**¹, **Wolfgang Löffelhardt**³, **Brona Brejova**⁴, and **Juraj Krajčovič**¹,*

Institute of Cell Biology and Biotechnology, Faculty of Natural Sciences, Comenius University, Mlynská dolina, 842 15 Bratislava, Slovakia¹; Pflanzenphysiologie, Institut für Biologie MLU, Halle-Wittenberg, Weinbergweg 10, Halle/Saale 06120, Germany²; Max F. Perutz Laboratories, Department of Biochemistry and Cell Biology, University of Vienna, Doktor-Bohr-Gasse 9, 1030 Vienna, Austria³ and Department of Computer Science, Faculty of Mathematics, Physics, and Informatics, Comenius University, Mlynská Dolina, 824 48 Bratislava, Slovakia⁴

*To whom correspondence should be addressed. Tel. +421 2 602-96-657. Fax. +421 2 602-96-288.
Email: krajcovic@fns.uniba.sk

Edited by Satoshi Tabata

(Received 9 February 2010; accepted 18 May 2010)

**Abstract**

The chloroplasts of *Euglena gracilis* bounded by three membranes arose via secondary endosymbiosis of a green alga in a heterotrophic euglenozoan host. Many genes were transferred from symbiont to the host nucleus. A subset of *Euglena* nuclear genes of predominately symbiont, but also host, or other origin have obtained complex presequences required for chloroplast targeting. This study has revealed the presence of short introns (41–93 bp) either in the second half of presequence-encoding regions or shortly downstream of them in nine nucleus-encoded *E. gracilis* genes for chloroplast proteins (Eno29, GapA, PetA, PetF, PetJ, PsaF, PsbM, PsbO, and PsbW). In addition, the *E. gracilis* Pbgd gene contains two introns in the second half of presequence-encoding region and one at the border of presequence-mature peptide-encoding region. Ten of 12 introns present within presequence-encoding regions or shortly downstream of them identified in this study have typical eukaryotic GT/AG borders, are T-rich, 45–50 bp long, and pairwise sequence identities range from 27 to 61%. Thus single recombination events might have been mediated via these cis-spliced introns. A double crossing over between these cis-spliced introns and trans-spliced introns present in 5'-UTRs of *Euglena* nuclear genes is also likely to have occurred. Thus introns and exon-shuffling could have had an important role in the acquisition of chloroplast targeting signals in *E. gracilis*. The results are consistent with a late origin of photosynthetic euglenids.

**Key words:** exon-shuffling; chloroplast-targeting; presequence; secondary endosymbiosis

1. Introduction

*Euglena gracilis* belongs to the order Euglenida, the protist phylum Euglenozoa, and the eukaryotic super-group Excavata. The phylum Euglenozoa includes also the orders Kinetoplastida (including suborders Trypanosomatina and Bodonina) and Diplonemida. The monophyly of Euglenozoa has been suggested based on various common morphological features, e.g. discoidal mitochondrial cristae and a characteristic feeding apparatus,¹,² and on molecular phylogenies.³ Moreover, Euglenozoa share the presence of the modified base ‘J’ in the nuclear DNA.⁴ There is little evidence for the presence of signalling pathways regulating nuclear gene expression at the transcriptional level.⁵,⁶ The addition of non-coding capped spliced-leaders to nuclear pre-mRNAs via trans-splicing is also common among Euglenozoa.⁷–¹²

*Euglena gracilis* and other phototrophic euglenids possess chloroplasts surrounded by three
membranes. These arose by a secondary endosymbiotic event in which an euglenozoan host engulfed a green alga. Chlorarachniophytes (belonging to the supergroup Rhizaria) possess complex green plastids with four envelope membranes and nucleomorph, obtained via an independent secondary symbiosis. While plastids of euglenids descended from a prasinophyte, chlorarachniophyte plastids most likely descended from an ulvophyte green algal endosymbiont. Many *Euglena* nuclear genes, mostly of symbiont (i.e. resulting from endosymbiotic gene transfer from the nucleus of the primary host cell to the nucleus of the secondary host cell), but also of host or other origin have acquired presequences for chloroplast targeting. Most presequences required for chloroplast import in *Euglena* are tripartite, comprising in order: N-terminal signal peptide for targeting to ER, the S/T-rich region resembling transit peptides of organisms possessing primary plastids, and the stop-transfer sequence serving as a membrane anchor (class I pro- teins, comprising also thylakoid-lumen-targeted class IB proteins possessing an additional hydrophobic thylakoid transfer domain). Therefore, the major part of the protein precursor stays 'outside' while passing through ER, Golgi apparatus, and membrane vesicles prior to their fusion with the outermost chloroplast membrane. A recent in-depth analysis of *E. gracilis* presequences revealed another set, the class II of nucleus-encoded plastid protein precur- sors. These lack the putative stop-transfer sequence and possess only a signal sequence at the N-terminus, followed by a transit-peptide-like sequence.

The complete sequence of the *E. gracilis* chloroplast genome disclosed an unusually high number of introns: groups II and III introns, and even twintrons (introns within introns). However, little is known about introns in nuclear genes of euglenids, as only few genomic sequences from euglenids are available. Introns in the *E. gracilis* Lhcbm1 gene (according to the nomenclature of Koziol and Durnford, encoding light-harvesting chlorophyll *a/b* binding protein of photosystem II), *RbcS* genes (encoding small subunit of RuBisCo), and *GapC* (encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase) do not possess consensus splicing borders (5'-GT/AG-3') and structural characteristics of group I and II introns, and many of them are flanked by short direct repeats. These introns can form secondary structure, which could potentially bring together 5'- and 3'-ends, probably without the involvement of spliceosomes. However, *E. gracilis* contains also canonical introns, e.g. the 16 introns of the *TubC* genes (two gene copies encoding gamma-tubulin) or the introns in the fibrillarin gene. The 5'-ends of these introns can potentially base pair with U1 snRNA, suggesting that they are excised in a spliceosome-dependent manner. Introns with GT/AG borders are present also in the beta-tubulin gene of the non-photosynthetic euglenoid flagellate *Entosiphon sulcatum*. Furthermore, introns in *E. gracilis* *TubA* and *TubB* genes (encoding alpha- and beta-tubulin, respectively) are of conventional as well as of non-conventional type. Recombination events and exon-shuffling have been discussed by various authors as possibly involved in the addition of sequences encoding transit peptides (mito- chondrial targeting signals) to nuclear genes for mito- chondrial proteins. In an analogous manner, sequences encoding stroma-targeting peptides might have been added to nucleus-encoded genes for chloro- plast proteins in organisms (Archaeplastida) possessing primary chloroplasts of cyanobacterial origin. Such exon-shuffling could occur via recombination processes mediated by introns. However, the identification of introns originally involved in exon-shuffling is problematic for nuclear genes encoding mitochondrial proteins, and for nuclear genes for proteins targeted to primary chloroplasts. The mitochondria arose via an alpha-proteobacterial endosymbiosis, which perhaps dates back to the origin of eukaryotes, and the cyanobacterial ancestry of primary plastids dates back to the origin of the Archaeplastida. Since then many intron integration/ excision events occurred in various lineages making it almost impossible to identify introns, which were ancestrally involved in the acquisition of transit peptides. However, the secondary chloroplasts are the results of relatively recent endosymbioses of red and green algae in eukaryotic hosts (for reviews see refs 41–45). It has been suggested that recombination processes might have led to addition of presequences (or at least their parts) to nuclear genes for chloroplast proteins in organisms possessing secondary plastids. Perhaps the best evidence so far for the involvement of recombination processes mediated by introns in the acquisition of presequences and/or their parts came from the study of Kilian and Kroth, which revealed the presence of a single intron either within the presequence region or shortly downstream of it in seven nucleus-encoded genes for plastid proteins (AtpC, FbaC1, PetJ, PsbM, PsbO, PsbU and Tpt1) in the diatom *Phaeodactylum tricornutum* possessing four-membrane-bounded plastids of red algal origin. In this study, we decided to extend this hypothesis to the flagellate *E. gracilis* possessing secondary chloro- plastids of green algal origin.

2. Materials and methods

*Euglena gracilis* (Pringsheim strain Z, SAG 1224–5/25 Collection of Algae, Göttingen, Germany) was
cultivated in 100 ml Erlenmeyer flasks containing 50 ml of a modified Cramer and Myers medium \(^4\) supplemented with ethanol (0.8%) and adjusted to pH 6.9. Medium was inoculated with \(5 \times 10^4\) cells per ml. Cells were grown at 27°C with continuous illumination (30 \(\mu\)mol photons \(\text{m}^{-2} \text{s}^{-1}\)). Cultures in the exponential growth phase were used for DNA isolation.

The protocol for genomic DNA isolation was used as described in the chapter 2.3.1. (Preparation of Genomic DNA from Plant Tissue) of Current Protocols in Molecular Biology \(^5\) with following modification: cells were harvested by centrifugation at \(100 \times g\) (3 min), then washed twice with ice-cold ddH2O, and resuspended with buffer (100 mM Tris–Cl, pH 8; 100 mM EDTA, pH 8; 250 mM NaCl) containing 8 \(\mu\)l of protease K (Merck, 20 mg/ml) per 1 ml of buffer. 20% \(N\)-lauroylsarcosine (Sigma) was added and the mixture was incubated in water-bath at 55°C for 1 h. After the steps of extractions, centrifugation (6000 \(\times g\), 30 min, 4°C), DNA precipitation (2-propanol), centrifugation (7500 \(\times g\), 15 min, 4°C) and solubilization (TE buffer, pH 8), RNA was removed (RNase A, 15 min). Thereafter, phenol:chloroform (1:1) and chloroform:isoamylalcohol (24:1) extractions were performed each followed by centrifugation (7500 \(\times g\), 7 min). One-tenth volume of 3 M sodium acetate (pH 5.2) was added to the top phase, and DNA was precipitated with 96% ethanol at \(-20^\circ\text{C}\), centrifuged (8000 \(\times g\), 15 min, 4°C) and washed (70% ethanol). DNA was resuspended in the TE buffer (pH 8).

Primers were derived from six \(E.\) gracilis nuclear mRNA sequences encoding chloroplast proteins. Table 1 contains the accession numbers of these mRNAs (see refs \(^{19, 26, 51–54}\)) and the corresponding positions of primer sequences. Another four pairs of primers were derived from four \(E.\) gracilis nuclear EST sequences (see ref. \(^{22}\)) encoding chloroplast proteins: PetF (ferredoxin), PsAf subunit of photosystem I, and the PsbM and PsbW subunits of photosystem II. All these four ESTs possessed SL-leader sequence (TTTTTTTTCG) at the 5’-end, and were used in previous analysis of presequences of \(E.\) gracilis. \(^{22}\) Table 2 contains the \(e\)-values, accession numbers of these ESTs used for the design of primers, and the positions corresponding to primer sequences in these ESTs.

Primers were designed using Primer-BLAST (primer 3 and BLAST) to obtain similar melting temperature (60°C) for all primers. The effort was made to design primers such as to be able to amplify the whole presequence-encoding region and short part downstream of it (or as long part of this region as possible following our stringent primer design criteria).

The PCRs were performed in 50-\(\mu\)l reaction volume with the final concentration of Mg \(^2+\), primers and dNTPs as 2 mM, 0.2 \(\mu\)M and 0.5 mM, respectively. 100 ng of total \(E.\) gracilis DNA and 2.5 Units of Taq DNA polymerase (Invitrogen) were used per reaction. Samples were denatured by heating for 5 min at 94°C, subjected to 34 cycles of 30 s denaturation at 94°C, 1 min annealing at 58°C, and 2 min extension at 72°C, and a final cycle of 8 min at 72°C. PCR products were visualized by 1% agarose gels (TAE), purified using QIAquick PCR Purification Kit (Qiagen), and sequenced twice (using forward as well as reverse primers) using ABI 3130xl Genetic Analyzer (Applied Biosystems) and the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) according to suppliers’ protocols. The services of the Department of Molecular Biology (Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia) were used for sequencing of PCR products.

The sequence data were analyzed using Chromas, BLAST and CLUSTALW. Sequence identity of the intron sequences was computed by the global alignment (Needle tool from the EMBOSS suite with the default settings). \(^{55}\) Since the introns have unusual nucleotide composition, which may have inflated the scores, the statistical significance of each alignment score was computed by a permutation test. For each pair of introns, 100 000 random permutations of their bases were computed and compared to the observed alignment score.

### Table 1. List of primers derived from \(E.\) gracilis mRNA sequences

| mRNA\(^a\) | Reference\(^b\) | Accession\(^c\) | Forward primers\(^d\) | Reverse primers\(^e\) |
|---------|----------------|----------------|---------------------|---------------------|
| Eno29   | 51             | AJ272112       | 65–84               | 298–279             |
| GapA    | 26             | L21904         | 23–42               | 468–449             |
| Pbgd    | 52             | X15743         | 183–202             | 511–492             |
| PetA    | 53             | AF443625       | 49–68               | 422–403             |
| PetP    | 19             | AJ130725       | 89–108              | 386–367             |
| PsbO    | 54             | D14702         | 40–59               | 674–655             |

\(\text{a}\)Primers were derived from mRNA sequences of nucleus-encoded genes (\(E\)no29, \(\text{Gap}\)A, \(\text{Pbgd}\), \(\text{Pet}\)A, \(\text{PetP}\) and \(\text{Psbo}\)) for chloroplast proteins (eno1ase, glyceraldehyde-3-phosphate dehydrogenase, porphobilinogen deaminase, cytochrome \(f\), cytochrome \(c_{6}\), and 30 kDa protein of the oxygen-evolving complex, respectively).

\(\text{b}Number of reference in the reference list in which the corresponding mRNA was characterized.

\(\text{c}Accession numbers of mRNAs.

\(\text{d}The numbers of primers correspond to the positions in mRNA sequences that can be found under the accession numbers (accession number) listed in the third column. For example, forward primer 65–84 (first row, fourth column) is identical to positions 65–84 of \(E\)no29 mRNA sequence, which can be found under accession number AJ272112 and reverse primers 298–279 (first row, fifth column) is complementary to the sequence 279–298 of \(E\)no29 mRNA.
were aligned, and the empirical distribution of scores was computed. Sequences were permuted by Shuffleseq from the EMBOSS suite, and the consensus splice sites (GT/AG) were kept in their original position in each permutation.

3. Results and discussion

The PCR products amplified using all primers were listed in Tables 1 and 2 (except those for Pbgd, PsbO, and PsbW) and total E. gracilis DNA as a template were about 50 bp longer than those expected for cDNA templates. In the cases of Pbgd, PsbO, and PsbW, PCR products were about 150, 90, and 250 bp longer, respectively.

Sequencing of seven PCR products revealed that each contained one 41–50 bp intron. The Pbgd PCR product contained three introns (48, 46, and 50 bp), the PsbO PCR product contained one 93 bp intron, and the PsbW PCR product contained two introns (48 and 195 bp). It is noteworthy, that the 195 bp psbW-i2 intron is present downstream of the stop codon in the 3'-UTR of PsbW gene. Thus the total number of introns identified in this study was 13. Except for the PsbO, PetJ, and the intron present in PsbW 3'-UTR, all introns are 45–50 bp in size and contain typical eukaryotic GT/AG consensus splicing borders (see Table 3 which also includes the accession numbers of the partial gene sequences containing introns identified in this study).

It was impossible to determine borders and phase of the 93 bp-long intron in the PsbO gene, because it does not contain consensus borders, and TG sequence is present in mRNA, but also on both

---

**Table 2.** List of primers derived from *E. gracilis* EST sequences

| EST producta | Accession numberb | Organism with the best BLASTX hit | E-value | Forward primersc | Reverse primersd |
|--------------|-------------------|-----------------------------------|---------|-----------------|-----------------|
| PetF         | EG565162          | *Euglena viridis*                 | 100E-40 | 57–76           | 627–608          |
| PsaF         | EG565174          | *Chlamydomonas reinhardtii*       | 900E-36 | 30–49           | 499–480          |
| PsbM         | EG565161          | *Ostreococcus tauri*              | 700E-11 | 85–104          | 575–556          |
| PsbW         | EG565140          | *Bigelowiella natans*             | 300E-14 | 68–86           | 560–543          |

aThe name of protein product of ESTs from which primers were derived. PetF is plastid-targeted ferredoxin, PsaF is subunit F of photosystem I, and PsbM and PsbW are subunits M and W of photosystem II.
bAccession numbers of ESTs.
c,dThe numbers of primers correspond to the positions in corresponding ESTs. E.g. forward primer 57–76 (first row, fifth column) is identical to positions 57–76 of EST with accession number EG56162, and reverse primer 627–608 (first row, sixth column) is complementary to nucleotides 627–608 of this EST.

**Table 3.** Introns in *E. gracilis* nucleus-encoded genes for chloroplast proteins identified in this study

| Intron   | Accession number | Length | Borders | Percent AT | Percent T | Nucleotide position | Phase |
|----------|------------------|--------|---------|------------|-----------|---------------------|-------|
| eno29-i1 | GQ925702         | 48     | GT/AG   | 62.50      | 37.50     | 166                 | 1     |
| gapA-i1  | GQ925704         | 50     | GT/AG   | 54.00      | 44.00     | 243                 | 0     |
| pbgd-i1  | GQ925705         | 48     | GT/AG   | 60.42      | 33.33     | 276                 | 1     |
| pbgd-i2  | GQ925705         | 46     | GT/AG   | 67.39      | 45.65     | 377                 | 0     |
| pbgd-i3  | GQ925705         | 50     | GT/AG   | 68.00      | 46.00     | 462                 | 1     |
| petA-i1  | GQ925706         | 45     | GT/AG   | 64.44      | 37.78     | 261                 | 0     |
| petF-i1  | GQ925703         | 46     | GT/AG   | 63.04      | 47.83     | 423                 | 1     |
| petJ-i1  | GQ925707         | 41     | GT/TC or TT/CG | 63.41 | 34.15 | 304 or 305 | 1 or 2 |
| psaF-i1  | GQ925708         | 47     | GT/AG   | 72.34      | 44.68     | 307                 | 0     |
| psbM-i1  | GQ925709         | 47     | GT/AG   | 55.32      | 38.30     | 411                 | 1     |
| psbO-i1  | GQ925710         | 93     | ?       | 68.82      | 44.09     | 517, 518 or 519 ?   | ?     |
| psbW-i1  | GQ925711         | 48     | GT/AG   | 64.58      | 39.58     | 198                 | 0     |
| psbW-i2  | GQ925711         | 195    | GA/GT   | 54.36      | 31.79     | 505                 | 1     |

All introns, except for psbW-i2, are present either in the presequence-encoding regions or shortly upstream of them. The table includes accession numbers of partial gene sequences containing corresponding introns, intron length (in nucleotides), intron borders, AT- and T-content of introns, and intron phase. Nt position is the position downstream of which the intron is inserted into the corresponding mRNA or EST sequence (for the accession numbers of mRNAs and ESTs see Tables 1 and 2).
intron–exon borders. The splicing borders of intron in PsbO may be TG/TT, GA/TT or AC/TG. Similar problems with the determination of intron borders have been described for the Lhcb gene-encoding LHCPII protein, and for GapC-encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase, because the introns in these genes are flanked by short direct repeats (2–5 bp) and do not possess consensus splicing borders. The 41 bp intron in the PetJ gene also does not show consensus splicing borders. A guanine nucleotide is present on both its intron–exon borders, thus its splicing borders might be GT/TC or TT/CG, and it is either in phase 1 or 2.

With the exception of the 195 bp intron in PsbW (psbW-i2, with GA/GT splicing borders and no direct repeat on intron–exon borders), all introns identified in this study are present either within the second half of the presequence-encoding region or shortly downstream of it. The 48 bp intron in Eno29 gene (eno29-i1) is present between the amino acid positions 166 and 167 of Eno29 mRNA (accession number AJ272112), while presequence-encoding region ends with the position 161 of this mRNA sequence. The 50 bp gapA-i1 was localized between the codons for aa 90 and 91 of the 127 aa GapA presequence. Sharif et al. reported a 139 aa presequence for Pbgd, whereas Durnford and Gray predicted a length of 151 aa. The C-terminus of the presequence region accounts for the difference between these two studies: the 48 bp pbgd-i1 is present in the codon for aa 85, the 46 bp pbgd-i2 is inserted between the codons for aa 119 and 120, and the 50 bp pbgd-i3 localizes to the codon for aa 144 of Pbgd preprotein (i.e. either within the end of the presequence region or shortly downstream of it). The 45 bp petA-i1 is present downstream of the codon 87 of the 147 aa presequence region. The 46 bp petF-i1 was found to be inserted into the codon-specifying aa 131 of the 138 aa presequence region of the EST-encoding ferredoxin (accession number EG565162). The 41 bp petf-i1 localizes downstream of either nt 304 or 305 of PetF partial mRNA sequence (accession number AJ130725), with nt 267 representing the end of the presequence-encoding region. The predicted 144 aa Psaf presequence harbours the 47 bp psaf-i1 downstream of codon 94. The 47 bp psbm-i1 is inserted into codon 131 of the PsbM presequence region (predicted to 154 aa). The 93 bp psbo-i1 was identified about 60 nt downstream of the PsbO presequence-encoding region. Finally, the 48 bp intron psbw-i1 localizes between the codons 66 and 67 of the predicted 82 aa Psbw presequence.

Taken together, in this study, 13 new intron sequences present in E. gracilis nuclear genes encoding chloroplast proteins have been described. In genes encoding chloroplast-targeted proteins Eno29, GapA, PetA, PetF, PetJ, Psaf, Psbo, Psbm and Psbw, one intron has been identified within the second half of presequence-encoding region or shortly downstream of it, while in gene encoding Pbgd, two introns were identified within the presequence and one at the presequence-mature peptide border encoding region. Importantly, the BLAST search revealed no significant primary sequence similarity of the introns identified in this study to either introns present in the E. gracilis chloroplast genome, or to any introns from other organisms in public databases.

Ten of 13 introns identified in this study are conventional, and are 45–50 bp long. Introns of similar size (44–53 bp) have been already described in some other E. gracilis nuclear genes, while some of them are conventional. The only shorter introns in euglenoid species known so far are three introns (27, 29 and 31 bp-long) present in hsp90 gene of the phagotrophic euglenid Peranema trichorum. Of these, only one is conventional. Nevertheless, it should be mentioned that E. gracilis introns can widely vary in size, and the largest one identified so far is the conventional intron i1 (9.2 kb) in one of the two copies of the gamma-tubulin gene.

Interestingly, the E. gracilis nuclear gene encoding chloroplast protein Rbcs also contains an intron within the second half of presequence region. The size of this intron is 53 bp, it is in phase 0, but does not possess GT/AG borders. In the nuclear gene Lhcb (Lhcbm1), a 86 bp intron roughly separates presequence and mature peptide coding regions. This intron is also non-conventional, and it is impossible to determine its phase due to TG dinucleotide present on both intron–exon borders. Likewise, the 93 bp intron in the PsbO presequence is also flanked by TG dinucleotide and shares 46% primary sequence identity with the 86 bp intron in Lhcb.

Importantly, 10 of 14 E. gracilis introns known to be present in the second half of presequence-encoding regions or shortly downstream of them share various characteristic features: the length (45–50 bp), consensus GT/AG splicing borders, they are AT- and especially T-rich, and possess characteristic pyrimidine tracks at the 3’-ends. Moreover, the primary sequence identity of each two of these 10 introns ranges from 27 to 61% (Table 4). Notably, the 44 and 46 bp introns of conventional type present in the E. gracilis fibrillarin gene share 58% primary sequence identity, and the primary sequence identity of these 2 introns and 10 45–50 bp introns found in this study ranges from 32 to 55% (Table 4). Although not all alignment scores are statistically significant (Table 4), the sequence similarity together with other characteristics of these 44–50 bp E. gracilis introns suggests that recombination
events between these introns can potentionally occur. In comparison, conventional introns present within or shortly downstream of presequence regions of nuclear-encoded plastid proteins from the diatom *Phaeodactylum tricornutum* are 183–410 bp long and their pairwise sequence comparison did not reveal significant sequence similarity.48

Kilian and Kroth48 suggested ‘semi-exon shuffling’ as a possible mechanism for the acquisition of presequence parts (e.g. signal peptides) in diatoms. The intron present within the presequence-encoding region of the donor gene might have recombined either with 5'-UTR of acceptor gene or with its transit peptide (likely transferred from the red algal symbiont host to the host nucleus with the acceptor gene), while new 3’-AG intron border in the acceptor gene might have been generated by utilizing random AG nucleotides.48 However, the primary sequence similarity of 10 45–50 bp introns present either in the presequence-encoding regions or shortly upstream of them. The primary sequence identity was calculated as the number of identical nucleotide oppositions of two introns in a pairwise alignment divided by the length of the alignment. Statistically significant alignments with *P*-value ≤0.05 are shown in bold (see section Methods).

**Table 4.** Primary sequence identity (top-right half) and alignment *P*-value (bottom-left half) of the selected introns from *E. gracilis* (44–50 bp long, with consensus GT/AG borders)

|        | eno29-i1 | gapA-i1 | pbgd-i1 | pbgd-i2 | pbgd-i3 | petA-i1 | petF-i1 | psaF-i1 | psbM-i1 | psbW-i1 | nop1p-i1 | nop1p-i3 (%) |
|--------|----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|
| eno29-i1 | 45%      | 60%     | 27%     | 42%     | 36%     | 61%     | 56%     | 55%     | 47%     | 55%     | 42%     |           |
| gapA-i1  | 0.07     | 46%     | 37%     | 52%     | 49%     | 49%     | 39%     | 54%     | 55%     | 45%     | 44%     | 37%       |
| pbgd-i1  | 0.01     | 0.39    | 55%     | 40%     | 43%     | 51%     | 53%     | 49%     | 42%     | 46%     | 32%     |           |
| pbgd-i2  | 0.32     | 0.35    | 0.20    | 55%     | 33%     | 53%     | 46%     | 29%     | 47%     | 41%     | 41%     |           |
| pbgd-i3  | 0.22     | 0.03    | 0.93    | 0.23    | 42%     | 47%     | 50%     | 37%     | 53%     | 37%     | 47%     |           |
| petA-i1  | 0.20     | 0.02    | 0.25    | 0.27    | 0.25    | 38%     | 56%     | 34%     | 46%     | 38%     | 38%     |           |
| petF-i1  | 0.01     | 0.21    | 0.15    | 0.02    | 0.90    | 0.63    | 44%     | 55%     | 46%     | 48%     | 47%     |           |
| psaF-i1  | 0.33     | 0.15    | 0.05    | 0.74    | 0.13    | 0.03    | 0.05    | 38%     | 40%     | 46%     | 39%     |           |
| psbM-i1  | 0.03     | 0.09    | 0.16    | 0.83    | 0.37    | 0.28    | 0.11    | 0.91    | 47%     | 50%     | 39%     |           |
| psbW-i1  | 0.01     | 0.06    | 0.20    | 0.34    | 0.05    | 0.24    | 0.26    | 0.07    | 0.27    | 50%     | 46%     |           |
| nop1p-i1 | 0.04     | 0.00    | 0.26    | 0.44    | 0.83    | 0.28    | 0.36    | 0.02    | 0.01    | 0.35    | 58%     |           |
| nop1p-i3 | 0.38     | 0.31    | 0.34    | 0.95    | 0.58    | 0.65    | 0.23    | 0.62    | 0.25    | 0.93    | 0.04    |           |

Except for nop1p-i1 and nop1p-i3 (introns present in the gene-encoding nucleolar protein fibrillarin), all these introns are present either in the presequence-encoding regions or shortly upstream of them. The primary sequence identity was calculated as the number of identical nucleotide oppositions of two introns in a pairwise alignment divided by the length of the alignment. Statistically significant alignments with *P*-value ≤0.05 are shown in bold (see section Methods).
the second half or shortly downstream of *Euglena* pre-sequence-encoding regions is indicative of a relatively recent acquisition of chloroplast-targeting signals in *Euglena*. This is consistent with, and adds additional support for a relatively recent origin of euglenoid secondary plastids, later than the endosymbiosis of the evolutionarily ancient red algae leading to diatoms. Anyway, the repertoire for creating novel targeting sequences or for replacing the transit sequences from the primary host cell by bi- or tripartite presequences did already exist. This applies for the α-proteobacterial endosymbiosis leading to mitochondria and the above-mentioned secondary endosymbiosis leading to chromophytes, respectively: exon-shuffling at the DNA level *via* appropriately placed introns enabling recombination. Our data suggest that euglenoids also made use of this mechanism, probably as the last in a row.

Although nuclear gene sequence data of euglenoids are fragmentary, it seems that nuclear genes of euglenids possess many cis-spliced introns. In contrast, wide-scale genome data from parasitic kinetoplastids are available, but very few cis-spliced introns from trypanosomes were reported so far, including a 11 bp intron in the gene for tRNA(tyr) of *Trypanosoma cruzi* and *Trypanosoma brucei*,61 and 653 and 302 bp introns in the gene for poly(A) polymerase of *T. brucei* and *T. cruzi*, respectively.62 One might argue that almost complete loss of cis-spliced introns in trypanosomes arose through parasitic life style, as did the overall compaction of nuclear genomes of trypanosomes including fairly short intergenic spacers with polycistronic transcription63,64 and overlapping genes.65 However, cis-spliced introns seem to be rare in both parasitic and free-living kinetoplastids, and this general condition could pre-date the adoption of parasitism by the trypanosomatid lineage.66 The euglenid lineage with numerous cis-spliced introns—as opposed to the kinetoplastid lineage—likely was better pre-adapted for the acquisition of chloroplast-targeting presequences, and thus for the successful integration of an algal symbiont.

**Funding**

This work was supported by grants from the Ministry of Education of the Slovak Republic (VEGA 1/0416/09, to J. K.; and VEGA 1/0118/08, to R. V.), by grants from Comenius University, Bratislava, Slovakia (UK/144/2007, and UK/208/2009 to M. V.), and by grant P19683 from the Austrian ‘Fonds zur Förderung der wissenschaftlichen Forschung’ to W. L.

**References**

1. Triemer, R.E. and Farmer, M.A. 1991, An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids, *Protoplasma*, 164, 91–104.
2. Simpson, A.G.B. 1997, The identity and composition of the Euglenozoa, *Arch. Protistenkd.*, 148, 318–328.
3. Simpson, A.G.B. and Roger, A.J. 2004, Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa, *Mol. Phylogenet. Evol.*, 30, 201–212.
1. Campbell, D.A., Thomas, S. and Sturm, N.R. 2003, Analysis of Euglena gracilis plastid-targeted proteins reveals different classes of transit sequences, *Eukaryot. Cell*, 5, 2079–2091.

2. Dornford, D.G. and Gray, M.W. 2006, Analysis of Euglena gracilis plastid-targeted proteins reveals different classes of transit sequences, *Eukaryot. Cell*, 5, 2079–2091.

3. Muchhal, U.S. and Schwartzbach, S.D. 1994, Characterization of the unique intron-exon junctions of Euglena gene(s) encoding the polypeptide precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II, *Nucleic Acids Res.*, 22, 5737–5744.

4. Henze, K., Badr, A., Wetter, M., Cerff, R. and Martin, W. 1995, A nuclear gene of eubacterial origin in Euglena gracilis reflects cryptic endosymbioses during protist evolution, *Proc. Natl. Acad. Sci. U.S.A.*, 92, 9122–9126.

5. Tessier, L.H., Keller, M., and Imbault, P. 1999, Structure and expression of Euglena gracilis nuclear rbcS genes encoding the small subunit of the ribulose 1,5-bisphosphate carboxylase/oxygenase: A novel splicing process for unusual intervening sequences, *J. Mol. Biol.*, 245, 22–33.

6. Canaday, J., Tessier, L.H., Imbault, P. and Paulus, F. 2001, Analysis of Euglena gracilis alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation, *Mol. Genet. Genomics*, 265, 153–160.

7. Breckenridge, D.G., Watanabe, Y-I., Greenwood, S.J., Gray, M.W. and Schnare, M.N. 1999, U1 small nuclear RNA and spliceosomal introns in Euglena gracilis, *Proc. Natl. Acad. Sci. USA.*, 96, 852–856.

8. Russell, A.G., Watanabe, Y., Charette, J.M. and Gray, M.W. 2005, Unusual structure of fibrillarin cDNA and gene structure in Euglena gracilis: evolutionary conservation of core proteins and structural predictions for methyl-ation-guide box C/D snoRNPs throughout the domain Eukarya, *Nucl. Acids Res.*, 33, 2731–2791.

9. Wischmann, C. and Schuster, W. 1995, Transfer of rps10 from the mitochondrion to the nucleus in Arabidopsis thaliana: evidence for RNA-mediated transfer and exon shuffling at the integration site, *FEBS Lett.*, 374, 152–156.

10. Long, M., de Souza, S.J., Rosenberg, C. and Gilbert, W. 1996, Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor, *Proc. Natl. Acad. Sci. USA*, 93, 7727–7731.

11. Patthy, L. 1999, Genome evolution and the evolution of exon-shuffling—a review, *Gene*, 238, 103–114.
34. Adams, K.L., Daley, D.O., Whelan, J. and Palmer, J.D. 2002, Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts, Plant Cell, 14, 931–943.
35. Martin, W. and Müller, M. 1998. The hydrogen hypothesis for the first eukaryote, Nature, 392, 37–41.
36. Vesteg, M. and Krajčovič, J. 2008, Origin of eukaryotic cells as a symbiosis of parasitic α-proteobacteria in the periplasm of two-membrane-bounded sexual pre-karyocytes, Commun. Integr. Biol., 1, 104–113.
37. McFadden, G.I. and van Dooren, G.G. 2004, Evolution: red algal genome affirms a common origin of all plastids, Curr. Biol., 14, R514–R516.
38. Rodríguez-Ezepeleta, N., Brinkmann, H., Burey, S.C., et al. 2005, Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes, Curr. Biol., 15, 1325–1330.
39. Jeffares, D.C., Mourier, T. and Penny, D. 2006, The biology of intron gain and loss, Trends Genet., 22, 16–22.
40. Basu, M.K., Rogozin, I.B., Deusch, O., Dagan, T., Martin, W. and Koonin, E.V. 2008, Evolutionary dynamics of introns in plastid-derived genes in plants: saturation nearly reached but slow intron gain continues, Mol. Biol. Evol., 25, 111–119.
41. Cavalier-Smith, T. 2003, Genomic reduction and evolution of novel genetic machinery and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae), Phil. Trans. R. Soc. Lond. B, 358, 109–134.
42. Palmer, J.D. 2003, The symbiotic birth and spread of plastids: How many times and whodunit?, J. Phycol., 39, 4–11.
43. Gould, S.B., Waller, R.F. and McFadden, G.I. 2008, Plastid evolution, Annu. Rev. Plant Biol., 59, 491–517.
44. Keeling, P.J. 2009, Chromalveolates and the evolution of plastids by secondary endosymbiosis, J. Eukaryot. Microbiol., 56, 1–8.
45. Vesteg, M., Vacula, R. and Krajčovič, J. 2009, On the origin of chloroplasts, import mechanisms of chloroplast-targeted proteins, and loss of photosynthetic ability, Folia Microbiol., 54, 303–321.
46. Waller, R.F., Keeling, P.J., Donald, R.G.K., et al. 1998, Nuclear-encoded proteins target to the plastid in Toxoplasma gondii and Plasmodium falciparum, Proc. Natl. Acad. Sci. USA, 95, 12352–12357.
47. Schaap, D., van Poppel, N.F. and Vermeulen, A.N. 2001, Intron invasion in protozoal nuclear encoded plastid genes, Mol. Biochem. Parasitol., 115, 119–121.
48. Kilian, O. and Kroth, P.G. 2004, Presequence acquisition during secondary endosymbiosis and the possible role of introns, J. Mol. Evol., 58, 712–721.
49. Cramer, M. and Myers, J. 1952, Growth and photosynthetic characteristics of Euglena gracilis, Arch. Mikrobiol., 17, 384–403.
50. Ausubel, F.M., Brent, R., Kingston, R.E., et al. 1996, Current Protocols in Molecular Biology, John Wiley & Sons, Inc., Boston, Massachusetts.
51. Hannoert, V., Brinkmann, H., Nowitzki, U., et al. 2000, Enolase from Trypanosoma brucei, from amitochondriate protist Mastigamoeba batamuthi, and from the chloroplast and cytosol of Euglena gracilis: pieces in the evolutionary puzzle of the eukaryotic glycolytic pathway, Mol. Biol. Evol., 17, 989–1000.