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

MATEJ Vesteg1, ROSTISLAV Vacula1, JÜRGEN M. Steiner2, BLANKA Mateášiková1, WOLFGANG Löffelhardt3, BRONA Brejová4, and JURAJ Krajčovič1,*

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 Euglena genes for chloroplast proteins (Eno29, GapA, PetA, PetF, PetJ, PsaF, PsbM, PsbO, and PsbW). In addition, the Euglena 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 Euglena 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 supergroup 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 splice-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 proteins, 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 precursors. 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 (mitochondrial targeting signals) to nuclear genes for mitochondrial proteins. In an analogous manner, sequences encoding stroma-targeting peptides might have been added to nucleus-encoded genes for chloroplast 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 39–45). 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 chloroplasts 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 \(49\) 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 m\(^{-2}\) 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 \(50\) with following modification: cells were harvested by centrifugation at 1000 \(\times\) g (3 min), then washed twice with ice-cold dd\(H_2O\), and resuspended with buffer (100 mM Tris–Cl, pH 8; 100 mM EDTA, pH 8; 250 mM NaCl) containing 8 \(\mu\)l of proteinase 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°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 on 1.5% agarose gels (TAE). 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 \(Eno\)29 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 \(Eno\)29 mRNA.

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

<table>
<thead>
<tr>
<th>mRNA</th>
<th>Reference</th>
<th>Accession number</th>
<th>Forward primers</th>
<th>Reverse primers</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Eno)29</td>
<td>51</td>
<td>AJ272112</td>
<td>65–84</td>
<td>298–279</td>
</tr>
<tr>
<td>GapA</td>
<td>26</td>
<td>L21904</td>
<td>23–42</td>
<td>468–449</td>
</tr>
<tr>
<td>(Pbgd)</td>
<td>52</td>
<td>X15743</td>
<td>183–202</td>
<td>511–492</td>
</tr>
<tr>
<td>(Pta)</td>
<td>53</td>
<td>AF443625</td>
<td>49–68</td>
<td>422–403</td>
</tr>
<tr>
<td>(Ptf)</td>
<td>19</td>
<td>AJ130725</td>
<td>89–108</td>
<td>386–367</td>
</tr>
<tr>
<td>(Psb)</td>
<td>54</td>
<td>D14702</td>
<td>40–59</td>
<td>674–655</td>
</tr>
</tbody>
</table>

\(\dagger\) Primers were derived from mRNA sequences of nucleus-encoded genes (\(Eno\)29, \(GapA\), \(Pbgd\), \(Pta\), \(Ptf\) and \(Psb\)) for chloroplast proteins (enolase, glyceraldehyde-3-phosphate dehydrogenase, porphobilinogen deaminase, cytochrome \(f\), cytochrome \(c\), and 30 kDa protein of the oxygen-evolving complex, respectively).

\(\ddagger\) Number of reference in the reference list in which the corresponding mRNA was characterized.

\(\ddagger\) Accession numbers of mRNAs.

\(\dagger\) 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 \(Eno\)29 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 \(Eno\)29 mRNA.

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 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 3'0-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>
<thead>
<tr>
<th>Intron</th>
<th>Accession number</th>
<th>Length</th>
<th>Borders</th>
<th>Percent AT</th>
<th>Percent T</th>
<th>Nucleotide position</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>eno29-i1</td>
<td>GQ925702</td>
<td>48</td>
<td>GT/AG</td>
<td>62.50</td>
<td>37.50</td>
<td>166</td>
<td>1</td>
</tr>
<tr>
<td>gapA-i1</td>
<td>GQ925704</td>
<td>50</td>
<td>GT/AG</td>
<td>54.00</td>
<td>44.00</td>
<td>243</td>
<td>0</td>
</tr>
<tr>
<td>pbgd-i1</td>
<td>GQ925705</td>
<td>48</td>
<td>GT/AG</td>
<td>60.42</td>
<td>33.33</td>
<td>276</td>
<td>1</td>
</tr>
<tr>
<td>pbgd-i2</td>
<td>GQ925705</td>
<td>46</td>
<td>GT/AG</td>
<td>67.39</td>
<td>45.65</td>
<td>377</td>
<td>0</td>
</tr>
<tr>
<td>pbgd-i3</td>
<td>GQ925705</td>
<td>50</td>
<td>GT/AG</td>
<td>68.00</td>
<td>46.00</td>
<td>462</td>
<td>1</td>
</tr>
<tr>
<td>petA-i1</td>
<td>GQ925706</td>
<td>45</td>
<td>GT/AG</td>
<td>64.44</td>
<td>37.78</td>
<td>261</td>
<td>0</td>
</tr>
<tr>
<td>petF-i1</td>
<td>GQ925703</td>
<td>46</td>
<td>GT/AG</td>
<td>63.04</td>
<td>47.83</td>
<td>423</td>
<td>1</td>
</tr>
<tr>
<td>petj-i1</td>
<td>GQ925707</td>
<td>41</td>
<td>GT/TC or TT/CG</td>
<td>63.41</td>
<td>34.15</td>
<td>304 or 305</td>
<td>1 or 2</td>
</tr>
<tr>
<td>psaF-i1</td>
<td>GQ925708</td>
<td>47</td>
<td>GT/AG</td>
<td>72.34</td>
<td>44.68</td>
<td>307</td>
<td>0</td>
</tr>
<tr>
<td>psbM-i1</td>
<td>GQ925709</td>
<td>47</td>
<td>GT/AG</td>
<td>55.32</td>
<td>38.30</td>
<td>411</td>
<td>1</td>
</tr>
<tr>
<td>psbO-i1</td>
<td>GQ925710</td>
<td>93</td>
<td>?</td>
<td>68.82</td>
<td>44.09</td>
<td>517, 518 or 519</td>
<td>?</td>
</tr>
<tr>
<td>psbW-i1</td>
<td>GQ925711</td>
<td>48</td>
<td>GT/AG</td>
<td>64.58</td>
<td>39.58</td>
<td>198</td>
<td>0</td>
</tr>
<tr>
<td>psbW-i2</td>
<td>GQ925711</td>
<td>195</td>
<td>GA/GT</td>
<td>54.36</td>
<td>31.79</td>
<td>505</td>
<td>1</td>
</tr>
</tbody>
</table>

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).

---

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

<table>
<thead>
<tr>
<th>EST producta</th>
<th>Accession number</th>
<th>Organism with the best BLASTX hit</th>
<th>E-value</th>
<th>Forward primersb</th>
<th>Reverse primersd</th>
</tr>
</thead>
<tbody>
<tr>
<td>PetF</td>
<td>EG565162</td>
<td><em>Euglena viridis</em></td>
<td>100E-40</td>
<td>57–76</td>
<td>627–608</td>
</tr>
<tr>
<td>PsaF</td>
<td>EG565174</td>
<td><em>Chlamydomonas reinhardtii</em></td>
<td>900E-36</td>
<td>30–49</td>
<td>499–480</td>
</tr>
<tr>
<td>PsbM</td>
<td>EG565161</td>
<td><em>Ostreococcus tauri</em></td>
<td>700E-11</td>
<td>85–104</td>
<td>575–556</td>
</tr>
<tr>
<td>PsbW</td>
<td>EG565140</td>
<td><em>Bigelowiella natans</em></td>
<td>300E-14</td>
<td>68–86</td>
<td>560–543</td>
</tr>
</tbody>
</table>

*The 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.
intron–exon borders. The splicing borders of intron in \( PsbO \) may be TG/TG, GA/GA or AC/AC. Similar problems with the determination of intron borders have been described for the \( Lhcb \) gene-encoding LHCP-II protein,\(^{25} \) and for GapC-encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase,\(^{26} \) 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/GT, and it is either in phase 1 or 2.

With the exception of the 195 bp intron in \( PsbW \) (\( psbW-i2 \), with GA/GA 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.\(^{51} \) reported a 139 aa presequence for \( Phgd \), whereas Durnford and Gray\(^{22} \) 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 \( petJ-i1 \) localizes downstream of either nt 304 or 305 of \( PetJ \) 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.\(^{26–29} \) 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.\(^{56} \) Of these, only one is conventional. Nevertheless, it should be mentioned that \( E. gracilis \) introns can widely vary in size,\(^{25–30} \) 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.\(^{28} \)

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.\(^{27} \) In the nuclear gene \( Lhcb (Lhcbm1) \), a 86 bp intron roughly separates presequence and mature peptide coding regions.\(^{25} \) This intron is also non-conventional, and it is impossible to determine its phase due to TG dinucleotide present on both intron–exon borders.\(^{25} \) 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\(^{29} \) 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 potentially 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 Kroth\(^{48}\) 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 been recombined either with 5'‑UTR of acceptor gene or with its transit peptide (likely transferred from the red algal symbiont nucleus 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)

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

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 downstream 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 euglenids also made use of this mechanism, probably as the last in a row.

Although nuclear gene sequence data of euglenids 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(try) of Trypanosoma cruzi and Trypanosoma brucei, and 653 and 302 bp introns in the gene for poly(A) polymerase of T. brucei and T. cruzi, respectively. 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 transcription and overlapping genes. 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. 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

Acquisition of Presequences in *Euglena* [Vol. 17, 230


60. Leander, B.S. 2004, Did trypanosomatid parasites have photosynthetic ancestors?, Trends Microbiol., 12, 251–258.


