Characterization of the unique intron - exon junctions of *Euglena* gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II

Umesh Shyamsunder Muchhal  
*University of Nebraska - Lincoln*

Steven D. Schwartzbach  
*University of Nebraska - Lincoln*

Follow this and additional works at: [http://digitalcommons.unl.edu/bioscifacpub](http://digitalcommons.unl.edu/bioscifacpub)

Part of the [Biology Commons](http://digitalcommons.unl.edu/bioscifacpub)

Muchhal, Umesh Shyamsunder and Schwartzbach, Steven D., "Characterization of the unique intron - exon junctions of *Euglena* gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II" (1994). *Faculty Publications in the Biological Sciences*. 585.  
[http://digitalcommons.unl.edu/bioscifacpub/585](http://digitalcommons.unl.edu/bioscifacpub/585)
Characterization of the unique intron–exon junctions of Euglena gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II

Umesh S. Muchhal and Steven D. Schwartzbach*
School of Biological Sciences, University of Nebraska-Lincoln, Lincoln, NE 68588-0343, USA

Received July 7, 1994; Revised and Accepted November 12, 1994 EMBL accession nos: X61361 and L29644

ABSTRACT

The precursor to the Euglena light harvesting chlorophyll a/b binding protein of photosystem II (LHCPII) is a polyprotein containing multiple copies of LHCPII covalently joined by a decapetide linker. cDNA and genomic clones encoding the 5′ and 3′ end of a 6.6 kb LHCPII mRNA were sequenced. A 3.1 kb genomic region encoding 1.05 kb of the 5′ end of LHCPII mRNA contains 4 introns. A 7.6 kb genomic region encoding 3.3 kb of the 3′ end of LHCPII mRNA contains 10 introns. The 5′ and 3′ ends of the 14 identified Euglena introns lacked the conserved dinucleotides (5′-GT and AG-3′) found at the termini of virtually every characterized nuclear pre-mRNA intron. A common consensus splice site selection sequence could not be identified. The Euglena introns do not have the structural characteristics of group I and group II introns. The only structural feature common to all Euglena introns was the ability of short stretches of nucleotides at the 5′ and 3′ ends of the introns to base pair, forming a stable stem-loop with the 5′ and 3′ splice site juxtaposed for splicing but displaced by 2 nucleotides. The 26 nucleotide sequence at the 5′ end of LHCPII mRNA is absent from the genomic sequence and identical to the 5′ end of one of the small Euglena SL-RNAs indicating that it is post-transcriptionally added by trans-splicing.

INTRODUCTION

The light harvesting chlorophyll a/b binding proteins of photosystem II (LHCPII) are a group of abundant, highly conserved thylakoid proteins. LHCPIIs are encoded by a nuclear multigene family comprising approximately 3–20 members depending upon the organism studied (1). LHCPIIs have been classified into three types based on the amino acids found at 14 positions and the presence or absence of introns within their genes (2). Type I LHCPIIs are encoded by intronless genes while a single intron is present in genes encoding type II LHCPIIs. Type III LHCPIIs can not be classified as type I or II based on amino acid sequence and their genes often contain multiple introns. Euglena gracilis, a unicellular protist, has a LHCPII precursor (pLHCPII) that is a polyprotein containing multiple copies of LHCPII covalently joined by a conserved decapetide linker (3–5). The amino acid sequence of Euglena pLHCPII was deduced from the nucleotide sequence of a genomic clone (GC18) containing 7.4 kb of the 3′ end of a LHCPII gene (4). Nine exons were identified in this clone based on sequence homology with Arabidopsis LHCPII. They can be assembled into a continuous open reading frame encoding 113 amino acids of the C-terminus of an LHCPII, followed by 4 complete LHCPIIs (4). Individual LHCPIIs derived from this polyprotein precursor (1PEP, 114PEP, 351PEP 575PEP and 811PEP) are named relative to their distance from the N-terminal amino acid of the encoded polyprotein (4). The Euglena LHCPIIs are 60–70% homologous to higher plant and green algal LHCPII and contain three hydrophobic membrane spanning α-helical domains as found in other LHCPIIs (4). Maintenance of amino acid co-linearity between Euglena and Arabidopsis LHCPII identified potential intron–exon junctions (4) lacking the highly conserved 5′-GT and AG-3′ intron boundaries found in virtually all eukaryotic organisms (6–8). The choice of alternative intron–exon junctions would have resulted in major amino acid insertions or deletions within otherwise highly homologous co-linear regions. The Euglena nuclear gene (rbcS) encoding the ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit (SSU) also lacks GT-AG intron boundaries (9), suggesting that Euglena uses splice recognition sequences differing from those used by all other organisms.

The LHCPII intron–exon junctions were identified solely through maintenance of amino acid homology in the encoded protein (4). Only five Euglena rbcS introns have been characterized by direct comparison between genomic and cDNA sequences (9). In an attempt to identify a novel consensus splice site for Euglena nuclear pre-mRNA introns, a Euglena genomic clone encoding the 5′ half of an LHCPII polyprotein gene was isolated. RNA-PCR was used to obtain cDNAs corresponding

*To whom correspondence should be addressed
to the transcripts encoded by the *Euglena* LHCP II genomic clones. Direct comparison of the genomic and cDNA sequences identified 14 introns in the two genomic clones. These unique introns do not contain the highly conserved GT and AG dinucleotides present at the 5' and 3' ends, respectively, of almost all the nuclear pre-mRNA introns characterized to date from eukaryotic organisms as diverse as humans, plants and ciliates (6–8). A brief report of this work has appeared (10).

**MATERIALS AND METHODS**

**Isolation and characterization of the LHCP II genomic clone**

The genomic clone, GC7, was obtained by screening a *Euglena* library containing 4×10^5 recombinant phage (4). The 0.68 kb *BstXI* fragment of the genomic clone GC18, encoding most of LHCPII unit 114PEP (4), was used as a probe for plaque hybridizations. The LHCP II hybridizing region of GC7 was subcloned as a 5.2 kb (7E52) *SalI*-EcoRI and a 4.8 kb (7SE48) *EcoRI*-SalI fragment in pBluescript II KS (Fig. 1A). A 3.1 kb region of the plasmid 7E52 was further subcloned in pBluescript II KS as a 1.6 kb *PstI* and an overlapping 2.0 kb *SalI*-BglII fragment (Fig. 1A) for sequence analysis.

**Characterization of *Euglena* LHCP II cDNAs**

The synthesis and subcloning of NLH1, the cDNA clone corresponding to the 5' end of LHCP II mRNA (Fig. 1A), by RNA-PCR has been described previously (11). The cDNAs corresponding to the LHCP II transcript encoded by GC18 (Fig. 1B) were synthesized by RNA-PCR using oligonucleotide primers corresponding to unique regions within each LHCP II type encoded by GC18 (4). First strand cDNA was synthesized from 2 μg of *Euglena* poly A+ RNA using Superscript (GIBCO-BRL) reverse transcriptase as described by the supplier. The PCR amplifications were carried out using *Taq* DNA polymerase (Perkin-Elmer Cetus) for 30 cycles of 1 min at 95°C, 30 s at 56°C and 3 min (last cycle 15 min) at 72°C, essentially as described by the manufacturer. The 0.34 kb cDNA product, LH34 (Fig. 1B), was synthesized using the oligonucleotides 1a (5'-CCGGATTCCTCAGGAAAGGCCCTGTG-3') and 2a (5'-GTGACCACCCCCAAGCCGCCCA-3'). The 1.38 kb cDNA, LH138 (Fig. 1B), was synthesized using the oligonucleotides 1a and 3b (5'-GTGGTCCAGACCCAGCTGGACAAACACAG-3'). The synthesis of a 0.64 kb cDNA, LH64 (Fig. 1B), utilized oligonucleotides 2b (5'-CCGCTCTGCAAACAAATCTTTCAGC-3') and 3b. For the cDNA clones, CLH22 and CLH9 (Fig. 1B), corresponding to the 3' end of the *Euglena* LHCPII transcript, an anchor-(dT)17 oligonucleotide (5'-GGG-GATTCTGTCAGACAGTTTTTTTTTTTTTTTTT-3') was used for first strand synthesis. The amplification reactions contained the anchor oligonucleotide (5'-GGGAATTCCTGACAAGTTTTTTTTTTTTTTTTT-3') and either oligonucleotide 3a (5'-GCTGTCACCCCTGTTGCTTTGAC-3') for CLH22, or 2b for CLH9 (Fig. 1B). LH34 was subcloned as a blunt-end insert, LH138 as an *EcoRI*-SalI insert, LH64 as a blunt end-SalI insert, and both CLH22 and CLH9 as *EcoRI/HindIII*-blunt ended inserts into the polylinker of pBluescript II KS.

**Sequencing, Southern and Northern blots**

The nucleotide sequence of all clones was determined and analyzed as described (4). Southern and Northern blot analysis were performed as described previously (4).

**Primer extension RNA sequencing**

The oligonucleotide Pex2 (5'-CTTATGCGATCGGTTGTTTTTG-3') used for primer extension-RNA sequencing, was gel purified and end labelled with γ-32P-ATP using T4 polynucleotide kinase as described (12). For sequencing, 20 μg of *Euglena* total RNA and 0.05 pmol of labelled oligonucleotide were heated at 90°C for 3 min in 5 μl of buffer containing 50 mM Tris–HCl, pH 8.3 and 50 mM KCl, and quickly transferred to a waterbath for annealing at 68°C. After 10 min, reverse transcription was carried out for 20 min at 45°C in a 20 μl reaction volume containing 50 mM Tris–HCl, pH 8.3, 50 mM KCl, 10 mM MgCl2, 0.5 mM spermidine, 10 mM DTT, 8 U of AMV reverse transcriptase and 0.4 mM of each dNTP. For sequencing, either 0.4 mM ddATP or 0.3 mM ddGTP or 0.6 mM ddTTP or 0.4 mM ddTTP was included in the appropriate tube. Extension reactions were stopped by adding 1 μl of 0.5 M EDTA and 1 μl of DNase-free pancreatic RNAse (100 μg/ml) and incubating at 37°C for 30 min. The reaction mix was phenol:chloroform extracted after adding 100 μl of 3M...
ammonium acetate and the products precipitated with 3 volumes of ethanol. The pellet was dried and resuspended in 10 µl of loading buffer (50% formamide, 10 mM EDTA, bromophenol blue and xylene cyanol). Products were analyzed by electrophoresis through 8% polyacrylamide sequencing gels containing 7 M urea and visualized by autoradiography.

**Figure 2.** Nucleotide sequence of the LHCPII genomic clone GC7, the LHCPII cDNA clone NLH1 (11) and the deduced amino acid sequence of the N-terminal region of pLHCPII. The cDNA sequence is specified by capital letters. The amino acid sequence (single letter code) is specified below each of the codons. A double arrow indicates the acceptor site for the 26 nucleotide splice-leader sequence that is post-transcriptionally added to LHCPII pre-mRNA by trans-splicing. A 'TATA' box like sequence is underlined. The end of the N-terminal presequence and start of mature LHCPII is separated by a slash (/). A line above the sequence indicates the oligonucleotide primers used for primer extension RNA sequencing (Pex2), and for the RNA-PCR synthesis (Con1 and Con2) of NLH1 (11). The GC7 sequence (nucleotides 2897–2935) absent from NLH1 which is thought to be the 3’ end of exon 5 based on homology to the GC18 sequence encoding 573PEP (4) is in italicized capital letters.
RESULTS

A 16 kb genomic clone, GC7, containing the 5' end of a LHCP II gene (Fig. 1A) was isolated by screening a Euglena genomic library containing \(4 \times 10^6\) recombinant phage with a fragment of the Euglena LHCP II genomic clone, GC18, encoding LHCP II unit 114PEP (4). Hybridization of strand specific in vitro transcripts of GC7 subclones (Fig. 1A) to total Euglena RNA identified the coding strand (data not shown). The nucleotide sequence of a 3.1 kb region of GC7, appearing at the 5' most end of the encoded LHCP II gene, was determined and is presented in Figure 2 along with the derived amino acid sequence of the encoded protein. The coding regions (exons) in GC7 were determined by comparison with the sequence of a cDNA clone, NLH1 (Fig. 1A), containing the 5' end of an Euglena LHCP II mRNA (11). NLH1 was isolated by RNA-PCR (11) using a 5' primer corresponding to the conserved 26 nt sequence added to the 5' end of all Euglena mRNAs by transsplicing (13), and two nested 3' primers (Con1 and Con2, Fig. 2) corresponding to a highly conserved sequence in the C-terminal region of Euglena LHCP II (4). The NLH1 sequence was completely homologous to the corresponding sequence of GC7, suggesting that it was derived from the LHCP II transcript encoded by GC7.

The five exons identified in GC7 can be combined into an open reading frame encoding the 33 kDa N-terminal portion of pLHCP II consisting of a 141 amino acid N-terminal extension (11) linked to an LHCP II unit, N1PEP (Fig. 2). The sequence preceding the initiation codon (TTCAAAATG, Fig. 2) is characteristic of translation initiation sites in eukaryotes (14,15), suggesting that this is in fact the translation start site. The four introns range in size from 86 to 436 nt while the exons range in size from 42 to 496 nt (Fig. 2). The introns have a higher A+T content (56%) than the exons (41%). An analysis of codon usage indicates a bias (64%) for codons ending with G or C as found for other Euglena nuclear genes (4,16-18). This bias is more pronounced in the N1PEP coding region (72%) than in the region encoding the N-terminal extension (55%). The amino acid sequence of N1PEP is 93% homologous to GC18 encoded LHCP II units 114PEP and 575PEP (4). At the nucleic acid level, the sequence of N1PEP encoding exons is 87% homologous to that of the 575PEP and 114PEP encoding exons of GC18 (4). The codon bias for G/C at the third position is 90% in the 575PEP encoding region of GC18 (4) suggesting that much of the nucleic acid sequence divergence between N1PEP and 114PEP/575PEP is the result of changes at the third position. Due to the redundancy of the triplet code, these changes have little effect
on the protein sequence. The 3' end of NLH1 was defined by the 3' PCR primer (Con 2, Fig. 2) used for its synthesis and does not represent the 3' end of the last GC7 exon (exon 5). Based on sequence homology between the 575PEP encoding region of GC18 (4) and exon 5 of GC7, nucleotide 2935 appears to be the 3' end of exon 5 (Fig. 2).

A single intron is found in identical positions in the regions of Euglena genomic clone GC18 encoding LHCP II units 114PEP and 575PEP (4). An intron is also present at the same position (Fig. 2, nucleotides 1528-1613) in the N1PEP coding region of GC7. The N1PEP coding region is however interrupted by at least four additional introns (Fig. 2) not found in the comparable coding regions of GC18 (4). The 5' and 3' ends of the 4 introns present in the LHCP II gene encoded by GC7, as identified by a comparison of the genomic and cDNA sequence, do not contain the conserved GT and AG found at the ends of virtually all eukaryotic pre-mRNA introns (Fig. 2) (6-8).

Direct primer extension RNA sequence analysis was used to determine the sequence at the 5' end of the LHCP II mRNA encoded by GC7 (Fig. 3). The primer extension product contained a 26 nt sequence (5'-NNNTTTCTGAGTGTCTATATTTTTTTTCG-3') at the 5' end that was not present within approximately 1 kb of the GC7 sequence upstream from the translation initiation site at nt 1063 (Fig. 2). The 5' end of the mature Euglena rbcS mRNA contains this identical sequence which is not encoded by the rbcS gene (13,16). This same sequence is present at the 5' end of a group of small Euglena RNAs called spliced-leader RNAs (SL-RNAs) and is probably added post-transcriptionally to the 5' end of most Euglena pre-mRNAs by trans-splicing (13,19).

A sequence comparison between the primer extension product and GC7 identifies at 1032 as the 5' end of the pLHCP II mRNA encoded by GC7 (Figs. 2 and 3). A TATA box (TATATAA) (20) is found 264 nt upstream of the 5' end of GC7 encoded pLHCP II mRNA suggesting that the actual transcription start site may be located approximately 240 nt upstream of this 5' end (Fig. 2). The GC7 encoded primary transcript (pre-mRNA) probably contains this approximately 240 nt sequence which is removed when the 26 nt splice-leader sequence present at the 5' end of SL-RNA is added by trans-splicing to the pre-mRNA.

RNAs of 6.6 and 9.5 kb have been identified as the LHCP II mRNAs of Euglena (4). A 0.68 kb BarX fragment of genomic clone GC18 encoding most of 114PEP (4) hybridized to both LHCP II mRNAs (Fig. 4a, lane 1). A 0.45 kb BamHI-BglII fragment from the 5' end of cDNA clone NLH1 (Fig. 1A) encoding only the N-terminal extension region of the polypeptide and none of the mature LHCP II (N1PEP) sequence hybridized to only the 6.6 kb message (lane 2). The 0.45 kb BamHI-BglII 5' end probe hybridized with approximately equal intensity to 3 fragments on Euglena genomic DNA Southern blots (Fig. 4b). A probe containing the 3' untranslated region of the GC18 encoded LHCP II gene also hybridizes only to the 6.6 kb mRNA and to 2-3 fragments on Euglena genomic DNA Southern blots (4). A sequence comparison of the 3' end of the 4.8 kb EcoRI-SalI fragment (7SE48) of GC7 (Fig. 1A) and the 5' end of GC18 (4) found no overlap between the two genomic clones. The 4.3 kb Euglena rbcS mRNA is transcribed from an approximately 15 kb gene (9,16). The exons in GC7 and GC18 (4) comprise approximately 45% of the sequenced regions. Based on these estimates of coding versus noncoding sequences within Euglena genes, the combined coding capacity of the approximately 9 kb LHCP II encoding region of GC7 and the 7.6 kb LHCP II encoding region of GC18 (4) is 5.0-7.4 kb. It can not be determined whether GC7 and GC18 are the 5' and 3' ends of the same gene or whether they represent different members of the 3 gene family encoding the 6.6 kb LHCP II message.

The previously characterized genomic clone GC18 encodes the C-terminal half of a Euglena LHCP II polypeptide (4). Based on the criteria of maintaining maximum amino acid sequence homology between Euglena and Arabidopsis LHCP II, 8 introns were tentatively identified (4). None of the 8 introns contained the universally conserved GT→AG intron borders (4). Euglena LHCP II could however contain numerous aminoacid insertions and deletions resulting in the misidentification of intron→exon junctions. A comparison of cDNA and genomic sequences was performed to identify the true intron→exon junctions of GC18 and thus demonstrate the uniqueness of the intron borders in the Euglena LHCP II gene. GC18 encodes 3 types of LHCP II having more than 90% nucleotide sequence homology within a type (4).

---

**Figure 5.** Nucleotide sequence of the 3' end of the LHCP II polypeptide genomic clone GC18 (4), the LHCP II cDNA clones CLH09 and CLH22, and the deduced amino acid sequence of the C-terminal region of pLHCP II. The cDNA sequence is specified by capital letters. The amino acid sequence (single letter code) is specified below the first base of each codon. An asterisk (*) indicates the stop codon. The sequence begins at the 5' end of the ninth exon (Ex9) of the previously published GC18 encoded LHCP II gene (4) and extends beyond the 3' end of the LHCP II mRNA. The SalI site marking the end of the previously published GC18 sequence (4) is underlined. A double arrow indicates the G at nucleotide 7258 that is absent from the published GC18 sequence (4).
Oligonucleotides specific to unique coding sequences within each LHCP-II type were used for cDNA synthesis and PCR amplification (Fig. 1B). Due to the presence of more than one unit of each LHCP-II type within a given polyprotein gene, each oligonucleotide could potentially prime DNA synthesis at more than one site. The oligonucleotide pairs for RNA-PCR were therefore chosen so that a single product of known size spanning one or more introns would be produced based on the known linear order of LHCP-II types within the characterized region of GC18 (Fig. 1B). Priming within other regions of the mRNA would produce either no product or products differing in size from the desired cDNA.

Five overlapping cDNAs (LH34, LH64, LH138, CLH90 and CLH22), completely spanning the coding region of GC18, were obtained and sequenced (Fig. 1B). All but one (spanning 3n6, the only intron in 57SPEP) of the intron—exon junctions were identified by comparing the genomic sequence to the sequence of two or more cDNA clones synthesized using different primer pairs. The sequence of these cDNA clones is completely homologous to the exon sequences from GC18 (data not shown) confirming that they were derived from mRNA transcribed from this gene. Alignment of the cDNA and GC18 sequences identified 10 introns (Figs 5 and 6). Eight introns (3n11—3n8) corresponded to those previously identified by the criteria of preservation of coding sequence (4). Two previously unidentified introns (3n9, 3n10) were identified at the 3' end of GC18 (Figs 5 and 6). All introns lacked the GT—AG border sequences found in all other eukaryotes (Fig. 6). The intron—exon junctions of 3n3 and 3n5 differ from those previously reported (4) based on the criteria of conserved amino acid sequence. The actual intron—exon junction of 3n3 inserts an alanine at amino acid 25 (nucleotides 1653—1655 of GC18 sequence) of 351PEP. The actual intron—exon junction of 3n5 changes amino acid 95 of 351 PEP from alanine to threonine. The GenBank sequence (accession number X61361) has been updated to reflect the correct intron-exon junctions and complete the 3' end sequence of the encoded LHCP-II gene.

cDNAs CLH90 and CLH22 (Fig. 1B) contain the 3' end of the GC18 encoded transcript. Comparison between the cDNAs and GC18 sequence (4) identified a single nucleotide difference between the two sequences in this region. The cDNA contained a G that was not found during the initial sequencing of GC18 (4). This region of GC18 was resequenced and the G present in the cDNA was found in GC18. The sequence of the 3' end of GC18 revised by insertion of a G at nucleotide 7258 and the corresponding cDNA sequence are presented in Figure 5. The previously published GC18 sequence (4) ended at the SalI site.

A comparison of the genomic and cDNA sequences identified two additional introns, 3n19 and 3n10, at the 3' end of GC18 (Figs. 5 and 6). The region downstream of the SalI site contains the last exon (Ex11) of the polyprotein gene (Fig. 5). The GC18 encoded LHCP-II mRNA contains a 109 nucleotide non-translated 3' end (Fig. 5). The penultimate exon (Ex10) encodes the C-terminus of the LHCP-II polyprotein and LHCP-II unit 811PEP (4), as well as the first 22 nt of the 3' non-translated end of LHCP-II mRNA (Fig. 5). The high degree of nucleic acid sequence homology between the 811PEP and 1PEP coding regions (97%) (4) extends to the presence of introns (3n1 and 3n9) at the same position within both coding regions and seven identical nucleotides at the 5' end of the introns (Fig. 6). This high degree of sequence homology is not found at the 3' end of the introns (Figs. 5 and 6). The last exon (Ex11) encodes most of the 3' non-translated end of the LHCP-II mRNA (Fig. 5). A polyadenylation signal (AATAAA) found 11—30 nt upstream of the site of poly (A) addition in many but by no means all eukaryotic nuclear transcripts (20) was not found in the 3' non-translated region.

The 10 GC18 and 4 GC7 intron—exon junction sequences are presented in Figure 6 with the consensus nucleotides within the intron shaded for ease of comparison. Except for introns 3n19 and 5n14, the presence of identical nucleotides at the 5' and 3' intron—exon junctions make splice site selections ambiguous. Splice sites were chosen to maximize homology within the 5' end of the intron. Of the 14 introns in the LHCP-II gene, only two contain the 5' GT and none contain the 3' AG (Fig. 6) found at the 5' and 3' ends, respectively, of virtually every nuclear pre-mRNA intron (6—8). A common sequence was not apparent at the 5' or 3' intron—exon junction even in the case of the two intron pairs (5n1 and 3n2, 3n1 and 3n9) that are at identical positions within highly conserved coding regions (Fig 6).

The sequence alignment identifies a number of positions with nonrandom nucleotide usage. All introns contain a purine (A or G) at the 5' end and all but one, 3n19, contain a pyrimidine (C or T) at the 3' end (Fig. 6). All exons contain a purine at the 5' end and all but one, Ex5, contain a pyrimidine at the 3' end (Fig. 6). One other highly conserved intron feature is a CA at nt positions 4 and 5 from the 5' end of 11 introns and a complementary TG at nt positions 7 and 6 from the 3' end of 10 introns (Fig. 6). This complementary CA, TG pair is replaced by a complementary AG, CT pair in intron 3n10 and a complementary CT, AG pair in intron 3n3. The two remaining introns, 5n1 and 3n14, contain respectively a CA, GG and TG, CC pair where only one of the two nucleotides are complementary. Three of the five sequenced Euglena rbcS introns, 12 (CA, TG), 11e (AG, TT) and 3e (GG, CC), have complementary dinucleotide pairs at nt positions 4 and 5 from the 5' end and at nt positions 7 and 6 from the 3' end of the intron (9). Although the sample set is small, the lack of a consensus splice site is apparent and implies that structural rather than sequence information is used for splice site selection in Euglena.

The Euglena LHCP-II introns could not be folded into the highly conserved secondary structures characteristic of group I (21) and

![Figure 6. Alignment of the 14 intron—exon junctions in the Euglena LHCP-II genomic clones GC7 and GC18 (4). Intron—exon junctions were identified by comparing the cDNA and genomic sequences. The introns in Euglena genomic clone GC7, encoding the 5' tail of LHCP-II mRNA, are labelled at Sall as 3n1—4; 5n1—4; 3n5—5; and 5n1—5 being the 5' most intron. The introns in Euglena genomic clone GC18 (4) encoding the 3' end of LHCP-II mRNA are labelled as 3n1—10; 3n1 being the 5' most intron. In a number of cases, splice site identification is ambiguous due to sequence duplication at the intron junctions. Splice sites were chosen and sequences aligned to obtain maximum 3' splice homology at the 5' end of the intron. A consensus sequence for the 5' and 3' intron ends was determined and the consensus nucleotides within the intron are shaded for ease of comparison.](image-url)
group II (22) introns. The introns in the Euglena rbcS gene fold into stem-loop structures with the splice sites juxtaposed but offset by 1 or 2 nt (9). Complementarity between short stretches of nucleotides within the 5' and 3' ends of all the sequenced Euglena LHCPII introns allowed the introns to be folded into similar stem-loop structure with the 5' and 3' splice sites juxtaposed but offset by 2 nt, except for intron 5ln1 where the offset is 3 nt. Inspection of the intron 5' and 3' ends forming the stem loop structures (Fig. 6) shows that except for the conserved complementary C-A and TG dinucleotides, the distribution of base pairs within the stem varied from intron to intron. The positional variation in the base paired nucleotides explains the lack of a conserved sequence at the 5' and 3' ends of the intron. The 'fold' and 'mfold' programs of the GCG sequence analysis package (23) failed to reveal more complex shared structural features among the Euglena introns. Most RNA regions can be folded into stable stem loop structures. Whether base pairing between the 5' and 3' end of the intron is required for splicing of these unique class of introns remains to be seen.

DISCUSSION

The Euglena LHCPII gene encodes a polyprotein composed of multiple copies of at least three types of LHCPII (4). Within a type, there is over 90% nucleotide and amino acid sequence conservation while between types, there is approximately 60–70% sequence homology; the same amount of homology found between Euglena LHCPII and higher plant LHCPII (4). The conservation of coding sequences is in marked contrast to the divergence of intron position and sequence. The GC7 region encoding N1PEP is 87% homologous to the GC18 regions encoding 114PEP and 575PEP (4). GC7 and GC18 either encode portions of two different LHCPII genes or if they are portions of the same gene, the N1PEP coding region is approximately 6–8 kb 5' to the 114PEP coding region. The N1PEP coding region contains at least 5 introns while the 114PEP and 575PEP coding regions (4) contain a single intron in the same position as the first of the 5 introns within the N1PEP coding region. The respective introns exhibit less than 40% sequence identity even though the flanking exons are greater than 90% homologous. The maintenance of a high degree of sequence homology among physically dispersed exons suggests selection pressure on the exon encoded domains. Higher plants also contain multiple LHCPII (1,2) types but it remains to be established whether they are functionally different. Functional differences among the multiple LHCPIIs found in all chlorophyll a/b containing organisms would provide the selection pressure needed to maintain sequence homology among physically dispersed exons encoding different LHCPII types.

Sequence comparisons between the genomic and cDNA clones revealed that Euglena LHCPII pre-mRNA processing requires both cis- and trans-splicing. The 5' end of LHCPII mRNA contains a 26 nt sequence that is not encoded by the immediately upstream genomic sequence. This sequence is identical to the non gene encoded 26 nt sequence found at the 5' end of rbcS mRNA (13) and is also identical to the 26 nt sequence at the 5' end of cytoplasmic EF-1α mRNA (24) and chloroplast IF-3 mRNA (25). This 26 nt sequence appears to be trans-spliced from the 5' end of one of the Euglena SL-RNAs (13,19) to the pre-mRNA.

Trypanosomes and nematodes are the only other organisms known to form the 5' end of their mRNAs by trans-splicing (26). Trans-splicing occurs in trypanosomes and nematodes with removal of an intron and formation of the 5' end of the mRNA by transfer of a leader sequence from a small SL-RNA, to the pre-mRNA (26—29). Euglenoids are phylogenetically related to trypanosomes (30—32). Trypanosome SL-RNAs contain a UUC/GUA exon/intron boundary (29.33.34) while Euglena SL-RNAs contains a UCG/GUA exon/intron boundary (13,19). Similarities between trans-splicing in Euglena, trypanosomes (26,29), and nematodes (27,28) such as an AG dinucleotide at the 3' splice acceptor site and removal of an intron (approximately 240 nts in case of Euglena LHCPII) from the pre-mRNA during trans-splicing suggest that the mechanism of trans-splicing in Euglena is similar to that used by both trypanosomes and nematodes.

Nuclear pre-mRNA cis-splicing appears to be an evolutionarily conserved process that has been most extensively characterized in yeast and mammals (6,35). The yeast and mammalian splicing complex, the spliceosome, is composed of the pre-mRNA, five small nuclear RNAs (snRNAs) and a large number of proteins (6,35). Virtually all eukaryotic pre-mRNA introns contain invariant 5'-GT and AG-3' borders (6—8). In addition to the invariant GT—AG borders, yeast introns contain a G 5' nucleotides downstream from the 5' splice site and a consensus branch point sequence about 30—40 nucleotides from the 3' splice site (35) while many mammalian introns contain a polypyrimidine tract at the 3' end of the intron (6,7). Splice site selection in both yeast and mammals involves interactions between the snRNAs and conserved sequences at both the splice site and within the intron (36—39). A non-Watson—Crick interaction between the terminal guanosines of the invariant GT—AG pair appears to be essential for positioning the 3' splice site for cleavage and exon ligation (40). The fourteen Euglena LHCPII mRNA introns and five Euglena rbcS mRNA introns (9) lacked GT—AG borders and there were no identifiable conserved intron sequences as found in cis-spliced nuclear pre-mRNA introns of other eukaryotes. This suggests that cis-splicing in Euglena is fundamentally different from the well characterized evolutionarily conserved spliceosome system of other eukaryotes.

The complete sequencing of the Euglena chloroplast genome has identified a large number of introns which include at least 74 group II introns, 64 group III introns and 15 tRNA introns, the introns within introns (41,42). The LHCPII introns and rbcS introns (9) could have evolved prior to transfer of the respective gene from the chloroplast to the nucleus during the reduction of a photosynthetic endosymbiont to a chloroplast. Group II introns are identified by conserved sequence elements and conserved secondary structures (22,42). Group III introns have fewer conserved features but they appear to be characterized by a group II intron-like domain VI at the 1st end (43—44). Secondary structures or conserved sequence elements characteristic of group II or group III introns could not be identified in the LHCPII introns. The only structural feature common to the Euglena LHCPII and rbcS introns (9) is the formation by hydrogen bonding of a stem between the 5' and 3' ends of the intron leaving the splice site juxtaposed but offset by 1—3 nt. Although the sample set is restricted to LHCPII and rbcS (9) pre-mRNA introns, Euglena nuclear encoded pre-mRNA introns appear to represent an entirely new, hitherto uncharacterized intron type.

Morphological criteria (30), cytoplasmic rRNA sequence comparisons (31) and trans-splicing of nuclear pre-mRNA (13,26) indicate that euglenoids are most closely related to
trypanosomes. The *Euglenoid*-trypanosome lineage is thought to be the most ancient protistan lineage having diverged prior to separation of all other protistan (32). *Cis*-spliced introns are absent from trypanosome genes (45), suggesting that the spliceosome dependent *cis*-splicing system found in most organisms evolved after their separation from the *Euglenoid*-trypanosome lineage. The ancestral *Euglenoid* is generally accepted to have been a phagotrophic trypanosome like organism that engulfed a eukaryotic algae (46,47). The transfer of intron containing chloroplast genes from the endosymbiotic algae to the nucleus of the trypanosome like host during reduction of the endosymbiont to a chloroplast would have required development of a *cis*-splicing system. The independent evolution of *Euglena* *cis*-splicing would explain the absence of conserved GT→AG intron borders and suggests that fundamental differences exist between the well characterized spliceosome *cis*-splicing system found in most organisms and the as yet uncharacterized *cis*-splicing system of *Euglena*.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation Grant MCB-9118721 and by funds from the University of Nebraska-Lincoln Research Council. U.S.M. was supported by a University of Nebraska-Lincoln Center of Biotechnology Graduate Research Fellowship. We thank Amy Siebert for technical assistance and the personnel at UNL Biotechnology Sequencing Facility for their assistance in sequence determination.

REFERENCES