Nucleic Acids Research, 1994, Vol. 22, No. 25 5737–5744

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 decapeptide 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 decapeptide 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 (*rbc*S) 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 rbc*S 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* LHCPII 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 LHCPII genomic clone

The genomic clone, GC7, was obtained by screening a *Euglena* library containing 4×10^5 recombinant phage (4). The 0.68 kb *Bst*XI fragment of the genomic clone GC18, encoding most of LHCPII unit 114PEP (4), was used as a probe for plaque hybridizations. The LHCPII hybridizing region of GC7 was subcloned as a 5.2 kb (7ES52) *Sal*I–*Eco*RI and a 4.8 kb (7SE48) *Eco*RI–*Sal*I fragment in pBluescript II KS (Fig. 1A). A 3.1 kb region of the plasmid 7ES52 was further subcloned in pBluescript II KS as a 1.6 kb *Pst*I and an overlapping 2.0 kb *Sal*I–*Bgl*II fragment (Fig. 1A) for sequence analysis.

Characterization of *Euglena* LHCPII cDNAs

The synthesis and subcloning of NLH1, the cDNA clone corresponding to the 5′ end of *Euglena* LHCPII mRNA (Fig. 1A), by RNA-PCR has been described previously (11). The cDNAs corresponding to the LHCPII transcript encoded by GC18 (Fig. 1B) were synthesized by RNA-PCR using oligonucleotide primers corresponding to unique regions within each LHCPII 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′-CCGGAATTCTCACCGGAAAGGGCCCTGT-3′) and 2a (5′-GTGACCACCCCCAAGGCGCCCA-3′). The 1.38 kb cDNA, LH138 (Fig. 1B), was synthesized using the oligonucleotides 1a and 3b (5′-GTGGTCGACACCAGCCTG-AACAAACAG-3′). The synthesis of a 0.64 kb cDNA, LH64 (Fig. 1B), utilized oligonucleotides 2b (5′-CCGTCTGCAAA-CAACATCTTCAGC-3′) and 3b. For the cDNA clones, CLH22 and CLH09 (Fig. 1B), corresponding to the 3′ end of the *Euglena* LHCPII transcript, an anchor-(dT)17 oligonucleotide (5′-GGG-AATTCGTCGACAAGCTTTTTTTTTTTTTTTT-3′) was used for first strand synthesis. The amplification reactions contained the anchor oligonucleotide (5′-GGGAATTCGTCG-ACAAGC-3′) and either oligonucleotide 3a (5′-GCTGCTC-ACCCTGGTGCTTGAGC-3′) for CLH22, or 2b for CLH09 (Fig. 1B). LH34 was subcloned as a blunt-ended insert, LH138 as an *Eco*RI–*Sal*I insert, LH64 as a blunt end-*Sal*I insert, and both CLH22 and CLH09 as *Eco*RI/*Hin*dIII-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′-CTTCATGGCATCGGCGTTG-TTTGGC-3′) used for primer extension-RNA sequencing, was gel purified and end labelled with γ^{32}P-ATP using T4 polynucleotide kinase as described (12). For sequencing, 20 μg of *Euglena* total RNA and 0.05 pmoles 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 $MgCl_2$, 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 ddCTP 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

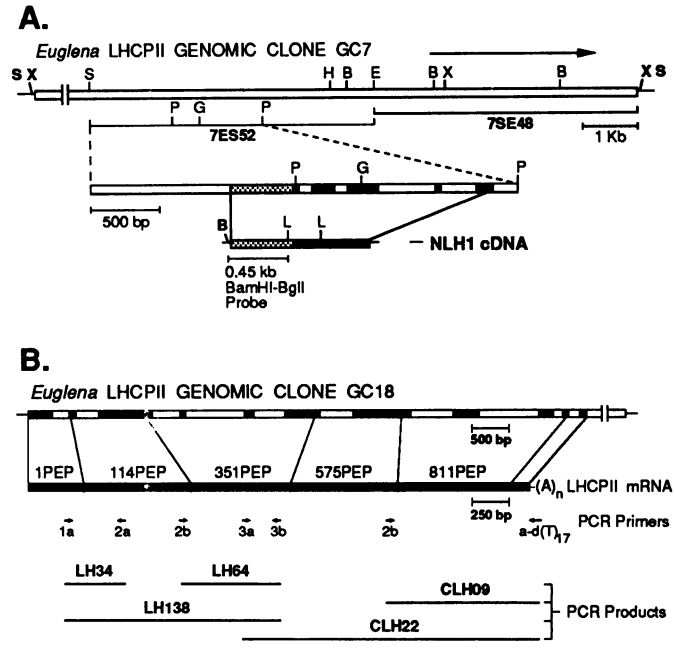

Figure 1. (A) Restriction map of *Euglena* genomic (GC7) and cDNA (NLH1) clones encoding the 5′ end of LHCPII mRNA. The *Euglena* insert is presented as the boxed area while phage DNA is represented by a solid line. The arrow indicates the direction of transcription of encoded LHCPII gene. The five exons in the 3.1 kb sequenced segment of subclone 7ES52 are shaded. The lightly shaded region identifies the portion of exon 1 encoding the N-terminal extension of pLHCPII while the more heavily shaded areas indicate exon regions encoding LHCPII unit N1PEP. Restriction sites are defined as: B—*Bam*HI, E—*Eco*RI, G—*Bgl*II, H—*Hin*dIII, L—*Bgl*I, P—*Pst*I, S—*Sal*I, X—*Xho*I. Vector derived restriction sites are in boldface. (B) RNA-PCR strategy for obtaining cDNAs corresponding to the transcript encoded by the *Euglena* LHCPII genomic clone GC18 (4). Oligonucleotides (1a, 2a, 2b, 3a and 3b) specific to unique regions of GC18 and an anchored-d(T)17 (a-d(T)17) were utilized as primers for RNA-PCR. The 11 exons in GC18 are shaded. The individual LHCPII polypeptide units encoded by the LHCPII mRNA are specified above their coding regions. Arrows indicate the orientation of each primer. The cDNAs obtained with each primer pair are aligned to the mRNA. All but one intron–exon junction is spanned by at least two cDNAs.

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.

```
gtcgacagcgcactatggcctgcctccaaggggcctacccggggcccggatgacgctccaggagcgtggtacaggagactagtccgcgggtggggcggtg   100
gcgtgggggttctctgacccatccaacaccggctgggtcagaggaaagaggaggaaatggaggaaagggaacacgttcttcttcgttctcctgcggtgag   200
gaggtactttgtttaaggacaaaagcaggggggcgtttgaggctccctgctgctgacccagcatatgccattactgaggggcttcttgttgtcacggtggg   300
tagcacagcttcacccaaaccgtgccccgggcccagggaaacgtctcttttggccctcgccacaagaggaaccccgggtttttggtttttgtgtgctttg   400
cttctttccgcgtctttgtgacccgttgatttgtctttggctcggcgcgtagttctgctcaaggggtcactaggggttaattgaggtgcaaacacagcag   500
gtgagagttcaatgttggtgatttttgtggagagtttgtgctcaaagtttgtgctcagagggatcacggtgggtagcacaacttcaccctaacgtgcccc   600
gggcccagggaaacgtctcttttggccctcgccacaaggagaacctcctcatgactccctgtgccgccacaggtcgcgggttgctatactaagcagtttt   700
aaaagaggtgttcggtgactggtgagaggtaagtggaaaatgagcgaggctcatcttcaatatataagagcgaggggcgtgtgtgttctggtgttgcct   800
tctttggcgtgttctggtgttgccttctttgtgggcaccaaatgcagaggaaattcttactgacgtgcacccccgagacgacacccaaaaccgcccacac   900
tacatcacacactcaaataccacacatattccatcatccgattctggtgttgtcagcttgactttgcaataaggaacccagatcgctcagaaactcagca  1000
                                                                                  Pex2
attgcgagaaaaatcttccactttctcacagACATTCTTTGCTTTCATCCACTCACTTCAAAATGCCAAACAACGCCGATGCCATGAAGTTTGGTCTCGC  1100
                                                                  M  P  N  N  A  D  A  M  K  F  G  L  A
TGCTGGGGCAGCCATGGGTGTCATTGTGTACGTGCTTGCTGGTGCGGCAAGTTCGACATCTTTGGCTGCAACTCATGTTAACATACAGCAGGCTCCCGCT  1200
 A  G  A  A  M  G  V  I  V  Y  V  L  A  G  A  A  S  S  T  S  L  A  A  T  H  V  N  I  Q  Q  A  P  A
GTTATTCCTCGGATGGCTTCCGTCCCATCCGCCTATACCATTGCCACCAACCCCATTGGTGCCAGTGCCCGTGTCGTGGATGCCAATTACGAATCCACGG  1300
V  I  P  R  M  A  S  V  P  S  A  Y  T  I  A  T  N  P  I  G  A  S  A  R  V  V  D  A  N  Y  E  S  T  D
ACTATCTCACGCTTCCTGCAACTGAGAAGTCAACTATGGGCAGCCTTCTGATGATTGCTGCCGCTGGTGTCGCCGCTGCCGTTGCGTTCGTCTGGAAATC  1400
  Y  L  T  L  P  A  T  E  K  S  T  M  G  S  L  L  M  I  A  A  A  G  V  A  A  A  V  A  F  V  W  K  S
CGTTCCTCGGCAACAAGACTCAGTTATCAATGTTCCACTGCTGCCTGTTTCTGTCGCCACCATGGCAACCAGCGGGAAGAAGAGCAAAGCACCTGCTGCA  1500
 V  P  R  Q  Q  D  S  V  I  N  V  P  L  L  P  V  S  V  A  T  M  A  T  S  G  K  K  S /K  A  P  A  A
GACAACCTTTCTCAGTGGTATGGCCCTgagcacaactgccttcagacttgtataaaagcatttttccaatttttctttttttcaatgaaagagcgaagac  1600
D  N  L  S  Q  W  Y  G  P
ggttgtggaacatGACCGTGCAAAGTGGTTGGGTCCCCTGACTGGACAGGTTCCCGCTTACCTGACTGGAGAGCTGCCAGGTGATTACGGGTGGGATACT  1700
            D  R  A  K  W  L  G  P  L  T  G  Q  V  P  A  Y  L  T  G  E  L  P  G  D  Y  G  W  D  T
GCTGGTCTGGGTGCTGATCCAGTAACACTGGCCCGCTACCGTGAGGCTGAAGTCATCCATGCCCGTTGGGCAATggtcagattcctgtgcccaggatgtc  1800
A  G  L  G  A  D  P  V  T  L  A  R  Y  R  E  A  E  V  I  H  A  R  W  A  M
caatttttctttccaacgttaaaaggtaaaaatgacatggcaaaggacatctggcagtGCTGGGCGCCTTGGGGGTGGTCACCCCCGAGCTGCTGGCG  1900
                                                                  L  G  A  L  G  V  V  T  P  E  L  L  A
GGCAACGGCGTGCCGTTTGGCGAGGGCGCCGTCTGGTACAAGGCCGGCGCGCAGATCTTCTCCGCGGACGGCCTGAACTACCTGGGCAACCCCAGCCTGA  2000
G  N  G  V  P  F  G  E  G  A  V  W  Y  K  A  G  A  Q  I  F  S  A  D  G  L  N  Y  L  G  N  P  S  L  I
TCCACGCCCAGTCCGTGGTGCTGACGTTCCTGTCCACCCTGGCCATCATGGGTGCTGTTGAGGGTTACCGCTATGGTGGTggccagggtagaacgggaat  2100
  H  A  Q  S  V  V  L  T  F  L  S  T  L  A  I  M  G  A  V  E  G  Y  R  Y  G  G
cggccaaaacagccagagtagtgccccccaaaagttacgagtcaggtgaaccagcaacaccgaggaggagggttagggattggccctgctgtaccttgg  2200
gcagggcaattttcccccggactccaaaatttcaacattttctcgaaatttcaaaggagaagtttccaaaagaaaaaatgtcagtataatttcactctgg  2300
caagtttggtacatttttttttcctatttgaccgtatttcacccccctgcgaaaaactacaaatttttcaaaccggaaaatttttgttggaaaacgtgtctt  2400
tttatcagaaatcatgtgttttgggttcatttcagcacaagtccccgttttccacaattggtgaaaatgccaaccccactccccaccatcaagccatccc  2500
gttctgcgctggcggcGGCTGGCCACTGGGTGATGACCTTGACCGCCTGTACCCTGGCatccatggcgaagttgttctgttcttttcaggtacaatccca  2600
               G  W  P  L  G  D  D  L  D  R  L  Y  P  G
atcaagcactcacgacagcaacaaactcccacgctgtttcagccaggccaagaaccagacttcctctgcacatgccattgctcccatcattatcccaaat  2700
gtgtaagcaggccacttcaaagtttttcttacgactcccatattggactgtcacggcatgccagaaattaacaattttccatttttttcctgtgtattgat  2800
                                                                                  Con2
                                                                                                  Con1
tggactgcgccctgcttttGGCCCGTTCGACCCCCTGGGTCTGGCCAACGACCCCGACGCCTTCGCTGAGCTCAAGGTGAAGGAGCTGAAGAACGGCCGT  2900
                  G  P  F  D  P  L  G  L  A  N  D  P  D  A  F  A  E  L  K  V  K  E  L  K  N  G  R
CTGGCCATGGTGGCCATGCTGGGCTTCTACGTCCAagcaggaagtttttgtttggaagttttgtgttcctgccgggcttggtgtggtttcatttcttcca  3000
L  A  M  V  A  M  L  G  F  Y  V  Q
gaggaaggcagctccacctagctcaaaaaaacgtttttttttttaattgtacagccccacctgggccaggcatttgcttaagggattgcagtacccccc  3100
cggagggggggctggactgcag  3122
```

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 the first base of each codon. 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 575PEP (4) is in italicized capital letters.

RESULTS

A 16 kb genomic clone, GC7, containing the 5′ end of a LHCPII gene (Fig. 1A) was isolated by screening a *Euglena* genomic library containing 4×10^5 recombinant phage with a fragment of the *Euglena* LHCPII genomic clone, GC18, encoding LHCPII 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 to contain the 5′ most end of the encoded LHCPII 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* LHCPII 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 *trans*-splicing (13), and two nested 3′ primers (Con1 and Con2, Fig. 2) corresponding to a highly conserved sequence in the C-terminal region of *Euglena* LHCPII (4). The NLH1 sequence was completely homologous to the corresponding sequence of GC7, suggesting that it was derived from the LHCPII 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 pLHCPII consisting of a 141 amino acid N-terminal extension (11) linked to an LHCPII 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 nts while the exons range in size from 42 to 496 nts (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 LHCPII 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

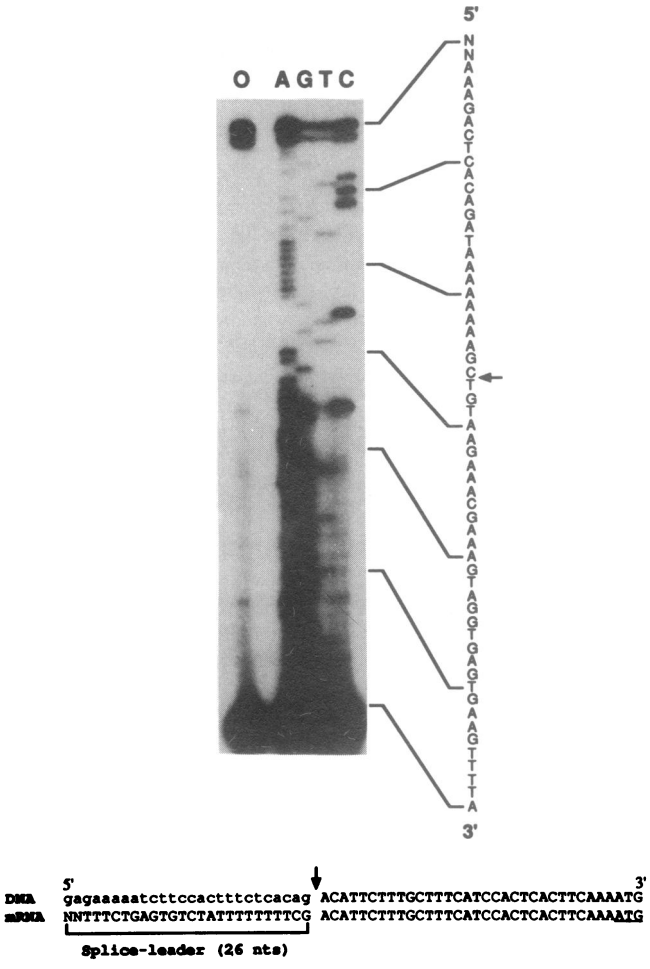

Figure 3. Primer-extension RNA sequencing of the 5′ end of LHCPII mRNA. The oligonucleotide Pex2 was used to prime DNA synthesis from 20 μg total *Euglena* RNA using AMV reverse transcriptase. Reactions were performed in the absence of any dideoxynucleotides (O) or in the presence of the indicated dideoxynucleotide (A, G, T or C). The sequence of the extended product (complimentary to the mRNA sequence) is shown to the right of the sequencing ladder. The primer-extension derived sequence of the 5′ end of the LHCPII transcript (mRNA) and the sequence of the corresponding region of the GC7 encoded LHCPII gene (DNA) are aligned to indicate the presence of a 26 nucleotide sequence in the message which is absent from the genomic sequence. An arrow indicates the site of addition of this 26 nucleotide splice-leader sequence.

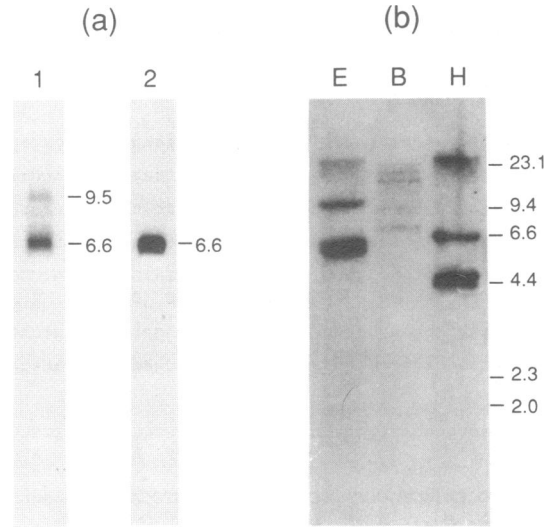

Figure 4. Northern blot analysis of *Euglena* total RNA and Southern blot analysis of *Euglena* genomic DNA. (**a**) Total cellular RNA (20 μg) was size fractionated on 0.8% agarose–formamide gels, transferred to nitrocellulose filters and the filters hybridized to a 0.68 kb *Bst*XI fragment of genomic clone GC18 encoding LHCPII unit 114PEP (4) (Lane 1) or a 0.45 kb *Bam*HI–*Bgl*I fragment of cDNA clone NLH1 encoding only the N-terminal extension of pLHCPII (Lane 2). (**b**) *Euglena* DNA (10 μg) was digested to completion with *Eco*RI (E), *Bam*HI (B) or *Hin*dIII (H), size fractionated on 0.7% agarose gels, transferred to nitrocellulose filters and the filters hybridized to the 0.45 kb *Bam*HI–*Bgl*I fragment of cDNA clone NLH1. Marker fragment sizes in kb are indicated to the right of each panel.

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 LHCPII 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 LHCPII 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 LHCPII mRNA encoded by GC7 (Fig. 3). The primer extension product contained a 26 nt sequence (5′-NNTTTCTGAGTGTCTATTTTTTTT-CG-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 rbc*S mRNA contains this identical sequence which is not encoded by the *rbc*S 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 nt 1032 as the 5′ end of the pLHCPII 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 pLHCPII 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 addded by *trans*-splicing to the pre-mRNA.

RNAs of 6.6 and 9.5 kb have been identified as the LHCPII mRNAs of *Euglena* (4). A 0.68 kb *Bst*XI fragment of genomic clone GC18 encoding most of 114PEP (4) hybridized to both LHCPII mRNAs (Fig. 4a, lane 1). A 0.45 kb *Bam*HI−*Bgl*I fragment from the 5′ end of cDNA clone NLH1 (Fig. 1A) encoding only the N-terminal extension region of the polyprotein and none of the mature LHCPII (N1PEP) sequence hybridized to only the 6.6 kb message (lane 2). The 0.45 kb *Bam*HI−*Bgl*I 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 LHCPII 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 *Eco*RI−*Sal*I 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 rbc*S 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 LHCPII encoding region of GC7 and the 7.6 kb LHCPII 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 LHCPII message.

The previously characterized genomic clone GC18 encodes the C-terminal half of a *Euglena* LHCPII polyprotein (4). Based on the criteria of maintaining maximum amino acid sequence homology between *Euglena* and *Arabidopsis* LHCPII, 8 introns were tentatively identified (4). None of the 8 introns contained the universally conserved GT−AG intron borders (4). *Euglena* LHCPII 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* LHCPII gene. GC18 encodes 3 types of LHCPIIs having more than 90% nucleotide sequence homology within a type (4).

```
GACCCCCTGTACCCCGGTGGCCCGTTCGACCCCCTGGGTCTGGCCGATGACCCCGAGGCATTCCTGGAGCTGAAGGTGAAGGAGGTCAAGAACGGCCGGC   6992
D  P  L  Y  P  G  G  P  F  D  P  L  G  L  A  D  D  P  E  A  F  L  E  L  K  V  K  E  V  K  N  G  R  L

TGGCCATGGTGGCCATCTTCGGCTTCTTCGTGCAGGGCATCCTCACCGGAAAGGGCCCCGTGGAGAACTGGGTGGACCACTTGACCGACCCATTCGTAAA   7092
  A  M  V  A  I  F  G  F  F  V  Q  G  I  L  T  G  K  G  P  V  E  N  W  V  D  H  L  T  D  P  F  V  N

Cgcccagcaattgttgtttttgaggttttcggtcactttgttttgatttttttttttggcgagtgtcccgtgctgtttgcttttctttgccaaataatatt   7192
                                                                  ⬇
tgtgcactctctttttttttgtttcaccaccatttgctgcctgaAACATCTTCCAGCTGACCCCTGGCTTCGCCATGTTCTAGTGATGTGGACAACGCAAC   7292
                                          N  I  F  Q  L  T  P  G  F  A  M  F  *  *

TTCTaaaaggtcttccatggtttgtctgtccattgcaggtgtgtattttggatgctcactatgcaactcaatgtctgccacctttcaagttattggtttt   7392

ttcgctcggtgtggactcgctcggtgtcgacttggtgtggaattgccgccccctggaaggcctgacctGCTTGCATCTGCTGGCTGTATTATTACACGCT   7492
                          SalI
CACTCACCTGCCCCTCTTTTTATTATGGACAGGACCATGAAAATGGTGCTTTATATtctgtctcctgctcgtctagagtttttgctgaagggtgattgga   7592

ttgaaatcaggcacgtttgtaattttggttgtctgaccaagcaactacggctagaagt  7650
```

Figure 5. Nucleotide sequence of the 3′ end of the LHCPII polyprotein genomic clone GC18 (4), the LHCPII cDNA clones CLH09 and CLH22, and the deduced amino acid sequence of the C-terminal region of pLHCPII. 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 LHCPII gene (4) and extends beyond the 3′ end of the LHCPII mRNA. The *Sal*I 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 LHCPII type were used for cDNA synthesis and PCR amplification (Fig. 1B). Due to the presence of more than one unit of each LHCPII 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 LHCPII 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, CLH09 and CLH22), completely spanning the coding region of GC18, were obtained and sequenced (Fig. 1B). All but one (spanning 3In6, the only intron in 575PEP) 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 (3In1−3In8) corresponded to those previously identified by the criteria of preservation of coding sequence (4). Two previously unidentified introns (3In9, 3In10) 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 3In3 and 3In5 differ from those previously reported (4) based on the criteria of conserved amino acid sequence. The actual intron−exon junction of 3In3 inserts an alanine at amino acid 25 (nucleotides 1653−1655 of GC18 sequence) of 351PEP. The actual intron−exon junction of 3In5 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 LHCPII gene.

cDNAs CLH09 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, 3In9 and 3In10, 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 LHCPII mRNA contains a 109 nucleotide non-translated 3′ end (Fig. 5). The penultimate exon (Ex10) encodes the C-terminus of the LHCPII polyprotein and LHCPII unit 811PEP (4), as well as the first 22 nt of the 3′ non-translated end of LHCPII 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 (3In1 and 3In9) 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 LHCPII 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 3In9 and 5In4, 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 LHCPII 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 (5In1 and 3In2, 3In1 and 3In9) 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, 3In9, 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 3In10 and a complementary CT, AG pair in intron 3In3. The two remaining introns, 5In1 and 3In4, contain respectively a CA, GG and TG, CC pair where only one of the two nucleotides are complementary. Three of the five sequenced *Euglena rbc*S introns, i2 (CA, TG), i1e (AG, TT) and i3e (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* LHCPII introns could not be folded into the highly conserved secondary structures characteristic of group I (21) and

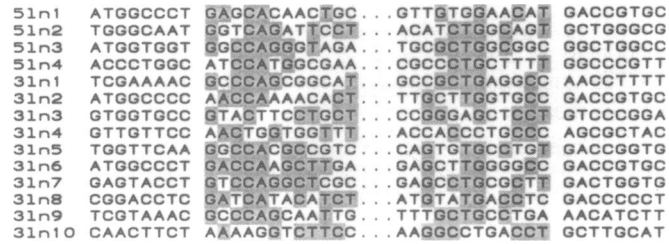

Figure 6. Alignment of the 14 intron−exon junctions in the *Euglena* LHCPII genomic clones GC7 and GC18 (4). Intron−exon junctions were identified by comparing the cDNA and genomic sequences. The introns in *Euglena* genomic clone GC7 encoding the 5′ end of LHCPII mRNA are labelled as 5In1−4; 5In1 being the 5′ most intron. The introns in *Euglena* genomic clone GC18 (4) encoding the 3′ end of LHCPII mRNA are labelled as 3In1−10; 3In1 being the 5′ most intron. In a number of cases, splice site identification is ambiguous due to sequence duplication at the intron exon junctions. Splice sites were chosen and sequences aligned to obtain maximum nt sequence 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.

group II (22) introns. The introns in the *Euglena rbc*S 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 5In1 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 *rbc*S 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 UUG/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 rbc*S 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 twintrons, the introns within introns (41,42). The LHCPII introns and *rbc*S 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 their 3′ end (42−44). 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 *rbc*S 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 *rbc*S (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 protistans (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

1. Buetow, D.E., Chen, H., Erdos, G. and Li, L.S.H. (1988) Photosynth. Res., **18**, 61−97.
2. Jansson, S. and Gustafsson, P. (1990) Plant Mol. Biol. **14**, 287−296.
3. Rikin, A. and Schwartzbach, S.D. (1988) Proc. Natl. Acad. Sci. U.S.A., **85**, 5117−5121.
4. Muchhal, U.S. and Schwartzbach, S.D. (1992) Plant Mol. Biol., **18**, 287−299.
5. Schiff, J.A., Schwartzbach, S.D., Osafune, T. and Hase, E. (1991) J. Photochem. Photobiol. B, **11**, 219−236.
6. Green, M.R. (1991) Annu. Rev. Cell Biol., **7**, 559−599.
7. Csank, C., Taylor, F.M. and Martindale, D.W. (1990) Nucleic Acids Res., **18**, 5133−5141.
8. Jackson, I.J. (1991) Nucleic Acids Res., **19**, 3795−3798.
9. Tessier, L.H., Chan, R.L., Keller, M., Weil J.H. and Imbault, P. (1992) F.E.B.S. Letters, **304**, 252−255.
10. Muchhal, U.S. and Schwartzbach, S.D. (1991) 3rd Int. Congress Soc. Plant Mol. Biol., Tucson, 6−10 October, 1991, Poster Abstract **232**.
11. Kishore, R., Muchhal, U.S. and Schwartzbach, S.D. (1993) Proc. Natl. Acad. Sci. U.S.A., **90**, 11845−11849.
12. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor University Press, Cold Spring Harbor.
13. Tessier, L.H., Keller, M., Chan, R.L., Fournier, R., Weil, J.H. and Imbault, P. (1991) EMBO J., **10**, 2621−2625.
14. Cavener, D.R. and Ray, S.C. (1991) Nucleic Acids Res., **19**, 3185−3192.
15. Kozak, M. (1993) Microbiol. Rev., **47**, 1−45.
16. Chan, R.L., Keller, M., Canaday, J., Weil, J.H. and Imbault, P. (1990) EMBO J., **9**, 333−338.
17. Houlne, G, and Schantz, R. (1988) Mol. Gen. Genet., **213**, 479−286.
18. Sharif, A.L., Smith, A.G. and Abell, C. (1989) Eur. J. Biochem., **184**, 353−359.
19. Keller, M., Tessier, L.H., Chan, R.L., Weil, J.H. and Imbault, P. (1992) Nucleic Acids Res., **20**, 1711−1715.
20. Lewin, B. (1990) Genes IV. Cell Press, Cambridge.
21. Cech, T.R. (1988) Gene, **73**, 259−271.
22. Michel, F., Umesano, K. and Ozeki, H. (1989) Gene, **82**, 5−30.
23. Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res., **12**, 387−395.
24. Montandon, P.E. and Stutz, E. (1990) Nucleic Acids Res., **18**, 75−82.
25. Lin, Q., Ma, L., Burkhart, W. and Spremulli, L.L. (1994) J. Biol. Chem., **269**, 9436−9444.
26. Agabian, N. (1990) Cell, **61**, 1157−1160.
27. Nilsen, T.W. (1989) Exp. Parasitol., **69**, 413−416.
28. Blumenthal, T. and Thomas, J. (1988) Trends Genet., **4**, 305−308.
29. Kapotas, N. and Bellofatto, V. (1993) Nucleic Acids Res., **21**, 4067−4072.
30. Kivic, P.A. and Walne, P.L. (1984) Origins of Life, **13**, 269−288.
31. Sogin, M.L., Elwood, H.J. and Gunderson, J.H. (1986) Proc. Natl. Acad. Sci. U.S.A., **83**, 1383−1387.
32. Sogin, M.L. (1991) Curr. Opin. Genet. Dev., **1**, 457−463.
33. Agami, R. and Shapira, M. (1992) Nucleic Acids Res., **20**, 1804−1804.
34. Bellofatto, V., Cooper, R. and Cross, G.A.M. (1988) Nucleic Acids Res., **16**, 7437−7456.
35. Ruby, S.W. and Abelson, J. (1991) Trends Genet., **7**, 79−85.
36. Newman, A.J. and Norman, C. (1992) Cell, **68**, 743−754.
37. Steitz, J.A. (1992) Science, **257**, 888−889.
38. Lesser, C.F. and Guthrie, C. (1993) Science, **262**, 1982−1988.
39. Sontheimer, E.J. and Steitz, J.A. (1993) Science, **262**, 1989−1996.
40. Parker, R. and Siliciano, P.G. (1993) Nature, **361**, 660−662.
41. Hallick, R.B., Hong, L., Drager, R.G., Favreau, M.R., Monfort, A., Orsat, B., Spielmann, A. and Stutz, E. (1993) Nucleic Acids Res., **21**, 3537−3544.
42. Copertino, D.W. and Hallick, R.B. (1993) Trends Biochem. Sci., **18**, 467−471.
43. Christopher, D.A. and Hallick, R.B. (1989) Nucleic Acids Res., **17**, 7591−7608.
44. Copertino, D.W., Hall, E.T., Van Hook, F.W., Jenkins, K.P. and Hallick, R.B. (1994) Nucleic Acids Res., **22**, 1029−1036.
45. Pays, E. (1993) in Broda, P., Oliver, S.G. and Sims, P.F.G. (eds), The Eukaryotic Genome - Organization and Regulation. Cambridge University Press, Cambridge, pp. 127−160.
46. Gibbs, S.P. (1981) Ann. N.Y. Acad. Sci., **361**, 193−208.
47. Gibbs, S.P. (1993) In Lewin, R.A. (ed.) Origin of Plastids. Chapman & Hall. New York, pp. 107−121.