The Euglena gracilis Chloroplast Ribulose-1,5-bisphosphate Carboxylase Gene

I. COMPLETE DNA SEQUENCE AND ANALYSIS OF THE NINE INTERVENING SEQUENCES*

Jeffrey C. Gingrich§ and Richard B. Hallick§

From the Department of Chemistry, University of Colorado, Boulder, Colorado 80309

The nucleotide sequence of 6225 base pairs (bp) of Euglena gracilis chloroplast DNA including the complete DNA sequence of the chloroplast-encoded ribulose-1,5-bisphosphate carboxylase large subunit gene along with the flanking DNA sequences is presented. The gene is >5.5 kilobase pairs in length and is organized as 10 exons coding for 475 amino acids, separated by 9 introns. The exons range in size from 45 to 438 bp, while the introns range in size from 382 to 568 bp. The introns have highly conserved boundary sequences with the consensus, 5'-N GTGTTGGATT... (intron)...TAAATTAT T N-3'. The introns are 82-85 mol% AT, with a pronounced T > A > G > C base bias in the RNA-like strand. They do not appear to encode any polyadenylate. In addition, the introns have a conserved sequence 30-50 bp from their 3'-ends with the consensus, 5'-TACAGTTTGAAAATGA-3'. The 5'-TACA sequence bears some homology to the 5'-end of the TACTAACA sequence found in a similar location in yeast nuclear mRNA introns. The conserved sequences of the Euglena rbcL introns may be indicative of a splicing mechanism similar to that of eucaryotic nuclear mRNA introns and group II mitochondrial introns.

Many features of chloroplast transcription units are closely related to their bacterial counterparts. The size and organization of chloroplast mRNA operons are homologous to bacterial rRNA operons. Transcription initiation and termination signals in chloroplast DNA may be quite similar to the bacterial sequences (Whitfield and Bottomley, 1983). Translation from chloroplast ribosomes is inhibited by bacterial translation inhibitors such as chloromphenicol and lincomycin (Ellis, 1981). Chloroplast genomes also appear to have some eucaryotic features, in particular, the presence of introns in some tRNA genes (Koch et al., 1981; Steinmetz et al., 1982; Deno et al., 1982; Deno and Sugira, 1984) and rRNA genes (Allet and Rochaix, 1979). Some of these introns appear to be related to mitochondrial mRNA and nuclear rRNA gene introns (Michel and Dujon, 1983). The first example of a chloroplast protein gene with introns was in the eucaryotic alga Euglena gracilis in which the gene for the large subunit of ribulose-1,5-bisphosphate carboxylase (rbcL) was shown to contain at least nine introns (Stiegler et al., 1982; Koller et al., 1984). Since this first report, intervening sequences have also been found in the genes for the 32-kDa photosystem II polypeptide (psbA) (Karabin et al., 1984; Keller and Stutz, 1984) and elongation factor Tu (tufA) (Montandon and Stutz, 1983) from Euglena, the psbA gene from Chlamydomonas reinhardtii (Erickson et al., 1984), and the gene for the ribosomal protein L2 (rp12) from Nicotiana debneyi (Zurawski et al., 1984).

In this paper, we report the DNA sequence of the E. gracilis chloroplast gene for the large subunit of ribulose-1,5-bisphosphate carboxylase (rbcL). The gene has nine intervening sequences of length 382-568 bp. Of the three major classes of introns described for eucaryotic nuclear genes (Cech 1983, 1985), the Euglena rbcL introns appear to be most like the introns found in Nuclear mRNA genes and the group II mitochondrial introns.

MATERIALS AND METHODS

Molecular Cloning and Sequencing of the Euglena rbcL Gene Locus—The rbcL gene locus from E. gracilis Z chloroplast DNA has been cloned into three recombinant plasmids. Two of these plasmids, pEZC38 and pEZC738, have been described previously (Stiegler et al., 1982; Koller et al., 1984). These plasmids contain overlapping chloroplast DNA fragments which map at one end of the EcoRI fragment A (EcoA) of E. gracilis Z chloroplast DNA (Stiegler et al., 1982; Koller et al., 1984). The 5'-end of rbcL is contained within a previously unmapped EcoRI fragment adjacent to EcoA. This fragment was identified among a library of chloroplast EcoRI fragments cloned into the plasmid vector pMB9 (Rodriguez et al., 1976; Gray, 1978). Radioactive labeling of the probe and Southern hybridization versus E. gracilis chloroplast DNA were performed as previously described (Rigby et al., 1977; Southern, 1975; Chelm et al., 1977).

Bacteriophage M13 clones for DNA sequencing were generated by ligation of chloroplast DNA from the three recombinant plasmids described above into the universal cloning sites of the M13 vectors M18, M19, MP18, or MP19 (Messing, 1983; Norgaard et al., 1983). DNA sequence analysis was accomplished by the dideoxy chain termination method developed by Sanger et al., (1977) and modified by Biggin et al. (1983). Either 32P- or 35S-labeled nucleotides were used.

Synthetic Oligonucleotides and Primer Extension Sequencing—Synthetic dideoxy oligonucleotide primers were designed to be complementary to rbcL mRNA, as determined from the DNA sequence. The primers were synthesized manually by a modification of the method of Matteucci and Caruthers (1981) using phosphoramidites purchased from New England Biolabs. Alternatively, some of the primers were synthesized with an applied Biosystems model 380A

* This work was supported by National Institutes of Health Grant GM21351 to R. B. H., and J. C. G. was the recipient of a National Institutes of Health postdoctoral fellowship. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ Current address: Department of Molecular, Cellular, and Developmental Biology, Campus Box 347, Porter Biosciences Building, University of Colorado, Boulder, Colorado 80309.

§ To whom correspondence should be addressed: University Department of Biochemistry, Biological Sciences West 624, University of Arizona, Tucson, AZ 85721.
DNA synthesizer. The synthetic tetradecanucleotides were isolated from urea polyacrylamide gels and were 5'-32P-end labeled as previously described (Nickoloff and Hallick, 1982).

Primer extension dyeoxy sequencing of chloroplast RNA and DNA was based on the procedure of Burke and Re Bhandary (1982) as modified by Taox et al. (1984). Chloroplast DNA and RNA were isolated as described (Chytil et al., 1977; Gray and Hallick, 1979). Two \( \mu \)g of chloroplast RNA or DNA were annealed with 0.5 pmol of 5'-end-labeled primer (1-5 \( \times \) 10^6 cpm/pmol) in 12 \( \mu \)l of annealing buffer (50 mM Tris-Cl, pH 8.3, 60 mM NaCl, 10 mM dithiothreitol). The samples were heated to 90 °C for 3 min and allowed to cool slowly to 37 °C (generally 2-3 h), at which time the samples were placed on ice. Primer extension sequencing of the DNA or RNA was performed as described (Zaug et al., 1984) using AMV reverse transcriptase purchased from Life Sciences Inc. and nucleotides (80 \( \mu \)M for each deoxynucleotide and 80 \( \mu \)M of one deoxynucleotide/reaction) purchased from P-L Biochemicals. Reactions were incubated at 37 °C for 30 min. Then deionized formamide was added to 33%, and the reactions were heated for 3 min in a 90 °C water bath. The reactions were electrophoresed on a 6% denaturing polyacrylamide gel.

Data Analysis—DNA sequence analysis was performed on an IBM PC-XT computer using versions 3.2 and 3.8 of the DNA and protein analysis programs of Mount and Conrad, provided by D. W. Mount, University of Arizona, and the programs described by Pustell and Kafatos (Pustell and Kafatos, 1984).

RESULTS

Molecular Cloning and Sequencing of the E. gracilis Ribulose-1,5-biphosphate Carboxylase Large Subunit Gene—The E. gracilis rbcL locus has previously been mapped to one end of the largest EcoRI fragment (EcoA) of Euglena chloroplast DNA (Koller et al., 1984). It was located via Southern hybridizations using heterologous probes derived from internal regions of the rbcL genes of C. reinhardtii and maize (Stiegler et al., 1982). At least one intervening sequence was predicted for the Euglena rbcL gene based on the distance between the sites of hybridization of 5'- and 3' specific rbcL probes within Euglena chloroplast DNA and the lack of hybridization of the probes to a portion internal to the hybridizing region. When heteroduplexes between the Euglena and spinach chloroplast rbcL loci and Euglena rbcL DNA-mRNA hybrids were examined by electron microscopy, at least nine intervening sequences within the Euglena chloroplast rbcL gene were detected (Koller et al., 1984). The organization of this locus as previously deduced by electron microscopy and limited sequence analysis is shown in Fig. 1B. Since this gene has such complex and novel features for a chloroplast gene, we undertook a DNA sequence analysis of the gene to determine the exact number and sequence of the introns, the relationships of the introns to each other and to introns in other genes, and the organization of the coding region.

To complete the sequence analysis of the 5'-end of rbcL, some additional mapping and cloning of Euglena chloroplast DNA between the EcoA and EcoO fragments was required. It was apparent from the electron microscopy of mRNA-chloroplast DNA hybrids (Koller et al., 1984) that the 5'-end of the Euglena rbcL gene was not contained in EcoA. This was confirmed at the DNA sequence level when codon 16 (subsequently identified as the beginning of exon 2) was found starting 27 bp from the end of EcoA (described below). Based on electron microscopy of the rbcL mRNA-DNA hybrids, we expected another exon (exon 1) to be approximately 500 bp away from EcoA within an unmapped EcoRI fragment located in the gap between EcoA and EcoO. A subsequent search of a chloroplast DNA EcoRI fragment library and additional restriction fragment mapping revealed two EcoRI fragments within this gap: EcoX (950 bp) and EcoAA (450 bp). Upon DNA sequence analysis of EcoX the coding region expected for the amino-terminal 15 amino acids (exon 1) of rbcL was found. It was located approximately in the center of this 950-bp fragment, consistent with the earlier electron microscopy data (Fig. 1B). EcoX was later confirmed as being adjacent to EcoA by sequencing of chloroplast DNA through the EcoRI site separating fragments EcoA and EcoX using a synthetic tetradecanucleotide primer complementary to exon 2 (data not shown). Therefore, the EcoRI restriction map of the region of Euglena chloroplast DNA of interest is EcoO-EcoAA-EcoX-EcoA, with EcoX and EcoA coding for the amino and carboxyl termini of rbcL, respectively.

DNA Sequence of rcbl—The complete DNA sequence of the Euglena rbcL locus was determined by the dyeoxy chain termination method. The rbcL sequencing strategy and map of the introns and exons are shown in Fig. 1. C and D. All restriction enzyme cloning sites have been overlapped. Approximately three-quarters of the gene has been sequenced from more than one cloning site. Of this, approximately one-third of the gene was sequenced from both strands. The map of exons and introns is co-linear with the earlier map derived from electron microscopy, except that the relationship between the length of the electron microscopy images and chloroplast DNA base pairs was overestimated by approximately 12%.

Fig. 2 shows 6225 bp of Euglena chloroplast DNA, including the 969-bp EcoX fragment and 5256 bp from one end of EcoA. This sequence encompasses 548 bp of leader sequence before the ATG initiator codon, the 5561-bp coding region, comprised of 10 exons separated by nine introns, and 115 bp of 3'-distal sequence containing the sequences coding for the 3'-end of the mRNA. The exons are underlined in this figure. The properties and coding potential of the exons are described in detail in the accompanying paper (Gingrich and Hallick, 1985).

Primer Extension Sequence Analysis of mRNA Splice Boundaries—Initially, mRNA splice sites and, therefore, the boundaries between exons and introns in the rbcL sequence...
were predicted based on maintaining maximum derived protein sequence homology and no insertion or deletion of amino acids compared with the non-intron-containing rbcL genes from other alga and plant sources. When the transition regions were compiled, the introns were found to contain closely related or identical boundary sequences (see Table I). Eight of the nine introns contain an in-frame termination codon of TAA or TAG within the 3'-conserved splice boundary (Table I). This is an indication that the proposed exon-intron boundaries must at least be very near the sites chosen. This compilation is an extension of our previous splice boundary predictions based on a very limited amount of DNA sequence data concerning 4 of the 9 splice sites (Koller et al., 1984).

In order to confirm the putative splice sites, primer extension sequence analysis with the spliced mRNA as template was employed to determine the sequence of the rbcL mRNA at selected exon boundaries. For this type of experiment, tetradecanucleotides were synthesized such that they would complement a sequence within the mRNA 3’ distal to a splice donor sequence analysis with the spliced mRNA as template. The primers used were complementary to exon 2 (bases 1027-1040, Fig. 2), exon 5 (bases 2815-2828, Fig. 2), and exon 6 (bases 4645-4658, Fig. 2). With the latter two primers 5’-end labeled, annealed to selected exon boundaries. For this type of experiment, tetradecanucleotides were synthesized such that they would complement a sequence within the mRNA 3’ distal to a splice donor sequence analysis with the spliced mRNA as template.
interm 1–9 are, respectively, 404, 514, 513, 568, 413, 479, 382, 420, and 441 bp. These compare closely in size to introns 1–4 of the Euglena psbA locus, which are 435, 443, 434, and 617 bp, respectively (Karabin et al., 1984). Second, the rbcL introns are very AT rich and closely homogenous in base composition. The introns range from 82–86 mol% A + T. By contrast, the rbcL exons are only 61 mol% A + T. Third, the distribution of bases within the intron is decidedly nonrandom. In every intron except intron 9, the transcribed strand has a T > A > G > C base bias. The range for the composition of bases within individual introns is 42–48% T, 33–43% A, 8–10% G, and 4–8% C, with average values of 47, 37, 9, and 7%, respectively. For intron 9 only, A (45%) > T (42%) > G > C. Such base bias could either be a characteristic property of Euglena introns, or it could reflect a DNA strand specific base bias in the chloroplast genome. Introns of the psbA locus have the same type of base bias as rbcL, i.e., 43–47% T, 34–38% A, 10–12% G, and 7–8% C. Three other introns from the Euglena psbC and psaA loci also shown this same base bias. Since psbA, psbC, and psaA are all coded on the opposite strand of Euglena chloroplast DNA as rbcL, we conclude that the T > A > G > C base composition is a characteristic feature of the introns per se. This property may have useful predictive value in identifying the coding strand in other regions of this chloroplast genome.

The introns were examined for potential open reading frames. None of the introns contain long open reading frames that are a 3′ continuation of the exons. The longest is a 34-codon extension of exon 3. Nor do the introns display long internal open reading frames. The longest beginning with an initiator ATG is of 51 codons within exon 2. There are no other orfs starting with a Met codon longer than 25 codons. In view of the high AT base content of the introns and the lack of significant open reading frames, it is unlikely that the introns code for any polypeptides.

**DISCUSSION**

The *E. gracilis* gene for the large subunit of ribulose-1,5-bisphosphate carboxylase was the first chloroplast protein gene demonstrated to contain introns. It is presently the most complex of the known chloroplast genes, containing 10 exons, separated by 9 AT-rich introns with highly conserved boundary sequences ranging in size from 382 to 568 bp. The 5559-bp coding region occupies approximately 4% of the 145-kbp *Euglena* chloroplast genome. By contrast, the non-intron-containing rbcL loci of higher plants require only approximately 1% of the coding capacity of the chloroplast genomes.

Introns with boundary sequences similar to the rbcL introns have been found in other *Euglena* chloroplast protein genes. The psbA gene (which codes for the 32-kDa photosystem II...
herbicide-binding protein) contains four introns of similar size as the rbcL introns and with essentially identical boundary sequences (Karabin et al., 1984). The gene for elongation factor Tu (tufA) contains two introns (Montandon and Stutz, 1983). Genes for both photosystem I (psaA) and photosystem II (psbC) reaction center polypeptides also have introns with similar boundary sequences (Hallick et al., 1984). At present there are at least 18 known introns of Euglena chloroplast protein genes characterized at the DNA sequence level for these five genes. In a global analysis of the Euglena chloroplast genome, based on the electron microscopy of chloroplast DNA-mRNA hybrids, Koller and Delius (1984) were able to detect up to 50 introns scattered around the 145-kbp circular genome. We would consider this to be a very realistic, and perhaps conservative, estimate. The Euglena introns all seem to have similar properties. A possible exception to this generalization may be the tufA locus. The two introns in this gene are quite short (105 + 110 bp) (Montandon and Stutz, 1983). The actual splice points for the transcript of this gene have not yet been reported, but at least one of the introns does not seem to have the potential conserved boundaries described for the other Euglena chloroplast protein genes.

Recently, introns in some higher plant chloroplast mRNA and tRNA genes have also been found to contain boundary sequences similar to the Euglena mRNA introns (Hallick et al., 1984). Among these are the ribosomal protein L2 from spinach (Zurawski et al., 1984) and the atpF gene from wheat (Bird et al., 1985). Therefore, Euglena is not unique among chloroplast-containing organisms in having introns with conserved boundary sequences in protein genes. A novel new feature in the plant chloroplast examples is the observation that protein genes and some tRNA genes seem to have the same type of introns.

Conserved boundary sequences were first described for introns of eucaryotic nuclear mRNA genes. In most cases, the boundary nucleotides are 5'-GTGAG...PyGAG3' (Breathnach and Chambon, 1981). As discussed earlier (Koller et al., 1984) the E. gracilis rbcL introns fit well with this consensus except for one important base. The two bases at either ends of eucaryotic mRNA introns are essentially invariant. The Euglena rbcL introns, however, have a T instead of a G at their 3' termini. For rbcL intron 7, the last base is an A. Two of the psaA introns have as the final intron base a C. The chloroplast introns also lack the conserved bases at the ends of the exons at the exon-intron boundaries that are characteristic of the eucaryotic nuclear genes. Therefore, the Euglena chloroplast introns are related to their counterparts in the protein genes of the eucaryotic nucleus but have distinctive features of their own.

More recently, Cech (1985) has noted that the boundaries of all group II mitochondrial introns might have the consensus sequence 5'-GTGCC...YAT, where the Ys is a pyrimidine-rich sequence that contains 4–10 residues. This sequence is more closely related to the chloroplast than the nuclear mRNA intron boundary sequences. Keller and Michel (1985) have also described a potential relationship between group II mitochondrial introns and Euglena chloroplast psaA introns. This relationship between nuclear, mitochondrial and chloroplast introns may be suggestive of a common origin for the introns and possibly common features in the splicing mechanism.

In addition to conserved boundary sequences, we were interested in other aspects of the rbcL intron sequences or structure that might relate either to their evolutionary origin or the features of the RNA-splicing mechanism. A number of recent reports have described a conserved sequence within nuclear mRNA introns which facilitates the splicing of this class of introns. In yeast nuclear introns, a conserved sequence (TACTAACA) is located 20–50 bp from the 3'-end of the introns (Langford and Gallwitz, 1983). Deletion or mutation of this sequence prevents splicing (Langford et al., 1984).

Similar sequences have been reported in introns from other organisms (Keller and Noon, 1984). During in vitro intron splicing a “lariat” is formed via a 2'-5' phosphodiester bond between the pentultimate A in the yeast TACTAACA (GraROWSKI et al., 1984; RUSKIN et al., 1984) sequence and the 5' G of the intron (Keller, 1984). This may be facilitated by small nuclear ribonucleoproteins (Keller and Noon, 1984). A compilation of the sequences as the near the 3'-ends of the Euglena rbcL introns is given in Table II. The most highly conserved sequence we observe in this region is the sequence GTTGGAAAA starting 35–55 bases upstream from each of the splice junctions. Although this sequence bears little resemblance to the yeast TACTAACA sequence, immediately 5' to the GTTGGAAAA sequence is a consensus sequence TACA. This is similar to the first 4 bases of the TACTAACA sequence. Note that the rbcL intron TACA sequence is not 100% conserved as is the yeast TACTAACA sequence. In other organisms, the consensus sequence analogous to the yeast TACTAACA is also not 100% conserved (Keller and Noon, 1984). We have observed sequences similar to the consensus shown in Table II in other introns from Euglena and higher plants (data not shown). The possibility can now be considered that splicing of at least one category of chloroplast introns might involve a mechanism like group II mitochonial introns, perhaps via a lariat intermediate.

Were the Euglena rbcL introns added after the evolution of the gene? Since there is very high sequence conservation between the procaroyte (Arababna variabilis), dicot (spinach), monocot (maize), eucaryotic algal (Chlamydomonas), and prokaryote (Euglena) ribulose-1,5-bisphosphate carboxylase/oxygenase polypeptides (Gingrich and Hallick, 1985), it is reasonable to conclude that all of these enzyme subunits evolved from a common ancestral gene. If the introns now found in the Euglena rbcL gene were present in such an ancestral gene, it would be necessary to postulate that the lack of introns in all the other known rbcL loci is the result of precise and identical deletion events that occurred independently in all organisms except Euglena. A more likely hypothesis is that the ancestral gene common to plants, algae, and cyanobacteria lacked introns and that the Euglena introns were added to the gene in more recent evolutionary times. This concept of the insertion of introns to a chloroplast gene following evolutionary descent from an ancestral gene is further strengthened by consideration of the chloroplast psaA locus, coding for the highly conserved 32-kDa protein of photosystem II. The gene lacks introns in monocots, dicots, and cyanobacteria but contains four introns in both Euglena (Karabin et al., 1984) and Chlamydomonas (ERICKSON et al., 1984). However, unlike the Euglena introns which resemble nuclear mRNA and mitochondrial group II introns the Chlamydomonas introns have been described as group I introns as defined by Michel et al. (1982) and Davies et al., (1982), and they occur at completely different sites within the psaA locus (ERICKSON et al., 1984).

Acknowledgments—We wish to thank Michael Farley for mapping of the EcoX chloroplast DNA fragment and Will Downs for technical services.
assistance. We also wish to thank Jeri Beltman and Marv Caruthers for some of the synthetic DNA primers used in these experiments.

REFERENCES

Alet, B., and Rochaix, J.-D. (1979) Cell 18, 55-60