Chloroplast gene for M_r 32000 polypeptide of photosystem II in *Euglena gracilis* is interrupted by four introns with conserved boundary sequences

Gerald D. Karabin, Michael Farley and Richard B. Hallick*

Department of Chemistry, University of Colorado, Campus Box 215, Boulder, CO 80309, USA

Received 18 May 1984; Accepted 28 June 1984

ABSTRACT

The gene for the M_r 32000 herbicide binding polypeptide of photosystem II has previously been mapped to the 5 kbp *Eco*RI fragment *Eco* I of *Euglena gracilis* chloroplast DNA. The nucleotide sequence of 3324 bp of *Eco* I, containing the *psbA* locus, has been determined. This locus encodes a polypeptide of 345 amino acids which is co-linear with, and has 86% derived amino acid sequence homology to sequences derived from four higher plants chloroplast *psbA* loci. The *Euglena* *psbA* gene contains four introns of size 435, 443, 434, and 617 bp. The four introns have conserved boundary sequences of the type previously described in the *Euglena* chloroplast gene (*rbcL*) for the large subunit of ribulose-1,5-bisphosphate carboxylase (Koller et al., Cell 36, 545-553. 1984).

INTRODUCTION

A chloroplast thylakoid membrane polypeptide with an apparent M_r 32000-36000 has been variously designated peak D (1), D2 (2), 32 kd polypeptide (3), and photogene polypeptide (4). It is believed to function in electron transport on the reducing side of photosystem II (5,6). The 32 kd polypeptide is a target receptor for inhibition of photosynthetic electron transport by triazine and diuron herbicides (5,6). The polypeptide is highly conserved among seven different angiosperms, and *Chlamydomonas* as judged by partial proteolysis digestion patterns, and protease sensitivity of membrane exposed polypeptide domains (3). The 32 kd polypeptide is derived from 33.5-34.5 kd precursors (4,7). In *Spirodela*, polypeptide processing is known to involve cleavage of COOH-terminal amino acids (8). Although the 32 kd polypeptide is a major, rapidly labeled translation product of chloroplast ribosomes in light-grown plants (reviewed in 9), it has a high turnover rate and does not accumulate at high levels.

The gene for the 32 kd polypeptide is coded in chloroplast DNA in photosynthetic eucaryotes (10). The gene, recently designated psbA (11,12), has been mapped to a locus adjacent to one of the inverted repeats in the large single copy region of chloroplast DNA from maize (10), mustard (13), spinach (14), and other angiosperms (15), to the inverted repeat in *Chlamydomonas reinhardtii* (16), and to the single copy region of *Euglena gracilis* (see 17). In the cyanobacteria *Anabena*, several copies of closely related psbA loci are found (18). The complete DNA sequence of psbA genes from spinach (11), *Nicotiana debeynei* (11), triazine sensitive and resistant biotypes of *Amaranthus hybridus* (19), soybean (20), mustard (21), and two variants from *Anabena* (18) have been determined. The triazine resistant *A. hybridus* biotype correlates with a single amino acid substitution in psbA (19).

We have been studying psbA and its RNA transcripts in *Euglena*. The RNA transcripts from the *Euglena* chloroplast gene are unusual in that in addition to a mature 14S mRNA characteristic of the other known genes (9,11,14,15), mRNA precursors of 3.1, 2.8, 2.3, and 1.8 kb were also detected in RNA blot hybridization experiments. These precursors could be interpreted as unspliced precursor transcripts of a gene with multiple introns (17). To determine the nature of the amino acid coding and intervening sequence regions of *Euglena* psbA, we have determined the complete nucleotide sequence of the gene. It consists of five exons separated by four introns ranging in size from 434-617 bp.

MATERIALS AND METHODS

Plasmid DNA.

pEZC514 plasmid DNA consists of the 5 kbp *EcoRI* fragment *Eco* I ligated into the *EcoRI* site of pBR322 (17). Plasmid DNA was isolated by a cleared lysis procedure (22) and purified by isopycnic centrifugation in CsCl-ethidium bromide (23).

DNA Sequence Determination.

Most of the psbA sequence was determined by the chemical cleavage methods of Maxam and Gilbert (24). DNA was labeled either at 5'-phosphates via a polynucleotide kinase reaction, or

at 3'-ends using α - 32 PdNTP(s) and the Klenow fraction of *E. coli* DNA polymerase I. Dideoxy, chain termination DNA sequence analysis was employed to determine the DNA sequence of psbA exon 2, and adjacent intron sequences. The two tetradecanucleotides, 5'-CTAGGAACAATTTC and 5'-GATGTTATTCTTTC were chemically synthesized using the Applied Biosystems DNA Synthesizer. They were used to prime DNA synthesis on electrophoretically separated strands of *Eco* I DNA. Hybridization of primers to templates, and chain termination sequence reactions were as described (25, 26).

RESULTS AND DISCUSSION

DNA Sequence of *Euglena* Chloroplast psbA Locus

The psbA gene has previously been mapped to a region of approximately 3 kbp within the 5 kbp *Eco*RI fragment *Eco* I of the *Euglena gracilis* chloroplast genome (17). A restriction map of this locus, and the DNA sequencing strategy employed in this study are shown in Figure 1. A continuous sequence of 3324 bp, which includes 86 bp of leader sequence preceeding the putative ATG initiator codon (see below), the 2967 bp psbA locus, and 269

Fig. 1. Restriction endonuclease map of the psbA coding locus of the *Eco*RI fragment *Eco* I of *Euglena* chloroplast DNA. Top: The organization of the psbA locus. Exons are shown in filled boxes, and introns in open boxes. Middle: Restriction map for several enzymes determined by standard procedures. Bottom: DNA sequence strategy. Vertical lines, 5'-labeling; Open circles, 3' labeling; Closed circles, primed dideoxy sequencing. Of the 345 psbA codons, 245 were determined for both DNA strands, and 68 additional codons were determined from multiple sequencing of the same strand.

bp 3'-distal to the gene is shown in Figure 2. The entire sequence was determined from DNA restriction fragments isolated from the plasmid pE2C514. A tRNA^{Leu} gene is located immediately 5'-proximal to the sequence shown in Figure 2. It has the same polarity as psbA (J. A. Nickoloff, G. D. Karabin, and R. B. Hallick, in preparation). The trnL-psbA spacer length is 86 bp.

Identification of psbA Exons and Introns

When the amino acid coding potential of the psbA locus was examined, five regions of high amino acid sequence homology with the spinach psbA locus were identified (Figure 2). These were interpreted as the five exons of a split psbA gene, which is interrupted by four introns. The four introns range in size from 434 to 617 bp and are punctuated by multiple termination codons in each reading frame. The introns are very AT rich (81 to 84 mol%).

The five putative exons range in size from 13 to 193 codons. When the longest possible open reading frame following an ATG codon and involving the five exons is considered, the resulting polypeptide contains 345 amino acids compared to 353 in the spinach chloroplast gene. This length heterogeneity is due to one additional codon at the amino terminus and nine fewer codons at the carboxyl terminus. A comparison of the amino acid sequences of the polypeptide products of the psbA loci of spinach and *Euglena* is given in Figure 3. The overall amino acid sequence homology with the spinach polypeptide is 86%. The length of the individual exons and their homology with spinach psbA are as follows: exon 1, 79 codons, 73%; exon 2, 28 codons, 79%; exon 3, 193 codons, 90%; exon 4, 33 codons, 97%; exon 5, 13 codons, 100% (excluding the nine additional spinach codons). The homology progressively increases toward the COOH-terminal end of the molecule.

Conserved Intron Boundary Sequences

Two considerations were involved in the identification of exon-intron boundaries. These are the very high conservation of amino acid sequence (as determined from DNA sequence analysis) of this polypeptide from sources as diverse as *Euglena*, higher plants, and cyanobacteria, and the conserved exon-intron boundary sequences of psbA. The proposed *Euglena* exon-intron boundaries,

-86	-80	-70	-60	-50	-40
TATATA	ATAAAATATT	TAATTTTSTA	TGTTATAATT	TTATCAACAG	TCGCGTTTGC
-30	-20	-10		15	
GACTGGGAGA	TTAATTTAAT	TAATTTATTG	ATG ATT	TCA CCT GTT	TTA AAG AAA TAT
			Met Ile	Ser Pro Val	Leu Lys Lys Tyr
30	45		60	75	
GCA AGA CCT	AGT CTT TGG	TAT CGT TTT	TGT GCT TGG	GTA GCT TCA	AAG AAA AAT
Ala Arg Pro	Ser Leu Trp	Tyr Arg Phe	Cys Ala Trp	Val Ala Ser	Lys Lys Asn
90	105	120	135		
CGT CTT TAT	GTA GGA TGG	TTC GGT GTT	TTG ATG ATT	CCA ACT TTA	CTT ACA GCT
Arg Leu Tyr	Val Gly Trp	Phe Gly Val	Leu Met Ile	Pro Thr Leu	Leu Thr Ala
150	165	180			
GCA ACT GTA	TTT ATT ATT	GCT TTC ATT	GCA GCT CCG	CCT GTT GAT	ATC GAT GGT
Ala Thr Val	Phe Ile Ile	Ala Phe Ile	Ala Ala Pro	Pro Val Asp	Ile Asp Gly
195	210	225	240		
ATT CGT GAG	CCT GTG TCT	GGA TCT CTT	TTT TAT GGA	AAT AAT ATA	A GTGCG
Ile Arg Glu	Pro Val Ser	Gly Ser Leu	Phe Tyr Gly	Asn Asn Ile	I
250	260	270	280	290	300
TAACCTTATCT	AATTGCCTAC	TATCAATTTG	AAAAATCCTC	ATTTGTTTCT	TCTATTTTAT
310	320	330	340	350	360
TTTAAAGAGG	TGAATTTTAA	AAATTTAATA	AAAACCTTTA	TTAGTAAATT	TCTAATTATA
370	380	390	400	410	420
TTATATAATA	TTTTAATGTT	TTTTATGTAA	AGTTTAATTA	TTTATTTCTT	GATTTTTTAAA
430	440	450	460	470	480
ATAAAAGTAT	TAAAAAAGGT	TTTATATATA	TAAAGCNAAA	AAAAGTTATA	GAGCTTTTAAA
490	500	510	520	530	540
ATATGTTTTA	TTTTAAATTC	AAAAAATTTA	AATTTAATAA	GGCTAGGAAC	AATTTCTAAT
550	560	570	580	590	600
TTTAATAATT	TTATTTCTTT	ATTTAAAGTT	TTTTTAGATA	GCCGAAGAGT	TTTTTTTTTAA
610	620	630	640	650	660
TTATTTTCGG	TTAGAAAAAA	GGTACGTTTT	ATATAACGAT	AAAAGTAAAG	TTATTTAGTA
670	675	690	705		
TCTATTTTAC	TT ACT GGT	GCT GTA GTT	CCA ACT TCA	AAT GCT ATT	GGA TTG CAT
	le Thr Gly	Ala Val Val	Pro Thr Ser	Asn Ala Ile	Gly Leu His
720	735	750	760		
TTT TAT CCT	ATT TGG GAA	GCA ACT TCA	TTA GAT GAA	TGG GTGTGTC	
Phe Tyr Pro	Ile Trp Glu	Ala Thr Ser	Leu Asp Glu	Trp	
770	780	790	800	810	820
AATTAGAATT	TAATTCGTTT	AATAACTTTG	AACTACATTT	GATTAGACGA	AAAAATTAAC
830	840	850	860	870	880
AAAGTATAGT	TTTTTCATTA	AGGTTAAATT	TTAAGCGTAA	TAAGTAATAT	TTCTTCTTTT
890	900	910	920	930	940
AAGTTCITTT	TAAAGTATAT	TTTAGTAAAT	TATTTTCATAT	TTTTTAATAG	CATTNNNAAA
950	960	970	980	990	1000
AACTTTAATT	TTTGTTTTAA	GTAAATTTTA	TCGTTGATTT	TTTAAGGAAA	AATATTAGTA
1010	1020	1030	1040	1050	1060
GTAGAAAGAA	ATTTTATTATG	AAAGAATAAC	ATCAATAAAT	TCGTAGTAAA	ATGAAATGAT
1070	1080	1090	1100	1110	1120
TAGACACTTG	GTCTAGTAGA	TTTCAATTTA	TATATAAAGA	GCTTTATATT	TGGTAACAAA
1130	1140	1150	1160	1170	1180
TTTTCACAGT	TCGTAAATGA	AGTAATTAAG	TATATAACTT	TTTAAGAACG	TTAATAGATT

Fig. 2.

```

      1190      1196      1205      1220
TTTTACTTTA  GTTTAT   TTG TAT AAT GGT GGT CCT TAT CAA TTA ATC GTA TGC
                Leu Tyr Asn Gly Gly Pro Tyr Gln Leu Ile Val Cys
1235      1250      1265      1280
CAT TTC TTT ATC GGT ATT TGT TCT TAT ATG GGA AGA GAA TGG GAG CTT TCA TTC
His Phe Phe Ile Gly Ile Cys Ser Tyr Met Gly Arg Glu Trp Glu Leu Ser Phe
      1295      1310      1325      1340
CGA TTA GGA ATG CGC CCG TGG ATT GCA GTT GCT TAT TCT GCT CCT GTT GCT GCT
Arg Leu Gly Met Arg Pro Trp Ile Ala Val Ala Tyr Ser Ala Pro Val Ala Ala
      1355      1370      1385
GCA AGT GCT GTA TTT ATT GTT TAT CCT TTA GGT CAA GGT TCT TTT TCA GAT GGT
Ala Ser Ala Val Phe Ile Val Tyr Pro Leu Gly Gln Gly Ser Phe Ser Asp Gly
1400      1415      1430      1445
ATG CCT TTA GGT ATT TCA GGT ACT TTT AAC TTT ATG ATT GTT TTC CAA GCT GAA
Met Pro Leu Gly Ile Ser Gly Thr Phe Asn Phe Met Ile Val Phe Gln Ala Glu
      1460      1475      1490
CAT AAT ATT TTA ATG CAT CCA TTT CAT ATG CTT GGT GTT GCT GGT GTT TTT GGT
His Asn Ile Leu Met His Pro Phe His Met Leu Gly Val Ala Gly Val Phe Gly
1505      1520      1535      1550
GGT TCT CTT TTC TCT GCT ATG CAT GGT TCT TTG GTA ACA TCT ACT TTG CTT CGT
Gly Ser Leu Phe Ser Ala Met His Gly Ser Leu Val Thr Ser Ser Leu Leu Arg
      1565      1580      1595      1610
GAA ACG ACT GAA AAT GAA TCA ATA AAT GTT GGT TAC AAG TTT GGT CAA GAA GAA
Glu Thr Thr Glu Asn Glu Ser Ile Asn Val Gly Tyr Lys Phe Gly Gln Glu Glu
      1625      1640      1655
GAA ACA TAT AAT ATT ATT GCT GCT CAC GCT TAT TTT GGT CGT TTA ATC TTC CAA
Glu Thr Tyr Asn Ile Ile Ala Ala His Ala Tyr Phe Gly Arg Leu Ile Phe Gln
1670      1685      1700      1715
TAT GCG TCG TTC AAT AAT TCA CGT TCA TTG CAT TTC TTT TTA GCT GTT TGG CCT
Tyr Ala Ser Phe Asn Asn Ser Arg Ser Leu His Phe Phe Leu Ala Val Trp Pro
      1730      1745      1760
GTT GTT GGT ATT TGG TTT ACA GCA TTA GGT GTT TCA ACT ATG GCA TTT AAT TTA
Val Val Gly Ile Trp Phe Thr Ala Leu Gly Val Ser Thr Met Ala Phe Asn Leu
1775 1780      1790      1800      1810      1820      1830
AAC GTGCG TAATTTGATC TAATGGTTTT TTTAGTATTG TAAGTTATGT TTTTTTACTG
Asn
      1840      1850      1860      1870      1880      1890
AAAAATAATT TAATAACATT ATAATTACAT TATATATTTT ATGTTTTTAT TTTAGTATTT
1900      1910      1920      1930      1940      1950
TATTTTAATT CAAAAATTTT TTAATTTAAA TATTTTTTGA TTGTAAATTA CGTTATTGTT
1960      1970      1980      1990      2000      2010
TTCATTATTT CCATTATGAA AAATGATAAT TCAACAGTAG TTTTAGTTGA AGAGTTGTAA
2020      2030      2040      2050      2060      2070
GAATAAGTTT ATTATATTGT GTGTTTTTTT GTTTTTAGAA AAGGCGTTTG CAAATTTTCAT
2080      2090      2100      2110      2120      2130
AAATTTTAT TTTTAGATTT CGTAAAAAAA CAATTTAAAA CTAAACTCT CAAATTTGAA
2140      2150      2160      2170      2180      2190
ATAAAAAATA AAAGCTTTGT ATATAGAAAT ATATTCTCGT AGTTTGATATC CAGCAAACTA
2200      2209      2215      2230      2245
TTTTGTTTGC TAGTGTTAT GGT TTT AAT TTT AAT CAA TCT GTT ATT GAT TCA CAA
                Gly Phe Asn Phe Asn Gln Ser Val Ile Asp Ser Gln
      2260      2275      2290
GGT CGC GTT ATC AAT ACT TGG GCT GAT ATT ATT AAT AGA GCA AAT TTA GGT ATG
Gly Arg Val Ile Asn Thr Trp Ala Asp Ile Ile Asn Arg Ala Asn Leu Gly Met

```

2305	2320	2330	2340	2350
GAA GTT ATG GT	GCGACAAGTT	TAAAGTAAAA	AAGCGTATTT	TTAACATTTT
Glu Val Met				
2360	2370	2380	2390	2400
TATTTACTTT	AAAAATATGA	TTCTTTATGA	AGTATTTTAA	TTGTATAGTC
2420	2430	2440	2450	2460
AATCCTTCAG	TCTTTCCTTT	TTTAAAGTCT	GATATTTATT	TTTAATATAG
2480	2490	2500	2510	2520
AAATTAGCAA	AAAATATAAA	TTTATTTGAT	AAAAATTATT	TATTTTATT
2540	2550	2560	2570	2580
AATTTATGGA	TTTTGCTTCA	AATAGTCGTT	TTTTTAAGAG	GGTAATATCA
2600	2610	2620	2630	2640
CAATTTTAT	TTATTACTTA	ATACTTCTTA	TCGTAAAGTC	CATAAATTTT
2660	2670	2680	2690	2700
TTTGAGTGTA	CTATATGGAA	CTTGGTAAAA	CGTTATATTT	GATGAAGTTT
2720	2730	2740	2750	2760
TATTAATAAT	ATAATAACGT	AGGNAAGTTA	TAAAAAAATT	AATAAAAAATA
2780	2790	2800	2810	2820
TATTTTTAAA	TTTTTCAAT	TACAGACTTT	GATTTTCGTT	TTCTTACGTT
2840	2850	2860	2870	2880
ATTAATAATT	AAATTAAAGT	TTTGAGCCTT	GTGATATTAA	AATTTTCATGC
2900	2910	2920	2925	2940
GACGATATTT	TTAGAATTGA	AAATTTTAGT	TTAAC	CAC GAA CGT AAT GCA CAT AAT
				His Glu Arg Asn Ala His Asn
2955	2970	2980	2990	3000
TTC CCA TTA	GAT TTG GCT	TAA TTT	TATATACTTG	TTTTTATCTT
Phe Pro Leu	Asp Leu Ala	Ter		TTTTACTAAC
3010	3020	3030	3040	3050
AAGTATATTT	TTTGTTTTGT	TAATTATTTT	TTAGTAAGTA	GATTTTAAAT
3070	3080	3090	3100	3110
AACGTTTTTA	ATAATAAAAA	TCGAATTATT	TGGTAAAAAG	CTTTTTTTTA
3130	3140	3150	3160	3170
CTTTTTTGA	AAATTATATT	AAATTTCTTT	CGATTTTAAA	AATATTTAAC
3190	3200	3210	3220	3230
ACATAATTTT	TTTCTTATTT	TATATTTTTT	ACTACTTCAG	TTTTTTTAAT
				TTCTTT

Fig. 2. DNA sequence of the *Euglena* chloroplast psbA locus. The RNA-like strand of the DNA is shown in the 5' to 3' direction. Amino acid translation of the psbA exon sequences is written below the DNA sequence.

which were chosen to maintain amino acid sequence homology to all known genes, are within codon 79 for exon 1-2, and between codons 106-107, 299-300, and 332-333 for exons 2-3, 3-4, and 4-5, respectively. If these proposed boundaries are correct, the *Euglena* polypeptide is co-linear with the higher plant and cyanobacterial polypeptides, with no internal insertion or deletion of amino acids. In addition, these exon boundaries would not result in a single amino acid substitution at the splice junctions compared to any of the other five known genes.

The conserved nature of the sequences at the 5' and 3' ends

```

      10      20      30      40      50      60      70
MISPV LKKYARPSLWYR FCAWVASKKNRLYVGV FGLMIPTLLTAA TVFIIAFIAAPPVDIDGIREPVSGSLFYG
MTAILERRESESLWGRFCNWITSTENRLYIGWFGV LMIPTLLTATSVFIIAFIAAPPVDIDGIREPVSGSLLYG
*** ***** * * * * * * * * * *
      80      90      100     110     120     130     140     150
NNIITGAVVPTSNAIGLHFYPIWEATSLDEWLYNGG PYQLIVCHFFIGICSYMGREWELSFRLGMRPWIAVAYSA
NNIISGAIPTSA AIGLHFYPIWEAASVDEWLYNGG PYELIVLHFLLG VACVYMGREWELSFRLGMRPWIAVAYSA
* * * * * * * * * * * * * * * *
      160     170     180     190     200     210     220
PVAASAVFIVYPLQGGSFSDG MPLGISGTFNFMIVFQAEHNILMHPF HMLGVAGVFGGSLFSAMHGSLVTSSLL
PVAATAVFLIYPIGQGSFSDG MPLGISGTFNFMIVFQAEHNILMHPF HMLGVAGVFGGSLFSAMHGSLVTSSLI
* * * * * * * * * * * * * * * *
      230     240     250     260     270     280     290     300
RETTENESINVGYKFGQEEETYNIIAAHAYFGR LIFQYASFNNRSRLHFFLAVWPVVG IWF TALGVSTMAFNLNG
RETTENESANEGYRFGQEEETYNIVAAHGYFGR LIFQYASFNNRSRLHFFLA AWPVVG IWF TALGISTMAFNLNG
* * * * * * * * * * * * * * * *
      310     320     330     340     350
FNFNQSVIDSQGRVINTWADIINRANLGM EVMH ERNAHNFP L DLA
FNFNQSVVDSQGRVINTWADIINRANLGM EVMH ERNAHNFP L DLA AIEAPSTNG
* * * * *

```

Fig. 3. Comparison of the amino acid sequences of the polypeptide product of the psbA locus of *Euglena* and spinach chloroplast DNAs. The sequence for *Euglena* is shown on the top line and spinach on the bottom line in the single letter amino acid code. Differences in sequence are noted with asterisks. Codon numbers are given for the *Euglena* sequence.

of the four psbA introns is illustrated in Figure 4. The consensus sequence 5'-GTGYG is found at the 5'-end of each intron. This is essentially the same sequence that we reported earlier for introns in the *Euglena* chloroplast rbcL locus (27). The sequence 5'-GTGYG does not occur within any of the psbA exons. The only other occurrence of this sequence in the 3324 bp sequence shown in Fig. 2 in either polarity is within intron 3 (Fig. 2, 2029-2033). The 3'-ends of the introns are slightly more variable. The psbA consensus sequence of NTT^APutTTTAY-3' is however very similar to the rbcL consensus sequence of TTTTAPuTTTAT. There is no apparent conservation of DNA sequence within the exons near the exon-intron boundaries. The proposed splice junction for exon 2-3 has been confirmed by primer extension DNA sequence analysis of spliced psbA mRNA (U. Johanningmeier and R. B. Hallick, unpublished observation).

Protein Molecular Weight and Codon Usage

The calculated molecular weight of the polypeptide encoded by the open reading frame of 345 codons (Fig. 3) is 38321. This is considerably larger than the predicted 32000 MW of the mature

	EXON	INTRON		EXON
psbA 1	ATAATATAA	GTGCGTAACTTATC	 AGTATCTATTTTAC	TTACTGGTG
psbA 2	GATGAATGG	GTGTGTCAATTAGA	 TTACTTTAGTTTAT	TTGTATAAT
psbA 3	AATTTAAAC	GTGCGTAATTTGAT	 TTTGCTAGTGTTAT	GGTTTTAAT
psbA 4	GAAGTTATG	GTGCGACAAGTTTA	 AATTTTAGTTTAAAC	CACGAACGT
Consensus		GTGTGGATTTT	TTTARTTTTAT	

Fig. 4. Conserved exon-intron boundary sequences in *Euglena* chloroplast psbA locus. The psbA introns (1-4) are closely related to the previously reported *Euglena* chloroplast rbcL intron boundaries (27). A consensus set of 5'- and 3'- intron boundary sequences for the psbA and rbcL introns is given at the bottom of the figure.

polypeptide in higher plant chloroplasts (3). Some of this size discrepancy may be the consequence of post-translational processing of the primary translation product, including COOH terminal proteolysis (8). The relatively low amino acid sequence homology at the NH₂-termini among various psbA polypeptides may be an indication of processing here also.

The codon usage in the *Euglena* chloroplast psbA locus is given in Figure 5. There are two noteworthy features about the pattern of codon usage. First, the *Euglena* gene encodes five lysine residues. Four of these are found in exon 1 among the first 27 residues. If translation of *Euglena* psbA mRNA begins with the first ATG as was proposed for spinach (11), and not the

TTT-Phe	18	TCT-Ser	10	TAT-Tyr	13	TGT-Cys	2
TTC-Phe	10	TCC-Ser	0	TAC-Tyr	1	TGC-Cys	1
TTA-Leu	14	TCA-Ser	12	TAA-Ter	1	TGA-Ter	0
TTG-Leu	7	TCG-Ser	1	TAG-Ter	0	TGG-Trp	10
CTT-Leu	8	CCT-Pro	10	CAT-His	8	CGT-Arg	7
CTC-Leu	0	CCC-Pro	0	CAC-His	2	CGC-Arg	2
CTA-Leu	0	CCA-Pro	4	CAA-Gln	7	CGA-Arg	1
CTG-Leu	0	CCG-Pro	2	CAG-Gln	0	CGG-Arg	0
ATT-Ile	21	ACT-Thr	9	AAT-Asn	19	AGT-Ser	3
ATC-Ile	5	ACC-Thr	0	AAC-Asn	2	AGC-Ser	0
ATA-Ile	2	ACA-Thr	4	AAA-Lys	2	AGA-Arg	3
ATG-Met	12	ACG-Thr	1	AAG-Lys	3	AGG-Arg	0
GTT-Val	18	GCT-Ala	21	GAT-Asp	7	GGT-Gly	24
GTC-Val	0	GCC-Ala	0	GAC-Asp	0	GGC-Gly	0
GTA-Val	7	GCA-Ala	10	GAA-Glu	12	GGA-Gly	6
GTG-Val	1	GCG-Ala	1	GAG-Glu	2	GGG-Gly	0

Fig. 5. Codon Usage in *Euglena* psbA.

Fig. 6. Sequence of the *Euglena* trnL-psbA intercistronic spacer region. The first in frame ATG codon of psbA exon 1 is numbered +1. The stem-loop structure centered at -34 to -35 and the ribosome binding site centered at -24 are proposed for the non-translated leader of psbA mRNA (see text). The calculated free energy for folding an RNA into the secondary structure as shown is approximately -16 kcal/mole (31). A tRNA^{Leu} gene (trnL-UAA) of the same polarity as psbA terminates at -87. The bracketed sequence centered at -55 resembles a conserved sequence from *Euglena* chloroplast tRNA gene 5'-leader regions (30).

second ATG at codon 37, as has also been suggested for higher plants (19), it might be possible to specifically label the nascent amino terminus with radioactive lysine during *in vivo* or *in vitro* translation. Second, 15 of 61 codons of the universal genetic code are not used. Nine of these absent codons terminate in C. No codons of the type GNC are used. There is a strong bias toward codons ending in T. T-ending codons are used for 57% of the amino acids, whereas C-ending codons occur only 7% of the time. Among codon families with purine redundancy, there is 4 to 1 preference for A-ending over G-ending codons. Such bias in codon usage is not found in the higher plant chloroplast psbA genes.

psbA 5'-Flanking Sequence

The 86 bp spacer sequence between the 3'-end of the tRNA^{Leu} gene and the first ATG codon in psbA has a 20% GC content. The predominant *in vivo* psbA mRNA species has been characterized by primer extension sequence analysis to begin at position -42 ± 2 nt with respect to the first, in frame ATG codon (U. Johanninger and R. B. Hallick, unpublished observation). Two features of this non-translated leader sequence are illustrated in Figure 6. First, the sequence 5'-GGGAG is present, centered at position -24 with respect to the ATG. This sequence is exactly complementary to the 3'-end of *Euglena* chloroplast 16S rRNA (28). Oligonucleotides containing the sequence 5'-GGGAG have been shown to specifically bind to *Euglena* chloroplast 30S ribosome subunits. We would therefore predict that this sequence in the psbA leader

region functions as the ribosome binding site. If so, it is located farther from the start codon than most ribosome binding sites of *E. coli* mRNAs. Second, the psbA leader region contains a pronounced dyad symmetry. The 3'-end of the symmetric sequence includes the putative ribosome binding site (Figure 6). Such a structure in an mRNA leader region could possibly have a role in translational regulation.

The psbA promoter sequence has not yet been identified. We are currently attempting to determine if the 5'-RNA start -42 $\pm$ 2 nt is the site of transcription initiation, or, alternatively is an mRNA processing site. The promoter could be in the -45 to -86 region of the trnL-psbA spacer, perhaps involving the sequence 5'-GTTATAATT centered at -55 (Figure 6) which resembles a conserved sequence motif in *Euglena* chloroplast tRNA gene 5'-leader regions (30).

ACKNOWLEDGMENTS

This work was supported by the U.S.P.H.S., NIH Grant GM21351. We wish to thank our colleague Jon Narita for many helpful discussions, and Udo Johannningmeier for the communication of data prior to publication.

*To whom correspondence should be sent

REFERENCES

1. Eaglesham, A. R. J. and Ellis, R. J. (1974) Biochim. Biophys. Acta 335, 396-407.
2. Chua, N.-H. and Gilham, N. W. (1977) J. Cell Biol. 74, 441-452.
3. Hoffman-Falk, H., Mattoo, A. K., Marder, J. B., Edelman, M., and Ellis, R. J. (1982)
4. Grebanier, A. E., Coen, D. M., Rich, A., and Bogorad, L. (1978) J. Cell Biol. 78, 734-746.
5. Pfister, K., Steinback, K. E., Gardner, G., and Arntzen, C. J. (1981) Proc. Natl. Acad. Sci. USA 78, 981-985.
6. Mattoo, A. K., Pick, U., Hoffman-Falk, H., and Edelman, M. (1981) Proc. Natl. Acad. Sci. USA 78, 1572-1576.
7. Edelman, M. and Reisfeld, A. (1980) in Genome Organization in Plants, Leaver, C. J., ed. pp 353-362, Plenum, New York.
8. Marder, J. B., Goloubinoff, P., and Edelman, M. (1984) J. Biol. Chem. 259, 3900-3908.
9. Edelman, M., Marder, J. B., and Mattoo, A. K. (1983) in Structure and Function of Plant Genomes, Ciferri, O. and Dure, L. eds., pp. 1887-192, Plenum, New York.

10. Bedbrook, J. R., Link, G., Coen, D. M., Bogorad, L., and Rich, A. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 3060-3064.
11. Zurawski, G., Bohnert, H. J., Whitfield, P. R., and Bottomley, W. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 7699-7703.
12. Hallick, R. B. and Bottomley, W. (1983) *Plant Mol. Biol. Reporter* **1**, 179-183.
13. Link, G. (1982) *Planta* **154**, 81-86.
14. Driesel, A. J., Spiers, J., and Bohnert, H. J. (1980) *Biochim. Biophys. Acta* **610**, 297-310.
15. Palmer, J. D. (1982) *Nucleic Acids Res.* **10**, 1593-1605.
16. Rochaix, J.-D. (1981) *Experientia* **37**, 323-332.
17. Hollingsworth, M. J., Johanningsmeier, U., and Hallick, R. B. (1984) *Nucleic Acids Res.* **12**, 2001-2007.
18. Curtis, S. and Haselkorn, R. (1984) *Plant Mol. Biol.*, in press.
19. Hirschberg, J. and McIntosh, L. (1983) *Science* **222**, 1346-1349.
20. Spielmann, A. and Stutz, E. (1983) *Nucleic Acids Res.* **11**, 7157-7167.
21. Link, G. and Langridge, U. (1984) *Nucleic Acids Res.* **12**, 945-958.
22. Guerry, P., LeBlanc, D. J. and Falkow, S. (1973) *J. Bacteriol.* **116**, 1064-1066.
23. Cohen, S. N. and Miller, C. A. (1970) *J. Mol. Biol.* **50**, 671-687.
24. Maxam, A. M. and Gilbert, W. (1980) *Methods Enzymol.* **65**, 499-560.
25. Sanger, F., Nicklen, S. and Coulson, R. A. (1977) *Proc. Natl. Acad. Sci. USA* **77**, 201-204.
26. Karabin, G. D., Narita, J. O., Dodd, J. R. and Hallick, R. B. (1983) *J. Biol. Chem.* **258**, 14790-14796.
27. Koller, B., Gingrich, J. C., Stiegler, G. L., Farley, M. A., Delius, H. and Hallick, R. B. (1984) *Cell* **36**, 545-553.
28. Orozco, Jr., E. M., Rushlow, K. E., Dodd, J. R. and Hallick, R. B. (1980) *J. Biol. Chem.* **255**, 10997-11003.
29. Steege, D. A., Graves, M. C. and Spremulli, L. L. (1982) *J. Biol. Chem.* **257**, 10430-10439.
30. Gruissem, W., Narita, J. O., Greenberg, B. M., Prescott, D. M. and Hallick, R. B. (1983) *J. Cellular Biochem.* **22**, 31-46.
31. Tinoco, Jr., I., Borer, P. N., Dengler, B., Levine, M. D., Uhlenbeck, O. C., Crothers, D. M. and Gralla, J. (1973) *Nature New Biol.* **246**, 40-41.