Nucleic Acids Research, 1993, Vol. 21, No. 15 3537–3544

Complete sequence of *Euglena gracilis* chloroplast DNA

Richard B.Hallick*, Ling Hong, Robert G.Drager[1], Mitchell R.Favreau, Amparo Monfort[2,+], Bernard Orsat[2,§], Albert Spielmann[2] and Erhard Stutz[2]

Department of Biochemistry and [1]Department of Molecular and Cellular Biology, University of Arizona, Tucson, AZ 85721, USA and [2]Laboratoire de Biochimie Végétale, Université de Neuchâtel, Chantemerle 18, CH-2000 Neuchâtel, Switzerland

Received March 17, 1993; Revised and Accepted June 15, 1993 EMBL accession no. X70810

ABSTRACT

We report the complete DNA sequence of the *Euglena gracilis*, Pringsheim strain Z chloroplast genome. This circular DNA is 143,170 bp, counting only one copy of a 54 bp tandem repeat sequence that is present in variable copy number within a single culture. The overall organization of the genome involves a tandem array of three complete and one partial ribosomal RNA operons, and a large single copy region. There are genes for the 16S, 5S, and 23S rRNAs of the 70S chloroplast ribosomes, 27 different tRNA species, 21 ribosomal proteins plus the gene for elongation factor EF-Tu, three RNA polymerase subunits, and 27 known photosynthesis-related polypeptides. Several putative genes of unknown function have also been identified, including five within large introns, and five with amino acid sequence similarity to genes in other organisms. This genome contains at least 149 introns. There are 72 individual group II introns, 46 individual group III introns, 10 group II introns and 18 group III introns that are components of twintrons (introns-within-introns), and three additional introns suspected to be twintrons composed of multiple group II and/or group III introns, but not yet characterized. At least 54,804 bp, or 38.3% of the total DNA content is represented by introns.

INTRODUCTION

Euglena gracilis is a unicellular facultative photosynthetic organism which is phylogenetically related to flagellate protists (1, 2). Although *Euglena gracilis* chloroplasts share many common structural and functional features with chloroplasts of chlorophytes and land plants, notably the chlorophyll content of the photosynthetic apparatus, the phylogenetic position of euglenoid plastids remains uncertain (3, 4). Euglena chloroplast DNA (cpDNA) was among the first well characterized organellar genomes (5), largely due to its rather low GC content (buoyant density) which allowed clear discrimination between nuclear and plastid DNA. Highly purified chloroplast DNA preparations amenable to molecular analysis could be obtained. Euglena cpDNA was the first known example of a circular chloroplast genome (6). In subsequent studies it became evident that the overall organization of Euglena cpDNA is quite different from cpDNA of green algae and land plants (7), but it is rather similar with respect to number and kind of genes. Unique features of Euglena cpDNA include a region containing a variable number of short, tandem repeats which may qualify as an origin of DNA replication (8, 9, 10), some extremely large and complex introns (twintrons) found in some of the genes involved in PSII synthesis (11, 12), and a unique class of very small introns designated group III which appear to be streamlined group II introns (13, 14). The sequence of the Euglena chloroplast genome discussed in this report is the first complete sequence from a unicellular organism, and the fourth example (following tobacco, liverwort, and rice) of a complete chloroplast sequence (15, 16, 17). A complete sequence of the plastid DNA of the non-photosynthetic epiphyte *Epifagus virginiana* has also been reported (15).

MATERIALS AND METHODS

Euglena gracilis (Pringsheim, strain Z) was grown and harvested following standard procedures. Cell growth, plastid isolation, and protocols for chloroplast DNA isolation, restriction, cloning and sequencing have been described (7, 16).

The DNA sequence for a number of Euglena chloroplast genes had previously been reported. In order to complete the entire sequence, all known regions were compiled and annotated, several corrections to earlier data were made and annotated, and all unknown regions were identified, cloned with appropriate overlaps, and sequenced on both strands. This information is provided in EMBL Accession No. X70810. The last 54 bp of the sequence X70810 represent a single copy of a sequence element that is repeated in variable copy number in different DNAs isolated from the same culture of cells. It can be formally described as a 'variable number of tandem repeat' or 'VNTR'-sequence. Individual Euglena cpDNAs will have more than 143,170 bp, depending on the number of 54 bp repeated

* To whom correspondence should be addressed

Present addresses: +Centro di Investigacion y Desarrollo CSIC, Dept. Genetica Molecular, Jordi Girona 18-26, 08034 Barcelona, Spain and §Department of Chemistry, MIT, Cambridge, MA 02139, USA

segments. We have previously shown that the 16S rRNA, *trnA*, *trnI*, and 23S rRNA genes of *rrnA, rrnB, and rrnC* cannot be distinguished by analysis with any restriction enzymes (7). Thus it was not possible to determine the DNA sequence of each rRNA operon individually. We have made the assumption that these regions are identical in preparing the DNA sequence compilation.

Details of sequencing procedures for new genes will be provided in subsequent publications. Sequence data were compiled and evaluated using the software from Genetics Computer Group, 575 Science Drive, Madison, Wisconsin, USA 53711 (17). Gene identification was based on screening of the GenBank Release 75.0, EMBL Release 30.0, PIR-Protein Release 33.0, PIR-Nucleic Release 36.0, and SwissProt Release 22.0 databases with the FASTA and BLITZ algorithms from EMBL, Heidelberg, and the BLAST algorithm available through the BLAST network service at the National Center for Biotechnology Information (NCBI), USA. Chloroplast gene nomenclature follows previous recommendations (18, 19). Genes encoding open reading frames conserved in chloroplasts of other species are designated with the prefix '*ycf*' here, and in the SwissProt database (R.B. Hallick, manuscript in preparation). These designations are temporary chloroplast gene names pending identification of the function of the gene product. Hypothetical genes of unknown function unique to Euglena chloroplasts are designated 'orfs' followed by the length of the reading frame in codons.

RESULTS AND DISCUSSION

Chloroplast genome organization

A physical map of the circular chloroplast DNA (143,170 bp) is shown in Figure 1. The sequence is numbered from the first nucleotide after the VNTR-region (position 1) clockwise to the last nucleotide before the VNTR-region (position 143,116), followed by one copy of the 54 nt VNTR sequence (positions 143,117–143,170). Data and annotations are reported in EMBL accession no. X70810. The single origin of DNA replication maps in close proximity to the VNTR region (9, 10). Overall base composition is 26.1% G+C and 73.9% A+T.

There are three copies of a tandemly repeated 5918 nt ribosomal RNA operon. The exactly duplicated DNA is from positions 115,663 to 132,813 (2.9 repeats). When regions with small insertions and deletions are included (from 115,606 to 133,549), and a fourth, partial operon encoding a complete 16S

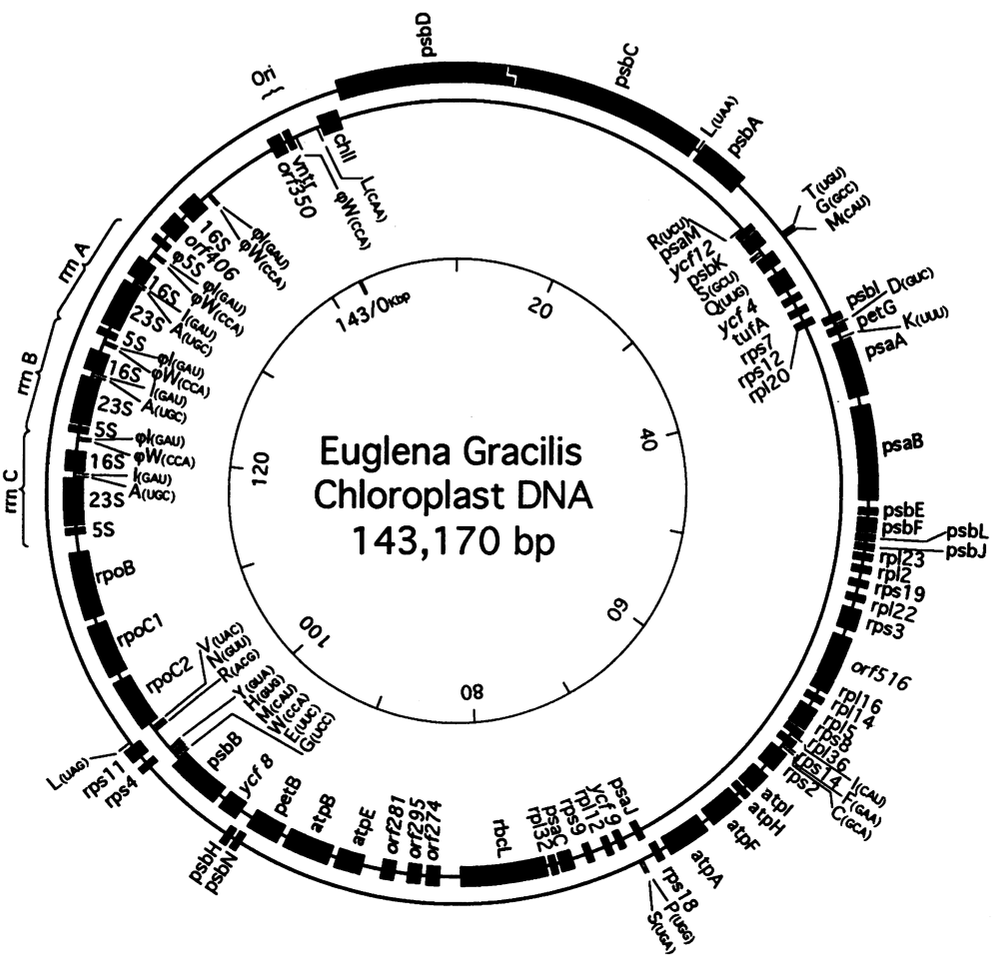

Figure 1. Circular map of *Euglena gracilis* chloroplast DNA. Genes are represented by filled boxes which are proportional to gene length, including exons and introns. For intron content of individual genes, see Table 3. Genes on the outer circle are transcribed clockwise. Genes on the inner circle are transcribed counter-clockwise. Chloroplast gene nomenclature has been previously described (18, 19), (see Table 1). Transfer RNA genes are identified by the single-letter code for the cognate amino acid, with the anticodon in parentheses.

Table 1. *Euglena gracilis* chloroplast genes

a) Ribosomal RNAs and Proteins	
23S rRNA	23S ribosomal RNA
16S rRNA	16S ribosomal RNA
5S rRNA	5S ribosomal RNA
rpl2	ribosomal protein L2
rpl5	ribosomal protein L5
rpl12	ribosomal protein L12
rpl14	ribosomal protein L14
rpl16	ribosomal protein L16
rpl20	ribosomal protein L20
rpl22	ribosomal protein L22
rpl23	ribosomal protein L23
rpl32	ribosomal protein L32
rpl36	ribosomal protein L36
rps2	ribosomal protein S2
rps3	ribosomal protein S3
rps4	ribosomal protein S4
rps7	ribosomal protein S7
rps8	ribosomal protein S8
rps9	ribosomal protein S9
rps11	ribosomal protein S11
rps12	ribosomal protein S12
rps14	ribosomal protein S14
rps18	ribosomal protein S18
rps19	ribosomal protein S19
b) Transfer RNAs	
trnA	ALA-tRNA-UGC (3-copies)
trnC	CYS-tRNA-GCA
trnD	ASP-tRNA-GUC
trnE	GLU-tRNA-UUC
trnF	PHE-tRNA-GAA
trnG	GLY-tRNA-GCC
trnG	GLY-tRNA-UCC
trnH	HIS-tRNA-GUG
trnI	ILE-tRNA-CAU
trnI	ILE-tRNA-GAU (3 copies)
trnK	LYS-tRNA-UUU
trnL	LEU-tRNA-CAA
trnL	LEU-tRNA-UAA
trnL	LEU-tRNA-UAG
trnM	MET-tRNA-CAU (elongator)
trnM	MET-tRNA-CAU (initiator)
trnN	ASN-tRNA-GUU
trnP	PRO-tRNA-UGG
trnQ	GLN-tRNA-UUG
trnR	ARG-tRNA-UCU
trnR	ARG-tRNA-ACG
trnS	SER-tRNA-GCU
trnS	SER-tRNA-UGA
trnT	THR-tRNA-UGU
trnV	VAL-tRNA-UAC
trnW	TRP-tRNA-CCA
trnY	TYR-tRNA-GUA
c) Transcription/Translation	
rpoB	RNA polymerase β subunit
rpoC1	RNA polymerase β' subunit
rpoC2	RNA polymerase β'' subunit
tufA	translation elongation factor EF-Tu
d) Photosynthetic Proteins	
psaA	photosystem I P700 apoprotein A1
psaB	photosystem I P700 apoprotein A2
psaC	photosystem I subunit VII (FA/FB containing)
psaJ	photosystem I 5 kDa protein
psaM	photosystem I M-polypeptide
psbA	photosystem II core 32 kDa protein
psbB	photosystem II CP47 chlorophyll apoprotein
psbC	photosystem II CP43 chlorophyll apoprotein
psbD	photosystem II core 34 kDa protein
psbE	photosystem II cytochrome b559 α subunit
psbF	photosystem II cytochrome b559 β subunit
psbH	photosystem II 10 kDa protein
psbI	photosystem II I polypeptide
psbJ	photosystem II J protein
psbK	photosystem II 3.9 kDa protein
psbL	photosystem II L protein
psbN	photosystem II N protein (tentative identification)
petB	cytochrome b6
petG	cytochrome b6/f complex subunit V
rbcL	RuBisC/O large subunit
atpA	ATPase α subunit
atpB	ATPase β subunit
atpE	ATPase ϵ subunit
atpF	ATPase subunit I
atpH	ATPase subunit III
atpI	ATPase subunit IV
chlI	chlorophyll biosynthesis (=ccsA)
e) ORFs identified by similarity to other chloroplast orfs	
ycf8	(orf31) hydrophobic, transcribed with psbB
ycf12	(orf33) similar to M. polymorpha ycf12
ycf9	(orf65) hydrophobic; occurs in land plants
ycf4	(orf206) polar; transcribed with tufA
ycf13	(ycf13) in psbC intron 4; occurs in Astasia
f) Other ORFS or unknown function	
orf177	encoded in psbC intron 2
orf241	encoded in psbC intron 2
orf274	in atpE-rbcL intercistronic DNA
orf281A	encoded in psbD intron 8
orf281B	in atpE-rbcL intercistronic DNA
orf295	in atpE-rbcL intercistronic DNA
orf350	encoded near origin of replication
orf406	within rDNA repeat
orf506	encoded in psbD intron 8; C2H2-type zinc finger
orf516	highly basic; in rpl23 operon

rRNA gene (from 135,492 to 137,229) is also added, there are 19.6 kb of repeated rDNA sequence, accounting for 13.7% of the genome. This region is GC-rich (41.0% G+C) compared to the entire DNA.

The remainder of the chloroplast DNA, other than the VNTR region, is single copy sequence, densely packed with genes for polypeptides and tRNAs. The overall gene arrangement is shown in Figure 1. The relative sizes of the genes on the map include both exons and introns. Although none of the tRNA genes contain introns, all genes for known polypeptides except eight of 21 ribosomal protein genes and six of 27 photosynthesis related genes are interrupted by one or more intervening sequences.

The most notable feature of genome organization may be the arrangement of coding and non-coding DNA strands with respect to the origin of replication (Figure 1). Euglena chloroplast DNA is believed to be replicated bidirectionally from a single replication origin to a terminator (10) on the opposite side of the circular DNA. Most gene clusters are transcribed away from the origin bidirectionally toward the presumptive terminator. Exceptions include the *rps4-11* operon, *psbN-psbH*, several tRNAs and a cluster of genes beginning with *rpl20* (Figure 1). The strong bias of gene polarity away from the origin of replication could be an indication that replication and transcription are closely linked in Euglena chloroplasts.

Genes for components of the chloroplast translation and transcription apparatus

A summary of the 55 known genes for components of the chloroplast 70S ribosomes, tRNAs, and translation factors is given in Table 1. Included are the 16S, 23S, and 5S rRNAs, 27 different tRNA species, 11 ribosomal proteins of the 30S subunit, 10 ribosomal proteins of the 50S subunit and the gene for elongation factor EF-Tu. All these genes are constitutively expressed. Their gene products are present in light- or dark-grown Euglena cells.

Table 2. Summary of codon usage frequency in identified Euglena chloroplast protein genes, and corresponding tRNA anticodons encoded in chloroplast DNA

Phe	UUU	627		Ser	UCU	320		Tyr	UAU	335		Cys	UGU	98	
Phe	UUC	92	*trnF-GAA*	Ser	UCC	43	*trnS-UGA*	Tyr	UAC	56	*trnY-GUA*	Cys	UGC	35	*trnC-GCA*
Leu	UUA	677	*trnL-UAA*	Ser	UCA	189		End	UAA	40		End	UGA	2	
Leu	UUG	214	*trnL-CAA*	Ser	UCG	52		End	UAG	6		Trp	UGG	189	*trnW-CCA*
Leu	CUU	231		Pro	CCU	295		His	CAU	234		Arg	CGU	206	
Leu	CUC	3	*trnL-UAG*	Pro	CCC	33	*trnP-UGG*	His	CAC	28	*trnH-GUG*	Arg	CGC	47	*trnR-ACG*
Leu	CUA	75		Pro	CCA	142		Gln	CAA	311	*trnQ-UUG*	Arg	CGA	89	
Leu	CUG	12		Pro	CCG	23		Gln	CAG	47		Arg	CGG	9	
Ile	AUU	620		Thr	ACU	294		Asn	AAU	497		Ser	AGU	173	
Ile	AUC	58	*trnI-GAU*	Thr	ACC	28	*trnT-UGU*	Asn	AAC	105	*trnN-GUU*	Ser	AGC	28	*trnS-GCU*
Ile	AUA	372	*trnI-CAU*	Thr	ACA	277		Lys	AAA	771	*trnK-UUU*	Arg	AGA	233	*trnR-UCU*
Met	AUG	236	*trnM-CAU*	Thr	ACG	65		Lys	AAG	139		Arg	AGG	58	
f-Met	AUG	48	*trnM-CAU*												
Val	GUU	475		Ala	GCU	383		Asp	GAU	379		Gly	GGU	480	
Val	GUC	31	*trnV-UAC*	Ala	GCC	39	*trnA-UGC*	Asp	GAC	70	*trnD-GUC*	Gly	GGC	65	*trnG-GCC*
Val	GUA	233		Ala	GCA	233		Glu	GAA	470	*trnE-UUC*	Gly	GGA	319	*trnG-UCC*
Val	GUG	43		Ala	GCG	61		Glu	GAG	107		Gly	GGG	60	

Three genes encode subunits of chloroplast DNA-dependent RNA polymerase. The *rpoB-rpoC1-rpoC2* genes are organized as a tricistronic operon. Notably absent from the gene list is *rpoA*, an RNA polymerase subunit gene which is ubiquitous in land plant chloroplast DNA, but absent in *E. virginiana*. This gene may be located in the nucleus in Euglena. Since *rpoA* is not well conserved in amino acid sequence in different species, another possibility is that *rpoA* might be present but not detectable without cDNA analysis. The high density of introns in Euglena chloroplast DNA can mask the location of protein coding regions, such that cDNA sequence analysis is often necessary to identify chloroplast genes. All of the exons reported for known RNA polymerase subunit genes and ribosomal proteins (except *rps9*) have been confirmed by cDNA analysis. Many of these exons are very small. Of 168 exons for known, intron-containing genes, 54 encode less than 20 amino acids. Database searches with these small exons as query sequences often yield false negative results. Thus it is likely that additional genes and introns will be identified as cDNA analysis is extended to as yet uncharacterized regions of the cpDNA.

The multiple copies of the 5S ribosomal RNA genes are not all identical. The 5S rRNA gene of the third complete operon (*rrnC*) differs in five of 116 positions from the corresponding genes in the *rrnA/B* operons. There is also a pseudo-5S rRNA gene, identical in 109 of 116 positions to the *rrnA/B* 5S rRNA gene. The fourth 16S rRNA gene in the incomplete rRNA operon differs in 21 of 1491 positions from the remaining three genes. By contrast, multiple copies of rRNAs of land plants are all identical. Although all genes are believed to be expressed in Euglena, it is not known if different alleles have different functions.

Genes for transfer RNAs and pseudo-transfer RNAs

All 61 code words of the universal genetic code are found in known chloroplast protein genes. A list of the 27 tRNA genes and the corresponding anticodons is given in Table 2. Transfer RNA loci are shown in Figure 1. The *trnI-trnA* genes are co-transcribed with the rRNA operons, and are the only tRNA genes present in multiple gene copies. Are 27 tRNAs sufficient for chloroplast protein synthesis? If expanded codon-anticodon pairing rules are assumed, allowing for U:N (or modified A:N) pairing between the first base of the anticodon and the third position of the codon for six codon families, these 27 tRNAs would represent a complete set for protein synthesis within the organelle. A codon usage table for the identified Euglena chloroplast protein genes, and the corresponding tRNA anticodon for translation of each codon is shown in Table 2. The proposed two out of three pairings would occur for seven out of eight codon families with four base redundancy at the third codon position. The tRNAs with potential U:N pairing are *trnA*-UGC, *trnL*-UAG, *trnP*-UGG, *trnR*-ACG, *trnS*-UGA, *trnT*-UGC, and *trnV*-UAC. Isoaccepting tRNAs are present only for leu, ile, arg, ser and gly codons.

The codon usage frequency shown in Table 2 reflects the high A + U base content of this genome. There is a 4.8:1 ratio of codons ending in either A or U compared to G or C. In codons ending with purines, there is a 3.6:1 bias of A over G. In codons ending in pyrimidines, there is a 7.4:1 bias of U over C. Although all 61 codons of the universal genetic code are used, some are very rare, including Leu-CUC, Leu-CUG, and Arg-CGG, used three, twelve, and nine times, respectively.

The locations of nine pseudo-tRNA genes are also shown in Figure 1. Of particular interest are the five copies of the previously described pseudo-*trnW*-CCA genes (7), which immediately precede the transcription start site of all four 16S rRNA genes. This pseudogene is also present at or near the origin of replication, adjacent to the VNTR sequences. There are also four copies of a pseudo-*trnI*-GAU gene, one preceeding each 16S rRNA gene. The pseudo-*trnW*-CCA genes are very similar to the single, intact *trnW*-CCA gene. The pseudo-*trnI*-GAU genes are derived from the *trnI*-GAU of the 16S-23S rRNA intercistronic region.

Genes for chloroplast ribosomal proteins

Euglena chloroplast DNA encodes at least 21 chloroplast ribosomal protein genes (Table 1), including 11 for the 30S small subunit and 10 for the 50S large subunit. Ten of these genes are present in a single ribosomal protein operon (20). Ribosomal protein coding capacity is similar to that of land plant chloroplast genomes (18 of 21 genes). Euglena has the small subunit genes *rps2, rps3, rps4, rps7, rps8, rps11, rps12, rps14, rps18*, and *rps19* that are found in nearly all known chloroplast genomes.

Present in land plants but absent in Euglena and *E. virginiana* are *rps15* and *rps16*. Euglena has the large subunit genes *rpl2, rpl14, rpl16, rpl20, rpl22, rpl23, rpl32, and rpl36. rpl33* which is found in land plant chloroplasts has not been detected. Genes present in Euglena but absent in land plants include *rpl5*, also found in chloroplast DNAs of *Astasia longa* (21), the red alga *Porphyra purpurea* (22), and cyanelle DNA of *Cyanophora paradoxa* (23), and *rpl12*. Four exons of an *rps9* locus have been identified in the *psaC-rpl12* intergenic region, but the exact splice boundaries are not yet known. *rps9* is also present in chloroplasts of Cryptomonas (24) and cyanelle DNA of *C. paradoxa* (25). Two other differences with respect to land plant chloroplast gene content are the presence of the *tufA* gene for elongation factor EF-Tu, and the absence of the *infA* gene for initiation factor IF-1.

Genes involved in photosynthesis

The Euglena chloroplast genome encodes at least 27 genes for components of the thylakoid membranes, the chloroplast ATP synthase complex, or the CO_2-fixing enzyme RUBISCO. Photosynthesis-related genes are listed in Table 1. There are 5 known genes for photosystem I polypeptides (designated *psaA−C, J, M*), 10 for photosystem II (designated *psbA−F, H−L*), and 2 for the cytochrome b_6/f complex (*petB, petG*). The *psaM* gene which was first described for cyanobacteria is also present in the liverwort, *Marchantia polymorpha*, chloroplast genome. The six Euglena ATP synthase subunit genes are organized in two operons similar to those of land plants. *atpF-atpH-atpF-atpA* are linked in the *rps2* operon, and *atpB-atpE* are co-transcribed. Notably absent from Euglena are any genes for subunits of a NADH dehydrogenase complex, present in land plant chloroplast genomes. Also present in land plants, but not detected in Euglena are the genes *psaI, psbM*, and *petD*. The Euglena *psbN* gene is located between *psbH* and *petB*, but lacks an AUG or GUG initiator codon. Euglena would not be expected to have a *petA* gene since cytochrome f is absent in this protist. Euglena contains a gene (*chlI*), (26) absent in the chloroplast genomes of land plants, but present in the red alga *P. purpurea* (22), that is most likely necessary for chlorophyll biosynthesis.

Other genes for proteins of known and unknown function

There are a number of protein genes of known function generally encoded in chloroplast DNA of land plants that are not detected in Euglena. These include *infA, clpP, frxB, ndhA−K, petA, petD, psaI, psbM, rpl32, rpoA, rps15* and *rps16*. Euglena has a reduced content of chloroplast genes for photosynthetic and non-photosythetic activities relative to the land plants. By contrast, various non-green alga such as Cryptomonas and *Porphyra purpurea* and the cyanelle genome of *Cyanophora paradoxa* have increased organelle DNA coding capacity when compared to land plants, and may encode genes for fatty acid biosynthesis, amino acid biosynthesis, the light harvesting proteins, chaperonins, and additional components of the transcriptional and translational apparatus (22, 25).

Several open reading frames (ORFs) encoding proteins of unknown function are conserved between chloroplasts of plants and algae, or between cyanobacteria and chloroplasts. Chloroplast genes that code for proteins of unknown function, and are conserved in more than one organism are now designated with the gene prefix '*ycf*' (Recommendation of the International Society for Plant Molecular Biology, Commission on Plant Gene Nomenclature). Of the genes *ycf1−ycf11*, Euglena has only *ycf4, ycf8*, and *ycf9* (Table 1). Representative examples of these genes from the tobacco chloroplast genome (identified by the SwissProt Accession No.) are *ycf4* (orf184, P12207), *ycf8* (orf34, P12184), and *ycf9* (orf62, P09974). The Euglena *ycf4* locus encodes a basic polypeptide of 206 amino acids rich in polar residues located distal to and co-transcribed with *tufA*. The land plant homologue has 184−185 codons. The Euglena *ycf8* locus encodes a short, hydrophobic protein of 31 amino acids that is co-transcribed with *psbB*. The Euglena *ycf9* gene encodes a polypeptide of 65 amino acids rich in hydrophobic residues.

Also listed in Table 1 are several additional hypothetical Euglena chloroplast protein genes identified as open reading frames that are found only on the Euglena chloroplast genome. Only orfs longer than 100 codons are included in Table 1 and Figure 1. Orf406 has previously been described (27). Orf516 is a very basic polypeptide encoded in the *rpl23* ribosomal protein operon, and interrupted by 4 introns. Antibodies directed against two different epitopes in this polypeptide cross-react with a soluble Euglena chloroplast protein of the expected size (K.Jenkins and R.B.Hallick, manuscript in preparation). orf281a and orf506 are encoded within *psbD* intron 8. orf506 has a C2H2-type zinc finger domain. orf177 and orf241 are located within *psbC* intron 2. orf274, orf281b, and orf295 are all located in the 5.8 kb *rbcL-atpE* intercistronic region that is not yet characterized by cDNA analysis. This list of potential protein genes is not comprehensive. As previously noted, the location of protein genes can be masked due to the high density of introns, the relatively small size of many exons, and the low amino acid sequence identity between some chloroplast genes from different organisms.

Comparison to *Astasia longa* plastid DNA

Astasia longa is a colorless, non-photosynthetic protist that is phylogenetically related to *Euglena gracilis* (28, 29). Astasia has a plastid DNA of size 73 kb. More than 25 kbp of Astasia plastid DNA sequence has been determined. No genes for photosynthetic function have been found except *rbcL*. Identified genes include 7 tRNAs, 3 rRNAs, 6 ribosomal proteins, *rpoB*, and *tufA*, all present in Euglena. Astasia has a gene cluster with the gene order *rpl5-rps8-rpl36-trnI-trnF-trnC-rps2* (EMBL Ac. X16004). Not only does this same gene cluster occur in Euglena, but three group II and five group III introns occur in the same positions in the same genes in both Euglena and Astasia. Another gene combination found in both organisms is *rbcL-rpl32*. Astasia *rbcL* has seven of the nine group II introns in the same positions as Euglena *rbcL* (28). Astasia *rpoB* also has at least seven group III introns, but their positions differ from Euglena *rpoB*.

Euglena has a locus designated *ycf13* for a protein of 458 amino acids, absent in land plants, but also found in plastid DNA of Astasia longa (30). The Euglena gene is encoded within a group III twintron internal to the *psbC gene* (D. W. Copertino and R. B. Hallick, in preparation), but lacks reverse transcriptase motifs often characteristic of intron-encoded polypeptides. The Astasia *ycf13* homologue for a 456 amino acid polypeptide is not intron-encoded (30). Assuming deletions of the *psbC* and *psbA* genes, the Astasia *ycf13* gene is on the same strand and in relatively the same location on the genome as its Euglena homologue. Since the plastid genes of Astasia can contain group III introns, and *ycf13* is encoded within a group III intron in Euglena, the *ycf13* gene product may be required for group III intron excision in both Euglena and Astasia.

Surprisingly, Astasia has two large orfs, designated orf211 and orf167 (30) that are absent in Euglena. It has been proposed that

Table 3. Introns of *Euglena gracilis* chloroplast DNA by location, category, and size in nucleotides (nt.)

No.	Gene	Intron	Type	Nt.	No.	Gene	Intron	Type	Nt.
1	atpA	1	II	603	76	psbK	2	III-Ex	93
2	atpA	2	II	551	77	psbK	2	III-In	111
3	atpB	1	II	374	78	rbcL	1	II	404
4	atpB	2	II	431	79	rbcL	2	II	514
5	atpB	3	II	326	80	rbcL	3	II	513
6	atpB	4	II	480	81	rbcL	4	II	568
7	atpE	1	II-Ex	355	82	rbcL	5	II	413
8	atpE	1	II-In	402	83	rbcL	6	II	479
9	atpE	2	II	661	84	rbcL	7	II	382
10	atpF	1	II	613	85	rbcL	8	II	420
11	atpF	2	II	361	86	rbcL	9	II	441
12	atpF	3	II	632	87	rpl12	1	III	104
13	atpI	1	III	108	88	rpl14	1	III	108
14	atpI	2	III	108	89	rpl14−5	intcis.	III	112
15	atpI	3	III	102	90	rpl16	1	III	91
16	atpI	4	II	323	91	rpl16	2	II	356
17	atpI	5	III	112	92	rpl16	3	III-In	112
18	atpI	6	III	106	93	rpl16	3	III-Ex	96
19	ccsA	1	II	332	94	rpl22	1	II	347
20	ycf4	1	II	297	95	rpl23	1	III	106
21	ycf12	1	III	107	96	rpl23	2	III	99
22	ycf8	1	II-In	601	97	rpl23	3	III	103
23	ycf8	1	II-In	393	98	rpl23−2	intcis.	III	100
24	ycf8	1	II-Ex	358	99	rpoB	1	III	93
25	orf516	1	II	349	100	rpoB	2	III	95
26	orf516	2	III	97	101	rpoB	3	III	94
27	orf516	3	II	325	102	rpoB	4	III	99
28	orf516	4	II	438	103	rpoB	5	III	101
29	pet B	1	II-Ex	399	104	rpoB	6	III	110
30	pet B	1	II-In	404	105	rpoB	7	III	99
31	pet B	1	III-In	106	106	rpoB	8	II	309
32	petB	2	II	535	107	rpoC1	10	III	103
33	petG	1	II	372	108	rpoC1	11	III-Ex	102
34	psaA	1	II	490	109	rpoC1	11	III-In	96
35	psaA	2	II	542	110	rpoC1	1	III-Ex	114
36	psaA	3	II	361	111	rpoC1	1	III-In	96
37	psaB	1	II	441	112	rpoC1	2	III	107
38	psaB	2	II	525	113	rpoC1	3	III-Ex	111
39	psaB	3	II	508	114	rpoC1	3	III-In	102
40	psaB	4	II	590	115	rpoC1	4	III	100
41	psaB	5	II	579	116	rpoC1	5	III	119
42	psaB	6	II	570	117	rpoC1	6	II	349
43	psaC	1	II	320	118	rpoC1	7	III	97
44	psaC	2	II	391	119	rpoC1	8	III	110
45	psbA	1	II	433	120	rpoC1	9	III	102
46	psbA	2	II	447	121	rpoC2	1	II	580
47	psbA	3	II	434	122	rpoC2	2	II	514
48	psbA	4	II	616	123	rps11	1	III	107
49	psbB	1	II	501	124	rps11	2	III	100
50	psbB	2	III	104	125	rps14	1	III	106
51	psbB	3	II	572	126	rps18	1	III	101
52	psbB	4	II	567	127	rps18	2a	III-Ex	107
53	psbC	1	II	543	128	rps18	2b	III-In	110
54	psbC	10	II	423	129	rps18	2c	III-In	106
55	psbC	3	II	671	130	rps18	2d	III-In	112
56	psbC	4	III-Ex	101	131	rps19	1	III	100
57	psbC	4*	III-In	1504	132	rps19	2	III	97
58	psbC	5	II	590	133	rps2	1	III	101
59	psbC	6	II	448	134	rps2	2	III	112
60	psbC	7	II	668	135	rps2	3	III	99
61	psbC	8	II	621	136	rps2	4	II	390
62	psbC	9	II	305	137	rps3	1	III-Ex	99
63	psbD	10	II	543	138	rps3	1	II-In	310
64	psbD	2	II	364	139	rps3	2	III	102
65	psbD	3	II	605	140	rps4−11	intcis.	III	95
66	psbD	4	II	651	141	rps7-tufA	intcis.	III	96
67	psbD	5	II	498	142	rps8	1	II	327
68	psbD	6	II	606	143	rps8	2	III	95
69	psbD	7	II	580	144	rps8	3	II	277
70	psbD	9	II	373	145	tufA	1	III	103

71	psbE	1	II	350	146	tufA	2	III	110
72	psbE	2	II	326	147	psbD	1	n.d.	1098
73	psbF	1	II-Ex	424	148	psbD	8*	n.d.	3658
74	psbF	1	II-In	618	149	psbC	2*	n.d.	4143
75	psbK	1	III	105		Total			54804

Data were extracted from annotations of EMBL Accession X70810. II and III refer to group II and group III introns, respectively. II-ex, III-ex, II-in, and III-in refer to external and internal group II and III introns that are constituents of twintrons. 'nd' refers to suspected twintrons not yet characterized by cDNA analysis. 'intcis' is for intercistronic introns. Asterisk (*) indicates orf(s) within intron.

maintenance of plastid DNA in the non-photosynthetic parasite *E. virginiana* is due to the expression of an essential plastid gene or gene(s) required for survival of the organism (15). By contrast, Euglena and Astasia may lack essential, non-photosynthetic genes, since Euglena mutants containing little or no plastid DNA are known (7).

Introns

Unlike land plant chloroplast genomes, there are no introns in the Euglena chloroplast rRNA or tRNA genes. Nevertheless, Euglena chloroplast DNA has at least 149 introns, the most introns of any known organelle genome. As cDNA analysis of chloroplast mRNAs and partially spliced mRNAs is extended, additional introns will be added to this list, including three or more introns in *rps9*, and introns predicted for uncharacterized twintrons. A list of all introns by gene, size, and intron category is given in Table 3. The sum of all intron lengths is 54,804 nt, representing 38.3% of the genome. Since introns only occur outside of the repeated rDNA sequences, introns account for at least 44.4% of non-rDNA sequences. The contrast between the high intron content of non-rDNA and the absense of introns in the repeated rDNA is very striking in Euglena chloroplasts. There are no known group II or group III introns in rRNA genes from any organism. It is possible that group II introns are not found in rRNA genes because structural features required for splicing are not compatible with rRNA secondary structure.

Euglena chloroplast introns fall into two categories. Group II introns are similar to introns of fungal and plant mitochondria, and plant and algal chloroplasts. The most characteristic features are the conserved 5'-boundary sequence motif of 5'-GTGYG, and the structural domains 5 and 6 at the 3'-end of the introns (31). Group III introns appear to be abbreviated versions of group II introns. Group III introns have a size of approximately 100 nt, a consensus boundary sequence of 5'-NUNNG, and a group II intron-like domain 6 (14, 20). Group III introns also occur in *Astasia longa* plastid DNA (30).

There are 72 individual group II introns, and ten additional group II introns that are components of twintrons (introns-within-introns). Sixty seven of these 82 group II introns occur in photosynthesis related genes. The size range for these 67 introns is 305–671 nt, with an average size of 483 nt. The remaining 15 group II introns are in genes for the transcription and translation systems, with an average size of only 368 nt, and a size range of 277–588 nt. The Euglena group II introns are small by comparison to those found in other chloroplasts, and in plant and fungal mitochondria. The smaller group II introns have abbreviated domain 1 structures, and some of them lack parts of domains 3 and 4. All group II introns appear to have domains 5 and 6, and the core stem for domain 1 as defined by Michel et al (31). The ten known group II introns of Astasia range in size from 270–421 nt (28).

Euglena chloroplast DNA also contains 46 individual group III introns and 18 more group III introns that are components of twintrons. Group III introns are predominately located within genes for components of the transcription and translation systems. Only 13 of 64 occur in photosynthsis related genes. The size range of group III introns is 91 to 119, with an average size of 103 nt.

In addition to numerous group II and group III introns, Euglena cpDNA has many twintrons, which are introns-within-introns. Twelve twintrons have been characterized via cDNA cloning of partially spliced pre-mRNAs. Three additional twintrons are predicted to occur from their size and an analysis of potential intron secondary structure. Twintrons fall into different categories. Among the simple twintrons, where one intron is inserted into another, examples include a group II internal to another group II intron (11), a group II intron internal to a group III (14), and four cases of group III introns internal to group III introns (32). Other introns are more complex, including 2 or more introns inserted into a third (33), and open reading frames within the internal intron of a twintron. Some introns are very large, and are putative twintrons, but they have not yet been fully characterized (*psbD* introns 1 and 8, *psbC* introns 2) (12). The designations 'II-ex', 'II-in', 'III-ex' and 'III-in' are used in Table 3 to signify the individual external (ex) and internal (in) group II introns which are components of twintrons.

Origin and evolution of introns

The description of 149 introns is an important new data set for the ongoing debate on the evolutionary origin of introns. In the 'introns early' view (34, 35, 36) ancient genes are viewed as a mosaic of functional domains that are assembled from smaller bits of information. Introns are proposed to have facilitated the assembly of ancient genes from these individual domains. The recent report of the identification of a novel intron (37), predicted by Gilbert (34), in the triosephosphate isomerase gene from a mosquito can be viewed as evidence of the assembly of ancient genes by exon shuffling. An alternative hypothesis is that introns are mobile genetic elements that have been added to ancestral genes during the evolutionary descent from a common, intron-less ancestral gene (14, 38–42). All of the known Euglena chloroplast genes encode ancient proteins, such as those involved in RNA synthesis, protein synthesis, ATP synthesis, and photosynthesis. All of these genes arose before the evolutionary divergence between eubacteria and eukaryotes. Do the sites of insertion of the 149 or more introns in Euglena chloroplast genes provide an evolutionary road map for ancient gene rearrangements or are these introns of more recent origin? We believe that the Euglena chloroplast introns are descendants of

mobile genetic elements that have invaded this genome. The evidence in support of this conclusion is that the genome contains introns in unique locations not found in other chloroplast DNAs, in intercistronic spacers, and within other introns. The genome also lacks introns conserved in other chloroplasts.

Prospects

The complete nucleotide sequence of the *Euglena gracilis* chloroplast genome is a significant addition to the existing chloroplast data set and will facilitate several important lines of investigation. The Euglena sequence is especially important because it is the first complete sequence from outside the land plants and adds much needed diversity to the knowledge of plastid genomes. The complete sequences of plastid genomes are very useful for detailed analysis of plastid genome rearrangements as well as gene-by-gene comparisons of plastid genome contents. These data may contribute new information to the ongoing controversy of whether plastids have mono- or polyphyletic origins (43). The *Euglena gracilis* plastid sequence will also be useful in testing the hypothesis that euglenoid plastids are chimaeric in origin (44).

Information from the complete sequence of *Euglena gracilis* chloroplast DNA will be a basis for future studies on the origin of chloroplasts, the development of the photosynthetic apparatus in eukaryotes, and the evolution of chloroplast genes and introns. Although Euglena contains some chloroplast genes such as *rps9* and *psaM*, and five putative new genes internal to introns, the overall coding capacity is the most restricted of any photosynthetic eukaryote. The group II introns, although clearly related to their fungal mitochondrial, plant mitochondrial, and chloroplast counterparts, are unique in their relatively small size, and potential evolutionary progenitor relationship with the group III introns. Although there is now a complete DNA sequence for Euglena chloroplasts, we anticipate that many new insights on mechanisms of RNA transcription, RNA processing, and splicing in Euglena chloroplasts will be forthcoming.

ACKNOWLEDGEMENTS

We wish to acknowledge all who have contributed to this project. Many former colleagues with published sequences are recognized in the annotations of EMBL accession no. X70810. In addition, from Tucson we thank Donald Copertino, Alexander Dvorak, Gloria Yepiz-Plascencia, Catherine Radebaugh, Barry Roth, Jennifer Stevenson, and Thomas Tubman. From Neuchâtel, we thank Philippe Chatellard, Thomas Lemberger, Sophie Marc-Martin, Charareh Pourzand, and Yves Stauffer. This work has been supported by U.S.P.H.S. National Institutes of Health (to R.B.H.) and by Fonds national suisse de la recherche scientifique (to E.S.), and Roche Research Foundation, Basel (to E.S.).

REFERENCES

1. Sogin, M. L., Elwood, H. J. & Gunderson, J. H. (1986) *Proc. Natl. Acad. Sci. U S A* **83,** 1383–1387.
2. Kivic, P. A. & Walne, P. L. (1984) *Origins of Life* **13,** 269–288.
3. Morden, C. W. & Golden, S. S. (1991) *J. Mol. Evol.* **32,** 379–395.
4. Douglas, S. E. & Turner, S. (1991) *J. Mol. Evol.* **33,** 267–273.
5. Brawerman, G. & Eisenstadt, J. M. (1964) *Biochim. Biophys. Acta* **91,** 477–485.
6. Manning, J. E., Wolstenholme, D. R., Ryan, R. S., Hunter, J. A. & Richards, O. C. (1971) *Proc. Natl. Acad. Sci. USA* **68,** 1169–1173.
7. Hallick, R. B. & Buetow, D. E. (1989) In Buetow, D. E. (ed), *The Biology of Euglena.* Academic Press, Inc., San Diego, California, 352–414.
8. Schlunegger, B., Fasnacht, M., Stutz, E., Koller, B. & Delius, H. (1983) *Biochim. Biophys. Acta* **739,** 114–121.
9. Schlunegger, B. & Stutz, E. (1984) *Curr. Genet.* **8,** 629–634.
10. Koller, B. & Delius, H. (1982) *EMBO J.* **1,** 995–998.
11. Copertino, D. W. & Hallick, R. B. (1991) *EMBO J.* **10,** 433–442.
12. Orsat, B., Chatellard, P. & Stutz, E. (1992) In Murata, N. (ed), *Research in Photosynthesis.* Kluwer Academic Publishers, Dordrecht, 255–258.
13. Christopher, D. A. & Hallick, R. B. (1989) *Nucleic Acids Res.* **17,** 7591–7608.
14. Copertino, D. W., Christopher, D. A. & Hallick, R. B. (1991) *Nucleic Acids Res.* **19,** 6491–6497.
15. Wolfe, K. H., Morden, C. W. & Palmer, J. D. (1992) *Proc. Natl. Acad. Sci. U S A* **89,** 10648–10652.
16. Hallick, R. B., Richards, O. C. & Gray, P. W. (1982) In Edelmann, M., Hallick, R. B. & Chua, N.-H. (ed), *Methods in chloroplast molecular biology.* Elsevier Biomedical, New York, 281–294.
17. Devereux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
18. Hallick, R. B. & Bottomley, W. (1983) *Plant Mol. Bio. Rep.* **1,** 38–43.
19. Hallick, R. B. (1989) *Plant Mol. Biol. Reporter* **7,** 266–275.
20. Christopher, D. A. & Hallick, R. B. (1990) *Plant Cell* **2,** 659–671.
21. Siemeister, G., Buchholz, C. & Hachtel, W. (1990) *Curr. Genet.* **18,** 457–464.
22. Reith, M. & Munholland, J. (1993) *Plant Cell* **5,** 465–475.
23. Michalowski, C. B., Pfanzagl, B., Loffelhardt, W. & Bohnert, H. J. (1990) *Mol. Gen. Genet.* **224,** 222–231.
24. Douglas, S. E. (1991) *Curr. Genet.* **19,** 289–294.
25. Loffelhardt, W. & Bohnert, H. (1993) In Bryant, D. A. (ed), *The Molecular Biology of the Cyanobacteria.* Kluwer Academic Publishers, Dordrecht,
26. Orsat, B., Montfort, A., Chatellard, P. & Stutz, E. (1992) *FEBS Lett.* **303,** 181–184.
27. Roux, E. & Stutz, E. (1985) *Curr. Genet.* **19,** 221–227.
28. Siemeister, G. & Hachtel, W. (1990) *Plant Mol. Bio.* **14,** 825–833.
29. Siemeister, G. & Hachtel, W. (1990) *Curr. Genet.* **17,** 433–438.
30. Siemeister, G., Buchholz, C. & Hachtel, W. (1990) *Mol. Gen. Genet.* **220,** 425–432.
31. Michel, F., Umesono, K. & Ozeki, H. (1989) *Gene* **82,** 5–30.
32. Copertino, D. W., Shigeoka, S. & Hallick, R. B. (1992) *EMBO J.* **11,** 5041–5050.
33. Drager, R. G. & Hallick, R. B. (1993) *Nucleic Acids Res.* **21,** 2389–2394.
34. Gilbert, W., Marchionni, M. & McKnight, G. (1986) *Cell* **46,** 151–154.
35. Darnell, J. E. & Doolittle, W. F. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 1271–1275.
36. Dorit, R. L., Schoenbach, L. & Gilbert, W. (1990) *Science* **250,** 1377–1382.
37. Tittiger, C., Whyard, S. & Walker, V. (1993) *Nature* **361,** 470–472.
38. Gingrich, J. C. & Hallick, R. B. (1985) *J. Biol. Chem.* **260,** 16156–16161.
39. Rogers, J. H. (1989) *TIGS* **5,** 213–216.
40. Cavalier-Smith, T. (1991) *TIGS* **7,** 145–148.
41. Palmer, J. D. & Logsdon, J. M., Jr. (1991) *Curr. Opinion Genet. Dev.* **1,** 470–477.
42. Palmer, J. D. (1991) In Bogorad, L. & Vasil, I. K. (ed), *Molecular biology of plastids.* Academic Press, San Diego, 5–53.
43. Gray, M. W. (1991) In Bogorad, L. & Vasil, I. K. (ed), *Molecular biology of plastids.* Academic Press, Inc., San Diego, 303–330.
44. Wolfe, K. H., Morden, C. W. & Palmer, J. D. (1991) *Curr. Opinion Genet. Dev.* **1,** 523–529.