1144

Université de Neuchâtel
Faculté des sciences
Institut de Botanique
Laboratoire de Biochimie Végétale

Contribution à l'étude du génome chloroplastique de l'algue *Euglena gracilis* : séquençage des gènes psbD, ccsA et d'une partie du psbC

Version réduite de la thèse présentée à la Faculté des Sciences
par

Bernard Orsat

Chimiste diplômé de l'Université de Neuchâtel pour l'obtention
du grade de docteur ès sciences

1992

IMPRIMATUR POUR LA THÈSE

Contribution à l'étude du génome chloroplastique de l'algue Euglena gracilis: Séquençage des gènes psbD, ccsA et d'une partie du psbC

de M. Bernard Orsat

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel
sur le rapport des membres du jury,

Messieurs E. Stutz, P. Schürmann et R. Schantz (Strasbourg),

autorise l'impression de la présente thèse.

Neuchâtel, le 25 août 1992

Le doyen:

A. Robert

Liste des Publications

- Orsat B., Monfort. A., Chatellard. Ph. & Stutz. E., " Mapping and sequencing of an actively transcribed *Euglena gracilis* chloroplast gene (ccsA) homologous to the *Arabidopsis thaliana* nuclear gene cs (ch-42) ", FEBS Lett., **303**, 181 (1992).

- Orsat, B., Chatellard, Ph. & Stutz, E., " *Euglena gracilis* chloroplast DNA : Anatomy and transcription of a DNA segment coding for the genes ccsA, psbD and psbC " in : Proceedings of the IXth International Congress on Photosynthesis Research, Kluwer Academic Publishers, Nagoya, 1992, in press.

Le texte complet de la thèse est déposé chez Monsieur le Professeur E. Stutz au Laboratoire de Biochimie Végétale de l 'Université de Neuchâtel.

Volume 303, number 2,3, 181–184 FEBS 11088 June 1992

Mapping and sequencing of an actively transcribed *Euglena gracilis* chloroplast gene (*ccs*A) homologous to the *Arabidopsis thaliana* nuclear gene *cs(ch-42)**

Bernard Orsat, Amparo Monfort*, Philippe Chatellard and Erhard Stutz

Laboratoire de Biochimie végétale, Université de Neuchâtel, Chantemerle 18, CH-2000 Neuchâtel, Switzerland

Received 2 March 1992; revised version received 9 April 1992

We mapped and sequenced a novel chloroplast gene encoding a protein (348 amino acids) which shows a high sequence identity with both the decoded nuclear *cs(ch-42)* gene product of *Arabidopsis thaliana*, and the C-terminal half of the decoded '*crt*A' gene product of *Rhodobacter capsulatus*. The chloroplast gene (*ccs*A) is split (two exons) and transcribed into a stable mRNA of about 1200 nucleotides. The putative protein may be involved in the biosynthesis of photosynthetic pigments.

Euglena gracilis; *ccs*A chloroplast gene; *Arabidopsis thaliana*; *cs(ch-42)* nuclear gene

1. INTRODUCTION

During our studies on structure and function of the *E. gracilis* chloroplast genome we mapped and sequenced a DNA segment with a split ORF coding for a protein of 348 amino acids. Using the FASTA service offered by EMBL, Heidelberg, we found that the amino acid sequence was very similar to a reeently published decoded sequence of the nuclear gene *cs(ch-42)* of *A. thaliana* [1]. According to this study *cs* is a light-regulated gene encoding a chloroplast protein which is imported into chloroplasts as shown by in vitro experiments. In case of *E. gracilis* the corresponding gene is located and expressed within the chloroplast. We propose to call this gene *ccs*A. It represents a chloroplast gene homologous to the *A. thaliana* nuclear *cs* gene which most likely is involved in chloroplast pigment biosynthesis.

2. MATERIALS AND METHODS

The previously described DNA fragment Bgl Z [2] was cloned into the *Bam*HI site of the vector Bluescript KSII- (pEgKS-Z). We subcloned a *Hin*dIII-*Bgl*II fragment (2935 bp) by digestion of the construct pEgKS-Z with *Hin*dIII. Fragments were separated on agarose gels (1%) and fragments of the appropriate length (the *Hin*dIII map of Bgl Z is known) were eluted from the gel (Biotrap, Schleicher & Schuell) and religated. The clone pEgKS-2.9 with the 2935 bp insert was totally sequenced (B. Orsat, Ph.D. Thesis, Neuchâtel, 1992) following standard protocols (STRATAGENE). Overlapping smaller fragments were generated by cutting with *Hin*dIII and *Kpn*I, followed by selective degradation with exonuclease III (BRL), blunt ending with mung bean nuclease (Promega) and religation. The 2935 bp insert carries at one end the 5'-terminal part of a tRNA-Leu (CAA) gene which is cut by *Hin*dIII as published [3].

Chloroplast RNA was isolated and purified as published [4]. Northern hybridization was done in 5 × SSPE based solutions, 50% formamide at 42°C. Filters (Schleicher & Schuell, BA83) were washed twice with buffer 2 × SSC, 0.1% SDS, 42°C for 15 min.

Nucleotide and amino acid sequence data were analysed using the sequence analysis software package of Genetics Computer Group (GCG), Wisconsin.

3. RESULTS AND DISCUSSION

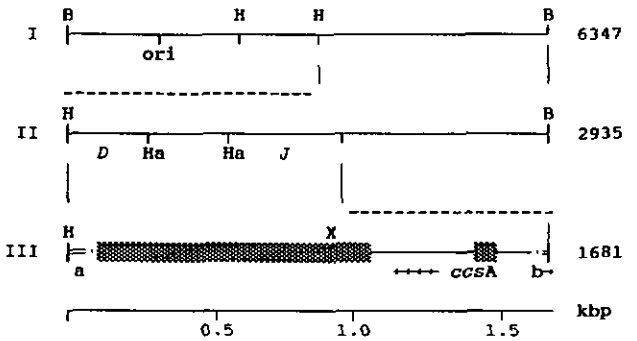
We show in Fig. 1(I,II,III) the position of the *ccs*A gene on the Bgl Z fragment which was previously mapped on the chloroplast genome [2]. The *ccs*A gene is situated between the *trn*L (CAA) [3] and the *psb*D gene-sharing transcription polarity with the *trn*L but not with the *psb*D gene. The coding part (ORF348) is split into exon 1 and 2 with 21 and 327 codons, respectively. The intron (332 bp) has canonical 5'- and 3'-termini and features of a chloroplast group II intron [5].

The *Euglena ccs*A gene encodes a protein (M_r = 39,307) having a high sequence identity (70%) with the *cs* nuclear gene of *A. thaliana* [1] and the C-terminal half of the '*crt*A'gene of *R. capsulatus* (Fig. 2) which according to a personal note (G.A. Armstrong, ETHZ, Zurich) and contrary to the published data [6] represents an independent ORF (*bch*I). Accordingly, we used in the alignment study only the C-terminal part of the

*The DNA sequence given in this article has received the EMBL data Library Accession number: X65484.

**Present address:* Centro de Investigacion y Desarrollo, CSIC, Depto. Genetica Molecular, Jordi Girona 18-26, 08034 Barcelona, Spain.

Correspondence address: E. Stutz, Laboratoire de Biochimie végétale, Université de Neuchâtel, Chantemerle 18, CH-2000 Neuchâtel, Switzerland. Fax: (41) (38) 242 695.



←

Fig. 1. Mapping of ORF 348. I: Bgl Z (6347 bp); II: *Bgl*I-*Hin*dIII fragment (2935 bp); III: subfragment of II (1681 bp, arbitrary cut). B, H, Ha, X respectively, are *Bgl*I, *Hin*dIII, *Hae*II and *Xba*I cleavage sites. *D,J* are *Hae*II fragments as published [9]. Note, however, that the two fragments *D* and *J* are separated by a small *Hae*II fragment (495 bp) not noticed previously. a,b represent, respectively, 5'-terminal part of *trn*L(CAA) and N-terminal part of exon 1 of *psb*D. ▩ exons of *ccs*A interrupted by intron. ← polarity of transcription. ori, origin of DNA replication [10].

```
                                    1
A   1   MASLLGTSSSAIWASPSLSSPSSKPSSSPICFRPGKLFGSKLNAGIQIRPKKNRSRYHVS

                                    3
E                 2     MNKKTNERPVFPFTSIVGQEEMKLSLILNVIDPKIGGVMIMGD
A   61  VMNVATEINSTEQVVGKFDSKKSARPVYPFAAIVGQDEMKLCLLLNVIDPKIGGVMIMGD
R               MTTAVARLQPSASGAKTRPVFPFSAIVGQEDMKLALLLTAVDPGIGGVLVFGD
                                  ***.**  ****..*** *.*  .** ****.. **

E   44  RGTGKSTIVRALVDLLPPIDVIENDPYNSDPYDTELMSDDVLEKIKKNEKVSIIQVKTPM
A  121  RGTGKSTIVRSLVDLLPEINVVAGDPYNSDPIDPEFMGVEVRERVEKGEQVPVIATKINM
R   54  RGTGKSTAVRALAALLPEIEAVEGCPVSSPNVEMIPDWATVLSTNV--------IRKPTPV
        ******* ** *  *** *  .   *  *   .       *             *     .

E  104  VDLPLGGTEDRVCGTIDIEKAISEGKKAFEPGLLAQANRGILYVDEVNLLDDHLVDVLLD
A  181  VDLPLGATEDRVCGTIDIEKALTEGVKAFEPGLLAKANRGILYVDEVNLLDDHLVDVLLD
R  107  VDLPLGVSEDRVVGALDIERAISKGEKAFEPGLLARANRGYLYIDECNLLEDHIVDLLLD
        ******  **** * .***.*.  * ********* **** **.** ***.**.**.***

E  164  SAASGWNTVEREGVSICHPARFILVGSGNPEEGELRPQLLDRFGMHAQIKTLKEPALRVK
A  241  SAASGWNTVEREGISISHPARFILIGSGNPEEGELRPQLLDRFGMHAQVGTVRDADLRVK
R  167  VAQSGENVVERDGLSIRHPARFVLVGSGNPEEGDLRPQLLDRFGLSVEVLSPRDVETRVE
        * ** * ***.*.** *****.*.********.**********.  ..   ..   **

E  224  IVQQRELFEKSPKEFKEKYKEEQNKLMEKIINARKKLKNIIIKYELLEKISQICSELNVD
A  301  IVEERARFDSNPKDFRDTYKTEQDKLQDQISTARANLSSVQIDRELKVKISRVCSELNVD
R  227  VIRRRDTYDADPKAFLEEWRPKDMDIRNQILEARERLPKVEAPNTALYDCAALCIALGSD
        ..  *  ..  ** * .  .  .  .  *  **  *  .           .*  *  *

E  284  GLRGDMVTSRAAKALVAFEDRTEVTPKDIFTVITLCLRHRLRKDPLESIDSGYKVQETFK
A  361  GLRGDIVINRAAKALAALKGKDRVTPDDVATVIPNCLRHRLRKDPLESIDSGVLVSEKFA
R  287  GLRGELTLLRSARALAALEGATAVGRDHLKRVATMALSHRLRRDPLDEAGSTARVARTVE
        ****..   * *.** *.     *    .  *    * ****.***.   *   *

E  344  KVFNY
A  421  EIFS-
R  347  ETLP-
```

Fig. 2. Amino acid sequence alignment of decoded genes. E: *E. gracilis* chloroplast *ccs*A, A: *A. thaliana cs*, R: *R. capsulatus* C-terminal part of '*crt*A' (*bch*I, 350 codons) starting with MTT.., as suggested by G.A. Armstrong (see text). (*) exact matches, (·) conservative matches across all sequences; 1 and 2 mark intron positions in nuclear *cs* gene; 3 marks intron position in *ccs*A gene. (–) gap.

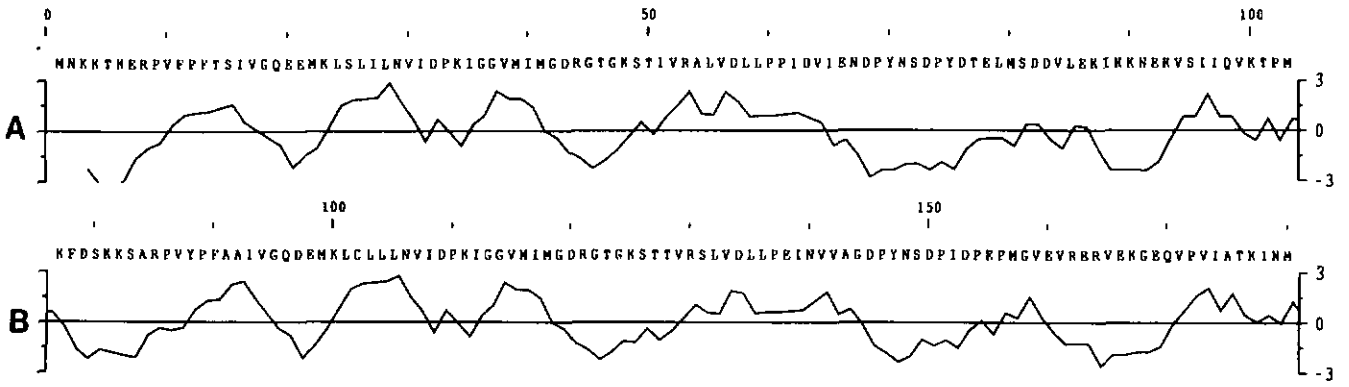


Fig. 3. Hydropathy plot (Kyte and Doolittle) of the N-terminal part (positions 1–103) of the decoded *ccs*A gene (profile A) and the equivalent segment (positions 77–200) of the decoded *A. thaliana cs* gene (profile B); numbering is according to [1].

published sequence starting with MTTA. We notice large domains of exact and conservative matches across all three sequences what strongly suggests that these proteins not only have a common evolutionary origin but most likely have an equivalent function.

The *A. thaliana cs* protein sequence has a long N-terminal part which qualifies as transit peptide [1] and therefore has no equivalent sequence in the chloroplast and bacterial counterpart. According to the result of Fig. 2, a first conservative domain (RPV..) starts at position 85 (line A), i.e. the transit sequence most likely ends upstream of that domain and not at position 93 (line A) as tentatively assumed [1]. In that context it was of interest to compare the hydropathy plot of the *ccs*A N-terminal part (103 amino acids) with the equivalent sequence of the *cs* gene (Fig. 3). Certainly the two profiles are congruent from position 80 on (*A. thaliana*) strongly suggesting that the processed *cs*-protein starts in that region.

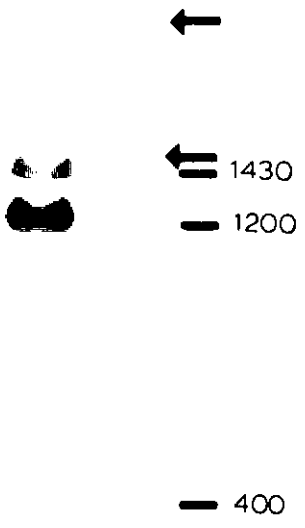


Fig. 4. Northern blot. Purified chloroplast RNA is hybridized with a DNA probe (*Hin*dIII-*Xba*I, 925 bp). Arrows mark the position of 23 and 16S rRNA; RNA fragment size is given in nucleotides.

The *ccs*A gene is transcribed in light grown cells, as shown in Northern hybridization experiments (Fig. 4). The stable transcript is about 1200 nucleotides long. In addition to the major band a precursor of about 1430 nucleotides interacts with the probe and a very faint band around 400 nucleotides can be detected on the radiograph. From previous studies (A. Monfort, Ph.D. thesis, Neuchâtel, 1990) we know that the *trn*L gene is co-transcribed with upstream elements and all indications are that the *ccs*A gene and the downstream *trn*L gene are part of a primary transcript which undergoes several steps of processing including the splicing of exon 1 with exon 2.

The function of the *ccs*A gene is presently unknown. Mutations in the *cs* gene, or, e.g. a T-DNA insertion in the 3'-end of the coding part lead to loss of chloroplast pigments (pale mutant) [1]. The same holds for mutations in the C-terminal part of the '*crt*A' gene (*bch*I) of *R. capsulatus*: such mutants show loss of bacteriochlorophyll accumulation [7,8].

Considering the close structural relationship between the *Euglena* chloroplast *ccs*A gene with both the plant and bacterial counterparts we postulate that the chloroplast gene is also involved in chloroplast pigment biosynthesis. If such is the case then *ccs*A is the first identified chloroplast gene participating in chlorophyll accumulation.

The *ccs*A gene represents another example for genes transferred to the nuclear DNA of higher plants but retained in the algal chloroplast genome as was shown for the *tuf*A gene [11,12]. Higher plant plastoms also show differences in gene composition. It was, e.g., reported that the tobacco and rice chloroplast genome contain the *rps*16 but lack the *rpl*21 gene while the opposite is true for *Marchantia polymorpha* [13].

Acknowledgements: We are grateful to Drs. G.A. Armstrong, ETHZ, Zurich and C. Koncz, Max-Planck-Institute, Cologne for valuable information about unpublished data. This work represents part of the Ph.D. thesis of B.O. and receives support from the Fonds national suisse de la recherche scientifique, to E.S.

REFERENCES

[1] Koncz, C., Mayerhofer, R., Koncz-Kalman, Z., Nawrath, C., Reiss, B., Redei, G.P. and Schell, J. (1990) EMBO J. 9, 1337–1346.
[2] Schlunegger, B. and Stutz, E. (1984) Current Genetics 8, 629–634.
[3] Monfort, A., Rutti, B. and Stutz, E. (1986) Nucleic Acids Res. 14, 3971.
[4] Montandon, P.E., Knuchel-Aegerter, C. and Stutz, E. (1987) Nucleic Acids Res. 15, 7809–7822.
[5] Michel, F., Umesone, K. and Ozeki, H. (1989) Gene 82, 5–30.
[6] Armstrong, G.A., Albert, M., Leach, F. and Hearst, J.E. (1989) Mol. Gen. Genet. 216, 254–268.
[7] Zsebo, K.M. and Hearst, J.E. (1984) Cell 37, 937–947.
[8] Armstrong, G.A., Schmidt, A., Sandmann, G. and Hearst, J.E. (1990) J. Biol. Chem. 265, 8329–8338.
[9] Hallick, R.B. and Buetow, D.E. (1989) in: The Biology of *Euglena*, vol. 4 (Buetow, D.E. ed.) pp. 351–414. Academic Press, London.
[10] Koller, B. and Delius, H. (1982) EMBO J. 1, 995–998.
[11] Montandon, P.E. and Stutz, E. (1983) Nucleic Acids Res. 11, 5877–5891.
[12] Baldauf, S.L. and Palmer, J.D. (1990) Nature 344, 262–265.
[13] Sugiura, M., Torazawa, K. and Wakasugi, T. (1991) in: NATO ASI Series, The Translational Apparatus of Photosynthetic Organelles, vol. H55 (Mache, R., Stutz, E. and Subramanian, A.R. eds.) pp. 59–69.

Euglena gracilis chloroplast DNA: Anatomy and transcription of a DNA segment coding for the genes *ccs*A, *psb*D and *psb*C

B.Orsat, Ph. Chatellard and E. Stutz. Laboratoire de Biochimie végétale, Université de Neuchâtel, CH-2000 Neuchâtel, Switzerland

The *Euglena gracilis* chloroplast genome has an average A+T content of 75%. This high value is essentially due to large A+T rich intergenic spacers and to numerous introns within protein coding genes (1). We have cloned and sequenced a 22 kb DNA segment containing the genes *ccs*A (2), *psb*D and *psb*C (3). We describe in the following some structural and transcriptional aspects of the *psb*D gene (D2-protein) which is particularly rich in introns,i.e., the coding part (10 exons) represents only 12% of the DNA sequence between start and stop codon.

In Fig.1 we show the relative positions of the *ccs*A and *psb*D genes. Pertinent restriction sites are indicated. The *psb*C gene is downstream of *psb*D (not shown). *ccs*A and *psb*D have opposite transcription polarity, *ccs*A being the first gene in a long row of genes (same strand) covering about one third of the genome (includes, e.g. *rrn* operons A,B,C). *psb*D is the first gene in a row of multiply split genes including *psb*C (3) and *psb*A (4).

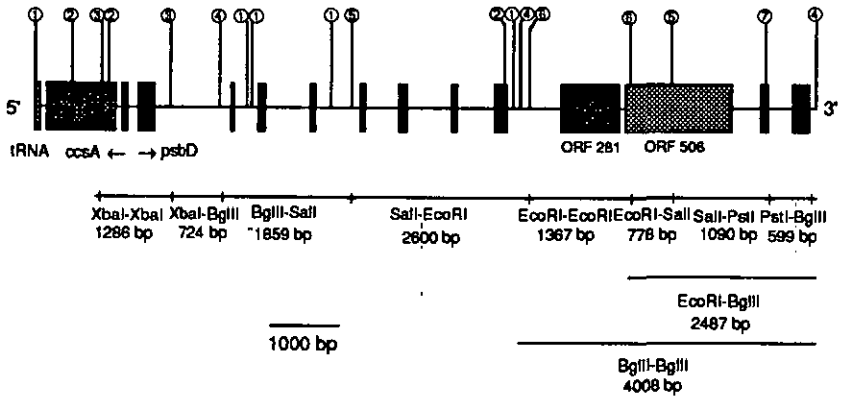


Fig.1. Map position and general structures of the genes *ccs*A (2) and *psb*D and position of restriction fragments used as probes in Northern experiments. 1) HindIII; 2) HaeII; 3) XbaI;4) BglII; 5) SalI; 6) EcoRI; 7) PstI.Two large orfs of unknown function are marked in intron-8; ⟷ transcription polarity.

The nucleotide sequence of the spliced coding part of *psb*D is given in Fig.2 along with the D2 amino acid sequence and a short decoded sequence potentially coding for the 5'-terminal part of *psb*C (5). The first 352 positions of the D2 protein are highly conserved (about 86% sequence identity with other D2 proteins,) but the reading frame continues for another 31 codons not seen in any D2 protein sequenced so far. We can exclude any sequencing error but it could be that an additional sequence (trailor) with a stop codon at the appropriate site is spliced to the 3'-end cutting off the 31 C-terminal codons. Such an assumption is supported by the electron-microscopic analysis of this region: two and not

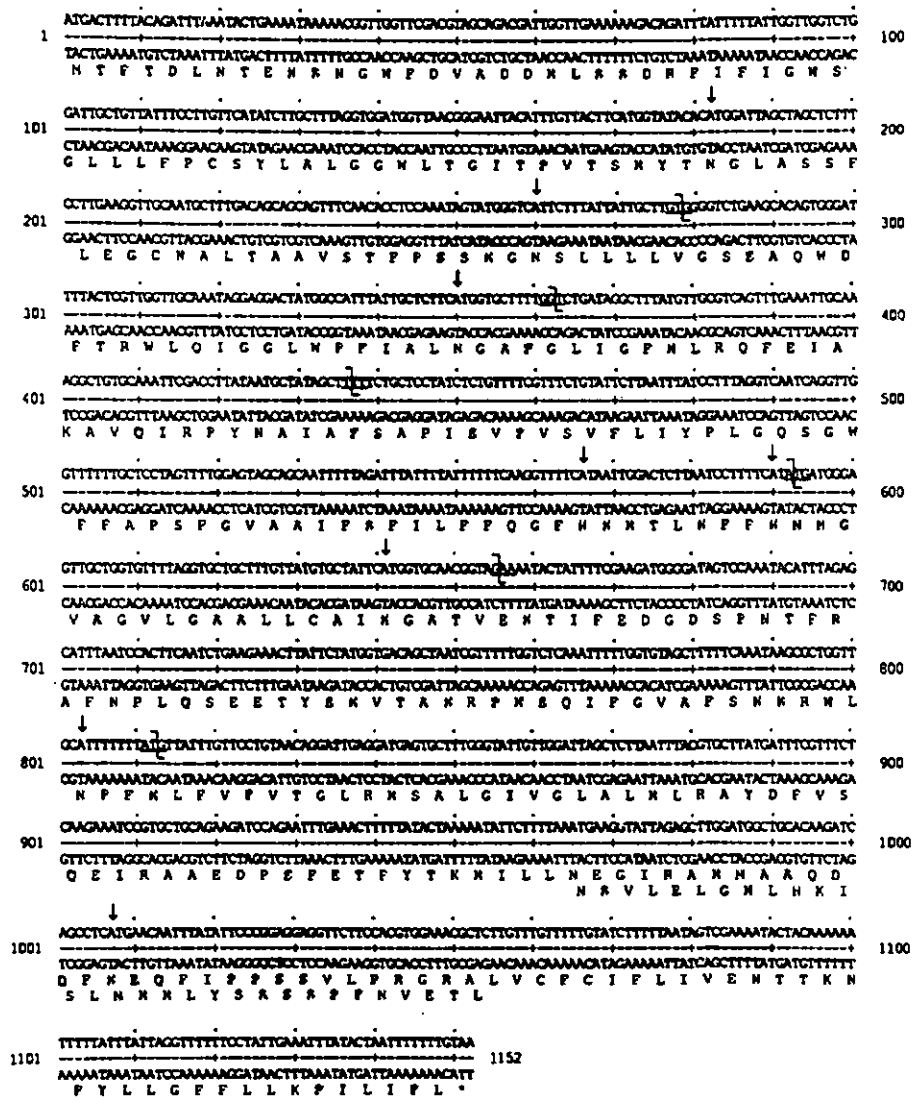


Fig.2. Nucleotide sequence of the coding part of *psb*D and the deduced aminoacid sequence of D2 protein. Splice sites are marked and split codons underlined; ↓ marks crucial histidines; the orf starting at position 965 may contain the N-terminus (VETL.) of the 44kD protein of PSII (*psb*C, 5).

only one small R-loops are visible downstream of the large intron-8 (6). If such is the case the final mRNA would be the product of 10 splicing events.

Table 1 Characteristics of introns and exons of the *psb*D gene

intron	size bp	5'-end	3'-end	exon	size bp
1	1098	-/GTGTG	TCTATTCTTT/-	1	243
2	364	-/GTGTC	TAGTTTTATT/-	2	35
3	605	-/GTGTG	ATTTATTCTT/-	3	84
4	651	-/GTGTG	CGACTTTGAC/-	4	74
5	498	-/GTGCG	TTAATTCTAT/-	5	41
6	606	-/GTGTG	CTACTTTAAC/-	6	115
7	580	-/GTGCT	CAATTTTCTC/-	7	63
8	3658	-/GGGTA	TCACCCTCAC/-	8	157
9	373	-/GTGTG	CGACTTCTAC/-	9	118
				10	222

In Table 1 some characteristics of *psb*D introns and exons are summarized. The intron 5'-ends follow with minor exceptions the canonical sequence 5'-GYGYG- and the 3'-ends have typical short pyrimidine clusters and they always terminate with a pyrimidine. The 3'-terminal part of all nine introns can form loop V and VI structures (not shown) typical for class II introns. Intron-8 (3658 bases) is the largest chloroplast intron (twintron?, 7) yet reported.

We used several DNA probes from the *psb*D gene (see Fig.1) to identify in Northern experiments the processed stable mRNA and its precursors (Fig.3 and 4). We obtained very complex patterns as expected, but certain bands dominate allowing the following interpretations: 1) Whenever the DNA probe contains one or more exons a strong band of about 1.4 kb appears (panels 3A,B,D,E,F,4A,D,E). We consider this to be the functional mRNA coding for the D2 protein. 2) The 1.4 kb band is absent in panels 3C and 4B,C which were obtained with intron specific probes. This result corroborates the interpretation given under point 1. 3) The largest precursor (top faint band) is about 9.3 kb what approximately equals the sum of the 10 exons and 9 introns listed in Table 1 (however, see above). 4) The intron-1 specific probe strongly interacts with a transcript of about 1.1 kb (bottom band of 3C) what most likely is the excised intron-1 since this band is not discernible in any of the other panels. 5) We notice a relatively strong band

of 5.3 kb in most profiles and in particular with the intron-8 specific probes (4B,C). Obviously, this p-mRNA still contains the largest intron and all but one of the smaller introns are excised [5300 - 3658 ≈ 1.8 kb].

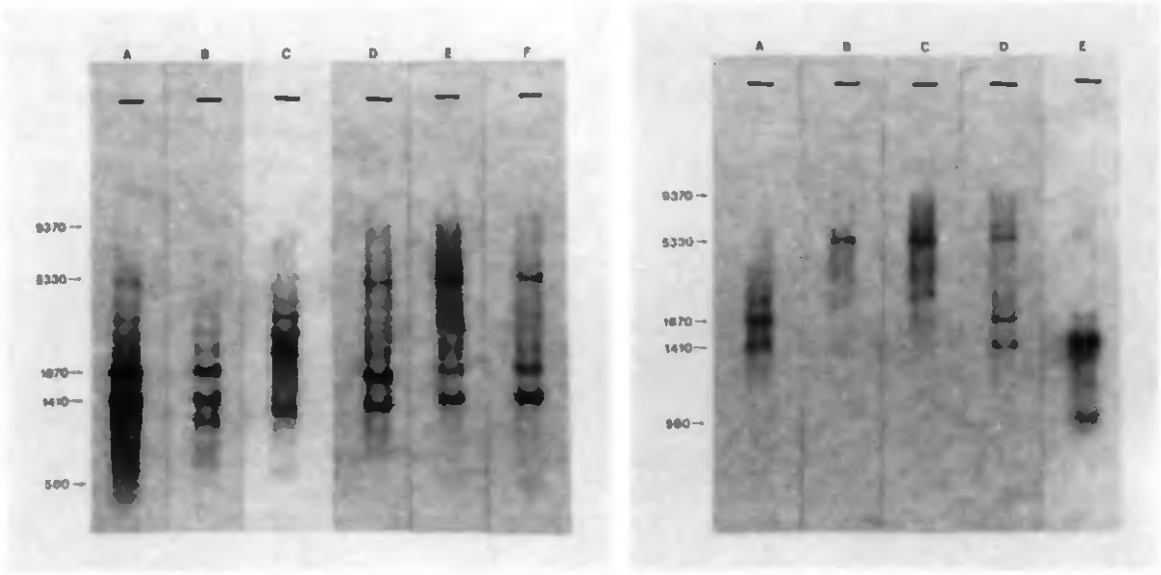


Fig.3. Northern blots of purified chloroplast RNA hybridized with DNA fragments from *psb*D, consult Fig.1. A,B) XbaI-XbaI 1286; C) XbaI-BglII 724; D) SalI-EcoRI 2600; E) BglII-BglII 4008; F) EcoRI-BglII 2487. The size (bases) of major bands is given in the margin.

Fig.4. see legend of Fig 3. A) BglII-SalI 1859; B) EcoRI-EcoRI 1367; C) EcoRI-SalI 778; D) SalI-PstI 1090; E) PstI-BglII 599.

References

1. Hallick,R.B. and Buetow,D.E.(1989) in The Biology of Euglena (Buetow,D.E.,ed.) vol.4, pp.351-414, Academic Press, San Diego
2. Orsat,B. Montfort,A. Chatellard,P. and Stutz,E. (1992) FEBS Lett.303, 181-184
3. Montandon,P.E. Vasserot,A. and Stutz,E. (1986) Curr. Genet.11,35-39
4. Keller,M. and Stutz,E. (1984) FEBS Lett.175,173-177
5. Holschuh,K. Bottomley, W. and Whitfeld, P.R. (1984) Nucleic Acids Res.12, 8819-8834
6. Koller,B. and Delius, H. (1984) Cell 36, 613-622
7. Copertino,D.W. and Hallick,R.B. (1991) EMBO J.10, 433-442

Acknowledgment: We are grateful to J.D. Rochaix, University of Geneva for giving us a *psb*D DNA probe (*C. reinhardtii*). This work is part of the Ph.D. thesis of B.O. and receives support from the *Fonds nationale suisse de la recherche scientifique*, to E.S.