Structure of the *Euglena gracilis* chloroplast gene *(psbA)* coding for the 32-kDa protein of Photosystem II

Mario Keller and Erhard Stutz+

Institut de Biologie Moléculaire et Cellulaire, 15 rue Descartes, F-67084 Strasbourg, +Laboratoire de Biochimie, Université de Neuchâtel, Chantemerle 18, CH-2000 Neuchâtel

Received 18 June 1984; revised version received 23 July 1984

An *Eco*RI fragment from *Euglena gracilis* chloroplast DNA, called Eco.I (4.9 kbp), contains the gene for the '32-kDa' thylakoid membrane protein of Photosystem II (*psbA* gene), the $tRNA_2^{Leu}$ gene and additional sequences which possibly code for two proteins of unknown function. The transcription polarity is the same for all these coding sequences. The *psbA* gene is split: 4 introns (size range, 433–616 bp) separate 5 exons (size range, 39–579 bp) which code for a protein of M_r 38380 (345 amino acids). The *Euglena* protein is about 87% homologous with the higher plant counterparts but it contains 5 lysine residues.

Euglena gracilis *Chloroplast genome* psbA trnL

1. INTRODUCTION

We have shown [1] that fragment Eco.I of the chloroplast genome of *Euglena gracilis* contains the *psbA* gene, which codes for the 32-kDa membrane protein of Photosystem II. This protein seems to be involved in the electron flow of Photosystem II [2] and in the binding of the urea and triazine herbicides [3,4]. We noticed that the Eco.I fragment, or at least parts of it, are transcribed at all stages of chloroplast development, and we observed, in 'Northern' type hybridization experiments, that Eco.I interacts with two major (14 S, 17 S) and several minor but larger (>17 S) transcription products. These observations suggested that Eco.I may contain, in addition to the *psbA* gene, other genes and that the transcription products of these genes may undergo several defined processing steps including splicing. Split protein-coding genes have not been found so far in chloroplast genomes of higher plants, but *Euglena* chloroplast DNA contains split *rbcL* [5,6] and *tufA* [7] genes. Furthermore, it was shown that Eco.I contains the *trnL2* gene [8], coding for $tRNA_2^{Leu}$. Very recently, it was reported that the *Euglena psbA* gene is transcribed into a large precursor (3.1 kb), which undergoes several processing steps including splicing, yielding finally a 1.2-kDa mature mRNA [9] probably identical to the 14 S RNA described in our studies, which directs the in vitro synthesis of the 32-kDa protein [1].

To determine the structure of the *Euglena psbA* gene and its exact position relative to the *trnL2* gene, it was necessary to sequence the entire Eco.I fragment. The *psbA* genes from several higher plants, including spinach, *Nicotiana debneyi* [10], soybean [11], *Amaranthus hybridus* [12] and mustard [13], and from the alga *Chlamydomonas reinhardii* [14] have been sequenced. It became evident that the corresponding gene product (the 32-kDa protein) has a highly conserved amino acid composition.

We report here the nucleotide sequence of the entire Eco.I fragment which, in addition to the *psbA* and *trnL2* genes, contains sequences coding for two proteins of unknown function.

2. MATERIALS AND METHODS

Euglena chloroplast DNA restriction fragment Eco.I (4.9 kbp) was inserted into the *Eco*RI site of pBR322. For large-scale isolation of recombinant plasmid DNA, we used the SDS lysis procedure

Published by Elsevier Science Publishers B.V.
00145793/84/$3.00

[15] followed by two cycles of CsCl/ethidium bromide density gradient centrifugations.

DNA fragments were labeled with ^{32}P at their 5′-ends, using T_4 polynucleotide kinase and labeled ends were separated either following a digestion with a second restriction enzyme or by strand separation [16]. The base-specific cleavage reactions were performed according to [16]. Enzymes were purchased from Boehringer-Mannheim and used as recommended by the supplier. [γ-^{32}P]ATP (3000 Ci/mmol) was obtained from Radiochemical Center, Amersham.

3. RESULTS AND DISCUSSION

3.1. *Arrangement of genes on DNA fragment Eco.I*

In fig.1a, we show the relevant restriction endonuclease sites on Eco.I and in fig.1b, the overall arrangement of the coding regions as deduced from the sequencing data given in fig.2 and explained in section 3.2. DNA fragment Eco.I which was previously mapped [17] has a total length of 4904 nucleotides. In fig.1b, one may see, from left to right (i) the 3′-terminal part coding for 108 amino acids of an open reading frame (URF2); (ii) another open reading frame (URF1), which has an ATG start and a TAA stop codon, coding for 132 amino acids; (iii) the *trnL2* gene (81 nucleotides) and (iv) the split *psbA* gene with its 5 exons and 4 introns. The transcription polarity is the same for all these coding sequences. The URF2 region may code for the C terminal part of a 46-kDa stroma polypeptide, which was previously identified as one of the translation products obtained using mRNA that interact with Eco.I in hybrid-select translation experiments [1]. Nothing is known about the possible translation product (M_r 14600) of URF1. The *trnL2* gene (anticodon TAA) of Eco.I is 55% homologous with the *trnL3* gene (anticodon TAG) mapped on Eco.G and sequenced in [18].

3.2. *Analysis of the* psbA *gene*

The nucleotide sequence of the entire Eco.I fragment is given in fig.2. We have only shown the RNA-like strand and have added, where appropriate, the corresponding amino acid sequence. In the *psbA* region, we have also shown those amino acid residues of the soybean 32-kDa protein which diverge from the *Euglena* chloroplast protein sequence. Using the deduced amino acid sequences of the two proteins as a guideline, we were able to identify 5 exons consisting of 78, 28, 193, 33 and 13 codons, and 4 introns of 433, 447, 434 and 616 nucleotides, respectively.

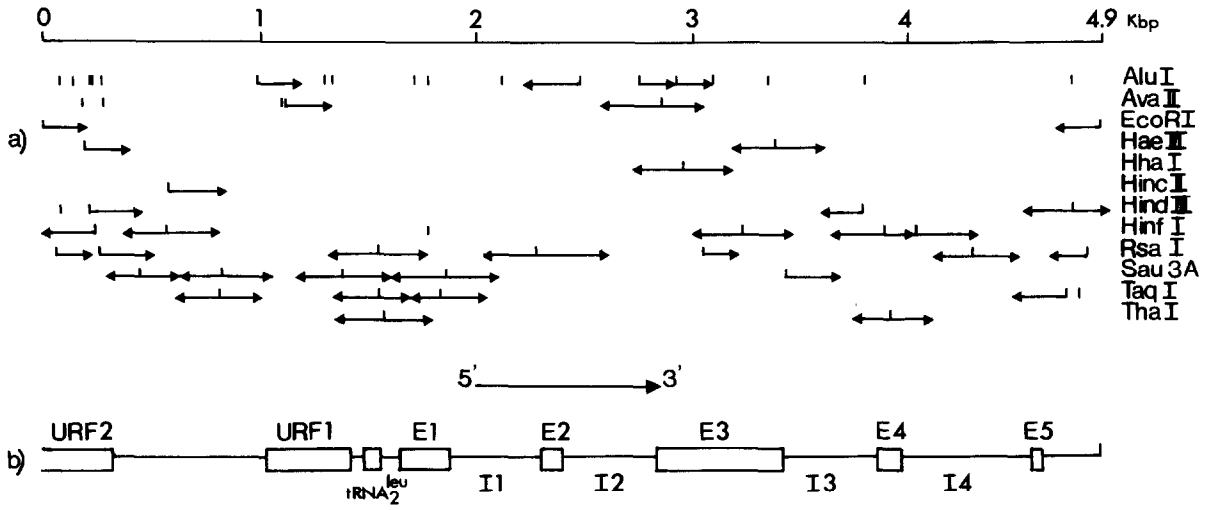


Fig.1. Restriction site map and gene arrangement of DNA fragment Eco.I. (a) Position of the relevant restriction sites and sequencing strategy. (b) Gene arrangement; URF, unidentified reading frame; the *psbA* gene consists of 5 exons (E_1–E_5) and 4 introns (I_1–I_4). The arrow indicates polarity of transcription.

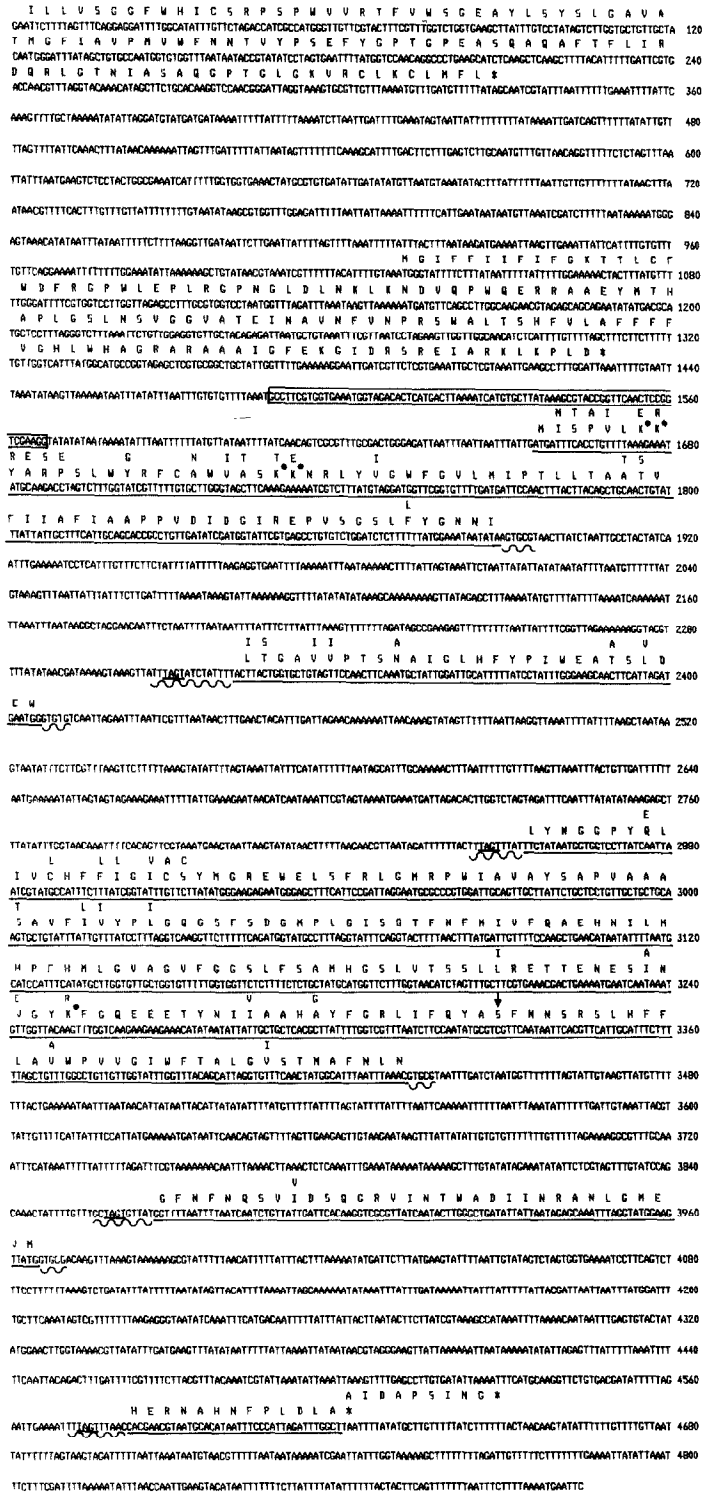
The translation start (exon 1) is 87 bp away from the 3′-end of the *trnL2* gene. The N-terminal part (78 amino acids) is only 74% homologous to the soybean counterpart, and this is far below the usual sequence homology of almost 100% found in higher plant 32-kDa proteins. This rather strong divergence raises the question of whether the *Euglena* 32-kDa protein gene actually starts at the first methionine residue, or at the second one located 38 codons downstream, which would reduce the size of the translation product from 38.38 kDa to about 34 kDa, as already suggested in the case of the *A. hybridus* 32-kDa protein [12]. This problem has also been studied by authors in [10], who have observed a high degree of homology in the nucleotide sequences between the two methionine codons in spinach and *N. debneyi*, and have concluded, on the basis of this sequence conservation, that the gene starts at the first methionine codon. However, studying the dipeptides synthesized in an in vitro system [19], authors in [20] have concluded that the translation of the 32-kDa protein in maize, tobacco and pea starts at the second methionine residue. Since a similar study has not been made in *Euglena*, the exact translation start of the 32-kDa protein from *Euglena* is not known. Between *Euglena* and soybean, there is only about 58% homology (22 out of 38 amino acids) between the first two methionine residues, and to align the *Euglena* properly with the soybean protein amino acid sequence, we had to put the N-terminus of the *Euglena* protein one residue ahead of that of the soybean start codon. Most remarkable are the 4 codons for lysine (positions 7, 8, 25 and 26), as no lysine codon has been found in any of the *psbA* genes of higher plants or of *C. reinhardii*.

Exon 2 and exon 3 code for 28 and 193 amino acids, respectively, representing the central part of the 32-kDa protein. This part is very similar to the soybean protein (about 90%). This region also contains the serine (Ser 265) which is replaced by glycine in the atrazine-resistant *Amaranthus* mutant [12] and by alanine in the *Chlamydomonas* mutant [21]. A fifth lysine is found at position 239 instead of an arginine in soybean. It is interesting to note that the *psbA* gene of the cyanobacterium *Anabaena* also contains a lysine residue at the same position [22].

Exon 4 and exon 5 are very short (33 and 13 codons, respectively). Sequence homology with the soybean protein is again very high (about 98%). However, the C-terminal part of the *Euglena* protein is shorter, lacking the last 9 amino acids found in the C-terminal part from higher plants. It is important to note that (i) exon 5, in contrast to the 4 other exons, is terminated by a stop codon (TAA) and (ii) that the missing 9 amino acids were not found further downstream. In the case of *C. reinhardii* [14], the last 8 residues also diverge very strongly from those of the higher plant 32-kDa proteins [10–13].

According to this study and others [1,9] the transcript of the *E. gracilis psbA* gene must undergo several splicing events to become a translatable mRNA. Intron-exon and exon-intron boundaries of the *Euglena rbcL* gene have been analysed and a 5′-intron consensus sequence 5′-GPyGPyG- and a 3′-intron consensus sequence -PyPyTAPuTTTTAT-3′ have been proposed [6]. From fig.2, it can be seen that the 4 introns have the 5′-intron consensus sequence proposed, but the first intron has an additional A between the last codon of exon 1 and the consensus sequence. However, none of the 4 introns terminates with the 3′-intron consensus sequence proposed. Nevertheless, 10–15 nucleotides upstream of the 3′-end the sequence 5′-TAGT (underlined twice on fig.2) may function as a recognition site for splicing. The two introns of the *Euglena tufA* gene [7] do not contain the proposed consensus sequences, suggesting that more than one splicing mechanism may be operating in *Euglena* chloroplasts. All 4 introns are very rich in AT and contain several stop codons, as the introns found in the *rbcL* and the *tufA* genes of *Euglena* [6,7].

Recently, the amino acid sequence of the *C. reinhardii* 32-kDa protein (as deduced from sequencing of the corresponding *psbA* gene) has been published [14]. This gene is also split and is composed of 5 exons and 4 introns. However, the positions of the introns do not coincide with those in the *Euglena* gene. Furthermore, the *Chlamydomonas* protein contains no lysine and is more homologous (about 94%) than the *Euglena* protein (87%) to the 32-kDa protein of higher plants.



ACKNOWLEDGEMENTS

We are grateful to Professor J.H. Weil for helpful discussions, encouragement and for critical reading of the manuscript. This research was supported in parts by the Fonds National Suisse de la Recherche Scientifique, grant 3.183.82 (to E.S.).

NOTE ADDED IN PROOF

A computer analysis performed by F. Michel on the nucleotide sequence upstream of the *trnL2* gene revealed the presence of two putative introns (intron 1, nucleotide 297–601 and intron 2, nucleotide 654–1076; fig.2). These two introns separate 3 exons of 98.17 and 117 codons, respectively, exon 1 (nucleotide 1–296) and exon 3 (nucleotide 1077–1426) being shortened versions of URF2 and URF1, respectively (fig.2). According to this hypothesis, the 3 exons and the 2 introns correspond to the 3′-terminal part of a split gene that may code for the 46-kDa stroma polypeptide [1].

REFERENCES

[1] Keller, M., Rutti, B. and Stutz, E. (1982) FEBS Lett. 149, 133–137.
[2] Mattoo, A.K., Pick, U., Hoffmann-Falk, H. and Edelman, M. (1981) Proc. Natl. Acad. Sci. USA 78, 1572–1576.
[3] Steinback, K.E., McIntosh, L., Bogorad, L. and Arntzen, C.J. (1981) Proc. Natl. Acad. Sci. USA 78, 7463–7467.
[4] Pfister, K., Steinback, K.E., Gardner, C. and Arntzen, C.J. (1981) Proc. Natl. Acad. Sci. USA 78, 981–985.
[5] Stiegler, G.L., Matthews, H.M., Bingham, S.E. and Hallick, R.B. (1982) Nucleic Acids Res. 10, 3427–3444.
[6] Koller, B., Gingrich, J., Farley, M., Delius, H. and Hallick, R.B. (1984) Cell 36, 545–553.
[7] Montandon, P.E. and Stutz, E. (1983) Nucleic Acids Res. 11, 5877–5892.
[8] Kuntz, M., Keller, M., Crouse, E.J., Burkard, G. and Weil, J.H. (1982) Curr. Genet. 6, 63–69.
[9] Hollingworth, M.J., Johanningen, U., Karabin, G.D., Stiegler, G.L. and Hallick, R.B. (1984) Nucleic Acids Res. 12, 2001–2017.
[10] Zurawski, G., Bohnert, H.J., Whitfeld, P.R. and Bottomley, W. (1982) Proc. Natl. Acad. Sci. USA 79, 7699–7703.
[11] Spielmann, A. and Stutz, E. (1983) Nucleic Acids Res. 11, 7157–7167.
[12] Hirschberg, J. and McIntosh, L. (1983) Science 222, 1346–1349.
[13] Link, G. and Langridge, U. (1984) Nucleic Acids Res. 12, 945–957.
[14] Erickson, J., Schneider, M., Vallet, J.M., Dron, M., Bennoun, P. and Rochaix, J.D. (1984) in: Advances in Photosynthesis Research (Sybesma, ed.) vol.4, 493–500, Martinus Nijhoff–Dr W. Junk, The Hague.
[15] Godson, G.N. and Vapnek, D. (1973) Biochim. Biophys. Acta 299, 516–521.
[16] Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol. 65, 499–560.
[17] Hallick, R.B. (1983) Chloroplast DNA. In: The Biology of Euglena (Buetow, D.E. ed.) vol.4, Academic Press, New York.
[18] Orozco, E.M. jr., and Hallick, R.B. (1982) J. Biol. Chem. 257, 3265–3275.
[19] Cenatiempo, Y., Twardowski, T., Redfield, B., Reid, B.R., Dauerman, H., Weissbach, H. and Brot, H. (1983) Proc. Natl. Acad. Sci. USA 80, 3223–3226.
[20] Cohen, B.N., Bloom, M.V., Coleman, T. and Weissbach, H. (1984) in: Molecular Biology of the Photosynthetic Apparatus (Arntzen, C. et al. eds) p.11, Cold Spring Harbor, New York.
[21] Erickson, J.M., Rahire, M., Bennoun, P., Delepelaire, P., Diner, B. and Rochaix, J.D. (1984) Proc. Natl. Acad. Sci. USA 81, 3617–3621.
[22] Curtis, S. and Haselkorn, R. (1984) Plant Mol. Biol., in press.

←

Fig.2. Nucleotide sequence of DNA fragment Eco.I. Only the RNA-like strand is given, along with the deduced amino acid sequence in the coding regions. For the *psbA* gene, the amino acid sequence of the soybean 32-kDa protein is aligned with that of the *Euglena* protein but only diverging amino acids are printed. The *trnL2* is boxed. The 5 exons of *psbA* are underlined. The 5′- and 3′-intron boundaries are underlined with wavy lines. The black points indicate the 5 lysine residues and the arrow the serine residue which is replaced by glycine in an atrazine-resistant mutant of *Amaranthus hybridus* [12] and by an alanine in an atrazine-resistant mutant of *Chlamydomonas* [21]. The asterisks indicate stop codons.