The introns of the *Euglena gracilis* chloroplast gene which codes for the 32-kDa protein of photosystem II

Evidence for structural homologies with class II introns

M. Keller and F. Michel*

Institut de Biologie Moléculaire et Cellulaire (CNRS), Université L. Pasteur, 15 rue Descartes, 67084 Strasbourg, and **Centre de Génétique Moléculaire du CNRS, Laboratoire Associé à l'Université Pierre et Marie Curie, 91190 Gif-sur-Yvette, France*

Received 18 September 1984; revised version received 31 October 1984

The four introns of the *Euglena* psbA gene, and in particular intron psbA4, show at their 5′- and 3′-terminal parts structural homologies with structured introns of class II. Two putative introns belonging to this class were also found in a URF located upstream of the psbA gene.

Euglena gracilis *psbA gene* *Chloroplast intron* *Intron class* *RNA secondary structure*

1. INTRODUCTION

While introns have not yet been found in the protein-coding genes of higher plant chloroplasts, the *Euglena gracilis* chloroplast DNA comprises at least 15 split protein genes, as judged from electron microscopy observation of heteroduplexes [1]. Three of them have been or are being sequenced, namely: (i) the rbcL gene coding for the large subunit of ribulose 1,5-bisphosphate carboxylase, which is interrupted by 9 introns of average size 460 bp [2,3]; (ii) the tufA gene, coding for the elongation factor EF-Tu [4], with two small introns approx. 100 bp each; (iii) the psbA gene, coding for a 32 kDa thylakoid membrane protein of photosystem II, which includes 4 introns with sizes ranging from 433 to 616 bp [5–7].

None of these introns fully obeys the GU...AG rule to which the termini of eukaryotic pre-mRNA introns conform. As shown [8], most of the other introns that do not follow this rule can be arranged into two families or 'classes', each of which is defined by the presence of distinctive sequence stretches and potential secondary structures. Class I includes mitochondrial-, chloroplast- and nuclear-encoded members [8–11], while members of class II remain confined so far to organelles.

Here, we present evidence that the 4 introns in the *Euglena* psbA gene belong to the second class of structured introns. The possible presence of two additional class II introns in an unidentified reading frame located next to the psbA gene is also discussed.

Abbreviations: Y, pyrimidine base; R, purine base

2. MATERIALS AND METHODS

Computer programs for primary sequence comparisons and secondary structure cataloguing have been described in [9]. The construction of secondary structure models from comparative data was carried out according to the strategy described in [12].

3. RESULTS AND DISCUSSION

The most distinctive features of class II introns have been indicated in fig.1 by aligning 6 mitochondrial and 4 chloroplast introns, 8 of which

Published by Elsevier Science Publishers B.V.
00145793/85/$3.30

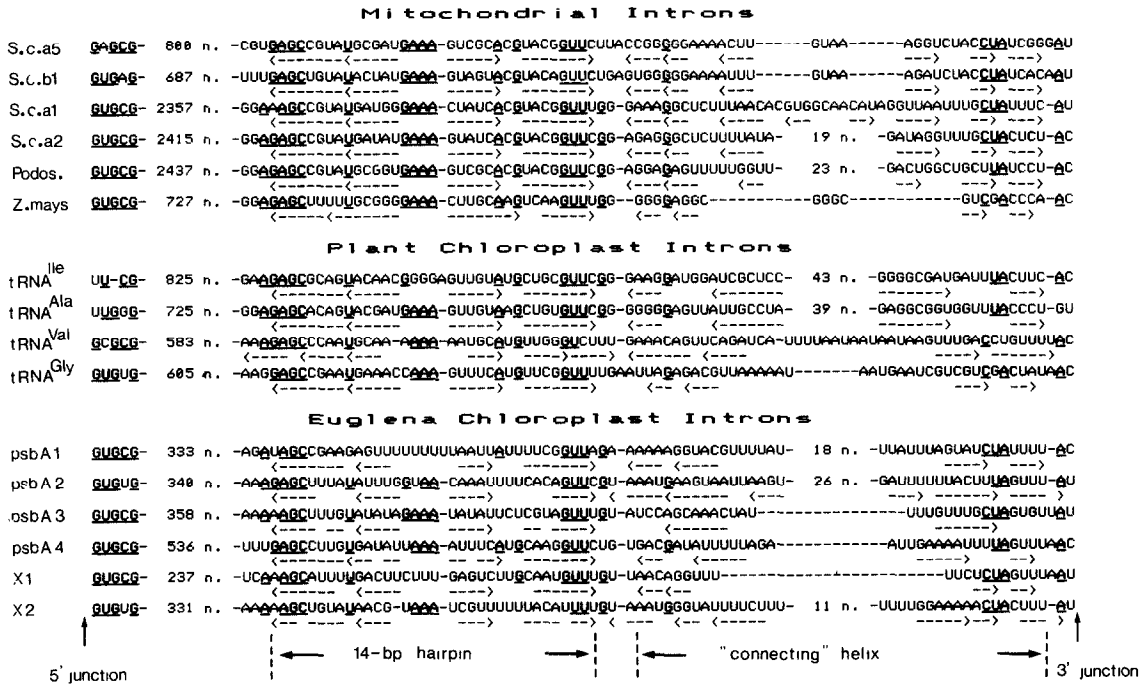

Fig.1. Conserved features in the 5′- and 3′-terminal segments of class II introns. Nucleotides underlined are common to at least 10 introns, including at least 2 in each subset. Divergent arrows show potential base-pairings. Sequences aligned are those of introns: a1, a2 and a5 o the cytochrome oxidase subunit I gene of *Saccharomyces cerevisiae* [21]; the first intron (b1) of the *S. cerevisiae* apocytochrome *b* gene [13]; an intron of the cytochrome oxidase subunit I gene of *Podospora anserina* [14]; the intron of the cytochrome oxidase subunit II gene of *Zea mays* [20]; introns in the $tRNA^{Ile}$ and $tRNA^{Ala}$ genes of *Zea mays* chloroplasts [16]; the introns in the $tRNA^{Val}$ and $tRNA^{Gly}$ genes of *Nicotiana tabacum* chloroplasts [18,22]; the four introns in the psbA gene and two putative introns in an unidentified gene (see text and fig.2) of *Euglena gracilis* chloroplast DNA [6]. The two vertical arrows indicate the 5′- and 3′-extremities of the introns.

had already been reported to belong to class II. These features include GUGCG and YUAYY.Y (.)AY consensus sequences at their 5′- and 3′-extremities respectively, and a characteristic 14-bp potential hairpin with a .G bulge on its 3′-side (see the hairpin structure with the bulge on fig.3) located within the last 100 residues of the 3′-terminal part of the intron [8]. This 14-bp hairpin, whose consensus sequence was reported in fig.1 of [8] to be GAGC...RUR..R.gaaa.U..YAygY... GUUY (unpaired nucleotides in lower case), can always be connected to the 3′ intron-exon junction by means of another base-pairing scheme, whose physical existence, like that of the major hairpin, is supported by the presence of many compensatory nucleotide changes in closely related sequences (see fig.1 of [8]).

As shown now in the lower part of fig.1 of this report, all 4 psbA introns fit fairly well into this picture. The extremities of introns of the rbcL and psbA genes have already been reported to follow a consensus [3,6,7] and this consensus can now be seen to agree with that previously derived for class II introns. As for the major hairpin, it is clearly recognizable, especially in intron psbA4, where its structure is fully orthodox and its sequence closely related to that of the corresponding element in other class II introns (cf. the $tRNA^{Gly}$ intron for instance).

rom the alignment in fig.1, a set of 'rules' can be derived that any class II intron is likely to obey. Provided such rules are sufficiently distinctive, they could allow identification of new class II introns in primary sequence data. When this ap-

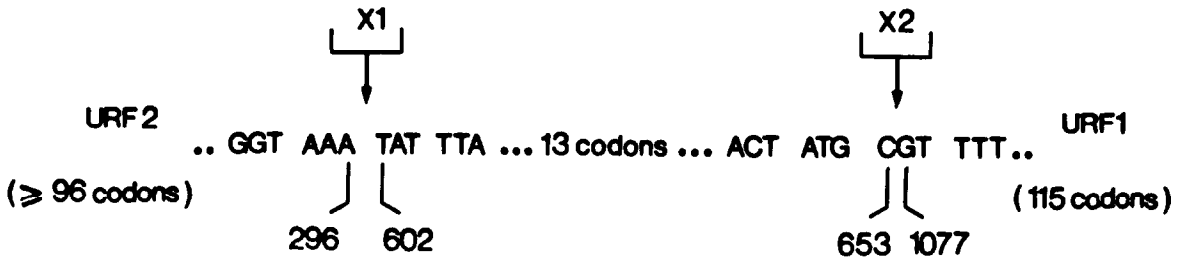

Fig.2. Insertion sites of putative introns X1 and X2 in *Euglena gracilis* chloroplast DNA. Coordinates are those of fig.2 in [6].

proach was used on the 1600-nucleotide sequence upstream of the psbA gene (reported in [6]), two sequence stretches emerged as good candidates for 3′-terminal sections of class II introns (X1 and X2 in fig.1; as a control, a search of organelle sequences in the EMBL Nucleotide Sequence Data Library, Release 3, only yielded the terminal segments of known class II members). Remarkably, when these putative X1 and X2 introns are removed from the sequence published in [6], using the first upstream GTGYG oligonucleotide as a possible 5′ exon–intron juncton, the URF1 and URF2 unidentified reading frames reported by these authors then merge into a single, much larger URF, as shown in fig.2.

The potential base-pairings shown in fig.1 are part of the complete secondary structure model of class II introns. This model, whose original version is in [9], has since been extended to all other introns recognized as belonging to class II ([8,13,14], unpublished). Its core consists of 6 base-paired stems, that bring the intron–exon junctions into relatively close proximity. The length of the unpaired stretches connecting successive stems is constant and these stretches have rather well-conserved primary sequences. As seen in fig.3 (cf. fig.2 of [8]), the sequence of intron psbA4 lends itself to the building of a model which essentially meets these requirements. Primary sequence homologies are especially striking with the first intron of the cytochrome *b* gene of yeast, whose corrected sequence and complete secondary structure model are found in [13]. It must be stressed however, that we were unable to fold the sequences of the 3 other psbA introns or the putative introns X1 and X2 into any convincing core structure. These 5 introns stand apart from the rest of class II by several criteria: (i) they are shorter than all other class II members; (ii) they seem to evolve much more rapidly. Indeed, even though they derive from the same genome, they are hardly more related to one another than to members of the other subsets. This is to be contrasted with the situation in the other class II subsets, which offer ample evidence of evolutionary conservatism: the $tRNA^{Ile}$, $tRNA^{Ala}$ and $tRNA^{Val}$ introns each have very closely related sequences in tobacco, a dicot, and maize, a monocot [16–19]; and the intron in the cytochrome oxidase subunit II gene of *Zea mays* [20], which has virtually the same core as fungal mitochondrial introns [8], also shares several peripheral structures and sequences with them (unpublished); (iii) even in the well-conserved 3′-terminal section, base-pairings tend to be weaker than in other class II members. Thus, the informational content of most *Euglena* introns is probably much lower than that of the other class II members, as could have been suspected from their extremely high (A + T) content.

The strong conservation of both primary sequences and secondary structures in each of the two classes of structured introns is generally thought to reflect the fact that these introns are actively involved in their own excision. Evidence in favor of this is 2-fold: (i) the 'self-splicing' *Tetrahymena* intron is a class I member [9,10]; (ii) even though most of them are known to require diffusible factors for their excision, yeast mitochondrial introns from both classes are the source of many cis-dominant splicing mutations. Remarkably, the mutations located in members of class I affect precisely those segments and structures that these introns have in common with their *Tetrahymena* relative (review [15]). This strongly suggests that

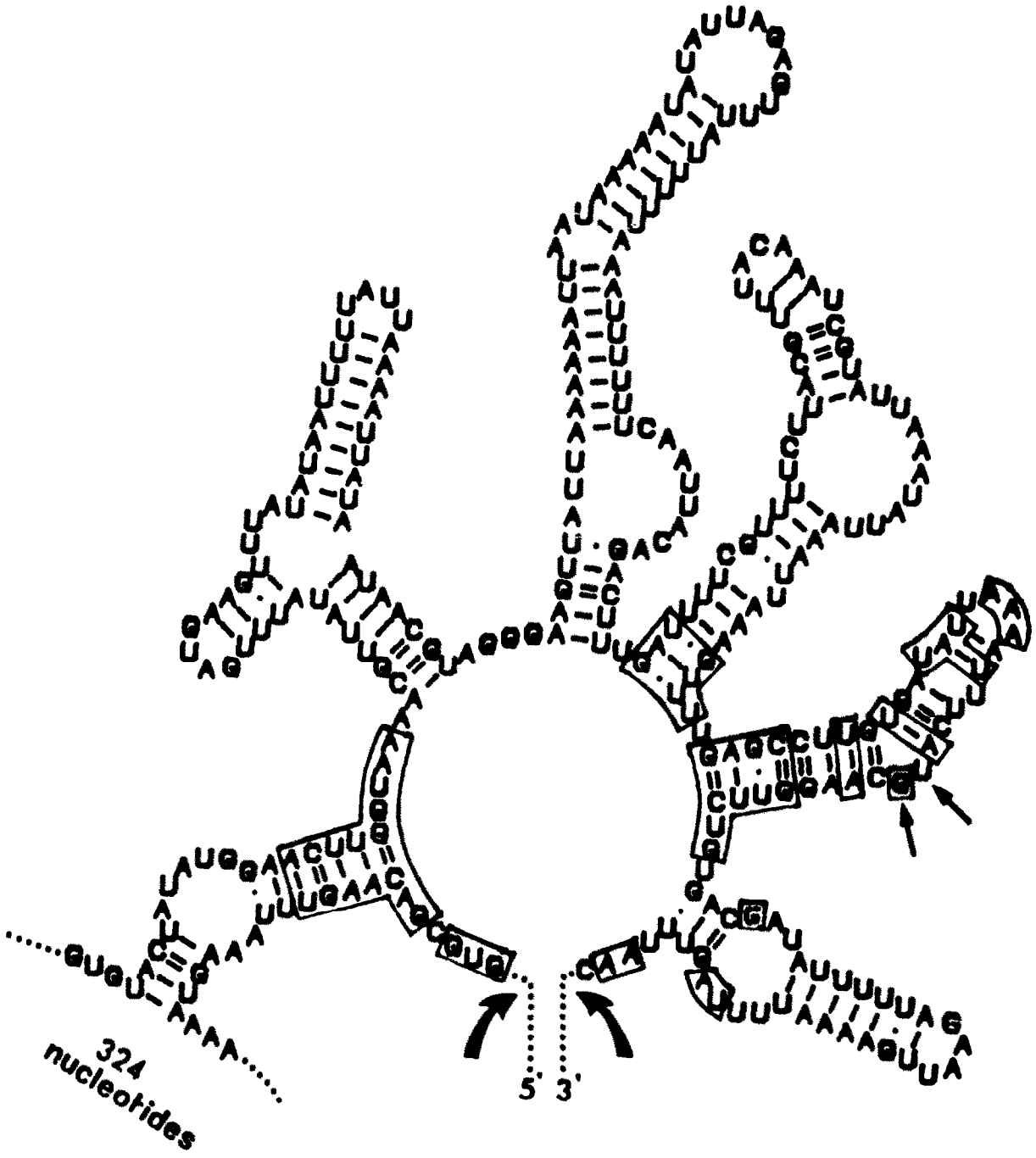

Fig.3. A tentative secondary structure model of intron psbA4. Curved arrows point to presumed intron–exon junctions. Boxed nucleotides are in common with the first intron in the apocytochrome *b* gene of *S. cerevisiae* (see [13] for a secondary structure model of the latter intron). The two small arrows indicate the .G bulge on the 3′-side of the 14-bp hairpin.

the mechanics of excision and splicing are basically the same for all class I introns, and by analogy, that there might be self-splicing introns in class II as well. However, self-splicing entities are unlikely to be found among *Euglena* introns: assuming the loss of information apparently suffered by these introns is real, they may best be regarded as well-advanced intermediates in a process of information transfer which, starting from a self-splicing intron, would end with one that merely specifies the cuts and ligations to be performed by an externally encoded splicing machinery.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Professors E. Stutz and J.H. Weil for encouragement and continued interest in our project and to Dr A. Steinmetz for critical reading of the manuscript.

REFERENCES

[1] Koller, B. and Delius, H. (1984) Cell 36, 613-622.
[2] Stiegler, G.L., Mathews, H.M., Bingham, S.E. and Hallick, R.B. (1982) Nucleic Acids Res. 10, 3427-3444.
[3] Koller, B., Gingrich, J., Farley, M., Delius, H. and Hallick, R.B. (1984) Cell 36, 345-353.
[4] Montandon, P.E. and Stutz, E. (1983) Nucleic Acids Res. 11, 5877-5892.
[5] Hollingworth, M.J., Johanningen, V., Karabin, G.D., Stiegler, G.L. and Hallick, R.B. (1984) Nucleic Acids Res. 12, 2001-2017.
[6] Keller, M. and Stutz, E. (1984) FEBS Lett., in press.
[7] Karabin, G.D., Farley, M. and Hallick, R.B. (1984) Nucleic Acids Res. 12, 5801-5812.
[8] Michel, F. and Dujon, B. (1983) EMBO J. 2, 33-38.
[9] Michel, F., Jacquier, A. and Dujon, B. (1982) Biochimie 64, 867-881.
[10] Waring, R.B., Scazzochio, C., Brown, T.A. and Dawies, R.W. (1983) J. Mol. Biol. 167, 595-605.
[11] Bonnard, G., Michel, F., Weil, J.-H. and Steinmetz, A. (1984) Mol. Gen. Genet. 194, 330-336.
[12] Noller, H.F., Kop, J., Wheaton, V., Brosius, J., Gutell, R.R., Kopylov, A.M., Dohme, F. and Herr, W. (1981) Nucleic Acids Res. 22, 6167-6189.
[13] Schmelzer, C., Schmidt, C., May, K. and Schweyen, R.J. (1983) EMBO J. 2, 2047-2052.
[14] Osiewacz, H.D. and Esser, K. (1984) Curr. Genet. 8, 299-305.
[15] Cech, T.R. (1983) Cell 34, 713-716.
[16] Koch, W., Edwards, K. and Kössel, H. (1981) Cell 25, 203-213.
[17] Takaiwa, F. and Sugiura, M. (1982) Nucleic Acids Res. 10, 2665-2676.
[18] Deno, H., Kato A., Shinozaki, K. and Sugiura, M. (1982) Nucleic Acids Res. 10, 7511-7520.
[19] Krebbers, E., Steinmetz, A. and Bogorad, L. (1984) Plant Mol. Biol. 3, 13-20.
[20] Fox, T.D. and Leaver, C.J. (1981) Cell 26, 315-323.
[21] Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E. and Tzagoloff, A. (1980) J. Biol. Chem. 255, 11927-11941.
[22] Deno, H. and Sugiura, M. (1982) Proc. Natl. Acad. Sci. USA 9, 405-408.