J Mol Evol (1997) 44:33–42

JOURNAL OF MOLECULAR EVOLUTION

The Reverse-Transcriptase-Like Proteins Encoded by Group II Introns in the Mitochondrial Genome of the Brown Alga *Pylaiella littoralis* Belong to Two Different Lineages Which Apparently Coevolved with the Group II Ribosyme Lineages

Jean-Marc Fontaine, Didier Goux, Bernard Kloareg, Susan Loiseaux-de Goër

Centre d'Etudes d'Océanographie et de Biologie Marine, Station Biologique de Roscoff, CNRS (UPR 9042), UPMC, BP 74, 29682 Roscoff Cedex, France

Received: 21 May 1996 / Accepted: 4 July 1996

Abstract. The mitochondrial genome of the brown alga *Pylaiella littoralis* contains two different types of group II introns. They each encode complete complex proteins, i.e., with a reverse transcriptase domain, a maturase or X domain, and an endonuclease or H-N-H/ zinc finger domain. To our knowledge, this is the first example of the presence in the same genome of introns belonging to subgroups IIA and IIB which both contain multidomained RT-like proteins. We describe the group IIA introns that interrupt the *cox1* gene. The RT-like proteins contained in these introns were compared to those of the LSU rDNA group IIB introns. The phylogenetic relationships of these intron ORFs were investigated and the possible evolution of group II introns is discussed.

Key words: Brown alga — *cox1* gene — Group IIA and IIB introns — Phylogeny — Reverse transcriptase-like genes

Introduction

Group I and group II introns are RNA enzymes (ribozymes) that catalyze their own splicing by different mechanisms. All group II introns show a characteristic secondary structure, which consists of six double helical domains, numbered from I to VI, radiating from a central wheel (see Fig. 2). The group II introns are classified into two different subgroups, IIA and IIB, and distinguished by specific structural features (Michel et al. 1989). Group II introns often encode open reading frames (ORFs) which consist of several domains, including a reverse-transcriptase (RT) domain related to the RT domains of the ORFs encoded in non-long-terminal repeat (LTR) transposable elements, other cellular transposable elements, retroviruses, and various DNA viruses (Toh et al. 1983). It is not clear how all these complex proteins are related and which are the most ancient (for review see McClure 1991).

The origin of group II introns is unknown and their idiosyncratic distribution seems to reflect efficient mechanisms for dispersal which may have led to lateral transfers from one organism to another. In fact they have been shown to be mobile genetic elements (Yang et al. 1996). The recent discovery of group II introns in cyanobacteria and proteobacteria (Ferat and Michel 1993; Ferat et al. 1994) has raised the hypothesis of a vertical transmission of these introns from prokaryotes to the derived eukaryotic organelles, mitochondria, and plastids.

While studying the mitochondrial genome of the brown alga *Pylaiella littoralis* (L.) Kjellm., we found four group IIB introns in the large-subunit (LSU) rDNA. Two of them contain a reverse-transcriptase-like gene in their region IV (Fontaine et al. 1995a). This was a rather

Correspondence to: S. Loiseaux-de Goër

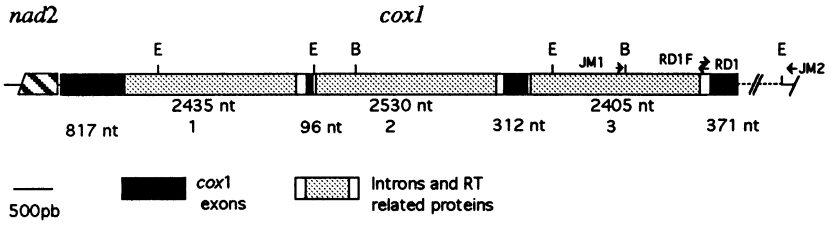

Fig. 1. Physical map of the mitochondrial *cox1* gene from the brown alga *P. littoralis.* The *cox1* gene is split into four exons by three group IIA introns, numbered *1–3*. Each of the three introns contains an ORF encoding for an RT-related complex protein. Fragments between *arrows* were amplified by polymerase chain reactions as indicated in Materials and Methods. *E, Eco*RI; *B, Bam*HI.

surprising finding as most group IIB introns have been described in eubacteria (cyanobacteria and γ proteobacteria; Ferat and Michel 1993) and in plastids (Kück et al. 1990; Maier et al. 1995). A few group IIB introns have been described from mitochondrial genomes but these do not contain ORFs, and sometimes some of the helical domains are missing (Kück et al. 1990; Bonitz et al. 1980; Leblanc et al. 1995).

In this paper we describe the *cox1* gene from the mitochondrial genome of *P. littoralis* as well as the three group II introns which interrupt this gene. Based upon their secondary structure they all belong to subgroup IIA and contain RT-related ORFs. The presence in the same mitochondrial genome of both group IIA and IIB introns with RT-like proteins provides the opportunity to compare the domain organization of these ORFs and to investigate their evolutionary history. In contrast with the ORFs found in group I introns, which often are unrelated (see Michel and Ferat 1995), the group II ORFs have a common ancestor and seem to have coevolved with their introns.

Materials and Methods

The *Bam*HI and *Eco*RI restriction fragments containing part of the *cox1* gene came from an organellar gene bank (Loiseaux-de Göer et al. 1988). Missing parts (the 3′ end, see Fig. 1) were amplified from organellar DNA extracted from axenic cultures of *P. littoralis* as in Dalmon and Loiseaux (1981). In a first amplification an intron-specific degenerated intermediate primer, RD1, nearly identical to the RID-1 primer designed by Ferat et Michel (1993), was designed from the conserved domain V (RD1: 5′ ACCGAACCGTACGTG(AC)(AGCT)A(GC)TTTC(AC)C 3′). The other primer (JM1: 5′ GAAAAGGGCACTCGTCGGAAGGCAA 3′) consisted of a known sequence from a cloned border *Bam*HI fragment. A second PCR was then performed with the reverse complement of the RD1 degenerate primer (RD1F: 5′ G(GT)GAAA(GC)T(AGCT)(GT)CACGTACGGTTCGGT 3′) and with a primer derived from the known sequence of a cloned *Eco*RI fragment (JM2: 5′ GCATCTACCGCAATTATAGGACTCA 3′). Cycle times were as follows: denaturation for 5 min at 94°C, followed by 40 thermal cycles (denaturation, 1 min at 94°C; annealing, 2 min at 68°C; elongation, 3 min at 72°C) and a last elongation step of 10 min at 72°C. For the second PCR, the hybridization temperature was 40°C and the duration of elongation was 5 min. The 100μl reaction mixture contained 2.5 units of Taq DNA polymerase (Promega), 0.2 μM of each primer, 2 mM each of dATP, dCTP, dGTP, and dTTP, 10 μl of 10× buffer (Promega), 10 mM of $MgCl_2$, and two drops of mineral oil (Perkin).

The entire amplified fragments were sequenced in both directions, using a dideoxy sequencing kit (Pharmacia) and synthetic oligonucleotide primers (Eurogentec). The sequence of the intron domain V was then verified by amplification and direct sequencing of this region. Nucleotide sequences were analyzed by the McMolly program (Karoi-Kerlag Reiner Bornemann Bielefeld) to search for open reading frames. Secondary structures of group II introns were constructed according to the consensus models of Michel et al. (1989). The amino acid sequences were aligned by eye with those of homologous genes found in group IIA and IIB introns. Alignments were used to generate distance and parsimony phylogenetic trees, as described in Fontaine et al. (1995a).

Results

The cox1 *Gene of* Pylaiella littoralis

The mitochondrial *cox1* gene of the brown alga *P. littoralis* spans over 8,956 nt. It is interrupted by three introns (Fig. 1). The exons code for a very conserved, 528-amino-acid protein. Using the universal genetic code all conserved tryptophans are correctly translated, as for *cox3* and other genes from this genome (Fontaine et al. 1995b). The tryptophan codon is generally the first to be modified in mitochondria when modifications do occur in the genetic code; it is the case, for example, of the protozoan *Acanthamoeba castellanii* (Burger et al. 1995) and of the red alga *Chondrus crispus* (Leblanc et al. 1995). The only particularity of the *P. littoralis cox1* gene is the presence of a long NH_2-terminal region (Fig. 2), 14 amino acids longer than that of *A. castellanii*, which already is longer than most of the other homologous genes. As can be seen in Fig. 2, most mutations in the *cox1* genes are silent or result in conservative substitutions of amino acids, with an overall identity in eukaryote genes of 74–82% (Table 1) and similarity ranging from 92.5 to 97.7%. Insufficient significant lineage signatures are present to build reliable phylogenies from these genes. Nevertheless, as in the case of the LSU rRNA gene (Fontaine et al. 1995a), it appears that, compared to other mitochondrial *cox1* genes, the *cox1* gene of *P. littoralis* is slightly more closely related to the homologous α-proteobacterial gene from *Paracoccus denitrificans* (Raitio et al. 1987; Table 1). The COXI amino acid sequences most closely related to that of *P. littoralis* are those from land plants, green and red algae, and from *Acanthamoeba castellanii*. Surprisingly the COXI amino acid sequence of the oomycete *Phytophtora megasperma* (Sachay et al. 1993) is more distant even

```
                                                                                M.pol 2
                                     M.pol 1                          P.ans 1   S.cer 1
                                        *                                 *        *
S.cer.                   MV--QRWLYSTNAKDI AVLYFMLAIFSGMAGTAMSLIIRLELAAPGSQYLH GNSQLFN VLVVG
M.pol.             MNNFA-----QRWLFSTNHKDM GTLYLIFGAIAGVMGTCFSVLIRMELAQPGNQILG GNHQLYN VLMTA
P.wic.                   MVT--RWLYSTNHKDI GTMYLIFGAFSGVLGTVFSLLIRMELAQPGNQILN GNHQLYN VIITA
A.cas.        MINRLLNNLTSFFTDNRWLFSTNHKDI GTLYLIFGGFSGIIGTIFSMIIRLELAAPGSQILS GNSQLYN VIITA
C.cri.        ......QSFFTQWIS-RWIFSTNHKDI GTLYLIFGAFSGVLGGCMSMLIRMELAQPSNHLLL GNHQIYN VLITA
P.meg.           MNFKNINKCQL-D--WLFSTNHKDI GTLYLIFSAFAGIVGTTLSLLIRMELAQPGNQIFM GNHQLYN VVVTA
P.lit.  MSYSINNIYNKDLVSFFSFSITSLSDFCS-RWLFSTNHKDI GTLYLIFGGFSGVLGTAMSVLIRLQLASPGNQFLG GNHQLYN VIVTA

         K.lac 1                       A.macr 3
     N.cra 1   S.cer 2                 P.ans IA
           *  *  *                        *
S.cer HAVLM IF F LVMPALIGGFGNYLLPLMIGATDTAFPRINNIAFW VLPMGLVCLVTSTLVESGAGTGWTVYPPLSSIQAHSGPSVDLAIFA
M.pol HAFLM IF F MVMPAMMGGFGNWFVPILMGSPDMAFPRLNNISFW LLPPSLLLLLSSALVEVGCGSGWTVYPPLSGMTSHSGGSVDLAIFS
P.wic HAFLM IF F MLMPALMGGFGNWFLPILIGAPDMAFPRLNNISFW LLPPSLLLLVSSALVEVGAGTGWTVYPPLASIASHSGGSVDLAIFS
A.cas HAFVM IF F FVMPVMIGGFGNWFVPLMIGAPDMAFPRLNNISFW LLPPSLFLLLCSSLVEFGAGTGWTVYPPLSSIVAHSGGSVDLAIFS
C.cri HAFLM IF F MVMPVMIGGFGNWLVPIMIGSPDMAFPRLNNISFW LLPPSLCLLLMSALVEVGVGTGWTVYPPLSSIQSHSGGAVDLAIFS
P.meg HAFIM VF F LVMPALIGGFGNWFVPLMIGAPDMAFPRMNNISFW LLPPALLLLVSSAIVESGAGTGWTVYPPLSSVQAHSGPSVDLAIFS
P.lit HAFLM IF F MVMPVLIGGFGNWFVPLMIGAPDMAFPRMNNISFW LLPPSLILLLASSLVESGAGTGWTVYPPLSGIQAHSGPSVDLAIFS

                 M.pol 5
                    *
S.cer.  LHLTSISSLLGAINFIV TTLNMRTNGMTMHKLPLFVWSIFITAFLLLLSLPVLSAGITMLLLDRNFNTSFFEVAGGGDPILYEHLFWFFGH
M.pol.  LHLSGVSSILGSINFMT TIFNMRAPGLTMHRLPLFVWSVLVTAFLLLLSLPVLAGAITMLLTDRNFNTTFFDPAGGGDPILYQHLFWFFGH
P.wic.  LHLAGVSSILGAINFIC TVFNMRAPGMSMLDL-LFVWAVFITAWLLLLCLPVLAGGITMLLTDRNFNTSFFDPAGGGDPILYQHLFWFFGH
A.cas.  LHLAGISSLLGAINFIT TIFNMRVPGLSMHKLPLFVWSVLITAFLLLFSLPVLAGAITMLLTDRNFNTSFFDPSGGGDPILYQHLFWFFGH
C.cri.  LHISGASSILGAVNFIS TILNMRSPGQSMYRIPLFVWSILVTAFLLLLAVPVLAGAITMLLTDRNFNTSFFDASGGGDPILYQHLFWFFGH
P.meg.  LHLTGISSLLGAINFIS TIYNMRAPGLSFHRLPLFVWSVLITAFLLLLTLPVLAGAITMLLTDRNLNTSFYDPSGGGDPVLYQHLFWFFGH
P.lit.  LHLSGAASILGAINFIT TIFNMRAPGMTMDRLPLFVWSVLITAFLLLLSLPVLAGGITMLLTDRNFNTTFFDPAGGGDPVLYQHLFWFFGH

          P.lit 1            P.lit 2
            *                  *
S.cer.   PEVYI LIIPGFGIISHVVSTYSKKPVFGEISMVY AMASIGLLGFLVWSHHMYIVGLDADTRAYFTSATMIIAIPTGIKIFSWLATIYGG
M.pol.   PEVYI LILPGFGIISHIVSTFSRKPVFGYLGMVY AMISIGVLGFIVWAHHMFTVGLDVDTRAYFTAATMIMAVPTGMKIFSWIATMWGG
P.wic.   PEVYI LIIPGFGIISHVIATFSKKPIFGYLGMVY AMCSIGILGFIVWAHHMYVVGLDIDTRAYFTAATMIIAVPTGIKIFSWVATMWGG
A.cas.   PEVYI LILPAFGIVSQIIGTFSNKSIFGYIGMVY AMLSIAVLGFIVWAHHMYTVGLDVDTRAYFTAATMMIAVPTGIKIFSWIATLWGG
C.cri.   PEVYI LILPGFGMISHIVSTFSRKPVFGYIGMVY AMVSIGVLGFIVWAHHMYTVGLDVDTRAYFTAATMIIAVPTGIKIFSWIATIWEG
P.meg.   PEVYI LILPAFGIISQVAAAFAKKNVFGYLGMVY AMLSIGLLGCIVWAHHMFTVGLDVDTRAYFSAATMIIAVPTGIKIFSWLATLWGG
P.lit.   PEVYI LILPGFGIVSHILSTFARKPVFGYLGMVY AMLSIGILGFIVWAHHMFTVGLDIDTRAYFTAATMIIAVPTGIKIFSWIATLWGG

                                       S.cer 5*  P.lit 3
                                              *  *
S.cer.   SIRLATPMLYAIAFLFLFTMGGLTGVALANASLDVAFHDTYYVVGHF HY VLSMGAIFSLFAGYYYWSPQILGLNYNEKLAQIQFWLIFI
M.pol.   SMQYKTPMLFAVGFMFLFTVGGLTGMVLANSGVDIALHDTYYVVAHF HY VLSMGAVFALFAGFYYWMGKMTGLQYPETLGQIHFWITFF
P.wic.   SIELRTPMLFAVGFLFLFTVGGLTGVVLANSGLDVAFHDTYYVVAHF HY VLSMGAVFALFSGFYYWIGKITGLQYPETLGQIHFWLMFL
A.cas.   QIVRKTPLLFVIGFLILFTLGGLTGIVLSNAGLDIMLHDTYYVVAHF HY VLSMGAVFAFFAGFYYWFWKISGYTYNEMYGNVHFWLMFI
C.cri.   SIHLKTPMLFAIGFIFLFTIGGLTGIVLANSGLDISLHDTYYVVAHF HY VLSMGAVFAIFAGFYYWFGKITGLQYPETLGQIHFWSTFI
P.meg.   SLKFETPLLFVLGFILLFVVGGVTGVAMSNSGLDIAIHDTYYIVGHF HY VLSMGAVFGIFTGFYFWIGKISGRKYPEILGQIHFWLFFI
P.lit.   SIRLKTPMYFPIGISFLFTIGGLTGVVLANSGVDIALHDTYYVVAHF HY VLSMGAAFTMFAAFYFWIGKMTGLAYPEVLGQIHFWLMFI


S.cer.   GANVIFFPMHFLGINGMPRRIPDYPDAFAGWNYVASIGSFIATLSLFLFIYILYDQ-LV--NNKS---VIYAKAP-DFVESNTIFNLNTVK
M.pol.   GVNLTFFPMHFLGLAGMPRRIPDYPDAYAGWNAFSSFGSYVSVVGIFCFFVVVF-LTLTS-ENKCAPS-PWAT--LEWM----VPSPPAFH
P.wic.   GVNITFFPMHFLGLAGMPRRIPDYPDCYAGWNAVASYGSYLSITAVLFFFYVVYK-TLTS-NEVCPRN-PWET--LEWM----LPSPPAFH
A.cas.   GVNLTFFPMHFVGLAGMPRRIPDYPDNYYYWNILSSFGSIISSVSVIVFFYLIY-LAF--NNNNTPKIKLVHSIFAPYI--NTLLSKN-LL
C.cri.   GVNLTFMPMHFLGLAGMPRRIPDYPDAYAGWNLIASYGSYIALFSTLFFFYIVF-VSLTS--NNPCTNFPWET--LEWI----VTSPPAYH
P.meg.   GVNITFFPMHFLGLAGMPRRIPDFPDAMSGWNAVSSFGSYISFFSALFFFYIVY-VTLVYG-KK-TEN
P.lit.   GVNLTFFPMHFLGLAGMPRRIPDYPDSYAGWNGLASLGSIMSSLASLFFFFVVY-ITLTKGVEE--AN-PWVKG-------RGLPSPPLPR

S.cer.   SSSIEFLLTSPPAVHSFNTPAVQS
M.pol.   --TFEELPAIKESI
P.wic.   --TFEEIQV
A.cas.   --TFASIKSTSDSSFFKFSKFFIFFM
C.cri.   --TFEE
P.lit.   RS
```

Fig. 2. Alignment of several *cox1* genes. The amino acid sequence of the *cox1* gene from *P. littoralis* is aligned with those from the oomycete *Phytophthora megasperma, P. meg.* (Sachay et al. 1993); the red alga *Chondrus crispus, C. cri* (Leblanc et al. 1995); the protozoan *Acanthamoeba castellanii, A. cas* (Burger et al. 1995); the green alga *Prototheca wickerhamii, P. wic* (Wolff et al. 1994); the land plant *Marchantia polymorpha, M. pol* (Oda et al. 1992); and the yeast *Saccharomyces cerevisiae, S. cer* (Bonitz et al. 1980). The locations of all known group II introns which interrupt *cox1* genes are indicated. They all belong to subgroup IIA, with the exception of *S. cer.* 5, which belongs to subgroup IIB and has no ORF. *A. macr. Allomyces macrogynus* (Paquin and Lang 1996); *K. lac: Kluyveromyces lactis* (Hardy and Clark-Walker 1991); *P. ans: Podospora anserina* (Cumming et al. 1989); *N. cra: Neurospora crassa* (Burger et al. 1982).

Table 1. Percentage of identical amino acids in *cox1* genes (478 amino acids compared)

	M.p.	*P.w.*	*A.c.*	*C.c.*	*P.m.*	*P.l.*	*P.d.*
Marchantia polymorpha	100						
Prototheca wickerhamii	82.2	100					
Acanthamoeba castellanii	75.9	75.9	100				
Chondrus crispus	82.2	79.9	76.8	100			
Phytophthora megasperma	75.1	76.6	75.3	74.3	100		
Pylaiella littoralis	81.2	80.3	77.8	79.3	76.6	100	
Paracoccus denitrificans	60.0	61.1	58.2	58.8	60.5	61.7	100

though oomycetes form a sister group to brown algae (Gunderson et al. 1987; Bhattacharya and Druel 1988). This latter protein, however, shows more random mutations and is equally distant from all of the other COXI sequences (Table 1).

Structural Organization of the Group IIA and IIB Intron ORFs

The three introns of the *cox1* gene are 2,435, 2,530, and 2,405 nt long, respectively. Introns were folded into a characteristic secondary structure (Fig. 3), with a central core from which six major domains radiate, referred to as I–VI (Michel and Dujon 1983). Based upon their secondary structures and tertiary interactions (Fig. 3), the three introns which interrupt the *cox1* gene belong to the subgroup IIA. They are delimited by the consensus GTGCG sequence at their 5′ end and by a consensus CYAC tetramer at the 3′ end. All of the three introns interrupt a codon specifying a conserved amino acid. In contrast with most known group II introns, which interrupt the *cox1* genes within the first 200 amino acids of the sequence, all three *P. littoralis* introns are located in the second half of the coding *cox1* sequence. This is also the case of the fifth *cox1* intron of *Saccharomyces cerevisiae*, which is one of the group IIB introns found in mitochondrial genomes (Fig. 2). On the whole, *cox1* genes contain more group I (see Wolff et al. 1993) than group II introns, but none of these interrupt the gene at the same location as the *P. littoralis* group II introns.

The three introns of *P. littoralis cox1* gene each contain an open reading frame coding for a multidomained protein, including one domain related to the reverse transcriptase regions of non–long-terminal-repeat transposable elements and of various viruses (Michel and Lang 1985; Doolittle et al. 1989; McClure 1991). In contrast with most group IIA introns, these open reading frames are not in frame with the 5′ exons, yet they span most of intron domain I. Interestingly, all three ORFs start in the first stem loop of domain I (see arrows in Fig. 3), referred to as subdomain IA by Michel et al. (1989).

An alignment of the ORFs of the *P. littoralis* mitochondrial group IIA and IIB introns with several homologous ORFs from other organisms is shown in Fig. 4. These ORFs fall into two different lineages, lineage *a* and lineage *b*. All of the ORFs encoded in characterized group IIA introns, including those of *P. littoralis*, belong to lineage *a* whereas all ORFs encoded in characterized group IIB introns fall into the other lineage, lineage *b*. The structural characteristics of these two different ORF lineages are delineated in Fig. 5. In the complex RT proteins from lineage *b*, the 5′ end region (noted A in Fig. 5), 50–100 amino acids in length and located in the intron region IV of the introns, is fairly conserved. In lineage *a*, this region, 180–280 aa long, which spans the intron regions I–III, is not readily alignable. Moreover, in contrast with the ORFs from lineage *b*, most of the RT amino acid sequences from lineage *a* are in continuity and in frame with the 5′ exons, implying a chimerically translated protein and subsequent processing events (Michel and Ferat 1995). In both ORF types this region is followed by approximately 50 rather well-conserved amino acids (referred to as region B in Fig. 5; the first 50 aa in Fig. 4). This clearly identifiable (conserved) region of the group II intron ORFs perhaps corresponds to the Z region of non-LTR transposable elements, which is said to be ''a unique but unidentifiable sequence'' (Doolittle et al. 1989). In the reverse transcriptase domain the main difference between the two intron sets is the addition of ca. 60 aa in the lineage *a* ORFs, between conserved regions RT4 and RT5. This additional segment is poorly conserved and difficult to align, with the exception of the ORFs in the three *cox1* introns of *P. littoralis*, the *cox1* intron 2 of *Marchantia polymorpha* (Oda et al. 1992) and the *cox1* intron 1 of *Neurospora crassa* (Burger et al. 1982). The RT part of the gene is followed by a very poorly conserved region (denoted C in Fig. 5), approximately 50 aa long in the *b* lineage and 75–300 aa long in the *a* lineage. The subsequent X domain is well conserved in all of the group II ORFs, albeit with an addition of 25–35 well-conserved amino acids in those from lineage *a*. This domain has been shown to have a maturase function, since the intron no longer splices when this region is mutated (Mohr et al. 1993; Moran et al. 1994, 1995). A very variable region follows, which is longer in the *b* than in the *a* lineage. Finally, the 3′ zinc finger-like/H-N-H domain (3′ end of Fig. 4) is fairly well conserved in both types of ORFs. It has been suggested that this region might participate in the endonuclease activity

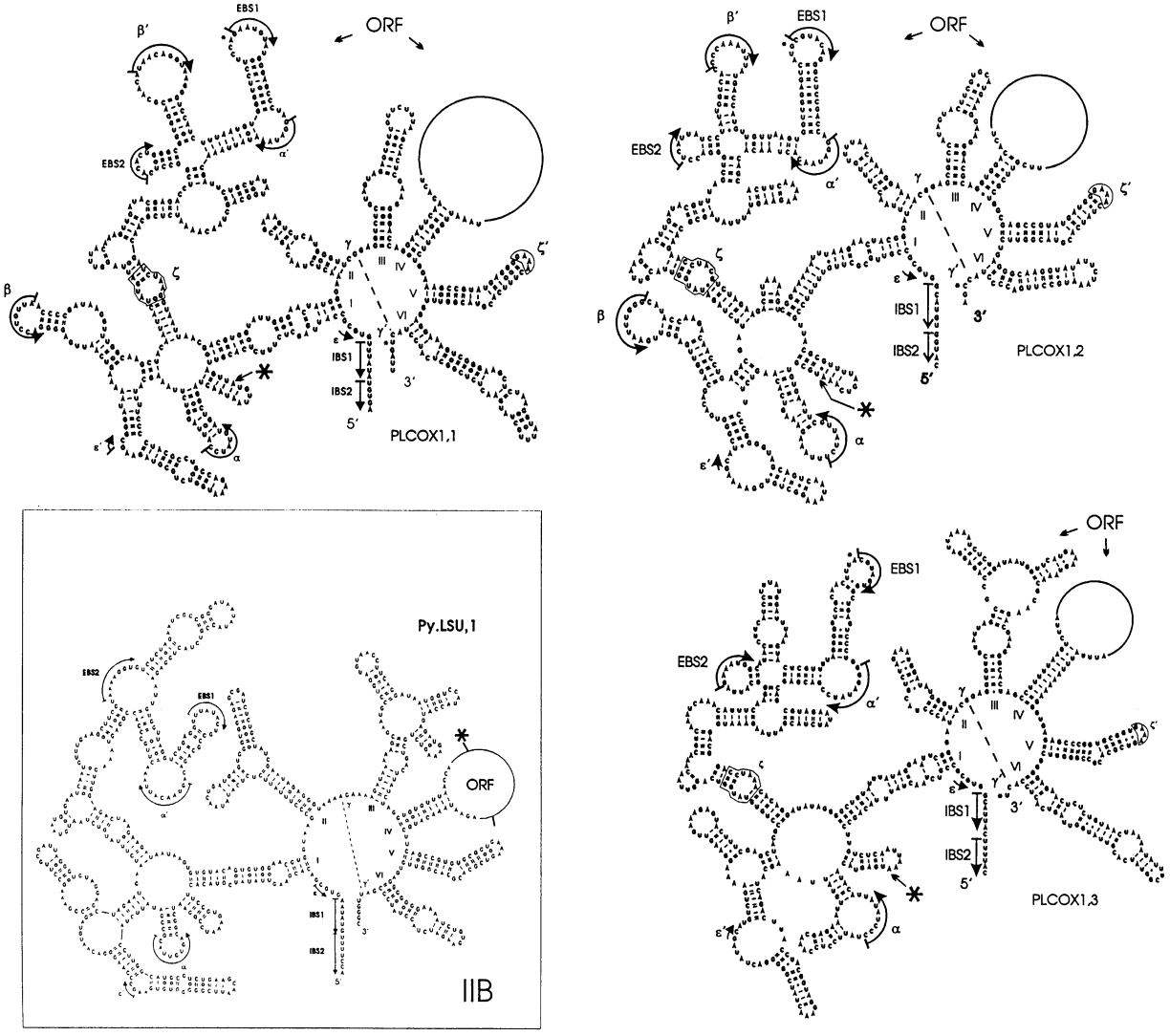

Fig. 3. The mitochondrial group IIA introns of *P. littoralis.* A group IIB intron from the mitochondrial LSU rRNA of *P. littoralis* (Fontaine et al. 1995a) is shown for the comparison. The tertiary base-pairing sequences are indicated by EBS1-EBS2, IBS1-IBS2, α-α′,β-β′ γ-γ′, ϵ-ϵ′, ζ-ζ′ (Michel and Jacquier 1987; Michel et al. 1989; Jacquier and Michel 1990; Costa and Michel 1995). *Arrows with a star* indicate the start of the ORF. The main differences between group IIA and IIB introns are as follows: a bulging ''A'' on the 3′ side of domain VI 7 nt (exceptionally 8 nt) upstream of 3′ intron-exon junction for the group IIA introns and 8 nt for group IIB introns. Group IIA introns usually end with YAY and group IIB with RAY. In group IIA the tertiary interaction ϵ′ is carried by an internal loop of 9/11 nt and of 4/5 nt for group IIB introns. Some exceptions do exist, as can be seen in the *P. littoralis cox1* intron 1. The tertiary interaction indicated by a *small star* is potentially present in all group IIA introns whereas it is found only in a minority of group IIB introns. In the group IIB introns the open reading frame is free-standing in domain IV, whereas the ORF in group IIA introns spans domains I–IV, and can even be in frame with the 5′ exons. For the other differences see Michel et al. (1989).

of group II introns (Zimmerly et al. 1995; Shub and Goodrich-Blair 1994; Gorbalenya 1994; Yang et al. 1996).

Phylogenetic Relationships of the Lineage a *and* b *Intron ORFs*

As exemplified in Fig. 6, various phylogenetic trees were constructed from the alignment of the group II intron complex RT proteins (complete alignment available upon request). Positions which were difficult to align were suppressed as well as the region D and the H-N-H domain, which is absent in some of the genes. Nearly identical results were obtained from distance and parsimony trees and whether 267 or 302 amino acid positions were taken into consideration. Two main branches are clearly delineated, separated by a bootstrap value of 100: all characterized group IIA intron ORFs group together in lineage *a;* all characterized group IIB intron ORFs group together in lineage *b.* In each lineage some of the ORFs are inserted in introns which have not been characterized, either because their sequence is incomplete (*S. obliquus*) or their secondary structure is unknown (*M. polymorpha*) or because they are too modified to be classified in group II subgroups (*Pyrenomonas salina cpn60* ORF).

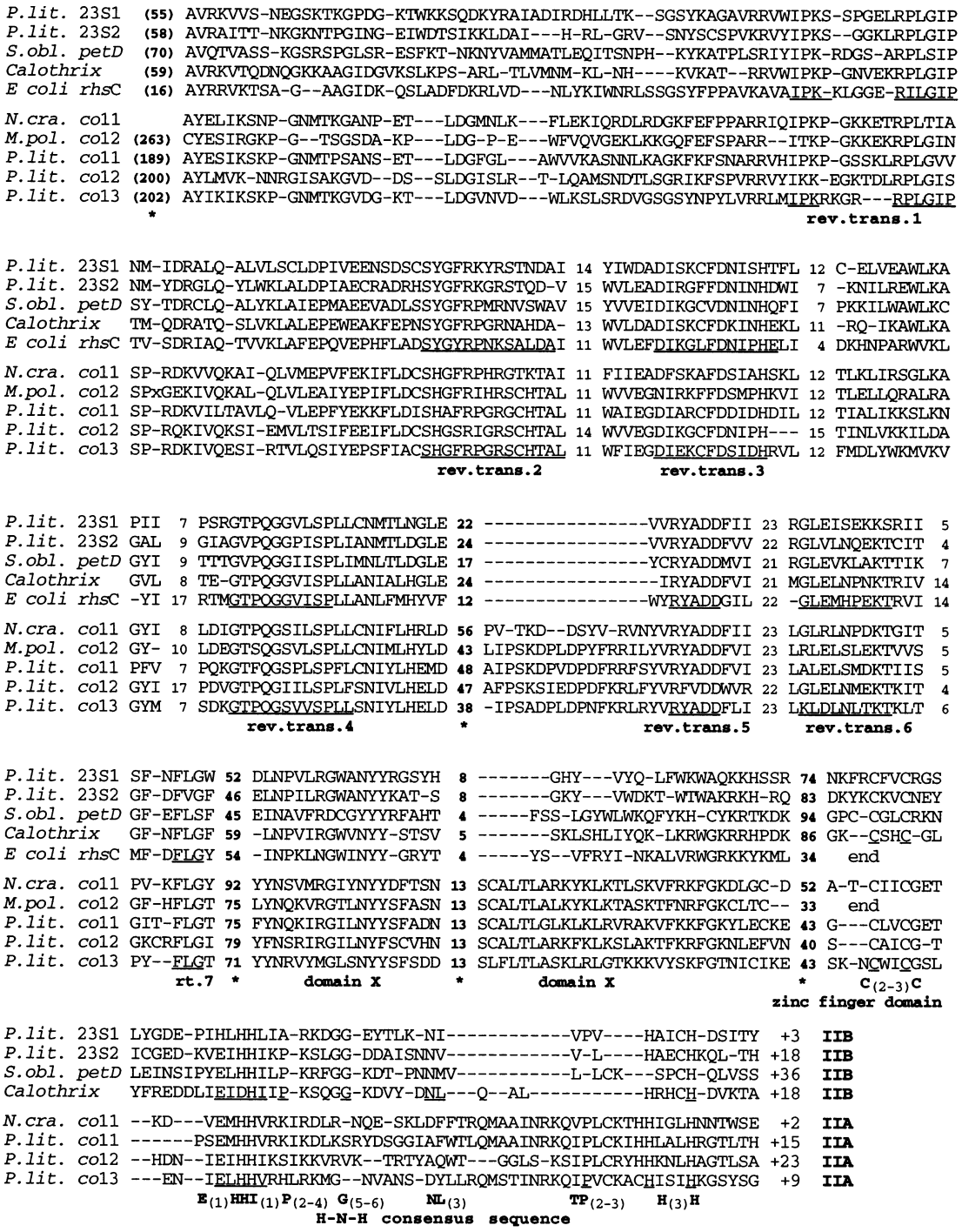

Fig. 4. Alignment of the deduced complex RT-like proteins encoded by the group IIA and IIB introns of *P. littoralis* with other homologous genes. Only the most closely related ORFs (see tree, Fig. 6) are shown. These ORFs fall into two groups, one group containing RT-like proteins encoded in group IIB introns and the other including the RT-like proteins encoded in group IIA introns. Regions which could not be aligned are indicated by the number of amino acids which have not been figured. *Stars* (*) refer to regions where a large addition (or deletion) exists in only one of the groups. These complex proteins contain a reverse transcriptase domain, a maturase domain (Lambowitz and Belfort), and an endonuclease domain (zinc finger or H-N-H, Shub and Goodrich-Blair 1994), shown under the alignment. The species included in this figure are *P. littoralis* (this work), *M. polymorpha* (Oda et al. 1992), *N. crassa* (Burger et al. 1982), *S. obliquus* (Kück 1989), *Calothrix* PCC7601 (Ferat and Michel 1993), and *E. coli* (Ferat et al. 1994). *23S1*, *23S2* refer to the first and second intron of the 23S rDNA; *co 1 = cox1* gene. The intron of *Calothrix* sp. is inserted in an unidentified gene (Ferat and Michel 1993).

In the *b* lineage, the intronic ORF 439 from the cyanobacterium *Anabaena* sp. PCC 7120 (U13767) seems to be inserted into a recognizable group II intron. Its relatedness to that of the plastidial *cpn60* gene of the cryptophyte *P. salina* (Maier et al. 1995) is probably artefactual even though it is found in all trees and with good bootstrap values (see Fig. 6). Alignments show that these two sequences are more heavily and randomly mutated

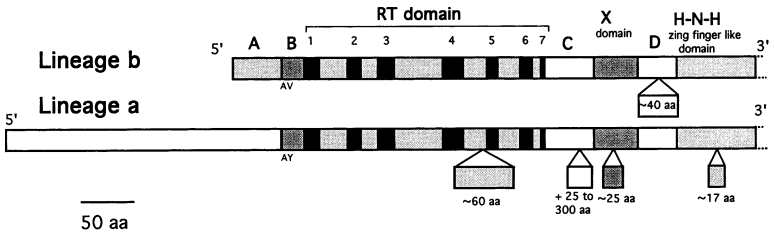

Fig. 5. Schematic representation and comparison of the multidomained proteins encoded in group II introns. The lineage *b* ORFs are free-standing in region IV. The lineage *a* ORFs often are in phase with the 5′ exon or begin in region I of the intron. They all end in the 3′ end of region IV. Regions are *shaded from white to black* to indicate increasing degrees of conservation. Domains are referred to as in Fig. 4. All of the ORFs found in IIB introns which have been characterized by their secondary structure belong to the *b* lineage whereas those found in characterised IIA introns belong to the *a* lineage.

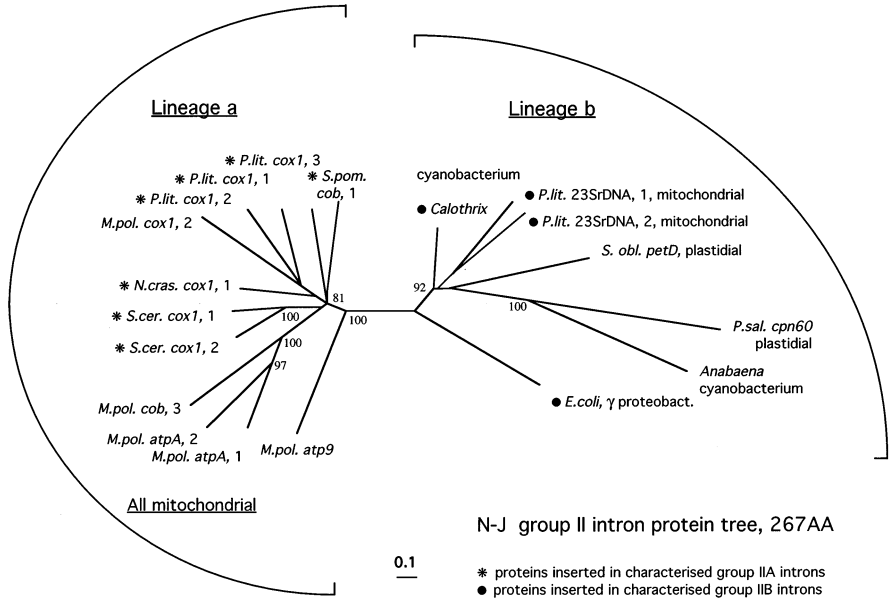

Fig. 6. Unrooted neighbor-joining tree (Saitou and Nei 1987) for values of divergence between amino acid sequences of group II intron-encoded genes. Twelve lineage *a* ORFs and all published lineage *b* ORFs were aligned manually, taking into account their conserved regions only. A Dayhoff distance matrix was generated by the PROTDIST program of the PHYLIP package, v. 3.5c (Felstenstein 1985). The *scale bar* represents 0.1 (corrected) amino acid substitution per site. Only the bootstrap values above 80 are indicated (100 replicates). Species are as in Fig. 2 and 4, with the addition of RT-like proteins from *Schizosaccharomyces pombe (S. pom)* in the cob gene (*x02819*), from *Anabaena* sp. PCC7120 (Bauer et al. 1994), from cryptophyte, *Pyrenomonas salina (P. sal)* in the plastid *cpn60* gene (Maier et al. 1995), and from *E. coli* (Ferat et al. 1994).

than the others, and this result is therefore likely to be due to a long-branch attraction phenomenon (see Philippe et al. 1995). The IIB ORF from *Escherichia coli* is the first to emerge in lineage *b*, well separated from all others, as in the tree published by Ferat et al. (1994).

In lineage *a* the intron ORF from the *M. polymorpha atp9* gene also consistently emerges first. The grouping of *cox1* intron 1, 2, and 3 ORFs from *P. littoralis* with those of *M. polymorpha cox1* intron 2 and *N. crassa cox1* intron 1 was found in all trees, consistent with the visual analysis of alignments. In contrast, the *Schizosaccharomyces pombe cob* intron 1 ORF does not appear as closely related to the latter proteins and, in some trees, branches together with the *S. cerevisiae cox1* ORFs.

Discussion

The Group II RT-Like Proteins Probably Coevolved with the Ribozyme Component of the Introns

Unlike other organelles, the *Pylaiella littoralis* mitochondrial genome possesses both group IIA and group IIB introns that contain complete RT complex proteins, i.e., with an RT domain, an X domain, and a zinc-finger-like domain. The recent discovery of two novel RT-like proteins contained in group IIB introns (Fontaine et al. 1995a) increased the number of available sequences for analysis of the complex proteins from this intron II lineage. As shown in Fig. 6, most of the proteins have coevolved with their respective introns: characterized subgroup IIA introns contain lineage *a* ORFs whereas known subgroup IIB introns contain lineage *b* ORFs. The distinctness of the two ORF lineages is well supported by bootstrap values in consensus trees inferred from both distance matrices and parsimony. Since the lineage *a* and *b* ORFs all have the same overall organization, they obviously originated from the duplication of a common ancestral multidomained gene. In most domains the lineage *a* ORFs are longer than those of the *b* lineage (Fig. 5), suggesting that they may have retained ancestral characteristics. In contrast with group I introns no shuffling between these ORF lineages and ribozyme components has yet been observed, suggesting that these have coevolved together.

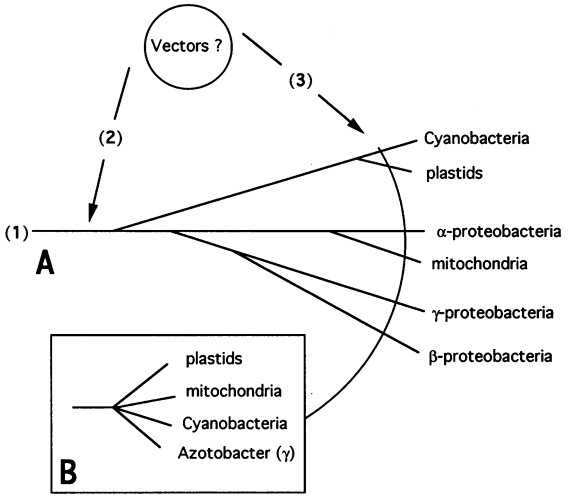

Fig. 7. Schematic drawing showing the expected phylogenetic branching of the RT-like intronic ORFs, according to the period of invasion by group II introns. Subtree *A* shows the consensus branching order of cyanobacteria, proteobacteria, and their derived organelles, mitochondria, and plastids, based upon ribosomal RNA gene sequences. Complete group II introns could have arisen in ancestral eubacteria (*1*), or have been propagated by unknown vectors (viruses?) into ancestral eubacteria (*2*) and then transmitted vertically. In such a case the phylogenetic tree based upon their ORFs should match the phylogeny of the other eubacterial genes, as shown in tree *A.* In contrast, if group II introns were transferred laterally (*3*) after the endosymbioses leading to plastids and mitochondria, one would expect a topology as in *B.* This is what is seen in Fig. 6 and in Ferat et al. (1994). The *E. coli* introns could have arisen from a previous contamination.

Distribution of Group II Introns

The two intron lineages (IIA, IIB) have representatives in both mitochondria and plastids. All of the complete group IIA introns found so far, i.e., containing multidomained ORFs, are inserted in mitochondrial genomes. In contrast, the group IIA known introns found in plastids do not contain complete ORFs. No group IIA intron has yet been found in free-living eubacteria. The complete group IIB introns have been so far identified in eubacteria, and in the *P. littoralis* mitochondrial LSU rDNA gene. Their ORFs group with those of two unclassified plastid group II introns. The other group IIB introns found in plastids do not contain ORFs and the other IIB introns reported in mitochondria do not contain RT-like complex proteins. This is, for example, the case for the mitochondrial *cox1* intron 5 of *Saccharomyces cerevisiae* (Bonitz et al. 1980), for the LSU rDNA intron of the chlorococcalean green alga *Scenedesmus obliquus* (Kück et al. 1990), and for the intron in a mitochondrial *trn* gene of the red alga *Chondrus crispus* (Leblanc et al. 1995). It is likely, however, that the introns that invaded these organellar genomes initially were mobile and contained the complete transposing machinery, i.e., with a reverse transcriptase, a maturase, and an endonuclease domain, which enabled them to insert at different locations in these genomes. For some unknown reason, the IIA introns have been preserved intact (with their ORFs) in mitochondria only, whereas intact IIB introns are much more rare in these organelles. As far as we know the only exceptions are the IIB introns from the mitochondrial LSU rDNA gene of *P. littoralis.* Once more, this finding may be thought of as another piece of evidence showing that this mitochondrial molecule has retained ancestral characteristics (Fontaine et al. 1995a; Delaroque et al. 1996).

The tendency of group II introns to invade one gene from place to place (Ferat et al. 1994) is particularly clear in the mitochondrial genome of *P. littoralis.* In this molecule the LSU rDNA gene displays three closely related group IIB introns, the third having lost its ORF while the *cox1* gene is split by three closely related group IIA introns. The fourth intron in the LSU rDNA gene, although belonging to subgroup B, is not of the same immediate origin (Fontaine et al. 1995a). Why these introns multiply in a given gene and do not invade other nearby genes is not known.

Origin of the Group II Introns

The scattered distribution of group II introns throughout various prokaryotic and eukaryotic lineages raises the question of their origin and transmission. Since they have been found in eubacteria (Ferat et al. 1994), a possible origin involves arising in these or from an early retroviral invasion of ancestral eubacteria. This would have been preceded or followed by an intron duplication, leading to complete IIA and IIB introns. Both intron lineages would then have been transmitted through successive endosymbioses of the prokaryotes which gave rise to mitochondria and plastids, then differentially and progressively eliminated from these organelles, losing their ORFs first.

Two findings, however, are inconsistent with the above hypothesis of a vertical transmission of group II introns. First, the detailed topology within both branches of the RT phylogenetic tree (Fig. 6) is not congruent with the 16S phylogeny of eubacteria and of the derived organelles nor of the eukaryotic tree for the *a* lineage (Fig. 7). In particular, with the exception of the *E. coli* intron, the *b* lineage branch shows a crown diversification of the RT-like complex proteins that encompasses intronic ORFs found in cyanobacteria, plastids, the mitochondria of *P. littoralis,* as well as in *Azotobacter vinelandii,* a γ-proteobacterium phylogenetically close to *E. coli* (see Ferat et al. 1994). Second, group II introns have not been found in the mitochondria of several eukaryotic lineages or groups such as animals and many protists (*A. castellanii, Prototheca wickerhamii, Chlamydomonas reinhardtii, Phytophthora infestans, Paramecium aurelia*). This distribution could be interpreted either as evidence for differential loss or independent insertion of these introns, but is probably more easily explained by independent insertion as in the case of nuclear spliceosomal introns (Stoltzfus et al. 1994).

If various prokaryotic and eukaryotic organisms were infected with group II introns, this must have occurred more or less at the same time (Fig. 7), i.e., at the period of the diversification of early eukaryotes. Retroviral-like group II introns may have been progressively invaded various cell genomes. Such a hypothesis of a late, lateral transfer of group II introns is consistent with both the distribution and the phylogenetic relationships of the two ORF lineages. Up to now, no group II introns have been found in nuclei. However, based upon similarities in their splicing mechanisms it has been suggested that nuclear mRNA introns may have arisen from organellar group II introns (Jacquier 1990; Sharp 1991), but such splicing similarity may only be due to a convergent evolution (Weiner 1993). It follows that the nuclear introns of direct (viral?) origin would have either been eliminated or transformed into nuclear-type introns which have retained only a very distant relatedness with the group II introns.

Acknowledgments. We wish to thank Dr. François Michel (CGM-CNRS, Gif/Yvette) for his help in constructing the intron secondary structure and his valuable comments on this paper. The authors are grateful to Claude Leroux and Olivier Collin for their help in the construction of phylogenetic trees. This work was supported by GDR 1002 ''Biology, biochemistry and genetics of marine algae'' (CNRS—IFREMER).

References

Bauer CC, Ramaswamy KS, Endley S, Golden JW, Haselkon R (1994) Suppression of a heterocyst differentiation in *Anabaena sp.* strain. (unpublished, U13767)

Bhatthacharya D, Druel LD (1988) Phylogenetic comparison of the small-subunit ribosomal DNA sequence of *Costaria costata* (Phaeophyta) with those of other algae, vascular plants and oomycetes. J Phycol 24:539–543

Bonitz SG, Coruzzi G, Thalenfeld BE, Tzagoloff A (1980) Assembly of the mitochondrial membrane system structure and nucleotide sequence of the gene coding for subunit 1 of yeast cytochrome oxidase. J Biol Chem 255:11927–11941

Burger G, Scriven C, Marchleidt W, Werner S (1982) Subunit 1 of cytochrome oxidase from *Neurospora crassa:* nucleotide sequence of the coding gene and partial amino acids sequence of the protein. EMBO J 1:1385–1391

Burger G, Plante I, Lonergan KM, Gray MW (1995) The mitochondrial DNA of the amoeboid protozoan, *Acanthamoeba castellanii:* complete sequence, gene content, and genome organization. J Mol Biol 245:522–537

Cummings DJ, Michel F, McNally KL (1989) DNA sequence analysis of 24.5 kilobase pair cytochrome oxydase subunit I mitochondrial gene from *Podospora anserina:* a gene with sixteen introns. Curr Genet 16:381–406

Costa M, Michel F (1995) Frequent use of the same tertiary motive by self-folding RNAs. EMBO J 14:1276–1285

Dalmon J, Loiseaux S (1981) The deoxyribonucleic acids of two brown algae *Pylaiella littoralis* L. (Kjellm) and *Sphacelaria sp.* Plant Sci Lett 21:241–251

Delaroque N, Fontaine JM, Kloareg B, Loiseaux-de Goër S (1996) Putative s-70 like promoters in a brown algal mitochondrial genome. C R Acad Sci 319:763–768

Doolittle RF, Feng DF, Johnson MS, Mc Clure MA (1989). Origin and evolutionary relationships of retrovirus. Q Rev Biol 64:1–30

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Ferat JL, Michel F (1993) Group II self-splicing introns in bacteria. Nature 364:358–361

Ferat JL, Le Gouar M, Michel F (1994) Multiple group II self-splicing introns in mobile DNA from *Escherichia coli.* C R Acad Sci Paris 317:141–148

Fontaine JM, Rousvoal S, Leblanc C, Kloareg B, Loiseaux-de Goër S (1995a) The mitochondrial LSU rDNA of the brown alga *Pylaiella littoralis* reveals α-proteobacterial features and is split by four group IIB introns with an atypical phylogeny. J Mol Biol 251:378–389

Fontaine JM, Rousvoal S, Delaroque N, Loiseaux-de Goër S (1995b) Characterisation of the *cox3, nad7* and *atp6* genes from the mitochondrial genome of the brown alga *Pylaiella littoralis.* Plant Physiol Biochem 33:605–609

Gorbalenya AE (1994) Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. Protein Sci 3:1117–1120

Gunderson JH, Elwood H, Ingold A, Kindle K, Sogin ML (1987) Phylogenetic relationships between chlorophytes, chrysophytes, and oomycetes. Proc Natl Acad Sci USA 84:5823–5827

Hardy CM, Clark-Walker GD (1991) Nucleotide sequence of the *cox1* gene in *Kluyveromyces lactis* mitochondrial DNA: evidence for recent horizontal transfert of a group II intron. Curr Genet 20:99–114

Jacquier A (1990) Self-splicing group II and nuclear pre-mRNA introns: how similar are they? Trends Biochem Sci 15:351–354

Jacquier A, Michel F (1990) Base-pairing interactions involving the 5′ and 3′-terminal nucleotides of group II introns. J Mol Biol 213: 437–447

Kück U (1989) The intron of a plastid gene from a green alga contains an open reading frame for a reverse transcriptase-like enzyme. Mol Gen Genet 218:257–265

Kück U, Godehardt I, Schmidt U (1990) A Self-splicing group II intron in the mitochondrial large subunit rRNA (LSUrRNA) gene of the eukaryotic alga *Scenedesmus obliquus.* Nucleic Acids Res 18: 2691–2697

Leblanc C, Boyen C, Richard O, Bonnard G, Grienenberger JM, Kloareg B (1995) Complete sequence of the mitochondrial DNA of the rhodophyte *Chondrus crispus* (Gigartinales). Gene content and genome organization. J Mol Biol 250:484–495

Loiseaux-De Goër S, Markowicz Y, Dalmon Y, Audren H (1988) Physical maps of the two circular plastid DNA molecules of the brown alga *Pylaiella littoralis* (L.) Kjellm. Location of the rRNA genes and of several protein coding regions on both molecules. Curr Genet 14:155–162

Maier UG, Rensing SA, Maerz GLIM (1995) Twintrons are not unique to *Euglena* chloroplast genome: structure and evolution of a plastome *cpn60* gene from a cryptomonad. Mol Gen Gent 246:128–131

McClure M (1991) Evolution of retroposons by acquisition or deletion of retrovirus-like genes. Mol Biol Evol 8:835–856

Michel F, Dujon B (1983) Conservation of RNA secondary structures in two introns families including mitochondrial, chloroplast and nuclear-encoded members. EMBO J 2:33–38

Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64:435–461

Michel F, Jacquier A (1987) Long-range intron-exon and intron-intron pairings involved in self-splicing of class II catalytic introns. Cold Spring Harb Symp Quant Biol 52:201–212

Michel F, Lang BF (1985) Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. Nature 316:641–643

Michel F, Umesono K, Oseki H (1989) Comparative and functional anatomy of group II catalytic introns. Gene 82:5–30

Mohr G, Perlman PS, Lambowitz AM (1993) Evolutionary relation-

ships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. Nucleic Acids Res 21:4991–4997

Moran JV, Mecklenburg KL, Sass P, Belcher SM, Mahnke D, Lwin A, Perlman PS (1994) Splicing defective mutants of the *cox1* gene of yeast mitochondrial DNA: initial definition of the marurase domain of the group II intron AI2. Nucleic Acids Res 22:2057–2064

Moran JV, Zimmerly S, Eskes R, Kennell JC, Lambowitz AM, Butow RN, Perlman PS (1995) Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. Mol Cell Biol 15:2828–2838

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Akashi K, Kanegae T, Ogura Y, Kohchi T, Ohyama K (1992) Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA, a primitive form of plant mitochondrial genome. J Mol Biol 223:1–7

Paquin B, Lang BF (1996) The mitochondrial DNA of Allomyces macrogynus: the complete genomic sequence from an ancestral Fungus. J Mol Biol 255:688–701

Philippe H, Germot A, Leguyadar H, Adoutte A (1995) Que savons nous de l'histoire évolutive des eucaryotes? 1. L'arbre universelle du vivant et les difficultés de la reconstruction phylogénéntique. m/s 11:1–12

Raitio M, Jalli T, Saraste (1987) Isolation of the gene for cytochrome c oxydase in *Paracoccus denetrificans.* EMBO J 6:2825–22833

Saitou N, Nei M (1987) The Neighbor Joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sachay DJ, Hudspeth DSS, Nadler SA, Hudspeth MES (1993) Oomycete mtDNA: *Phytophthora* genes for cytochrome c oxidase use an unmodified genetic code and encode proteins most similar to those of plants. Exp Mycol 17:7–23

Sharp PA (1991) Five easy pieces. Science 254:663

Shub DA, Goodrich-Blair H (1994) Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. Trends Biochem Sci 19:402–404

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolitle WF (1994) Testing the exon theory of genes: the evidence from protein structure. Science 265:202–206

Toh H, Hayashida H, Miyata T (1983) Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature 305:827–829

Weiner AM (1983) mRNA Splicing and autocatalytic introns: distant cousins or products of chemical determinism? Cell 72:161–164

Wolff G, Burger G, Lang BF, Kück U (1993) Mitochondrial genes in the colourless alga *Prototheca wickerhamii* resemble plant genes in their exons but fungal genes in their introns. Nucleic Acids Res 21:719–726

Wolff G, Plante I, Lang BF, Kuck U, Burger G (1994) Complete sequence of the mitochondrial DNA of the chlorophyte alga *Prototheca wickerhamii.* Gene content and genome organization. J Mol Biol 237:75–86

Yang J, Zimmerly S, Perlman PS, Lambowitz M (1996) Efficient integration of an intron RNA into double-stranded DNA by reverse splicing. Nature 381:332–335

Zimmerly S, Guo H, Eskes R, Yang J, Perlman PS, Lambowitz A (1995) A group II RNA is a catalytic component of a DNA endonuclease involved in intron mobility. Cell 83:529–538