RNA (2001), *7*:1142–1152. Cambridge University Press. Printed in the USA.

DOI: 10.1017.S1355838201010251

Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases

NAVTEJ TOOR,[1] GEORG HAUSNER,[1,2] and STEVEN ZIMMERLY[1]

[1]Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N 1N4, Canada

ABSTRACT

Catalytic RNAs are often regarded as molecular fossils from the RNA World, yet it is usually difficult to get more specific information about their evolution. Here we have investigated the coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases (RTs). Unlike group I introns, there has been no obvious reshuffling between intron RNA structures and ORFs. Of the six classes of intron structures that encode ORFs, three are conventional forms of group II A1, B1, and B2 secondary structures, whereas the remaining classes are bacterial, are possibly associated with the most primitive ORFs, and have unusual features and hybrid features of group IIA and group IIB intron structures. Based on these data, we propose a new model for the evolution of group II introns, designated the retroelement ancestor hypothesis, which predicts that the major RNA structural forms of group II introns developed through coevolution with the intron-encoded protein rather than as independent catalytic RNAs, and that most ORF-less introns are derivatives of ORF-containing introns. The model is supported by the distribution of ORF-containing and ORF-less introns, and by numerous examples of ORF-less introns that contain ORF remnants.

Keywords: catalytic RNA; intron-encoded protein; molecular evolution; retroelement

INTRODUCTION

Catalytic RNAs are usually thought to be remnants from the RNA World, but there is little evidence for the evolution of individual ribozymes. In general, evolution of catalytic RNAs is difficult to address because of the rapid change of RNA sequences and structures, combined with horizontal transfers for some catalytic RNAs. Perhaps the only catalytic RNAs whose evolution can be readily followed are rRNAs (Gutell, 1992) and RNase P RNAs (Frank & Pace, 1998), as they are present in all cells and are inherited vertically.

Group II introns are self-splicing RNAs commonly believed to have been ancestors of spliceosomal introns (Michel & Ferat, 1995; Nilsen, 1998). Structurally, group II introns are divided into two major classes, each with two variants: A1, A2, B1, and B2 (Michel et al., 1989; Michel & Ferat, 1995; Qin & Pyle, 1998). The major differences between A and B structures are indicated in Figure 1 by red and blue boxes, and include the ϵ' region (internal loop in domain IC1), the length of domain ID(iv), the EBS2 region, domain ID(iii)2, the domain III internal loop, and linkers between domains I–VI. A2 structures differ from A1 structures by an insertion in the 3′ strand of domain I(i)/I(ii) plus minor sequence differences (green boxes in Fig. 1A). B2 structures are distinguished from B1 structures by an insertion in the 5′ strand of domain I(i)/I(ii) and by interdomain linkers (green boxes in Fig. 1B).

A minority of group II introns encode reverse transcriptases (RTs) and are active retroelements (Lambowitz et al., 1999). The RTs of group II introns are related to RTs of non-LTR retroelements, and both classes of retroelements are mobile through a mechanism termed target primed reverse transcription (Luan et al., 1993; Zimmerly et al., 1995b). There have been two major speculations addressing the origin of group II intron retroelements. The favored model has been that an RT ORF (possibly retron-like) inserted into a preexisting self-splicing RNA structure to form a transposable element (Lambowitz & Belfort, 1993; Wank et al., 1999). If true, such an event may have occurred more than once, as RT ORFs are found in both group IIA and IIB intron structures. An alternative idea first conjectured by Curcio and Belfort (1996) is that the group II intron RNA structure could have developed at the termini of a retroelement to prevent the element from inactivating host genes. This second idea is inconsistent with an origin of the catalytic RNA structure in the RNA World.

Reprint requests to: Dr. Steven Zimmerly, Department of Biological Sciences, University of Calgary, 2500 University Drive NW, Calgary, Alberta T2N 1N4, Canada; e-mail: zimmerly@ucalgary.ca.

[2]Present address: Department of Botany, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.

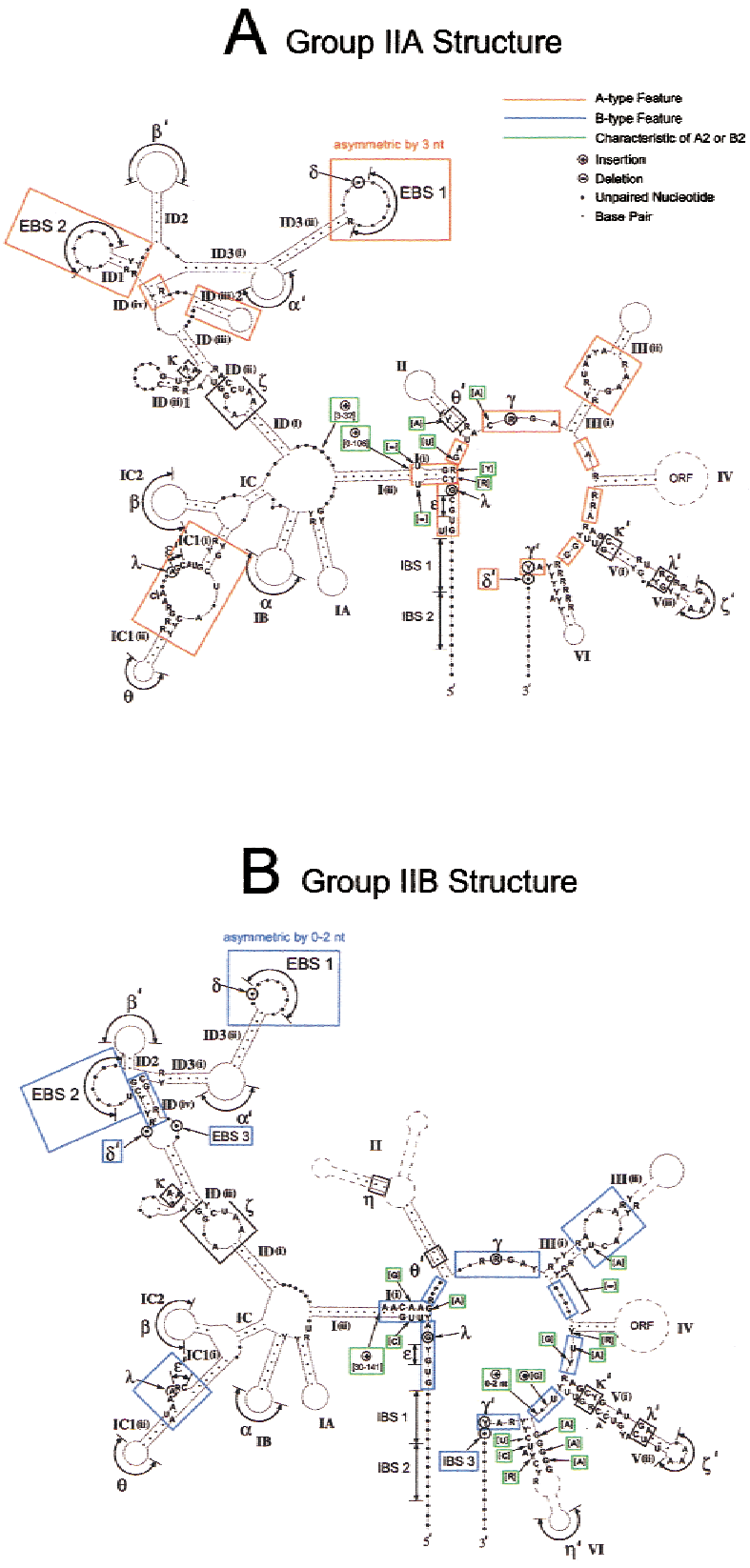

FIGURE 1. Consensus structures for group IIA and IIB introns (adapted from Michel et al., 1989; Qin & Pyle, 1998). Features characteristic of A and B structural forms are indicated by red and blue boxes, respectively, and green boxes show differences between A1 and A2, or B1 and B2 forms. Nomenclature of group II intron structures is according to Michel et al. (1989) and Qin & Pyle (1998).

Previously, we addressed the evolution of retrotransposable group II introns by analyzing the phylogeny of the intron-encoded RTs (Zimmerly et al., 2001). The inferred phylogenetic relationships revealed two major classes of ORFs, denoted the mitochondrial and chloroplast-like lineages. The mitochondrial lineage contains all known fungal mitochondrial introns, as well as plant and algal mitochondrial introns and bacterial introns, whereas the chloroplast-like lineage is a very heterogeneous group containing chloroplast, algal mitochondrial, and bacterial introns (Zimmerly et al., 2001). The majority of bacterial group II introns were positioned at the bases of the mitochondrial and chloroplast-like lineages. Rooting of the tree with non-LTR RT outgroups suggested that the bacterial families of ORFs were the earliest branching, but statistical support for their placement was very weak. The resulting evolutionary model predicted patterns of horizontal versus vertical inheritance of group II intron ORFs, and was consistent with an origin of mobile group II introns in bacteria with subsequent spread to mitochondria and chloroplasts.

This model for the history of mobile group II introns, however, was based entirely on ORF sequences and excluded information about intron RNA structures. Coevolution of intron RNAs and intron-encoded ORFs is not necessarily expected because group I intron RNAs and ORFs clearly did not coevolve (Loizos et al., 1994; Dalgaard et al., 1997). At least four classes of ORF have inserted into group I introns, and in multiple intron locations (Lambowitz et al., 1999). The group I intron of *Neurospora ND1*, for example, encodes two unrelated proteins in two different strains of *Neurospora* (Mota & Collins, 1988). The group I intron in the LSU rRNA of *Cryphonectria parasitica* encodes a fusion of the S5 ribosomal protein with a typical group I intron ORF of the LAGLIDADG family, whereas the intron in other fungal LSU rRNAs encode only the S5 ribosomal protein (Hausner et al., 1999). Group II introns, on the other hand, encode one class of ORF that is always located in intron domain IV (Lambowitz et al., 1999), a pattern suggestive of coevolution. One study of group II introns has suggested coevolution between intron RNAs and ORFs, but the analysis was not detailed and was based on a limited sample of 11 group II intron structures (Fontaine et al., 1997).

In this study, we have investigated the degree of coevolution between group II intron RNA structures and ORFs through detailed analysis of the intron RNA structures corresponding to the ORF phylogenetic tree. The analysis has identified six discreet classes of intron RNA structure associated with ORFs. Three of these are standard group II A1, B1, and B2 structures, and the remaining three classes are bacterial and have somewhat unusual structures. Because the RNA structural classes coincide with ORF phylogenetic groupings, coevolution between RNA structures and ORFs appears to be the general rule. The combined data support a new model for the evolution of group II introns, designated the retroelement ancestor hypothesis. This hypothesis predicts that currently known group II introns (both ORF-containing and ORF-less) are descendants of bacterial ORF-containing group II introns, and that the major structural forms of group II intron RNAs (group IIA and IIB) developed through coevolution with the intron-encoded ORFs in bacteria and organelles rather than as independent catalytic RNAs in the RNA World. The model is supported by the identification of many ORF-less group II introns that contain remnants of RT ORFs.

RESULTS

Figure 2 presents an unrooted phylogenetic tree derived from neighbor-joining analysis of the group II in-

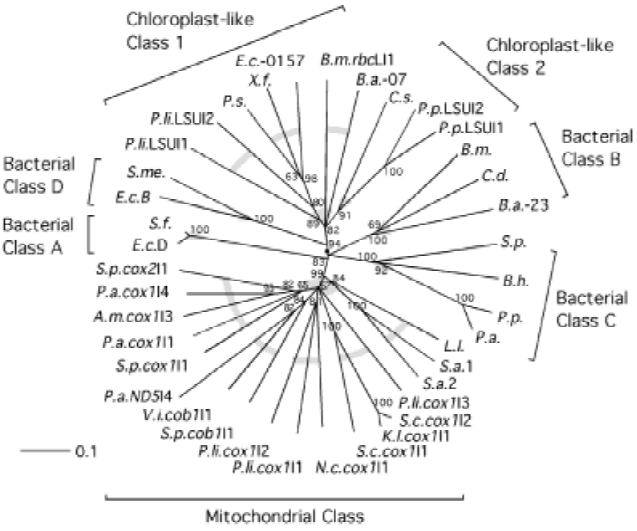

FIGURE 2. Phylogenetic relationships among group II intron ORFs. An unrooted phylogenetic tree was inferred by neighbor-joining analysis (PHYLIP, see Materials and Methods) based on 260 amino acids of RT subdomains 0–7 and X. Values for 1,000 bootstrap replicates are shown in percent, and nodes with less than 50% support were collapsed. Maximum parsimony analysis gave the same branching pattern with slightly lower bootstrap support. The black dot in the center marks the position of the root suggested by non-LTR RT outgroups, although support is quite weak (see text and Zimmerly et al., 2001). Gray sectors and arcs specify phylogenetic subgroupings referred to in the text. Abbreviations are: *A.m.*: *Allomyces macrogynus*; *B.a.*: *Bacillus anthracis*; *B.h.*: *Bacillus halodurans*; *B.m.* (no intron specified): *Bacillus megaterium*; *B.m.* (*rbcL*I1): *Bryopsis maxima*; *C.s.*: *Calothrix* sp.; *C.d.*: *Clostridium difficile*; *E.c.*: *Escherichia coli*; *K.l.*: *Kluyveromyces lactis*; *L.l.*: *Lactococcus lactis*; *N.c.*: *Neurospora crassa*; *P.a.* (*cox1*, *ND5*): *Podospora anserina*; *P.li.*: *Pylaiella littoralis*; *P.p.* (LSUI1 or LSUI2): *Porphyra purpurea*; *P.a.* (no intron specified): *Pseudomonas alcaligenes*; *P.p.* (no intron specified): *Pseudomonas putida*; *P.s.*: *Pseudomonas* sp.; *S.a.*: *Sphingomonas aromaticivorans*; *S.c.*: *Saccharomyces cerevisiae*; *S.f.*: *Shigella flexneri*; *S.me.* (no intron specified): *Sinorhizobium melliloti*; *S.p.* (*cob1*, *cox1*, *cox2*): *Schizosaccharomyces pombe*; *S.p.* (no intron specified): *Streptococcus pneumoniae*; *X.f.*: *Xylella fastidiosa*. Abbreviations are consistent with Zimmerly et al. (2001) and omit host gene names for bacterial introns.

tron ORFs considered in this article (see Materials and Methods). The tree recapitulates the previous analysis (Zimmerly et al., 2001) with two primary lineages of mitochondrial and chloroplast-like ORFs that are separated by four classes of bacterial ORFs. The black dot at the center signifies the suggested position of the root based on outgroups of non-long terminal repeat RTs; however, low bootstrap support (<55%) and slightly differing results using other outgroups and/or algorithms prevent definite conclusions about the position of the root (Zimmerly et al., 2001).

Intron RNA sequences corresponding to ORFs in the tree were folded into group II intron structures (see Materials and Methods and Fig. 1). Of great usefulness, it was found that introns associated with neighboring ORFs in the phylogenetic tree almost always shared sequence identity (38–75%; not shown) that provided covariation support for RNA helices. Sequence identity for introns corresponding to distantly related ORFs was significantly lower (e.g., *S.c.cox1*I2 (see Fig. 2 for organism abbreviations) and *E.c.*-0157 RNAs are 20% identical, based strictly on homologous positions in the secondary structures; not shown). Introns that were included in the previous analysis (Zimmerly et al., 2001) but which are excluded here because the RNAs could not be folded into standard structures are: *nad1*I4 introns of higher plants, *Marchantia polymorpha* (*M.p.*) *cox1*I1, *M.p.cob1*I3, *M.p.atpA*I1, *M.p.atpA*I2, *M.p.cox1*I2, *M.p.cox2*I2, and *M.p.atp9*I1. The failure of these introns to fold into consensus structures suggests that general splicing factors in plants (Vogel et al., 1999) may relieve the introns from the need to self-splice efficiently, or that there is extensive RNA editing (Wissinger et al., 1991).

Figure 3 summarizes the folded RNA structures as six consensus secondary structures. Each consensus structure corresponds to a phylogenetically defined grouping of ORFs as defined in Figure 2. Introns of the mitochondrial lineage of ORFs have uniform group IIA1 RNA structures containing all major features of group II A and A1 structures (red boxes). There is little sequence identity among the entire set of introns, but the consensus structure is strongly supported by covariation (yellow shading). Covariation is defined as sequences that: (1) are conserved at least 50% in primary sequence for at least two members; (2) contain at least four base changes preserving pairing (including G-U base pairs) for those two members; and (3) have the potential to form the pairing for all other members of the group. RNA structural variations that deviate from the consensus are in all cases consistent with (but not always identical to) phylogenetic groupings of ORFs. For example, the neighboring ORFs *N.c.cox1*I1 and *P.li.cox1*I1 have the group IIB form of the ϵ' region in domain IC1, and also have variants of intron domain II. The intron RNA structures of *S.p.cox2*I1, *P.a.cox1*I4, *A.m.cox1*I3, and *P.a.cox1*I1 lack domain ID(ii)1 along with the κ-κ' interaction. Interestingly, intron domain ID(iii) exists as five variants that are consistent with phylogenetic subgroupings (Fig. 3A, bottom). Together the data suggest coevolution of RNA structures and intron-encoded ORFs within the mitochondrial lineage.

Intron RNA structures associated with chloroplast-like class 1 ORFs are group IIB1 intron structures. B-like features (blue boxes) include domains I(i), the ϵ' region, domain ID(iv), EBS1, EBS2, domain III, and interdomain linkers; B1 features include the absence of an insertion between domains I(i) and I(ii), the domain III proximal loop, and two interdomain linkers. Again, RNA structural variations are consistent with phylogenetic groupings of ORFs. The introns *E.c.*-0157, *P.s.,* and *X.f.* have additional stems in intron domain IC2 and between domains ID2 and ID3(i), but lack the internal loop of domain VI.

Intron structures of chloroplast-like class 2 ORFs are group IIB2 in structure. B-like features are the ϵ' region, domain ID(iv), EBS2, domain III, and interdomain linkers; B2-like features are domain I(i), the insertion between domains I(i) and I(ii), and some interdomain linkers. The domain I/II linker and domain III are similar for both chloroplast-like classes, which reinforces the possibility of common ancestry between the two classes of intron structures.

The remaining classes of intron RNA structures are bacterial and are less typical. We have omitted the intron RNA structures of bacterial class A because its two sequences are 99% identical and the inferred structures have no support from covariation; however, the intron structural model contains a mixture of group IIA and group IIB features. Intron structures of bacterial class B resemble the group IIB2 class, but have many unique features. B-like features (blue) are domain I(i), the ϵ' region, ID(iv), EBS2, and some interdomain linkers. B2-like characteristics are the insertion in domain I(i)/I(ii) and the domain V/VI interdomain linker. Unusual features (green) are the absence of domain IA, an extended domain IB stem (shared among bacterial groups B, C, and D), a variant of the ϵ' region, domain ID(ii)1 and adjacent loops, the α' region, domain III, and the sequence NCGC in domain 5 in place of the highly conserved RAGC (NCGC is shared with bacterial class C).

Intron structures of the bacterial class C ORFs are by far the most unusual structures and differ significantly from both group IIA and IIB forms. The most singular feature is domain 5, which is normally highly conserved in sequence, but for this class is shortened by two base pairs and has a 5′ CCGC instead of the typical RAGC. The ϵ' region of domain IC is very different compared to other group II introns, but is conserved within the class. Missing domains are IC2, ID(iv), ID1, and ID2, and there is no obvious EBS2-IBS2 pairing. Three of the intron structures have anomalous inser-

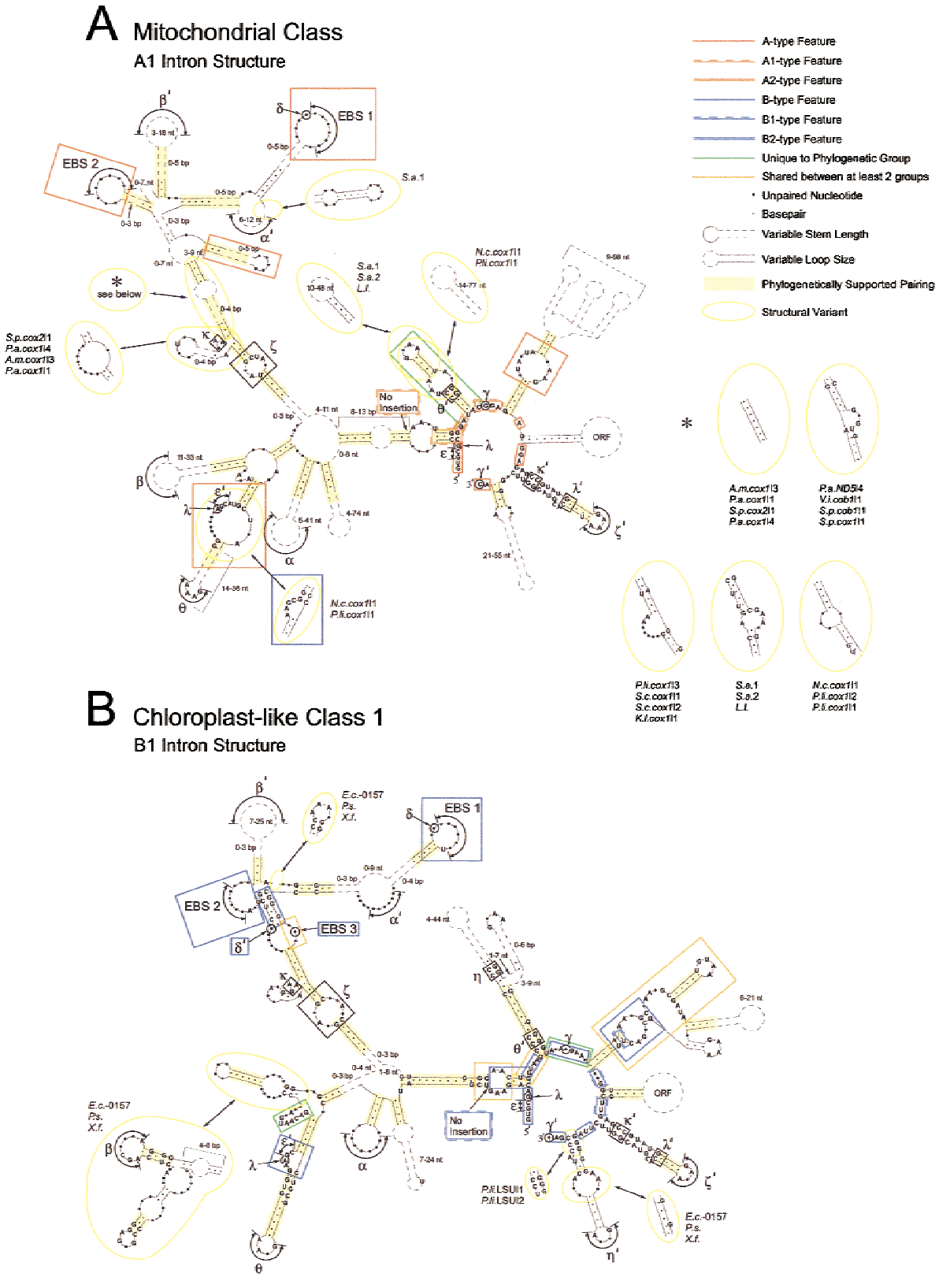

FIGURE 3. Consensus structures for ORF-encoding group II introns. Consensus structures are shown for six phylogenetic groupings of ORFs as defined in Figure 2. Consensus structures represent 100% identity for chloroplast-like class 2 and bacterial classes B, C, and D. For the mitochondrial class and chloroplast-like class 1, which are larger groups, consensus structures were derived as follows. First 100% consensus structures were made for subclasses of ORFs in the group (five and three subclasses for mitochondrial and chloroplast-like introns, respectively, as marked with gray arcs in Fig. 2). The overall consensus then represents a 3/5 or 2/3 consensus derived from the subconsensus structures. All deviations from the overall consensus are indicated, and consensus structures are accurate to ±1 bp in a paired region and ±2 nt in an unpaired region, with greater variations indicated by dotted lines and numbers. (*Figure continued on facing page.*)

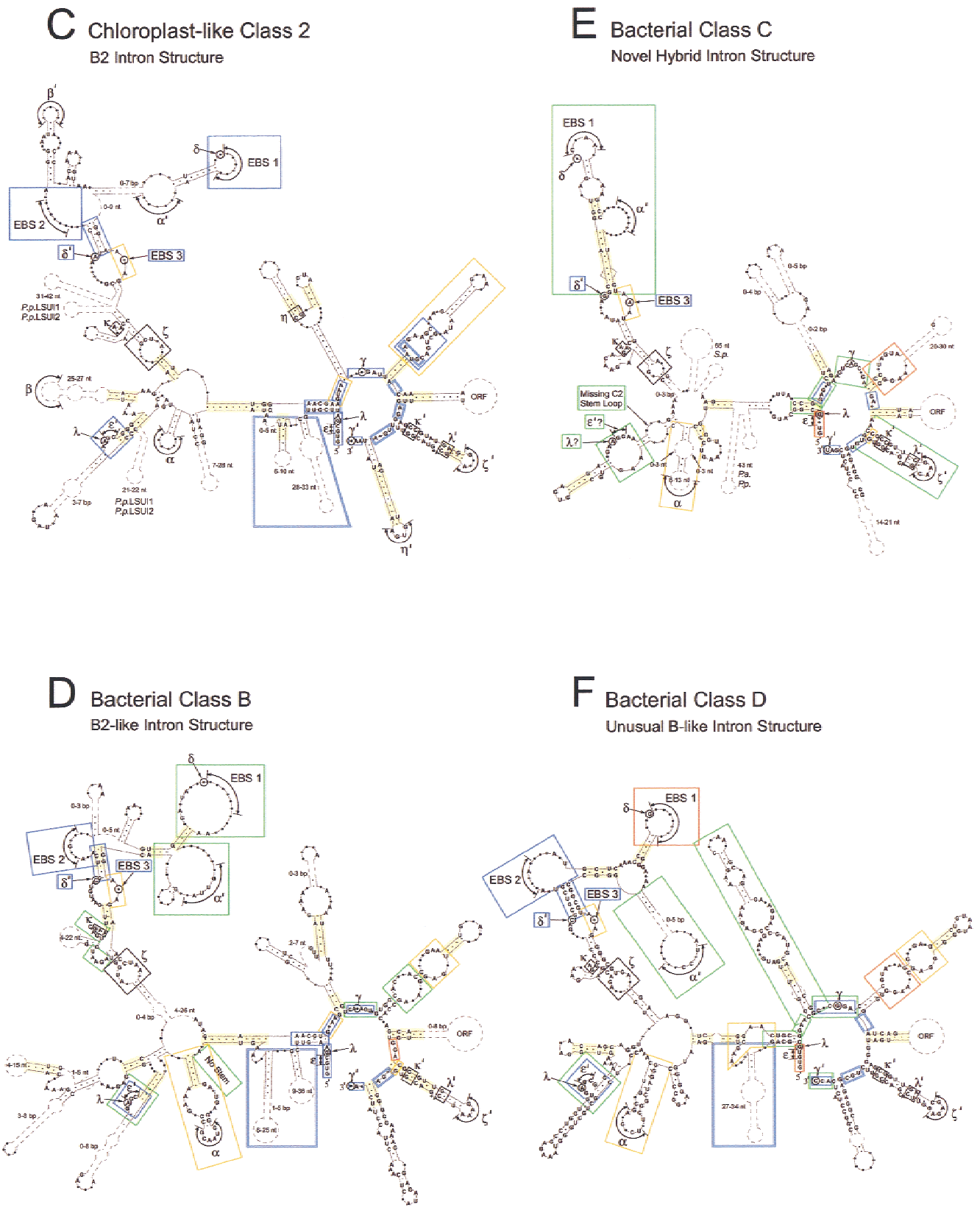

FIGURE 3 (*continued***).** Predicted tertiary pairings are marked (EBS1, EBS2, EBS3, α, β, γ, δ, ε, ζ, η, θ, κ, λ; Michel et al., 1989; Qin & Pyle, 1998; Boudvillain et al., 2000; Costa et al., 2000), although the potential interactions have not been experimentally demonstrated for all structural classes. Yellow highlighting indicates stems with covariation support (see text). **A:** Mitochondrial class, A1 intron structure. **B:** Chloroplast-like class 1, B1 intron structure. **C:** Chloroplast-like class 2, B2 intron structure. **D:** Bacterial class B, B2-like intron structure with unusual features (green). **E:** Bacterial class C, highly unusual structure with many unusual features (green), some A-like features (red), and some B-like features (blue). **F:** Bacterial class D, B2-like structure with A-like (red) and unique features.

tions in two different locations. Because of the unusual features, we cannot rule out that the introns are folded incorrectly or are mutated from their active forms. Nevertheless, the most unusual feature, domain 5, appears to be correctly folded, and the *P.a.* intron is reportedly mobile in vivo (Yeo et al., 1997), arguing that at least one of these introns is functional.

Introns of bacterial class D are group IIB-like but have some IIA features. B-like features are the ϵ' region, domain ID(iv), EBS2, and several interdomain linkers; B2-like features are the I(i)/I(ii) insertion and two interdomain linkers. A-like features are domain III, EBS1, and the 5′ GUGUG. Unusual features are an extended domain IB (shared with bacterial groups B and C), domain I(i), a variant ϵ' region, a stem-loop in the α' region, and the domain I/II linker. Although the double loops in domain III resemble those of chloroplast-like classes 1 and 2 and bacterial class B, and may perform similar functions in all four classes, only the chloroplast-like motifs match the previously defined consensus sequences (Michel et al., 1989).

In summary, we conclude that group II intron RNA structures encoding ORFs fall into six structural classes (seven if bacterial class A is included). Each RNA structural class corresponds to a phylogenetic grouping of ORFs, indicating a primary pattern of coevolution. Coevolution is supported by the observation of sequence identity between RNA structures corresponding to related ORFs, and conversely, low sequence identity for RNAs of distantly related ORFs. There are no examples of individual RNA structures that resemble intron secondary structures from another region of the ORF tree. Although the six classes of RNA structure were grouped initially because of ORF phylogenetic groupings, they also stand alone as RNA structural groupings because they have novel variations of conserved motifs (green and orange boxes in Fig. 3D–F), whereas subgroupings within the mitochondrial and chloroplast-like classes have comparatively minor deviations in less conserved regions.

If coevolution between intron and ORF were completely continuous, important ramifications would follow. Continuous coevolution for the entire set would predict that the intron structures developed in conjunction with the RT ORF rather than as independent RNAs, and in bacteria and organelles rather than in the RNA World. Further, it would predict that it would be possible in principle to follow the evolution of the catalytic RNA structure by linkage to ORF evolution. This possibility of continuous coevolution (alternatives discussed below) is the basis for a new model for group II intron evolution that we term the retroelement ancestor hypothesis. According to this hypothesis, mobile group II introns may have arisen in bacteria (Michel & Ferat, 1995; Lambowitz et al., 1999; Zimmerly et al., 2001), and the oldest intron structures corresponding to the data set may have been bacterial introns that encoded ORFs and had novel or "hybrid" intron RNA structural features. The "standard" group IIA and IIB families of intron RNA structures evolved subsequently in the mitochondrial and chloroplast-like lineages, codifferentiating with RT ORF structural types (Zimmerly et al., 2001). Finally, because most known group II introns are ORF-less and of group IIA and IIB forms, these ORF-less introns would be predicted to be the products of ORF loss from the mitochondrial and chloroplast-like lineages of mobile introns. This model organizes group II intron structural forms into a phylogenetic framework, and accounts for all of the major structural forms of group II introns except A2, possibly due to its low abundance in nature (Michel et al., 1989), or divergence after ORF loss. The hypothesis does not specify the ultimate origin of the catalytic RNA structure, that is, whether a primordial group II intron existed prior to association with an RT ORF and was perhaps a descendant of the RNA World. Neither does the hypothesis specify the source of the putative retroelement ancestor, that is, whether it was formed by insertion of an RT ORF into a group II intron structure, or from a retroelement that evolved a self-splicing intron structure at its termini, or from some other source.

The retroelement ancestor hypothesis is based on data from 39 group II introns, a small fraction of over 1,000 reported group II introns. To address plausibility of the hypothesis, we compiled a list of group II introns, drawing from data available in the Comparative RNA Web site (http://www.rna.icmb.utexas.edu/), the NCBI Organelle Genome Resource (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_euk_o.html), the Organelle Genome Database (http://megasun.bch.umontreal.ca/gobase/) and from published data in Ohyama et al. (1986), Shinozaki et al. (1986), Oda et al. (1992), Unseld et al. (1997), Ehara et al. (2000), and Zimmerly et al. (2001). By far the majority of reported introns are homologous introns in different plants. For example, there are over 1,000 reports alone of plant *trnK*I1 introns (matK family) and over 100 reported plant *nad1*I4 (matR) introns. In addition to redundant plant introns, a number of group II introns are "degenerate" and lack agreement with highly conserved structural features such as domain V and regions of domain I. Such degenerate group II-like introns and related group III introns are exceptionally abundant in euglenoids ($>$100 in the *Euglena* chloroplast genome; Copertino & Hallick, 1993). These introns have not been shown to self-splice, likely because of structural irregularities. Similarly, *Chlamydomonas reinhardtii* contains degenerate group II introns that are spliced *in trans* and lack conserved structural motifs (Rochaix, 1996). When our compilation of group II introns is corrected for homologous introns and highly degenerate introns, the data set is reduced to 142 group II introns: 20 introns in bacteria, 40 in lower eukaryotic mitochondria, 9 in lower eukaryotic chloroplasts, 25 in liverwort mitochondria, 19 in

liverwort chloroplasts, 21 in higher plant mitochondria, and 21 in higher plant chloroplasts (of which 13 are putatively homologous to liverwort chloroplast introns). An examination of published structures for these introns has not identified any as being related to bacterial class B, C, and D; however, it is not possible to exclude that relationships are overlooked, as highly accurate secondary structure models are required to make convincing conclusions.

The resulting compilation of introns (available upon request) supports the retroelement ancestor hypothesis by suggesting progressive ORF loss. All bacterial group II introns identified to date contain ORFs (Martinez-Abarca & Toro, 2000; Zimmerly et al., 2001). (However, it is important to recognize that the bacterial introns were identified solely by their encoded ORFs, and ORF-less introns would be difficult to identify.) In lower eukaryotes, about half of group II introns contain ORFs (27 ORF-containing versus 23 ORF-less introns in fungi and algae). In liverwort, a primitive land plant, most group II introns are ORF-less (9 ORF-containing versus 35 ORF-less) whereas in higher plants, virtually all group II introns are ORF-less (2 ORF-containing versus 40 ORF-less). This pattern suggests that bacterial and lower eukaryotic introns primarily contained ORFs, but the ORFs were lost in higher plants.

To test for specific evidence for ORF loss, we analyzed mitochondrial introns of liverwort, a primitive land plant that might represent a stage of ORF loss. Of the eight ORF-containing introns in liverwort, four have frameshifts and/or premature stop codons (*M.p.atp9*I1, *M.p.cox1*I1, *M.p.cox1*I2 *M.p.*SSUI1; Zimmerly et al., 2001). Seven of the remaining 17 ORF-less introns encode fragments of group II intron ORFs ranging from 36% identity over 596 amino acids to 61% over 49 amino acids (*nad2*I1, *nad3*I1, *nad7*I1, *nad7*I2, *nad4L*I2, *cox1*I5, *cox3*I1). Four additional introns (*nad4*I1, *rpl2*I1, *rps*I4, *cob1*I1) are related in sequence to other liverwort introns that contain ORFs or ORF fragments ($\geq$87% identity over $\geq$161bp in BLASTN match alignments), suggesting complete loss of the ORFs. Further, an examination of group II introns in *Arabidopsis thaliana* (*A.t.*) (higher plant) mitochondria revealed two introns that have no apparent ORF remnants yet are related in sequence to liverwort introns containing ORFs or ORF remnants. *A.t.nad2*I3 is 93% identical to *M.p.nad2*I1 (which encodes an ORF remnant) over 235 alignable base pairs based on BLASTN matches, and the introns are located in homologous exon sequences (Unseld et al., 1997). *A.t.nad1*I2 is 92% identical over 130 alignable base pairs to *M.p.atp9*I1 (ORF-containing). Thus, some group II introns with no traces of ORFs nevertheless appear to be derived from ORF-containing introns.

Additional ORF remnants in ORF-less introns are found in other early branching land plants: *Notothylas breutelii cox1*I4 (RT subdomains 2-7,X), *Sphagnum recurvum nad1*I4 (domain X), *Notothylas breutelii cox2*I3 (X, Zn domains), *Sphagnum recurvum cox2*I4 (RT subdomains 0-2), and *Andreaea rothii cox2*I4 (N-terminal amino acids). The only chloroplast intron with an identifiable ORF remnant was the algal intron *Codium fragile rbcL*I1 (RT subdomain 7, X), which is closely related to *Bryopsis maxima rbc*LI1, whose ORF has three frame shifts. No ORF remnants were identified in introns of liverwort or higher plant chloroplasts, or fungal mitochondria.

Additional examples of introns that have no detectable ORF remnants, but which are related in sequence to ORF-containing introns include: *Scenedesmus obliquus* LSUI1 (65% sequence identity to *P.li.*LSUI2, based on homologous positions in the intron secondary structures), *Pedinomonas minor* LSUI1 (50% identity to *P.li.*LSUI2), *S. obliquus* SSUI1 (49% identity to *P.li.*LSUI2), and *P.li.*LSUI4 (44% identity to *P.p.*LSUI1). By comparison, related ORF-containing introns have less identity to each other (*P.li.*LSUI2 is only 54% identical to *P.li.*LSUI1, 47% identical to *E.c.*-0157, 46% identical to *B.a.*-07; *P.p.*LSUI1 is 38% identical to *C.s.*) Of course, for these examples the direction of evolution is unclear, but it would appear that either an ORF was lost or gained. It is also notable that two biochemically characterized introns, yeast *cox1*I5γ and *cob1*I1, are also related to *P.li.*LSUI2 (46% and 44% identical, respectively) and might be products of ORF loss. Together, these examples demonstrate that at least some ORF-less group II introns were derived from ORF-containing introns, and the phenomenon may have been widespread.

DISCUSSION

We have compared group II intron RNA structures with the predicted phylogenetic relationships of the ORFs encoded within them and have observed a primary pattern of coevolution. Based on the possibility of continuous coevolution for the entire data set, we propose a new model for evolution of group II introns, denoted the retroelement ancestor hypothesis. This hypothesis predicts that the ancestral group II intron for the data set was a bacterial group II intron RNA structure containing "nonstandard" or hybrid structural features and encoding a compact reverse transcriptase ORF (Zimmerly et al., 2001). The "standard" A and B structural forms of group II introns are predicted to have originated subsequently by coevolution with ORFs in the mitochondrial and chloroplast-like lineages. ORF-less introns, which are mainly A and B forms, are predicted to be the result of ORF loss from mobile introns of the mitochondrial and chloroplast-like lineages. This scenario is not inconsistent with an origin for group II introns in the RNA World, but merely holds that the currently known forms of group II intron RNA structures differentiated in bacteria and organelles rather than in the RNA World.

The foundation of the hypothesis is the possibility of continuous coevolution between intron structure and RT ORF among all six phylogenetic classes. This is a very important point because if coevolution had no exceptions, it would then allow one to follow evolution of the catalytic RNA by associating it with ORF phylogeny. Although the data show a general pattern of coevolution, it is not possible to conclude global coevolution without exceptions because of uncertainties among the ORF relationships, and because of differences in RNA structural features that obscure relationships based on RNA alone.

Perhaps the strongest evidence for coevolution is our knowledge of the biochemical interactions between the intron RNA and RT protein. The *L.l.* RT (ltrA) binds very tightly to its intron (K_d = 0.25 pM), with a primary binding site in intron domain IV and additional contacts with other intron domains (Wank et al., 1999). Moreover, both intron and RT subunits are required for each reaction of the RNP particle, including forward splicing, reverse splicing into DNA, DNA cleavage, and template-specific reverse transcription (Zimmerly et al., 1995a, 1995b, 1999; Matsuura et al., 1997; Wank et al., 1999). This high degree of biochemical cooperation between intron and RT would present a barrier to the reshuffling of introns and ORFs while retaining full splicing and mobility functions. The close cooperation contrasts with group I introns, for which the ORF's mobility activity (DNA nuclease) is biochemically independent of the intron's self-splicing activity (Lambowitz et al., 1999), and this functional distinction may provide a rationale for why group II intron RNAs and ORFs predominantly coevolved whereas group I intron RNAs and ORFs did not.

Other lines of evidence supporting coevolution among the six groupings include: (1) the one-to-one correspondence of intron structures with ORFs (i.e., many random insertions would predict multiple ORF types per intron structural class, or multiple intron structures per ORF class); (2) the seemingly nonrandom distribution of three "hybrid" intron structures located phylogenetically between A and B structures; and (3) the invariant location of the ORF in intron domain IV despite multiple intron locations that can support insertions (Michel et al., 1989).

The only evidence for discontinuous coevolution comes from bacterial class B, because its intron structures resemble the chloroplast-like structures at least as much as bacterial class D, yet bacterial class B is not positioned next to the chloroplast-like class based on ORF phylogeny. Still, this is not considered to be a clear example of discontinuous evolution because of the many unique features of the bacterial class B intron RNAs and possibilities of convergent evolution for individual motifs (e.g., the ϵ' region in *N.c.cox1*I1).

Alternatives to the retroelement ancestor hypothesis assume some degree of discontinuous coevolution. Probably the most likely alternative is that each of the six classes of mobile group II introns had an independent origin by the insertion of an RT ORF into a different intron RNA structure. This explanation, however, would appear not to account for the one-to-one correspondence between ORF and introns types and other nonrandom patterns mentioned above. In other possible alternatives, the intron structures and ORFs might have freely reshuffled among the putatively most primitive bacterial introns, but after establishment of the mitochondrial and chloroplast-like lineages, coevolution predominated. Or, reshuffling between distantly related ORFs and intron RNAs might have necessitated convergent evolution of the RNA and/or protein motifs due to the functional constraints. Finally, intron RNA structures and ORFs could have reshuffled frequently among closely related introns, which would be undetectable in the data set.

Perhaps the major question regarding the credibility of the hypothesis is whether all ORF-less introns could have been derived from ORF-containing introns. Our compilation of group II introns suggests a pattern of progressive ORF loss from bacteria to plants, and furthermore, of the 79 compiled ORF-less introns, 11 were identified that contain ORF remnants. The actual frequency of ORF loss would be expected to be higher, as at some point the evidence of a former ORF would be lost completely. In agreement with this scenario, we identified 12 ORF-less introns that are related in sequence to ORF-encoding introns but that have no ORF remnants. ORF loss was also detected in a lower eukaryotic chloroplast (*C. fragile psbC*I4), supporting the occurrence in both organelles.

Interestingly, RT ORF remnants were not found in plant chloroplast introns or in fungal mitochondrial introns. One explanation might be the known differences in evolution of plant mitochondrial and chloroplast genomes (Palmer, 1990; Clegg et al., 1994). Plant mitochondrial genomes have a low rate of point mutation and appear to retain extraneous DNA, while chloroplast genomes have a several-fold higher rate of point mutations and appear to quickly lose extraneous DNA, at least in some cases. For example, in the nonphotosynthetic plant *Epiphagus*, the chloroplast genome has lost all its photosynthetic genes because they are no longer necessary, yet some chloroplast-derived photosynthetic genes are retained in the mitochondrial genome (dePamphilis & Palmer, 1990; Palmer, 1990). This evolutionary difference between organelles might help explain the retention of the ORF remnants in plant mitochondria but not in plant chloroplasts. Similarly, fungal mitochondrial genomes have a high rate of insertions, deletions, and recombinations (Clark-Walker, 1992) that might eliminate nonfunctional ORFs.

The evolutionary pattern observed in this study was a consequence of the selective consideration of RT-encoding group II introns. Might this focus on a subset

of examples distort perception of the true evolutionary history? This limitation cannot be dismissed, but there are plausible reasons to consider the opposite—that the narrow focus may have refined the data set and allowed observation of a previously unnoticed evolutionary pattern. Unlike group I introns, group II introns have generally been viewed as poor catalytic RNAs. Only a small number of group II introns are catalytic in vitro, and the reaction typically requires highly nonphysiological conditions (Michel & Ferat, 1995). The generalization that few group II introns are fully catalytic was primarily derived from the earliest introns identified, which were mainly ORF-less plant mitochondrial and chloroplast introns (Michel et al., 1989). In fact, plant organellar introns do not fold into good secondary structures (Michel et al., 1989; N. Toor, unpubl.), and none have been demonstrated to self-splice in vitro. The poor catalytic efficiency is thought to be compensated for by host splicing factors, and in plant chloroplasts, several splicing factors have been identified that appear to be general to either IIA or IIB intron structures (Vogel et al., 1999). The combination of marginal secondary structures and catalytic deficiencies suggest that many plant organellar introns have come to rely on host splicing factors, and have consequently lost catalytic proficiency, intron structural motifs, and possibly intron-encoded ORFs.

In contrast, it is expected that mobile group II introns should retain splicing proficiency, as splicing and reverse splicing are intrinsic to the mobility pathway. Focusing on mobile group II introns in this study would therefore be expected to enrich for catalytically proficient introns. Although it is true that a narrow focus on RT-encoding group II introns might erroneously bias the conclusions, it is also possible that eliminating ORF-less and degenerate introns may strip away "noise" from less catalytic intron structures and expose the backbone of evolution for the first time. Hopefully, all of the possibilities discussed here will eventually be resolvable. As more genomes are sequenced, more intron data points will be produced to either fill the gaps and confirm the proposed evolutionary pathway, or produce a new pattern and suggest a different evolutionary model.

MATERIALS AND METHODS

Phylogenetic analysis

GenBank accession numbers were listed previously (Zimmerly et al., 2001) except for *Xylella fastidiosa* (AE003999), *Notothylas breutelii nad1*I4 (AF068932), *Sphagnum recurvum nad1*I4 (AF068936), *Notothylas breutelii cox2*I3 (AF068930), *Sphagnum recurvum cox2*I4 (AF068935), *Andreaea rothii cox2*I4 (AF068934), *Codium fragile rbcL*I1 (M67453), *Scenedesmus obliquus* LSUI1 (NC_002254), *Pedinomonas minor LSUI1 (NC_000892), Scenedesmus obliquus* SSUI1 (AF204057), and *Pylaiella littoralis* LSUI4 (Z48620). Amino acid sequence alignment and phylogenetic analysis were described previously (Zimmerly et al., 2001) and were based on 260 amino acids from RT domains 0–7 and X. Analysis was with PHYLIP version 3.573c using the programs SEQBOOT, PROTDIST, NEIGHBOR, PROTPARS, and CONSENSE (Felsenstein, 1985, 1995). Rooting of the trees with non-LTR or retron RTs gave essentially the same results as previously obtained with a larger data set (Zimmerly et al., 2001).

RNA folding

RNAs were folded by Mfold version 3.0 (Mathews et al., 1999; Zuker et al., 1999) combined with the requirement for known structural motifs of group II introns, including the global structure and tertiary pairings (see Fig. 1). Structures were imported into RNAdraw (Matzura & Wennborg, 1996) for manual editing. Conserved sequences among related introns were required to fold into the same local structures. In cases where alignable primary sequence could pair in only one of two related structures, the bases were left unpaired to maintain agreement. For introns that could not be folded into a satisfactory group II intron structure (see text), we thoroughly eliminated the possibility that they were, in fact, similar to structures in another part of the tree.

ACKNOWLEDGMENTS

We thank Ken Sanderson for critical reading of the manuscript. This work was supported by National Science and Engineering Research Council Grant 203717-98. Salary support for S.Z. was from Alberta Heritage Foundation for Medical Research.

Received February 23, 2001; returned for revision
April 9, 2001; revised manuscript received May 8, 2001

REFERENCES

Boudvillain M, de Lencastre A, Pyle AM. 2000. A tertiary interaction that links active-site domains to the 5′ splice site of a group II intron. *Nature 406*:315–318.

Clark-Walker GD. 1992. Evolution of mitochondrial genomes in fungi. *Int Rev Cytology 141*:89–127.

Clegg MT, Gaut BS, Learn GH Jr, Morton BR. 1994. Rates and patterns of chloroplast DNA evolution. *Proc Natl Acad Sci USA 91*:6795–6801.

Copertino DW, Hallick RB. 1993. Group II and group III introns of twintrons: Potential relationships with nuclear pre-mRNA introns. *Trends Biochem Sci 18*:467–471.

Costa M, Michel F, Westhof E. 2000. A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J 19*: 5007–5018.

Curcio MJ, Belfort M. 1996. Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell 84*:9–12.

Dalgaard JZ, Klar AJ, Moser MJ, Holley WR, Chatterjee A, Mian IS. 1997. Statistical modeling and analysis of the LAGLIDADG family of site-specific endonucleases and identification of an intein that encodes a site-specific endonuclease of the HNH family. *Nucleic Acids Res 25*:4626–4638.

dePamphilis CW, Palmer JD. 1990. Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature 348*:337–339.

Ehara M, Watanabe KI, Ohama T. 2000. Distribution of cognates of group II introns detected in mitochondrial cox1 genes of a diatom and a haptophyte. *Gene 256*:157–167.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution 39*:783–791.

Felsenstein J. 1995. PHYLIP (Phylogeny Inference Package) version 3.55c. Distributed by the author. Department of Genetics, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip/getme.html.

Fontaine JM, Goux D, Kloareg B, Loiseaux-de Goër S. 1997. 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. *J Mol Evol 44*:33–42.

Frank DN, Pace NR. 1998. Ribonuclease P: Unity and diversity in a tRNA processing ribozyme. *Annu Rev Biochem 67*:153–180.

Gutell RR. 1992. Evolutionary characteristics of 16S and 23S rRNA structures. In: Hartman H, Matsumo K, eds. *The origin and evolution of the cell.* Singapore: World Scientific. pp 243–309.

Hausner G, Monteiro-Vitorello CB, Searles DB, Maland M, Fulbright DW, Bertrand H. 1999. A long open reading frame in the mitochondrial LSU rRNA group-I intron of *Cryphonectria parasitica* encodes a putative S5 ribosomal protein fused to a maturase. *Curr Genet 35*:109–117.

Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. *Annu Rev Biochem 62*:587–622.

Lambowitz AM, Caprara M, Zimmerly S, Perlman PS. 1999. Group I and group II ribozymes as RNPs: Clues to the past and guides to the future. In: Gesteland RF, Cech TR, Atkins JF, eds. *The RNA world*, 2nd ed. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 451–485.

Loizos N, Tillier ERM, Belfort M. 1994. Evolution of mobile group I introns: Recognition of intron sequences by an intron-encoded endonuclease. *Proc Natl Acad Sci USA 91*:11983–11987.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: A mechanism for non-LTR retrotransposition. *Cell 72*:595–605.

Martinez-Abarca F, Toro N. 2000. Group II introns in the bacterial world. *Mol Microbiol 38*:917–926.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol 288*:911–940.

Matsuura M, Saldanha R, Ma H, Wank H, Yang J, Mohr G, Cavanagh S, Dunny GM, Belfort M, Lambowitz AM. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: Biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes & Dev 11*:2910–2924.

Matzura O, Wennborg A. 1996. RNAdraw: An integrated program for RNA secondary structure calculation and analysis under 32-bit Microsoft Windows. *CABIOS 12*:247–249.

Michel F, Ferat J-L. 1995. Structure and activities of group II introns. *Ann Rev Biochem 64*:435–461.

Michel F, Umesono K, Ozeki H. 1989. Comparative and functional anatomy of group II catalytic introns—a review. *Gene 82*:5–30.

Mota EM, Collins RA. 1988. Independent evolution of structural and coding regions in a *Neurospora* mitochondrial intron. *Nature 332*:654–656.

Nilsen TW. 1998. RNA–RNA interactions in nuclear pre-mRNA splicing. In: Simons RW, Grunberg-Manago M, eds. *RNA structure and function.* Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press. pp 279–307.

Oda K, Yamato K, Ohta E, Nakamura Y, Takemura M, Nozato N, Kohchi T, Ogura Y, Kanegae T, Akashi K, 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.

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, Aota S, Inokuchi H, Ozeki H. 1986. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature 322*:572–574.

Palmer JD. 1990. Contrasting modes and tempos of genome evolution in land plant organelles. *Trends Genet 6*:115–120.

Qin PZ, Pyle AM. 1998. The architectural organization and mechanistic function of group II intron structural elements. *Curr Opin Struct Biol 8*:301–308.

Rochaix J-D. 1996. Post-transcriptional regulation of chloroplast gene expression in *Chlamydomonas reinhardtii*. *Plant Mol Biol 32*: 327–341.

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh, N, Shimada H, Sugiura M. 1986. The complete nucleotide sequence of tobacco chloroplast genome: Its gene organization and expression. *EMBO J 5*:2043–2049.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genet 15*:57–61.

Vogel J, Börner T, Hess WR. 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Res 27*:3866–3874.

Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM. 1999. A reverse transcriptase/maturase promotes splicing by binding at its own coding segment in a group II intron RNA. *Mol Cell 4*:239–250.

Wissinger B, Schuster W, Brennicke A. 1991. Trans splicing in *Oenethera* mitochondria: nad1 mRNAs are edited in exon and *trans*-splicing group II intron sequences. *Cell 65*:473–482.

Yeo CC, Tham JM, Yap MW-C, Poh CL. 1997. Group II intron from *Pseudomonas alcaligenes* NCIB 9867 (P25X): Entrapment in plasmid RP4 and sequence analysis. *Microbiology 143*:2833–2840.

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

Zimmerly S, Guo H, Perlman PS, Lambowitz AM. 1995b. Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell 82*:545–554.

Zimmerly S, Hausner G, Wu X. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res* 29:1238–1250.

Zimmerly S, Moran JV, Perlman PS, Lambowitz AM. 1999. Group II intron reverse transcriptase in yeast mitochondria: Stabilization and regulation of reverse transcriptase activity by the intron RNA. *J Mol Biol 289*:473–490.

Zuker M, Mathews DH, Turner DH. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In: Barciszewski J, Clark BFC, eds. *RNA biochemistry and biotechnology*. Dordrecht, The Netherlands: Kluwer Academic Publishers. pp 11–43.