

Mitochondrial introns: a critical view

B. Franz Lang, Marie-Josée Laforest and Gertraud Burger

Robert Cedergren Centre, Program in Evolutionary Biology, Canadian Institute for Advanced Research, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, C.P. 6128, Montréal, Québec H3T 1J4, Canada

Although group I and group II introns were discovered more than 25 years ago, they are still difficult to identify. Modeling their RNA structure also remains particularly challenging for organelle sequences, owing to their great diversity. In fact, accelerated evolution in organelles often results in a reduced RNA structure and a loss of autocatalytic splicing and intron mobility. We set out to identify all mitochondrial group I and II introns in published sequences, and, to this end, we developed and applied a new search approach: RNAweasel. On the basis of the results, we focus here on building a comprehensive picture of mitochondrial group I introns, including a modified (reduced) consensus RNA secondary structure and a concise phylogeny-based subclassification.

Introduction

Introns – sequences that interrupt the coding region of genes – have been divided into four main types on the basis of their splicing mechanism: spliceosomal introns, nuclear and archaeal tRNA introns, group I introns and group II introns. Group I and group II introns are the subject of this opinion article. These introns are characterized by their distinct, conserved RNA secondary structure and were first recognized in the early 1980s in fungal mitochondria [1–3]. Subsequently, this classification was extended to include introns in nuclear, bacterial and plastid genomes [4–7]. It was further recognized that intron distribution is highly uneven. For example, group I introns are common in nuclear rRNA-encoding genes and abundant in fungal mitochondrial genes (with preference for the genes *cox1*, *cob* and *rnl*), but they are absent from most animal and protist mitochondrial genomes. By contrast, in general, group II introns are uncommon, except in plant mitochondrial genomes, in which they are the predominant intron type. Knowledge of the distinctive structural features of group I and group II introns (i.e. the constituents of the secondary and tertiary structure of the ‘intron cores’, including the regions surrounding the P1, P3, P4, P6 and P7 helices of group I introns, and domains I to VI of group II introns [8–10]) is essential for the prediction of their catalytic properties and for their classification into major subgroups (of which there are at least six for group I introns and two for group II introns). Unfortunately, electronic versions of secondary (and tertiary) structure models are unavailable in most instances. Barely a dozen explicit secondary structure diagrams are accessible on the World Wide Web [11,12], despite comprehensive alignments of intron sequences being published more than 15 years ago (for group I introns, see Ref. [13]).

Certain group I introns and group II introns (Box 1) are mobile and proliferate through a mechanism termed intron homing (see Glossary) [14,15]. In addition, some of these introns carry out autocatalytic splicing *in vitro* [7,16–18]. These features have become the main focus of most reviews of introns [5,15,19], contributing to the common misconception that group I and group II introns are genuinely mobile and self-splicing. By contrast, publications that document the absence of these features [20–23] have received much less attention, although such introns use intricate mechanisms in excision and exon ligation.

Here, we highlight the less well-known properties of mitochondrial group I introns. We discuss why it is difficult to identify these introns with the available bioinformatics tools. By using a new and efficient intron-prediction tool, we compile a comprehensive and

Glossary

Autocatalytic splicing: accurate splicing reactions carried out *in vitro* by intron precursor RNAs in the absence of protein factors. (The term protein-assisted autocatalytic splicing is, therefore, misleading.) It should be noted that *in vitro* splicing is usually observed under nonphysiological conditions, in the presence of substances that stabilize RNA structure (e.g. polyethylene glycol, dimethylsulfoxide, bovine serum albumin and high concentrations of magnesium). *In vivo* splicing of mitochondrial group I and group II introns usually involves proteins that assist in RNA folding. These proteins can be encoded in the nucleus and/or by the introns themselves (modified homing endonucleases).

GNRA tetraloop: a compact four-base loop with the conserved sequence GNRA (where N denotes any nucleotide and R denotes A or G), which participates in tertiary folding of RNA molecules [42]. The most frequent GNRA tetraloops in introns have the sequence GAAA (Figure 1) and GUGA.

Homing endonucleases: enzymes that facilitate intron homing. Group I introns often contain long open reading frames (ORFs) encoding endonucleases that are involved in intron mobility and have rare recognition sites. Mitochondrial group I intron ORFs mostly fall into two of the four main families of homing endonucleases [38,39], which are designated according to their conserved sequence motifs: LAGLIDADG and GIY-YIG [40]. The LAGLIDADG endonuclease family is so diverse that global sequence alignment of this protein family – even within its most conserved sequence motif – is problematic. (For an example, see Ref. [40].) Placing one of the deviant types of LAGLIDADG group ORFs into a separate group, previously known as omega (named after the omega intron in the mitochondrial large subunit rRNA gene of *Saccharomyces cerevisiae* [41]) might resolve this issue. Nevertheless, a reclassification of all known intron ORFs by phylogenetic methods, including a rigorous statistical analysis, would be preferable.

Intron homing: the mechanism by which introns move into previously intronless genes. Group I introns and group II introns use different intron-homing mechanisms. Group I introns rely on highly specific endonucleolytic cleavage of intronless genes by an enzyme called homing endonuclease. This cleavage triggers activation of the DNA repair machinery of the cell, resulting in precise insertion (i.e. ‘homing’) of the intron sequence into the target gene. (Group II introns are discussed in Box 1.) In an organelle, intron-containing genes are a template for repeated transfers until all chromosomal copies contain the intron.

Intron maturases: intron-encoded proteins (usually derived from homing endonucleases) that are required for the correct splicing of non-autocatalytic introns.

Pseudoknot: a nested RNA structure in which a stretch of bases in a loop pairs with bases that are external to that loop.

Corresponding author: Lang, B.F. (franz.lang@umontreal.ca). Available online 5 February 2007.

Box 1. Common features of group II introns

Mitochondrial group II introns are generally less frequent than group I introns, except in land plants, in which they are the predominant intron type. Some group II introns are autocatalytic *in vitro* (e.g. the first intron in the *cob* gene of *Saccharomyces cerevisiae* [17]), whereas others require protein factors to mediate splicing (e.g. an intron in the *cob* gene of *Schizosaccharomyces pombe* [20]). Group II introns can participate in *trans*-splicing of discontinuous genes, a process that involves several intron segments encoded at distant genome locations. The corresponding exons plus adjacent intron portions are transcribed separately, and the typical RNA secondary and tertiary structure that is required for splicing arises through interactions between the intron parts. In an intron in the *nad1* gene of *Oenothera*, *trans*-splicing further depends on preceding RNA-editing steps, which confer autocatalysis *in vitro* [43]. Some group II introns are mobile, but the mechanism by which they propagate differs from that of group I introns. Group II intron RNAs reverse splice into a DNA target, and this is followed by reverse transcription mediated by intron-encoded proteins [15,44]. Curiously, mitochondrial group II introns of several ascomycete and basidiomycete fungi (e.g. *Cordyceps* spp. [45]) and of the protist *Nuclearia* sp. (B.F.L., unpublished) do not encode reverse-transcriptase-type proteins but encode endonucleases that are characteristic of group I introns. These exceptions

suggest that intron classification based on intron open reading frames is unreliable and that features of the conserved RNA primary and secondary structure should be used instead.

The most distinctive and best-recognized secondary structure element of group II introns is domain V, a bulged stem-loop that often carries a GNRA-tetraloop motif (see Glossary; Figure 1). Domain V (sometimes in conjunction with the adjacent, less well-conserved, domain VI) is used for identifying group II introns (e.g. in Rfam). Most other structural features are more variable, and comprehensive alignments of group II introns are unavailable. Indeed, group II introns in tRNA or rRNA genes persistently lack one or other of the otherwise well-conserved intron-exon interactions [intron-binding site 1 (IBS1) and IBS2, and exon-binding site 1 (EBS1) and EBS2] and tend to be structurally reduced [46]. The reduction in the intron RNA secondary structure might be compensated by the RNA structure of flanking exons.

Previously, the restricted number of group II intron sequences available has limited comparative analyses and the development of more-sophisticated intron search strategies. Now, hundreds of group II intron sequences are known for mitochondria, plastids and bacteria. An update on their structural classification, and the development of more-accurate approaches for intron identification and secondary structure modeling (as discussed here for group I introns), is highly desirable.

up-to-date set of mitochondrial group I introns. On the basis of this data set, we infer the size range of these introns and the variations in their RNA secondary structure. This broad comparison enables us to redefine the universal core structure and to specify a rigorous subgroup classification using phylogenetics. Our analyses caution against popular generalizations about intron structure and function, and emphasize the astounding biological diversity of group I introns.

How many mitochondrial group I introns are there?

Surprisingly, we do not have a more precise answer to this question than 'a few hundred'. The identification of mitochondrial sequences in public sequence repositories is inherently difficult in the absence of reliable information that documents the subcellular origin (i.e. the source) of genomic sequences. This shortcoming has been addressed in GOBASE (<http://megasun.bch.umontreal.ca/gobase>) [24], a relational database that contains the mitochondrial and plastid sequence subsets of GenBank (<http://www.ncbi.nlm.nih.gov/Genbank>) and that provides curated and standardized information on genes, gene products and taxonomy, in addition to structure diagrams of RNAs. As of January 2007, the total number of complete mitochondrial group I or group II introns stored in GOBASE was 1230. However, we do not know how many of these intron sequences have been identified correctly and how many remain undetected in the database. In addition, most introns retrieved from GenBank are not annotated with respect to the group or subgroup to which they belong.

An assessment of current tools for intron prediction

Among the bioinformatics approaches to intron identification, the most widely used are CITRON [25] and Rfam (<http://www.sanger.ac.uk/Software/Rfam>) [26]. CITRON was designed to predict only group I introns. It uses a nucleotide consensus matrix of the intron core plus a hierarchical search of peripheral secondary structure

elements (helices and junctions). CITRON searches are based on the information inferred from 143 introns that were known when the program was developed in 1994 [25]; at that time, CITRON had a 92% success rate at correctly predicting group I introns. By contrast, the tools of the Rfam service are generic, using covariance to build statistical models from training sets of aligned RNA sequences. Unfortunately, Rfam does not consider RNA pseudoknots, which are distinctive and often highly conserved.

We tested the performance of these two applications using nine carefully annotated complete mitochondrial DNA (mtDNA) sequences (Table 1), which contained 112 group I introns and 14 group II introns; we generated secondary structure diagrams of most of these introns and deposited these in GOBASE. CITRON (at the default parameter settings) predicted more group I introns (63% sensitivity; Table 1) than did Rfam. However, the interpretation of results generated by CITRON is complicated by many (58) false positives and by an overwhelming number of redundant solutions (455). The identification of false positive and redundant solutions is cumbersome, often requiring the generation of secondary structure diagrams to assess which of the solutions might be valid. It is not surprising that the sensitivity of CITRON predictions is particularly low (32%) for *Rhizophyidium brooksianum* introns, because the tool was developed before the availability of these unusually derived structures (see also mini-introns, discussed later).

By contrast, using Rfam to predict group I and group II introns generated readily interpretable and nonredundant results, but sensitivity was weak (48% sensitivity overall; 41% for group I introns; and 100% for domain V of group II introns, which is well conserved). The number of false positives is low, and this is achieved by setting a conservative cut-off score at the cost of sensitivity, which is most apparent in the failure to detect most *R. brooksianum* introns (15% sensitivity). In addition, the sequences used in the training set do not span the full range of diversity, in part explaining the low sensitivity for predicting group I

Table 1. Comparison of bioinformatics approaches for intron identification^a

Species	GenBank accession number	Group I ^b	CITRON group I predictions ^c	Rfam group I predictions ^c	Group II ^b	Rfam group II predictions ^c
<i>Acanthamoeba castellanii</i>	U12386	3	2 (15, 2)	1	0	0
<i>Allomyces macrogynus</i>	AMU41288	26	24 (179, 23)	12	2	2
<i>Podospira anserina</i>	X55026	31	19 (109, 9)	20	3	3
<i>Porphyra purpurea</i>	AF114794	0	0	0	2 (3) ^d	3
<i>Prototheca wickerhamii</i>	PWU02970	5	2 (31, 9)	2	0	0
<i>Rhizophyidium brooksianum</i>	AF404306	34	11 (8, 9)	5	2	2
<i>Saccharomyces cerevisiae</i>	AJ011856	9	8 (29, 2)	6	4	4
<i>Schizosaccharomyces pombe</i>	X54421	2	2 (10, 4)	0	1	1
<i>Metridium senile</i>	MSAF000023	2	2	0	0	0
Total	NA	112	70 (455, 58)	46	14 (15)	15

^aAbbreviation: NA, not applicable.

^bNumber of introns in complete mtDNAs of the given species.

^cNumber of introns (group I or group II) identified by using a given approach. CITRON only predicts group I introns; the numbers in parentheses are the total number of redundant solutions, followed by the number of false positives. Rfam does not have false positives, owing to rigorous cut-off values.

^dIdentification of group II introns is mainly based on domain V, which is highly conserved. This results in a false positive for *P. purpurea* mtDNA, which has a partial (pseudo) intron in an intergenic region.

introns using this tool. In particular, mitochondrial group I introns vary widely in size, owing to large insertions that occur at various points in the RNA secondary structure [10]. To capture the full extent of length variation, the multiple alignment of the training set needs to include examples of the longest insertions. This would extend the length of global group I intron alignments to several thousand bases, much longer than the current group I intron alignments in Rfam. Searching with such extended-length structural models would increase the computational load markedly. Because using Rfam tools already requires both patience and access to super computers, even for the prediction of moderately sized RNA structures, the average user cannot use Rfam for the identification of the full range of mitochondrial group I introns. At present, the sequence length limit for searches at the Rfam website is 2000 bases, shorter than some mitochondrial group I introns and much shorter than many intron-containing mitochondrial genes. (Certain *cox1* genes are longer than 20 kb: e.g. in *Podospira anserina* [27].)

A new tool for group I and group II intron identification

To identify group I and group II introns in published collections of organelle sequences, we developed an alternative intron prediction tool – RNAweasel (<http://megasun.bch.umontreal.ca/RNAweasel>) – that makes use of the computationally efficient search engine ERPIN (Easy RNA Profile Identification; <http://tagc.univ-mrs.fr/erpin>) [28]. The search algorithm of ERPIN utilizes RNA primary and secondary structure profiles, which are computed from training sets of RNA sequence alignments plus user-defined secondary structure information. Much of the efficiency of the algorithm stems from the definition of precisely delimited structural elements that can be searched individually or in combination. In addition, the option to use a specific search order (i.e. search strategy) can reduce execution time substantially. Finally, ERPIN provides expectation value cutoffs for both individual and combined sequence elements, facilitating the development of secondary structure models. But, similar to the covariance approach underlying Rfam, ERPIN is difficult to use, because it does not provide tools that would facilitate the

compilation and manipulation of training-set sequences. Therefore, we have added a suite of new functionalities (B. Franz Lang, unpublished) that enable the following: (i) easy visualization and editing of alignments and structure definitions (using the Genetic Data Environment (GDE) sequence editor [29]); (ii) automatic alignment of ERPIN results; (iii) normalization of training-set sequences to increase the sensitivity of searches; and (iv) a reiterative mode of searching.

A major failure of CITRON and Rfam is that they consider only common core features of group I introns, although intron subgroups differ considerably in both conserved motifs and peripheral elements of their overall secondary structure [10]. Therefore, a substantial increase in sensitivity is expected when several subgroup-specific intron models are defined (a bioinformatics approach termed divide and conquer), in addition to models based on only core features. From a user's perspective, an additional benefit is that intron and intron subgroup predictions are carried out simultaneously. Therefore, we generated separate training sets for as many subgroups as was required to reach an average sensitivity of >95%, with a negligible number of false positives. We obtained a total of nine predictor models: one each for subgroup IA, IA3, IB, IC1, IC2 and ID, plus three models that were restricted to several core features, to capture the remainder of derived introns with ambiguous subgroup features. The predictor for mitochondrial group II introns was obtained from a single training set of domain V sequences (~98% sensitivity). The consensus sequence and secondary structure of this domain is shown in Figure 1. It should be noted that this consensus sequence has more conserved sites than have been established by other research groups [5].

RNAweasel correctly predicted all but six of the most highly derived *R. brooksianum* introns in the test set. When applying searches to all 1230 mitochondrial intron sequences in GOBASE (i.e. those annotated as being complete), the sensitivity is even higher (~98%; Table 2). Careful manual verification of the 59 GOBASE records in which ERPIN did not identify introns reveals that most of these sequences contain only partial introns

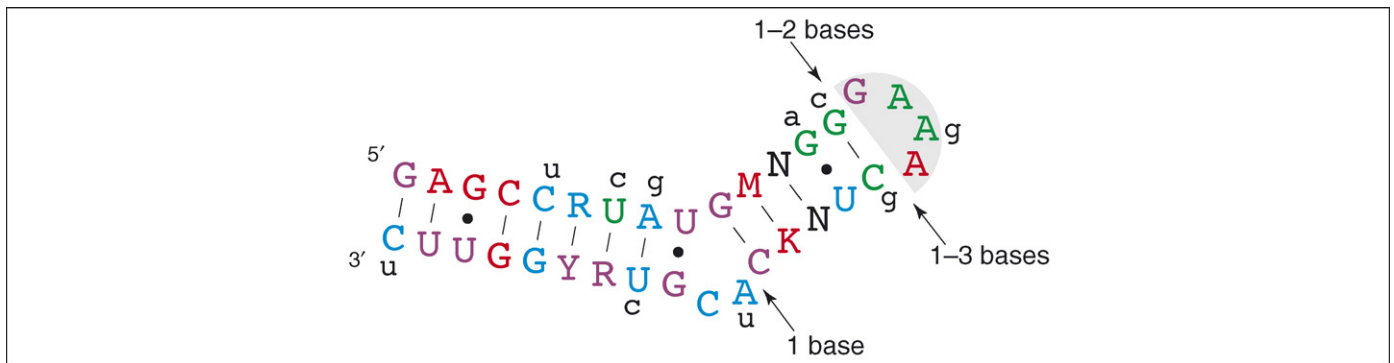


Figure 1. Secondary structure model of domain V of mitochondrial group II introns. The consensus structure is based on compilation of the 597 mitochondrial introns identified by ERPIN searches (Table 2). Regarding positional sequence conservation, bases that are conserved in $\geq 99\%$ of introns are shown in red; $\geq 95\%$, purple; $\geq 80\%$, blue; and $\geq 60\%$, green. Bases indicated in lower case are alternatives that occur at frequencies of $\geq 10\%$. Recurrent insertions of up to three bases are indicated by arrows. Filled circles indicate GU wobble interactions. It should be noted that in some introns, the conserved GAAA-tetraloop motif (shaded gray), which interacts with a conserved structural motif in domain I, is absent; the size of the loop can vary by a few bases, and the number of bases in its helical region can be reduced. Abbreviations: K, U or G; M, A or C; N, A, C, G or U; R, A or G; Y, C or U.

(and had not been annotated as such by the sequence submitters). However, 16 bona fide group I introns remain undetected, among them the six highly derived *R. brooksonianum* introns and introns containing unusual insertions in otherwise highly conserved domains (insertions that are not considered in our training sets).

The main advantage of RNAweasel is the use of subgroup-specific intron predictors that can be easily developed and updated. On the basis of the RNAweasel intron alignments, the same divide-and-conquer approach could now be used to improve CITRON and Rfam. Although RNAweasel is the most efficient tool at present, it has the limitation that only intron core elements are considered. To accommodate the peripheral secondary structure elements (e.g. P1, P2, P5, P9, P10 plus the corresponding loops, junctions and tertiary interactions), which are more variable, more-sophisticated search algorithms that combine the most valuable features of the covariance model implemented in Rfam, and the modular search strategies of ERPIN, need to be developed. Such an approach should enable optional searching for the presence of conserved elements and helices of variable length, and it should enable the recognition of motifs, noncanonical base pairs and complex tertiary interactions (e.g. tetraloops and

their receptor sequences). Finally, for the identification of group II introns, elements in addition to domain V and domain VI need to be included in the search strategies, and separate models are also required to predict intron subgroups.

The current version of RNAweasel (<http://megasun.bch.umontreal.ca/RNAweasel>) can be used as part of the sequence analysis workbench AnaBench [30]. Similarly, the secondary structure diagrams of intron RNA that we used for building and testing the intron search models are accessible through GOBASE (<http://gobase.bcm.umontreal.ca/searches/intron.php>; check 'Yes' next to 'Secondary Structure Available' and retrieve).

Group I intron sizes: minuscule to extra large

On the basis of the expanded and validated set of mitochondrial group I introns described here, we can now reassess the typical features and structural diversity of these introns. A frequent misconception in the literature concerns the length of group I introns. For example, a recent review characterizes group I intron sizes as moderate with little variability (250–500 bases) [19]. Although this might be valid for group I introns residing in nuclear genes, those in mitochondria have an approximately tenfold-greater size range (142 bases to >3000 bases). At one extreme, mitochondrial introns contain long insertions that include either an open reading frame or consist of noncoding sequences similar to intergenic spacers. At the other extreme, introns are reduced to miniature versions with highly compact secondary structures that lack numerous structural elements of 'typical introns' (see the next section for a redefinition of the minimum consensus structure). These mini-introns occur in fungi (e.g. in *Spizellomyces punctatus* and *R. brooksonianum*).

Another frequent misconception is that all group I introns are mobile and splice autocatalytically. None of the mini-intron RNAs that we have tested experimentally can splice autocatalytically, and, in the absence of encoded homing endonucleases, they cannot propagate by intron homing. It should be emphasized that a small size does not necessarily preclude autocatalytic activity. Recently described group-I-intron-like structures residing in nuclear rRNA genes of the slime mold *Didymium iridis*

Table 2. Group and subgroup distribution of mitochondrial introns, based on ERPIN predictions^{a,b}

Intron group	No. of introns	E-value cutoff ^c
II	597	$6e^{-2}$
IA	73	$8e^{-9}$
IA3	5	$5e^{-20}$
IB	346	$8e^{-2}$
IC1	4	$4e^{-18}$
IC2	21	$5e^{-20}$
ID	53	$3e^{-12}$
Group-I-derived	72	$1e^{-2}$
Not found: group II	8	NA
Not found: group I	15	NA
Incorrect records ^d	36	NA
Total	1230	NA

^aAbbreviation: NA, not applicable.

^bStatus of records listed in GOBASE (January 2007).

^cThe listed E-values correspond to searching a target sequence of 1 Mb on both strands.

^dFour of 36 records contain two different introns; in all others, no intron was identified because sequences were either partial or did not contain an intron.

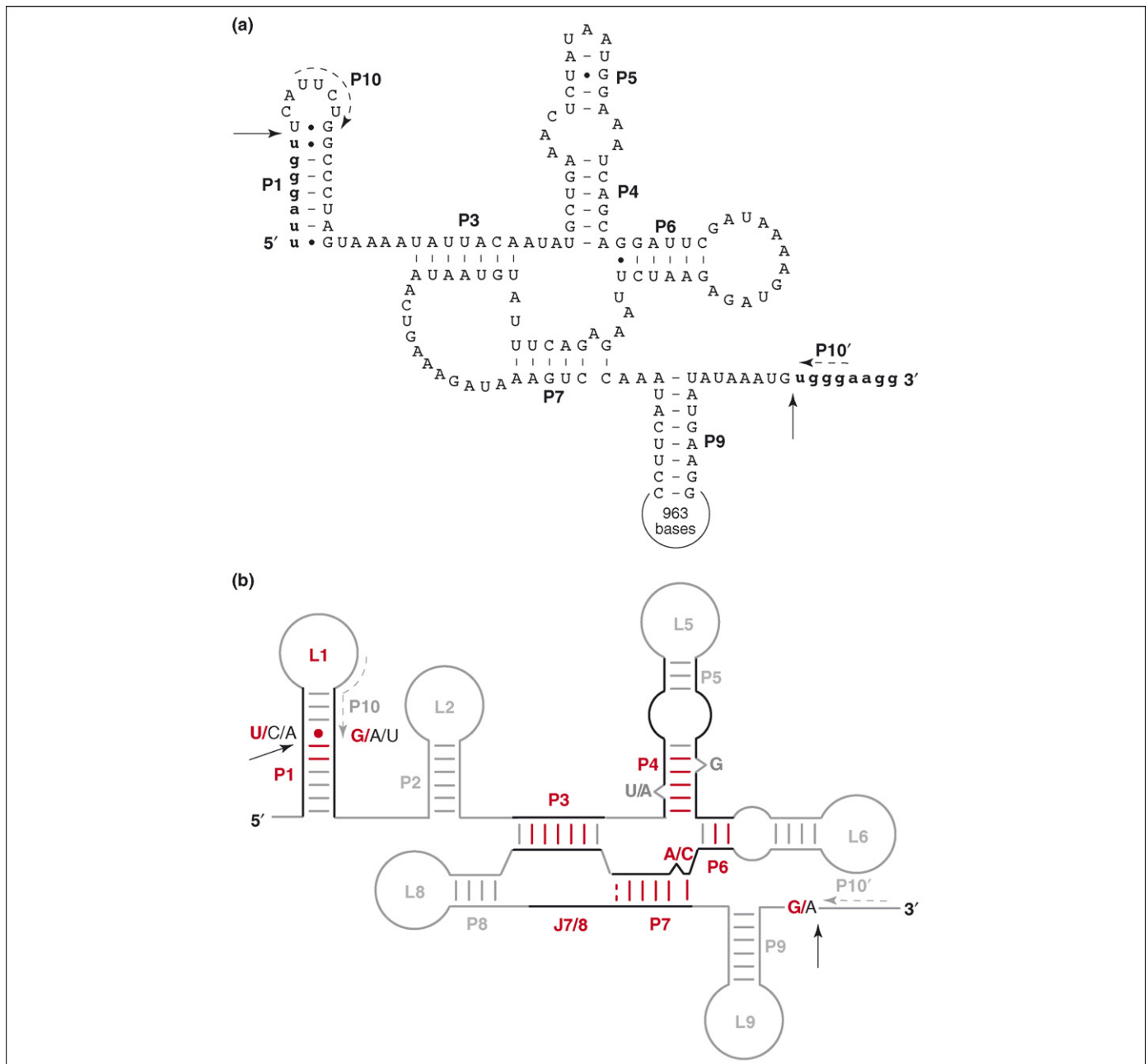


Figure 2. Secondary structure of mitochondrial group I introns. **(a)** Example of an unusual, highly reduced, group I intron lacking P2 and P8 pairings (the first intron of *nad5* in the fungus *Spizellomyces punctatus*). Intron sequence is shown in upper case. Exon sequence is shown in lower case. Unbroken arrows point to 5' and 3' splice sites. Broken arrows delineate P10. Filled circles indicate GU wobble interactions. **(b)** Mitochondrial consensus structure. The consensus structure is based on the mitochondrial structures identified in this survey. Black lines indicate the conserved intron core. Pairings (P), Loops (L) and junctions (J) are numbered according to previous conventions that emphasize secondary structure [6]. Universally conserved structural elements in mitochondrial group I introns are shown in red, and variable portions are shown in gray. The minimum number of conventional base pairs (A-U, G-C and G-U) in helices is shown in red. One base pairing in P7 (broken) is absent in several instances. Two recurrent bulged U/A and G bases (in group IA and IC1, and IC2 introns, respectively) in P4 are indicated. Solidi indicate alternative residues. Splicing typically occurs downstream (5') of a U that is paired with a G, indicated by unbroken arrows, and the last base of the intron is usually a G; these bases are shown in red, and rare exceptions to these rules are indicated in black. It should be noted that J3/4 and J3/7 vary in size and have, therefore, not been listed as part of an invariant core structure. P10 and P10' are the two halves of a helical interaction.

and various species of the amoeba *Naegleria* have small sizes (160–190 bases), similar to the fungal mini-introns, but are autocatalytic [31,32]. Conversely, normal or large intron size (several hundred bases and above) does not imply autocatalytic splicing activity. Loss of autocatalytic properties for both mitochondrial group I and group II introns has been reported in two fungi (for all five introns in *Schizosaccharomyces pombe* mtDNA and for at least seven of 13 introns in *Saccharomyces cerevisiae* mtDNA [2,20–23,33]). Apparently, these species compensate for

loss of structural RNA elements by recruiting auxiliary protein factors (either intron- or nucleus-encoded). In one example, the helper protein is a tRNA synthase [34]. In another, more complex, example, a unique nucleus-encoded protein (so far found only in *S. cerevisiae* and its close relative *Saccharomyces douglasii*) operates together with a mtDNA-encoded intron maturase protein [23,33]. We predict that mini-introns require a larger number of such helper proteins to compensate for their more advanced reduction of intron RNA structure.

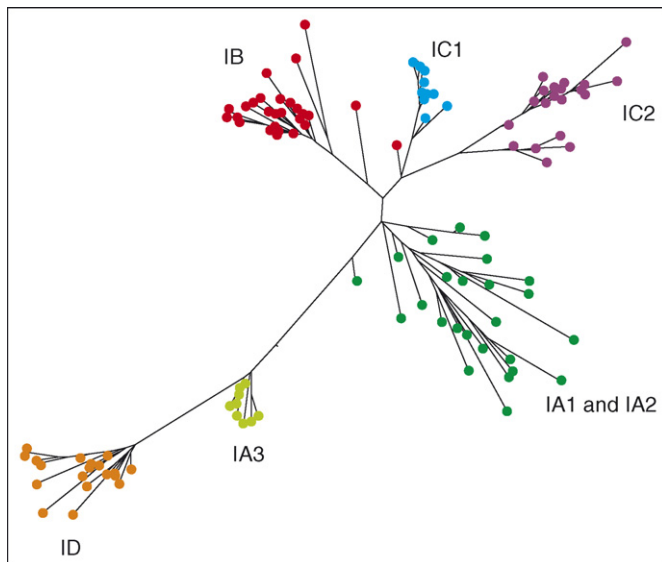


Figure 3. Phylogenetic intron classification using the sequence of the conserved intron core. Members of subgroup IA, IB, IC and ID introns were chosen to represent the diversity within subgroups fully (Table 2). Dots represent individual sequences. The color indicates the subgroup affiliation, according to the traditional classification system of Michel and Westhof [13]. Intron core sequences were aligned, and a phylogenetic analysis was carried out using a maximum-likelihood approach, the Hasegawa–Kishino–Yano (HKY) model of evolution and gamma-distributed among-site variation [47]. ‘Derived introns’, as listed in Table 2, were not included, because they lack sequence elements used in this analysis and are difficult to align in general. The analysis confirms the coherent grouping of IA1, IA3, IC1, IC2 and ID introns, but the separation of IA introns and IB introns is imperfect. (Even less resolution is obtained when sequences from the class of highly derived introns are included, data not shown.).

Minimum consensus structure of mitochondrial group I introns revisited

Another misconception pertains to the structural elements that characterize a ‘typical’ intron and its invariable core. Many of the mitochondrial introns reported in the past 20 years lack certain structural elements, including P2, P8, P9 and P10 [10] (Figure 2a). (For more examples, see the collection of secondary structure diagrams in GOBASE.) But these elements have been considered by some researchers [19] to be integral to ‘typical’ group I introns; P8 was even thought to be a constituent of the intron core. A synthesis of the data compiled here contradicts previous views of what defines a group I intron: a modified consensus secondary structure of mitochondrial group I introns includes P1 but lacks P8 (Figure 2b). It should be noted that many introns are structurally reduced to a degree that their classification into the established subgroups remains ambiguous (Table 2, Group-I-derived row), although they clearly belong in group I.

Classification of mitochondrial group I introns using phylogenetics

Traditionally, group I introns have been divided into four main subgroups (A, B, C and D) plus further subdivisions (indicated by numerals) [13]; recently, an additional subgroup (E), which occurs in nuclear rRNA genes, has been proposed [35]. Despite this, intron classification has been rather subjective, based on sequence variation of the intron core and supplemented by diverse architectural features of peripheral domains selected according to the authors’ view of their importance. The original method for measuring

sequence variation was principle component analysis (PCA) clustering based on a sequence distance matrix [13]. Subsequently, likelihood methods, which have more resolution power than distance methods, became available [36]. In fact, the application of a maximum-likelihood approach, especially when including among-site variation of evolutionary rates, provides superior separation into intron subgroups. Our likelihood tree (Figure 3) is based on only the intron core sequence. It is consistent with the original classification, except for the separation of subgroup IA3 from IA1 and IA2: mitochondrial introns are divided into the established four (and potentially more) main subgroups. Subgroup IB contains most mitochondrial group I introns, and subgroup IA and subgroup IB are less well distinguished than subgroups IA3, IC1, IC2 and ID (Figure 3).

This phylogenetic analysis confirms that primary structure conservation is an important factor in classifying group I introns. However, subgroup division is not supported by statistically rigorous likelihood ratio tests such as AU [37], owing to the small number (123) of available informative sequence positions. It is evident that we need an objective way of adding higher-order features of the RNA structure (e.g. helices and absence or presence of structural elements) to this phylogenetic analysis, and this is a promising avenue for future development.

Conclusions

Since their initial identification in 1982, much progress has been made in understanding group I introns, in particular with regard to autocatalysis and mobility. But not enough attention has been paid to the structural diversity of group I introns, especially those found in mitochondria. The occurrence of structurally streamlined introns and mini-introns raises the question of whether these are exceptions that have evolved in only a few species only or whether these reflect an overall trend in intron evolution. We favor the latter view for the following reasons: first, the fungal taxa in which group I mini-introns are found are phylogenetically distant from one another and have close relatives that carry typical introns; second, introns that have reduced secondary structure but normal size occur in various fast-evolving eukaryotes, including fungi, amoebae and green algae. Given the apparent reductive trend in the evolution of mitochondrial group I introns, these introns provide a unique opportunity to investigate the progressive substitution of RNA structural elements with proteins.

Acknowledgements

We thank F. Michel and the anonymous reviewers for comments on the manuscript, and we acknowledge support from the Canadian Institutes of Health Research (grant numbers MOP 15331, 42475), the Canada Research Chairs Program (to B.F.L.) and the Canadian Institute for Advanced Research (to B.F.L. and G.B.).

References

- 1 Michel, F. *et al.* (1982) Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure. *Biochimie* 64, 867–881
- 2 Anziano, P.Q. *et al.* (1982) Functional domains in introns: *trans*-acting and *cis*-acting regions of intron 4 of the *cob* gene. *Cell* 30, 925–932

- 3 Waring, R.B. *et al.* (1982) Internal structure of a mitochondrial intron of *Aspergillus nidulans*. *Proc. Natl. Acad. Sci. U. S. A.* 79, 6332–6336
- 4 Michel, F. and Dujon, B. (1983) Conservation of RNA secondary structures in two intron families including mitochondrial, chloroplast- and nuclear-encoded members. *EMBO J.* 2, 33–38
- 5 Bonen, L. and Vogel, J. (2001) The ins and outs of group II introns. *Trends Genet.* 17, 322–331
- 6 Saldanha, R. *et al.* (1993) Group I and group II introns. *FASEB J.* 7, 15–24
- 7 Cech, T.R. *et al.* (1983) Secondary structure of the *Tetrahymena* ribosomal RNA intervening sequence: structural homology with fungal mitochondrial intervening sequences. *Proc. Natl. Acad. Sci. U. S. A.* 80, 3903–3907
- 8 Costa, M. *et al.* (1997) Multiple tertiary interactions involving domain II of group II self-splicing introns. *J. Mol. Biol.* 267, 520–536
- 9 Michel, F. and Ferat, J.L. (1995) Structure and activities of group II introns. *Annu. Rev. Biochem.* 64, 435–461
- 10 Michel, F. and Westhof, E. (1990) Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216, 585–610
- 11 Toor, N. *et al.* (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* 7, 1142–1152
- 12 Cannone, J.J. *et al.* (2002) The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3, 2 DOI: 10.1186/1471-2105-3-2 (www.biomedcentral.com/bmcbioinformatics)
- 13 Michel, F. and Westhof, E. (1990) Modeling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* 216, 585–610
- 14 Dujon, B. *et al.* (1986) Mitochondrial introns as mobile genetic elements: the role of intron-encoded proteins. *Basic Life Sci.* 40, 5–27
- 15 Lambowitz, A.M. and Zimmerly, S. (2004) Mobile group II introns. *Annu. Rev. Genet.* 38, 1–35
- 16 Van der Veen, R. *et al.* (1986) Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing *in vitro*. *Cell* 44, 225–234
- 17 Schmelzer, C. and Schweyen, R.J. (1986) Self-splicing of group II introns *in vitro*: mapping of the branch point and mutational inhibition of lariat formation. *Cell* 46, 557–565
- 18 Jacquier, A. and Rosbash, M. (1986) Efficient *trans*-splicing of a yeast mitochondrial RNA group II intron implicates a strong 5' exon–intron interaction. *Science* 234, 1099–1104
- 19 Haugen, P. *et al.* (2005) The natural history of group I introns. *Trends Genet.* 21, 111–119
- 20 Schäfer, B. *et al.* (1991) The mitochondrial genome of fission yeast: inability of all introns to splice autocatalytically, and construction and characterization of an intronless genome. *Mol. Gen. Genet.* 225, 158–167
- 21 Wallweber, G.J. *et al.* (1997) Characterization of *Neurospora* mitochondrial group I introns reveals different CYT-18 dependent and independent splicing strategies and an alternative 3' splice site for an intron ORF. *RNA* 3, 114–131
- 22 Bousquet, I. *et al.* (1990) Two group I mitochondrial introns in the *cob*-*box* and *coxI* genes require the same MRS1/PET157 nuclear gene product for splicing. *Curr. Genet.* 18, 117–124
- 23 Kreike, J. *et al.* (1987) A yeast nuclear gene, *MRS1*, involved in mitochondrial RNA splicing: nucleotide sequence and mutational analysis of two overlapping open reading frames on opposite strands. *EMBO J.* 6, 2123–2129
- 24 O'Brien, E.A. *et al.* (2006) GOBASE – a database of organelle and bacterial genome information. *Nucleic Acids Res.* 34, D697–D699
- 25 Lisacek, F. *et al.* (1994) Automatic identification of group I intron cores in genomic DNA sequences. *J. Mol. Biol.* 235, 1206–1217
- 26 Griffiths-Jones, S. *et al.* (2003) Rfam: an RNA family database. *Nucleic Acids Res.* 31, 439–441
- 27 Cummings, D.J. *et al.* (1989) DNA sequence analysis of the 24.5 kilobase pair cytochrome oxidase subunit I mitochondrial gene from *Podospira anserina*: a gene with sixteen introns. *Curr. Genet.* 16, 381–406
- 28 Gautheret, D. and Lambert, A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* 313, 1003–1011
- 29 Smith, S.W. *et al.* (1994) The genetic data environment: an expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.* 10, 671–675
- 30 Badidi, E. *et al.* (2003) AnaBench: a Web/CORBA-based workbench for biomolecular sequence analysis. *BMC Bioinformatics*, 4, 63 DOI: 10.1186/1471-2105-4-63 (www.biomedcentral.com/bmcbioinformatics)
- 31 Nielsen, H. *et al.* (2005) An mRNA is capped by a 2', 5' lariat catalyzed by a group I-like ribozyme. *Science* 309, 1584–1587
- 32 Einvik, C. *et al.* (1998) Group I-like ribozymes with a novel core organization perform obligate sequential hydrolytic cleavages at two processing sites. *RNA* 4, 530–541
- 33 Bassi, G.S. and Weeks, K.M. (2003) Kinetic and thermodynamic framework for assembly of the six-component b13 group I intron ribonucleoprotein catalyst. *Biochemistry* 42, 9980–9988
- 34 Lambowitz, A.M. and Perlman, P.S. (1990) Involvement of aminoacyl-tRNA synthetases and other proteins in group I and group II intron splicing. *Trends Biochem. Sci.* 15, 440–444
- 35 Li, Z. and Zhang, Y. (2005) Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup. *Nucleic Acids Res.* 33, 2118–2128
- 36 Kuhner, M.K. and Felsenstein, J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11, 459–468
- 37 Shimodaira, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508
- 38 Galburt, E.A. and Stoddard, B.L. (2002) Catalytic mechanisms of restriction and homing endonucleases. *Biochemistry* 41, 13851–13860
- 39 Belfort, M. and Roberts, R.J. (1997) Homing endonucleases: keeping the house in order. *Nucleic Acids Res.* 25, 3379–3388
- 40 Belfort, M. and Perlman, P.S. (1995) Mechanisms of intron mobility. *J. Biol. Chem.* 270, 30237–30240
- 41 Plessis, A. *et al.* (1992) Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. *Genetics* 130, 451–460
- 42 Costa, M. *et al.* (1997) Rules for RNA recognition of GNRA tetraloops deduced by *in vitro* selection: comparison with *in vivo* evolution. *EMBO J.* 16, 3289–3302
- 43 Borner, G.V. *et al.* (1995) RNA editing of a group II intron in *Oenothera* as a prerequisite for splicing. *Mol. Gen. Genet.* 246, 739–744
- 44 Eskes, R. *et al.* (2000) Multiple homing pathways used by yeast mitochondrial group II introns. *Mol. Cell. Biol.* 20, 8432–8446
- 45 Toor, N. and Zimmerly, S. (2002) Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. *RNA* 8, 1373–1377
- 46 Hausner, G. *et al.* (2006) Origin and evolution of the chloroplast *trnK* (*matK*) intron: a model for evolution of group II intron RNA structures. *Mol. Biol. Evol.* 23, 380–391
- 47 Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704

Reproduction of material from Elsevier articles

Interested in reproducing part or all of an article published by Elsevier, or one of our article figures? If so, please contact our *Global Rights Department* with details of how and where the requested material will be used. To submit a permission request online, please visit:

www.elsevier.com/locate/permissions