

A conserved 3' extension in unusual group II introns is important for efficient second-step splicing

Fredrik B. Stabell, Nicolas J. Tourasse and Anne-Brit Kolstø*

Laboratory for Microbial Dynamics (LaMDa), Department of Pharmaceutical Biosciences, University of Oslo, Oslo, Norway

Received November 19, 2008; Revised March 5, 2009; Accepted March 6, 2009

ABSTRACT

The *B.c.I4* group II intron from *Bacillus cereus* ATCC 10987 harbors an unusual 3' extension. Here, we report the discovery of four additional group II introns with a similar 3' extension in *Bacillus thuringiensis kurstaki* 4D1 that splice at analogous positions 53/56 nt downstream of domain VI *in vivo*. Phylogenetic analyses revealed that the introns are only 47–61% identical to each other. Strikingly, they do not form a single evolutionary lineage even though they belong to the same Bacterial B class. The extension of these introns is predicted to form a conserved two-stem-loop structure. Mutational analysis *in vitro* showed that the smaller stem S1 is not critical for self-splicing, whereas the larger stem S2 is important for efficient exon ligation and lariat release in presence of the extension. This study clearly demonstrates that previously reported *B.c.I4* is not a single example of a specialized intron, but forms a new functional class with an unusual mode that ensures proper positioning of the 3' splice site.

INTRODUCTION

Group II introns are self-splicing ribozymes that are able to excise themselves from precursor mRNA transcripts. They are also retroelements which encode a multifunctional reverse-transcriptase (RT) open reading frame (ORF) and through reverse-splicing they are able to invade new DNA locations (1–5). They are found in the genomes of bacteria, archaea and eukaryotic organelles. Phylogenetically, group II introns can be divided into several major subfamilies based on RNA secondary structure features and ORF sequences [Figure 3A; (1,5–7)]. The secondary structure of group II intron RNA consists of six domains that are linked by a network of tertiary in-

te-
reactions (2,8–10). In particular, domain I forms the scaffold for intron assembly and domain V is essential for catalysis. The other structural elements are important for compaction, stabilization and/or catalysis. Group II intron splicing proceeds through two transesterification reactions. The first reaction is mediated via nucleophilic attack on the 5' intron–exon junction either by the 2' hydroxyl group of the bulged adenosine in domain VI or by water. Subsequently the flanking exons are ligated and a branched intron lariat or a linear intron form is respectively released (1–3).

We previously showed that a group II intron *B.c.I4*, from *Bacillus cereus* ATCC 10987, has unusual properties by splicing 56 nt downstream of the predicted 3' splice site (11,12). *In vivo* and *in vitro* analyses revealed that this intron harbors a 3' extension that is a part of the RNA molecule that splices out. In addition, these studies showed that *B.c.I4* has adapted to splice with the extra element, as the splicing efficiency *in vitro* is slightly better than that of a construct deleted of the 3' extension, and this extra substructure has been referred to as a domain VII (13). In this study we report four new group II introns, *B.th.I5*, *B.th.I6*(a and b) and *B.th.I7* from *Bacillus thuringiensis* 4D1 that harbor a 3' extension similar to that of *B.c.I4*. *Bacillus cereus* and *B. thuringiensis* are genetically very closely related and are members of the *B. cereus* group of bacteria (14,15). The extensions of all these introns form two conserved stem-loop secondary structures and *in vitro* mutagenesis showed that the larger of the two stems is needed for an efficient second-step splicing with the extension.

MATERIAL AND METHODS

Bioinformatic searches

The unusual group II introns *B.th.I5*, *B.th.I6*(a and b) and *B.th.I7* in *B. thuringiensis kurstaki* 4D1 were identified in preliminary genome sequence data (Økstad, O.A. and Nederbragt, L., University of Oslo, Norway, unpublished

*To whom correspondence should be addressed. Tel: +47 22 85 69 23; Fax: +47 22 84 49 44; Email: a.b.kolsto@farmasi.uio.no

data) or in fragments from private sequence data (Papazisi, L. and Peterson, S.N., J. Craig Venter Institute, USA, unpublished data), using BLASTN (16) search with the *B.c.I4* group II intron and its 56-nt 3' extension as query. The extensions of *B.th.I5*, *B.th.I6(a and b)* and *B.th.I7* were also used to search the Genbank and EMBL nucleotide sequence databases. BLASTN was run using two sets of parameters, one using lowered gap penalties (opening cost $G = 2$ and extension cost $E = 1$), and the other using increased reward for nucleotide match (match reward, $r = 2$). Other parameters set to nondefault values were: word size of 7 ($W = 7$), E -value of 1 ($e = 1$), and no filtering of low-complexity regions ($F = F$).

For structural searches, domains V and VI of the bacterial group II intron sequences available at the Group II intron database [<http://www.fp.ucalgary.ca/group2introns/>; (17)] were used as queries to search the Genbank and EMBL databases using BLASTN. The hits were extracted along with 150 bases of downstream flanking sequence. This set of sequences was then scanned using the RNAMotif program (18) with a descriptor representing the 3' extensions of the unusual introns.

Secondary structure predictions

The secondary structures of the *B.th.I5*, *B.th.I6* and *B.th.I7* intron RNAs (ORF removed) were computationally predicted by constrained folding using the MFOLD 3.1 package (19) following the consensus structures of group IIB (B class) introns (6,11). That is, conserved and identifiable sequence motifs corresponding to the consensus structures were forced during the folding computation.

Phylogenetic analysis

Bacterial group II intron sequences were taken from the Group II intron database (17). Amino acid sequences of the ORFs of these introns and the new *B. thuringiensis kurstaki* 4D1 introns were aligned using CLUSTALW (20) followed by manual correction. An unrooted phylogenetic tree of a total of 221 introns was reconstructed using the maximum likelihood method as in ref. 21, based on all the RT domains, by means of the program RAXML 7.0.4 with the amino acid substitution model RtREV + Γ + F (22). After removing ambiguously aligned regions, the alignment contained 221 amino acid sites. Statistical support for the groupings in the tree was assessed using 1000 bootstrap replicates (23). The same procedure was employed to build a tree of B class introns only, except that in this case the full-length intron-encoded ORFs could be aligned (438 amino acid positions).

DNA and RNA isolation

Bacillus thuringiensis kurstaki 4D1 was grown on Luria Bertani (LB) agar plates at pH 7 and 30°C. An overnight culture (16 h) was inoculated for 3.5 h in 10 ml LB, and then cells were lysed with 10 mg/ml lysozyme. DNA isolation was performed using the Genomic DNA Midi kit (Qiagen) as described by the supplier. Total RNA isolation was conducted as in (12).

PCR and RT-PCR

PCR and RT-PCR were performed as described in (11), with the exception that the annealing temperature was set to 59°C for PCR. A listing of all the primers used in this study is given in Supplementary Table 1.

Cloning and site-directed mutagenesis

RT-PCR products, either taken directly or gel purified from 1× TAE gel (QIAquick gel extraction Kit, Qiagen), were cloned into TA cloning vector (Invitrogen) and subsequently sequenced.

B.th.I5 and *B.th.I6a* were cloned into pBluescript KS+ or TA-topo vector pCRII respectively, using primers *B.th.I5_right/left* and *B.th.I6a_right/left*, based on sequence fragments of *B. thuringiensis kurstaki* 4D1 and orthologous genes in *Bacillus anthracis* or *B. cereus* ATCC 10987. The intron-containing inserts were then amplified by PCR with outward primers, *B.th.I5dORF_right/left* and *B.th.I6adORF_right/left*, in order to remove the ORF encoded in domain IV.

Site-directed mutagenesis to generate point mutation and deletion constructs was performed with Quikchange II (Stratagene) according to the manufacturer's instructions using two complementary oligonucleotides (of ~40 bases) containing the desired mutation(s) with *B.c.I4*, *B.th.I5* or *B.th.I6a* Δ ORF as a template (11). Primers are listed in Supplementary Table 1. All constructs were verified by sequencing.

In vitro transcription

One microgram of plasmid construct was linearized by XhoI for transcription reactions with 30 U T7 or Sp6 RNA polymerase (Ambion) according to the manufacturer's instructions. Transcription and gel-purification of radiolabeled and unlabeled RNA were conducted as previously described (11).

In vitro self-splicing of ribozyme

In vitro generated transcripts were denatured and refolded using a GenAmp 2700 PCR machine (Applied Biosystems), by incubating the transcripts in 10 mM MOPS, pH 7.5 at 90°C for 1 min, 75°C for 5 min, and then slow cooling to the splicing temperature at 47°C. Intron transcripts were spliced with 70 000 c.p.m. RNA or ~0.1 μ g unlabeled transcripts in 40 mM MOPS, pH 7.5, 100 mM MgCl₂ and 500 mM (NH₄)₂SO₄ at 47°C. Reactions were initiated by adding prewarmed splicing buffer to the transcript RNA giving a total reaction volume of 40 μ l. At each time point, 2 μ l were taken out, quenched with loading buffer (Ambion) and storing samples on dry ice. Samples were then heated to 95°C and cooled on ice, before being separated on a 7 M or 8.5 M Urea 4% polyacrylamide gel. Gels were then vacuum dried, exposed and analyzed using a Molecular Dynamics Storm 860 Phosphorimager.

For subsequent RT-PCR and sequencing of these splicing products, either unlabeled spliced transcripts, purified with Nucleotide purification kit (Qiagen), or labeled

spliced transcript species, excised from gels were used as templates.

For kinetic analysis, the intensities of the radioactive bands were quantified using the ImageQuant 5.0 software and corrected for the number of uridines. The relative fractions of unspliced precursor and free lariat RNA were computed from the intensities of the radioactive bands of all intron-containing products. Data were fitted to a biphasic exponential kinetic model [Equations (6) and (8) in ref. 24] by the nonlinear least squares method using the GNU gretl 1.6.5 software (<http://gretl.sourceforge.net/>).

Sequence availability

The nucleotide sequences of the *B.th.*I5, *B.th.*I6a, *B.th.*I6b and *B.th.*I7 group II introns have been deposited in the EMBL database under accession numbers FM992108, FM992109, FM992110 and FM992111, respectively.

RESULTS

New group II introns with a 3' extension in *B. thuringiensis kurstaki* 4D1

Through a sequence similarity search of private sequence collections using BLAST (16) and the 3' extension of the unusual *B.c.*I4 intron of *B. cereus* ATCC 10987 (11,12) as query, four sequence fragments exhibiting similarity to the *B.c.*I4 extension were identified in *B. thuringiensis kurstaki* BGSC 4D1 (also known as AH248 or KB004). Further cloning, sequencing and computational secondary structure predictions revealed that all of these sequences contained a full group II intron with the six typical domains and a 3' extension (Figure 1). No additional group II introns with this unusual extension were identified in a similar search of public sequence databases, as well as in a structural search with RNAMotif (18) (see 'Material and Methods' section). Three of the

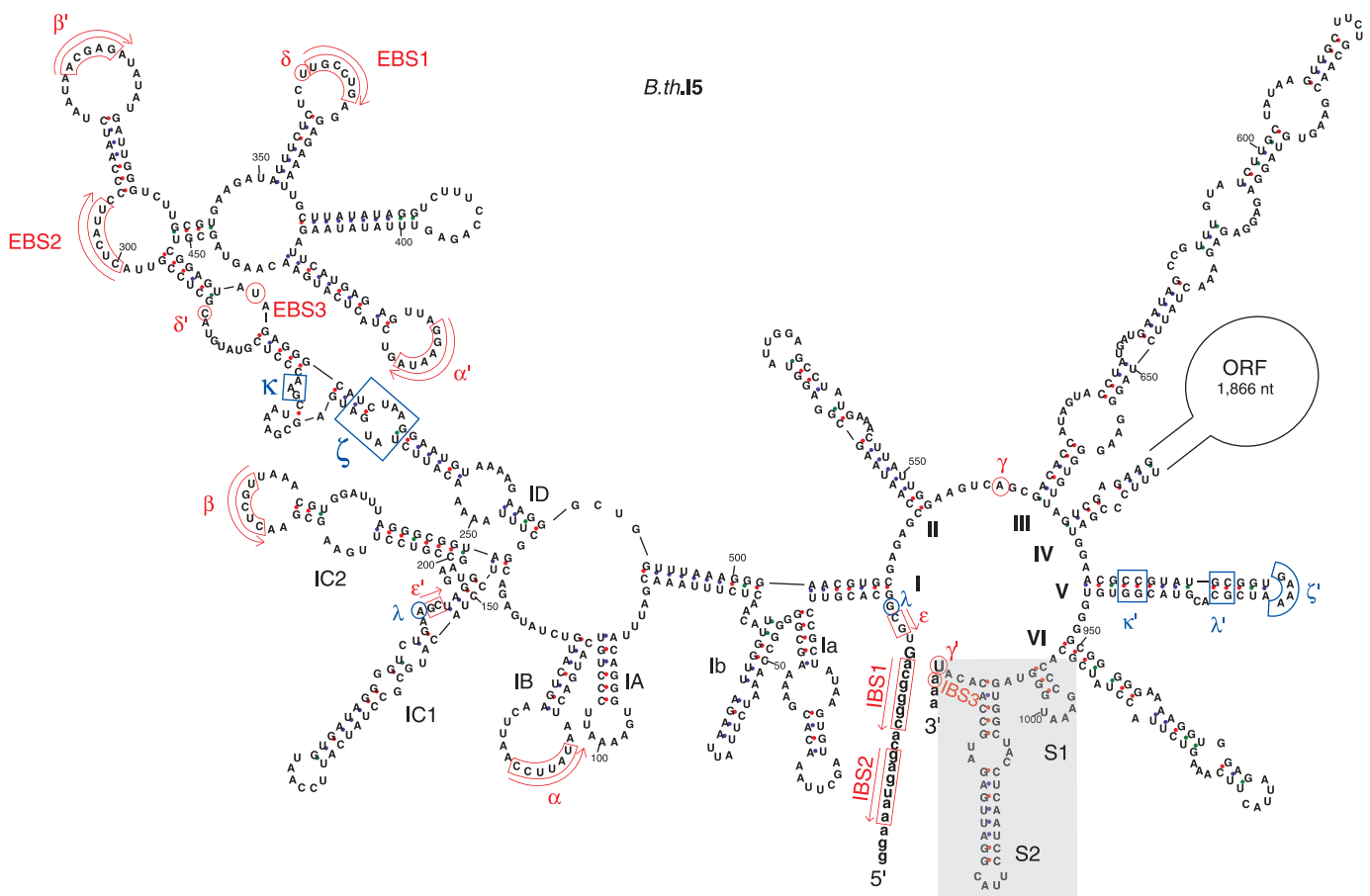


Figure 1. Predicted secondary structures of the *B.th.*I5, *B.th.*I6 and *B.th.*I7 group II introns from *B. thuringiensis kurstaki* 4D1. Exon nucleotides are in lowercase. Roman numerals (I to VI) indicate the six typical functional RNA domains, and subdomains within domain I are designated following the nomenclature of (40). The extra 53/54-nt 3' segment harbored by the three introns is boxed in gray. Sites corresponding to consensus positions involved in tertiary interactions (6,41) are indicated by pairs of Greek letters or EBS/IBS (exon/intron-binding sites). Sites of tertiary base-pairing interactions are boxed or circled in red with arrows indicating the orientation of complementarity. Sites implicated in other tertiary contacts are boxed or circled in blue. For *B.th.*I5 and *B.th.*I7, the δ' nucleotide was set at the expected location (C:290 and C:332, respectively) according to (26), however it is not complementary to the δ site, while the adenosine 5' of δ' is. Two copies of the *B.th.*I6 intron (*B.th.*I6a and *B.th.*I6b) were found in separate genomic locations and the nucleotide differences in *B.th.*I6b compared to *B.th.*I6a are shown by green boxes and letters. ORF, intron-encoded multifunctional open reading frame. Numbering of residues does not include the ORF. The lengths of the *B.th.*I5, *B.th.*I6a and *B.th.*I7 RNAs (excluding the ORF) are 1044, 936 and 905 nt, respectively (*B.th.*I6b has one extra adenosine compared to *B.th.*I6a). A-U, G-C and G-U base pairs are linked by blue, red and green dots, respectively.

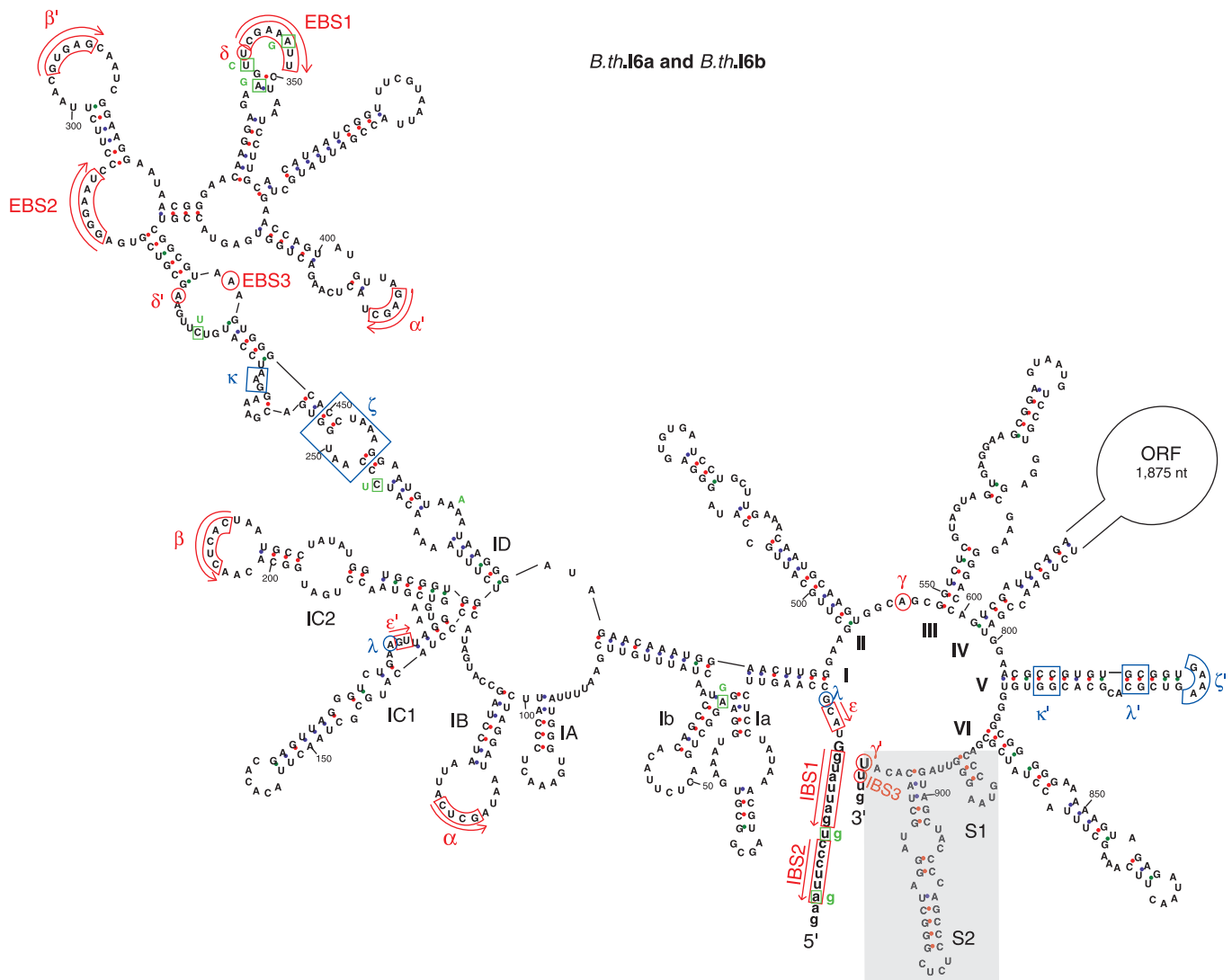


Figure 1. Continued.

B. thuringiensis kurstaki introns were located in genes homologous to genes from the pXO1 plasmid of *B. anthracis* (Table 1). These introns were named (according to the nomenclature used in ref. 25) *B.th.I5*, *B.th.I6a* and *b*. The latter two are inserted in different genomic contigs and homologs of a predicted conjugation gene (~80% nucleotide sequence identity). *B.th.I6a* and *b* are 98.4% identical overall and inserted in the same sites thereby representing two copies of the same intron *B.th.I6*. The remaining intron *B.th.I7* was inserted in a homolog of a hypothetical gene from the pBc10987 plasmid of *B. cereus* ATCC 10987 (Table 1). RT-PCR conducted on total RNA from *B. thuringiensis kurstaki* 4D1 with host gene-specific primers showed that the *B. thuringiensis* introns spliced and thus were functional *in vivo* (Figure 2 and Supplementary Table 1). Sequencing of the RT-PCR products confirmed that the 3' splice sites of *B.th.I5*, *B.th.I6* and *B.th.I7* were located respectively 53, 54 and 54 nucleotides downstream of domain VI, as opposed to the usual three or four bases.

In addition, *in vitro* splicing of *B.th.I5* and *B.th.I6a* confirmed the splice boundaries observed *in vivo* for these introns (see below). Furthermore, as in the *B.c.I4* intron, potential EBS3-IBS3 and γ - γ' base-pair interaction sites, which are important for 3' splice site selection (2,26,27) could be identified at the observed 3' splice site (Figure 1). *B.th.I5*, *B.th.I6* and *B.th.I7* therefore represent new examples of bacterial group II introns carrying a 3' extension. Overall, the four different *B. cereus* and *B. thuringiensis* introns are only 47–61% identical at the nucleotide level (31–52% amino acid sequence identity between the ORFs). Phylogenetic analysis of the ORFs of 221 bacterial group II introns available at the group II intron database (17) revealed that these four unusual introns belong to the B class (according to the nomenclature mentioned in ref. 6; Figure 3). However, they do not group in a single lineage but are located in two subgroups herein named α and β . *B.c.I4* and *B.th.I7* belong to subgroup α , while *B.th.I5* and *B.th.I6* are in subgroup β . The division of the four introns harboring an extra 3'

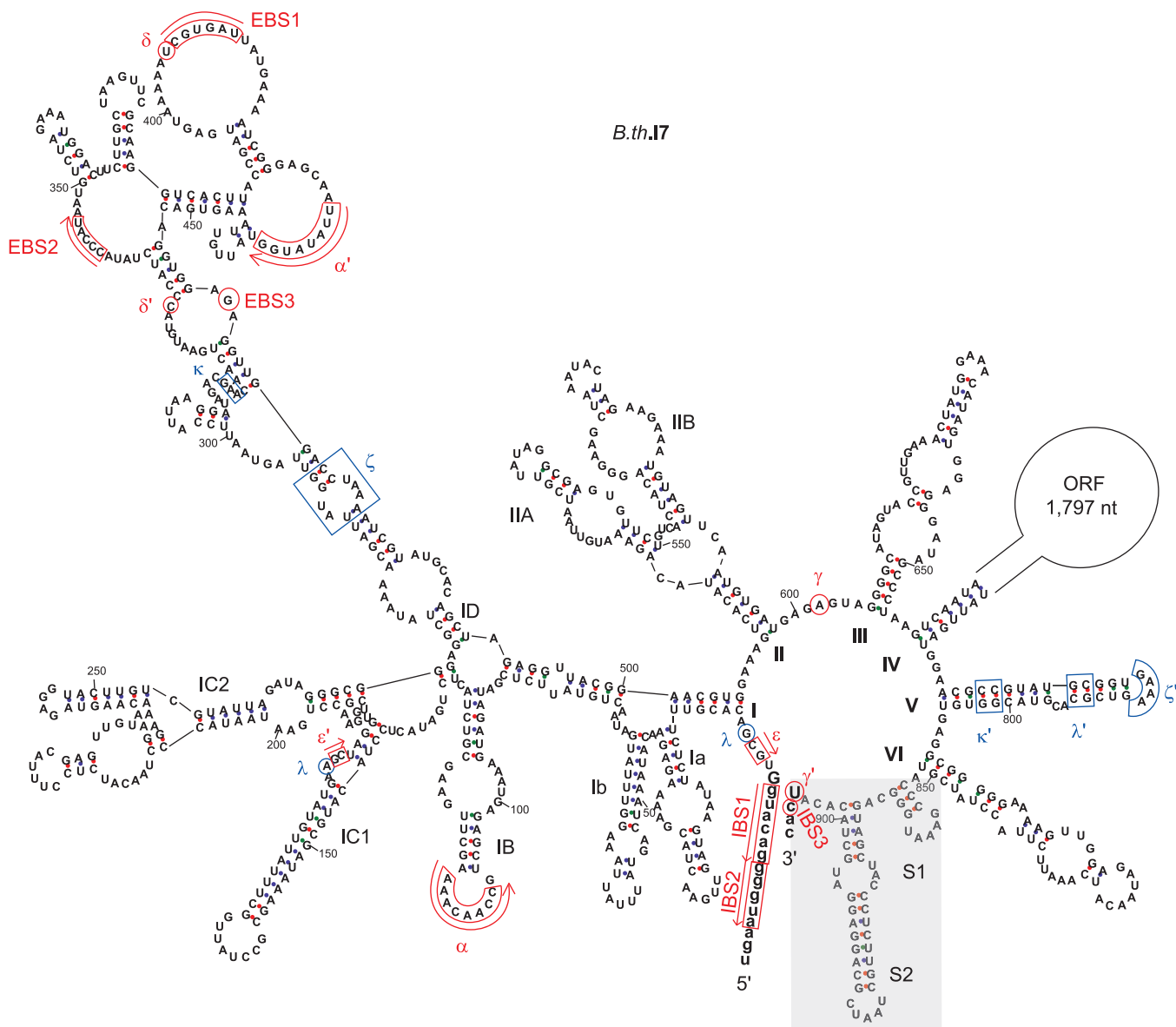


Figure 1. Continued.

segment based on ORF sequence relatedness is supported by the fact that the introns share structural features common to each subgroup (Supplementary Figure 1). These features are all located in domain I of the RNA secondary structure. Despite this divergence, sequence and secondary structure comparisons done both manually and using RNAForester (28,29) revealed that all four introns share several conserved regions (see Supplementary Figures 2 and 3). Besides domains V and VI, which are highly conserved, several nucleotides within the 3' extra segment are identical (marked in red in Figure 4A). The 3' extension of the *B. thuringiensis* introns could fold into two stem-loop structures (S1 and S2) similar to those in *B.c.I4* of *B. cereus*, where the most conserved sites in structure and sequence are within the small stem S1 and the asymmetric internal loop of the longer stem S2. The sequence and folding conservation, together

with the occurrence of compensatory mutations in S2, strongly suggest that the 3' extension forms a stable structure downstream of domain VI for these four unusual introns, and thus might indicate the importance of maintaining this structure for intron function. Remarkably, the S2 internal loop and its surrounding base pairs show a striking resemblance to, and matches the consensus of, the 11-nt tetraloop receptor motif 5' [CCUAAG...UAUGG] 3' (30). This is a common RNA motif that participates in the tertiary folding of several catalytic RNAs by interacting with tetraloops of the generic GN_n/RA family (30–32). In addition to the 3' end, there is a high sequence and structure conservation in the stem of subdomain IC1 in *B.c.I4*, *B.th.I5* and *B.th.I6*, while *B.th.I7* shows a lower sequence conservation (see Figures 1 and 4F). Intriguingly, the conserved area is contiguous to the bulged region containing the ϵ' - and λ sites (z-anchor)

Table 1. Unusual group II introns with 3' extensions in *B. cereus* and *B. thuringiensis*

Species and strains	Intron name	Intron ORF domains ^a	Intron ORF length (bp/aa)	Intron length (bp)	Intron ORF phylogenetic class ^b	Host gene	Closest intron relative (% aa identity) ^c
<i>B. cereus</i> ATCC 10987	<i>B.c.14</i>	RT-X-En	1884/627	2843	B	BCEA0036 + BCEA0033; hypothetical protein (DNA primase domain)	<i>Clostridium perfringens</i> CPE F4969 (BAE79013, 61%)
<i>B. thuringiensis</i> <i>kurstaki</i> 4D1	<i>B.th.15</i>	RT-X-En	1866/621	2910	B	Homolog of pXO1-08 from <i>B. anthracis</i> , hypothetical protein with two helicase domains	<i>Bacillus sp.</i> EA1 (ABN04186, 40%)
	<i>B.th.16a</i> ^d	RT-X-En	1875/624	2811	B	Homolog of pXO1-42 from <i>B. anthracis</i> , conjugation protein of the traG/traD family	<i>Geobacillus sp.</i> WCH70 (EDT35839, 41%)
	<i>B.th.16b</i> ^d	RT-X-En	1875/624	2812	B	Homolog of pXO1-42 from <i>B. anthracis</i> , conjugation protein of the traG/traD family	<i>Geobacillus sp.</i> WCH70 (EDT35839, 41%)
	<i>B.th.17</i>	RT-X-En	1797/598	2702	B	Homolog of BCEA0037 gene from <i>B. cereus</i> ATCC 10987, encoding a hypothetical protein.	<i>B. thuringiensis</i> <i>kurstaki</i> HD73 (AAZ06578, 47%)

^aRT, reverse transcriptase domain; X, maturase (splicing) domain; En, endonuclease domain.

^bAccording to the classification of Toor *et al.* (6). See also Figure 3.

^cGenbank accession numbers and amino acid sequence identity between intron ORFs are given in parentheses (top hit of BLAST search of the Genbank database).

^d*B.th.16a* and *b* are overall 98.4% identical at the nucleotide sequence level and represent two copies of the same intron, *B.th.16*, inserted in different genomic locations.

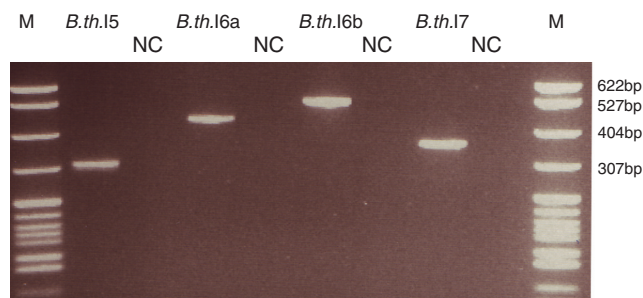


Figure 2. *In vivo* splicing of unusual group II introns in *B. thuringiensis* *kurstaki* 4D1. RT-PCR was conducted on total RNA with primers located in the flanking exons. The gel picture shows the RT-PCR products of the spliced exons (names of the products related to each intron are given on top) and the corresponding negative controls run without reverse transcriptase (lanes marked with NC). Lane M, pBR322 DNA digested with MspI (New England Biolabs), as marker. Samples were separated on a 2.8% NuSieve GTG agarose gel (Cambrex).

that form interactions with the 5' end of the intron and/or domain V (9,33).

Mutational analysis of the *B.c.14* intron's 3' extension

Since the extra 3' element is conserved in structure and partially in sequence between the four *B. cereus* and *B. thuringiensis* introns described here, we conducted an *in vitro* mutational analysis of this element in *B.c.14* in order to investigate whether it contributes to the splicing activity of the intron. *In vitro* splicing was conducted under the same conditions as in (11), i.e. at 47°C in

0.5 M ammonium sulfate ((NH₄)₂SO₄), 40 mM MOPS, pH = 7.5 and 100 mM MgCl₂ (see 'Material and Methods' section). First, two deletion mutants were made by removing separately each of the two stem-loop structures of the 3' extension (S1 and S2) from the *B.c.14* wild-type (WT) ΔORF construct (Figure 4B and C). For the mutant dS2 in which the longer S2 stem was deleted, there was a drastic reduction of the amount of free lariat formed after the second step of splicing, together with a clear increase of the first step intermediate 'lariat with 3' exon', compared to WT intron (Figure 5, two top bands). A time-course kinetic analysis showed that the fraction of unreacted dS2 precursor RNA was ~20% higher than that of WT, while the relative fraction of lariat released by dS2 was decreased by ~60% on average (Figure 6A and B). Altogether, these results indicate it is mostly the second splicing reaction that is severely slowed down when the S2 stem-loop structure of the 3' extension is removed from the intron. To support this argument no clear band corresponding to the ligated exons could be observed for this deletion mutant, even though RT-PCR analysis revealed that it did occur, suggesting that the efficiency of exon ligation was decreased. To determine whether the observed phenomenon applies for other introns carrying the extensions, corresponding deletions of S2 were performed on the *B. thuringiensis* *B.th.15* and *B.th.16a* intron constructs. For both introns, a drastic reduction in the efficiency of the second splicing step was also observed compared to the WT, thus pointing to a general importance of S2 for the splicing of the unusual introns

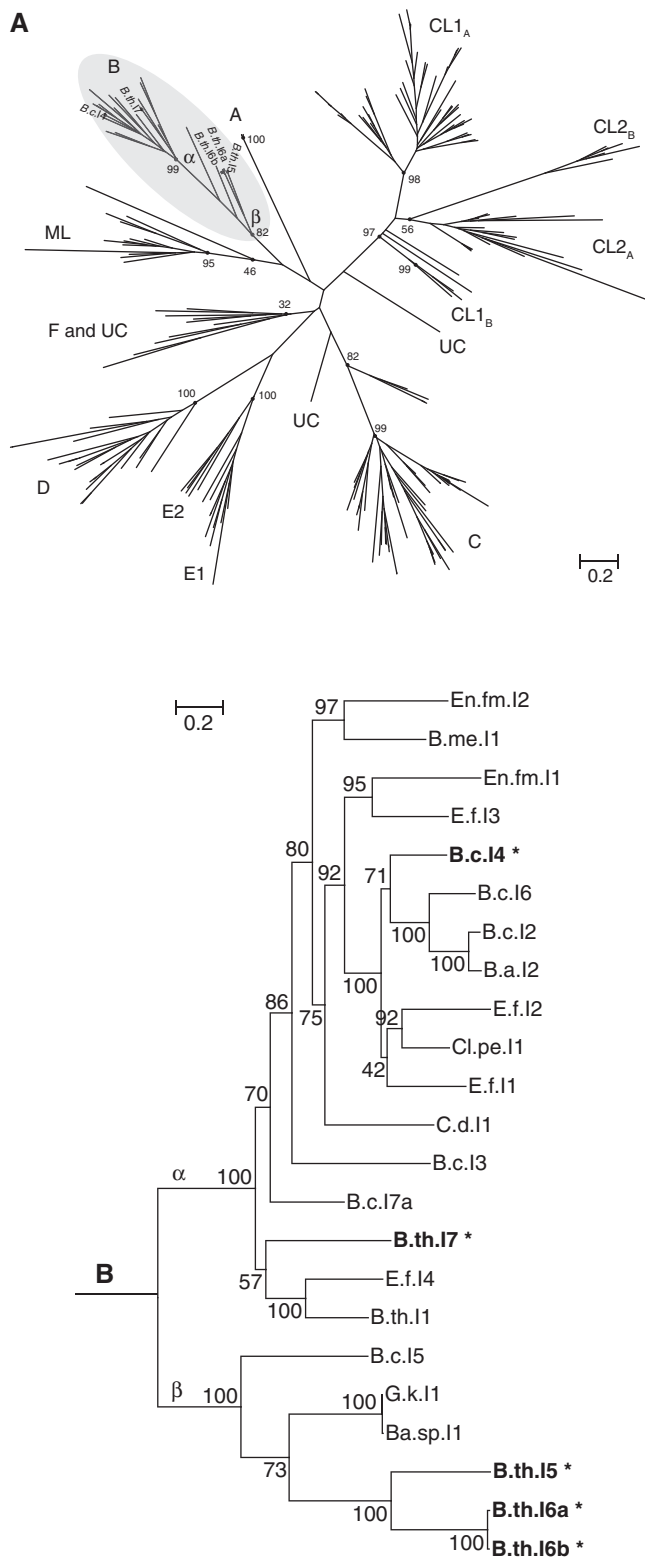


Figure 3. (A) Unrooted phylogenetic tree of 221 bacterial group II intron-encoded proteins. The tree was reconstructed using the maximum-likelihood method (RAxML program) and was based on the RT domains of the proteins. The major intron classes are named as in ref. 21: A-F, ML (mitochondrial-like), CL (chloroplast-like) and UC (unclassified). B-class introns are shaded, with the unusual *B. cereus* and *B. thuringiensis* group II introns carrying the 3' extension indicated by name and a black square. Unlike in the tree as in ref. 21, introns

(Figures 5 and 6A and B, and Supplementary Figure 4). In sharp contrast to the dS2 mutant, the *B.c.I4* construct in which the smaller S1 stem-loop structure was deleted, dS1, showed a splicing efficiency equal or better than that of the WT construct with respect to both the amount of precursor processed and lariat formed (Figure 6A and B). Furthermore, mutating the sequence of the terminal loop of S1 (mutant mS1, Figure 4C) revealed no negative effect on either of the two splicing steps. Therefore, the smaller S1 stem-loop part of the 3' extension does not appear to be critical for splicing under the conditions tested in this study.

To further investigate whether the deficiency in the second step of splicing observed for the dS2 constructs could be due to specific sites within the S2 stem, a number of modifications were made within that stem in *B.c.I4* (Figure 4B) in order to test for structural or sequence-dependent effects. Mutations of the conserved internal asymmetric loop within S2 were first considered. These included closing the internal loop by re-establishing base pairings (mIntclosed mutant), shuffling the sequence of either one or both sides of the loop while preserving its asymmetry (mutants mIntL, mIntR and mIntL/R), and substituting the conserved C922:G951 bp 5' to this loop to a A:U pair (mutant m922_951). Unlike the dS2 construct, none of these modifications had a major effect on the second step of splicing and the rate of lariat formation (Figure 6D). They led to slightly higher fractions of unreacted precursor suggesting a less efficient first step of splicing (Figure 6C). In contrast, mispairing C922:G951 by substituting C:922 with A (mutant m922), which would create a larger internal loop, had a drastic effect with respect to the second splicing step comparable to that of the dS2 construct (~30% decrease in free lariat fraction; Figures 5 and 6C). We then addressed the upper part of S2 possessing a 8–9-bp stem with a terminal (top) loop that appears to be conserved in structure by compensatory mutations in the four introns (Figure 4A). Changing the sequence of the terminal loop of S2 from AAAUA to CACGA (mS2_TL construct; Figure 4B) or shortening the upper stem by 5 bp with or without modifying the internal and terminal loop (dS2_SL and dS2_SS mutants) had little effect on the amount of spliced intron overall, compared to the WT construct (Figure 6C and D), although the rate of the first splicing reaction was

from the CL2_A and CL2_B subclasses are grouped together in the present tree. (B) Detailed rooted phylogenetic tree of the B class group II introns, built the same way as in (A), but based on amino acid sequences covering the full length of the intron encoded ORF. The unusual introns are shown in bold and indicated with asterisks. Information, sequences and secondary structure models of all other introns can be found in the Group II intron database [<http://www.fp.ualgary.ca/group2introns/>; (17)]. Species names are abbreviated as follows: *Ba.sp.*, *Bacillus sp.*; *B.a.*, *Bacillus anthracis*; *B.c.*, *Bacillus cereus*; *B.me.*, *Bacillus megaterium*; *B.th.*, *Bacillus thuringiensis*; *C.d.*, *Clostridium difficile*; *Cl.pe.*, *Clostridium perfringens*; *E.f.*, *Enterococcus faecalis*; *En.fm.*, *Enterococcus faecium*; and *G.k.*, *Geobacillus kaustophilus*. In (A) and (B) numbers next to branch nodes indicate bootstrap support values (in percentage out of 1000 replicates). Scale bars are in average numbers of amino acid substitutions per site. Proposed subgroupings within the B class are labeled α and β .

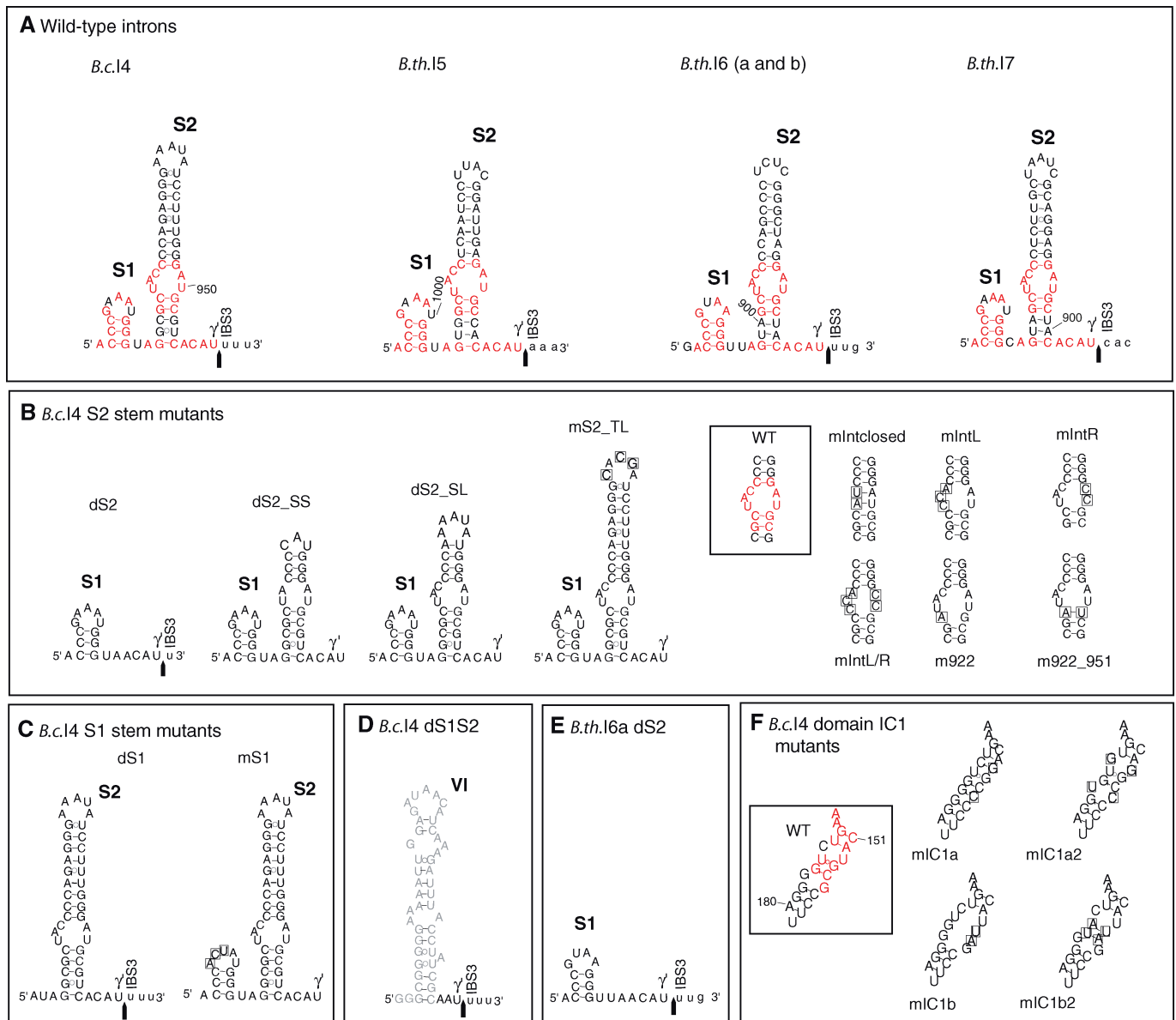


Figure 4. Mutational analysis of the group II introns *B.c.14* from *B. cereus* ATCC 10987 and *B.th.16a* from *B. thuringiensis kurstaki* 4D1. (A) Predicted secondary structure of the extra 53/56-nt 3' segment of *B.c.14* and the newly discovered group II introns *B.th.15*, *B.th.16* (a and b) and *B.th.17*, forming stems S1 and S2. Nucleotides within the 3' extension that are identical between all four introns are colored in red. The observed 3' splice junction is indicated by an arrow. Exon residues are in lowercase. (B) and (C) *B.c.14* constructs mutated in the S2 or S1 stem of the 56-nt 3' extension, respectively. Substituted nucleotides are boxed. (D) *B.c.14* construct deleted of the whole 3' extension (previously named d56; see ref. 11). Domain VI is drawn in gray. (E) *B.th.16a* construct deleted of the S2 stem. (F) *B.c.14* constructs mutated in the IC1 stem of domain I. Nucleotides that are identical between all four introns are colored in red, and substituted nucleotides are boxed. WT, wild-type *B.c.14*.

reduced for dS2_SL and dS2_SS (Supplementary Table 2). This could indicate that the upper section of the S2 stem is less important for splicing under the conditions tested here.

Mutational analysis of the subdomain IC1 of *B.c.14*

The area adjacent to the ϵ' and λ sites in subdomain IC1 is highly conserved in sequence and structure between the *B.c.14*, *B.th.15* and *B.th.16* unusual group II introns, and therefore it was mutated in *B.c.14* in order to assess

its possible role in intron function and whether there could be a relationship with the extra 3' segment. There was no clear effect on the efficiency of either splicing steps when base pairing or changing the conserved mispairs U153:C186 and G156:G183 in the IC1 stem (mutants mIC1a or mIC1a2; Figures 4F, 6E and F). These observations, combined with the fact that these nucleotide positions are not conserved in *B.th.17* imply that they are less important for splicing. However, when mutating both of the two pairs G154:U185 and C155:G184, which are shared between all four introns

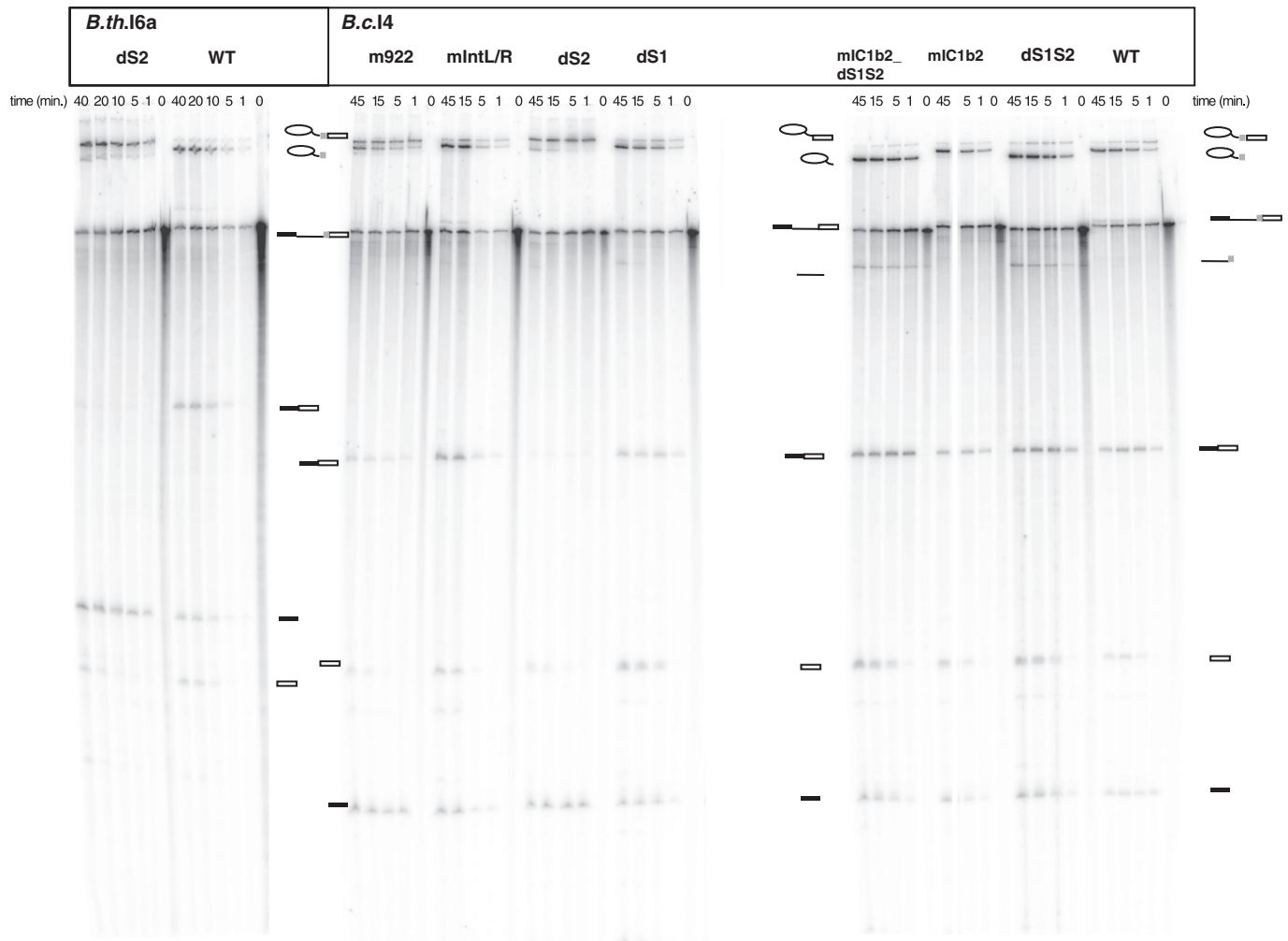


Figure 5. *In vitro* self-splicing of *B.c.I4* and *B.th.I6a* wild-type (WT) and mutant constructs. Splicing was performed in 40 mM MOPS (pH 7.5), 500 mM $(\text{NH}_4)_2\text{SO}_4$ and 100 mM MgCl_2 at 47°C. Samples were separated on a 7 M urea 4% polyacrylamide gel. Schematic drawings are shown next to the bands corresponding to the different splicing products. The light gray box represents the extra 54/56-nt element. Lariat-containing products, precursor and ligated exons were identified by gel excision, and subsequent RT-PCR and sequencing. The linear intron form and free exons were determined by size. The mutants are described in Figure 4.

into U:A pairs (mutants mIC1b and mIC1b2; Figure 4F), a visible reduction in the first splicing reaction was observed as indicated by the fraction of unreacted precursor RNA, which was ~25% higher than for the WT intron (Figure 6E). The fraction of released lariat was also decreased and this could be a consequence of the slower first splicing step. These results are similar to the results of the mutations within the internal loop of S2, although they are somewhat stronger (compare Figure 6C and E). Interestingly, the reduction in the first splicing reaction was abolished when the extra 56-nt 3' element was deleted from the mIC1b2 mutant (mIC1b2_dS1S2).

Finally, it should be noted that sequencing of the spliced exons generated by all the mutant introns used in this study confirmed that they all used the same 5' and 3' splice sites as the WT construct with the 56-nt extension, indicating that none of the mutations created affected the fidelity of the splicing process.

DISCUSSION

In this study we have discovered four additional group II introns from *B. thuringiensis kurstaki* 4D1 with a 53/54-nt 3' extension. This extension is similar in structure and partly in sequence to that of the unusual *B.c.I4* intron identified earlier in *B. cereus* ATCC 10987 (11,12). These four additional introns, named *B.th.I5*, *B.th.I6a* and *b* and *B.th.I7*, all splice *in vivo* in *B. thuringiensis* (Figure 2), and imply that the unusual 3' extension may be common to more group II introns than previously thought. Strikingly, all the introns carrying a 3' extension belong to the same phylogenetic class in terms of ORF sequence (B class); however they do not form a single evolutionary branch within that class (Figure 3). This subgrouping correlates with subtle differences in the RNA secondary structure and is in agreement with the demonstrated coevolution of group II intron structure and ORF (6) (Supplementary Figure 1). Thus, it is not

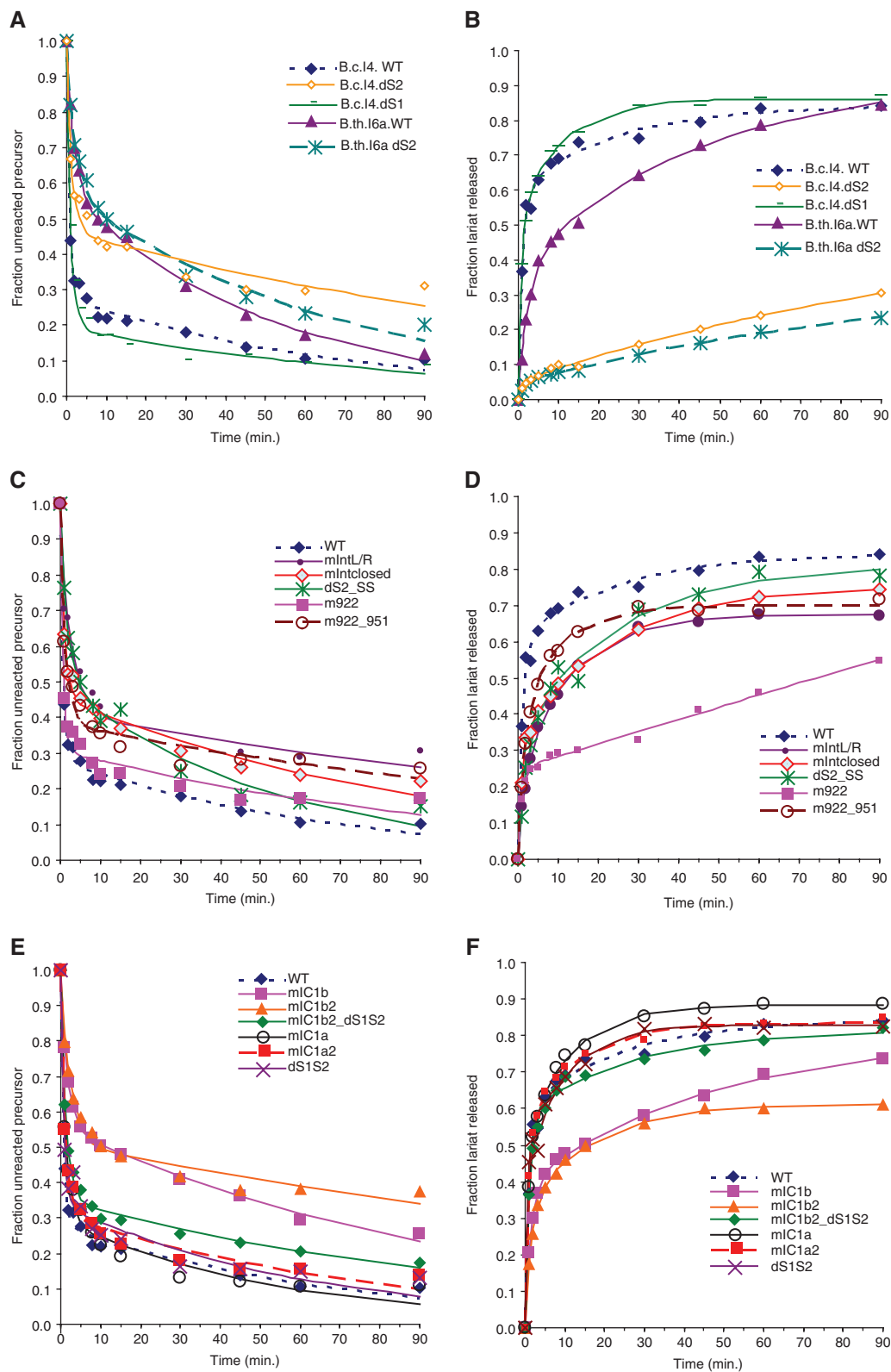


Figure 6. Time-course analysis of *in vitro* self-splicing of *B.c.14* and *B.th.I6a* wild-type (WT) and mutant constructs carrying changes in the 3' extension or in subdomain IC1 (see Figure 4). dS1S2 is a *B.c.14* construct lacking the entire 3' extension (previously named d56; see ref. 11). Splicing was performed in 40 mM MOPS, pH = 7.5, 100 mM MgCl₂ and 500 mM (NH₄)₂SO₄ at 47°C. The relative fractions of unspliced precursor RNA (A, C and E) and released lariat intron (B, D and F) were computed from the intensities of the radioactive bands using a phosphorimager. Reactions were repeated three times for each construct, and are expressed as averages. Data were fitted to a biphasic exponential kinetic model [Equations (6) and (8) in ref. 24, and rate constants are given in Supplementary Table 2].

clear whether the extension was acquired from a common ancestor. The extra segment might represent a kind of mobile element targeting group II introns. Alternatively, the extension may have been acquired independently by each of these introns, or one of the introns ancestral to the B class obtained the extra segment and spread in different environments and bacterial hosts, and still evolved like the other group II introns in the B class. Indeed, several introns in this class appear to be more flexible in their 3' splice-site selection, a property that could have enabled some of them to acquire and adapt to this extra segment (21,34). Interestingly, the fact that *B. thuringiensis kurstaki* 4D1 contains representatives of the two different subgroups of the B class (*B.th.17* and *B.th.15/16* in α and β subgroup, respectively) suggests that these introns have been acquired separately by the bacterium via different mobility events. Furthermore, the nucleotide sequence identity between the *B.th.16a* and *b* introns is significantly higher than the similarity between their homologous host genes (98.4% as opposed to $\sim 80\%$). Unless this is the result of a duplication event and a strong selective pressure to maintain intron sequence and structure, this observation implies a recent mobility event of *B.th.16* with the extra segment within *B. thuringiensis kurstaki* 4D1. It could therefore support the hypothesis of an ancestral origin and spread of a group II intron carrying this 3' extension. All B class introns known to date are in Gram-positive host bacteria and the unusual introns are only found in the *B. cereus* group of bacteria so far. However, the unusual elements are not typical of the group II introns found in this bacterial group as they only represent a small number of them; 5 out of 81 (35). No similar segment could be identified through sequence and secondary structure search of public databases. It cannot be excluded that additional variants of the extension were not detected by the search procedure used in this study. Therefore, it would be necessary and of great interest to find more examples to determine if this is just sampling bias (as half of the B class introns known are from the *B. cereus* group of bacteria; Figure 3B) or if this is due to a specific property of this type of introns or hosts.

The 53/56-nt 3' segment is predicted to fold into two stem-loop structures (S1 and S2) and the observation of compensatory substitutions between the *B. cereus* and *B. thuringiensis* introns strongly suggests that the 3' element forms a stable structure that must be maintained for intron structure and/or activity (Figure 4A). Indeed, deletion of the S2 stem from the *B. cereus B.c.14* intron led to an accumulation of the 'lariat with 3' exon' intermediate and very little exon ligation, i.e. a slower second step of splicing under the conditions tested here (ammonium sulfate splicing buffer; Figures 5 and 6A and B). Similar results were obtained for *B. thuringiensis B.th.15* and *B.th.16a* constructs indicating that this is a general effect. However, the complete deletion of the extension did not affect splicing (11). Therefore, while the 3' extension is not essential for splicing, the S2 stem is important for maintaining an efficient second step of splicing in presence of the extension.

The conserved asymmetric internal loop of S2, which resembles a tetraloop receptor could be a candidate for facilitating the second splicing step through interaction with other parts of the intron. However, mutational analysis of these nucleotides showed no visible impact on the second step of splicing, but rather a moderate reduction of the first step. Thus, these residues do not appear to be directly responsible for the effect observed when deleting S2. Interestingly, a more pronounced slowdown of the first step also occurred when mutating the two conserved base pairs in subdomain IC1 (mutant mIC1b2), an effect that was abolished when removing the whole 3' extension (mutant mIC1b2_dS1S2; Figure 6E and F). The faster first step for the latter mutant could be explained by more splicing through the hydrolytic pathway, as constructs without the extension (mIC1b2_dS1S2 and dS1S2) appear to release more linear form of the intron in addition to the lariat (Figure 5). Another possibility is that the two conserved base pairs in IC1 may be involved in accommodating the extension properly into the intron structure via some interactions. The negative effect on the first splicing step observed when mutating the residues within IC1 or in the S2 internal loop may be a consequence of disrupted interaction(s) and/or subtle changes in RNA structure or conformation which may interfere with elements involved in the first step of splicing, e.g. the coordination loop in domain I with the branch point in domain VI, and the *z*-anchor in subdomain IC1 with the 5' end of the intron (9,33,36–39). The S2 deletion also had an effect on the first step, but the clearly reduced exon ligation and lariat release strongly suggest that this stem is mainly required for efficient 3' splice site recognition with the extension. This could be due to specific interaction sites other than those investigated in this study, and/or to structural constraints of stem S2. The strong effect of mispairing basepair C922:G951, which is predicted to form a larger internal loop within S2, could point to the latter interpretation. Furthermore, in a *trans*-splicing assay conducted with a *B.c.14* construct containing the 5' exon and domains I to VI and one covering the 3' extension and the 3' exon, the ligated exon product was detected by RT-PCR, which may suggest that the 3' extra segment has interacted with the rest of the intron in a way permitting correct splicing (data not shown). Additional thorough mutagenesis and biochemical experiments will be needed in order to reveal and characterize any interaction partners.

The four new group II introns discovered in this study, which carry a 3' extension, show that the previously reported *B.c.14* is not a single example of a specialized intron, but forms a new functional class with an unusual mode of ensuring proper positioning of the 3' splice site. All these introns have a conserved two-stem-loop structure at the 3' end and splice at analogous positions 53/56-nt downstream of domain VI. Mutagenesis showed that the larger stem S2 is important for self-splicing, while the smaller stem S1 is not, and suggests that the S2 stem helps bring the 3' splice site close to the ribozyme's active site. These findings add support to the proposal of the extension as domain VII (13). A surprising

finding was that the five introns do not form a monophyletic group within class B introns. Therefore, the origin of this extension and why the introns maintain it (as it is not essential for splicing) are open questions. Further work is needed to elucidate how the introns have adapted to the extra segment, which would shed light on the structural and functional evolution of these ribozymes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Leka Papazisi and Scott N. Peterson, J. Craig Venter Institute, USA, and Ole Andreas Økstad and Lex Nederbragt, University of Oslo, Norway, for providing unpublished preliminary genomic sequence data. We are grateful to Murali S. Srinivasan for proofreading the text. We thank anonymous referees for their constructive comments.

FUNDING

The Norwegian Functional Genomics (FUGE II) and the Consortium for Advanced Microbial Sciences and Technologies (CAMST) platform of the Research Council of Norway. Funding for open access charge: Norwegian Functional Genomics (FUGE II) platform of the Research Council of Norway and the University of Oslo.

Conflict of interest statement. None declared.

REFERENCES

- Lambowitz, A.M. and Zimmerly, S. (2004) Mobile group II introns. *Annu. Rev. Genet.*, **38**, 1–35.
- Lehmann, K. and Schmidt, U. (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Crit. Rev. Biochem. Mol. Biol.*, **38**, 249–303.
- Pyle, A.M. and Lambowitz, A.M. (2006) In Gesteland, R.F., Cech, T.R. and Atkins, J.F. (eds), *The RNA World*. 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., Vol. Cold Spring Harbor monograph series; 43, pp. 469–505.
- Robart, A.R. and Zimmerly, S. (2005) Group II intron retroelements: function and diversity. *Cytogenet. Genome Res.*, **110**, 589–597.
- Toro, N., Jimenez-Zurdo, J.I. and Garcia-Rodriguez, F.M. (2007) Bacterial group II introns: not just splicing. *FEMS Microbiol. Rev.*, **31**, 342–358.
- Toor, N., Hausner, G. and Zimmerly, S. (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, **7**, 1142–1152.
- Zimmerly, S., Hausner, G. and Wu, X. (2001) Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.*, **29**, 1238–1250.
- Dai, L., Chai, D., Gu, S.Q., Gabel, J., Noskov, S.Y., Blocker, F.J., Lambowitz, A.M. and Zimmerly, S. (2008) A three-dimensional model of a group II intron RNA and its interaction with the intron-encoded reverse transcriptase. *Mol. Cell*, **30**, 1–14.
- Toor, N., Keating, K.S., Taylor, S.D. and Pyle, A.M. (2008) Crystal structure of a self-spliced group II intron. *Science*, **320**, 77–82.
- Waldsich, C. and Pyle, A.M. (2007) A folding control element for tertiary collapse of a group II intron ribozyme. *Nat. Struct. Mol. Biol.*, **14**, 37–44.
- Stabell, F.B., Tourasse, N.J., Ravnum, S. and Kolstø, A.B. (2007) Group II intron in *Bacillus cereus* has an unusual 3' extension and splices 56 nucleotides downstream of the predicted site. *Nucleic Acids Res.*, **35**, 1612–1623.
- Tourasse, N.J., Stabell, F.B., Reiter, L. and Kolstø, A.B. (2005) Unusual group II introns in bacteria of the *Bacillus cereus* group. *J. Bacteriol.*, **187**, 5437–5451.
- Fedorova, O. and Zingler, N. (2007) Group II introns: structure, folding and splicing mechanism. *Biol. Chem.*, **388**, 665–678.
- Rasko, D.A., Altherr, M.R., Han, C.S. and Ravel, J. (2005) Genomics of the *Bacillus cereus* group of organisms. *FEMS Microbiol. Rev.*, **29**, 303–329.
- Tourasse, N.J., Helgason, E., Økstad, O.A., Hegna, I.K. and Kolstø, A.B. (2006) The *Bacillus cereus* group: novel aspects of population structure and genome dynamics. *J. Appl. Microbiol.*, **101**, 579–593.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Dai, L., Toor, N., Olson, R., Keeping, A. and Zimmerly, S. (2003) Database for mobile group II introns. *Nucleic Acids Res.*, **31**, 424–426.
- Macke, T.J., Ecker, D.J., Gutell, R.R., Gautheret, D., Case, D.A. and Sampath, R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
- Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Simon, D.M., Clarke, N.A., McNeil, B.A., Johnson, I., Pantuso, D., Dai, L., Chai, D. and Zimmerly, S. (2008) Group II introns in eubacteria and archaea: ORF-less introns and new varieties. *RNA*, **14**, 1704–1713.
- Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
- Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39**, 783–791.
- Daniels, D.L., Michels, W.J. Jr. and Pyle, A.M. (1996) Two competing pathways for self-splicing by group II introns: a quantitative analysis of *in vitro* reaction rates and products. *J. Mol. Biol.*, **256**, 31–49.
- Dai, L. and Zimmerly, S. (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res.*, **30**, 1091–1102.
- Costa, M., Michel, F. and Westhof, E. (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. *EMBO J.*, **19**, 5007–5018.
- Jacquier, A. and Michel, F. (1990) Base-pairing interactions involving the 5' and 3'-terminal nucleotides of group II self-splicing introns. *J. Mol. Biol.*, **213**, 437–447.
- Höchsmann, M., Töller, T., Giegerich, R. and Kurtz, S. (2003) Local similarity in RNA secondary structures. *Proc. IEEE Comput. Soc. Bioinform. Conf.*, **2**, 159–168.
- Höchsmann, M., Voss, B. and Giegerich, R. (2004) Pure multiple RNA secondary structure alignments: a progressive profile approach. *IEEE ACM Trans. Comput. Biol. Bioinform. IEEE ACM*, **1**, 53–62.
- Costa, M. and Michel, F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J.*, **14**, 1276–1285.
- Abramovitz, D.L. and Pyle, A.M. (1997) Remarkable morphological variability of a common RNA folding motif: the GNRA tetraloop-receptor interaction. *J. Mol. Biol.*, **266**, 493–506.
- Hendrix, D.K., Brenner, S.E. and Holbrook, S.R. (2005) RNA structural motifs: building blocks of a modular biomolecule. *Q. Rev. Biophys.*, **38**, 221–243.
- De Lencastre, A. and Pyle, A.M. (2008) Three essential and conserved regions of the group II intron are proximal to the 5'-splice site. *RNA*, **14**, 11–24.

34. Robart,A.R., Montgomery,N.K., Smith,K.L. and Zimmerly,S. (2004) Principles of 3' splice site selection and alternative splicing for an unusual group II intron from *Bacillus anthracis*. *RNA*, **10**, 854–862.
35. Tourasse,N.J. and Kolstø,A.B. (2008) Survey of group I and group II introns in 29 sequenced genomes of the *Bacillus cereus* group: insights into their spread and evolution. *Nucleic Acids Res.*, **36**, 4529–4548.
36. Hamill,S. and Pyle,A.M. (2006) The receptor for branch-site docking within a group II intron active site. *Mol. Cell*, **23**, 831–840.
37. Toor,N., Rajashankar,K., Keating,K.S. and Pyle,A.M. (2008) Structural basis for exon recognition by a group II intron. *Nat. Struct. Mol. Biol.*, **15**, 1221–1222.
38. Boudvillain,M., de Lencastre,A. and Pyle,A.M. (2000) A tertiary interaction that links active-site domains to the 5' splice site of a group II intron. *Nature*, **406**, 315–318.
39. De Lencastre,A., Hamill,S. and Pyle,A.M. (2005) A single active-site region for a group II intron. *Nat. Struct. Mol. Biol.*, **12**, 626–627.
40. Michel,F., Umesono,K. and Ozeki,H. (1989) Comparative and functional anatomy of group II catalytic introns—a review. *Gene*, **82**, 5–30.
41. Qin,P.Z. and Pyle,A.M. (1998) The architectural organization and mechanistic function of group II intron structural elements. *Curr. Opin. Struct. Biol.*, **8**, 301–308.