2118–2128 *Nucleic Acids Research, 2005, Vol. 33, No. 7*
doi:10.1093/nar/gki517

Predicting the secondary structures and tertiary interactions of 211 group I introns in IE subgroup

Zhijie Li and Yi Zhang*

State Key Laboratory of Virology and Department of Biotechnology, College of Life Sciences, Wuhan University, Wuhan, Hubei 430072, China

Received November 29, 2004; Revised February 26, 2005; Accepted March 29, 2005

ABSTRACT

The large number of currently available group I intron sequences in the public databases provides opportunity for studying this large family of structurally complex catalytic RNA by large-scale comparative sequence analysis. In this study, the detailed secondary structures of 211 group I introns in the IE subgroup were manually predicted. The secondary structure-favored alignments showed that IE introns contain 14 conserved stems. The P13 stem formed by long-range base-pairing between P2.1 and P9.1 is conserved among IE introns. Sequence variations in the conserved core divide IE introns into three distinct minor subgroups, namely IE1, IE2 and IE3. Co-variation of the peripheral structural motifs with core sequences supports that the peripheral elements function in assisting the core structure folding. Interestingly, host-specific structural motifs were found in IE2 introns inserted at S516 position. Competitive base-pairing is found to be conserved at the junctions of all long-range paired regions, suggesting a possible mechanism of establishing long-range base-pairing during large RNA folding. These findings extend our knowledge of IE introns, indicating that comparative analysis can be a very good complement for deepening our understanding of RNA structure and function in the genomic era.

INTRODUCTION

Group I introns represent a large family of structurally complex catalytic RNA. The self-splicing activity of a group I intron consists of two consecutive transesterification reactions which result in the precise intron excision and exon ligation (1). Although the self-splicing activity was first revealed in study of the *Tetrahymena* intron (2) and was then confirmed by a number of studied group I introns, some group I introns can only undergo the same reactions in living cells, presumably because the requirement of protein factors for assisting the folding of the native structure of these introns (3–6).

Group I introns interrupt the coding sequence of rRNA genes, tRNA genes and mRNA genes in lower eukaryotes and bacteria, with the major population being located in the rRNA genes. Over 2000 group I introns are currently available in the Comparative RNA Website (CRW) database [http://www.rna.icmb.utexas.edu, (7)], which are classified into 5 major subgroups and 10 minor subgroups according to the conservation of the core sequences and structural motifs (8,9). The major subgroups include IA, IB, IC, ID and IE, and the minor subgroups are IA1, IA2, IA3, IB1, IB2, IB3, IB4, IC1, IC2 and IC3. Comparative sequence analysis reveals a common secondary structure for group I introns, including nine base-paired regions designated P1, P3–P10 (8,10). P1 and P10 helices constitute the substrate domain that contains the 5′ and 3′ splice sites, while the other conserved helices constitute the catalytic core structure organized by two separable helical domains P4–P6 and P3–P9 (8,11–12).

In addition to the conserved core helices, at least one, usually multiple variable peripheral base-paired regions are present in each group I intron. Although not conserved in sequence or structure, the universal presence of peripheral elements suggests its importance in the biological function of group I introns (13). Many peripheral elements are predicted to form tertiary interactions among themselves or with the conserved structural elements, establishing an extensive network of peripheral interactions for each subgroup of group I introns (14). Study of the *Tetrahymena* group I ribozyme, belonging to IC1 minor subgroup, showed that many peripheral interactions significantly contribute to the ribozyme activity at the physiological magnesium concentrations through stabilizing the ribozyme core structure (8,15–17). Some peripheral interactions were shown to modulate the folding pathway of the *Tetrahymena* ribozyme (17–19).

Although folding environments, such as temperature and ions, modulate the folding pathway of a group I intron, accumulated data lead to a view that the differences in the folding pathway of group I introns are intrinsic to the intron structure

*To whom correspondence should be addressed. Tel: +86 27 68756207; Fax: +86 27 68754945; Email: yizhang@whu.edu.cn

(20–23). Recent resolved crystal structures of several different group I introns display the similar ribozyme core structure, while the peripheral interactions are significantly different among these introns (24–27). Our study of the folding of a group I intron (subgroup IE) from the 26S rRNA gene of *Candida albicans* has demonstrated essential differences between folding of this group I ribozyme and that of the *Tetrahymena* group I ribozyme. This difference is attributed to the substantial differences in the peripheral interactions and in the tendency of mispairing of the P3 stem of those two introns (28–30). Consistently, an IC3 intron *Azoarcus* group I intron sharing a similar core sequence, but containing very different peripheral interactions from the *Tetrahymena* intron, folds very fast to the native structure, in contrast to the slow folding of the *Tetrahymena* intron (8,24–26,31).

Both biochemical and crystallographic study support that the major function of peripheral elements is to assist the formation and stabilization of the core structure of group I introns (13–15,26–27). Therefore, differences in the peripheral elements may reflect the difference in the intron core sequence, and/or in the splicing factors specifically interacting with the peripheral elements to assist intron folding in the host organisms. This study aims to understand the structure and folding of group I introns, as well as the relationship between the core sequence and peripheral elements, by systematic analysis of all 211 group I introns in the major subgroup IE. IE introns were first discriminated from other introns by their core sequence variation (9); the secondary structure of some IE introns have been proposed (7,9,32). Denotation of these introns from the GenBank nucleic acid sequence database is available on CRW (7). The 218 IE introns indicated in the CRW site were reduced to 211 because of the unavailability of two intron sequence in GenBank, misclassification of four IC1 introns and a non-group I intron sequence. Detailed secondary structures, including all conserved and variable regions, were proposed. Reliable alignments of all these introns were deduced as well. All the secondary structures and alignments are now available at http://www.rna.whu.edu.cn and are also provided as Supplemental Material of this paper.

The secondary structure of a large number of IC1 introns was analyzed before; however, the study was restricted to the introns in the order Bangials and was more focused on the phylogenetic analysis (33). The present study covers all IE introns denoted before July 2003, which distribute in 4 kingdoms and 67 genera. Although the sample number is not huge, the abundance of the analyzed sequences should be suitable for producing some statistically convincing data, which in turn gives a comprehensive description of the structural characteristics of IE introns.

MATERIALS AND METHODS

Sequence source collection and extracting the intron sequences

The accession numbers of the GenBank DNA sequence records containing IE introns were retrieved via the RDBMS system of the Gutell lab's CRW (http://www.rna.icmb.utexas.edu) (Supplementary Table S1). This information was used for downloading the sequence records from GenBank. The sequence files were downloaded in GenBank format. Then the intron sequence and their flanking 10 nt exon sequences were extracted following the insertion site information provided by the CRW database. Biochemical studies prove that group I introns are always cut out after a U at the 5′ splicing site, and after a G at the 3′ splicing site (1). This rule was employed during intron extraction in this study. Although the annotations for these introns in GenBank records do not always comply with it, no violation of the splicing site rule was found in the final sequence alignment, proving that the basic splicing mechanism is applicable to all IE introns.

Secondary structure prediction

In predicting the secondary structure of the group I intron, a divide-and-conquer strategy was used. The known secondary structure of *Candida* group I intron, Ca.LSU, was used as a starting reference (29). The sequence segments P7′ and J8/7 showed high conservation in preliminary examination of the intron sequences; prediction for a new sequence began with a search for P7′ and J8/7, respectively. The starting searching pattern for P7′ was the CGTGCC of Ca.LSU. The search pattern was expanded to CGTGCT and other less prevalent patterns as more patterns became available along with the deposition of newly predicated structures. Then the J8/7 segments were inspected to ensure the correctness of the P7′ prediction. Lying 50–80 nt upstream of P7′, the P7 segments were identified by their complementarities with P7′. The P3 and P3′ segments were also identified by their conserved sequences. Definition of P3 and P7 paired regions divided the whole sequence into four segments containing no pseudoknot structure, whose secondary structures were predicted by the software RNAstructure version 3.71 (34). All P1's were defined by finding the appropriate complementary sequence of the 5′ exon, known as the internal guiding sequence (IGS).

Secondary structure-favored alignment and refinement of the secondary structure prediction

Sequences in the paired regions of the predicted secondary structure of each intron were marked at each position. The alignment began with aligning the bases occupying the same position of different introns. For example, the second paired bases in P4 were aligned together. The sequence similarities are considered as the secondary criterion in aligning operation. G ⇔ A and U ⇔ C are considered to be more compatible than G ⇔ C, G ⇔ U, A ⇔ C and A ⇔ U when they are aligned to a same position, for two reasons. First, the transition mutations are more likely to happen in genome replication process than the transversion mutations. Second, the former mutations have less effect on altering the helical conformation or base-pairing. Introns inserted at positions corresponding to 516, 1199 and 989 of the small subunit rRNA and all in the large subunit rRNA of *E.coli* were aligned separately, and four alignments instead of one were thus generated.

After the alignments were produced, sequences in each alignment were further checked to identify conserved patterns. The unfit sequences in the alignment were rechecked for their secondary structure prediction. The secondary structure and alignment were iteratively adjusted to improve the structure prediction and alignment.

Computation of the distance matrix

To compute the phylogenetic distance of core sequences, the sequences in the segments P3–J3/4–P4, P4′–P6 (partial), P6′(partial)–J6/7–P7–P3–P8 (partial) and P8′(partial)–J8/7–P7′ corresponding to 95–112, 208–223, 250–282 and 296–312 sites of the sequence of group I intron from the large subunit rRNA of *T.thermophilia* were calculated.

Distance matrices were calculated using the DNAdist program in PHYLIP package (35), which takes nucleotide substitution rates as sequence distance measures. The distances between the core sequences of 211 introns constituted a 211-dimensional space. A Principle Component Analysis, using the Statistic Toolbox of MATLAB 6.5, reduced the 211 dimensions to 3 principle dimensions, which cover >90% of the variations in the distance matrix. Then the distances were plotted to a 3D Euclidean space to visualize the relation of the core sequences.

Computing mutual information

Mutual information is defined (in nits) as: $M(x,y) = \sum(f_{b_x b_y}) \ln(f_{b_x b_y}/f_{b_x} f_{b_y})$, for all bases b_x and b_y, where b_x and b_y refer to the identities of each possible base at positions x and y (A, G, C, U or gap corresponding to the ambiguous bases). f_{b_x} and f_{b_y} are the frequencies of each base at each position, and $f_{b_x b_y}$ is the frequency of each possible pair of bases at x and y (36). When two sites co-vary, a large mutual information number should be observed.

Mutual information was computed by using both a C program and the mutual information function in the program BioEdit (37). The resulting matrices were plotted to a 2D planar graph, in which the coordination of a dot are two sites in the alignment, and the darkness represents the level of mutual information between these two sites.

Supplementary Table S1 summarizes the general information of all IE introns analyzed in this study.

RESULTS AND DISCUSSION

The conserved secondary structure and tertiary architecture

According to CRW (7), all 211 IE introns are from the rRNA genes of lower eukaryotes, predominately in fungi; 92.4% of these introns are located in the small subunit rRNA gene and 7.6% in the large subunit rRNA gene. These IE introns insert at 6 and 4 positions on the small and large subunit rRNAs, respectively; these positions correspond to S516, S568, S651, S989, S1201, S1199 and L1923, L1926, L2066, L2563 of the 16S rRNA and 23S rRNA of *E.coli*. The number of IE introns at those sites are 82, 1, 1, 21, 1, 89, 4, 2, 6 and 4, respectively, demonstrating a strong insertion bias to positions S516, S1199 and S989. Introns inserted into the same rRNA site tend to share common structural features. Our preliminary examination of the secondary structure of each intron revealed four classes of IE introns, i.e. S516 introns, S989 introns, S1199 introns and LSU introns. Note that the S568 intron and S651 intron were classified into S516 intron and S1201 intron into the S1199 one, based on the observed structural similarity.

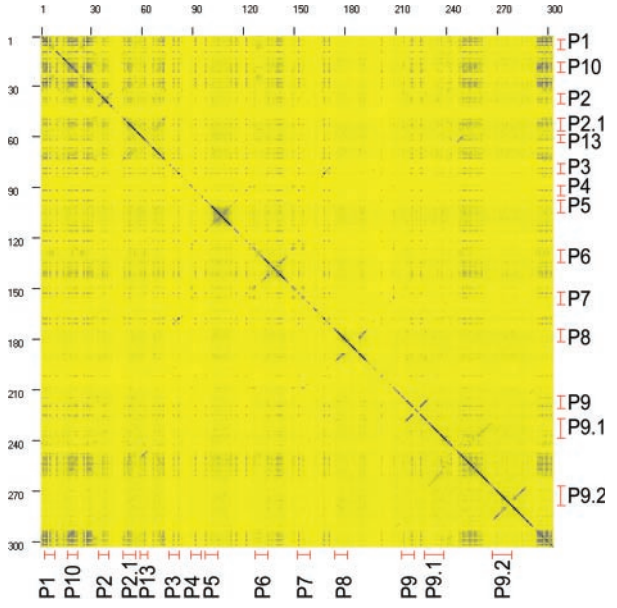


Figure 1. Mutual information computed from the alignment of 211 IE introns. Mutual information is defined (in nits) as $M(x,y) = \sum(f_{b_x b_y}) \ln(f_{b_x b_y}/f_{b_x} f_{b_y})$; the higher $M(x,y)$ indicates that the two corresponding sites are more likely to interact. Mutual information was computed for each pair of two sites in the alignment, and the resulting matrix was plotted in a 2D map, with the $M(x,y)$ value being represented by the density of darkness. Mutual information was computed using the program BioEdit (37).

Iterative adjustment of the secondary structures and alignments of these 211 introns yielded a reliable secondary structure for each individual intron and high-quality alignments for each class of IE introns (see http://www.rna.whu.edu.cn; Supplementary 2nd Structures and Supplementary Alignments). Mutual information computed from the alignments of IE introns indicated 14 conserved base-paired regions, which include P10–P1–P2–P2.1, P4–P5–P6, P8–P3–P7–P9–P9.1–P9.2 that form three separable domains of group I introns (Figures 1 and 2). The common secondary structures for IE introns were deduced according to the alignments (Figure 2A–C). Four of the conserved stems formed by long-range base-pairing were predicted; among which P3, P7 and P10 are conserved among all group I introns and P13 is shown to be conserved among IE introns. Interestingly, four peripheral stems including P2, P2.1, P9.1, P9.2 that are highly variable among other subgroups are exclusively conserved among IE subgroup, suggesting the importance of these peripheral elements in the function of IE introns. These conserved peripheral elements might endorse a tertiary architecture specific for IE introns. For example, the recently published crystal structures of *Azoarcus* tRNA intron (IC3) and a phage Twort intron (IA2) demonstrate very different overall architectures due to the significantly different peripheral interactions, although the core structures are similar (26–27).

Base pairs between nucleotides that are remote in the secondary structure contribute significantly to the stabilization of RNA tertiary structure in the large ribosomal subunit (38). The most striking feature of IE introns is the highly conserved P13 paired region (Table 1), a 6–8 bp stem resulting from long-range base-pairing of the two remote peripheral elements P2.1 and P9.1 (Figures 2 and 3) (14). Two incomplete S516

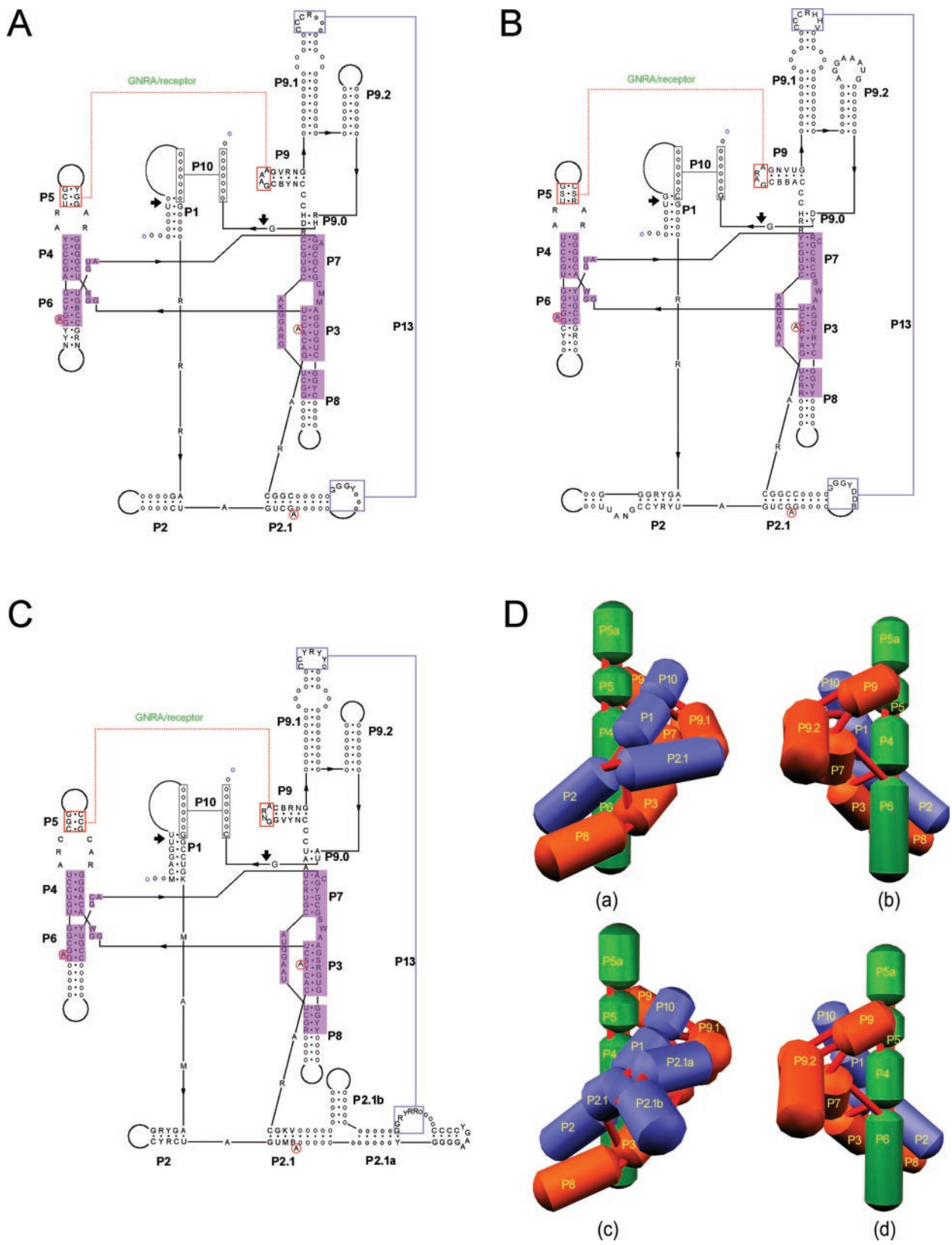
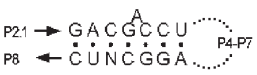
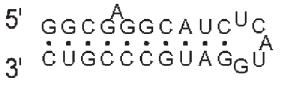
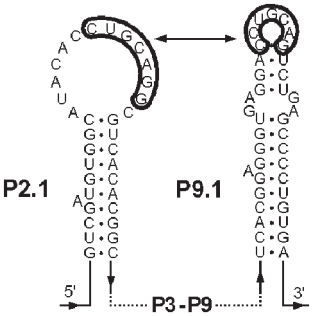
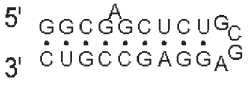
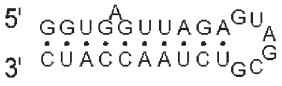
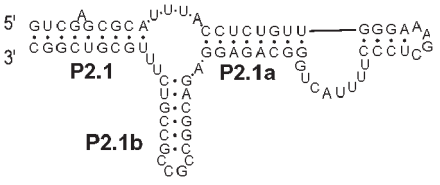
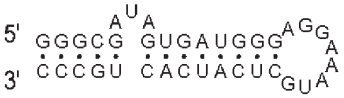
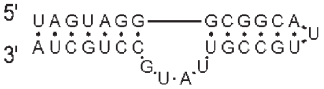


Figure 2. Predicted structures of IE introns. The common secondary structures of IE1 (**A**), IE2 (**B**) and IE3 (**C**) introns are shown. The highly conserved core sequences used in the distance computation are shaded in purple. The conserved sequences (90% or over) of each minor subgroup are indicated. When two or three bases collectively respond for the conservation at one site, they are denoted using degenerate codes: R: A/G; W: A/U; Y: C/U; M: A/C; S: C/G; K: G/U; B: not A; V: not U; H: not G; D: not C. Non-conserved sequences are denoted as circles. (**D**) The outlined tertiary structures. Upper row shows the models for IE1 and IE2 introns: (a) front view and (b) back view. Lower row shows the models for IE3 introns: (c) front view and (d) back view.

Table 1. Statistic analysis of the structural motifs present in different classes of IE introns

Conserved pattern	Location	Examples	Frequency of occurrence LSU	S516	S989	S1199
A-bulge	P3		16/16	82/84	20/21	88/90
A-bulge	P6		16/16	82/84	21/21	90/90
P13	P2.1, P9.1		16/16	80/84	21/21	88/90
GNRA loop	P6		8/16	46/84	0	4/90
GYNRCG loop	P6		0	0	20/21	82/90
P2.1a/P2.1b	P2.1		0	0	20/21	85/90
AGGAAAUG loop	P9.2		8/16	32/84	0	0
UANG loop	P2′		5/16	41/84	0	0

introns were found in viridiplantae, one completely lacking P3–P9 domain and one lacking P7–P8–P9 region. The accession numbers for these two introns are #185 and #191 in this study, respectively. Interestingly, although these two introns lack P9.1 structure, the remained terminal loop of P2.1 stem contain conserved sequences capable of pairing with the normal L9.1 (Supplementary 2nd Structures). The strong conservation of the P13 stem formed by base-pairing of two remote peripheral elements P2.1 and P9.1 indicates the importance of this tertiary interaction in forming the native structure of these introns.

Tertiary interaction between P5 and P9 stems is conserved among group I introns including subgroups IA, IB, IC and ID, either through a GNRA loop/helix interaction (L9/P5) or through forming new stem P17 by Watson–Crick base-pairing (14,39). We found that most IE introns from fungi are capable of forming the standard L9/P5 tetraloop–receptor interactions, in which the L9 terminal loop predominately consists of GAAA sequence and the receptor stem P5 contains the typical GNRA receptor motif (Supplementary Figure S1). Interestingly, most L9 loops of the viridiplantae IE introns contain >4 nt. Nonetheless, most of these loops contain a typical GNRA sequence, and nucleotides flanking the GNRA sequence are capable of forming an A–U pair. This A–U pair closes a typical L9 GNRA tetraloop, with the nucleotides adjacent to the A–U pair being purines that are capable of forming non-canonical base pair. On the other hand, the P5 stems of these viridiplantae introns contain the typical GNRA receptor motifs. Therefore, we suggest that these larger P9 loops form an isoform of the canonical GNRA tetra loop, and the L9/P5 interaction establishes in these introns (Figure 4).

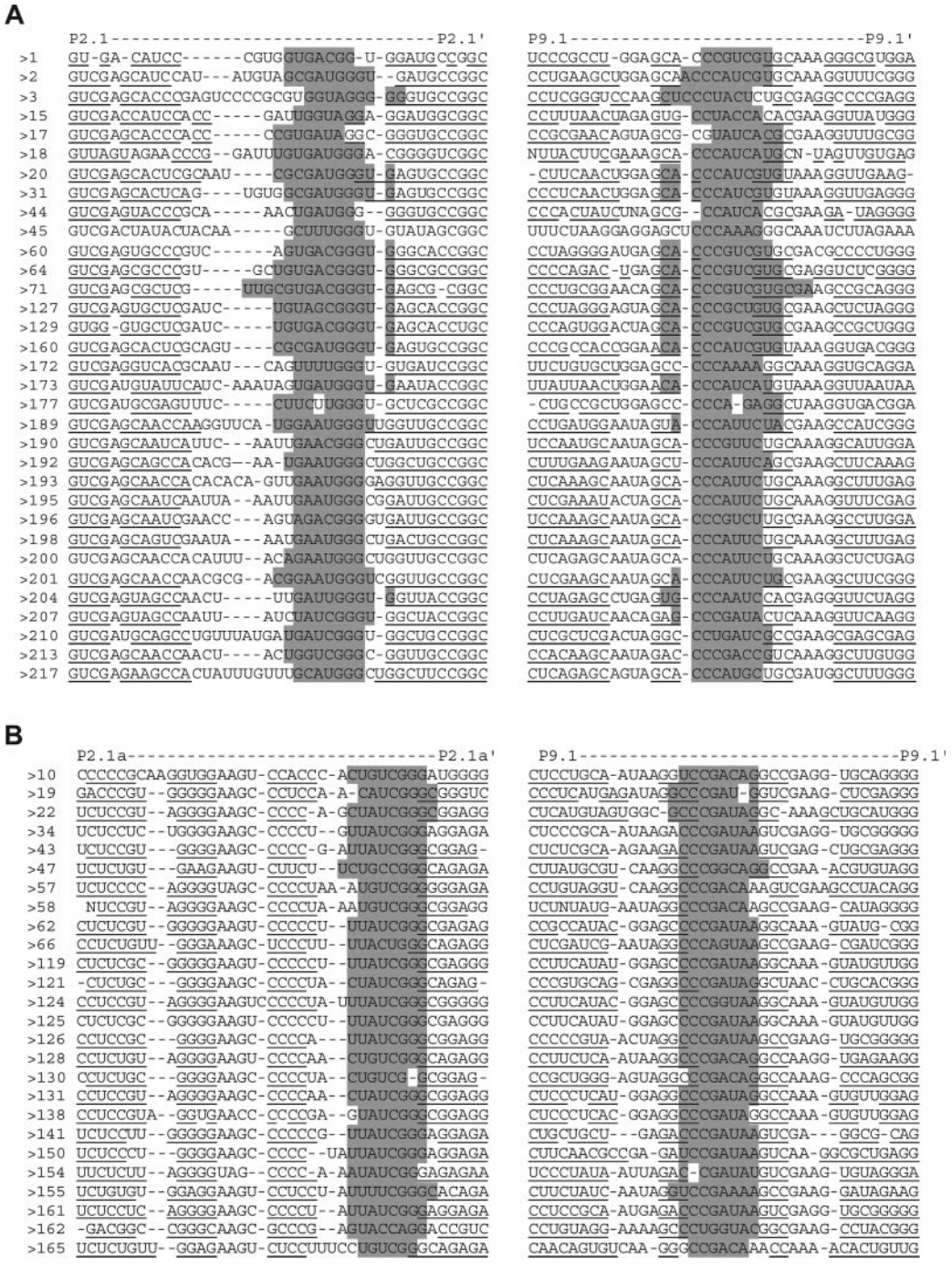


Figure 3. Phylogenetic evidence for the conserved long-range interaction P13 established by base-pairing between P2.1 and P9.1. Extracts from the alignments of IE introns inserted at SSU516 (**A**) and SSU1199 (**B**). The sequences participating in P13 base-pairing are shaded.

In the view of organizing the compact tertiary structure of IE introns that exerts the splicing function, the conserved tertiary interactions established between P2.1 and P9.1, P5 and P9 represent two important domain–domain interactions that bring three domains to a close proximity. The base triples formed between J3/4 and P6, J6/7 and P4 are highly conserved (Figure 2A–C), further tethering domains P4–P5–P6 and P8–P3–P7–P9–P9.1–P9.2. Formation of P3 by long-range base-pairing brings the domain P10–P1–P2–P2.1 close to the domain P8–P3–P7–P9P9.1–P9.2. These interactions reveal the outlined tertiary architecture of IE introns, which is illustrated in Figure 2D.

The conserved A residues not participating in the regular helix formation are more abundant than other bases in the 23 rRNA; these conserved A residues participate in A-minor motif interactions (40). The A-minor motif represents the most abundant tertiary structure interaction in 23 rRNA, significantly contributing to the structural stability of this extremely large structural RNA (40,41). Interestingly, an extremely high occurrence of the A-bulge motif separating the paired

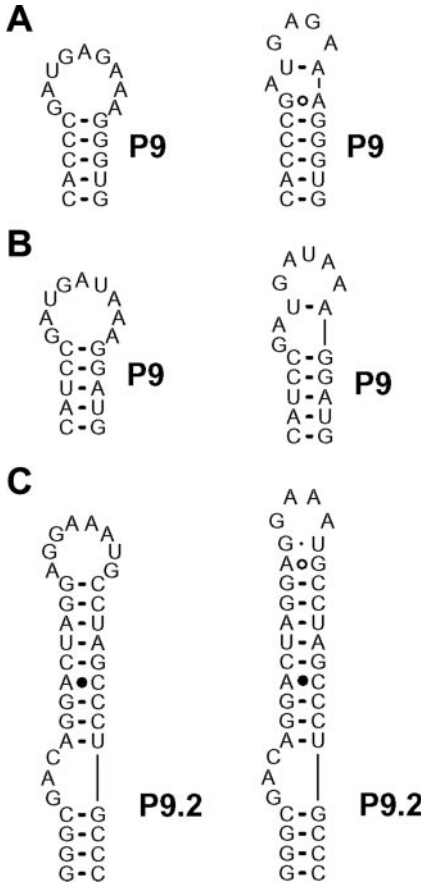


Figure 4. Predicted isoforms of canonical GNRA tetraloops presented in L9 and L9.2 of some IE introns. Standard base-pairing resulted in the loops (left). If non-canonical A–G base pair and bulges are allowed, the loops in (**A**) and (**C**) become GNRA tetraloops, while that in (**B**) turns to GAUAA. See Figure 2 and Supplementary Figure S1 for the receptor structure of the L9 GNRA loop, which is located in P5.

regions was observed among IE introns. In P3 and P6 stems, the conserved first 4 bp immediately followed by a single A-bulge, dividing these stems into two parts. This '4 bp stem-single A bulge-stem' motif is exclusively conserved in the P3 and P6 helical regions of IE introns, with a few exceptions. In these few exceptions, an A-bulge motif containing multiple unpaired A was frequently observed. The '4 bp stem-single A bulge-stem' motif also occurs frequently in P2.1 stem; in some cases, the single A bulge is replaced by an AU bulge. Moreover, base pairs abutting the A-bulge are usually G–C pairs (Table 1). The conserved A-bulge motifs might establish the A-minor tertiary interactions with the corresponding receptor helices, thus playing an important role in stabilizing the tertiary structure of IE introns. This prediction is consistent with the report of rich A-minor interactions in the *Azoarcus* group I ribozyme (25).

Another noteworthy feature of the secondary structure of IE introns is the relatively long P10 stem that is essential for precise ligation of the 5′ and 3′ exons. The P10s of IE introns are predominately 6–8 Watson–Crick base pairs. The typical P10 stem is present in the two naturally occurring incomplete IE introns (#185 and #191 in Supplementary 2nd Structures). However, the mechanistic reason for such a long P10 stem could not be readily proposed.

Variation in the core sequences divides IE introns into three distinct minor subgroups

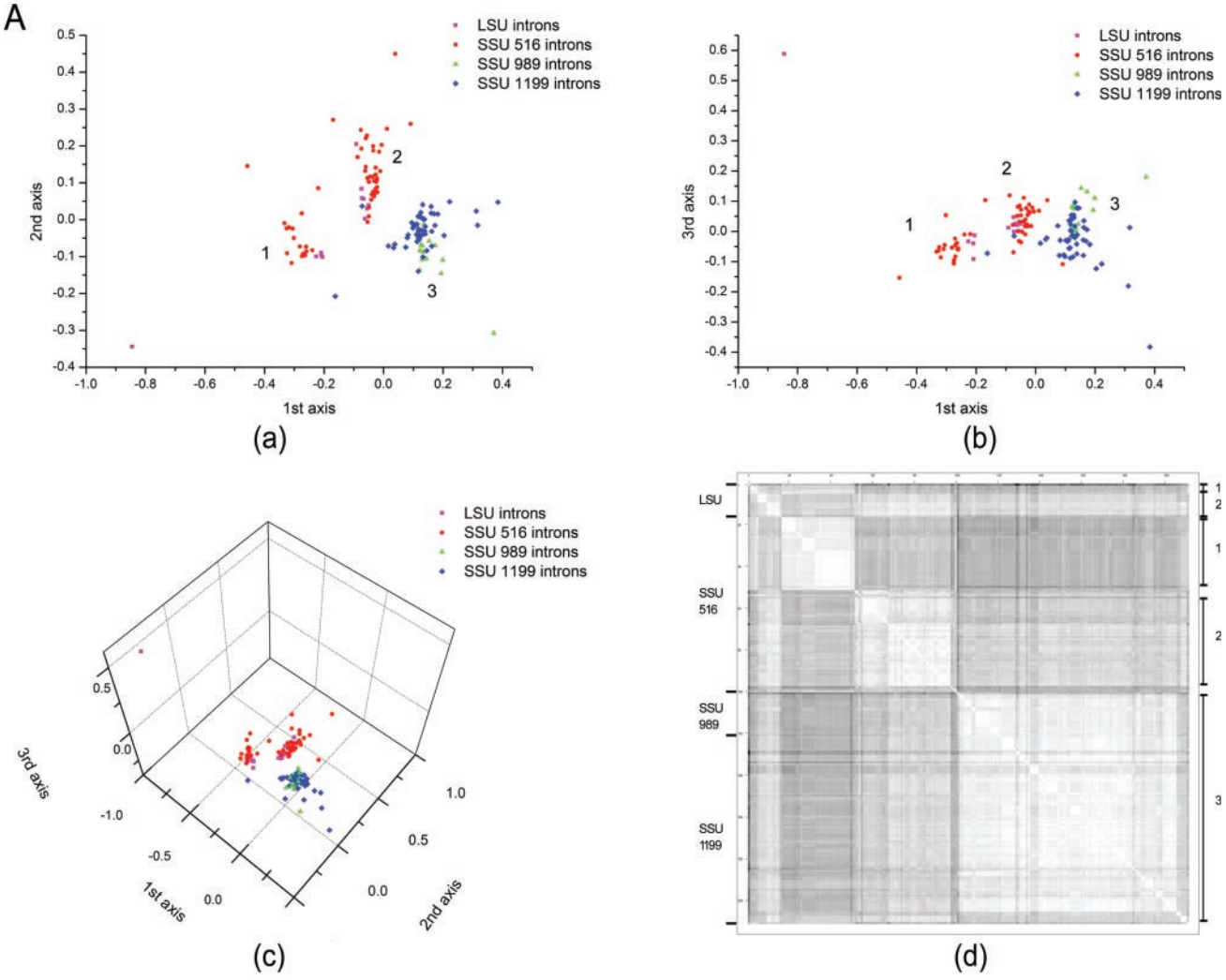
Group I introns vary greatly in their sequences, however, the nucleotide sequences in the core region of group I introns are relatively conserved, which include stems P3, P4, P7, a part of the stems P6, P8, and four joint segments J3/4, J6/7, J7/3 and J8/7. As expected, crystal structures of three group I introns demonstrate that these conserved core segments constitute the structure core as well (24–27). Nonetheless, the core sequences vary among the subgroups of group I introns, which is a major criterion used to categorize group I introns into different subgroups (8,9). On the other hand, most known minor subgroups are more distinguishable by the signature structural motifs but not by core sequence conservation (8).

Multiple sequence patterns in the conserved core of IE introns were noticed (Figure 5). Interestingly, when the nucleotide sequences in the core were calculated to compute the phylogenetic distance of IE introns, it was found that almost all IE introns fell into one of the three well-separated clusters (Figure 5A), representing three minor groups, namely IE1, IE2 and IE3. Therefore, we demonstrated ambiguously that the core sequence variation could be used as a criterion to distinguish the minor subgroups. The similar approach of minor subgroup classification has been used previously, which did not yield such a distinct separation of minor subgroups probably due to the lack of sufficient sequences to generate good statistical data (8). In the following section, we describe the relationship between the signature structural motifs and core sequence identity.

Co-variation of the peripheral structural motifs with the core sequences

The hypothesis that the major function of peripheral elements is to assist the folding of the core structure of group I introns predicts co-variation of the core sequences with the peripheral elements. Consistent with this prediction, IE introns, whose core sequence significantly diverges from that of the other major subgroups of group I introns, contain a number of characteristic peripheral elements. For example, the P13 stem formed by long-range base-pairing between the peripheral elements P2.1 and P9.1 is highly conserved among IE introns. In addition, the peripheral elements P2 and P9.2 are also conserved among IE introns (Table 1). Because IE1, IE2 and IE3 are distinctly different in their core sequences, we were interested to know if there are any signature structural motifs for each minor subgroup.

Interestingly, we noticed that a branched P2.1 is conserved among IE3 introns, with the large bulge within the P2.1a stem that participates in base-pairing with L9.1 to form P13 stem (Table 1). On the other hand, IE1 and IE2 introns contain a non-branched P2.1, in which the terminal loop participate in the P13 stem formation. Concomitant with the branched P2.1 motif, almost all IE3 introns contain a GYNRCG terminal loop at P6 stem. Nevertheless, conserved base-pairing between these two conserved motifs was not observed. Because the tertiary structural feature of GYNRCG loop has never been reported, it is difficult to predict the interaction site of L6 in IE3 introns. We suggest that the branched P2.1 stem does not directly interact with L6 because of the spatial restriction



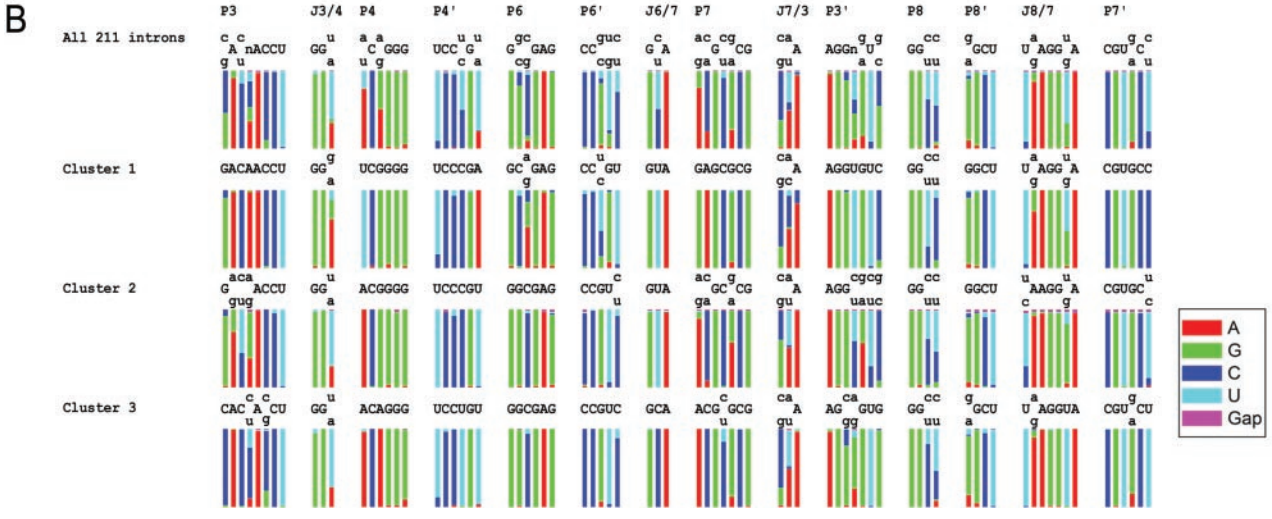


Figure 5. Sequence variation in the core region divides IE introns into three minor subgroups. (**A**) The core distance matrix was plotted to a 3D space (a) (b) (c), or a 2D planar graph (d). Each dimension in (d) represents the 211 introns; the lighter dots are the crosses of those sequences having higher similarity. (**B**) Sequence conservation for the core region of all IE introns and for each minor subgroup. For each group, the upper lines show the conserved sequence. The lower stacked bars show the frequency of each nucleotide (or gaps) observed in each site. The conventions used for defining the consensus sequence are from (47): an uppercase letter designates >90% conservation of the particular nucleotide; a pair of lowercase letters indicates that 2 nt frequently occupy a position and together account for >90% of the sequences; and an 'n' in a position indicates that no nucleotide is conserved at the level of these criteria.

imposed by base-pairing of L9.1 and L2.1a (Figure 2D). Strikingly, the GYNRCG terminal loop and branched P2.1 motifs are excluded from IE1 and IE2 introns inserted at position 516 of the small ribosomal rRNA and the large ribosomal rRNA, indicating that those structural motifs are only required for folding of the core structure of IE3 intron. On the other hand, co-occurrence of a L6 GNRA tetraloop and a UANG bulge at P2 were conserved among most IE2 introns (Table 1). No distinct structural motifs were found for IE1. These results are consistent with our hypothesis that variation of the core sequences requires the concomitant change in the peripheral structures.

As a frequently observed structural motif, GNRA tetraloop plays important roles in stabilizing the RNA tertiary structure (42,43). Co-occurrence of GNRA terminal loop in P6 and a UANG bulge in P2 dominates S516 IE2 introns, suggesting a possible tertiary interaction between L6 and P2.

Host-specific structural motifs

It was noticed that all IE3 introns are inserted at S1199 and S989 positions of the small subunit rRNA, and no other IE introns are found at these positions. On the other hand, examination of CRW database revealed that all except for a few of the 111 introns inserted at those two sites belong to IE3 subgroup. Such a strong restriction of intron type by the insertion site on rRNA suggests an intimate relationship between the intron core structure/folding and the neighboring rRNA structure/folding. In contrast, the position S516 harbors 84 of IE1 and IE2 introns, and 107 of IC1 introns, indicating that the interaction between the intron core and the folding context around S516 is much less intimate than that of the introns at S1199 and S989. Such a loose interaction may offer opportunities for trans-acting cellular factors to interact with the intron core to assist intron folding. This hypothesis is consistent with our finding that the core sequence and peripheral structural motifs of the S516 IE introns strongly correlate with the host species. First, all of the 30 S516 IE introns in viridiplantae belong to IE2 minor subgroup. Correspondingly, the structural motifs including a GNRA terminal loop in the P6 stem, a UANG bulge in P2 and a GNRA tetraloop motif in P9.2 are highly conserved in these viridiplantae introns (Table 1 and Supplementary 2nd Structures). The terminal loop sequence of P9.2 is highly conserved as AGGAAAUG, which contains a GAAA tetraloop abutted by the G–U and A–G pairs (Figure 4C). Interestingly, the conserved L9.2 motif is also found in 5 out of 6 non-Ascomycota fungal IE2 introns, but not in any Ascomycota fungal IE2. In contrast, co-occurrence of the GNRA terminal loop in P6 and a UANG bulge in P2 is present in 7 out of 18 fungal IE2 introns, in spite of the host species. These results led us to conclude that the L9.2 AGGAAAUG motif is more host-specific than the L6 GNRA and P2 UANG motifs.

The observation that the L9.2 AGGAAAUG sequence motif is strongly excluded from all Ascomycota IE2 introns, but highly conserved among non-Ascomycota IE2 introns suggests that this motif is specifically required for intron splicing in non-Ascomycota hosts. The L9.2 motif might exert its function through interaction with specific cellular factor(s) that promote intron folding. Although the mechanism is unknown, the presence of host-specific structural motifs suggests an intimate interaction between the intron RNA and host-specific cellular factors.

Competing base-pairing in formation of the long-range paired helices

Long-range base-pairing is frequently used in nature to organize the structurally complex ribozyme RNAs, and pseudoknot structure is a frequently observed long-range base-pairing (14,38,44–46). The P3–P7 pseudoknot is essential for formation of the core structure, while the P1–P10 pseudoknot containing the 5′ and 3′ splice sites is important for the precise ligation of exons of group I introns. Here, we showed that the P13 stem formed by long-range base-pairing of P2.1 and P9.1, is highly conserved among IE introns, indicative of its importance in the function of these group I introns.

Sequence analysis revealed that one or two bases at both ends of the P13 paired region are also capable of forming base pairs in P9.1 and/or P2.1, which is highly conserved among IE introns. Therefore, unwinding of at least 1 bp in P9.1 or P2.1 stems is usually required to form additional base pairs of P13 stem, which we named as competitive base-pairing (Figure 6).

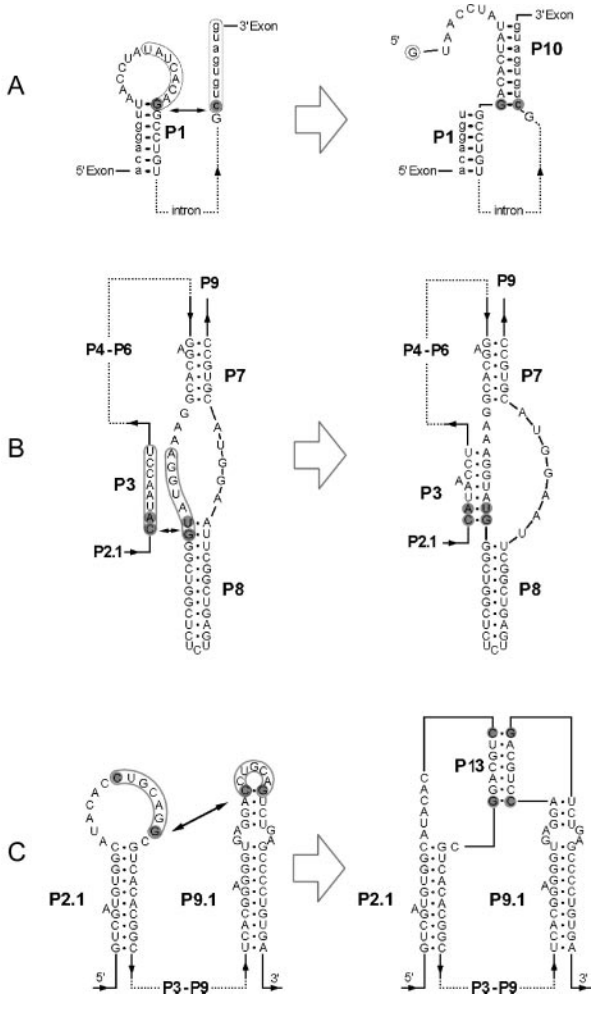


Figure 6. Competitive base-pairing at the junctions of distantly paired stems and local stems. (**A**) Competitive pairing at P1–P10 junction; (**B**) Competitive pairing at P3–P8 junction; (**C**) Competitive pairing at P2.1–P13–P9.1. Shaded regions indicate nucleotides participating in competitive base-pairing.

Such a competitive pairing at the junctions of P13 was reported previously (14). Competitive base-pairing is also present at the junction of P1 and P10, where the IGS sequence pairing with the first nucleotide of the intron in P1 is also capable of pairing with the first nucleotide of 3′ exon in P10; P10 sometimes is capable of extending one more pair by unwinding one P1 pair (Figure 6). Conserved competitive base-pairing takes place in the junction of P3 and P8 as well. The last 2 bp in P8 close to this junction are U–C, A–U or U–G, A–U pairs; these P8 pairs are supposedly opened during P3 formation and subsequently converted to C–G, A–U pairs of P3 stem (Figure 6).

The conserved competitive pairing at the junctions of long-range base-pairing regions of IE introns leads us to understand the mechanism of large RNA folding from a different aspect. Folding of the large RNAs to the 3D functional structure often requires distant tertiary contacts, such as helix–helix contact, tetraloop–helix receptor contact and long-range base-pairing. All these tertiary contacts should be established after the local secondary structure formation. Study of the folding of *Tetrahymena* intron *in vitro* has proven that rapid formation of some tertiary contacts results in metastable intermediates that fold to the active structure very slowly (17,18,21), leading to a view that a slower formation of the long-range base-pairing than the local one may prevent the ribozyme from entering the kinetically trapped folding pathway. Consistent with this hypothesis, the stability of the long-range base-pairing is usually less stable than or equivalent to that of the neighboring local base-pairing (Figure 6). We proposed a model for formation of a stable helix by long-range base-pairing. During large structural RNA folding, local stem–loop structures form rapidly; while long-range base-pairing contributing to tertiary structure formation and stability form relatively slowly. After the latter base-pairing is established, further tertiary stability may be achieved by forming extra base pairs through replacing the adjacent base pairs in the abutted local helices (Figure 6). This hierarchical folding strategy by competitive base-pairing may ensure that the intron RNA folds into the native tertiary structure without being trapped in the misfolded intermediates.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Profs Z. Tan and P. Shen for the useful discussion and critical comments. We owe thanks to our lab member Mr Chen Lu for help in the preparation of figures. This work is supported by the National Natural Science Foundation of China through grants 30270317 and 30330170, and by Wuhan University through grant 0000028 (Y.Z.). Funding to pay the Open Access publication charges for this article was provided by State Key Laboratory of Virology.

Conflict of interest statement. None declared.

REFERENCES

1. Cech,T.R. (1990) Self-splicing of group I introns. *Annu. Rev. Biochem.*, **59**, 543–568.
2. Kruger,K., Grabowski,P.J., Zaug,A.J., Sands,J., Gottschling,D.E. and Cech,T.R. (1982) Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of *Tetrahymena*. *Cell*, **31**, 147–157.
3. Herschlag,D. (1995) RNA chaperones and the RNA folding problem. *J. Biol. Chem.*, **270**, 20871–20874.
4. Lambowitz,A.M., Caprara,M.G., Zimmerly,S. and Perlman,P.S. (1999) Group I and group II ribozymes as RNPs: clue to the past and guides to the future. In Gesteland,R., Cech,T.R. and Atkins,J. (eds), *The RNA World, 2nd edn*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 451–485.
5. Schroeder,R., Barta,A. and Semrad,K. (2004) Strategies for RNA folding and assembly. *Nature Rev. Mol. Cell Biol.*, **5**, 908–919.
6. Schroeder,R., Grossberger,R., Pichler,A. and Waldsich,C. (2002) RNA folding *in vivo*. *Curr. Opin. Struct. Biol.*, **12**, 296–300.
7. Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V., Muller,K.M. *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.
8. 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.
9. Suh,S.O., Jones,K.G. and Blackwell,M. (1999) A group I intron in the nuclear small subunit rRNA gene of *Cryptendoxyla hypophloia*, an ascomycetous fungus: evidence for a new major class of group I introns. *J. Mol. Evol.*, **48**, 493–500.
10. Burke,J.M., Belfort,M., Cech,T.R., Davies,R.W., Schweyen,R.J., Shub,D.A., Szostak,J.W. and Tabak,H.F. (1987) Structural conventions for group I introns. *Nucleic Acids Res.*, **15**, 7217–7221.
11. Tanner,M.A., Anderson,E.M., Gutell,R.R. and Cech,T.R. (1997) Mutagenesis and comparative sequence analysis of a base triple joining the two domains of group I ribozymes. *RNA*, **3**, 1037–1051.
12. Golden,B.L., Gooding,A.R., Podell,E.R. and Cech,T.R. (1998) A preorganized active site in the crystal structure of the *Tetrahymena* ribozyme. *Science*, **282**, 259–264.
13. Doherty,E.A. and Doudna,J.A. (2000) Ribozyme structures and mechanisms. *Annu. Rev. Biochem.*, **69**, 597–615.
14. Lehnert,V., Jaeger,L., Michel,F. and Westhof,E. (1996) New loop–loop tertiary interactions in self-splicing introns of subgroup IC and ID: a complete 3D model of the *Tetrahymena* thermophila ribozyme. *Chem. Biol.*, **3**, 993–1009.
15. Engelhardt,M.A., Doherty,E.A., Knitt,D.S., Doudna,J.A. and Herschlag,D. (2000) The P5abc peripheral element facilitates preorganization of the *Tetrahymena* group I ribozyme for catalysis. *Biochemistry*, **39**, 2639–2651.
16. Ikawa,Y., Ohta,H., Shiraishi,H. and Inoue,T. (1997) Long-range interaction between the P2.1 and P9.1 peripheral domains of the *Tetrahymena* ribozyme. *Nucleic Acids Res.*, **25**, 1761–1765.
17. Treiber,D.K. and Williamson,J.R. (2001) Concerted kinetic folding of a multidomain ribozyme with a disrupted loop–receptor interaction. *J. Mol. Biol.*, **305**, 11–21.
18. Pan,J. and Woodson,S.A. (1999) The effect of long-range loop–loop interactions on folding of the *Tetrahymena* self-splicing RNA. *J. Mol. Biol.*, **294**, 955–965.
19. Russell,R., Zhuang,X., Babcock,H.P., Millett,I.S., Doniach,S., Chu,S. and Herschlag,D. (2002) Exploring the folding landscape of a structured RNA. *Proc. Natl Acad. Sci. USA*, **99**, 155–160.
20. Tinoco,I.,Jr and Bustamante,C. (1999) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
21. Treiber,D.K. and Williamson,J.R. (2001) Beyond kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.*, **11**, 309–314.
22. Hanna,R. and Doudna,J.A. (2000) Metal ions in ribozyme folding and catalysis. *Curr. Opin. Chem. Biol.*, **4**, 166–170.
23. Woodson,S.A. (2002) Folding mechanisms of group I ribozymes: role of stability and contact order. *Biochem. Soc. Trans.*, **30**, 1166–1169.
24. Guo,F., Gooding,A.R. and Cech,T.R. (2004) Structure of the *Tetrahymena* ribozyme: base triple sandwich and metal ion at the active site. *Mol. Cell*, **16**, 351–362.
25. Adams,P.L., Stahley,M.R., Gill,M.L., Kosek,A.B., Wang,J. and Strobel,S.A. (2004) Crystal structure of a group I intron splicing intermediate. *RNA*, **10**, 1867–1887.

26. Adams,P.L., Stahley,M.R., Kosek,A.B., Wang,J. and Strobel,S.A. (2004) Crystal structure of a self-splicing group I intron with both exons. *Nature*, **430**, 45–50.
27. Golden,B.L., Kim,H. and Chase,E. (2005) Crystal structure of a phage Twort group I ribozyme-product complex. *Nature Struct. Mol. Biol.*, **12**, 82–89.
28. Zhang,Y. and Leibowitz,M.J. (2001) Folding of the group I intron ribozyme from the 26S rRNA gene of *Candida albicans*. *Nucleic Acids Res.*, **29**, 2644–2653.
29. Xiao,M., Leibowitz,M.J. and Zhang,Y. (2003) Concerted folding of a *Candida* ribozyme into the catalytically active structure posterior to a rapid RNA compaction. *Nucleic Acids Res.*, **31**, 3901–3908.
30. Zhang,L., Xiao,M., Lu,C. and Zhang,Y. (2005) Fast formation of the P3–P7 pseudoknot: a strategy for efficient folding of the catalytically active ribozyme. *RNA*, **11**, 59–69.
31. Rangan,P., Masquida,B., Westhof,E. and Woodson,S.A. (2003) Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc. Natl Acad. Sci. USA*, **100**, 1574–1579.
32. Mercure,S., Montplaisir,S. and Lemay,G. (1993) Correlation between the presence of a self-splicing intron in the 25S rDNA of *C.albicans* and strains susceptibility to 5-fluorocytosine. *Nucleic Acids Res.*, **21**, 6020–6027.
33. Muller,K.M., Cannone,J.J., Gutell,R.R. and Sheath,R.G. (2001) A structural and phylogenetic analysis of the group IC1 introns in the order Bangiales (Rhodophyta). *Mol. Biol. Evol.*, **18**, 1654–1667.
34. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
35. Felsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
36. Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res.*, **20**, 5785–5795.
37. Hall,T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids. Symp. Ser.*, **41**, 95–98.
38. Ban,N., Nissen,P., Hansen,J., Moore,P.B. and Steitz,T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science*, **289**, 905–920.
39. Jaeger,L., Westhof,E. and Michel,F. (1996) Function of a pseudoknot in the suppression of an alternative splicing event in a group I intron. *Biochimie*, **78**, 466–473.
40. Gutell,R.R., Cannone,J.J., Shang,Z., Du,Y. and Serra,M.J. (2000) A story: unpaired adenosine bases in ribosomal RNAs. *J. Mol. Biol.*, **304**, 335–354.
41. Nissen,P., Ippolito,J.A., Ban,N., Moore,P.B. and Steitz,T.A. (2001) RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proc. Natl Acad. Sci. USA*, **98**, 4899–4903.
42. Jaeger,L., Michel,F. and Westhof,E. (1994) Involvement of a GNRA tetraloop in long-range RNA tertiary interactions. *J. Mol. Biol.*, **236**, 1271–1276.
43. Costa,M. and Michel,F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J.*, **14**, 1276–1285.
44. Ferre-D'Amare,A.R., Zhou,K. and Doudna,J.A. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–574.
45. 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.
46. Walker,S.C. and Avis,J.M. (2004) A conserved element in the yeast RNase MRP RNA subunit can participate in a long-range base-pairing interaction. *J. Mol. Biol.*, **341**, 375–388.
47. Cech,T.R. (1988) Conserved sequences and structures of group I introns: building an active site for RNA catalysis. *Gene*, **73**, 259–271.