



Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution

Maria Costa, Francois Michel

► To cite this version:

Maria Costa, Francois Michel. Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution. EMBO Journal, 1997, 16, pp.3289 - 3302. <10.1093/emboj/16.11.3289>. <hal-03914569>

HAL Id: hal-03914569

<https://hal.science/hal-03914569v1>

Submitted on 26 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Rules for RNA recognition of GNRA tetraloops deduced by *in vitro* selection: comparison with *in vivo* evolution

Maria Costa and François Michel

Centre de Génétique Moléculaire du CNRS, 91190 Gif-sur-Yvette, France

e-mails: costa@smigiris.cgm.cnrs-gif.fr
michel@smigiris.cgm.cnrs-gif.fr

Terminal loops with a GNRA consensus sequence are a prominent feature of large self-assembling RNA molecules. In order to investigate tertiary interactions involving GNRA loops, we have devised an *in vitro* selection system derived from a group I ribozyme. Two selections, destined to isolate RNA sequences that would recognize two of the most widespread loops (GUGA and GAAA), yielded variants of previously identified receptors for those loops, and also some yet unrecognized, high-affinity binders with novel specificities towards members of the GNRA family. By taking advantage of available crystal structures, we have attempted to rationalize these results in terms of RNA–RNA contacts and to expose some of the structural principles that govern GNRA loop-mediated tertiary interactions; the role of loop nucleotide 2 in ensuring specific recognition by receptors is emphasized. More generally, comparison of the products of *in vitro* and natural selection is shown to provide insights into the mechanisms underlying the *in vivo* evolution of self-assembling RNA molecules.

Keywords: GNRA terminal loops/*in vitro* selection/molecular evolution/RNA tertiary interactions/self-splicing introns

Introduction

Like protein enzymes, large RNA catalysts, such as group I and group II self-splicing introns or the RNA component of bacterial RNase P, fold into compact structures that are stabilized by a multiplicity of tertiary interactions (Latham and Cech, 1989; Cate *et al.*, 1996a). At least some of these interactions must correspond to recurrent structural motifs, whose identification and characterization should be essential to our understanding of the principles underlying RNA folding and catalysis and for future predictions of three-dimensional structures from RNA sequence.

That natural, self-assembling RNA molecules tend to make intensive use of a relatively small number of building blocks is suggested indeed by the state of our knowledge concerning terminal loops and their interactions. The sizes and sequences of terminal loops are extremely biased in large natural RNAs with a stable three-dimensional structure; loops with four nucleotides and a GNRA consensus sequence (R stands for a purine and N for any base) may constitute up to one third of the total in some molecules (e.g. Woese *et al.*, 1990). A first indication

that GNRA loops frequently participate in RNA tertiary interactions came from comparative sequence analyses of group I self-splicing introns (Michel and Westhof, 1990). These analyses revealed two cases of phylogenetic covariation in which a GUAA loop and a C:G pair in a distant helix had exchanged repeatedly during evolution with the combination of a GUGA loop and a U:A pair. These covariations were proposed to result from a direct contact between the third base of the loop and the shallow ('minor') groove side of the base pair. Since then, additional instances of the same type of covariation have been found in bacterial RNase P RNA (Brown *et al.*, 1996), ribosomal RNA (Gutell, 1996) and group II self-splicing introns (Costa *et al.*, 1997). Moreover, all the interactions proposed to exist in self-splicing introns have been checked to be compatible with biochemical evidence (Jaeger *et al.*, 1993, 1994; Murphy and Cech, 1994; Costa and Michel, 1995; Chanfreau and Jacquier, 1996; Costa *et al.*, 1997).

Our first atomic resolution picture of a GNRA loop interacting with RNA was brought by a packing contact between two consecutive C:G pairs and a GAAA loop in crystals of the hammerhead ribozyme (Pley *et al.*, 1994). In addition to confirming that GNRA loops dock into the shallow groove of RNA helices, this structure revealed a network of hydrogen bonded contacts involving 2' hydroxyl groups on both the loop and loop receptor sides. More recently, the crystal structure of a 160 nucleotide (nt) domain from a group I intron (Cate *et al.*, 1996a) has provided us with one instance of an intramolecular contact involving a GAAA loop. The receptor in that case does not consist of two C:G pairs, but of an 11 nt motif (CCUAAG..UAUGG) that had already been shown by Murphy and Cech (1994) and Costa and Michel (1995) to interact with GAAA loops in self-splicing introns (this motif is also present in the RNase P RNA of some Gram-positive bacteria; Tanner and Cech, 1995).

Despite these advances, our understanding of tertiary interactions involving GNRA loops remains fragmentary. Partners have not yet been identified for a majority of GNRA loops in large natural RNAs. Although a number of these loops will probably turn out to be recognized by proteins (e.g. Gluck *et al.*, 1992), others, especially in self-assembling molecules, must be contacted by yet unidentified RNA receptors. Neither has there been any comprehensive investigation into the specificity of currently known receptors towards the various members of the GNRA family. Published phylogenetic and biochemical evidence could be taken to argue that the smaller receptors, which seemingly consist of only two base pairs, are poor discriminators, while the (CCUAAG..UAUGG) sequence, which is so frequent in self-splicing introns, is highly specific for GAAA loops (Costa and Michel, 1995). However, specific partners are likely to exist for other GNRA loops as well. The question then is whether these motifs are also used by nature, and if not, why not?

We reasoned that *in vitro* selection of RNA motifs capable of recognizing GNRA loops would not only constitute the most powerful strategy to recover any missing receptors for these loops, but should also greatly help our understanding of the structural principles that govern these tertiary interactions. Accordingly, we have devised an *in vitro* selection system suitable for the isolation of RNA motifs that specifically bind terminal loops and we have used this system, which is based on mutual recognition of a group I ribozyme and its substrate, to look for molecules that would recognize the GUGA and GAAA loops. After seven rounds of selection and amplification, variants of previously identified receptor motifs were found to predominate among selected molecules. However, both final pools also contain a number of new receptor sequences. These sequences are in no way inferior to previously identified motifs in terms of efficiency of recognition, but show novel patterns of discrimination, especially between loops that differ from one another by their second nucleotide. Rationalization of these results with the help of available structural data leads to general principles regarding the interaction of GNRA loops with their receptors and provides some insights into the mechanisms that underlie the evolution of loop–receptor partnerships in nature.

Results

In vitro selection of GUGA- and GAAA-binding motifs

Our *in vitro* selection system is based on the *td* molecule, a group I intron that interrupts the thymidylate synthase gene of bacteriophage T4 (Belfort *et al.*, 1987). We have shown previously that in the wild-type intron, the GUGA loop at the tip of helix P2 specifically binds the CU:AG sequence of helix P8, at positions four and five when counting from the base of the latter helix (Figure 1A; Costa and Michel, 1995). The *td* intron was transformed into a bimolecular system (Figures 1 and 2) by splitting a molecule composed of the intron itself and a short 5' exon into a 'substrate', formed by hairpin structures P1 and P2, and a 'core', which comprises the rest of the intron (see Materials and methods). We verified that co-incubation of these two pieces in the absence of the guanosine cofactor of group I self-splicing is followed by attack of the normal 5' splice site within the P1 helix by the terminal G residue of the intron, resulting in a new covalent bond between the two ends of the intron sequence. This addition reaction, which rests on the ability of the core to recognize and properly position the substrate into its active site, forms the basis for our selection procedure, which is shown in Figure 2 and is very similar to the one used by Robertson and Joyce (1990). 'Chimeric' core–substrate products are reverse-transcribed with an oligonucleotide primer designed to ensure selective recovery of those core molecules having catalyzed nucleophilic attack at the proper 5' splice site. The next step consists of PCR amplification of cDNA molecules with a set of oligonucleotides that introduces a T7 promoter and deletes all remaining substrate nucleotides from the core. T7 transcription of the resulting DNA matrices yields a new population of core molecules ready for another round of selection.

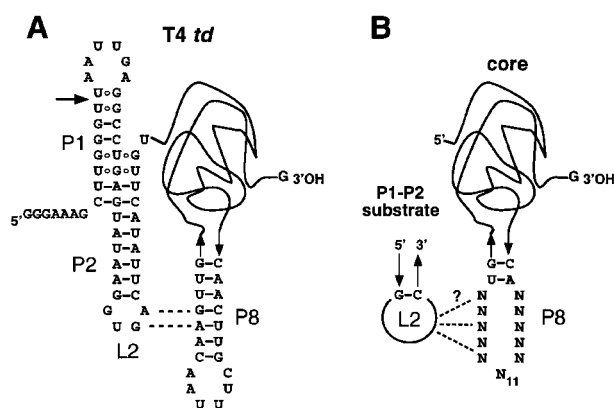


Fig. 1. Conversion of the *td* intron of bacteriophage T4 into a two-piece system to allow *in vitro* selection of co-adapted L2 and P8 sequences. **(A)** Our original *td* construct. The arrow points to the 5' splice site. The *td* 5' exon was replaced by a 14 nt sequence, the last seven bases of which belong to the original *td* sequence. Dashed lines indicate the interaction between the GUGA L2 loop and its receptor in the P8 helix. Note that the UGAG sequence at the P1–P2 junction can pair with either the sequence to its left (5' of P1) or the one to its right (3' of P2). **(B)** Constructs used for *in vitro* selection. The sequence of the P1–P2 'substrate' piece is that shown in (A): only the sequence of the L2 loop differed from one substrate to the next. The second piece was derived from the catalytic core of the intron, beginning immediately 3' of the P1–P2 sequence and ending with the intron 3' terminal G (a G was added at the 5' end of this piece, to allow efficient transcription; see also Costa and Michel, 1995). The distal section of the P8 hairpin structure was replaced by a random sequence of 21 nucleotides. Core molecules are selected for their ability to interact with the L2 loop of a P1–P2 substrate.

Preliminary experiments (not shown) allowed us to verify that just like attack by free guanosine (Costa and Michel, 1995), the efficiency of the addition step depends strongly on the nature of the L2 and P8 partners. Using a wild-type *td* core, the addition reaction was optimal with a P1–P2 substrate carrying a wild-type, GUGA L2 loop, much poorer with a GAAA loop, and barely detectable when the loop was UUCG (UUCG loops do not interact with known receptors for GNRA loops; Jaeger *et al.*, 1994; Murphy and Cech, 1994). We then constructed a population of intron core molecules (see Materials and methods) in which 16 of the 20 nucleotides of the original P8 hairpin of intron *td* had been replaced by a random sequence of 21 bases (Figure 1B). Five extra positions were added to the wild-type structure in order to avoid missing potential binding motifs larger than the wild-type one. On the other hand, the first 2 bp of P8 were left unchanged: we reasoned that maintaining base pairing at the base of P8 would further the formation of hairpin-like structures and reduce the risk of disturbing the overall architecture of the core. An additional reason for leaving the nucleotides at the base of P8 unaltered is that these residues tend to be well-conserved among relatives of the *td* intron (members of subgroups IA and IB; see Michel and Westhof, 1990) and may therefore be involved in tertiary contacts. In contrast, there is a complete lack of sequence conservation in the rest of P8 in those members of subgroups IA and IB which lack a P2 stem (Michel and Westhof, 1990): this observation, which strongly suggests that the section of P8 that was randomized, has no other function in *td* and related introns other than binding the L2 loop, vindicates the choice of P8 as a

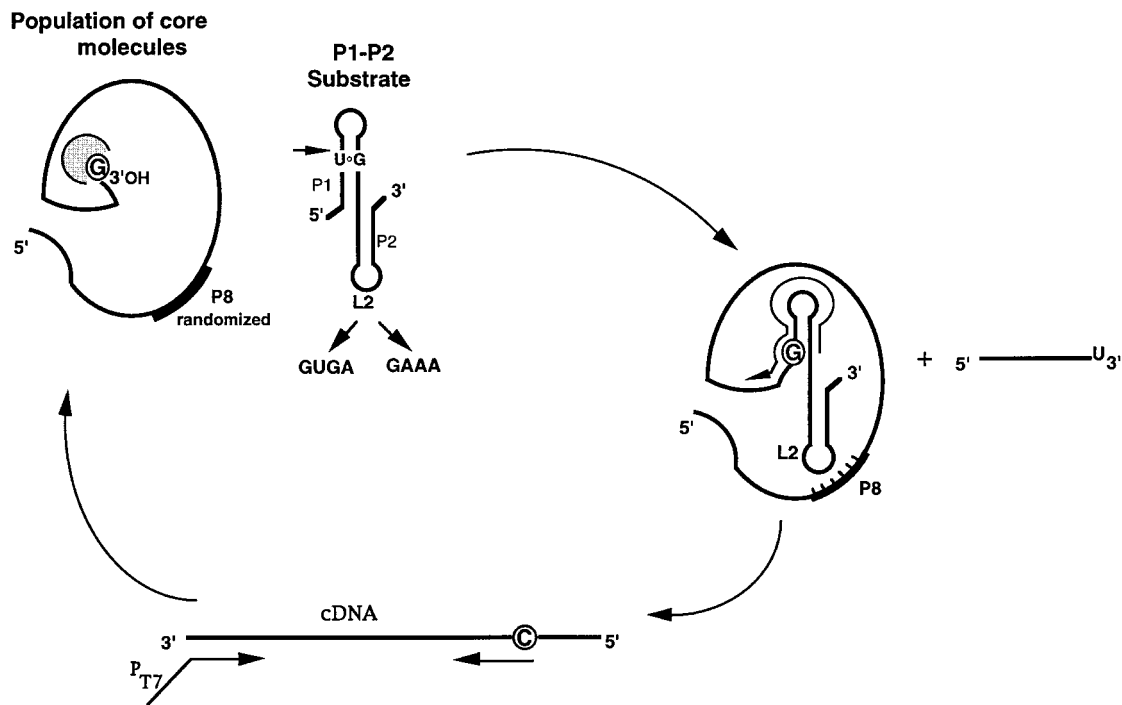


Fig. 2. *In vitro* selection procedure. The horizontal arrow above the U:G wobble pair of the P1–P2 substrate marks the site of attack by the intron terminal G (circled), which is shown sitting in the group I guanosine-binding site (symbolized by a dark grey sector). Binding of the substrate by the intron core is dependent on recognition of the L2 loop by an appropriate P8 sequence. An oligonucleotide complementary to the distal part of P1 and the last nucleotides of the intron (see Materials and methods) allows selective reverse transcription of reacted core molecules carrying a correct core–substrate junction. The same initial population of core molecules was used to carry out two selection experiments with substrates that differed only by the sequence of their L2 loop (GUGA or GAAA). See Materials and methods for experimental details.

target for the type of selection that was carried out in this work.

With a mass of 7.5 pmol, our initial pool of core molecules must have contained some 64% $[1 - (1 - 4^{-21})^n]$, with $n = 7.5 \times 10^{-12} \times 6.023 \times 10^{23}$ of all possible P8 sequences of 21 nucleotides. The same initial population of core molecules was used to carry out two selections in parallel: one for binding of the GUGA loop and the other one for binding of the GAAA loop. For each experiment, seven rounds of selection and amplification were performed, resulting in two final selected pools, designated as the 'GUGA' and 'GAAA' pools. Figure 3 shows the products of addition reactions of the *td* core, of the initial pool and of the final, selected pools. In each case, a major reaction product whose existence depends on the presence of a substrate molecule can be seen: its electrophoretic mobility is the one expected for a molecule consisting of the intron core and the 3' portion of the substrate. Comparison of the addition reactions of the initial pool (Figure 4A) with those of the final pools (Figure 4B) is indicative of the degree of improvement achieved by selection (note that the conditions used in Figure 4B, which are those of the last round of selection, are more stringent than the ones of Figure 4A, which correspond to the first round). Moreover, adaption was specific, since final pools were found to react much better with the substrate with which they had been confronted (Figure 4B). Interestingly, reaction of the *td* core with the wild-type (GUGA) substrate was significantly slower than that of the final GUGA pool, betraying the fact that the former molecule is not optimal in terms of substrate binding (see below). However, an even more efficient reaction was

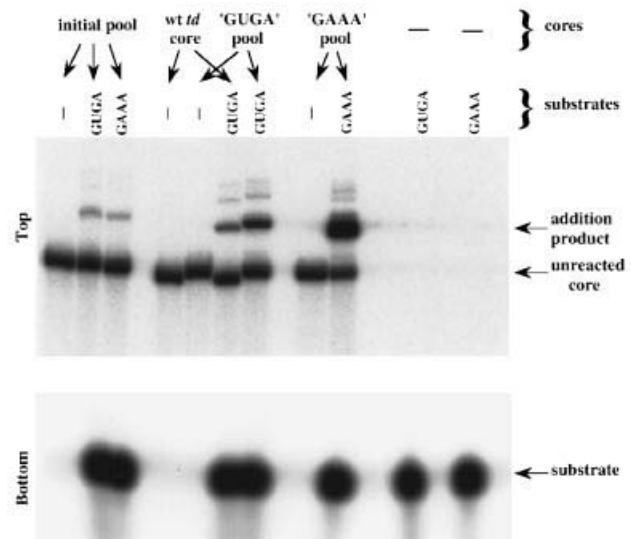


Fig. 3. Addition reaction products of the initial and final pools. Core molecules were incubated with a substrate carrying a GUGA or GAAA loop for either 20 min (initial pool) or 10 min (final GUGA and GAAA pools and wild-type *td* core); control reactions with no substrate or no core were included; see legend to Figure 4 for reaction conditions. Products were identified from their electrophoretic mobilities (expected sizes were 235 nt for core molecules of the various pools and 230 nt for the wild-type *td* core, with 36 additional nucleotides for addition products). Only the top and bottom of a 6% denaturing polyacrylamide gel are shown (no products were detected in the rest of the gel). The bottom part of the gel was less exposed than the top in order to compensate for the higher specific activity of the substrates (substrate over core labelling ratios ranged from 32 to 34.8).

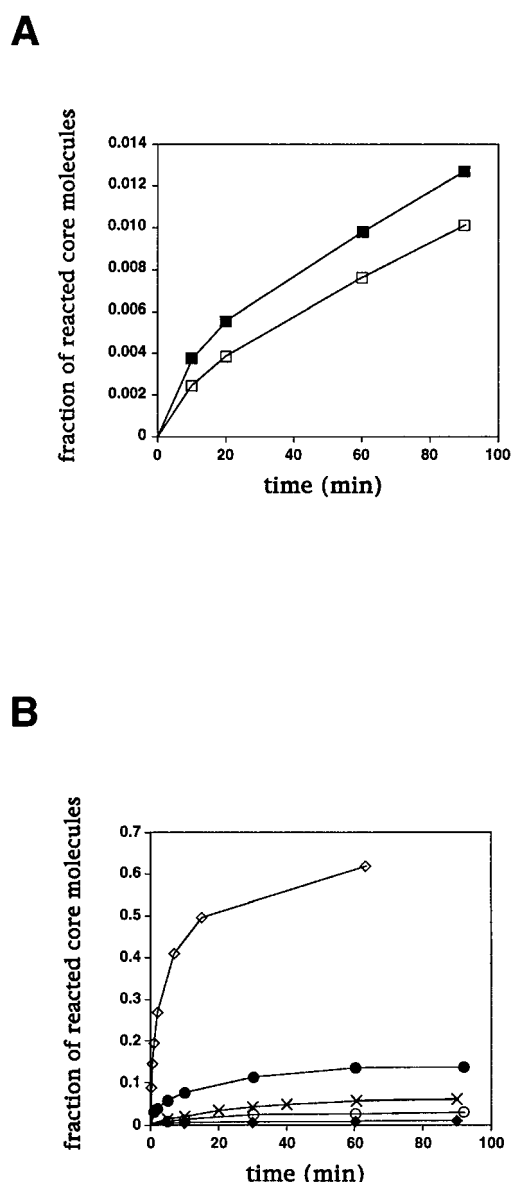


Fig. 4. Kinetics of addition reactions of P1–P2 substrates with intron core molecules. Filled symbols, GUGA substrate; empty symbols, GAAA substrate. (A) Time course of addition reactions resulting from incubation of the initial pool with each one of the substrates under the reaction conditions used for the first round of selection (45°C, 50 mM magnesium buffer, 4 μ M substrate, 0.4 μ M core molecules). (B) Time course of addition reactions of the final GUGA (circles) and GAAA (diamonds) pools and of the wild-type *td* core (crosses); the latter was incubated only with the wild-type, GUGA substrate. Reaction conditions were as in (A), except that the buffer contained 20 mM magnesium.

obtained for the combination of the final GAAA pool and GAAA substrate.

Because of competition between addition and hydrolysis at the core–substrate junction, all reactions in Figure 4B eventually level off. Importantly, by using chimeric core–substrate molecules carrying previously characterized, matched and mismatched L2 $\times$ P8 combinations (Costa and Michel, 1995), we checked (data not shown) that hydrolysis is indeed specific to the 3' end of the intron and that its rate is largely insensitive to the strength of the L2 $\times$ P8 interaction. Therefore, hydrolysis does not antagonize selection for binding efficiency.

Several families of receptors for GUGA and GAAA tetraloops

We sequenced the entire intron core of some 30 individuals from each of the final pools. As expected from our use of a high-fidelity DNA polymerase for all amplifications, few molecules were found to carry mutations outside the randomized P8 segment. Moreover, the positions affected were never the same, except for a recurrent G to A mutation at the first position of the J8/7 segment, immediately 3' of P8. This mutation is present in 18% of clones from the final GUGA pool (Figure 5A) and its role in the recognition of substrate molecules by the core will be described elsewhere. These data are consistent with the view that the L2 $\times$ P8 interaction was the main target for improvement of substrate binding.

Alignment of the P8 segments of selected individuals (Figure 5) revealed the presence of several classes of receptor for each loop. In the GUGA pool, up to five receptor families—defined on the basis of the sequence at positions four and five of P8—may be distinguished (Figure 5A). Much as expected, the most abundant family is the one with a CU:AG sequence: CU:AG helices were shown previously to be specific receptors for GUGA loops (Michel and Westhof, 1990; Jaeger *et al.*, 1994). Importantly, this sequence is most often located at the same place as in the *td* intron, which confirms that splitting the *td* molecule into two transcripts does not alter the relative positioning of the P2 and P8 helices. In fact, most sequences may be aligned in such a way that they share either base pair 4 or 5 with the *td* sequence. However, class IV molecules, which have a G[N₀₋₄]GCU:GGCC consensus sequence, appear unrelated to either the *td* molecule or other clones. Finally, the observation that 11 of the 16 clones in subclasses IA to IC have a C:G pair on the distal side of their CU:AG receptor motif suggests that this position could also be involved in some kind of contact with the GUGA loop.

Two families of receptors were recovered from the GAAA pool (Figure 5B). Aside from the five class II molecules, in which a CCC:GGG consensus sequence from positions 3 to 5 is followed by an asymmetric internal loop, the majority of the clones harbour either a canonical version or variants of an 11 nt motif (CCU-AAG...UAUGG), which we have shown previously to bind GAAA loops with remarkable affinity (Costa and Michel, 1995). Among minor variations that were selected, the most frequent ones are a C instead of an A at position 5 (see Figure 6 for numbering of the 11 nt receptor), and an A:C combination instead of G:U at the tip of the receptor. However, more divergent variants of the same 11 nt motif were also recovered. In subgroup IB, the UAA sequence of the 5' branch is replaced by UGY and in three of the four clones, various mismatches substitute for the G:U pair. In subgroup IC, the same UAA sequence becomes UGNA and is most often followed by a C:G pair.

Kinetic characterization of some selected motifs

In both final pools, the major class of sequences corresponds to a motif that was already known to be a specific receptor for the loop to which this pool was confronted. Therefore, it is clear that our selection system reproduces at least some of the evolutionary forces that shape receptors in nature. One possible difference, however, between

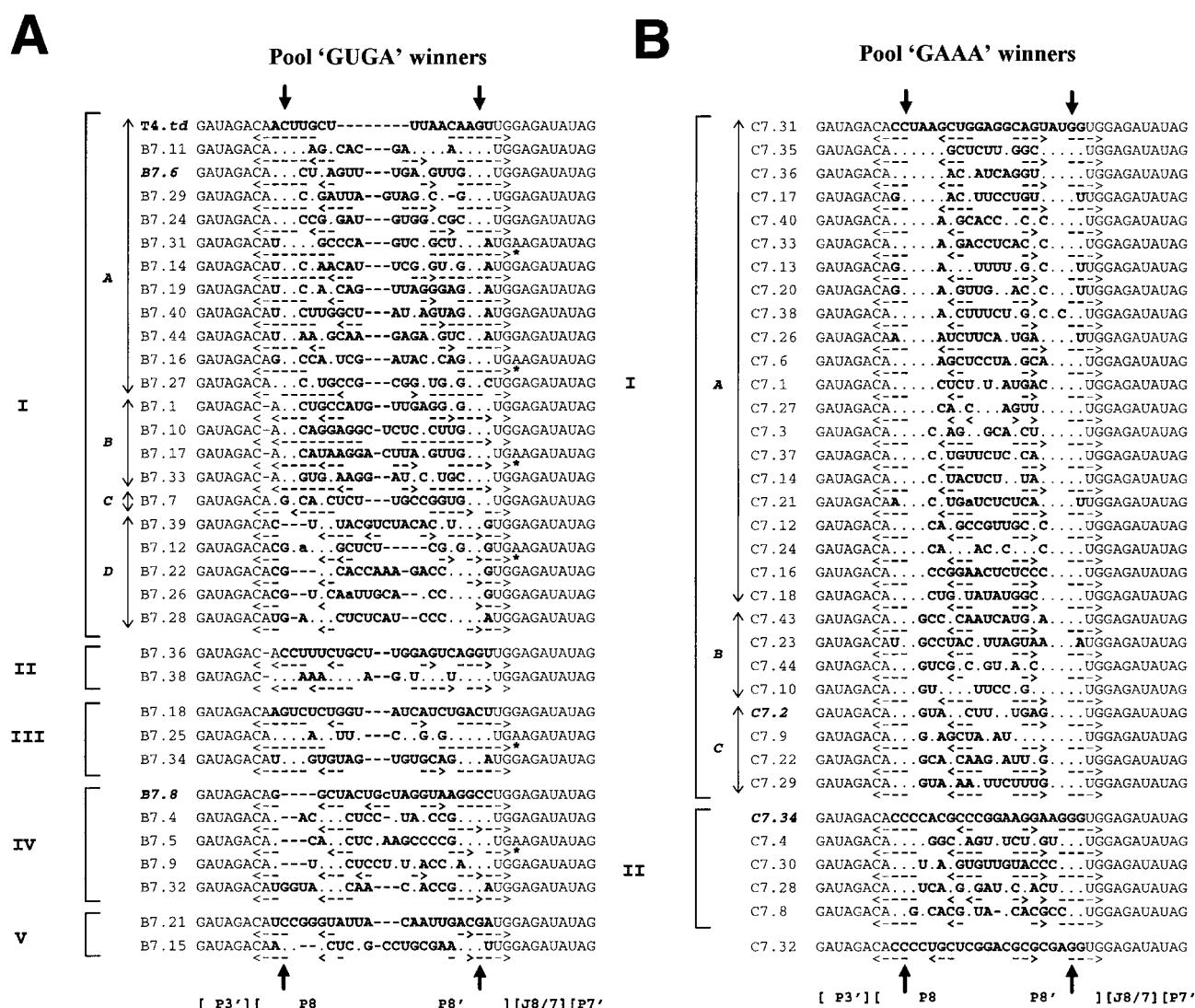


Fig. 5. Sequence alignment of the P8 domains of selected core molecules. The segment of sequence that was randomized is shown in bold type; for each class, one representative sequence is shown in full, with identical nucleotides being replaced by dots in the other sequences. Potential secondary structure pairings are indicated by dashes and divergent arrows below a sequence. Vertical arrows point to the C:G base pair that is common to most currently known receptors and was found to interact with the last A of GNRA loops in crystal structures (see text). Residues in lower case stand for dinucleotides of the corresponding residue. Roman numerals and letters designate classes and subclasses of P8 sequences (see Results). Code names in bold type are those of the clones that were characterized kinetically. (A) Alignment of 33 pool GUGA sequences with the wild-type T4.td sequence. In classes IB and II, a gap was introduced between the first and second nucleotides of P8 to force alignment of the receptor sequence with that of *td* (a number of natural group I sequences have a gap at the same position; see Michel and Westhof, 1990). A recurrent G to A mutation of the first nucleotide of the J8/7 segment is marked by an asterisk (see Results). (B) Alignment of 35 clones belonging to the final GAAA pool.

natural conditions and ours is the absence in our case of counterselection for cross-recognition of a receptor by other members of the GNRA family. In order to estimate not only the efficiency of recognition, but also its specificity, we have resorted to kinetic characterization of loop-receptor pairs: selected molecules were assayed for their ability to bind not only their cognate loop but also those among the other members of the GNRA family that differed from it by one nucleotide.

Rather than using addition reactions, the analysis of which is complicated by hydrolysis at the junction of the core and substrate, we went back to a reaction that mimics the first step of self-splicing. In the presence of excess guanosine, 3'-truncated intron core molecules act as true catalysts and promote specific cleavage of P1-P2 substrates at the 5' splice site (Costa and Michel, 1995).

Figure 7 shows time courses of cleavage in the presence of excess enzyme for five different L2 substrates incubated with the same ribozyme; all L2×P8 combinations that were tested yielded similar data, compatible with first-order kinetics. Single-turnover reactions were performed at ribozyme concentrations much lower than K_m (see Materials and methods) so as to estimate directly the value of k_{cat}/K_m (Table I) from rates of reaction. Importantly, all ribozyme-substrate combinations were checked by polyacrylamide gel electrophoresis (data not shown) to cleave at the correct 5' splice site.

We chose to characterize clones B7.6 and B7.8 from the GUGA pool (Figure 8). Like the majority of class I molecules, clone B7.6 differs from the *td* intron in that it has a C:G rather than a U:A pair on the distal side of its CU:AG receptor sequence. However, as shown in Table I,

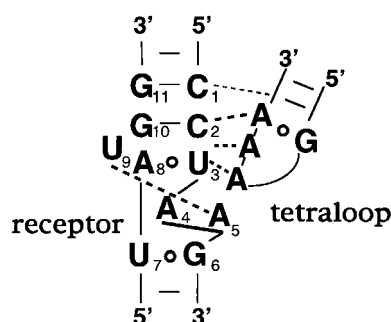


Fig. 6. Schematic diagram of the structure of the 11 nt receptor and its interactions with a GAAA loop. The figure, which was redrawn from Cate *et al.* (1996), is meant to evoke the actual crystal structure. Dashed lines stand for hydrogen bonded contacts. The first adenosine of the tetraloop contacts a reverse Hoogsteen U:A pair in the receptor (drawn as U3:A8). The second adenosine recognizes the receptor via 2' hydroxyl contacts. The third adenosine, which makes a non-canonical base pair with the first nucleotide of the loop, also contacts the C2-G10 pair of the receptor. Nucleotides A4 and A5 are drawn side by side, to indicate that they form a platform by stacking on U7 and G6, respectively (stacking extends in fact from G6 and A5 to the adenines of the loop). U9 bulges out of the middle part of the receptor and contacts A5 (see text). The thin dotted line indicates a 2' hydroxyl-mediated interaction between the first C:G pair of the receptor and the base pair above the tetraloop.

neither the affinity of the core for the GUGA substrate (as far as it can be inferred from values of k_{cat}/K_m ; see Discussion), nor its pattern of discrimination between different GNRA loops is significantly altered by this substitution. On the other hand, the rates measured for clone B7.8, which carries the highly divergent class IV consensus sequence, suggest a different network of molecular contacts.

In the GAAA pool, clone C7.2 was investigated in order to determine to what extent the replacement of UAA by UGNA in the 11 nt receptor motif might interfere with binding of GAAA and the other tetraloops. As can be seen in Table I, the answer is that this substitution has negligible effects on those parts of the receptor that directly contact the tetraloop. In contrast, the C7.34 molecule, which is typical of class II GAAA receptors, shows widely different preferences.

Role of the second loop nucleotide

The recently determined X-ray structures of two loop-receptor pairs (Pley *et al.*, 1994; Cate *et al.*, 1996a) provide a framework within which sequence and kinetic data may be discussed in terms of specific contacts between GNRA loops and their receptors. Confrontation of biochemical and structural data is especially illuminating in the case of loop nucleotide 2, whose role in loop-receptor recognition had been underestimated. Thus, the receptor for GUGA loops has until now been assumed to consist of only two consecutive base pairs with a CU:AG sequence (Michel and Westhof, 1990; Jaeger *et al.*, 1994). However, in 13 of the 15 clones in subclasses IA and IB of the GUGA pool, helix P8 extends beyond the CU:AG sequence at positions 4 and 5. When the terminal loop that caps the CC:GG receptor of Pley *et al.* (1994) is replaced by additional Watson-Crick base pairs so as to lengthen the receptor helix (see legend to Figure 9), it becomes apparent that selective pressure for base pairing at position 6 could be due to the formation of a hydrogen bond (Figure 9A)

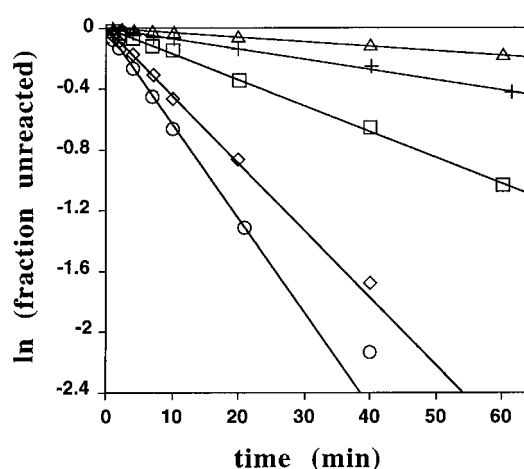


Fig. 7. Examples of time courses of cleavage reactions of P1-P2 substrates by a *td*-derived ribozyme. Reactions were carried out under single-turnover and k_{cat}/K_m conditions (see Results and Materials and methods) with 0.5 μM wild-type *td* ribozyme and 0.05 μM substrate (K_m was previously determined to be 13.5 μM for the combination of the *td* ribozyme and a GUGA substrate; see Costa and Michel, 1995). Ordinates are natural logarithms of the unreacted fraction. Plots could be fitted to single exponentials after correction for the presence of at most 12% of inactive substrate molecules. P1-P2 substrates differed only by the sequence of their L2 loop. Symbols: (○), GUGA; (◇), GCGA; (□), GUAA; (+), GAGA; (△), GGGA.

between the O2 acceptor of U (or C) at loop position 2 and the 2' hydroxyl group on the 5' side of P8 bp 6. Therefore, the CU:AG and CC:GG receptors for GYGA and GYAA loops should better be regarded as being CUN:N'AG and CCN:N'GG, with N:N' indicating canonical base pairing.

While modelling a continuous A-type RNA helix in front of GNRA loops, we also noticed a severe clash between the NH_2 group of a guanine at loop position 2 and the ribose on the 5' side of what would be bp 6 in P8 (Figures 9B and C). Accordingly, the GGGA loop is very poorly bound by the *td* and B7.6 ribozymes (Figures 7 and 8 and Table I). In fact, even with an A at loop position 2, a close contact still occurs, which probably explains the 7- to 11-fold preference of extended helices for GYRA over GARA and also, the lack of base pairing at position 6 in the GAAA pool (Figure 5B).

In nature as well, few GAAA loops are faced by continuous helices and a rather frequent substitute for the 11 nt motif is precisely the CCC:GGG interrupted helix of class II clones. In one extreme case, at the P5 site of subgroup IC self-splicing introns, a majority of sequences have a CCC:GGG helix, with two Us on top, in front of a GAAA L9 loop (Michel and Westhof, 1990; Damberger and Gutell, 1994); 11 nt motifs are missing altogether at that location in subgroup IC introns, probably because the backbone undergoes a sharp turn on the distal side of helix P5 (see Cate *et al.*, 1996a). Even when a GAAA (or GAGA) L9 loop happens to interact with a continuous helix, the base pair in front of loop position 2 is more often U:G than not (8 out of 11 continuous helices facing GARA loops in subgroup IB introns; data not shown): substitution of a U:G wobble pair for a Watson-Crick one pushes the uracil and its ribose away from the shallow (minor) groove, thus making room for the adenine at loop position 2.

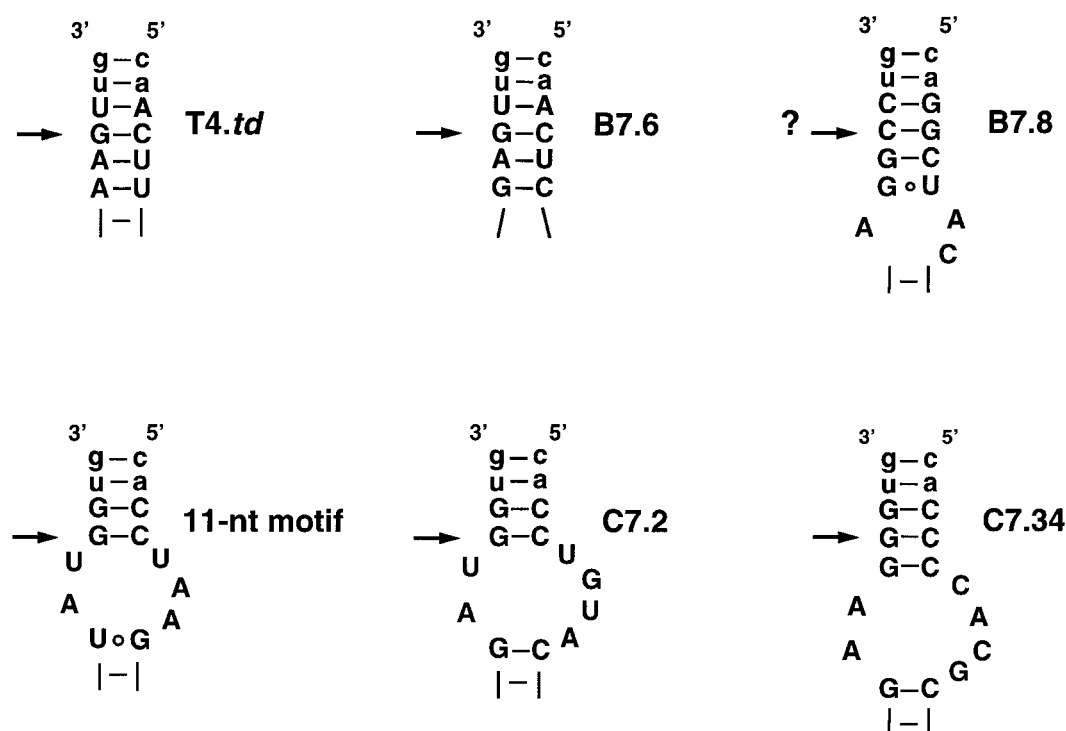


Fig. 8. Receptors tested for their ability to discriminate between GNRA loops. The relevant part of the P8 sequence of each one of the ribozymes tested is shown (we checked that these ribozymes differ only by the sequence of their P8 segment). The two base pairs in lower case are the ones that were not randomized. The base pair interacting with the last A of GNRA loops is indicated by an arrow. This pair is C:G in all currently known receptors, with the only exception of the class III and class IV clones of the final GUGA pool, which have a G:C pair at that position (indicated by a question mark in clone B7.8).

Table I. Values of k_{cat}/K_m ($\times 10^5/\text{min}/\text{M}$) for cleavage reactions of P1–P2 substrates by *td*-derived ribozymes

Ribozymes	Substrates (L2 loop)							
	GYRA				GRRR			
	GUGA	GCGA	GUAA	GCAA	GAGA	GGGA	GAAA	GGAA
P8 T4.td (CUU:AAG)	1.32	0.89	0.33	ND	0.13	0.048	0.09 ^a	ND
P8 B7.6 (CUC:GAG)	1.14	0.95	0.26	ND	0.10	0.042	ND	ND
P8 B7.8 (GCUAC..AGGC)	3.42	0.039	6.97	ND	0.37	0.05	ND	ND
P8 ^a (CCU:AGG)	0.36 ^a	ND	1.03 ^a	ND	ND	ND	0.12 ^a	ND
P8 11 nt motif (CCUAAG..UAUGG)	0.03 ^a	ND	0.38	0.13	0.79	ND	47.5	0.18
P8 C7.2 (CCUGUAC..GAUGG)	ND	ND	0.27	0.12	0.67	ND	47.5	0.22
P8 C7.34 (CCCCACGC..GAAGGG)	ND	ND	2.86	0.83	2.46	ND	6	17.6

Ribozymes are designated by the name of the clone in Figure 5 and the relevant section of their P8 receptor (P8 bp 4 is in bold type). Values of k_{cat}/K_m were determined by dividing first-order rates (determined from the type of data shown in Figure 7) by ribozyme concentrations. ND; not determined.

^aData from Costa and Michel (1995). These values, which were determined under the same experimental conditions and had been normalized to the k_{cat}/K_m value of the wild-type L2×P8 combination, were multiplied here by the k_{cat}/K_m of the *td*×GUGA combination (1.32).

Novel receptors with different specificities

Receptors that lack a base pair in front of the second base of GNRA loops may have been selected not only to avoid clashes, but because of the ability of some of them to interact directly with that base. Thus, clone C7.34 (Figure 10), a class II member of the GAAA pool, does not markedly discriminate between GAAA, GAGA and GUAA, but reacts more strongly with the GGAA sequence. As seen in fact from the values of k_{cat}/K_m (Table I), the combination of the C7.34 motif with a GGAA loop is second only to that of the 11 nt motif with GAAA. It should be interesting to investigate the other class II clones, which have essentially the same secondary structure as

clone C7.34 but differ from it by the sequence of their internal loop, for their ability to discriminate between different members of the GNRA family.

The case for direct recognition of the second loop nucleotide is even more convincing with the B7.8 receptor from the GUGA pool. The B7.8 molecule shows a 90-fold preference for GUGA over GCGA, in total contrast to the *td* and B7.6 ribozymes, which fail to distinguish between these two loops (Table I). There must, therefore, exist a direct contact between some nucleotide(s) in the internal loop of the B7.8 receptor (Figure 10) and either the N3-H or O4 groups of the U at loop position 2. Clone B7.8's best choice is in fact GUAA, which is readily

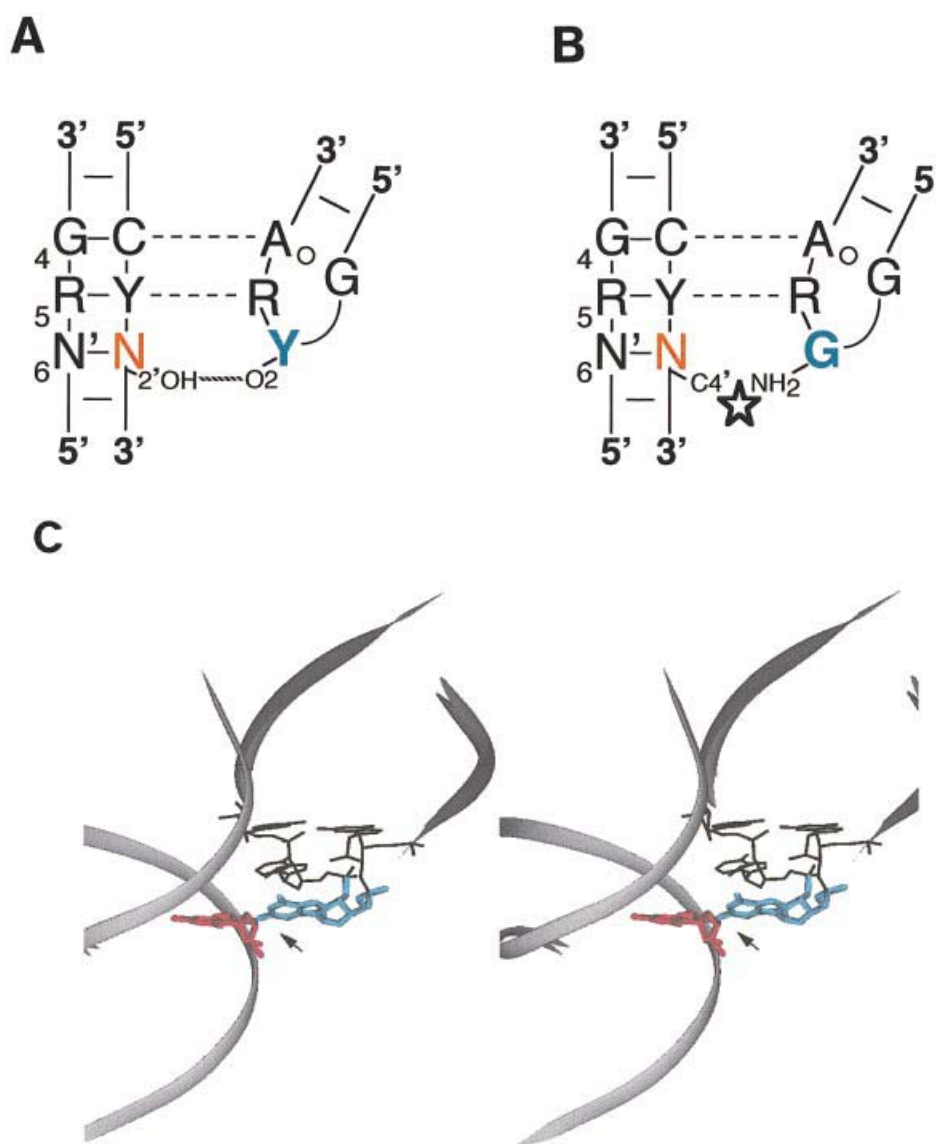


Fig. 9. Contacts between the second nucleotide of GNRA loops and receptors consisting of a continuous helix. (A) Schematic drawing of the interactions between a GYRA loop and a continuous helix. The heavy dotted line stands for a hydrogen bond between the O2 group of the pyrimidine at loop position 2 (in blue) and the 2' hydroxyl group of helix nucleotide N (in red, corresponding to position 6 on the 5' strand of P8). (B) Schematic drawing of a steric clash between the NH₂ group of the second G of a GGRA loop (in blue) and the ribose of helix nucleotide N, at position 6 on the 5' strand of P8 (in red). (C) Stereo drawing of the steric clash schematized in (B). The figure was drawn with DRAWNA (Massire *et al.*, 1994) from coordinates generated by grafting an A-type RNA helix on the 3' side of the CC:GG receptor of Pley *et al.* (1994) and replacing the second nucleotide of their GAAA loop by a G. The arrow points to the clash between the NH₂ group of that G (in blue) and the ribose of the nucleotide (in red) immediately 3' of the CC sequence (this corresponds to position 6 on the 5' strand of P8). Regular A-type helical backbones are drawn as grey ribbons and bases engaged in canonical base pairing have been omitted.

explained by its fifth P8 pair being C:G, instead of U:A, as in most of the other clones of the GUGA pool (see Table I, Costa and Michel, 1995 and references therein). However, the B7.8 sequence also differs from the consensus receptor for GUGA loops (and in fact from the vast majority of known receptors for GNRA loops), at position 4 of P8, where it has a G:C rather than a C:G pair; a C:G pair would seem preferable at that location, because its shallow groove NH₂ group is better oriented to interact with the N3 acceptor of the last adenine of GNRA loops (see figures in Pley *et al.*, 1994 and Cate *et al.*, 1996a). Judging from the data in Table I, this non-optimal combination is more than compensated by improved binding of the second loop nucleotide.

Variations on the 11 nt GAAA receptor

Such a large number of copies of the 11 nt receptor for GAAA loops were recovered from the GAAA pool (Figure 5B) that a rough estimate can be made of the variability between selected molecules of this motif, the structure of which is schematically drawn in Figure 6. We have attempted to compare the variations tolerated in subclass IA molecules—the ones with an A at position 4 of the motif (see Figures 5B and 6)—with those observed in nature, at three locations at which 11 nt motifs are particularly abundant (Figure 11). Such a comparison is all the more interesting since an X-ray structure of the 11 nt motif interacting with a GAAA loop was recently published by Cate *et al.* (1996a).

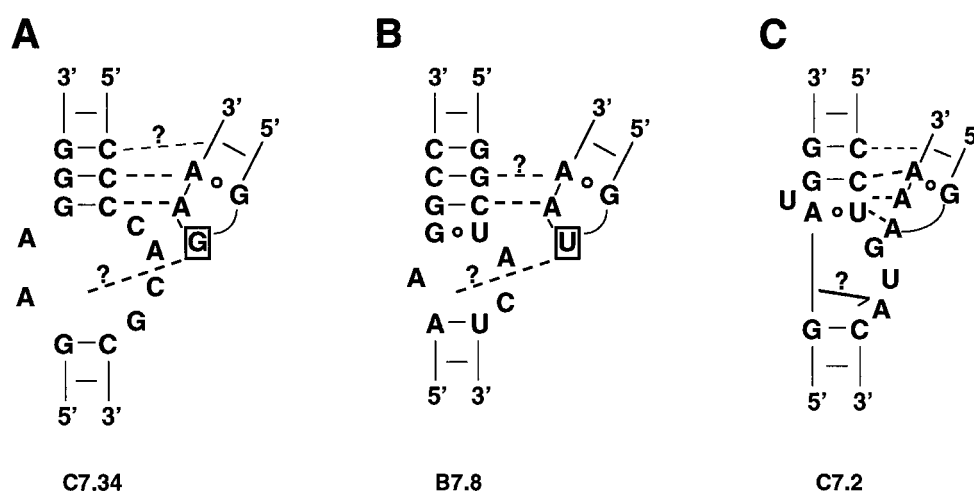


Fig. 10. New receptors for GNRA loops isolated by *in vitro* selection. Receptors are drawn according to their most probable secondary structure (Figure 5) and shown interacting with their preferred GNRA loop partner (see Table I). Bases in common with previously known receptors are assumed to interact with loop nucleotides in the same way as described previously (the only exception is the pair that faces the last A of the loop in clone B7.8). Bases specific to the C7.34 and B7.8 new receptors must be involved (in some unknown way) in recognition of the base at loop position 2 (see text). It is unknown whether the structure of the C7.2 receptor includes a platform similar to the one that exists in the 11 nt GAAA receptor (Figure 6 and Cate *et al.*, 1996).

The crystal structure of the receptor reveals that only four of its 11 nucleotides, the ones at positions 2, 3, 8 and 10, interact directly with nucleotides in the loop. Not too surprisingly, these nucleotides are invariant or nearly so, whether in *in vitro* selected sequences or natural ones. Conversely, most of the nucleotides that vary in selected molecules do so even more in nature (Figure 11). Among variable parts is the so-called A–A platform, which was certainly the least anticipated feature of the 11 nt receptor: by stacking respectively on the U and G of a wobble pair (Figures 6 and 12A), the As at positions 4 and 5 form a pseudo base pair that extends the distal helix towards the GAAA loop and the core of the receptor. Correct positioning of A5 would seem crucial for loop recognition, since that base is the one on which the adenines of the loop stack. Yet, both the bases on which the platform stacks and the platform itself are somewhat variable.

Of the three different instances of A–A platforms in the molecule crystallized by Cate *et al.* (1996b), two stack on G:U wobble pairs and the third one on a non-Watson–Crick A:U pair; as pointed out by the authors, these non-canonical geometries maximize stacking with the adenines of the platform (Figure 12A). Selection pressure against canonical base pairs and in favour of a wobble geometry is evident in our selected molecules, for an A:C combination is nearly as frequent as a G:U one at positions 6 and 7 and most of the remaining clones have pyrimidine–pyrimidine mismatches. As already noted by Tanner and Cech (1995), A:C substitutes frequently for G:U in natural sequences as well. Still, about as many of these sequences have A and U at positions 6 and 7 (Figure 11), as also does one of our selected clones; it seems reasonable to speculate that these bases do not form Watson–Crick pairs.

Concerning the platform itself, its most frequent variant in otherwise canonical motifs, whether from natural or subclass IA *in vitro* selected molecules, has a C at position 5. This substitution does not markedly alter stacking with G6 and should improve hydrogen bonding with the acceptor at N3 of A4 (Figure 12B). However, having a C

at position 5 entails the loss of a hydrogen bond with U9. U9 bulges out of the motif, but rather than pointing outwards, the base folds back towards A5, with which it forms at least one hydrogen bond (U9–N3 with A5–N1, see Figure 12A; a second bond may exist, between U9–O4 and A5–N6). U9 changes to C in clone C7.38 and is rather variable in natural sequences, being replaced not only by C, but also sometimes by A and G (Figure 11). However, these substitutions are not coordinated with the one from A5 to C5, which must mean that interaction between the bases at positions 5 and 9 is not an essential feature of the receptor.

In summary, not only variable positions, but also the nature of substitutions tend to be the same in natural and selected sequences. Still, the two sets differ in a major way at receptor position 4, where little variation exists in nature, whereas eight of the selected molecules have a G (subclasses IB and IC in Figure 5B). Moreover, in four of those eight clones, there is an extra nucleotide inserted between positions 4 and 6. Do platforms exist at all in those molecules (Figure 10C)? We attempted to address this issue by comparing clone C7.2 and a molecule with a canonical 11 nt motif for their ability to bind to, and discriminate between, different GNRA tetraloops (Table I). No significant difference could be observed, which we interpret as meaning that at least those bases that interact directly with the tetraloop and also probably the one (at position 5) on which the second base of the loop is stacked must be at roughly the same place in the two molecules. The sequence of clone C7.10, in which only positions 4 (G) and 5 (U) differ from the 11 nt consensus, also pleads in favour of these two nucleotides occupying the same locations as the two As of the canonical motif. In fact, whether or not platforms actually exist in clones of the IB and IC types, a deeper mystery is why such motifs should be so rare in nature. We know of only three instances of natural 11 nt receptors with a G at position 4. Interestingly, one of them, in the P5 stem of intron $\beta 22$.td (Bechhofer *et al.*, 1994), reads CCUGUAG..UAUGG, i.e.

has one extra nucleotide inserted 3' of G4, just as in the subclass IC clones.

Of the remaining bases, 1 and 11 are regarded as being part of the 11 nt receptor because of the strong preference for C and G at those positions in nature, at most sites of interaction with GAAA loops (e.g. Figure 11). The same

preference is manifest in selected molecules and since a backbone contact exists between the 2' hydroxyl groups of C1 on the one hand and the nucleotide 3' of the last A of the loop on the other, the need for a precise stacking geometry might be invoked to explain this sequence constraint. However, variations of the 1:11 base pair in natural molecules do not correlate with those of the pair that closes the tetraloop. It is also worth noting that although both classes of sequence in the GAAA pool show the same bias in favour of a C1:G11 pair, no such constraint exists in the GUGA pool.

Discussion

Compared with comparative analysis of natural sequences, *in vitro* selection is clearly a superior approach to problems of molecular structure and recognition. Molecules in nature are subject to a multiplicity of selective pressures and are the products of history: the world of sequences can only be explored step by step by natural selection. In contrast, we believe that the sequences in Figure 5A and B constitute largely unbiased samples of optimal or near-optimal solutions to the problem of binding tightly a terminal loop with a specific sequence and a precise location in three-dimensional space relative to the randomized segment. One reason for this is that the section of the P8 stem of the *td* intron that binds the L2 terminal loop, and which we chose to randomize, appears to make no additional contact with the rest of the intron: this was initially suggested to us by comparative sequence analysis of natural group I introns (see Results) and has been confirmed *a posteriori* by the variety of *in vitro* selected P8 sequences (an even greater diversity was observed among products of additional selections aimed at recognition of more divergent L2 sequences; M.Costa and F.Michel, in preparation). Also, adaption to substrate recognition did not involve sequences other than in P8 (with the exception of a single site 3' of that segment) and the process must have been essentially complete by the time the experiment was stopped, since there was no significant increase in pool reactivity during the ultimate round of selection (data not shown). Finally, care was taken to avoid population bottlenecks by making sure throughout the experiment that numbers of molecules of nucleic acids being manipulated were much larger than the estimated complexity of the pools undergoing selection: in fact, no two of the final sequenced clones are identical. In summary, there is every reason to believe that the ability to bind efficiently a given L2 loop and to position this loop correctly with respect to the intron active site were both necessary and sufficient conditions for a P8 sequence to

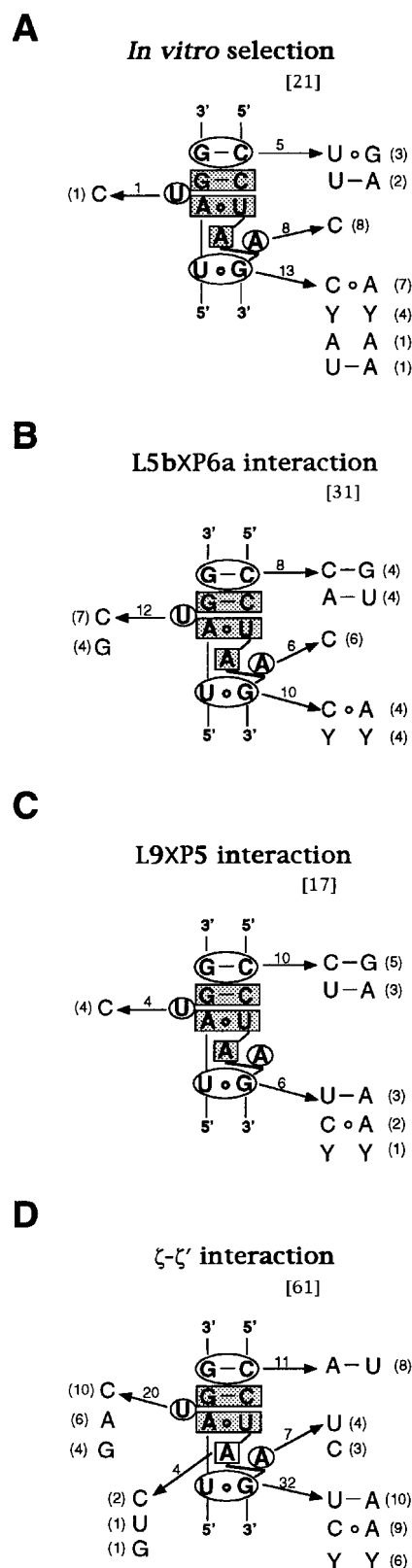


Fig. 11. *In vitro* versus *in vivo* variability of the 11 nt receptor of GAAA loops. Nucleotides framed with grey are invariant residues that constitute the receptor core. Numbers above each arrow correspond to the total number of changes observed at each position. Only the most frequent or relevant alternatives to the canonical sequence are shown. (A) Only subclass IA pool GAAA individuals (Figure 5B) were scored. (B) The L5b×P6a interaction is present only in subgroup IC1 and IC2 introns (Michel and Westhof, 1990). (C) The sequences scored are those of subgroup IA introns. (D) The ζ-ζ' interaction is present in most group II introns (Costa and Michel, 1995). Sequences of group I and group II introns came from our own databases.

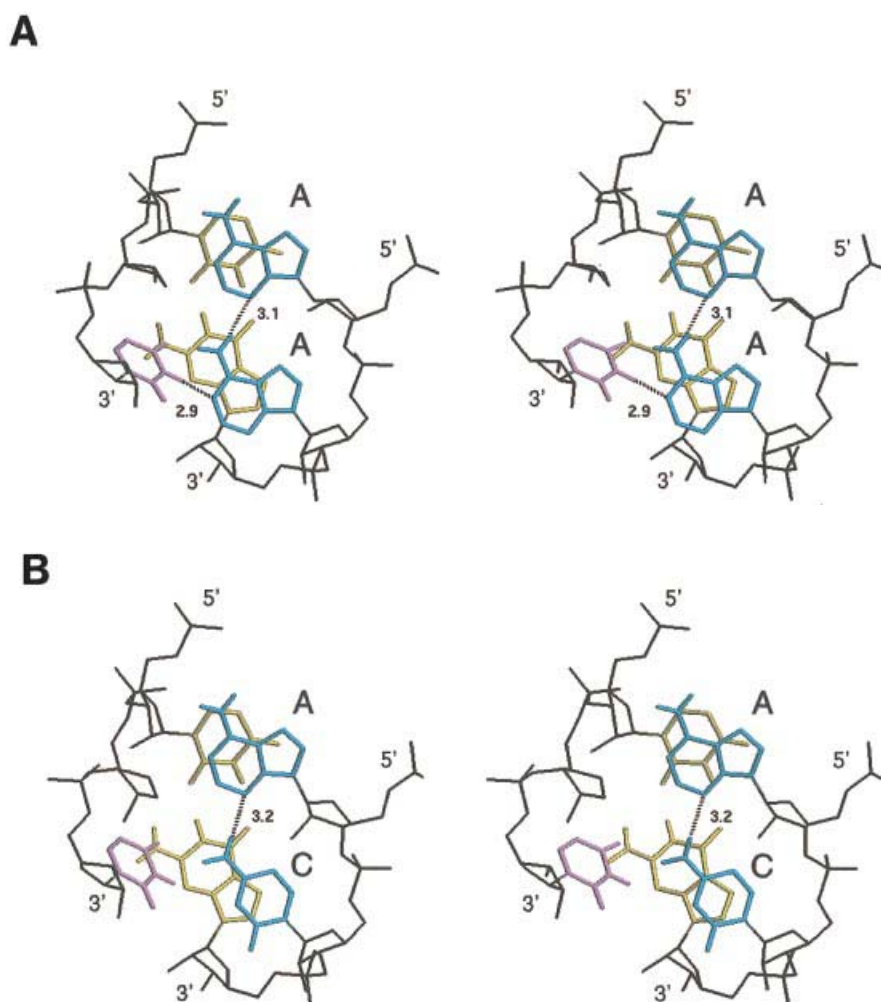


Fig. 12. Stereo drawings of the A–A platform and an *in vitro* selected variant of it. Views are from the top of the platforms in order to illustrate base stacking. Potential hydrogen bonds are indicated by dotted lines, together with the distance (in Å) between interacting atoms. (A) The A–A platform. Coordinates are from Cate *et al.* (1996a,b). Bases that form the platform are in cyan and the G:U wobble pair below them is shown in yellow. Receptor base U9 (in violet) contacts the second A of the platform. (B) An A–C platform. The structure shown was generated by replacing the two As of the original platform by a base combination frequently observed in clones from the GAAA pool (Figures 5B and 11) and then performing 100 rounds of energy minimization with version 3.1 of X-PLOR (Brunger, 1993). Note that stacking is not markedly altered compared with the structure shown in (A).

be selected (that is, as long as the folding of P8 did not interfere with that of the *td* ribozyme core).

Therefore, our finding that a majority of clones in our final GAAA pool carry either the same 11 nt motif or minor variants of it may be regarded as the first demonstration that this motif, which is unusually frequent in natural RNAs at sites that interact with GAAA loops (Costa and Michel, 1995), is indeed an optimal receptor for those loops. Aside from that, analysis of our data leads to general rules for binding of the second loop nucleotide. Thus, receptors consisting of a continuous A-type helix are shown to be most appropriate for binding loops with a pyrimidine at the second position and to clash instead with GGRA loops. The situation with GARA loops appears to be a somewhat intermediate one, since receptor bases in register with loop nucleotide two should not extend the receptor helix by forming a Watson–Crick pair, but a U:G pair is acceptable at that location. In fact, lacking base pairing in front of the second loop nucleotide can be advantageous, and we show that it may lead—depending on the sequence

of the resulting internal loop—to improved recognition of, and increased specificity towards, particular GNRA loop sequences.

In order to characterize some of the selected receptor motifs, we incubated receptor carrying, truncated intron cores with a diversity of L2 substrates: in the presence of the guanosine cofactor of group I splicing, the intron core acts as a true catalyst and specifically cleaves the substrate at the normal 5' splice site. We reported previously (Costa and Michel, 1995) that kinetic analysis of this cleavage reaction under either single- or multiple-turnover conditions yielded similar values for k_{cat} and K_m and also noted that two L2×P8 combinations with a 37-fold difference in k_{cat}/K_m had k_{cat} values that differed by only 1.5-fold. From these and other lines of evidence, we argued that differences in K_m or k_{cat}/K_m from one L2×P8 combination to the next were most likely to reflect differences in affinity. This statement should apply as well to the receptors characterized in this work, for not only was cleavage found to occur always at the correct site, but all

but one of the sequences that were characterized share a C:G base pair at position 4 of P8 and presumably, therefore, the ability to form with the last A of GNRA loops the same base triple that was observed in both available crystal structures: presence of this interaction should guarantee a correct positioning of the substrate into the intron core. In summary, it is reasonable to assume that the vast majority of the receptors that survived up to the final pools were selected because of their unusually high affinity for the L2 loop with which they had been confronted. However, even though a majority of our *in vitro* selected sequences resemble natural ones, affinity is not all that matters in nature. Although many group I introns position their P1 substrate by means of an L2(GNRA)–P8 interaction (Michel and Westhof, 1990), GAAA loops are rare at the tip of P2 and only two of the natural P8 sequences include an 11 nt receptor motif. In view of the abundance of 11 nt receptors in the GAAA pool, this cannot be a problem of stereochemistry, but must have to do with the necessity for the P1 substrate to successively dock and undock during the catalytic cycle of group I introns (see Cech and Herschlag, 1996): in nature, the interaction between L2 and P8 is most likely a dynamic one and high-affinity P8 receptors must actually be counter-selected (see Discussions in Costa and Michel, 1995). Cases of conformational rearrangements involving GNRA loops and their receptors have also recently been uncovered in group II self-splicing introns (Chanfreau and Jacquier, 1996; Costa *et al.*, 1997).

Even static interactions need not be energetically optimal. While 11 nt GAAA receptors predominate at sites such as P6a in subgroup IC introns, they form but a fraction of the P5 sequences of subgroup IA introns (see Costa and Michel, 1995). Both L5×P6a and L9×P5 are integral parts of the stable ribozyme core of those group I introns in which these interactions exist (Jaeger *et al.*, 1994; Murphy and Cech, 1994). However, because the entire tertiary structure of subgroup IA introns forms and melts in a cooperative manner (Jaeger *et al.*, 1993, 1994; P.Brion, F.Michel and E. Westhof, in preparation), it should be possible to compensate for a weak L9×P5 interaction by reinforcing some other contacts. In contrast, stable folding of the P4–P5–P6 domain of subgroup IC introns into its sharply bent shape rests critically on only two long-range contacts (Murphy and Cech, 1994; Cate *et al.*, 1996a), one of which consists of the L5×P6a interaction.

Nevertheless, many of the known interactions involving GNRA loops in natural molecules appear to be static ones and to have been selected primarily for the quality of binding. Why then should those sequences that we have designated as the ‘novel receptors’ (Figure 10), i.e. clones C7.34, B7.8 and C7.2, be so rare in nature, at least at currently known sites of interaction with GNRA loops? We know of only one single natural variant of the C7.2 receptor, of very few molecules with a structure similar to that of the C7.34 receptor and of no natural copy of the B7.8 receptor. Yet, these new receptors are in no way inferior to previously identified sequences in terms of efficiency of recognition. Cleavage of a substrate carrying a GUAA loop by the B7.8 molecule is seven times more rapid than by a CC:GG receptor (the preferred partner of GUAA loops in natural phylogenies) and there is only a 2.7-fold difference between the combination of the C7.34

receptor and a GGAA loop on the one hand and that of the 11 nt motif and a GAAA loop on the other (Table I). Nor is specificity likely to be the issue: the B7.8 ribozyme is actually more sensitive to the sequence of the loop facing it than the *td* or B7.6 molecules. In fact, our favoured explanation for the near absence of these receptors in nature is that they may not be ‘robust’ structural solutions, in the sense that good loop binders may be lacking or rare among closely related sequences (which is clearly not the case for motifs like the 11 nt GAAA receptor; see Table I and Figure 11). Natural selection works only step by step, so that of two energetically equivalent nucleotide combinations, it is bound to favour the one that is most likely to be reached because it is part of a vast network of tolerable solutions differing from their closest neighbours by no more than one or two substitutions. This is a major difference compared with *in vitro* selection, which, in the absence of ongoing mutagenesis, will screen the space of sequences in a uniform way.

As we examined only a few individuals, many more of the clones that we isolated must carry relatively small, high-affinity receptors with novel abilities to discriminate between GNRA loops, and a far greater number of those must lie unsequenced in our final pools. Does the fact that these motifs may have been largely ignored by nature, or so it seems, makes them less interesting? To those concerned with the principles underlying RNA structure, they could be a vast source of riddles, the solution of which would inevitably contribute to our understanding of RNA in general. And even if left not understood, the new receptors described in this work could still be used for the rational design of idiosyncratic tertiary interactions within and between RNA molecules.

Materials and methods

DNA constructs

Construction of the plasmid containing the DNA template for the P1–P2 substrate (which consists of hairpin structures P1 and P2 and a GGGAAAG 5' extension) is described in Costa and Michel (1995). Derivative plasmids with different L2 loops were obtained by subcloning PCR fragments generated with oligonucleotides that carried appropriate nucleotide substitutions. All constructs were verified by sequencing the entire length of the insert.

Pool construction

The initial DNA pool was obtained by ligation of two *Bbs*I-digested DNA fragments (I and II) that had been generated by PCR on sections of our *td* intron constructs (Costa and Michel, 1995). Product I was generated with gel-purified oligonucleotides MC3591 (5'-ATTTAATA-CGACTCACTATAGAATCTATCTAAACG; T7 promoter in bold type) and P8MK2 (5'-TGCGATGAAGACAGCAGACTATATCTCCA[N]₂₁TGTCTATCGTTTCG; nucleotides in bold type correspond to the first two base pairs of P8, which were left unchanged; the *Bbs*I site is underlined; [N]₂₁ corresponds to positions synthesized using an equimolar mixture of the four phosphoramidites). The resulting PCR product carries, downstream of the T7 promoter, a G followed by the *td* core sequence (except for its P8 segment which was replaced by 21 completely randomized positions) beginning 2 nt upstream of the 5' branch of helix P3 and ending 3 nt downstream of the 3' branch of helix P7, followed by a *Bbs*I site. PCR product II was obtained with oligonucleotides MCE (5'-ACGCTTGAAGACAGTCTGCTCTGCATGGTGA; the *Bbs*I site is underlined) and 24-mer (5'-CGCCAGGGTTTCCCAGTCACGAC). Product II carries, downstream of its *Bbs*I site, the *td* intron sequence from the second nucleotide of the 3' branch of P7 to the terminal intron G, followed by 23 nucleotides of the 3' exon and the sequence of pTZ19U (US Biochemicals) from the 3' half of its *Hinc*II site to the 24-mer priming site. In order to avoid introducing mutations outside the

randomized segment, a high-fidelity recombinant *Pfu* DNA polymerase (Stratagene) was used for all PCRs. The ligation product was gel-purified and quantified: pool complexity was estimated to be 7.5 pmol. In order to generate DNA templates suitable for the synthesis of the initial RNA pool of core molecules, the entire ligation product was amplified ~11-fold with gel-purified oligonucleotides MC3222 (5'-AGATTCCTGCA-GGTAATACGACTCACTATAG; T7 promoter in bold type) and MC3332 (5'-CATTATGTTTCAGATAA, this sequence is complementary to the last 16 nt of the *td* intron). An aliquot of the resulting PCR product (III) was cloned. Sequencing of the entire length of the core of 12 clones revealed that in all of them the randomized segment had been correctly incorporated and showed no major bias in nucleotide composition (G, 25.3%; A, 20.1%; T, 30%; C, 24.5%). In order to avoid losing pool complexity, up to 22 pmol of DNA product III were transcribed in the presence of [³²P]UTP and the resulting initial RNA pool was gel-purified, eluted and quantified as described in Costa and Michel (1995).

In vitro selection and amplification

Addition reactions were performed as follows: samples of the RNA pools were pre-incubated for 5 min at 45°C in 50 mM Tris-HCl, pH 7.5 (at 25°C), 50 mM NH₄Cl, 50 mM MgCl₂ (or 20 mM MgCl₂ during the last three rounds) and 0.02% SDS. Reactions were initiated by mixing these samples with a solution of the appropriate substrate that had been pre-incubated for 5 min at 45°C in the same buffer. Reactions were carried out for 20 min (or 10 min, for addition of the 'GAAA' pool to its substrate during the last two rounds) and stopped by adding an equal volume of urea loading buffer containing 75 mM Na₂EDTA. In order to retain maximal sequence diversity during the first round of selection, the first addition reactions were performed with 42 pmol of the initial pool and 420 pmol of each one of the substrates. During the entire selection, substrate and pool concentrations were kept at 4 μM and 0.4 μM, respectively. Extents of addition reactions (molar ratios of reacted intron cores over reacted and unreacted cores) were estimated with a PhosphorImager (Molecular Dynamics).

Reacted core molecules were separated from other RNA molecules by electrophoresis in 8% polyacrylamide-urea gels. After elution and precipitation, the entire RNA product was reverse transcribed using a large excess of gel-purified oligonucleotide MC3338 (5'-GCCT-CAATTACAT; bases in bold type are complementary to the last 3 nt of the intron). Reverse transcription was performed at 50°C with SuperScript™ RNase H⁻ reverse transcriptase (Gibco BRL), to ensure maximal specificity and to avoid generating self-complementary cDNA products, and [α-³²P]dCTP was used for trace labelling. After ethanol precipitation, pellets were resuspended in a 0.2 N NaOH solution and incubated at 37°C for 40 min in order to degrade RNA molecules. After neutralization and precipitation, cDNA molecules were purified on 8% polyacrylamide-urea gels, eluted and precipitated. The entire cDNA product was PCR-amplified with oligonucleotides MC3591 and MC3332 so as to generate templates suitable for T7 transcription of a new population of core molecules. After seven rounds of selection, an aliquot of this MC3591-MC3332 product was PCR-amplified with oligonucleotides that allowed cloning into the *Pst*I and *Bam*HI sites of vector pUC19. For each of the final pools, some 35 of the resulting clones were sequenced throughout the entire length of their insert (Figure 5). Note that *Pfu* polymerase was used for all amplification reactions during the course of selection.

RNA synthesis and purification of *td*-derived ribozymes and substrates

To generate the *td*-derived ribozymes used for kinetic measurements (Figures 7 and 8), the appropriate DNA plasmids were used as templates for PCR with oligonucleotide MC3591 and gel-purified oligonucleotide MC3419 (5'-TGTTCAGATAAGGTC): the resulting products end 5 nt upstream of the 3' end of the *td* intron. Templates for synthesis of P1-P2 substrates were generated by *Rsa*I digestion. PCR products and digests were transcribed with T7 RNA polymerase in the presence of [α-³²P]UTP, and the RNAs were purified and quantified as described in Costa and Michel (1995).

Kinetic measurements

For determination of k_{cat}/K_m (Table I), ribozyme samples were pre-incubated for 5 min at 45°C in 50 mM Tris-HCl, pH 7.5 (at 25°C), 50 mM NH₄Cl, 100 mM MgCl₂ and 0.02% SDS. Reactions were initiated by mixing the ribozyme solution with a substrate and GTP solution that had been incubated for 5 min at 45°C in the same buffer. The final concentration of GTP was 1 mM. All reactions were carried

out under single-turnover and ' k_{cat}/K_m ' conditions, i.e. at ribozyme concentrations both much higher than substrate concentrations and much lower than the estimated K_m (see also Costa and Michel, 1995). Specifically, ribozyme and substrate concentrations were 0.5 μM and 0.05 μM for combinations of the *td* and B7.6 ribozymes with all substrates and of the B7.8 ribozyme with the GCGA, GAGA and GGGA substrates; 0.1 μM and 0.02 μM, respectively, for the B7.8 ribozyme and GUGA and GUAA substrates; 0.04 μM (ribozyme) and 0.01 μM (substrate) for reactions of the C7.2 ribozyme and a ribozyme carrying a canonical 11 nt motif (the complete sequence of the P8 domain of this ribozyme is shown in Figure 2 of Costa and Michel, 1995); and 0.2 μM (ribozyme) and 0.04 μM (substrate) for reactions of the C7.34 ribozyme.

Reactions were stopped by adding an equal volume of urea loading buffer containing 130 mM Na₂EDTA. Samples were loaded on 10% polyacrylamide-8 M urea gels. After electrophoresis, unfixed and undried gels were quantified for radioactivity with a PhosphorImager (Molecular Dynamics). Extents of reaction were estimated from the molar ratio of the 3' piece of the cleaved P1-P2 substrates over cleaved and uncleaved molecules.

Acknowledgements

We are especially grateful to Jamie Cate and Jennifer Doudna for allowing us to use their coordinates, to Michel Kochoyan and Denise Menay for invaluable help in oligonucleotide synthesis, and to Eric Westhof and Bruno Sargueil for helpful discussions. This work was funded by grant Bio2CT-93-0345 of the EC.

References

- Bechhofer,D.H., Hue,K.K. and Shub,D.A. (1994) A novel intron in the thymidylate synthase gene of *Bacillus* phage β22: evidence for independent evolution of a gene, its group I intron, and the intron open reading frame. *Proc. Natl Acad. Sci. USA*, **91**, 11669–11673.
- Belfort,M., Chandry,P.S. and Pedersen-Lane,J. (1987) Genetic delineation of functional components of the group I intron in the phage T4 *td* gene. *Cold Spring Harbor Symp. Quant. Biol.*, **52**, 181–192.
- Brown,J.W., Nolan,J.M., Haas,E.S., Rubio,M.A.T., Major,F. and Pace,N.R. (1996) Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl Acad. Sci. USA*, **93**, 3001–3006.
- Cate,J.H., Gooding,A.R., Podell,E., Zhou,K., Golden,B.L., Kundrot,C.E., Cech,T.R. and Doudna,J.A. (1996a) Crystal structure of a group I ribozyme domain: principles of RNA packing. *Science*, **273**, 1678–1685.
- Cate,J.H., Gooding,A.R., Podell,E., Zhou,K., Golden,B.L., Szewczak,A.A., Kundrot,C.E., Cech,T.R. and Doudna,J.A. (1996b) RNA tertiary structure mediation by adenosine platforms. *Science*, **273**, 1696–1699.
- Cech,T.R. and Herschlag,D. (1996) Group I ribozymes: substrate recognition, catalytic strategies, and comparative mechanistic analysis. In Eckstein,F. and Lilley,D.M.J. (eds), *Nucleic Acids and Molecular Biology*. Springer-Verlag, Berlin, Germany, Vol. 10, pp. 1–17.
- Chanfreau,G. and Jacquier,A. (1996) An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J.*, **15**, 3466–3476.
- Costa,M. and Michel,F. (1995) Frequent use of the same tertiary motif by self-folding RNAs. *EMBO J.*, **14**, 1276–1285.
- Costa,M., Deme,E., Jacquier,A. and Michel,F. (1997) Multiple interactions involving domain II of group II self-splicing introns. *J. Mol. Biol.*, **267**, 520–536.
- Damberger,S.H. and Gutell,R.R. (1994) A comparative database of group I intron structures. *Nucleic Acids Res.*, **22**, 3508–3510.
- Gluck,A., Endo,Y. and Wool,I.G. (1992) Ribosomal RNA identity elements for ricin A-chain recognition and catalysis. Analysis with tetraloop mutants. *J. Mol. Biol.*, **226**, 411–424.
- Gutell,R.R. (1996) Comparative sequence analysis and the structure of 16S and 23S rRNA. In Zimmermann,R.A. and Dahlberg,A.E. (eds), *Ribosomal RNA: Structure, Evolution, Processing, and Function in Protein Biosynthesis*. CRC press, Boca Raton, FL.
- Jaeger,L., Westhof,E. and Michel,F. (1993) Monitoring of the cooperative unfolding of the sunY group I intron of bacteriophage T4. The active form of the sunY ribozyme is stabilized by multiple interactions with 3' terminal intron components. *J. Mol. Biol.*, **234**, 331–346.
- Jaeger,L., Michel,F. and Westhof,E. (1994) Involvement of a GNRA tetraloop in long-range tertiary interactions. *J. Mol. Biol.*, **236**, 1271–1276.

- Latham, J.A. and Cech, T.R. (1989) Defining the inside and outside of a catalytic RNA molecule. *Science*, **245**, 276–282.
- 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.
- Murphy, F.L. and Cech, T.R. (1994) GAAA tetraloop and conserved bulge stabilize tertiary structure of a group I intron domain. *J. Mol. Biol.*, **236**, 49–63.
- Pley, H.W., Flaherty, K.M. and McKay, D.B. (1994) Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature*, **372**, 111–113.
- Robertson, D.L. and Joyce, G.F. (1990) Selection *in vitro* of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature*, **344**, 467–468.
- Tanner, M.A. and Cech, T.R. (1995) An important RNA tertiary interaction of group I and group II introns is implicated in Gram-positive RNase P RNAs. *RNA*, **1**, 349–350.
- Woese, C.R., Winker, S. and Gutell, R.R. (1990) Architecture of ribosomal RNA: Constraints on the sequence of ‘tetra-loops’. *Proc. Natl Acad. Sci. USA*, **87**, 8467–8471.

Received on November 20, 1996; revised on January 21, 1997