Structurally Complex and Highly Active RNA Ligases Derived from Random RNA Sequences

Eric H. Ekland, Jack W. Szostak, David P. Bartel

Seven families of RNA ligases, previously isolated from random RNA sequences, fall into three classes on the basis of secondary structure and regiospecificity of ligation. Two of the three classes of ribozymes have been engineered to act as true enzymes, catalyzing the multiple-turnover transformation of substrates into products. The most complex of these ribozymes has a minimal catalytic domain of 93 nucleotides. An optimized version of this ribozyme has a k_{cat} exceeding one per second, a value far greater than that of most natural RNA catalysts and approaching that of comparable protein enzymes. The fact that such a large and complex ligase emerged from a very limited sampling of sequence space implies the existence of a large number of distinct RNA structures of equivalent complexity and activity.

Only seven different classes of catalytic RNAs (ribozymes) have been found in nature (*1*). We and others have used the method of in vitro selection to isolate new classes of ribozymes from partially (*2*) or completely (*3–5*) random sequences. The process of generating new types of ribozymes from completely random sequences has provided an indication of the density of catalytic RNA sequences within sequence space. Our previous isolation of 65 independent ligases showed that at least 1 in 20 trillion 220-nucleotide (nt) sequences can carry out a particular RNA self-ligation reaction. The study of these ribozymes allows us to address other questions related to the origins of biological catalysis, such as the size, complexity, and activity of catalysts in sequence space.

We now report that seven independent ligases can be grouped into three structural classes. These classes are represented by multiple isolates of two relatively simple structural motifs and by a single isolate of a large, complex structure. This last class illustrates that a limited sampling of random sequences can yield ribozymes with remarkably complex structures, implying that the number of distinct complex functional RNA structures is very large indeed. This ribozyme, when optimized by a combination of in vitro evolution and design, approaches the catalytic activity of protein enzymes that catalyze similar reactions.

Seven families of ligases. Our previously described selection experiment started with a pool of more than 10^{15} different RNA molecules, each with a common 5′ leader sequence followed by a segment containing 220 random-sequence positions (Fig. 1A) (*3*). In vitro selection was used to enrich RNAs in which the random-sequence region promoted the joining of a substrate oligonucleotide to the 5′ leader sequence, an example of intramolecular catalysis similar to the reverse of the first step of self-splicing of the group I introns. The ligation reaction is also similar to the reaction catalyzed by protein enzymes that synthesize RNA in that it involves the attack of a terminal hydroxyl of an RNA molecule on the 5′-triphosphate of another RNA molecule, resulting in the displacement of pyrophosphate and the joining of the two RNAs by means of a new phosphodiester bond (Fig. 1A). After four rounds of selection, we obtained a collection of 65 independent sequences capable of carrying out this self-ligation reaction at low but detectable levels. With mutagenesis and continued selection, descendants of some of the initial collection of ribozymes came to dominate the population, and the average self-ligation rate of the pool increased, becoming millions of times faster than that of the uncatalyzed ligation rate (*3*).

Ligases that had undergone a total of ten rounds of in vitro selection and evolution (pool 10 ligases) were cloned, and 66 random clones were sequenced. A family of 45 related sequences generated by mutation of a single ancestral sequence dominated the final pool of ligases and was designated family A. The 21 remaining clones represented six additional families (families B through G). One of the more active members of each of the seven sequence families was chosen for further analysis.

We wished to localize the "catalytic domain" of each family; that is, the subset of the random-sequence region residues necessary for self-ligation (Fig. 1A). Deletion mapping revealed that the catalytic domains spanned a portion of the random-sequence region that varied widely in length from less than 56 nt (ligase a4) to more than 191 nt (ligase e3), with most extending over more than half of the 220 random positions of the original pool molecules (Fig. 2).

The efficiency of the deletion construct a4-10 was more than ten times greater than that of its parental full-length ribozyme (Fig. 2). The bases responsible for lowering the activity of the full-length a4 ligase lie 3′ of the catalytic domain, suggesting that the increased activity of the deletion derivative is the result of the removal of nucleotides that make unfavorable contacts with the catalytic domain or substrates.

In the self-ligation reaction one of the substrates, the 5′ leader sequence, is part of the same molecule as the ligase. We tested the catalytic domains (Fig. 2) for their ability to catalyze true intermolecular ligation, that is, ligation in which neither of the substrates was covalently linked to the catalytic domain (Fig. 1B). The catalytic domains of representatives of all seven sequence families (the "−10t" series of ligases) were capable of catalyzing this type of ligation (Fig. 3A). The four most efficient ribozymes (a4-10t, b1-10t, c2-10t, d1-10t)

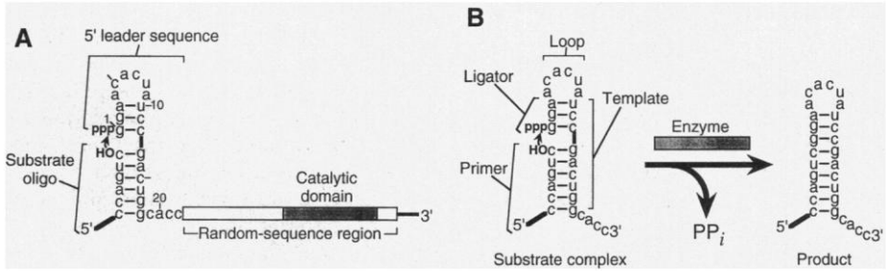


Fig. 1. Schematic of ligation reactions. (**A**) Self-ligation. The 5′-triphosphate of the leader sequence of the ribozyme is attacked by a terminal hydroxyl of a substrate oligonucleotide (*25*) to form a phosphodiester bond with the displacement of pyrophosphate. (**B**) Intermolecular ligation. Two external substrate RNAs form a complex and become ligated during incubation with a ligase ribozyme. Selected ligases utilize the designed base pairing scheme to limited and varied extents.

E. H. Ekland and D. P. Bartel are at the Whitehead Institute for Biomedical Research, Cambridge, MA, 02142, USA. J. W. Szostak is in the Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA.

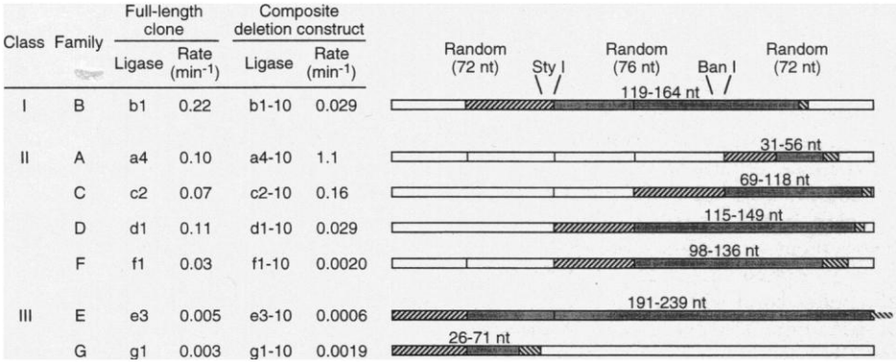


Fig. 2. Deletion analysis of representatives of the seven sequence families. Bars correspond to the random-sequence region of the original pool molecules, which consisted of three random-sequence segments (with 72, 76, and 72 random residues, respectively) linked by two restriction sites (Sty I, Ban I). Deletion analysis (*26*) localized the catalytic domains to the range indicated (////- segment containing 5′ boundary; \\\\-segment containing 3′ boundary). The 3′ boundary of the e3 catalytic domain was localized to the first 9 nt of the 20-nt 3′ primer-binding segment (narrow hatched bar). Composite deletion constructs (the −10 series) contained only bases common to the shortest active 5′- and 3′-deletion constructs and were assayed for self-ligation activity as described (*27*). All sequences have been submitted to GenBank (*10*).

catalyzed the intermolecular ligation reaction with multiple turnover (Table 1), demonstrating that in vitro selection can lead to the isolation, from completely random sequences, of RNA molecules with the classical properties of enzymes.

Examination of ligation regioselectivities, ligase secondary structures, and modes of substrate binding showed that the catalytic domains of the seven sequence families could be grouped into three classes. We describe each class of ligases individually.

The class I ligases. The product of the ligation reaction catalyzed by the family B ribozymes differed from the product generated by all of the other ligases; accordingly, the family B ligases were designated as class I ligases. In denaturing gels, the ligation product migrated as if it was nearly one nucleotide shorter than the products of the other six ligation reactions (Fig. 3A). Analysis of the ligation junctions (Fig. 3B) revealed that the difference in mobility can be attributed to formation of a different phosphodiester linkage. Clone b1 was distinctive among the seven ligases in that it catalyzed formation of a 3′,5′-phosphodiester linkage, whereas the other six ribozymes yielded 2′,5′-linkages (6).

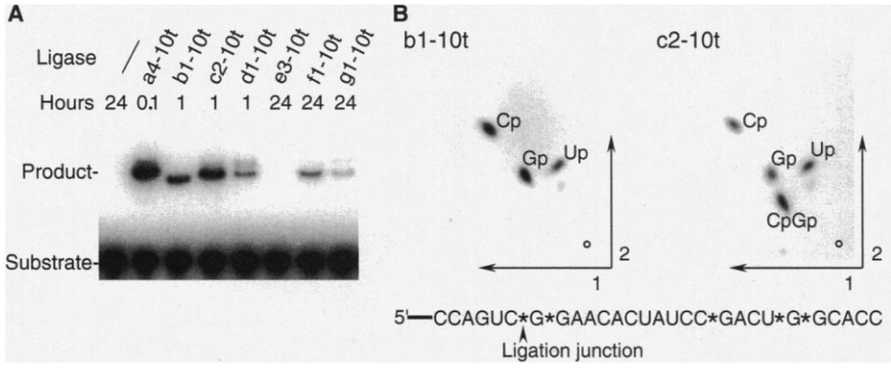


Fig. 3. Intermolecular ligation. (**A**) Substrate RNAs (5 μM) were incubated with each ligase (5 μM) as described (*23*). Labeled substrate oligonucleotide, *AAAccaguc (*, ^{32}P; upper-case, DNA; lowercase, RNA), was separated from product, *AAAccagucggaacacuauccgacuggcacc, on a 15 percent acrylamide–7 M urea gel. With the exception of construct f1-10t (*28*), each enzyme of the −10t series corresponds to the catalytic domain identified by deletion analysis (Fig. 2), preceded (if necessary) by one or two guanosine residues to facilitate in vitro transcription of the RNAs by T7 polymerase (*10*). The e3-10t ligation product cannot be detected in this autoradiogram, but phosphorimaging revealed a small amount of product, which co-migrated with the a4-10t, c2-10t, d1-10t, f1-10t, and g1-10t ligation products. (**B**) Ligation regiospecificities of b1-10t and c1-10t ribozymes. Ligation reactions were performed with a ^{32}P body-labeled 22-nt substrate. Gel-purified product with ^{32}P at the indicated (*) positions was digested with nuclease T2, which does not efficiently digest 2′,5′ RNA linkages. Digestion products were separated by thin-layer chromatography (TLC) in two dimensions (*3*). The unique labeled spot in the c2-10t TLC co-migrated with authentic unlabeled 2′,5′-linked CpGp. Similar analysis of the a4-10t, d1-10t and f1-10t ribozymes revealed that they also catalyze the formation of 2′ linkages.

Because the class I ligases were represented by only one of the seven sequence families, and because the four sequenced members of this family were closely related, comparative sequence analysis could not be used to deduce any significant features of the class I catalytic motif. We therefore sought to generate a set of sequences suitable for the detection of conserved and co-varying bases. A pool of 10^{14} variants of the b1-10 deletion construct was synthesized such that, on average, 79 percent of the 172 mutagenized positions were the same as the starting sequence, 20 percent were point substitutions, and 0.7 percent were deleted (*7*). Catalytically active variants isolated from this pool by in vitro selection were cloned and sequenced (*8*). Analysis of conserved and covarying residues led to the secondary structure model presented in Fig. 4. The class I catalytic motif, when complexed with its substrate, forms a nested pseudoknot secondary structure with seven stems and several important joining segments. Nucleotide identity within segment 20 to 24 and within two loops at positions 60 to 129 and 157 to 162 was not conserved. A 112-nt RNA (b1-206) in which segment 20 to 24 was deleted and the two loops were replaced by 5′-UUCG stable loops (9) was as active as the parental 185-nt construct. Most of the remaining bases were either participants in what appear to be critical base pairing interactions or were invariant and thus probably participate in essential tertiary interactions. It is doubtful that more than a few additional nucleotides could be deleted from the 93-nt catalytic domain of the b1-206 deletion construct.

In addition to revealing a large number of obligate Watson-Crick interactions, analysis of covarying residues also suggested a non–Watson-Crick interaction between positions 11 and 178 (8). Residue 11 was a C in the starting sequence; the substrate complex was designed such that this residue would pair with G2 (Fig. 1A). It appears that two of the designed Watson-Crick pairs of the ligator-template stem (Fig. 1, G2·C11 and A3·U10) have been displaced by interactions of residues 10 and 11 with residues near the 3′ terminus of the catalyst (Fig. 4, U10·A179 and A11·A178). Nucleotide C12 was conserved in all 25 variants and may still participate in a critical Watson-Crick interaction with the nonmutagenized G1 residue. However, this possibility requires further investigation, particularly because of the potential proximity of G1 to many other highly conserved residues.

We examined the activity of optimized class I derivatives in both the self-ligation and multiple turnover reactions (Fig. 5). The consensus construct b1-207 combined

the primary and secondary structural features conserved in the selected clones, including ten frequently occurring switches from the identity of the starting sequence (*10*). In self-ligation, 70 percent of the substrate was converted to product within 5 seconds (Fig. 5A). When the substrate complex was separated from the catalytic domain and the substrate-enzyme pairing was slightly weakened, the resulting enzyme (b1-210t) (*10*) ligated exogenous substrate with a k_{cat} of 100 min^{-1}, K_m of 9 μM (Fig. 5B). Comparison of the k_{cat} with the rate constant for the uncatalyzed ligation of preformed substrate complex (k_{uncat} = 1.2 × 10^{-7} min^{-1}, pH 8.0, 22°C) yields a rate enhancement approaching 10^9 (*11*). The b1-210t ligase can maintain high activity over long periods of time; more than 48,000 catalytic turnovers have been achieved during a 24-hour incubation.

The k_{cat} (100 min^{-1}) of the optimized class I ribozyme compares favorably to rates achieved by other RNA enzymes and by some protein enzymes. Derivatives of a group I intron and a ribonuclease P ribozyme have k_{cat} values of 6 min^{-1} and 37 min^{-1}, respectively (*12*). Like the class I RNA ligase, cellular DNA ligases join oligonucleotides within a pre-formed substrate complex. In the last step of ligation, DNA ligases promote the attack of a 3′-hydroxyl on a 5′-α-phosphate to form a 3′,5′-phosphodiester bond with concomitant release of adenosine monophosphate. *Escherichia coli* DNA ligase has a k_{cat} of 28 min^{-1} for this reaction (*13*), although the K_m of the *E. coli* DNA ligase (0.056 μM) is 160 times lower than the K_m of the class I ribozyme. Although natural ribozymes (and enzymes in general) are not usually optimized for maximal rates, it is perhaps surprising that new ribozymes with activities comparable to those found in nature can be accessed from such a limited population of random-sequence RNAs.

The class II and class III ligases. The regiospecificity of ligation set the class I (b1) catalyst apart from the other six ligases. Reaction kinetics, comparative sequence analysis, and site-directed mutagenesis experiments distinguished the class II ligases (a4, c2, d1, and f1) from the class III ligases (e3 and g1).

Multiple-turnover experiments (Table 1) suggested that at low temperatures three of the ligases (a4-10t, c2-10t, d1-10t) were rate-limited by product release; after an initial burst of product synthesis, the rate slowed considerably as the product concentration approached the enzyme concentration. Multiple turnover was, however, achieved at temperatures ≥ 37°C. The possibility of rate-limiting product release led us to search for potential base pairing between these ribozymes and their substrate RNAs. This search uncovered the potential for significant base pairing between these three enzymes and the unfolded form of what was designed to be the substrate com-

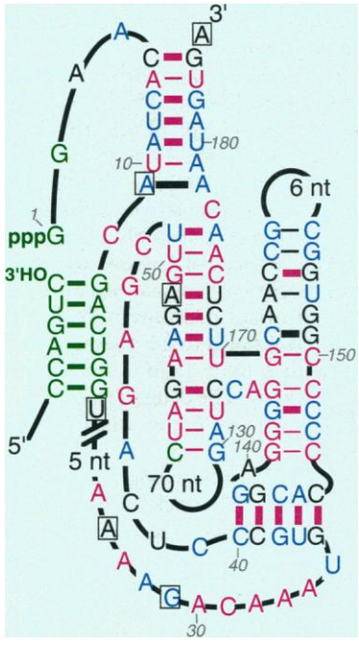


Fig. 4. Secondary structure of the class I ligase. The model is based on covariation analysis of 25 active sequence variants selected from a degenerate-sequence pool based on the b1-10 self-ligating construct (*8, 10*). Thick dash, evidence for pairing from covariation. Green residue, not mutagenized; pink residue, conserved in all 25 sequences (P = 0.0038); blue residue, identity constrained to the starting sequence and only one of the three point-substitution possibilities (P = 0.0029); pink dash, pairing conserved in ≥24 of the 25 sequences (P = 0.0022); black dash, pairing conserved in ≥22 of 25 sequences (P = 0.042). Probabilities (P), calculated with the cumulative binomial distribution, refer to the chance of a neutral base or pairing being conserved (or changed) to the extent indicated. Lines between U19 and A25 indicate where RNA was cut to convert the self-ligating configuration to the intermolecular ligation configuration.

Table 1. Multiple-turnover ligation catalyzed by four of the seven catalytic domains. Ribozymes were incubated with 5 or 30 μM substrate complex for 24 hours at 22°, 30°, 37°, 42°, and 45°C (*23*). Enzyme and substrate sequences were as in Fig. 3A. Data from the temperature which yielded the greatest number of turnovers is reported. An enzyme concentration was selected that generated a detectable amount of product in 24 hours without exhausting substrate RNAs. Enzyme concentrations of 0.05 μM (a4-10t, d1-10t), 0.5 μM (b1-10t, c2-10t, f1-10t), or 5 μM (e3-10t, g1-10t) were used.

Enzyme	Temperature °C	(Product]/[enzyme) in 24 hours		k_{obs} for 30 μM substrate (min^{-1})
		5 μM substrate	30 μM substrate	
		Class I		
b1-10t	22	0.6	1.7	1.2 × 10^{-3}
		Class II		
a4-10t	42	10	16	1.1 × 10^{-2}
c2-10t	37	2.1	7	5 × 10^{-3}
d1-10t	42	8	22	1.5 × 10^{-2}
f1-10t	37	0.27	0.6	4 × 10^{-4}
		Class III		
e3-10t	30	0.0014	0.005	3 × 10^{-6}
g1-10t	37	0.007	0.021	1.5 × 10^{-5}

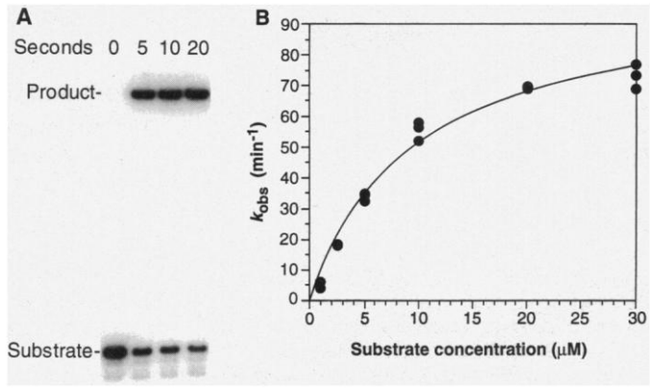


Fig. 5. Efficient self-ligation and multiple-turnover ligation of the class I ligase. (**A**) Time-course of the b1-207 ribozyme (*10*) ligating itself to the labeled substrate oligonucleotide *AAAccaguc (*27*). (**B**) Turnover rate of ligation (k_{obs}) as a function of substrate concentration. Multiple-turnover reactions were performed at 22°C as described (*23*), with the use of the enzyme b1-210t (*10*) and substrates *AAAccaguc and pppggaacgaaauacgacuggcacca (underlined bases differ in substrate shown in Fig. 1B). The line is the nonlinear least-squares fit of a Michaelis-Menten curve to the data and indicates a K_m of 9 μM and a k_{cat} of 100 min^{-1}.

plex (Fig. 6A). Isolate f1, a less efficient ligase, was also found to have the potential for similar base pairing to substrate RNAs (Fig. 6A). Site-directed mutagenesis experiments (*14*) support a model in which the "primer" substrate pairs with a 5′-GAGRCUGG segment, present within the catalytic domains of each of these ligases, rather than pairing with the substrate segment that was designed to serve as a template (R, A or G; underline indicates sequence complementarity with primer). In contrast, analogous experiments indicated that base pairing of the primer with the designed template region was required for ligation by class I and III ribozymes.

A surprising feature of the class II catalytic motif is that the continuous base pairing across the ligation junction has been replaced by an internal loop involving the ligator (G1-A3) and the first two nucleotides of the class II 5′-GAGRCUGG consensus segment (G71 and A72 of ligase a4-10 in Fig. 6A). The two 5′-GA segments in this loop may form tandem G·A mismatches. Tandem G·A mismatches with this polarity are unusually stable within Watson-Crick helices (*15*). This type of tandem G·A mismatch flanked by a 3′ purine is thought to form the binding site for a Mg^{2+} critical for the structural integrity of the hammerhead ribozyme (*16*). In the class II ligases, an analogous metal-binding site or sites a 3′ purine flanks both G·A pairs) may position a critical structural or catalytic Mg^{2+} ion(s).

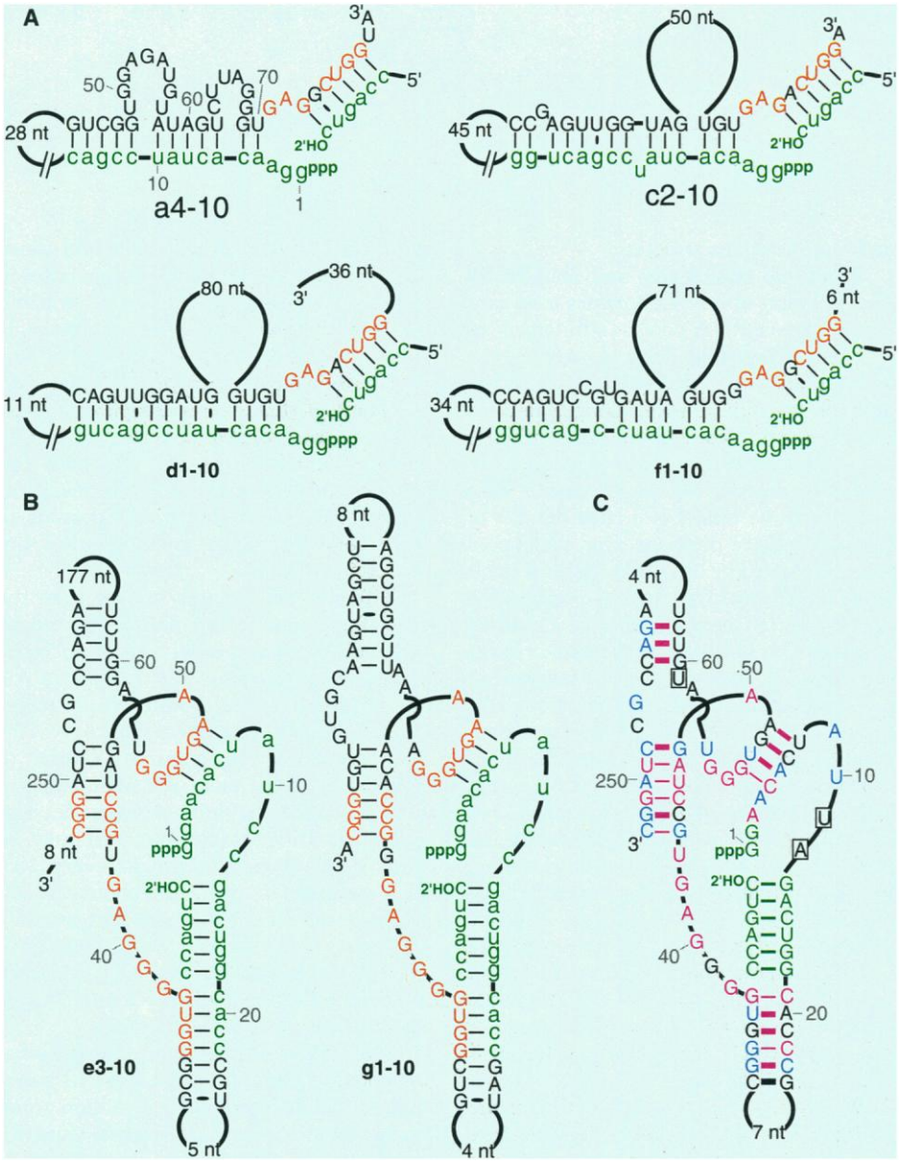


Fig. 6. Secondary structure models of class II and class III ligases. (**A**) The class II ligases. Nucleotides of catalytic domains (as mapped in Fig. 2) are shown in capital letters; nucleotides of what was designed to be the substrate complex (Fig. 1B) are shown in green lower-case letters. Significant stretches of shared primary sequence identity are indicated in orange. Lines within the loop at the left of each representation indicate where the RNA was cut to convert the self-ligating configuration to the intermolecular ligation configuration (Fig. 1). Watson-Crick pairing is indicated by lines; G·U pairing is indicated by ovals. (**B**) The class III ligases. Drawing conventions are as in (A). (**C**) A model of the class III secondary structure based on covariation analysis of 25 active sequence variants of the e3-11 construct (*18*). Drawing conventions are as in Fig. 4. Residue numbering is as in (B).

The importance of the altered ligation junction was tested directly (Table 2). When the designed substrate complex was incubated without enzyme, the terminal 3′-hydroxyl attacked ≥60 times more readily than did the 2′-hydroxyl (Table 2), despite the fact that in an unconstrained aqueous environment a nucleoside 2′-hydroxyl is about six times more reactive toward activated phosphate esters than is the 3′-hydroxyl (*17*). On the other hand, when the first three Watson-Crick partners of the ligator were replaced by 5′-GA to generate the GA·GA substructure seen in all class II ligases, the rate of attack by the 2′-hydroxyl increased dramatically (2000 times), while attack by the 3′-hydroxyl was only seven times faster (construct a4-20 in Table 2).

Although the GA·GA substructure can be thought of as the minimal class II catalytic motif, other parts of the class II ligases are required for more significant ligation activity. For instance, adding the remainder of the ligase a4 catalytic domain increased the rate of ligation by another millionfold (Table 2). Both bulged segments in the ligase a4-10 model (segments U49-U57 and C63-G68 in Fig. 6A) are likely to be required for catalysis because all but one of these 15 nucleotides were invariant in the 45-sequenced clones of family A. On the other hand, the sequence of the 28-nt loop of ligase a4-10 was not highly conserved, and it could be replaced with the 5′-UUCG stable-loop to form a 54-nt ligase with little loss of activity (construct a4-11 in Table 2). We have not yet started to define the substructures that, in addition to the minimal class II motif, are critical for the other class II ligases. Each of the class II ligases may in fact represent a distinct catalytic structure.

The other two ribozymes that catalyze the formation of 2′,5′-linkages, ligases e3 and g1, were grouped together as class III ligases. Site-directed mutagenesis experiments indicated that complementarity between the primer and the template segment was required for the class III ligases, but similar experiments indicated that ligator-template pairing was not required for either the e3 or the g1 ligase (*14*). Comparison of the sequences of the e3-10 and g1-10 deletion constructs indicated that these two ligases have the potential to fold into nearly

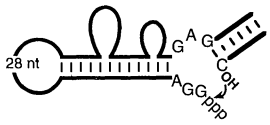
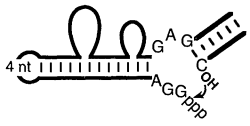
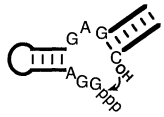
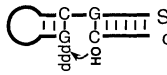
identical branched-pseudoknot secondary structures that share significant stretches of primary sequence identity (Fig. 6B). The major difference between the two ligases lies in the sequence and length of the 3′ terminal loop. Replacement of the 177-nt loop of ligase e3-10 with a 4-nt loop (5′-UUUU) and deletion of the last eight bases of e3-10, yielded a minimal class III construct (e3-11) with self-ligation activity comparable to that of parental constructs (0.003 min^{-1}).

In order to further define the class III ligase, we generated a set of active e3-11 ligase variants in a manner analogous to that described for the b1-10 class I ligase (*18*). Analysis of conserved and covarying residues confirmed the branched-pseudoknot secondary structure of the class III ligases and defined the residues most critical for class III activity (Fig. 6C).

The finding that ligases representing six out of seven sequence families from pool 10 catalyzed the synthesis of 2′ linkages was unexpected because no 2′-linked RNA was detected in our earlier analysis of product generated by the pool 6 ligases (*3*). However, these six families comprised only a small fraction (3 percent) of the pool 6 sequences (*19*). The RNA population of pool 6 was dominated by the sequence family B (50 percent), the family now known to catalyze only the formation of 3′ linkages, as well as several other families for which there are no sequenced pool 10 representatives (47 percent). Thus, it appears that the switch in ligation regioselectivity coincided with a dramatic change in population structure during the last four cycles of the previously reported selection. The reasons for this shift are unclear, but could include replication bias or differential effects of the increasingly stringent selection conditions.

None of the three ligases recognize the substrate complex in the designed configuration, in which primer and ligator are aligned by extensive Watson-Crick pairing across the ligation junction. Two factors may have contributed to the emergence of ribozymes with the observed modes of substrate binding. First, the steric environment in a continuous double helix may be so crowded that it is difficult for a ribozyme to position functional groups effectively for catalysis (5). Even protein DNA ligases prefer substrates with mismatches at or near the ligation junction (*20*), and group I introns require a U·G wobble base pair at the reactive site on the P1 stem. A second factor may be that it is difficult for a ribozyme (especially a small ribozyme) to bind a continuously paired helix. Substrate binding by base-pairing requires only the fixation of a small number of bases to form a complementary sequence, whereas binding an intact duplex may require a larger and more complex structure.

Table 2. Self-ligation rates and regioselectivities of isolate a4 derivatives compared to the rate and regioselectivity of uncatalyzed ligation (*24*).

Ribozyme	Ligation rate (min^{-1})	
	2′ linkages	3′ linkages
 a4-10	1.1	<0.015
 a4-11	1.0	<0.013
 a4-20	9 X 10^{-7}	2 X 10^{-7}
 Substrate complex	4 X 10^{-10}	2.9 X 10^{-8}

Ribozyme complexity and abundance. The isolation of novel ribozymes from random-sequence RNA pools is still largely an art. One frequently debated issue is the optimal length of the random-sequence region of the initial pool. Longer random sequences are more difficult to synthesize but increase the probability of finding complex structures by allowing segments of a ribozyme to be in any of a large number of different relative positions. Our initial pool was constructed by linking three short pools with 72, 76, and 72 random positions to create one longer pool with 220 random positions (*3*). The catalytic domain of clone g1, a class III catalyst, was entirely within the first of these random segments, whereas the a4 class II catalytic domain was confined to the third random segment. A selection done directly on the short pools that had 72 random positions could have led to the isolation of one of the four class II and one of the two class III ribozymes. However, the class I catalytic domain derived critical nucleotides from all three random-sequence segments, and even our shortest composite deletion construct has a catalytic domain of 93 nt. A pool with a long random-sequence region was necessary to obtain ribozymes of this class, which appears to be the most interesting of the three.

We started with only 1.4 × 10^{15} of the 10^{132} possible N_{220} sequences, and therefore were confronted with the question of how a structure as large as the class I ribozyme could have been selected from this very sparse sampling of sequence space. An upper estimate of the probability of finding a class I catalytic motif in a given random 220-nt sequence is ~4 × 10^{-19}, based on the number of conserved residues and base pairs, allowing three sequence segments to shift in register with each other, and allowing for the presence of two mismatches in conserved stems (*21*). The actual probability could be lower, but even this estimate gives a probability of 5 × 10^{-4} for finding such a ribozyme in a pool of ~10^{15} sequences. The fact that such a complex catalytic structure was isolated suggests that a great many different large RNA structures can catalyze the ligation reaction—so many different large motifs that it was possible to isolate one of them by sampling only 1.4 × 10^{15} N_{220} sequences.

The class III ligase is smaller than the class I ribozyme and was represented by descendants of two independent ancestral sequences, suggesting that most catalytic motifs of this or less complexity are represented in our initial pool of 1.4 × 10^{15} N_{220} sequences. Assumptions similar to those used to estimate the abundance of class I catalysts in random sequences suggest that this smaller motif should be present about eight times in every 10^{12} N_{220} sequences (*22*), or about 10,000 times in our initial pool. In that we only recovered about 65 active ribozymes altogether, this is an overestimate, suggesting that our calculated class I abundance is also likely to be an overestimate.

The above estimates of informational complexity suggest that the class III structure should be more than a million times more common in our random-sequence pool than the class I structure; the class II structures may be even simpler and correspondingly more common. However, in an admittedly small sample of seven characterized ribozymes, we found one class I, two class III, and four class II ribozymes. The simplest explanation for this discrepancy is that for every small motif, such as the class II and class III motifs, there are thousands,

if not millions, of distinct larger catalytic motifs, which raises the question of why this should be. There are many more distinct large structures than distinct small structures. Even if all distinct folded structures had an equal probability of being an active ribozyme, there would be many more large ribozymes than small ribozymes. Furthermore, a larger fraction of complex structures may be active catalysts. For example, a larger framework may allow the positioning of functional groups required for substrate binding and catalysis through a wider range of distances and angles, thus more effectively enveloping the reactive site. It may also be easier to subtly adjust the positioning of these functional groups in a large structure than in a small structure, where a single base-change has a proportionately larger effect. It is nevertheless remarkable, that these factors lead experimentally to the isolation of comparable numbers of large and small catalytic motifs. This suggests that even the most complex natural ribozymes, such as ribonuclease P and the group I and II self-splicing introns could have arisen in one step from long random sequences, and that complex ribozymes may have played an important role early in the RNA world.

REFERENCES AND NOTES

1. A. M. Pyle, *Science* **261**, 709 (1993).
2. T. Pan and O. C. Uhlenbeck, *Nature* **358**, 560 (1992); J. R. Lorsch and J. W. Szostak, *ibid.* **371**, 31 (1994); C. Wilson and J. W. Szostak, *ibid.* **374**, 777 (1995).
3. D. P. Bartel and J. W. Szostak, *Science* **261**, 1411 (1993).
4. J. R. Prudent, U. Tetsuo, P. G. Schultz, *ibid.* **264**, 1924 (1994); M. Illangasekare, G. Sanchez, T. Nickles, M. Yarus, *ibid.* **267**, 643 (1995).
5. K. B. Chapman and J. W. Szostak, *Chem. Biol.* **2**, 325 (1995).
6. The use of either guanosine triphosphate or inorganic monophosphate as a leaving group would also yield a product with faster than expected mobility. However, pyrophosphate was released during ligation catalyzed by all of the ribozymes, including family B (D. P. Bartel and J. W. Szostak, unpublished data).
7. The 6-nt segment (G13-G18 in Fig. 4) known to hybridize to the primer was not mutagenized. Eight other potentially relevant positions were not mutagenized for technical reasons: The DNA template for this RNA pool was made by linking two shorter pools; six positions (segment 59 to 64 in Fig. 4) were not mutagenized to facilitate this linkage. The first two nucleotides of the ligator (segment G1-G2 in Fig. 4) were not mutagenized because T7 RNA polymerase prefers guanosine at the first two positions of the transcript.
8. The details of the selection procedure and the analysis of selected variants are available. E. H. Ekland and D. P. Bartel, *Nucleic Acids Res.*, in press.
9. C. Tuerk *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1364 (1988).
10. GenBank accession numbers for the representative ligase isolates, a4 to g1, are U26406 to U26412. Each entry notes the span of the catalytic domain and the sequences of deletion constructs used for self-ligation (−10 series) and for intermolecular ligation (−10t series). The sequence in Fig. 4 differs from construct b1-10 at the boxed positions; b1-10 nucleotides at these positions are C11, C19, G26, U29, G52, G185. Construct b1-207 (accession number U26413) nucleotides that differ from Fig. 4 are C36, C56, G133, G145; 5′-UUCG also replaces the segment 60-129. The b1-210t enzyme corresponds to the b1-207 construct except the segment A3-U25 of b1-207 is deleted (to generate a construct for multiple-turnover ligation) and A181-U183 is replaced by UUC. When using b1-210t, A6-U8 of the substrate was changed to GAA. This replacement of three base pairs within stem C5-U10·A179-G184 is predicted to slightly weaken substrate-enzyme pairing [$\Delta\Delta G \approx 0.6$ kcal mole^{-1}; D. H. Turner, N. Sugimoto, S. M. Freier, *Annu. Rev. Biophys. Chem.* **17**, 167 (1988)].
11. The uncatalyzed ligation rate of the substrates used in Fig. 5B is the same, within error of the analysis, as that of the substrates used in Fig. 3A.
12. D. Herschlag and T. R. Cech, *Biochemistry* **29**, 10172 (1990); M. E. Harris *et al.*, *EMBO J.* **13**, 3953 (1994).
13. P. Modrich and I. R. Lehman, *J. Biol. Chem.* **21**, 7502 (1973); Michaelis-Menton parameters for the complete reaction were used because the reaction is somewhat more efficient when starting with saturating ATP and unadenylylated DNA (K_m = 0.056 μM and k_{cat} = 28 min^{-1}) than when starting with the adenylylated DNA (K_m = 0.11 μM and k_{cat} = 9.1 min^{-1}). The rate enhancement of the protein ligase is probably higher than that of the class I ligase because it promotes more difficult chemistry, that is, the attack of a 3′-hydroxyl of DNA rather than of RNA. For example, adenosine is 11 times more reactive in template-dependent condensation reactions than is deoxyadenosine [R. Lohrmann and L. E. Orgel, *J. Mol. Biol.* **113**, 193 (1977)]. The greater reactivity is attributed to the adenosine *cis*-(2′,3′)-diol. Using this factor of 11, the rate enhancement of *E. coli* DNA ligase is three times greater than that of *b1-210t.*
14. E. H. Ekland and D. P. Bartel, unpublished data.
15. J. Santa Lucia Jr., R. Kierzek, D. H. Turner, *Biochemistry* **29**, 8813 (1990).
16. H. W. Pley, K. M. Flaherty, D. B. McKay, *Nature* **372**, 68 (1994).
17. R. Lohrmann and L. E. Orgel, *Tetrahedron* **34**, 853 (1978).
18. To facilitate direct comparison of the class I and class III ligases, the active e3-11 ligase variants were generated by a procedure that closely paralleled the procedure used to generate active b1-10 ligase variants (Fig. 4) (*8*). Briefly, a pool of 10^{14} variants of the e3-11 construct was generated in which 72 positions were mutagenized at a rate of 20 percent per position, and eight positions (segments G1-G2 and G13-G18) were not intentionally mutagenized. Catalytically active variants isolated from this pool by four rounds of in vitro selection were cloned, and 31 clones were sequenced. Covariation analysis of the 25 most active sequences (self-ligation rates, 0.002 to 0.013 min^{-1}, with 10 mM Mg^{2+}) is reported (Fig. 6C).
19. We determined the relative abundances of sequence families in the selected ribozyme pools by inspection of a previous gel [figure 8 in (*3*)]. Predicted restriction sites within the consensus sequences of the sequence families under study allowed the assignment of the families to bands corresponding to labeled restriction fragments of the following lengths: family A, 76 nt; family B, 60 nt; family C, 130 nt; family D, 90 or 93 nt; family E, 133 nt; family G, 46 nt. Less than half of the members of each family have nonconsensus sites because of point mutations from the error-prone PCR. For the two most dominant families in pool 10, some of these minor bands are detectable and correspond to restriction fragments of the following lengths: family A, 50, 145, 150, 158, and 168 nt; family C, 186 nt.
20. K. Harada and L. E. Orgel, *Nucleic Acids Res.* **21**, 2287 (1993).
21. The abundance of the class I motif in 220-nt random sequences can be estimated from the comparative analysis of b1 variants summarized in Fig. 4. (i) There are 42 nt that are part of critical Watson-Crick (w-c) base pairs (Fig. 4, pink dashes involving segments G36-C40, U49-C59, G130-C133, G136-G139, G141-C150, G169-A176); if these can be any Watson-Crick base pair, then $P = (0.25)^{21}$. (ii) The stem involving segments G151-C156 and G163-C168 requires at least four base pairs; assuming these can be any Watson-Crick base pair, $P = 0.038$ (cumulative binomial). (iii) 19 of the remaining positions (nonpaired pink residues A25, A27, A30 to A34, G45 to C48, A135, C177 and the segment that pairs with the substrate loop A179-G184) are limited to one possibility, $P = (0.25)^{19}$. (iv) The seven remaining blue residues are limited to two possibilities; if one of these is a neutral base conserved by chance, $P = (0.5)^6$. The product of the probabilities (i) through (iv) is the probability of finding the consensus class I segments A25-C59, G130-C156, G163-G184 in a 92-nt random sequence, $P = 5 \times 10^{-28}$. (v) Finding the three segments of the motif in a 220-nt random sequence is nearly 360,000 times more likely than finding them in a 92-nt sequence if the segments preceding and following the catalytic domain can be any sequence or length and the two loops can be any sequence or any length ≥4 nt, making $P = 1.8 \times 10^{-22}$. (vi) Because the class I motif is so active, many sequences that do not perfectly match the consensus motif would have marginal activity sufficient for their initial enrichment; if any mismatch or wobble could be tolerated at any two of the 21 consensus base pairs (i), the probability of finding an acceptable sequence increases by a factor of about 2000, making $P = 3.6 \times 10^{-19}$. This is a crude estimate; any underestimate of the sequence flexibility is probably more than offset by other considerations such as the inability of all possible Watson-Crick pairs to substitute for all 21 of the consensus pairs (lowering factor i) or the inability of all possible sequences and sequence lengths to be tolerated at the loop and flanking segments (lowering factor v).
22. The abundance of the class III motif in 220-nt random sequences can be estimated from the comparative analysis of e3 variants (Fig. 6C) in a manner analogous to the way in which the class I abundance was estimated (*21*). (i) There are 14 nucleotides that are part of critical Watson-Crick base pairs (Fig. 6C, pink dashes involving bases C23 and G33, and segments G44-G49 and C248-C253); if these can be any Watson-Crick base pair, $P = (0.25)^7$. (ii) The stem involving segments 59-64 and 240-246 requires at least four base pairs; if these can be any Watson-Crick base pair, $P = 0.038$. (iii) Sixteen of the remaining positions (nonpaired pink residues G38, G40 to U43, A50, G55, G56, and residues that pair with the substrate, G34 to G37, and A51 to G54) are limited to one possibility, $P = (0.25)^{16}$. (iv) The two remaining blue residues are limited to two possibilities, $P = (0.5)^2$. The product of these four probabilities estimates the probability of finding the consensus class III motif (segments C23, G33-C62, G242-C253 separated by 4-nt loops of any sequence) within a 51-nt random sequence, $P = 1.4 \times 10^{-16}$. (v) Finding the three segments of the motif in a 220-nt random sequence is 58,000 times more likely than finding them in a 51-nt sequence, if the segment following the catalytic domain can be any sequence or any length, if the two loops can be any sequence or any length, and if joining segments A58-U59 and C145-C147 can be any sequence or any length from 2 to 3 nt. When taking factor (v) into account, the probability of finding an at least marginally active class III ligase is 8×10^{-12}. For the same reasons as discussed above (*21*), the true probability of finding the class III motif in 220-nt random sequences may be lower than this estimate.
23. Intermolecular ligation reactions were performed in 30 mM tris, pH 7.4, 200 mM KCl, 60 mM $MgCl_2$, 0.6 mM EDTA, except the reactions of Fig. 5B, where EPPS (*N*-[2-hydroxyethyl]piperazine-*N*′-[3-propanesulfonic acid], pH 8.0) replaced tris. Substrate molecules were incubated together in water at 80°C for 1 minute then cooled to 22°C, as was enzyme. Salt and buffer were added to substrate, reactions were started by addition of enzyme, and stopped by adding 2 volumes of 120 mM EDTA. Substrate and product were separated on gels as shown in Fig. 3A, and quantitated with a Phosphorimager. Catalytic turnover rates (k_{obs}) were calculated as [product]/([enzyme] × incubation duration).
24. Overall self-ligation rates were determined as described (*27*). For construct a4-20 and substrate complex, the relative amounts of 2′ linkages and 3′ linkages in isolated ligation product were determined as described (*3*). For substrate complex, this ratio

was determined by analysis of product from reactions done at pH 8.0 and 100 mM $MgCl_2$ (R. Rohatgi, D. P. Bartel, J. W. Szostak, in preparation). The rate of formation of 3′ linkages by ligase a4-10t was determined by quantitation of a gel similar to that shown in Fig. 3A, but the electrophoresis was over a longer distance to better separate the 2′- and 3′-linked product. The amount of 3′-linked product in the a4-10t lane was below the limits of detection by this analysis (<1.3 percent of 2′-linked product). This value was used to estimate the upper limit of rates of 3′-linked product formation in a4-10 and a4-11 self-ligation reactions.

25. The substrate oligonucleotide 5′-AAAccaguc (Fig. 3A) was usually used in this study. Self-ligation rates of the seven ligases did not systematically or substantially change when any of the three other substrates from a previous study (*3*) were used in place of this 9-nt substrate.
26. A library of 3′ deletions was generated by in vitro transcription of the RNA in the presence of the chain-terminator cordycepin (3′-deoxyadenosine) 5′-triphosphate [V. D. Axelrod and F. R. Kramer, *Biochemistry* **24**, 5716 (1985)]. After phenol extraction, desalting and precipitation, the deletion library was incubated with ^{32}P end-labeled substrate oligo. The self-ligation products were resolved on a sequencing gel. The shortest active 3′ deletion was identified by comparison to a marker lane containing a partial nuclease T1 digest of end-labeled ligated product. Each template for the 5′ deletion derivatives was individually constructed by PCR using a primer that hybridized to a site within the full-length clone. The primer also contained sequences corresponding to the T7 promoter and to the 22-nt leader sequence (Fig. 1A). The 5′ deletion derivatives indicated by the vertical lines shown within the bars in Fig. 2 were generated and tested. For example, three b1 derivatives, deleting the first 36, 78 (first 72-nt random segment plus Sty I site), or 116 nucleotides of the random-sequence region, were tested for self-ligation. Templates for the composite deletion constructs were also generated by PCR.
27. Self-ligation reactions were performed at 22°C in 30 mM tris, pH 7.4, 200 mM KCl, 60 mM $MgCl_2$, 0.6 mM EDTA. (Lowering $MgCl_2$ concentration from 60 mM to 10 mM typically led to five- to tenfold decreases in both catalyzed and uncatalyzed ligation rates.) Ribozyme (1.0 μM, final concentration) was incubated in water at 80°C for 1 minute, then cooled to 22°C (in 1 to 5 minutes) before simultaneous addition of salt, buffer, and ^{32}P-labeled substrate oligonucleotide (<100 nM, final concentration). In reactions with ligases that had the 3′ primer-binding segment of the original full-length ligase isolates (Fig. 2), a DNA oligonucleotide 5′-CGGGATCCTAATGACCAAGG (1.0 μM, final concentration), which is complementary to the 3′ primer-binding segment, was also included. (Exclusion of this DNA resulted in a 2- to 20-fold decrease in self-ligation rate, depending on the identity of the ligase.) Reactions were stopped by the addition of two volumes of 120 mM EDTA. Substrate and product were separated on gels (Fig. 5A). Gels were scanned with a Phosphorimager. Self-ligated rates were calculated as the fraction of initial labeled substrate converted to product (corrected for the depletion of substrate, if necessary) divided by the duration of the incubation. With some ribozymes, the rate of self-ligation slowed at later time points, so that the reported rates are based on the earliest time point that could be accurately measured.
28. In order to retain more of the activity of the full-length ligase, the f1-10t construct included the 46-nt segment immediately 5′ of the catalytic domain.
29. We thank E. Sun for assistance with DNA sequencing. Supported by a grant from the W. M. Keck foundation to the Whitehead Fellows program (E.H.E. and D.P.B.) and grants from NASA and Hoechst AG (J.W.S. and D.P.B.).

27 January 1995; accepted 8 June 1995