Nucleic Acids Research, 1995, Vol. 23, No. 16 3231–3238

The secondary structure and sequence optimization of an RNA ligase ribozyme

Eric H. Ekland and David P. Bartel*

Whitehead Institute for Biomedical Research, Cambridge, MA 02114, USA

Received May 5, 1995; Revised and Accepted July 5, 1995

GenBank accession nos U26407 and U26413

ABSTRACT

***In vitro* selection can generate functional sequence variants of an RNA structural motif that are useful for comparative analysis. The technique is particularly valuable in cases where natural variation is unavailable or non-existent. We report the extension of this approach to a new extreme—the identification of a 112 nt ribozyme secondary structure imbedded within a 186 nt RNA. A pool of 10^{14} variants of an RNA ligase ribozyme was generated using combinatorial chemical synthesis coupled with combinatorial enzymatic ligation such that 172 of the 186 relevant positions were partially mutagenized. Active variants of this pool were enriched using an *in vitro* selection scheme that retains the sequence variability at positions very close to the ligation junction. Ligases isolated after four rounds of selection catalyzed self-ligation up to 700 times faster than the starting sequence. Comparative analysis of the isolates indicated that when complexed with substrate RNAs the ligase forms a nested, double pseudo-knot secondary structure with seven stems and several important joining segments. Comparative analysis also suggested the identity of mutations that account for the increased activity of the selected ligase variants; designed constructs incorporating combinations of these changes were more active than any of the individual ligase isolates.**

INTRODUCTION

At least 65 independent catalytic RNAs (ribozymes) have been previously isolated by *in vitro* selection from random sequences (1). These ribozymes were isolated by virtue of their ability to promote a self-ligation reaction designed to be analogous to chain elongation by 1 nucleotide (nt) during RNA polymerization. With mutagenesis and continued selection descendants of some of the original ligases came to dominate the ribozyme population and ribozymes emerged with self-ligation rates millions of times faster than the uncatalyzed rate (1).

We recently described seven of the more successful ligase sequence families (2). Although each sequence family clearly descended from a different ancestral sequence within the original random sequence pool, the seven sequence families can be grouped into three structural classes. Among the three classes, the class I ligases are unique in their ability to generate a 3′,5′-phosphodiester bond at the junction of the ligated RNAs. The class II ligases (families A, C, D and F) are differentiated from class III ligases (families E and G) based on the substrate-binding modes, primary sequence motifs and proposed secondary structures. As with other structured RNAs (3–5), comparative sequence analysis was invaluable for predicting the defining features of the class II and class III ligases (2). However, because the class I ligases were represented by only a single family of four very similar isolates (sequence family B), this method could not be utilized to deduce the important features of the class I ligases. To compound the problem, the shortest active class I derivative obtained by simple deletion mapping and boundary experiments was still a relatively unwieldy 186 nt in length, at least twice as large as the shortest class II or class III truncation derivatives (which were 78 and 93 nt respectively; 2). *In vitro* selection protocols have been developed to isolate a diverse set of functional sequence variants suitable for detection of conserved and co-varying bases (reviewed in 6–8). We employed this approach to elucidate important features of the class I ligase.

Two important findings arise from comparative analysis of the new class I sequences and associated experiments. (i) The structure of the class I ligase is remarkably complex. The fact that a such a large ribozyme emerged from a very limited sampling of possible RNA sequences is significant, in that it implies the existence of a very large number of other distinct RNA structures of equivalent complexity and activity (2). Here we report experimental proof for the secondary structure model of this new ribozyme. (ii) Optimized versions of the class I ligase are very efficient; a version capable of multiple-turnover ligation has a k_{cat} of 100/min, far greater than that of most natural RNA catalysts and approaching that of comparable protein enzymes (2). Here we describe the means by which the dramatic improvement in class I ligation efficiency was achieved.

MATERIALS AND METHODS

Pool construction and oligonucleotide synthesis

A schematic of the pool construction is shown in Figure 1B. Degenerate 91 and 161 nt oligodeoxynucleotides were synthesized (Expedite 8900 nucleic acid synthesizer, standard 1 μmol synthesis protocol) using phosphoramidite mixtures at the mutagenized positions (91mer, AATACGACTCACTATAGGaacactatccGACTGGcaccgtagaatacaaatgtgccctcagagcttgggaagat***CCTGG***tcattaggatc; 161mer, CGGGACTCTGA***CCTTGG***taacatagctctaacctgagttcgaagtccgaatggaacaagctcaaatttctattaaacaccaggatcca-

* To whom correspondence should be addressed

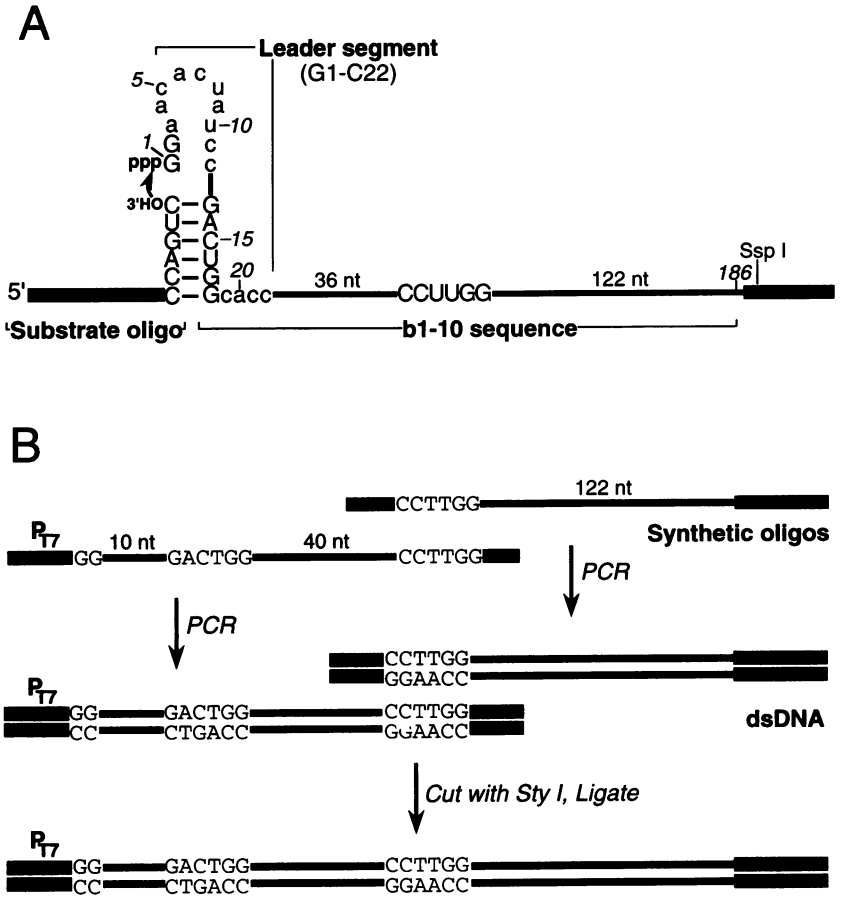


Figure 1. Design and construction of a degenerate pool based on the b1-10 ligase. (**A**) Schematic of a pool molecule hybridized to the substrate oligonucleotide. Lower-case letters and thin lines represent the 172 bases mutagenized during degenerate pool construction; upper-case letters represent the 14 potentially important positions that were not intentionally mutagenized. Primer-binding segments (thick lines) flank the substrate and ligase sequences. The b1-10 sequence is a deletion derivative of b1, a class I isolate from the original random-sequence selection (1,2) (database accession no. U26407). Segment 23–186 of construct b1-10 was derived from the domain of the original pool molecule that contained 220 random positions, whereas segment 1–22 was derived from the 5′ leader segment of the original pool molecule (1,2). The ligase promotes the attack (arrow) on the 5′-triphosphate of its leader segment by the 3′ hydroxyl of the substrate oligonucleotide to form a phosphodiester bond with displacement of pyrophosphate. The *Ssp*I site within the 3′ primer-binding site is indicated. (**B**) Schematic of construction of the DNA pool from which the degenerate RNA pool was transcribed. The T7 RNA polymerase promotor is indicated (P_{T7}). Other drawing conventions are as in (A).

ggggaggcacccccggtggctttaacgccaacgttctcaacaatagtggacaa***AATAT-T***CCTGTTGCTGTTG; mutagenized bases in lower case; *Sty*I and *Ssp*I sites in bold italics). Gel-purified DNA was used as template for large-scale PCR reactions (91mer, 10 ml reaction; 161mer, 15 ml reaction). Large-scale PCR was as typical small scale PCR except that reactions were performed in 15 ml screw-cap polypropylene tubes (Corning; ≤10 ml/tube) and temperature cycling was performed by manual transfer between three water baths (96°C, 4 min; 42°C, 5 min; 72°C, 7 min). To promote uniform heating and cooling the contents of the tubes were frequently mixed (at least once per minute by inversion). To minimize production of heteroduplex DNA caused by annealing of mismatched product strands, which occurs in late PCR cycles, cycling was stopped while the product was still increasing nearly 2-fold per PCR cycle (91mer, 8 cycles; 161mer, 12 cycles). Amplified DNA was phenol-extracted, ethanol-precipitated and digested with restriction endonuclease *Sty*I to trim off the undesired primer-binding fragments (6 ml digests using 2000 U enzyme, 2 h, 37°C). Digested DNA was heated to 60°C for 5 min, phenol-extracted, ethanol-precipitated and ligated [300 µl reaction, 120 000 U T4 DNA ligase (NEB), 1 h, 22°C]. The desired linking of the two degenerate pools to each other occurred at >50% yield, because the rate of re-ligation of the primer-binding fragments to the degenerate fragments was inhibited by the PCR primers. (Hybridization of the two strands of the excised primer-binding fragment is less stable than hybridization of the PCR primer with the longer strand because the PCR primer hybridizes to all of the nucleotides of the longer strand, including those of the 5′ overhang necessary for efficient ligation. The 60°C incubation of digested DNA was performed to promote this displacement of the shorter strand by the PCR primer.) Half of the ligation mix (1.3×10^{14} different full-length dsDNA molecules) was transcribed *in vitro* (5 ml reaction, 24 000 U T7 RNA polymerase) to yield the degenerate RNA pool.

Selection and amplification

A schematic of the first round of selection and amplification is shown in Figure 2. Gel-purified pool RNA (40 µg) was incubated with RQ1 DNase (Promega) to hydrolyze template DNA.

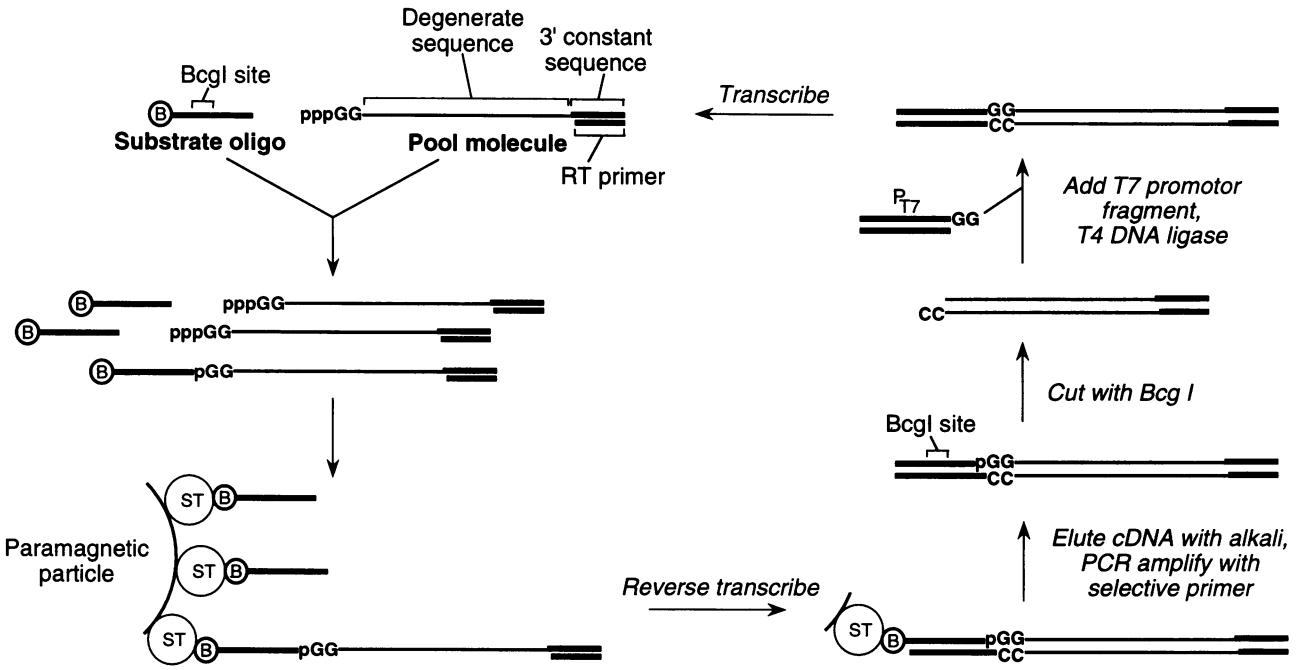


Figure 2. Selection scheme for rounds 1 and 3. Rounds 2 and 4 differed in that the ligation mix was reverse transcribed prior to biotin (B)–streptavidin (ST) affinity selection.

Following phenol extraction and precipitation the RNA was incubated in water (80°C, 3 min) in the presence of the reverse transcriptase primer (RT primer, AAGCAACAGCAACAGGAATA; the RT primer was annealed to the 3′ primer-binding segment at this early stage so that it would prevent residues within the primer-binding segment from interacting with the ribozyme or substrate bases). After brief cooling at room temperature buffer, salt and a synthetic substrate oligonucleotide AABAAGCATCTAA***GCA***TCTCAA***TCG***TAAACCAGUC were added (RNA bases underlined, *Bcg*I recognition site in bold italics, B indicates biotin–dT, introduced using a Glen Research phosphoramidite). Final concentrations of the reaction components were 0.5 μM pool RNA, 1 μM substrate oligonucleotide, 0.55 μM RT primer, 30 mM Tris, pH 7.4, 200 mM KCl, 0.1 mM EDTA, 10 mM $MgCl_2$. The reaction was stopped with 12 mM EDTA, ethanol-precipitated, resuspended in water and biotinylated molecules were captured with streptavidin paramagnetic particles (Promega), as recommended for mRNA isolation by the manufacturer. After washes with 0.1× SSC (twice) and water (twice) particles were incubated with additional RT primer (10 μM primer, 65°C, 2 min in water), cooled to room temperature and then washed with reverse transcription buffer. RNA retained on the particles was reverse transcribed using an RNase H-deficient enzyme (Superscript II; Gibco-BRL) while rotating in an incubator to keep the particles suspended [120 μl reaction, 1400 U enzyme, 30 U RNasin (Promega), 1 h incubation at 48°C, reagent and buffer concentrations as recommended by the enzyme manufacturer]. The reaction was stopped with EDTA, particles were washed with 0.1× SSC (once) and water (once) and cDNA was eluted in 100 μl alkali buffer (150 mM KOH, 20 mM Tris, 20 μM 37 nt oligodeoxynucleotide which served as carrier DNA during elution and as a primer during the subsequent PCR reaction). Eluent was neutralized with HCl (150 mM).

Eluted cDNA was amplified by PCR using the RT primer and the 37 nt primer (ATAAGCATCTAAGCATCTCAATCGTAAACCAGTCGG), which hybridizes to the substrate sequence of the cDNA (for a 'hot start' the PCR mixture was heated to 80°C prior to addition of *Taq* polymerase). After phenol extraction and ethanol precipitation amplified DNA was cut with *Bcg*I, then the digest was supplemented with 1 mM ATP, T4 DNA ligase and 28mer and 26mer oligodeoxynucleotides constituting the T7 promotor (5′-GCGGAATTCTAATACGACTCACTATAGG and pTATAGTGAGTCGTATTAGAATTCCGC, p indicates 5′-phosphate). Following incubation at 16°C the digestion–ligation mix was gel purified (4% low-melt agarose) and PCR amplified using the RT primer and the 28mer oligodeoxynucleotide with the T7 promotor sequence. This PCR product served as a template to generate RNA for the next round of selection.

Subsequent rounds of selection were performed in a similar manner, except that in rounds 2 and 4 the RNA was reverse transcribed prior to selection on the streptavidin particles. This change in procedure was instituted in alternate rounds of selection because we reasoned that selection for biotinylated RNA:cDNA heteroduplex molecules and subsequent elution with strong alkali would discourage enrichment of ssRNA sequences with fortuitous affinity for the substrate or streptavidin beads. The stringency of the ligation incubation was also gradually increased with each round of selection. Increased stringency was achieved by decreasing the incubation duration (round 1, 4 h; round 2, 30 min; round 3, 2 min; round 4, 0.5 min). In round 4 the Mg^{2+} concentration was also decreased to 3 mM. (The class I ligase requires Mg^{2+} and the self-ligation rate of round 3 RNA slows by >100-fold when

assayed in 2 mM rather than 10 mM $MgCl_2$.) After four rounds of selection the amplified cDNA was cloned (T-Vector kit; Novagen) and individual clones were sequenced.

RNA preparation

All cloned ligase RNAs were transcribed from PCR-generated DNA templates. DNA templates for b1, b1-10, b1-200 and Table 1 constructs were generated by amplification of the b1 cDNA clone, using primers that produce the desired base changes or truncations. The DNA templates for each of the selected variants (b1-101–b1-126) were generated by PCR amplification of the cDNA clones, followed by restriction digestion with *Ssp*I, which trims off >80% of the 3′ primer-binding segment from the PCR product (Fig. 1). DNA sequences corresponding to constructs b1-201–b1-207 were synthesized as two or three oligodeoxynucleotides that were assembled using T4 DNA ligase and synthetic splint oligodeoxynucleotides. Ligated DNA was PCR amplified, cloned and sequenced. This DNA was used to generate PCR DNA suitable as a template for T7 transcription.

Self-ligation assays

Self-ligation reactions were performed at 22°C in 30 mM Tris, pH 7.4, 200 mM KCl, 60 or 10 mM $MgCl_2$, 0.6 or 0.1 mM EDTA. Ribozyme (1.0 μM final concentration) was incubated in water at 80°C for 1 min, then cooled to 22°C (in 1–5 min) prior to simultaneous addition of salt, buffer and ^{32}P-labeled substrate oligonucleotide. A DNA–RNA chimera, 5′-AAACCAGUC (RNA bases underlined), served as the substrate oligonucleotide in the self-ligation assays (<100 nM final concentration). [The length and sequence of the DNA fragment of the chimera was not important; rates of ligase b1-10 self-ligation when using this 9 nt substrate or the three longer substrate oligonucleotides of the original selection (1) were similar (standard deviation, 23%). The three DNA bases (5′-AAA) were retained on the substrate to slow migration during gel electrophoresis, facilitating its resolution from the salt front.] In reactions using the b1 ligase, which had the 3′ primer-binding segment of the original full-length ligase isolates, a DNA oligonucleotide CGGGATCCTAATGACCAAGG (1.0 μM final concentration) complementary to the 3′ primer-binding segment was also included (1). (Excluding this DNA resulted in an 8-fold decrease in self-ligation rate.) Reactions were stopped by adding 2 vol 120 mM EDTA. Substrate and product were separated on denaturing 20% acrylamide gels. Gels were scanned using a phosphorimager. Self-ligation rates were calculated as the fraction of initial substrate converted to product (corrected for the depletion of substrate if necessary) divided by the duration of the incubation. The rate of self-ligation typically slowed at later time points; the reported rates were based on the earliest time point at which the product signal could be accurately measured (1). The time points chosen for rate determination ranged from 5 s (the most efficient constructs) to 2 min (b1-10 construct).

Calculations

Probabilities of a neutral base or pairing being conserved by chance were calculated using the cumulative binomial distribution. For example, given a mutation rate of 20%, the probability of fortuitous conservation of a non-functional Watson–Crick pair in ≥22 of 25 sequences (black dashes in Fig. 4A) was calculated as follows. The probability of the pairing being conserved (permitting G:U wobble pairs) in a single selected variant is the probability of retaining the starting residues, 0.8^2, plus the probability of retaining one of the residues while changing the other to generate the more probable wobble pair, 0.8×0.067, plus the probability of changing both starting residues to generate any of the other three Watson–Crick pairs or the other G:U wobble pair, $4 \times (0.067^2)$. This sum totals 0.71. The probability of this pairing being conserved in ≥22 of 25 isolates is $(0.71^{25}) + [(25!/24!) \times (0.71^{24}) (0.29)] + \{[25!/(23! \times 2!)] \times (0.71^{23}) (0.29^2)\} + \{[25!/(22! \times 3!)] \times (0.71^{22}) (0.29^3)\} = 0.0002 + 0.0020 + 0.0096 + 0.030 = 0.042$.

RESULTS AND DISCUSSION

Substrate alignment

The original class I ligase was derived from a pool of $>10^{15}$ different RNA molecules, each with a common 5′ leader sequence followed by a segment containing 220 random sequence positions (1). *In vitro* selection was used to isolate RNAs in which the random sequence region promoted ligation of a substrate oligonucleotide to its 5′ leader sequence (Fig. 1A). During self-ligation of the class I isolate the terminal 3′ hydroxyl of the substrate oligonucleotide attacks the 5′ triphosphate of the leader sequence, displacing pyrophosphate, with concomitant formation of a 3′,5′-phosphodiester bond (2).

The substrate oligonucleotide and the 5′-terminus of the ligase leader sequence were designed to be aligned by Watson–Crick pairing to U10–G18, a 9 nt segment within the leader sequence (Fig. 1A). Mutagenesis experiments indicated that Watson–Crick pairing of the substrate oligonucleotide with segment G13–G18 was important for efficient self-ligation of a class I isolate, clone b1 (Table 1). Mutations within the substrate oligonucleotide or within segment G13–G18 of clone b1 abolished activity; combining the mutant RNAs to restore predicted pairing restored activity to a detectable level, though not the starting level. A similar relationship between segments G1–A3 and U10–C12 could not be established (Table 1). Therefore, the only structural information available at the time of construction of the degenerate pool was the site of substrate oligonucleotide pairing.

Table 1. Relative self-ligation rates of mutant b1 ligases designed to test the importance of Watson–Crick pairing of the ligating segments (substrate oligonucleotide and G1–A3) to segment U10–G18

		Substrate oligo	
		5′-CCAGUC	5′-gCAuUC
Segment G13-G18	5′-GACUGG	1.0	<0.0005
	5′-GAaUGc	<0.0005	0.057

		Segment G2-A3	
		5′-GA	5′-cg
Segment U10-C11	5′-UC	1.0	0.012
	5′-cg	0.0008	0.004

Mutant residues are in lower case. See Figure 1 for residue numbering.

Pool design

Deletion mapping and boundary experiments revealed that the class I ligase spanned most of the 220 random sequence positions of the original pool (2). The minimal truncation construct generated by this analysis (construct b1-10) was 186 nt in length. This sequence served as a starting point from which to synthesize a degenerate pool in which 172 of the 186 positions were partially mutagenized (Fig. 1). Sequencing of randomly cloned DNA templates from this initial pool indicated that, on average, 79% of the 172 mutagenized positions were the same as the starting sequence, 20% were point substitutions and 0.7% were deletions. The 6 nt segment (G13–G18) known to hybridize with the substrate oligonucleotide was not mutagenized. Eight other potentially relevant positions were not mutagenized for technical reasons: the DNA template for this pool was made by linking two shorter pools (Fig. 1B) and segment C59–G64 was not mutagenized to facilitate this linkage. The first 2 nt of the pool (segment G1–G2) were also not mutagenized, primarily because T7 RNA polymerase strongly prefers guanosine at the first two positions of the transcript. A constant segment G1–G2 also facilitated attachment of T7 promotor sequences to the amplified cDNAs during each round of selection (Fig. 2).

Selection for active class I variants

Active ligases within the starting population of ~10^{14} different class I variants were enriched using the selection scheme illustrated in Figure 2. After incubation with the biotinylated substrate, oligonucleotide molecules that had ligated themselves to the substrate oligonucleotide were retained on streptavidin beads. Following reverse transcription and elution of cDNA from the column, cDNAs derived from ligated molecules were preferentially amplified using a PCR primer complementary to the substrate region of the cDNA. The amplified DNA was cut with restriction endonuclease *Bcg*I, which removed the substrate sequence, leaving a 3′ overhang that facilitated addition of oligonucleotides representing the T7 promotor. The ligated DNA was transcribed to produce RNA for another round of selection. After four rounds of selection the enriched pool of ribozymes ligated to the substrate oligonucleotide 20 times faster than the starting sequence. cDNAs were cloned and sequenced. A survey of 73 full-length clones revealed 26 sequences derived from independent members of the starting pool. Ligase ribozymes generated from these clones catalyzed self-ligation up to 700 times more efficiently than the starting sequence, b1-10 (Table 2).

Comparative sequence analysis

Sequences of the 25 ligases with at least the activity of the starting sequence were compared with the starting sequence (Fig. 3). All the conserved residues fell within the first third and within the last third of the ligase sequences, whereas the residues within segment 65–131 varied at the rate of initial mutagenesis. A high number of the mutagenized residues (33 of 172) were invariant (Fig. 3; Fig. 4A, pink residues). Because the chance occurrence of an invariant base was unlikely ($P = 0.8^{25}$ or 0.0038 per position), nearly all of the invariant residues must be important for catalysis. Interspersed among and flanking the invariant residues were 36 co-varying residues (broad-dashed pairs in Fig. 4A), which revealed the secondary structure of the ligase (Fig. 4A). The most striking co-variation occurred between residues of segment G36–C40 and those of segment G141–C145 (Fig. 3, pink regions). The pairing of these two segments to form the proposed P3 helix was supported by the fact that 24 of the 26 mutations in these segments involved a concerted change from a Watson–Crick base pair in the starting sequence to a different Watson–Crick pair in selected ligase variants (Fig. 3). Highly significant co-variation also supported pairing within stems P2 (Fig. 3, light purple regions), P4 (light blue), P5 (green) and P6 (orange). [As with other ribozyme secondary structure models (9), P indicates pairing segments, L indicates loop and J indicates a joining segment between pairing segments.]

Table 2. Self-ligation rates of the class I ligase variants

Clone	Ligation rate (/min)	
	10 mM Mg^{2+}	60 mM Mg^{2+} [a]
b1	0.072	0.22
b1-10	0.002	0.029
b1-105	1.4	10.3
b1-106	1.4	5.3
b1-102	1.1	3.7
b1-103	0.75	1.3
b1-107	0.27	
b1-114	0.24	
b1-111	0.15	
b1-101	0.10	
b1-110	0.10	
b1-108	0.09	
b1-120	0.066	
b1-125	0.045	
b1-104	0.036	
b1-119	0.033	
b1-124	0.026	
b1-123	0.025	
b1-122	0.023	
b1-117	0.022	
b1-109	0.016	
b1-126	0.014	
b1-112	0.012	
b1-113	0.011	
b1-116	0.009	
b1-121	0.004	
b1-118	0.002	
b1-115	0.0009	

[a]The four isolates most efficient in 10 mM Mg^{2+} were assayed in 60 mM Mg^{2+} for further comparison with the original class 1 isolate (b1) and its deletion derivative (b1-10).

The proposed requirement for the P7 helix (Fig. 3, gray regions) was less obvious, but still compelling. Within G151–C156:G163–C168, which is adjacent to the invariant C150:G169, most (17 of 25) of the selected clones had potential for at least five of six base pairs (Watson–Crick or G:U wobble pairs), as seen in the starting sequence. Furthermore, all except one of the selected clones (clone b1-112) had potential to make at least four base pairs. If only the identity of C150 and G169 was important and pairing of other bases within P7 was not important,

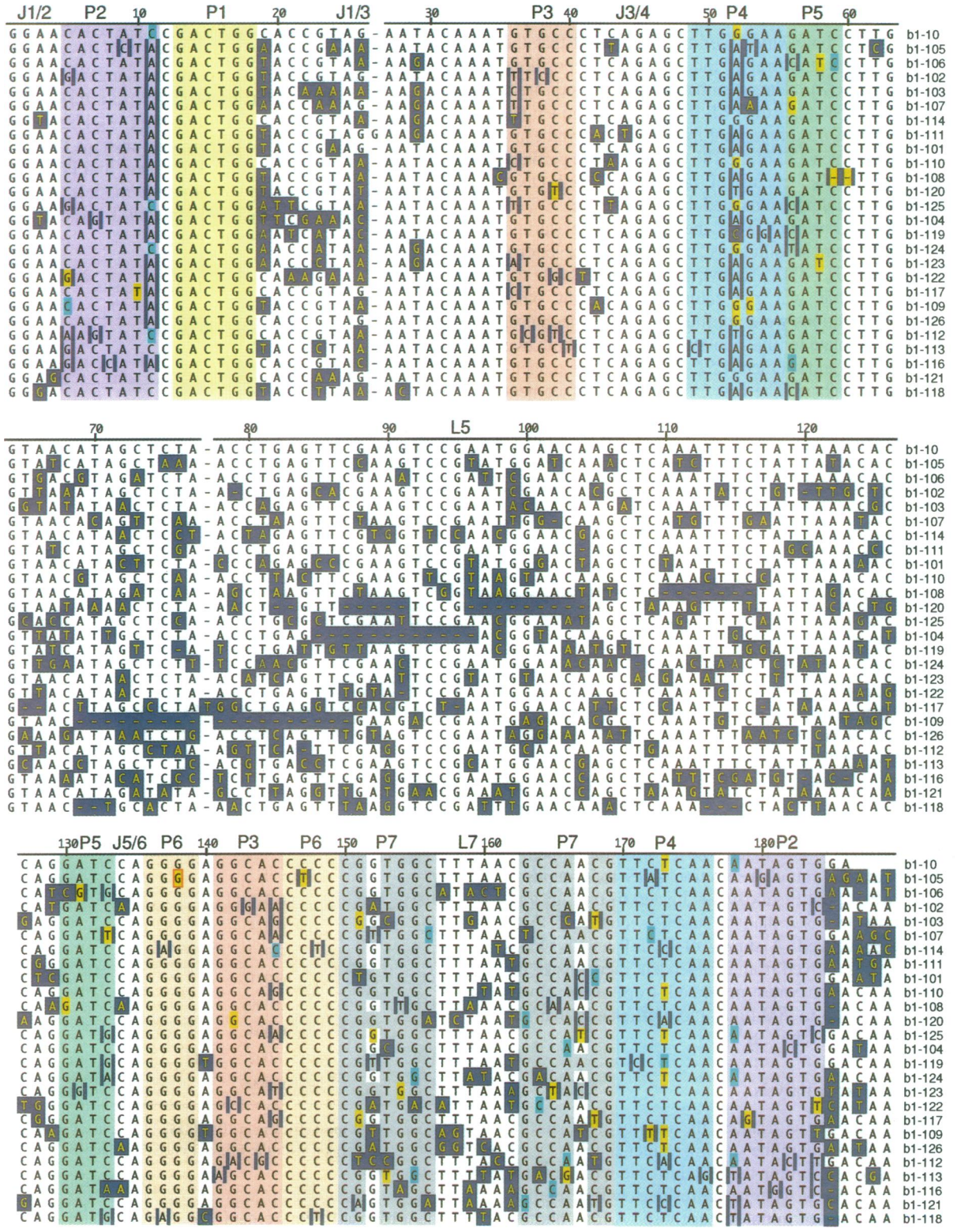
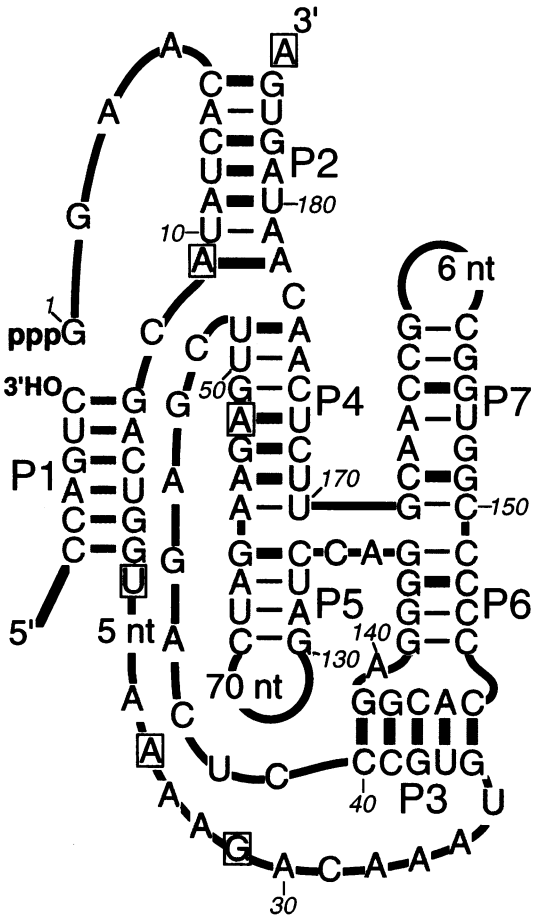


Figure 3. Comparative analysis of class I variant sequences. Segments with conserved potential for base pairing are indicated (large shaded regions) and labeled (P1–P7). Dark blue vertical bars flanking a base indicate that the base differs from the b1-10 starting sequence in a manner that maintains predicted base pairing. If this predicted pairing was maintained by a switch to a G:U wobble, then both bases of the wobble are shaded yellow. Bases that differ from the starting sequence in a manner that does not maintain conserved pairing are boxed (dark blue box). If such a change created a mismatch within proposed pairing segments then the non-mutated partner of the boxed base is also indicted (local blue shading).



Construct	Ligation rate (min^{-1})	
	10 mM Mg^{2+}	60 mM Mg^{2+}
b1-10	0.002	0.029
b1-200	0.0036	0.038
b1-201	0.45	6.2
b1-202	0.60	11.9
b1-203	0.29	2.7
b1-204	0.46	9.4
b1-206	0.46	5.0
b1-207	3.9	14.4

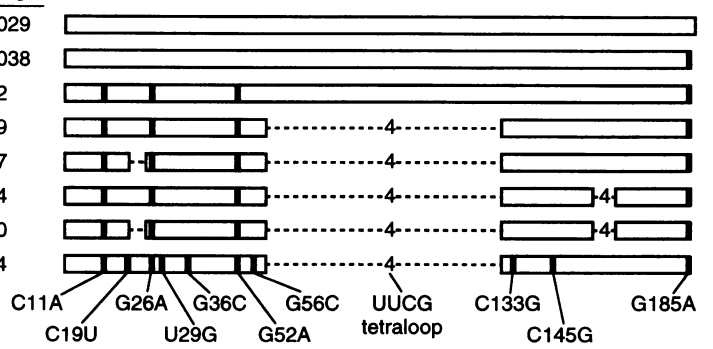


Figure 4. The class I catalytic motif. (**A**) Secondary structure of class I ligase. The model is based on the comparative analysis summarized in Figure 3. A thick dash indicates evidence for pairing from co-variation. Green residues were not mutagenized; pink residues were conserved in all 25 sequences ($P = 0.0038$); blue residues were semi-conserved (i.e. identity was constrained to the starting sequence and only one of the three point substitution possibilities, $P = 0.084$); boxed residues switched to the nucleotide shown in ≥ 7 of the 25 sequences ($P = 0.0029$). A pink dash indicates pairing was conserved in ≥ 24 of the 25 sequences ($P = 0.0022$); a black dash indicates pairing was conserved in ≥ 22 of 25 sequences ($P = 0.042$). Probabilities (P) refer to the chance of a neutral base or pairing being conserved (or changed) to the extent indicated. (**B**) Design and self-ligation rates of b1 derivatives. The b1-207 sequence has been submitted to the databases (accession no. U26413).

then ≥4 potential pairs in ≥ 24 of 25 clones would have been very unlikely to occur by chance ($P < 10^{-6}$, allowing G:U wobble base pairs). The ribozyme–substrate secondary structure model that emerged from this analysis involves seven critical stems that adopt a branched, nested, double pseudo-knot configuration (Fig. 4A).

Joining segments J1/3 (segment C19–U35) and J3/4 (segment C41–C48) had extensive stretches of conserved residues. These conserved residues most likely participate in important non-canonical contacts that define the tertiary structure of the ligase. Any residues making important contacts to the sugar–phosphate backbone or to a base within a helix could have been conserved, because

no isosteric alternative was possible or because the concerted changes that produce the isosteric alternative were too infrequent, given a mutation rate of only 20%. For example, a critical base triple that had a single isosteric alternative would have been expected to make the 3 nt concerted change in only one out of 1700 active variants (the probability of switching to a particular non-starting base is 0.067, the probability of retaining the starting base is 0.8, thus the fraction of active molecules with a particular 3 nt switch equals $0.067^3/0.8^3$). Similarly, the isosteric alternative of a non-canonical base pair would have been expected in only one in 144 active variants ($0.067^2/0.8^2$). A higher rate of global mutagenesis would facilitate detection of these types of interactions (10). Alternatively, now that a secondary structure model is in hand, it would be possible to choose specific patches of the ligase for complete randomization (11–18) or to select for second-site suppressors (19).

The only non-Watson–Crick interaction suggested by analysis of variation in the selected clones involved residues at the base of P2 (positions 11 and 178). In all but three of the selected clones C11:A178 of the starting sequence changed to A11:A178 (19 clones) or C11:T178 (three clones). The predominance of the C11A mutation suggested that the A11:A178 combination confered an increase in activity. Indeed, RNAs with the C11A change were typically 30 times more active than the otherwise identical parent molecules (results not shown).

Designed constructs

We sought to determine whether other changes seen at a high frequency in the selected clones could confer increased activity to the starting sequence. As with C11, the identity of G26, G52 and G185 changed to an adenosine very frequently (14, 15 and 14 clones respectively). The b1-201 construct, which incorporates all four changes (C11A, G26A, G52A and G185A), was 200 times more efficient than b1-10, the starting sequence (Fig. 4B). Constructs were also generated to determine whether non-conserved segments were expendable in the b1-201 background. Nucleotide identity within the 70 nt L5 loop was not conserved. Replacement of this large loop with a UUCG stable loop (20) led to a modest ($\leq$2-fold) increase in ligation rates (compare b1-202 with b1-200 in Fig. 4B). Segment 20–24 and L7 were also not conserved; deletion of segment 20–24 and replacement of L7 also had modest effects on ligation activity (Fig. 4B). An RNA (b1-206) in which all three non-conserved segments were replaced by tetraloops (L5 and L7) or deleted (segment 20–24) had nearly the same activity as its parent sequence, b1-201. Given the high level of conservation and non-random variation seen at the positions remaining in ligase b1-206, it is doubtful that more than a few additional nucleotides could be deleted from this 112 nt construct.

The most active designed construct, b1-207, differed from construct b1-202 at six positions that frequently change in the selected clones, including four positions involving critical base pairs. An essential Watson–Crick pairing that can be replaced by the other three alternatives with no consequence would have been replaced, on average, in one of 48 active sequences ($3 \times 0.067^2/0.8^2$). Yet in the 25 independent isolates pairs G36:C145 and G56:C133 changed to other Watson–Crick possibilities at a much higher rate (seven and five isolates respectively), suggesting that the original base pair was not the optimal Watson–Crick combination. Construct b1-207, which had alternative Watson–Crick pairs at these positions and a substitution at the hypervariable C19 and U29, was somewhat (six times) more active than the parent sequence, b1-202, when assayed at low Mg^{2+} concentration (Fig. 4B). A similar difference in activity was not seen at high Mg^{2+}, but at high Mg^{2+} self-ligation of both constructs was >60% complete by 5 s, the first time point in our assay. Techniques more sophisticated than manual pipetting will be required to accurately measure self-ligation rates of these efficient ligases under such permissive reaction conditions.

At 10 mM Mg^{2+} the observed rate of construct b1-207 self-ligation was 50 times faster than that of the b1 isolate (Table 2 and Fig. 4B) and >2000 times faster than the self-ligation rate of the ancestral sequence within the original random sequence pool that gave rise to the b1 isolate (D. P. Bartel and J. W. Szostak, unpublished results). It appears that ribozymes selected directly from random sequences are far from optimal and more active sequence variants are easily accessible by a combination of *in vitro* evolution and enzyme engineering. The increases in self-ligation rates could not have reflected improvement in substrate binding, because self-ligation rates were assayed using a ligase concentration (1 μM) that was far above the K_d of the substrate–ligase helix (1). It will be of interest to examine whether the rate increases can be attributed to a faster chemical step or to a faster conformational step. It will also be of interest to examine the extent to which the class I ligase is amenable to further sequence optimization, although this will require the development of selection strategies that differentiate between very efficient ligases.

ACKNOWLEDGEMENTS

We thank Pardis Sabeti for assistance with DNA sequencing and Charles Query and Bonnie Bartel for helpful comments on this manuscript. This work was supported by a grant from the W. M. Keck foundation to the Whitehead Fellows.

REFERENCES

1 Bartel,D.P. and Szostak,J.W. (1993) *Science*, **261**, 1411–1418.
2 Ekland,E.H., Szostak,J.W. and Bartel,D.P. (1995) *Science*, **269**, 364–370.
3 Levitt,M. (1969) *Nature*, **224**, 759–763.
4 Woese,C.R., Gutell,R., Gupta,R. and Noller,H.F. (1983) *Microbiol. Rev.*, **47**, 621–669.
5 Michel,F. and Westhof,E. (1990) *J. Mol. Biol.*, **216**, 585–610.
6 Burke,J.M. and Berzal-Herranz,A. (1993) *FASEB J.*, **7**, 106–112.
7 Joyce,G.F. (1994) *Curr. Opin. Struct. Biol.*, **4**, 331–336.
8 Bartel,D.P. and Szostak,J.W. (1994) In Nagai,K. and Mattaj,I.W. (eds), *RNA–Protein Interactions*. Oxford University Press, Oxford, UK, pp. 248–268.
9 Burke,J.M., Belfort,M., Cech,T.R., Davies,R.W., Schweyen,R.J., Shub,D.A., Szostak,J.W. and Tabak,H.F. (1987) *Nucleic Acids Res.*, **15**, 7217–7221.
10 Bartel,D.P., Zapp,M.L., Green,M.R. and Szostak,J.W. (1991) *Cell*, **67**, 529–536.
11 Tuerk,C. and Gold,L. (1990) *Science*, **249**, 505–10.
12 Green,R., Ellington,A.D. and Szostak,J.W. (1990) *Nature*, **347**, 406–8.
13 Tsai,D.E., Harper,D.S. and Keene,J.D. (1991) *Nucleic Acids Res.*, **19**, 4931–6.
14 Pan,T. and Uhlenbeck,O.C. (1992) *Biochemistry*, **31**, 3887–95.
15 Berzal-Herranz,A., Joseph,S. and Burke,J.M. (1992) *Genes Dev.*, **6**, 129–34.
16 Tuerk,C., Mac Dougal,S. and Gold,L. (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 6988–92.
17 Giver,L., Bartel,D., Zapp,M., Pawul,A., Green,M. and Ellington,A.D. (1993) *Nucleic Acids Res.*, **21**, 5509–5516.
18 Green,R. and Szostak,J.W. (1994) *J. Mol. Biol.*, **235**, 140–155.
19 Berzal-Herranz,A., Joseph,S., Chowrira,B.M., Butcher,S.E. and Burke,J.M. (1993) *EMBO J.*, **12**, 2567–2573.
20 Tuerk,C., Gauss,P., Thermes,C., Groebe,D.R., Gayle,M., Guild,N., Stormo,G., d'Aubenton-Carafa,Y., Uhlenbeck,O.C., Tinoco,I.,Jr, Brody,E.N. and Gold,L. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 1364–1368.