Tracing Sequence Diversity Change of RNA-Cleaving Deoxyribozymes under Increasing Selection Pressure during in Vitro Selection†

Kenny Schlosser and Yingfu Li*

Departments of Biochemistry and Chemistry, McMaster University, 1200 Main Street West, Hamilton, Canada L8N 3Z5

Received February 3, 2004; Revised Manuscript Received April 30, 2004

ABSTRACT: In vitro selection has been used extensively over the past 10 years to create functionally diverse DNA enzymes. The majority of in vitro selection experiments to date have focused on the outcome rather than the process itself, a process that remains to be fully elucidated. In vitro selection techniques rely on the probability that some DNA molecules in a random-sequence library will fold into an appropriate tertiary structure and catalyze a desired reaction. Thus, sufficient sequence diversity in the DNA pool (and hence more catalytic DNA sequences) is a prerequisite for the successful isolation of efficient deoxyribozymes. The catalytic sequence diversity established by in vitro selection is governed largely by the choice of selection pressures, one of which is the length of the reaction time. The objective of this study was to evaluate the sequence diversity change of a pool of RNA-cleaving deoxyribozymes as a function of the reaction time. Seventeen rounds of in vitro selection were performed, and the reaction time was progressively decreased from 5 h to 5 s. A representative population from each time class was subsequently cloned and sequenced. A decline in sequence diversity was observed with decreasing reaction time, and the relationship appears to be logarithmic. In contrast, a control selection performed with a constant reaction time during each round led to a linear and comparatively very slow decrease in sequence diversity. This study provides the first methodical examination of the change in catalytic sequence diversity that occurs through the course of a deoxyribozyme selection experiment. Moreover, it represents a first step toward fully understanding the intricate pathway that lies between the beginning and end of an in vitro selection experiment.

DNA is most commonly recognized for its preeminent role in the storage of genetic information. The inherent stability and duplex structure of this polymer make it perfectly suited to serve in this capacity. There is, however, an alternative function of DNA that is less commonly acknowledged. Masked only by its complementary strand, the latent catalytic prowess of DNA is slowly emerging as a rival to the more traditional protein and RNA biocatalysts (*1*−*6*).

Thus far, enzyme engineers have created a wealth of deoxyribozymes that catalyze a variety of chemical transformations, boasting rate enhancements between 4 and 10 orders of magnitude over the corresponding uncatalyzed reactions (*4*). Indeed, many of the reactions of contemporary biochemistry have become accessible to deoxyribozymes, including RNA cleavage (*7*−*19*), RNA ligation (*20*−*23*), DNA cleavage (*24*, *25*), DNA ligation (*26*−*29*), DNA phosphorylation (*30, 31*), DNA adenylylation (*32*), porphyrin metalation (*33*−*36*), and N-glycosylation (*37*).

Obviously much time and effort have been invested in the pursuit of new DNA enzymes, yet comparatively little time has been devoted to understanding the route by which they were created. The current process of in vitro selection is rather rudimentary, as it is based largely on intuition and a paucity of empirical data. To realize the full catalytic potential of deoxyribozymes, the process by which they are created must first be thoroughly elucidated. Only then can we begin to appreciate the intrinsic limitations of DNA to serve as an alternative molecular format for catalysis and be certain that any limitation is due to the small chemical repertoire of DNA and not simply the in vitro selection process.

In vitro selection techniques rely on the sheer probability that one or more DNA molecules in a vast random-sequence pool will possess some catalytic propensity for a given chemical transformation. Therefore, the extent of sequence diversity in the DNA pool may have a profound effect on the final outcome of the experiment. Without sufficient sequence diversity, the probability of finding a proficient DNA enzyme can be unfavorably low. The catalytic sequence diversity established by in vitro selection can be manipulated by a judicious choice of selection pressures, including the type of metal ion cofactors used and the length of the reaction time. In a previous research endeavor, our group employed a parallel in vitro selection format to assess the effect of four common divalent metal ions, Ca^{2+}, Cu^{2+}, Mg^{2+}, and Mn^{2+}, on the sequence diversity of self-phosphorylating DNA enzymes (*31*). Our results suggested that Cu^{2+}- and Mn^{2+}-utilizing sequences occur far more frequently in sequence space than their Mg^{2+} and Ca^{2+} counterparts.

In the current study, we wanted to assess the influence of yet another selection pressure, reaction time, on the outcome

† This work was supported by research grants from the Canadian Institutes of Health Research and Canada Foundation for Innovation. Y.L. is a Canada Research Chair.

* To whom correspondence should be addressed. E-mail: liying@mcmaster.ca. Tel: (905) 5259140 ext 22462. Fax: (905) 5229033.

10.1021/bi049757j CCC: $27.50
Published on Web 07/09/2004

of an in vitro selection experiment. More specifically, we investigated the change of sequence diversity of RNA-cleaving deoxyribozymes as a function of the cleavage reaction time. To this end, multiple rounds of in vitro selection were performed on a pool of random-sequence DNAs coupled to a long RNA substrate, and the reaction time was progressively decreased. Representative populations from each time class were then cloned and sequenced to reveal what relationship, if any, could be ascertained from the data.

EXPERIMENTAL PROCEDURES

Materials and Common Procedures. Standard oligonucleotides were prepared by automated DNA synthesis using cyanoethylphosphoramidite chemistry (Keck Biotechnology Resource Laboratory, Yale University; Central Facility, McMaster University). Random-sequence DNA libraries were synthesized using an equimolar mixture of the four standard phosphoramidites (produced by hand-mixing method). DNA oligonucleotides were purified by 10% preparative denaturing (8 M urea) polyacrylamide gel electrophoresis (PAGE), and their concentrations were determined by spectroscopic methods.

Nucleoside 5′-triphosphates, [γ-^{32}P]ATP and [α-^{32}P]dGTP, were purchased from Amersham Pharmacia. *Taq* DNA polymerase, T4 DNA ligase, and T4 polynucleotide kinase (PNK) were purchased from MBI Fermentas. All chemical reagents were purchased from Sigma.

The 50-nt R1 was produced by RNA transcription using T7 RNA polymerase and a double-stranded DNA template having the following sense sequence: 5′-GAATT C*TAAT ACGAC TCACT ATA*GG AGAGA GATGG GTGCG TTACG TAAAC TTACA TCTAC GAATC AGGTT CGA-3′ (italicized letters indicate the promoter sequence for T7 RNA polymerase). The double-stranded DNA template was generated using the above template and the following two primers: 5′-GAATT CTAAT ACGAC TCACT ATA-3′ and 5′-TCGAA CCTGA TTCGT AGA-3′. The transcription reaction was conducted under the following conditions: 0.25 μM DNA template, 10 mM each of GTP and ATP, 5.3 mM CTP, 8 mM UTP, 80 mM GMP, 1× transcription buffer (supplied by the manufacturer as the 10× stock), and 0.1 unit/μL T7 RNA polymerase. The concentration of each NTP is proportional to the relative composition within R1 (A = 30%: G = 30%: C = 16%: U = 24%). Excess GMP is included to produce a monophosphate at the 5′-end of R1, which is required for subsequent ligation of R1 to A1. The above solution was incubated at 37 °C for 30 min. The transcribed R1 was gel purified.

R1 was then ligated to A1 as follows: 2500 pmol of 50-nt R1, 7500 pmol of template T1, and 15000 pmol of DNA oligonucleotide A1 were combined and heated at 90 °C for 30 s. After cooling, 25 μL of 10× ligase buffer (supplied by the manufacturer) and 25 μL of T4 DNA ligase (5 Weiss units/μL) were added to initiate the ligation reaction. The reaction mixture (300 μL in total) was incubated for 4 h at room temperature. Ligated A1-R1 was recovered by ethanol precipitation and purified by 10% denaturing PAGE.

In Vitro Selection Procedure. Each round of in vitro selection consists of steps I−VIII illustrated in Figure 2 (sequences of DNA or RNA molecules are given in Figure 1).

Step I: The 104-nt library L1 was ligated to the 64-nt A1-R1 substrate to yield 168-nt A1-R1-L1 constructs. The reaction mixture containing 250 pmol of A1-R1, 250 of pmol of L1, and 300 pmol of template T2 was heated at 90 °C for 30 s and then cooled to room temperature. The 10× ligase buffer (12.5 μL) (supplied by the manufacturer) and 25 μL of T4 DNA ligase (5 Weiss units/μL) were added to the mixture (250 μL final volume). The resultant reaction mixture was incubated for 4 h at room temperature. It is noteworthy that the 10× ligase buffer was actually used as a 20× stock to minimize the activation of DNA catalysts by Mg^{2+} in the buffer, since we found that T4 DNA ligase was as effective in the reduced metal ion concentration.

Step II: A1-R1-L1 chimeras were recovered by ethanol precipitation and purified by 10% denaturing PAGE.

Step III: A 2× selection buffer was added to an equal volume of H_2O in which the DNA pellet from step 2 was dissolved. The resultant reaction mixture (with DNA concentration at ~0.1 μM) was allowed to stand at room temperature for the following designated times: 5 h for G0−G7, 30 min for G8, 5 min for G9−G11, 30 s for G12−G14, and 5 s for G15 and G16. In the control selection experiment, the reaction mixture was incubated for 5 h at room temperature for G0−G15. After the desired amount of time has elapsed, a 2× volume of 45 mM EDTA (pH 8.0) was added to the reaction mixture to stop the reaction. Selection buffer A (composed of 50 mM HEPES, pH 7.0 at 23 °C, 400 mM NaCl, 100 mM KCl, 7.5 mM $MnCl_2$, 50 μM $CuCl_2$, and 7.5 mM $MgCl_2$) was used for the first eight rounds, and selection buffer B (which was identical to selection buffer A with the exception that $CuCl_2$ was eliminated) was used for the remaining selection rounds.

Step IV: The above reaction mixture was subjected to 10% denaturing PAGE. The cleavage products of interest were excised from the gel and recovered by ethanol precipitation. The cleavage fragments with lengths between 104 and 154 nt were recovered during G0−G7 selection rounds. For the remaining selection rounds, the cleavage product of ~143 nt was excised instead.

Steps V and VI: Two successive rounds of PCR were used to amplify the recovered cleavage fragments. P1 and P2 were used as the primer set for the first PCR reaction and P2 and P3 for the second reaction (Figure 1). One-hundredth of the double-stranded DNA product from the first reaction was used as the template in the second reaction. Both reactions were performed on a SmartCycler (Cepheid) and monitored in real time using SYBR Green (Molecular Probes) as the reporter. The second PCR mixture contained 20 μCi of [α-^{32}P]dGTP to introduce radiolabels into the amplified DNA products for visualization purposes. The second PCR also used a ribo-terminated primer (P3) so that the 104-nt putative catalytic DNA molecules could be regenerated by alkaline digestion. PCR1 and PCR2 typically consisted of ~30 and ~12 cycles, respectively (30 s at 94 °C, 45 s at 50 °C, and 40 s at 72 °C).

In generation 8, a 15-nt extension was engineered onto the existing 3′ primer-binding site, and a new P2 primer (5′-CCA TCA GGA TCA GCT) was employed accordingly during subsequent rounds of PCR. The 15-nt extension was created during PCR2 using P3 and an elongated 30-nt primer composed of the old and new P2 primer sequences (5′-CCA TCA GGA TCA GCT ACT GCT GAT TCG ATG). The

same procedure was used for the control selection experiment using a different P2 primer (5′-TCA TCA GCT CCA GGA) for subsequent rounds of PCR and a different 30-nt primer for the first PCR2 reaction (5′-TCA TCA GCT CCA GGA ACT GCT GAT TCG ATG).

Step VII: The amplified DNA from the second reaction above was recovered by ethanol precipitation. Ninety microliters of 0.25 M NaOH was added to the DNA pellet, and the resultant solution was heated at 90 °C for 10 min, followed by addition of 10 μL of 3 M NaOAc (pH 5.5 at 23 °C). This alkaline treatment serves to cleave the embedded RNA linkage.

Step VIII: The 104-nt cleavage fragments, with length corresponding to that of the original library, were purified by 10% denaturing PAGE. The recovered DNA molecules were incubated with 10 units of PNK at 37 °C for 1 h for DNA phosphorylation in a 100 μL reaction mixture containing 50 mM Tris-HCl, pH 7.8 at 23 °C, 40 mM NaCl, 10 mM $MgCl_2$, 1 mg/mL BSA, and 0.5 mM ATP. The 5′-phosphorylated DNA (denoted G1) was used for the second round of selection. Steps I−VIII were repeated 17 times using the same procedure described for the first round of selection except that the ligation reaction in step I was conducted at $^1/_{10}$th scale (i.e., 25 μL final volume; however, all reaction components were maintained at the same concentrations as in round 1).

Cloning and Sequencing of Selected DNA Populations. DNA sequences from a relevant selection round were amplified by PCR and cloned into a vector by the TA cloning method. Plasmids containing individual catalysts were prepared using the Qiagen mini-prep kit. DNA sequencing was performed on a CEQ 2000XL capillary DNA sequencer (Beckman-Coulter) following the manufacturer's recommended procedures.

Cleavage Site Determination. Individual A1-R1-L1 constructs made from DNA-I, DNA-II, and DNA-III populations obtained in G7 (i.e., L1 = DNA-I, DNA-II, and DNA-III, respectively) were labeled with [γ-^{32}P]ATP using PNK and purified by PAGE. Each purified construct was allowed to self-cleave in the presence of the selection buffer for 10 h before each reaction was quenched by 30 mM EDTA. The stopped reaction solution was combined with a urea-containing denaturing gel loading buffer for PAGE analysis. A partial RNA ladder was generated from A1-R1-DNA-II by RNase T1 as follows (*38*): ^{32}P-Labeled A1-R1-DNA-II was dissolved in 7 M urea, 25 mM sodium citrate (pH 5.0 at 23 °C), and 1 mM EDTA (pH 8.0 at 23 °C). RNase T1 was added to the above solution to a final concentration of 10 units/μL. The mixture was incubated at 55 °C for 15 min and then combined with the gel loading buffer for PAGE analysis. A full RNA ladder that accounts for all ribonucleotides was also generated from A1-R1-DNA-II using the following procedure (*38*): ^{32}P-Labeled A1-R1-DNA-II was dissolved in 0.5 M Na_2CO_3 (pH 9.0 at 23 °C), and the mixture was then incubated at 90 °C for 5 min before being combined with the gel loading buffer for PAGE analysis.

Kinetic Analyses. A typical reaction involved the following steps: (1) heat denaturation of the A1-R1-DNA pool construct in water for 30 s at 90 °C, (2) incubation for RNA cleavage at room temperature in the selection buffer B for a designated time, (3) addition of EDTA to 30 mM to stop the reaction, (4) separation of cleavage products by denaturing 10% PAGE, and (5) quantitation using a PhosphorImager and ImageQuant software. For deriving the catalytic rate constants, aliquots of an RNA-cleavage reaction solution were collected at different reaction time points that were under ~30% completion, and the rate constant for the reaction was determined by plotting the natural logarithm of the fraction of DNA that remained unreacted vs the reaction time. The negative slope of the line produced by a least squares fit to the data was taken as the rate constant.

Quantification of Observed Sequence Identity. Sequences were aligned on the basis of the location of their fixed primer-binding sites and the aid of the Bioedit Sequence Alignment Editor computer program. Once aligned, individual sequences were grouped into common sequence classes based on an arbitrarily set, 10% allowable base mutation threshold. In other words, sequences that differed from one another by no more than eight mutations (10% of the 80-nt random-sequence domain = 8 bases) were grouped into a single sequence class. Multiple sequence classes were formed in each generation. A sequence identity matrix was generated for each sequence class in each generation, using Bioedit. The sequence identity matrix is simply an $N \times N$ matrix (where N = number of sequences in a given class) that shows the sequence identity (expressed as a number between 1 and 0, where 1 represents identical sequences and 0 represents no identical bases in their current alignment) for each pairwise permutation of all sequences being considered (in this case for all sequences within a class). The average sequence identity for each of the N sequences in a given sequence class, relative to all other $N - 1$ sequences in the same class, was calculated, and the minimum of these values was determined. This process was performed for each of the sequence classes in a given generation. The minimum average sequence identity observed for each of the sequence classes in a generation was then averaged to yield the "observed sequence identity within classes" values shown for each generation in Figure 10C. A sequence identity matrix was also generated for all of the sequences in a given generation. From this matrix, the maximum and minimum observed sequence identity between any two sequences from different sequence classes was identified and recorded as the "sequence identity between two closest classes" and "sequence identity between two farthest classes", respectively, in Figure 10C.

RESULTS

To investigate the relationship between the catalytic sequence diversity and reaction time, we devised a simple experimental protocol that consisted of three basic steps: several rounds of in vitro selection to establish a catalytic DNA population from a random-sequence pool, followed by multiple rounds of in vitro selection with incrementally decreasing reaction times to establish different time classes of deoxyribozymes, followed by the cloning and sequence analysis of a representative population from each time class. Although there are a number of ways in which our objective could have been approached, we adopted the following design parameters to facilitate this endeavor.

RNA Cleavage as Model Reaction. RNA cleavage was chosen as the model reaction for two reasons. First, RNA cleavage is a relatively facile reaction, which is a desirable

quality for a model system. A model reaction with a higher degree of difficulty might obscure fundamental relationships, simply because the random-sequence pool may have few catalytic sequences to begin with. Second, RNA-cleaving DNA enzymes are highly coveted for many innovative applications. A DNAzyme known as 10-23 (*10*, *39*) has been studied for use in the targeted destruction of disease-causing mRNAs in vivo (*40*−*42*). RNA-cleaving DNA enzymes are also highly amenable to biosensor technology (*40*, *43*−*49*). Many recent studies have explored the potential of DNA enzymes to serve as fluorescence-signaling probes that report molecular recognition for targets including nucleic acids (*44*−*47*, *50*), small molecules such as ATP (*18*), and metal ions such as lead (*43*) and manganese (*17*).

Library and Substrate Designs and Choice of Divalent Metal Cofactors. When evaluating the relationship between any two variables, there is always a risk that one or more confounding variables may exist. In this study we did not want to constrain the sequence diversity with anything but the reaction time. Therefore, we tried to account for any potential variables and limit their influence accordingly.

First, it was essential to use a DNA pool with the capacity to yield highly diverse catalytic DNA sequences. Second, it was desirable to adopt a substrate that could facilitate the selection of catalytic DNAs. On the basis of these two considerations, we decided to use 50 fixed-sequence ribonucleotides and 80 random-sequence deoxyribonucleotides as the substrate and putative catalytic domain, respectively. By comparison, the longest RNA motif used to date had been 12 nt in conjunction with a 50-nt random-sequence domain (*10*). The longest random-sequence domain (for RNA cleavage) to date had been 74 nt in conjunction with a single ribonucleotide cleavage domain (*9*). Whereas short substrate and random-sequence domains may restrict the sequence diversity, the relatively long substrate and random-sequence domains employed in this study should have the opposite effect. The 50-nt RNA substrate provides all 16 dinucleotide junctions (2−3 times each) as possible cleavage sites, as well as numerous sequence options that could serve as potential binding sites for prospective DNAzymes. The 80-nt random-sequence domain provides a large combinatorial advantage for accessing diverse structures and more catalytically active molecules. Of course, we do acknowledge that the benefits of such a large random-sequence domain may be limited by the small fraction of total available sequence space ($\sim 10^{14}$ of 10^{48} possible sequences) that can actually be sampled in any one experiment. Nevertheless, there is a sound theoretical motivation for employing long random-sequence domains (*51*).

The detailed design of our chimeric RNA/DNA library is illustrated in Figure 1. Each molecule in the library contains three key domains: a 14-nt DNA fragment (denoted A1) precedes the 50-nt RNA substrate (denoted R1), which in turn precedes another 104-nt DNA fragment (denoted L1). L1 contains 80 random-sequence nucleotides that serve as the putative catalytic domain, flanked by 9-nt and 15-nt fixed-sequence domains at the 5′ end and 3′ end, respectively. These two short DNA sequences serve as primer-binding sites for PCR amplification. In a preliminary assessment, we found that DNA primers as short as nine nucleotides could be used by *Taq* polymerase for efficient DNA synthesis as long as the annealing temperature was kept below 50 °C

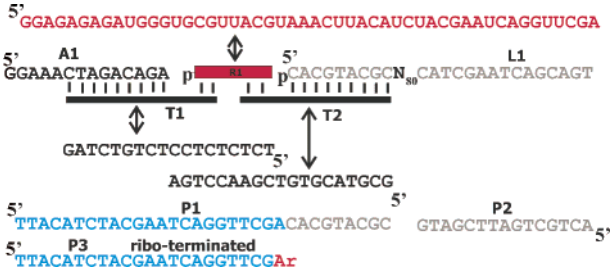


FIGURE 1: DNA sequences and their relationships. Each molecule in the library ($\sim 10^{14}$ DNA molecules) contains three key domains: a 14-nt DNA fragment (A1) precedes a 50-nt RNA substrate (R1), which in turn precedes another 104-nt DNA fragment (L1). L1 contains 80 random-sequence nucleotides (N_{80}) that serve as the putative catalytic domain, flanked by 9-nt and 15-nt fixed-sequence domains that serve as primer-binding sites for PCR amplification. A1, R1, and L1 are ligated by T4 DNA ligase in the presence of templates T1 and T2. Sequences denoted by P1, P2, and P3 serve as primers during PCR.

(data not shown). Therefore, we used a relatively short primer-binding site in favor of an extended random-sequence domain, which kept the overall length manageable (it is generally known that the yield of chemical DNA synthesis is significantly reduced for long DNA molecules). In addition, most cleavage products will contain some ribonucleotides at the 5′ end, which can also serve as part of the primer-binding site since primer P1 shares some common nucleotides with the 3′ end of the RNA substrate sequence. This design feature should increase the PCR efficiency. Therefore, PCR bias is not likely to have played a significant role in the amplification of the predominantly observed sequences. The G7 DNA pool exhibited extensive sequence diversity (see below), which provides some experimental evidence to suggest that PCR bias was not a factor. The RNA motif, R1, is an artificial sequence that was carefully designed to eliminate the occurrence of strong secondary structures that might hinder access to certain dinucleotide sites. Inclusion of the short DNA motif, A1, was intended for easy separation of the cleavage fragments in our gel-based selection strategy (see below). In addition, A1 could serve as a binding arm for any potential deoxyribozymes that chose the first few RNA dinucleotide junctions as cleavage sites.

In keeping with our desire not to impose limitations on the sequence diversity by anything except the reaction time, we selected Mg^{2+}, Cu^{2+}, and Mn^{2+} as metal cofactors. Previously, we demonstrated that sequence diversity was most abundant in deoxyribozyme populations derived from the transition metal ions Cu^{2+} and Mn^{2+}. In contrast, very few catalytic sequences were derived when the main group metal ions, Mg^{2+} and Ca^{2+}, were included in the selection buffer (*31*). These results were obtained in the context of a DNA phosphorylation reaction. However, it is conceivable that the observed trend may still be applicable to other types of chemical transformations as well. For instance, many other deoxyribozymes that catalyze RNA cleavage (*15*, *17*, *19*, *39*) or the synthesis of branched RNA (*23*) have been shown to be highly specific for Mn^{2+} or highly active in the presence of this metal cofactor. Mg^{2+} was included in the selection buffer, since deoxyribozymes must be able to function in a Mg^{2+}-rich environment to be biologically relevant, and so the inclusion of Mg^{2+} may facilitate the isolation of new RNA-cleaving DNAzymes suitable for in vivo applications.

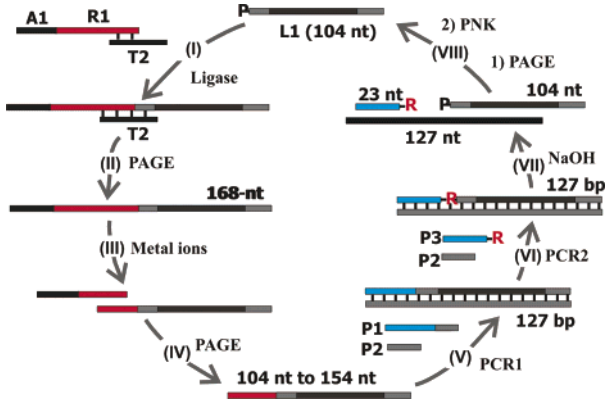


FIGURE 2: In vitro selection scheme. Each round of in vitro selection consists of steps I−VIII: (I) The A1-R1-L1 chimeras are constructed by DNA/RNA ligation. (II) The ligated 168-nt A1-R1-L1 is isolated by 10% denaturing PAGE. (III) The catalytic DNA molecules are activated by incubation with divalent metal ions for a designated period of time, followed by quenching with EDTA. (IV) The 104-nt to 154-nt cleavage fragments are isolated by PAGE. (V) The recovered DNAs are amplified by PCR using primers P1 and P2. (VI) The 127-bp PCR product from step V is reamplified using primers P1 and P3 to incorporate a ribonucleotide linkage within the DNA. (VII) The double-stranded DNA products are treated with NaOH to cleave the RNA linkage. (VIII) The 104-nt single-stranded DNA is purified by PAGE, phosphorylated at the 5′ end, and used to initiate the next round of in vitro selection.

In Vitro Selection Strategy. In vitro selection was conducted using the scheme shown in Figure 2. Each selection cycle consists of eight steps. In step I, the A1-R1-L1 chimeras were assembled by T4 DNA ligase in a reaction involving the use of A1-R1, 5′-phosphorylated L1, and a synthetic DNA template T2 (A1 was ligated to R1 prior to ligation with L1). The resultant DNA-RNA-DNA molecules were purified by 10% denaturing PAGE (step II), followed by incubation with divalent metal ion cofactors to promote catalytic activity (step III). The reaction mixture, in parallel with a sample containing two nucleic acid ladders (a 104-nt ssDNA and a full-length 50-nt RNA/104-nt DNA chimera), was subjected to 10% denaturing PAGE (step IV). The gel segment within the two ladders was excised and eluted so that all DNA molecules that cleaved a ribonucleotide linkage anywhere in the RNA motif could be recovered. The isolated DNAs were amplified by two consecutive PCR reactions (steps V and VI). The first PCR reaction used P1 and P2 as the primers. In addition to the 9-nt forward priming site, P2 contained 23 extra nucleotides at the 5′ end to introduce a new forward priming site for the second PCR reaction, which used primers P1 and P3. Since P3 was a ribo-terminated primer, the DNA product from the second PCR reaction could be digested under alkaline conditions to regenerate the single-stranded deoxyribozyme sequences (step VII). The DNA mixture was subjected to 10% denaturing PAGE and DNA phosphorylation (step VIII), and the resulting phosphorylated 104-nt DNA was used to initiate the next round of selection.

Establishment of Catalytic DNA Populations. Using the method described by Li and Breaker (*52*), the uncatalyzed rate of RNA cleavage under our selection criteria was estimated to be $\sim 3 \times 10^{-7}$ min^{-1}. Given the huge number of molecules in the initial library ($\sim 10^{14}$), we felt that it would be a reasonable expectation that many sequences would be capable of producing a rate enhancement of 10^4-fold. This rate enhancement would correspond to a catalytic population exhibiting a k_{obs} of $\sim 3 \times 10^{-3}$ min^{-1}. Therefore, as a tentative approximation, we chose 5 h as the initial reaction time in step III ($1 \div 3 \times 10^{-3}$ min$^{-1} \approx 333$ min or 5.5 h). Presumably, this reaction time would be stringent enough to select for catalytically active molecules but still allow for the establishment of a diverse pool of catalytic DNA sequences.

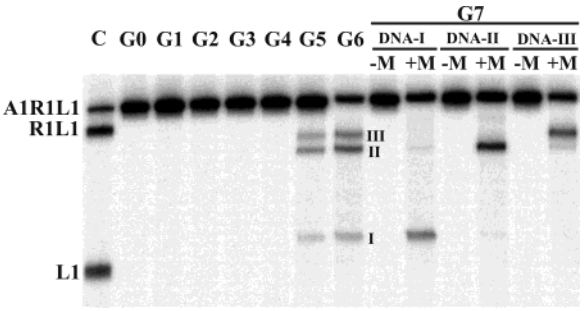


FIGURE 3: Establishment of catalytic DNA populations. The first lane represents a control indicating positions in a 10% denaturing PAGE gel corresponding to A1-R1-L1, R1-L1, and L1 (R1-L1 and L1 define the size boundaries of potential cleavage fragments). The remaining lanes illustrate the cleavage activity for generations 0−7. Three different cleavage bands were identified in G5, and the cleavage yields increased in G6. Each DNA band in G6 (labeled I, II, and III) was excised and treated separately for RNA cleavage in G7. All three DNA populations are dependent on divalent metal cofactors for catalytic activity. −M = without metal cofactors; +M = with metal cofactors. The RNA-cleavage reaction was performed in a selection buffer containing 50 mM HEPES, pH 7.0 at 23 °C, 400 mM NaCl, 100 mM KCl, 7.5 mM $MnCl_2$, 50 μM $CuCl_2$, and 7.5 mM $MgCl_2$.

In the first round (denoted generation 0 or G0), a pool of $\sim 10^{14}$ individual single-stranded RNA/DNA chimeras was used. The RNA-cleavage reaction was performed in selection buffer A, containing 7.5 mM Mg^{2+}, 7.5 mM Mn^{2+}, and 50 μM Cu^{2+} in addition to 50 mM HEPES, pH 7.0 at 23 °C, 400 mM NaCl, and 100 mM KCl. Since a high concentration of Cu^{2+} is known to be a denaturant to nucleic acid structures (*53*, *54*) and to exhibit strong inhibitory effects on the enzymatic activities of several deoxyribozymes (*26*, *31*, *32*, *55*), a relatively low concentration was used to minimize these effects.

Eight rounds of in vitro selection were initially conducted, and the progression of catalytic activity is illustrated in Figure 3. No cleavage product was observed in generations 0−4. However, three different cleavage bands were identified in G5, and the cleavage yields increased in G6. The data suggest that we established a catalytic DNA population that cleaved the attached RNA substrate at three different linkages or regions. Each DNA band in G6 (labeled as I, II, and III in Figure 3) was carefully excised, and the eluted DNA was amplified separately. The resulting three DNA populations, denoted DNA-I, DNA-II, and DNA-III (Figure 3), were treated separately for RNA cleavage in the eighth round of selection (G7). Each DNA population produced a primary cleavage band that corresponded well with the band that had been selected in G6, suggesting that DNA-I, DNA-II, and DNA-III cleaved the RNA substrate at either a defined linkage or one of the closely located linkages.

Metal Specificities of Selected Populations. As indicated in Figure 3, all three DNA populations from G7 required divalent metal cofactors for their cleavage activities. The next

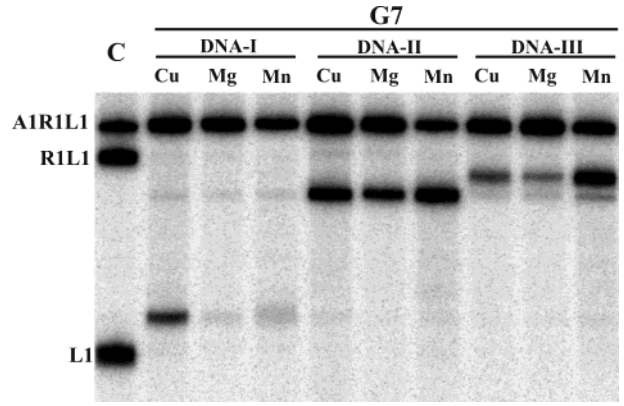


FIGURE 4: Metal specificities of selected populations. DNA-I, -II, and -III from G7 were assessed for their respective RNA-cleavage abilities in the presence of different divalent metal cofactors. The first lane represents a control indicating positions in a 10% denaturing PAGE gel corresponding to DNA/RNA sequences A1-R1-L1, R1-L1, and L1. DNA-I was most active in the presence of Cu^{2+} but only slightly active when Mg^{2+} or Mn^{2+} was used as the sole metal cofactor. DNA-II was able to use all three divalent metal ions, though it was most active with Mn^{2+} and the least active with Mg^{2+}. DNA-III was also active with all three metal ions, but again Mn^{2+} was preferred over the other two metal ions. Each reaction mixture contained 50 mM HEPES, pH 7.0 at 23 °C, 400 mM NaCl, and 100 mM KCl, in addition to 50 μM $CuCl_2$ and 15 mM $MgCl_2$ for Cu, 15 mM $MnCl_2$ for Mn, or 15 mM $MgCl_2$ for Mg.

logical step was to assess each DNA population for individual metal dependences (Figure 4). DNA-I was most active in the presence of Cu^{2+} but only slightly active when Mg^{2+} or Mn^{2+} was used as the sole metal cofactor. DNA-II was able to use all three divalent metal ions, though it was most active with Mn^{2+} and the least active with Mg^{2+}. DNA-III was also active with all three metal ions, but again Mn^{2+} was preferred over the other two metal ions. DNA-II and DNA-III exhibited the most robust activity in the presence of Mn^{2+} (15 mM $MnCl_2$), cleaving ~60% and ~50% of the attached RNA substrate in 5 h, respectively. DNA-I displayed the least activity, having cleaved just under 15% of the DNA/RNA chimera in the allotted 5 h, in a buffer containing 15 mM $MgCl_2$ and 50 μM $CuCl_2$.

Target Site Identification. A cleavage reaction mixture from G7 (produced in selection buffer A) was subjected to PAGE, and a long running time was used to promote optimal fragment separation. The gel fragments containing DNA-I, DNA-II, and DNA-III were carefully excised to avoid overlapping contamination. After PCR amplification in the absence of an internal ^{32}P label and ligation with the A1-R1 substrate sequence, each self-cleaving DNA/RNA construct was labeled with [γ-^{32}P]ATP using PNK and purified by PAGE. Experiments were then conducted to determine the RNA-cleavage site for each DNA population (Figure 5; lane 1 shows a partial RNA ladder generated by RNase T1 that identifies all Gs in the substrate domain; lane 2 is a full RNA ladder generated by alkaline digestion that accounts for all ribonucleotides; lanes 3−5 are reaction products of self-cleaving DNA-II, DNA-III, and DNA-I, respectively). The results indicate that each DNA population cleaved a specific site within the 50-nt RNA domain (shown in Figure 5C): DNA-I cleaved C_{42}-A_{43} (see the expanded gel section shown in Figure 5B), DNA-II cleaved the G_{11}-G_{12} junction, and DNA-III cleaved A_5-G_6 (Figure 5A). The precise reasons for these preferred cleavage sites have yet to be determined.

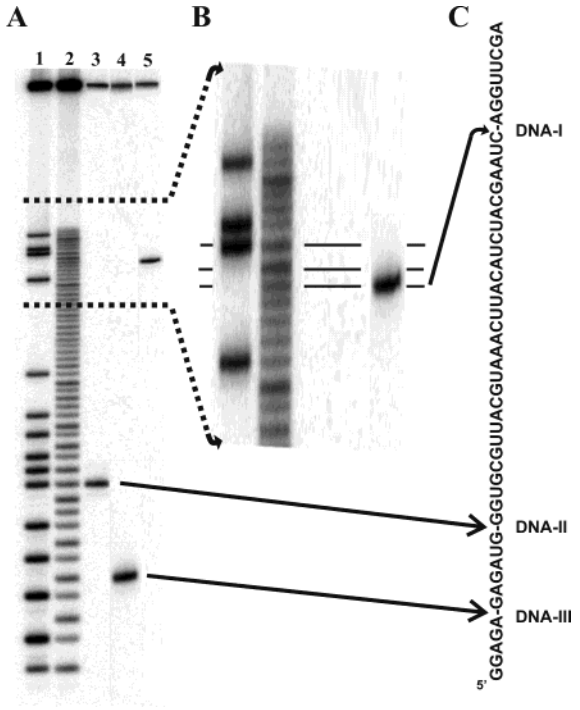


FIGURE 5: Target site identification. (A) Each A1-R1-L1 DNA-I, -II, and -III construct was 5′ end-labeled with [γ-^{32}P]ATP using PNK, purified by PAGE, and then used in experiments to determine the RNA-cleavage site for each DNA population. Lane 1 is a partial RNA ladder generated by RNase T1 that identifies all Gs in the substrate domain. Lane 2 is a full RNA ladder generated by alkaline digestion that accounts for all ribonucleotides. Lanes 3−5 are the reaction products of self-cleaving DNA-II, -III, and -I, respectively. (B) An expanded view of the gel section from (A) corresponding to the cleavage site for DNA-I. (C) Sequence of the 50-nt RNA substrate domain with specific cleavage site positions indicated by arrows. DNA-I cleaved C42-A43, DNA-II cleaved the G11-G12 junction, and DNA-III cleaved A5-G6.

However, we speculate that the preference was likely influenced by the intrinsic secondary or tertiary structure properties of the RNA motif, which may have made the three sites more prone to cleavage.

Selecting Catalytic DNA Populations Using Progressively Reduced Reaction Times. A preliminary sequence analysis indicated that each of the three DNA subpopulations contained a diverse pool of catalytic sequences (data not shown). We decided to subject DNA-II to additional rounds of in vitro selection under progressively reduced reaction times. The DNA-II population was specifically selected because it exhibited the highest cleavage activity in G7. At this stage, we felt it was no longer advantageous to include Cu^{2+} in the selection buffer, given its potential denaturant and inhibitory effect on deoxyribozymes. The DNA-II population had already been established and was most active in the presence of Mn^{2+}. Therefore, from generation 8 onward, we used a selection buffer that consisted of 7.5 mM $MnCl_2$, 7.5 mM $MgCl_2$, 100 mM KCl, 400 mM NaCl, and 50 mM HEPES.

In generation 8, a 15-nt extension (5′-AGCTGATCCT-GATGG) was engineered onto the existing, conserved 3′ primer-binding site, and a new P2 primer was employed in all subsequent rounds of PCR. This alteration would be used to assess the level of spontaneous mutagenesis occurring during PCR by measuring the number of acquired mutations in the "old" 15-nt P2 primer-binding site (designated as MP2

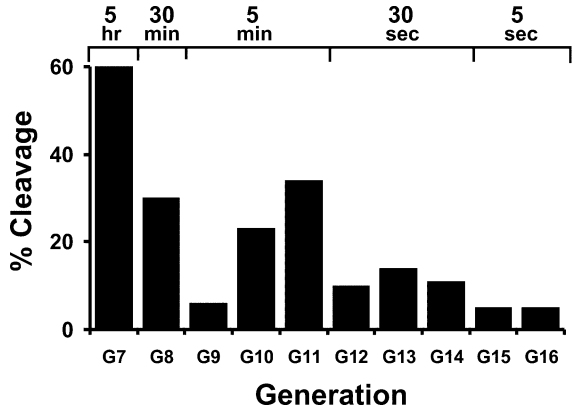


FIGURE 6: In vitro selection progress of DNA-II. Nine additional rounds of in vitro selection were performed on DNA-II after G7, while the reaction time was progressively reduced from 5 h (G0−G7), to 30 min (G8), to 5 min (G9−G11), to 30 s (G12−G14), and finally to 5 s (G15, G16). The reaction time was reduced when the population achieved a robust cleavage activity (∼30% cleavage) or exhibited a leveled activity over two to three selection rounds. The selection was discontinued at G16, since it would not be practical to reduce the reaction time below 5 s using manual quenching methods.

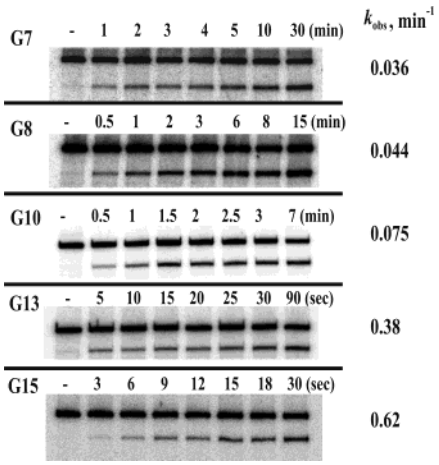


FIGURE 7: Kinetic characterization of selected generations from DNA-II. Time courses and corresponding k_{obs} values for DNA populations from selected generations representing each of the five selection times (G7 = 5 h, G8 = 30 min, G10 = 5 min, G13 = 30 s, and G15 = 5 s). The time points used for each time course are indicated above the relevant lanes. All populations exhibited significant rate enhancements ranging between $\sim10^5$- and $\sim10^6$-fold over the corresponding rate of uncatalyzed RNA cleavage ($\sim10^{-7}$ min^{-1}). The rate constant was taken as the negative slope of the line produced by a least squares fit to a plot of ln(fraction of DNA uncleaved) versus reaction time.

for convenience). Ultimately, this information would provide some quantitative criteria for the assignment of individual sequences into common sequence classes (see below).

Nine more rounds of in vitro selection were performed, while the reaction time was progressively reduced from 5 h (G0−G7), to 30 min (G8), to 5 min (G9−G11), to 30 s (G12−G14), and finally to 5 s (G15, G16). The in vitro selection progress is depicted in Figure 6. The reaction times were chosen somewhat arbitrarily, based largely on how the population responded to the given time challenge. The reaction time was reduced when the population achieved a robust cleavage activity (∼30% cleavage) or exhibited a leveled activity over two to three selection rounds. The selection was discontinued at G16, since it would not be practical to reduce the reaction time below 5 s using manual quenching methods.

Generations 7 (DNA-II population only), 8, 10, 13, and 15 were selected for analysis of their sequence diversity. These particular generations were chosen because they collectively represent all five different reaction times employed in the study, and each displayed fairly robust activity. A kinetic assay was conducted for each DNA population to confirm their catalytic ability (Figure 7). All populations exhibited significant rate enhancements ranging between $\sim10^5$- and $\sim10^6$-fold over the corresponding rate of uncatalyzed RNA cleavage. Generation 15 exhibited a k_{obs} of ∼0.62 min^{-1}. By comparison, the current repertoire of DNA enzymes exhibit rate constants that fall between 0.01 and 10 min^{-1}, though most do not exceed 1 min^{-1} (*4*).

Sequence Analysis. G7 (DNA-II population), G8, G10, G13, and G15 were cloned, and approximately 50 clones from each population were sequenced to reveal the diversity within (all sequences are given in Supporting Information, Figure 1, by generation). As a first approximation to facilitate the bulk sorting of sequences into tentative classes, we chose an arbitrary minimum requirement of 90% sequence identity between sequences (i.e., the bases in at least 72 locations within the 80-nt random-sequence domain are identical) and employed the Bioedit Sequence Alignment Editor computer software to aid in the alignment and grouping of sequences (Supporting Information, Figure 2, lists all sequences by class).

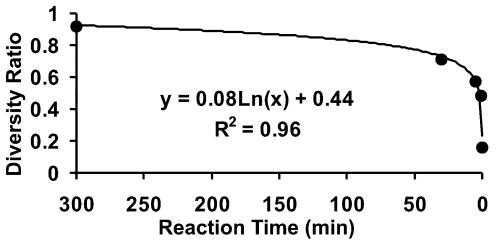
Figure 8 summarizes the data in both graphical and tabulated form. In total, 245 clones were sequenced and 110 classes were revealed. As a measure of the sequence diversity within each population, the "sequence diversity ratio" was established. The diversity ratio is defined as the ratio of observed sequence classes over the total number of sequenced clones (Figure 8A). A plot of the sequence diversity ratio versus the reaction time (Figure 8B) shows good adherence ($R^2 = 0.96$) to a *logarithmic* regression trend line. In general, the sequence diversity is observed to decrease with decreasing reaction time. For practical reasons, only ∼50 clones were sequenced per population; however, we anticipate that the relationship would be consolidated by even more sequence data. Figure 8C shows the frequency and distribution of individual clones within each generation. While most classes (39 out of 43) in G7 existed as a single copy (represented by "39 × 1" in Figure 8C, where "39" stands for 39 classes and "×1" stands for a single copy), most classes in G15 had 4−15 copies each. Figure 9 traces the evolution of all of the sequence classes from G7 to G15. All but one of the sequences in G15 are represented in a preceding generation. Interestingly, the two dominating sequence classes in G15, SC1 and SC2 (15 and 14 copies, respectively), did not dominate in G13 and are not even observed in preceding generations. On a similar note, the two dominating sequence classes in G10, SC6 and SC17 (8 copies each), appear to play very minor roles in the terminal population.

To assess the degree of sequence diversity present in each population, it was necessary to establish some criterion for the assignment of sequenced clones into common sequence classes. Although in vitro selection, rather than evolution (*56*), was intentionally employed throughout the course of

A

Generation	G7	G8	G10	G13	G15
Time (min)	300	30	5	0.5	0.083
Total Sequences	48	49	49	50	49
Unique Sequences	43	35	28	23	8
Diversity Ratio	0.90	0.71	0.57	0.46	0.16

B



C

Generation	G7	G8	G10	G13	G15
Sequence Distribution & Frequency	39x1	26x1	20x1	13x1	3x1
	3x2	7x2	5x2	5x2	1x4
	1x3	1x3	1x3	2x3	1x5
		1x6	2x8	1x5	1x8
				1x7	1x14
				1x9	1x15

FIGURE 8: Summary statistics relating sequence diversity to reaction time. (A) DNA populations from G7 (5 h), G8 (30 min), G10 (5 min), G13 (30 s), and G15 (5 s) were cloned and sequenced. As a measure of the sequence diversity within each population, the sequence diversity ratio was established. This diversity ratio is an arbitrary parameter defined as the ratio of unique sequence classes over the total number of sequenced clones. (B) The sequence diversity ratio vs reaction time. The sequence diversity ratio from (A) is illustrated graphically, along with the equation of the line of best fit. The correlation coefficient, *R*, indicates good adherence to a logarithmic regression trend line. (C) The sequence distribution and frequency for each generation are indicated as (number of different sequence classes) × (number of sequenced clones in each class).

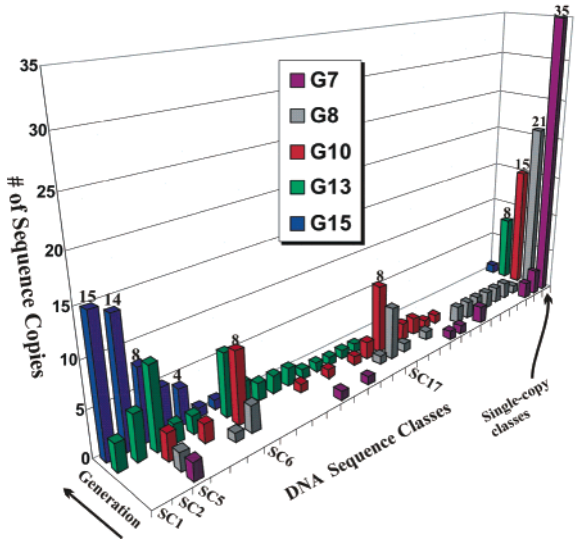


FIGURE 9: A catalytic DNA landscape. A catalytic DNA landscape was constructed to trace the rise and fall of individual sequence classes through time and sequence space. Generations 7, 8, 10, 13, and 15 are plotted against all of the catalytic DNA classes. The number of sequence copies in each DNA sequence class is illustrated in the third dimension. For simplicity, the total number of single-copy/single-generation sequence classes (those classes composed of only one sequence and occurring only in one generation) in each generation is depicted as a single bar at the far right of the graph.

this study, sequence classification would not be entirely straightforward since *Taq* polymerase does have a measurable error rate that would lead to some level of spontaneous mutagenesis during the PCR amplification steps.

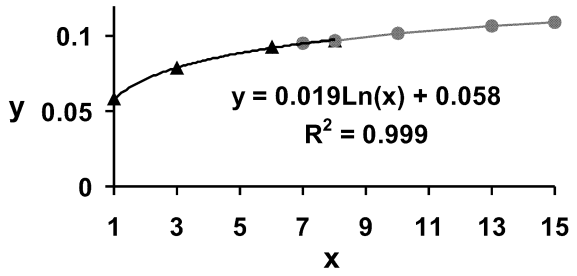
Next, we wanted to determine if the above classification criterion was vigorous enough to provide an accurate measure of the sequence diversity. To this end, mutational statistics were compiled to compare the theoretically expected, versus experimentally observed, degree of sequence identity between sequences within a class. Rather than use the published *Taq* polymerase error rate that is not specific to our PCR conditions, we used a more informative experimental approach to establish expected values for the sequence identity. Figure 10A summarizes the mutational data derived from the former P2 primer-binding site (MP2). As mentioned previously, starting from generation 8 this 15-nt sequence was no longer used as a primer-binding site and was therefore susceptible to PCR-induced mutations. Moreover, the mutational data gained from this region could be used to predict the maximum expected number of mutations in the random-sequence domain (Figure 10B).

The total number of mutations per generation (point mutations, deletions, and/or insertions) were counted and used in conjunction with the total number of sequences per generation to determine the overall mutational rate per nucleotide per generation, denoted as "*y*", where $y =$ (total number of mutations)/(15 × total number of sequences). Parameter "*x*" represents the number of rounds of selection that MP2 underwent, simply calculated as the overall generation number minus 7. The expected number of mutations in the 80-nt random-sequence domain (calculated as $y \times 80$) and the observed number of mutations in the SC5 sequence class are also illustrated for comparative purposes. SC5 is a unique sequence class in that it contained representatives from each generation, which allowed us to catalog the average number of acquired mutations in each generation relative to G7. In every case, the *observed* number of mutations is lower than the *expected* number of mutations. Figure 10B is a plot of the parameters *y* versus *x*. The graph shows exceptional adherence ($R^2 = 0.999$) to a logarithmic regression trend line and was used to predict the mutational rate per nucleotide for generations 8, 10, 13, and 15 through extrapolation of the data. Figure 10C compares the predicted and observed percent sequence identity within a sequence class and also provides some additional statistics for perspective. The observed within-class sequence identity values represent the minimum average value of all possible permutations (see Experimental Procedures). The predicted number of mutations (second row of Figure 10C) for G7 to G15 was calculated using the equation derived from Figure 10B, and the predicted sequence identity between sequences *within* a class (third row) was then determined from this value. In each generation, the *observed* within-class sequence identity (fourth row) is higher than the predicted value. The sequence identities between the two closest (fifth row) and two farthest sequence classes (sixth row) are provided as further validation of our proposed sequence classifications. The maximum observed identity *between* sequence classes does not exceed 53%, whereas the minimum observed sequence identity *within* sequence classes is ~93%. Considering that our original combinatorial DNA library covered

A

Generation	8	10	13	15
Total observed mutations	43	58	70	71
Total sequences	49	49	50	49
y (overall mutation rate, per nucleotide)	0.058	0.079	0.093	0.097
x (generation-7)	1	3	6	8
Expected mutations	4.7	6.3	7.5	7.7
Observed mutations in SC5	4.5	4.3	5.9	6.0

B



C

Generation	7	8	10	13	15
Predicted mutations	7.6	7.8	8.1	8.5	8.8
Predicted sequence identity within classes	90.5%	90.2%	89.8%	89.3%	89.1%
Observed sequence identity within classes	95.0%	95.7%	95.6%	96.5%	92.8%
Observed identity between the two closest classes	50.6%	50.0%	52.5%	51.2%	50.0%
Observed identity between the two farthest classes	12.5%	12.5%	20.0%	15.0%	18.7%

FIGURE 10: Mutational analysis and sequence classification statistics. (A) Summary of the mutational characteristics of the former P2 primer-binding site (denoted MP2). The total number of mutations in this 15-nt region per generation was counted and used in conjunction with the total number of sequences per generation to determine the overall mutational rate per nucleotide per generation, denoted as *y*. Parameter *x* is equivalent to the number of rounds of selection that MP2 underwent. The *expected* number of mutations in the 80-nt random-sequence domain and the *observed* average number of mutations in the SC5 sequence class (relative to G7) are also given for comparative purposes. (B) A correlational graph of parameters *y* versus *x*. Triangles represent real data points, while circles represent extrapolated data points. The equation of the line of best fit can be used to estimate the mutational rate per nucleotide for generations 8, 10, 13, and 15. (C) Sequence classification statistics. The predicted number of mutations per 80-nt random-sequence domain was calculated on the basis of the equation from (B), and the *predicted* sequence identity among sequences within a sequence class was determined from this value. The *observed* sequence identity among sequences within a class based on our proposed classification (see Experimental Procedures) is also shown. For comparison, the observed sequence identity *between* classes for the two closest and two farthest sequence classes in each generation is also detailed.

only a miniscule fraction of the entire sequence space available to 80 random nucleotides (which is 1.5×10^{48}, though only $\sim 10^{14}$ molecules were actually used), it seems highly probable that the individual sequences within a sequence class were derived from a single parental sequence in the original library. The preceding analysis suggests that our sequence classifications are adequate.

DISCUSSION

Even a decade after its inception, the underlying processes that govern the outcome of in vitro selection still remain somewhat of a mystery. Presumably, with a better understanding of the selection process we can devise better experimental designs and protocols that will ultimately yield better enzymes. The objective of this study was to help to increase our understanding of the pathway by which all DNA enzymes are created, by elucidating one aspect of in vitro selection. Specifically, we wanted to examine the relationship between the sequence diversity and the reaction time as a selection pressure. Sequence diversity is a primary consideration in deoxyribozyme selection studies because their success is predicated entirely on probability. In order for an in vitro selection experiment to succeed, at least one DNA molecule in a combinatorial library must possess some penchant for catalysis. Therefore, adequate sequence diversity is a prerequisite for success. Deoxyribozyme engineers control the reaction time and use it as one means to manipulate the sequence diversity in favor of more efficient DNA enzymes.

After 17 rounds of in vitro selection with DNA-II, we obtained a population of single-stranded DNA molecules that catalyze efficient ($k_{obs} \sim 0.62$ min^{-1}) cleavage of a G-G ribonucleotide junction within a 50-nt RNA substrate. The reaction time was reduced four times through the course of our experiment (from 5 h in G0−G7, to 30 min in G8, to 5 min in G9−G11, to 30 s in G12−G14, to 5 s in G15 and G16), and one representative population from each time class was cloned and sequenced. The relationship between sequence diversity and reaction time is illustrated in Figure 8B. In general, a decline in sequence diversity is observed with decreasing reaction time, and the relationship appears to be logarithmic. At least for the selection of RNA-cleaving deoxyribozymes, it seems that an initial reaction time greater than 5 h might be unnecessary. At 5 h the sequence diversity ratio is >0.90, and any sequence diversity gained from an even longer reaction time might not justify the additional time investment. Although it has not been experimentally confirmed, we suspect that the general logarithmic relationship is applicable to other chemical transformations as well. However, for more complex reactions exhibiting lower background rates, we expect the trend line to fall below that observed for the RNA-cleavage reaction in Figure 8B and perhaps shifted to the left. Presumably, fewer catalytic DNA sequences will exist in a given random-sequence pool, leading to a reduction in the observed sequence diversity. In this case, an initial reaction time greater than 5 h may be required.

It would be of interest to know the theoretical basis for this logarithmic dependence, although we cannot ourselves offer an adequate explanation at this point in time. The logarithmic relationship between the sequence diversity and reaction time may conceal some deeper fundamental relationship between the sequence diversity and catalytic activity of the deoxyribozyme inhabitants in a given selected pool. As a selection pressure, the reaction time must reflect the intrinsic catalytic rate of the population, so this hypothesis may have some merit. Graphical manipulation of the diversity ratio versus k_{obs} data (i.e., with and without log axes) yields some interesting relationships (data not shown), but whether they are meaningful remains to be answered.

The 5 h G7 population boasts 43 sequence classes in contrast to the 8 discrete sequences observed in the 5 s G15 population (Supporting Information, Figure 1). The 8 se-

A

DZ1	ATTTCCAGCGGATCGATTCTCTTTCCCGTCGTAGGTATGACCAGGGAAGAATAGGTGGACATAAATTGATGGTGTCGGG	(15)
DZ2	ACTTCCAGCGGATCGAAATCTTGAACGCAGCTAGGTCTCGGGTGTGGCGGTGAGTTGGCGTAGGCCATGCCTTCCGCTGG	(14)
DZ3	ACTGCTAGCAGCTCGAAATCGCTCTCTCAATATGGGCTTTCGGGGAAGACGGTAATAGGAGAAATGGTGCCTTGTCCGTG	(8)
DZ4	ACCTCAATAGCAGCGTTAACAAAAAGTTTCGAGAAAGCGAATCATCTAACGGTGGTGACTCCATTGGTTTTTTGGGTGGG	(5)
DZ5	ACTGCCAGCGGCGCGAGGCTCTTGATCGGGCGCAGGAGGGGACCGGTGATATCGGCATCCTCGATGTTAGACTGGATGGT	(4)
DZ6	ACTGCCAGTGGCGCGAATTCTCTGGGAGATCTGTATAGGGTTGCCTGCGAGTTGACAGGGATGGTGTGCAGTTTGTGTGG	(1)
DZ7	ACTGCCAGCGGCGCGATTCACTGTCGGAGACTAGTTGTTGCCTTCGGCTTGGAAGGACAAAACTTTTGTCATAGCGTGGG	(1)
DZ8	TACTCTCAGTGAGGCGAAATCTTCTCTCTGCGGGAACAATCGGGGGCGCAGTGATCAAGGGTGGAAATGGGGATGGGTG	(1)

B

FIGURE 11: G15 sequence classes and the putative 8-17 motifs. (A) The sequences of 49 clones from G15 grouped into eight discrete sequence classes, DZ1−DZ8. The number of clones that appeared in each sequence class is shown at the right of each sequence. Variants of the 8-17 motif are underlined for easy recognition. (B) Putative 8-17 secondary structure arrangements for DZ1−DZ8 (lower sequence) and their possible binding interactions with the substrate domain (upper sequence). For simplicity, only the 8-17 concerning region is depicted.

quence classes in G15 are somewhat surprising given the demanding selective pressure used in this generation. However, closer scrutiny shows that 92% of all sequences are distributed over 5 sequence classes, and 58% of all sequences are distributed over just 2 classes. Interestingly, the two most abundant sequences occur in approximately equal frequency and share the same 14-nt consensus sequence within the full-length 80-nt random domain. Thus, these two *sequence* classes may represent two very closely related *deoxyribozyme* classes.

Prevalence of 8-17. Prior to this study, an RNA-cleaving deoxyribozyme known as 8-17 had been identified in four independent in vitro selection studies (*9*, *10*, *14*, *19*). Not surprisingly, a thorough examination of all sequenced clones from generations 7, 8, 10, 13, and 15 revealed numerous representatives of this seemingly ubiquitous DNA motif (see Supporting Information, Figure 1). Figure 11A illustrates eight deoxyribozyme sequences from G15, one from each sequence class (they are named DZ1−DZ8, corresponding to the top sequence of each sequence class given in Supporting Information, Figure 1). All of the deoxyribozyme sequences appear to contain a variant of the 8-17 motif close to the 5′ end. Figure 11B illustrates the putative secondary structure arrangement of each discrete sequence observed in G15 as it might potentially interact with the substrate domain. These motifs are key suspects for the observed catalytic activity since they correspond very well with the original secondary structure description proposed by Santoro and Joyce (*10*) and contain the following four demonstrated features (*10*, *14*, *19*, *39*): (1) a G•T wobble at the cleavage site; (2) a 3-bp stem; (3) the conserved A and G in the bottom-left triloop (italicized), and (4) the conserved C and G in the single-stranded region opposite the cleavage site (italicized) except for DZ4, which only has a C in this region but has several adenines. Synthetic oligonucleotides corresponding to each putative deoxyribozyme exhibited strong RNA-cleavage activity toward the RNA substrate R1 in trans as between 68% and 96% of R1 was cleaved in 2 h under a single-turnover condition ([DNAzyme] = 5 μM, [R1] = 0.05 μM; data are shown in Supporting Information, Figure 3). This analysis suggests that the suspected 8-17 motifs are indeed responsible for the catalytic activity.

Interestingly, the prevalence of the small 8−17 motifs may account for the high degree of sequence diversity observed in our selection study. In other words, another independent in vitro selection experiment for RNA-cleaving DNA enzymes may potentially reveal a lower degree of sequence diversity for a given reaction time, if the occurrence of 8-17 is somehow avoided, whether it is by chance or by design. Nevertheless, we expect the logarithmic relationship between the sequence diversity and reaction time to remain intact, since the presence of 8-17 would represent a systematic increase in the frequency of observed catalytic sequences in each time class, which may affect the absolute position of the trend line in Figure 8B, rather than the nature of the relationship itself. This hypothesis will be tested using other in vitro selection studies in the future.

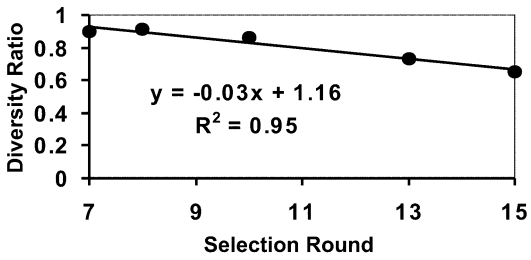
Though discouraging for those in search of new RNA-cleaving DNAzyme motifs, the recurrence of 8-17 is not surprising. In vitro selection is, after all, predicated on the old adage of "survival of the fittest". 8-17 has many favorable characteristics that make it highly fit, in an evolutionary context. Its large catalytic rate coupled with its ability to function under various metal ion conditions, as well as its capacity to cleave multiple dinucleotide junctions, allows 8-17 to readily survive and propagate under typical selection pressures imposed during most in vitro selection experiments. Furthermore, the catalytic core of 8-17 is quite small and the base composition highly variable, which means that this motif is likely to be present at a relatively high frequency in any given library. Taken together, these traits give 8-17 a highly competitive edge and a unique ability to "hijack" in vitro selections. Indeed, this motif likely represents a kind of evolutionary trap, a strong local optimum in the adaptive fitness landscape as described by Lehman in his discussion of the factors that influence the recurrence of nucleic acid enzymes (*57*).

Catalytic Fitness Landscape. Figure 9 illustrates the distribution and frequency of sequenced clones from G7 to G15. The fitness landscape is essentially flat in G7, due to the relatively permissive reaction time used in this generation. However, most of the sequences that flourished during earlier

A

Generation	G7	G8	G10	G13	G15
Time (min)	300	300	300	300	300
Total Sequences	48	45	44	48	46
Unique Sequences	43	41	38	35	30
Diversity Ratio	0.90	0.91	0.86	0.73	0.65

B



C

Generation	G7	G8	G10	G13	G15
Sequence Distribution & Frequency	39_{X1} 3_{X2} 1_{X3}	38_{X1} 2_{X2} 1_{X3}	34_{X1} 2_{X2} 2_{X3}	26_{X1} 6_{X2} 2_{X3} 1_{X4}	19_{X1} 8_{X2} 1_{X3} 2_{X4}

FIGURE 12: Summary statistics relating sequence diversity to selection round with constant selection pressure. (A) A control experiment was conducted to evaluate the effect of increasing the number of rounds of selection on the sequence diversity, using a constant reaction time of 5 h. DNA populations from G7, G8, G10, G13, and G15 of the control experiment were cloned and sequenced. The diversity ratio is an arbitrary parameter defined as the ratio of unique sequence classes over the total number of sequenced clones. (B) The sequence diversity ratio vs selection round. The sequence diversity ratio from (A) is illustrated graphically, along with the equation of the line of best fit. The correlation coefficient, *R*, indicates good adherence to a linear regression trend line. (C) The sequence distribution and frequency for each generation are indicated as (number of different sequence classes) × (number of sequenced clones in each class).

generations are absent by G15, presumably because they could not contend with the increasing stringency of the shorter reaction times imposed during in vitro selection. The ancestral sequences of most G15 sequences were not detectable in G7. These G15 sequences may be the progeny of previously inefficient catalysts that acquired advantageous mutations during the PCR step of subsequent selection rounds, or they may have simply escaped previous observation due to a highly diverse sequence pool and the limited sample set chosen for sequence analysis. Interestingly, the sequences that dominate in the middle pool (G10) do not appear to dominate in the terminal pool, which is consistent with the findings presented in a previous study involving the isolation of ligase ribozymes (*58*). Presumably, these sequences were outcompeted by other more proficient catalytic motifs under the increased selection pressure.

A Control Experiment. In our proposed experimental design we have simultaneously changed two independent variables (i.e., reaction time and the number of selection rounds) while trying to evaluate the effect of only one on the sequence diversity. Therefore, to clearly delineate the individual roles of each variable, we devised a control experiment to examine the effect of simply increasing the number of rounds of selection without changing the selection pressure (i.e., no reduction in reaction time). We started with the DNA-II population from G7 and conducted eight additional rounds of selection using a constant reaction time of 5 h. To help to prevent cross-contamination between the two selection experiments, a different 15-nt 3′ primer-binding site (5′-TCCTGGAGCTGATGA) was appended to the G8 control population. Both of the new primer-binding sites contained the same frequency and similar distribution of the four standard nucleotides to minimize unequal PCR amplification efficiency between primers. As an added barrier to cross-contamination the two selection experiments were conducted sequentially in time, rather than in parallel.

Figure 12A summarizes the results of the control experiment in the same format as was presented in Figure 8A for the original selection experiment. A plot of the sequence diversity ratio versus the number of selection rounds (Figure 12B) shows good adherence ($R^2 = 0.95$) to a linear regression trend line having a slightly negative slope of −0.03 diversity ratio units/selection round, which corresponds to a loss of ∼2 sequence classes per selection round (assuming a population sampling size of 50 clones). The *y*-intercept (i.e., 1.16) exceeds the theoretical maximum value for the diversity ratio, which is defined as 1 since the number of sequence classes (the numerator) can only be less than or equal to the number of sequenced clones (the denominator). This observation is likely an artifact of the limited population sampling size (i.e., ∼50 clones) used in this study. Interestingly, the diversity ratio is equal to 1 when the selection round is approximately 5, which corresponds to the appearance of the first detectable catalytic populations (see Figure 3). Figure 12C shows the frequency and distribution of individual clones within each generation. Despite many rounds of selection, the 5 h G15 control population does not exhibit any dominant sequence classes as was characteristic of the 5 s G15 population. There are no sequence classes with more than four clones, and the vast majority of classes have only one or two members.

The linear dependence observed between the diversity ratio and selection round in Figure 12B indicates that the logarithmic relationship observed in Figure 8B is largely due to the reaction time with little contribution from the number of selection rounds. Nevertheless, the slightly negative slope in Figure 12B cannot be ignored. We can suggest two possible reasons that may account for the modest decrease in sequence diversity that occurs with increasing rounds of selection, in the absence of any change in selection pressure. First, PCR bias may be involved, in which case the additional rounds of selection become somewhat counterproductive, since the sequence variants are being propagated on the basis of their amplification efficiency rather than their catalytic efficiency. Second, some sequences may become slightly enriched because of superior catalytic efficiency but will not have the opportunity to fully exploit that advantage and dominate the population, simply because there is no adequate selective pressure acting to eliminate the lower fitness sequence variants. Interestingly, if we presume this linear dependence is constant over future generations, extrapolation of the data suggests it would take approximately 30 rounds of selection to match the sequence diversity observed in the 5 s G15 population.

Perspective and Significance. The reaction time is a key parameter of in vitro selection experiments. However, a

review of the literature generally reveals a lack of consensus as to what reaction time(s) will yield the best results. This, of course, is consistent with the fact that no one really knows the answer. Ultimately, our study may add some measure of justification to a choice that is currently made rather arbitrarily.

Although multiple reaction times may be employed in a selection experiment, the length of the *initial* reaction time is probably the most important choice to make. This is because the initial library of DNA molecules (for $N \geq 26$ nt and library size $\leq 10^{16}$) contains only one or a few copies of each sequence variant, and so the sequence diversity is especially vulnerable to random DNA losses. For instance, only a fraction of the starting DNA sample (50−80%) is recovered following a typical PAGE purification step due to incomplete elution. Fortunately, as selection progresses, the catalytic DNA sequences become more resilient to random and inherent experimental losses, since multiple copies of each sequence variant will come to exist after PCR amplification.

Not surprisingly, deoxyribozyme engineers try to establish the catalytically active DNA population as quickly as possible. However, one must try to balance the "need for speed" with prudent judgment and be careful not to introduce too much selection pressure too fast. While it seems logical to use a highly stringent reaction time to select for highly efficient deoxyribozymes, doing so may not necessarily yield the best results. There could be some DNA molecules that possess *latent* catalytic ability, which would be lost under conditions that are initially too stringent. These DNA molecules may simply need to be optimized at a few base locations to expose their full catalytic capacity. Optimization could occur passively, during standard PCR from random mutations, or aggressively, by employing hypermutagenic PCR to introduce mutations at a high frequency.

The reaction time has varied widely from one in vitro selection study to another, reflecting different objectives and different opinions on the relative merits of long versus short reaction times. We recommend a long initial reaction time to establish a diverse catalytic DNA population to which further demanding selection pressure can later be applied. This is a conservative, but effective approach for the following reasons. First, a longer reaction time is equated with greater sequence diversity, which in turn provides more potential DNA motifs for catalysis. Second, extremely efficient DNA enzymes can still be obtained by imposing stringent reaction times during later rounds of selection and by employing mutagenic PCR to introduce favorable mutations. Third, it minimizes the loss of latent DNA enzymes during the initial rounds of selection when the population diversity is especially vulnerable to random experimental losses. Fourth, it will provide an opportunity to perform reselection on a diverse pool at a later date, perhaps to select for some other desirable feature. Finally, as a worst case scenario, the use of a long initial reaction time might entail the unnecessary expenditure of time and resources. By contrast, the worst case scenario of using a reaction time initially too stringent could be an in vitro selection experiment that completely fails.

This study serves as a preliminary effort toward the comprehensive elucidation of the factors that influence in vitro selection. Future experiments will involve different model reactions, as well as different aspects of study, such as the length of the random-sequence domain.

ACKNOWLEDGMENT

We thank Sylva Zyba for technical assistance and members of the Li laboratory for helpful discussions and critical comments on the manuscript.

SUPPORTING INFORMATION AVAILABLE

Sequence classes within generations (Figure 1), sequences of each generation grouped into sequence classes (Figure 2), and PAGE-based assay for assessing activities of the suspected 8-17 motifs embedded in the sequences of the eight deoxyribozyme classes in G15 (Figure 3). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Breaker, R. R. (2000) Making catalytic DNAs, *Science 290*, 2095−2096.
2. Breaker, R. R. (1997) DNA enzymes, *Nat. Biotechnol. 15*, 427−431.
3. Breaker, R. R. (1997) DNA aptamers and DNA enzymes, *Curr. Opin. Chem. Biol. 1*, 26−31.
4. Emilsson, G. M., and Breaker, R. R. (2002) Deoxyribozymes: new activities and new applications, *Cell. Mol. Life Sci. 59*, 596−607.
5. Li, Y., and Breaker, R. R. (1999) Deoxyribozymes: new players in the ancient game of biocatalysis, *Curr. Opin. Struct. Biol. 9*, 315−323.
6. Sen, D., and Geyer, C. R. (1998) DNA enzymes, *Curr. Opin. Chem. Biol. 2*, 680−687.
7. Breaker, R. R., and Joyce, G. F. (1994) A DNA enzyme that cleaves RNA, *Chem. Biol. 1*, 223−229.
8. Breaker, R. R., and Joyce, G. F. (1995) A DNA enzyme with Mg(II)-dependent RNA phosphoesterase activity, *Chem. Biol. 2*, 655−660.
9. Faulhammer, D., and Famulok, M. (1996) The Ca(II) ion as a cofactor for a novel RNA-cleaving deoxyribozyme, *Angew. Chem., Int. Ed. Engl. 35*, 2809−2813.
10. Santoro, S. W., and Joyce, G. F. (1997) A general purpose RNA-cleaving DNA enzyme, *Proc. Natl. Acad. Sci. U.S.A. 94*, 4262−4266.
11. Geyer, C. R., and Sen, D. (1997) Evidence for the metal-cofactor independence of an RNA phosphodiester-cleaving DNA enzyme, *Chem. Biol. 4*, 579−593.
12. Roth, A., and Breaker, R. R. (1998) An amino acid as a cofactor for a catalytic polynucleotide, *Proc. Natl. Acad. Sci. U.S.A. 95*, 6027−6031.
13. Santoro, S. W., Joyce, G. F., Sakthivel, K., Gramatikova, S., and Barbas, C. F., III (2000) RNA cleavage by a DNA enzyme with extended chemical functionality, *J. Am. Chem. Soc. 122*, 2433−2439.
14. Li, J., Zheng, W., Kwon, A. H., and Lu, Y. (2000) In vitro selection and characterization of a highly efficient Zn(II)-dependent RNA-cleaving deoxyribozyme, *Nucleic Acids Res. 28*, 481−488.
15. Feldman, A. R., and Sen, D. (2001) A new and efficient DNA enzyme for the sequence-specific cleavage of RNA, *J. Mol. Biol. 313*, 283−294.
16. Perrin, D. M., Garestier, T., and Helene, C. (2001) Bridging the gap between proteins and nucleic acids: a metal-independent RNAseA mimic with two protein-like functionalities, *J. Am. Chem. Soc. 123*, 1556−1563.
17. Liu, Z., Mei, S. H., Brennan, J. D., and Li, Y. (2003) Assemblage of signaling DNA enzymes with intriguing metal-ion specificities and pH dependences, *J. Am. Chem. Soc. 125*, 7539−7545.
18. Mei, S. H., Liu, Z., Brennan, J. D., and Li, Y. (2003) An efficient RNA-cleaving DNA enzyme that synchronizes catalysis with fluorescence signaling, *J. Am. Chem. Soc. 125*, 412−420.
19. Cruz, R. P. G., Withers, J. W., and Li, Y. (2004) Dinucleotide junction cleavage versatility of 8-17 deoxyribozyme, *Chem. Biol. 11*, 57−67.
20. Flynn-Charlebois, A., Prior, T. K., Hoadley, K. A., and Silverman, S. K. (2003) In vitro evolution of an RNA-cleaving DNA enzyme into an RNA ligase switches the selectivity from 3′-5′ to 2′-5′, *J. Am. Chem. Soc. 125*, 5346−5350.

21. Flynn-Charlebois, A., Wang, Y., Prior, T. K., Rashid, I., Hoadley, K. A., Coppins, R. L., Wolf, A. C., and Silverman, S. K. (2003) Deoxyribozymes with 2′-5′ RNA ligase activity, *J. Am. Chem. Soc. 125*, 2444−2454.
22. Ricca, B. L., Wolf, A. C., and Silverman, S. K. (2003) Optimization and generality of a small deoxyribozyme that ligates RNA, *J. Mol. Biol. 330*, 1015−1025.
23. Wang, Y., and Silverman, S. K. (2003) Deoxyribozymes that synthesize branched and lariat RNA, *J. Am. Chem. Soc. 125*, 6880−6881.
24. Carmi, N., Balkhi, S. R., and Breaker, R. R. (1998) Cleaving DNA with DNA, *Proc. Natl. Acad. Sci. U.S.A. 95*, 2233−2237.
25. Carmi, N., Shultz, L. A., and Breaker, R. R. (1996) In vitro selection of self-cleaving DNAs, *Chem. Biol. 3*, 1039−1046.
26. Cuenoud, B., and Szostak, J. W. (1995) A DNA metalloenzyme with DNA ligase activity, *Nature 375*, 611−614.
27. Levy, M., and Ellington, A. D. (2002) In vitro selection of a deoxyribozyme that can utilize multiple substrates, *J. Mol. Evol. 54*, 180−190.
28. Levy, M., and Ellington, A. D. (2001) Selection of deoxyribozyme ligases that catalyze the formation of an unnatural internucleotide linkage, *Bioorg. Med. Chem. 9*, 2581−2587.
29. Sreedhara, A., Li, Y., and Breaker, R. R. (2004) Ligating DNA with DNA, *J. Am. Chem. Soc. 126*, 3454−3460.
30. Li, Y., and Breaker, R. R. (1999) Phosphorylating DNA with DNA, *Proc. Natl. Acad. Sci. U.S.A. 96*, 2746−2751.
31. Wang, W., Billen, L. P., and Li, Y. (2002) Sequence diversity, metal specificity, and catalytic proficiency of metal-dependent phosphorylating DNA enzymes, *Chem. Biol. 9*, 507−517.
32. Li, Y., Liu, Y., and Breaker, R. R. (2000) Capping DNA with DNA, *Biochemistry 39*, 3106−3114.
33. Li, Y., Geyer, C. R., and Sen, D. (1996) Recognition of anionic porphyrins by DNA aptamers, *Biochemistry 35*, 6911−6922.
34. Li, Y., and Sen, D. (1996) A catalytic DNA for porphyrin metalation, *Nat. Struct. Biol. 3*, 743−747.
35. Li, Y., and Sen, D. (1997) Toward an efficient DNAzyme, *Biochemistry 36*, 5589−5599.
36. Li, Y., and Sen, D. (1998) The modus operandi of a DNA enzyme: enhancement of substrate basicity, *Chem. Biol. 5*, 1−12.
37. Sheppard, T. L., Ordoukhanian, P., and Joyce, G. F. (2000) A DNA enzyme with N-glycosylase activity, *Proc. Natl. Acad. Sci. U.S.A. 97*, 7802−7807.
38. Koizumi, M., and Breaker, R. R. (2000) Molecular recognition of cAMP by an RNA aptamer, *Biochemistry 39*, 8983−8992.
39. Santoro, S. W., and Joyce, G. F. (1998) Mechanism and utility of an RNA-cleaving DNA enzyme, *Biochemistry 37*, 13330−13342.
40. Cairns, M. J., Saravolac, E. G., and Sun, L. Q. (2002) Catalytic DNA: a novel tool for gene suppression, *Curr. Drug Targets 3*, 269−279.
41. Khachigian, L. M. (2000) Catalytic DNAs as potential therapeutic agents and sequence-specific molecular tools to dissect biological function, *J. Clin. Invest. 106*, 1189−1195.
42. Lowe, H. C., and Khachigian, L. M. (2002) Coating stents with antirestenotic drugs: the blunderbuss or the magic bullet?, *Circulation 105*, E29.
43. Li, J., and Lu, Y. (2000) A highly sensitive and selective catalytic DNA biosensor for lead Ions, *J. Am. Chem. Soc. 122*, 10466−10467.
44. Stojanovic, M. N., de Prada, P., and Landry, D. W. (2001) Catalytic molecular beacons, *ChemBiochem. 2*, 411−415.
45. Stojanovic, M. N., Mitchell, T. E., and Stefanovic, D. (2002) Deoxyribozyme-based logic gates, *J. Am. Chem. Soc. 124*, 3555−3561.
46. Stojanovic, M. N., and Stefanovic, D. (2003) A deoxyribozyme-based molecular automaton, *Nat. Biotechnol. 21*, 1069−1074.
47. Stojanovic, M. N., and Stefanovic, D. (2003) Deoxyribozyme-based half-adder, *J. Am. Chem. Soc. 125*, 6673−6676.
48. Lowe, H. C., Fahmy, R. G., Kavurma, M. M., Baker, A., Chesterman, C. N., and Khachigian, L. M. (2001) Catalytic oligodeoxynucleotides define a key regulatory role for early growth response factor-1 in the porcine model of coronary in-stent restenosis, *Circ. Res. 89*, 670−677.
49. Santiago, F. S., and Khachigian, L. M. (2001) Nucleic acid based strategies as potential therapeutic tools: mechanistic considerations and implications to restenosis, *J. Mol. Med. 79*, 695−706.
50. Stojanovic, M. N., de Prada, P., and Landry, D. W. (2000) Homogeneous assays based on deoxyribozyme catalysis, *Nucleic Acids Res. 28*, 2915−2918.
51. Sabeti, P. C., Unrau, P. J., and Bartel, D. P. (1997) Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool, *Chem. Biol. 4*, 767−774.
52. Li, Y., and Breaker, R. R. (1999) Kinetics of RNA degradation by specific base catalysis of transesterification involving the 2′-hydroxyl group, *J. Am. Chem. Soc. 121*, 5364−5372.
53. Eichhorn, G. L., and Shin, Y. A. (1968) Interaction of metal ions with polynucleotides and related compounds. XII. The relative effect of various metal ions on DNA helicity, *J. Am. Chem. Soc. 90*, 7323−7328.
54. Rifkind, J. M., Shin, Y. A., Heim, J. M., and Eichhorn, G. L. (1976) Cooperative disordering of single-stranded polynucleotides through copper cross-linking, *Biopolymers 15*, 1879−1902.
55. Carmi, N., and Breaker, R. R. (2001) Characterization of a DNA-cleaving deoxyribozyme, *Bioorg. Med. Chem. 9*, 2589−2600.
56. Beaudry, A. A., and Joyce, G. F. (1992) Directed evolution of an RNA enzyme, *Science 257*, 635−641.
57. Lehman, N. (2004) Assessing the likelihood of recurrence during RNA evolution in vitro, *Artif. Life 10*, 1−22.
58. Bartel, D. P., and Szostak, J. W. (1993) Isolation of new ribozymes from a large pool of random sequences, *Science 261*, 1411−1418.

BI049757J