Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool

Pardis C Sabeti, Peter J Unrau and David P Bartel

Background: In the past few years numerous binding and catalytic motifs have been isolated from pools of random nucleic acid sequences. To extend the utility of this approach it is important to learn how to design random-sequence pools that provide maximal access to rare activities. In an effort to better define the relative merits of longer and shorter pools (i.e. pools with longer or shorter random-sequence segments), we have examined the inhibitory effect of excess arbitrary sequence on ribozyme activity and have evaluated whether this inhibition overshadows the calculated advantage of longer pools.

Results: The calculated advantage of longer sequences was highly dependent on the size and complexity of the desired motif. Small, simple motifs were not much more abundant in longer molecules. In contrast, larger motifs, particularly the most complex (highly modular) motifs, were much more likely to be present in longer molecules. The experimentally determined inhibition of activity by excess sequence was moderate, with bulk effects among four libraries ranging from no effect to 18-fold inhibition. The median effect among 60 clones was fivefold inhibition.

Conclusions: For accessing simple motifs (e.g. motifs at least as small and simple as the hammerhead ribozyme motif), longer pools have little if any advantage. For more complex motifs, the inhibitory effect of excess sequence does not approach the calculated advantage of pools of longer molecules. Thus, when seeking to access rare activities, the length of typical random-sequence pools (≤ 70 random positions) is shorter than optimal. As this conclusion holds over a range of incubation conditions, it may also be relevant when considering the emergence of new functional motifs during early evolution.

Address: Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

Correspondence: David P Bartel
E-mail: dbartel@wi.mit.edu

Key words: *in vitro* selection, random sequence, ribozymes, RNA folding, SELEX

Received: **30 July 1997**
Revisions requested: **26 August 1997**
Revisions received: **15 September 1997**
Accepted: **17 September 1997**

Chemistry & Biology October 1997, **4**:767–774
http://biomednet.com/elecref/1074552100400767

Introduction

The isolation of molecules with unique activities from large pools of random nucleic acid sequences has been important for examining the functional abilities of RNA and other nucleic acid polymers. New ribozymes isolated from random sequences have extended the catalytic repertoire of RNA and DNA far beyond what has been found in contemporary biology [1,2], and numerous molecules with interesting and potentially useful binding activities have been isolated [3,4]. The starting point of an *in vitro* selection experiment is a large pool of random sequences, typically generated by combinatorial synthesis on a DNA synthesizer. (The pool is used directly for DNA selections or transcribed for RNA selections.) Then, sequences with a desired biochemical property are enriched through an iterative selection-amplification process until they dominate the pool.

An *in vitro* selection experiment can only succeed if a molecule with the desired activity resides within the initial random-sequence pool. For some types of activity, such as specific binding to a protein, generating a sufficiently diverse initial pool is not a major concern; these selections nearly always succeed with pools of modest sequence complexity and length [3]. Other types of activity, however, are thought to be more rare in random sequences. Selections for ribozymes and for aptamers that bind small molecules have a lower success rate. Although it is often difficult to determine the reason for a particular failure, it is likely that in some instances the desired activity is so rare that the initial pool does not contain any sequences with activity sufficient to survive the first round of selection. As the *in vitro* selection method is being summoned for increasingly challenging tasks it has become crucial to learn how to design pools that maximize the ability to access rare activities. The three considerations in the design of random-sequence pools are: the identity of the constant regions used for primer-binding sites; the nucleotide composition of the random-sequence domain; and the length of the random-sequence domain.

To the extent that primer-binding sites can become key parts of the ribozymes or aptamers, the identity of the primer-binding sites influences the outcome of the selection

[5,6]. Assimilation of primer-binding sites cannot be predicted for new RNA activities, however. A more important design consideration is to use primer-binding sites that yield clean amplification for the ≥ 200 polymerase chain reaction (PCR) cycles of a typical *in vitro* selection experiment [7]. In regard to the second consideration, synthesis of the random-sequence domain typically strives for a non-biased nucleotide composition in which each of the four nucleotides is equally represented. This goal should possibly be modified. The observation that the nucleotide composition of functional RNAs is consistently biased suggests that a similarly biased initial pool would harbor a greater number of functional RNA sequences [8].

The length of the random-sequence domain has varied widely from one *in vitro* selection study to another (from < 30 to > 200 nucleotides; [2–4]) indicating different goals of the practitioners as well as different opinions on the relative merits of long and short random sequences. Such differences of opinion are not surprising given the absence of data on the subject. In a previous selection for RNA ligase ribozymes we used molecules with a random segment as long as was practical [6]. We argued that, if one wants to generate a pool of molecules that contains an activity thought to be exceedingly rare, the chances of success increase dramatically with the length of the random sequence. This is because increasing the length of the random region provides a combinatorial advantage for finding the set of sequence segments within the molecule that must interact to form a given RNA structure. But this argument ignores a possible inhibitory effect of excess arbitrary sequence; at some length misfolding induced by additional random sequence may inactivate the functional portion of the molecules more than the added sequence increases the abundance of functional molecules.

Here, we examine in more detail the calculated advantages of long random sequences for finding different motifs. We also add excess arbitrary sequences to ribozyme motifs and directly measure their inhibitory effects. Determination of both the positive and the negative effects of molecule length permits a more informed discussion of the importance of biopolymer length in accessing rare activities by *in vitro* selection and of its importance in the emergence of new activities during early evolution.

Results and discussion

Calculated advantage of length

One of the larger functional motifs to be isolated from random sequences is the class I RNA ligase, a 92 nucleotide ribozyme motif isolated from a pool with 220 random positions (i.e. a pool with molecules containing 220 random positions; [9]). After doping this sequence and re-selecting more active variants, the resulting optimized motif was a much more efficient ribozyme than was the 51 nucleotide class III ligase motif, optimized in a similar manner [10]. If a goal of the selection is to isolate a motif as large as the class I ligase (e.g. if a motif this size is considered to be better than smaller motifs as a starting point for developing an efficient catalyst), then the advantage of longer pools is clear — a class I ligase could not have been isolated from a pool with fewer than 92 random nucleotides. If the desired motif could conceivably reside in a shorter pool, then the advantages of longer pools must be weighed-up against the disadvantages. First, we address the advantages that increasing pool length has on the probability of finding a particular RNA motif.

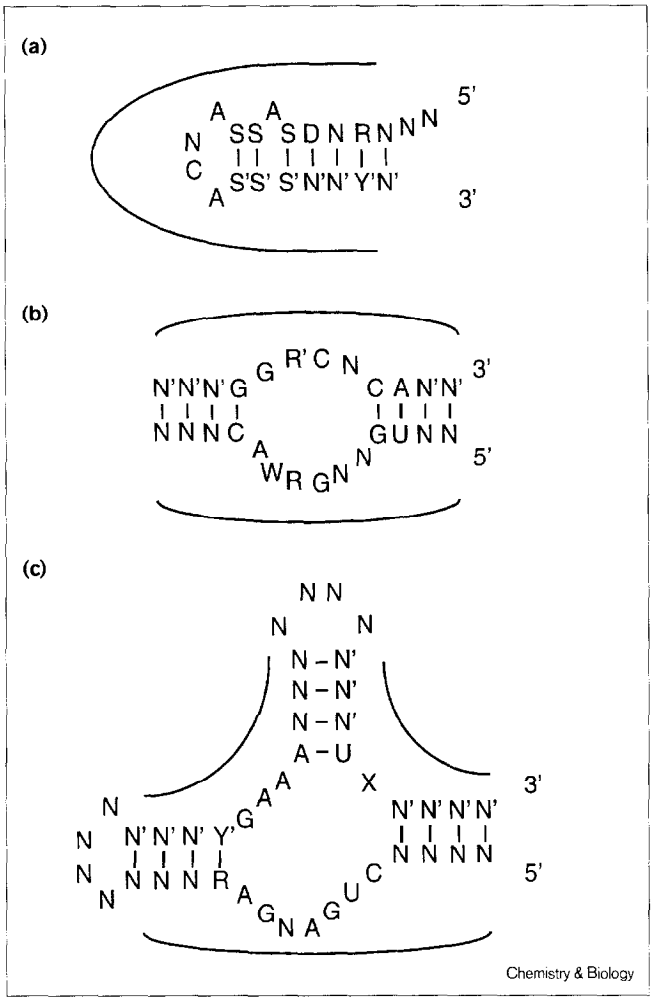
The probability P of finding a particular motif within a random sequence is a function of the minimal size of the motif n, the length of the random sequence l, the modularity m of the motif, and the redundancy r of the motif. For most motifs over a reasonable range of pool lengths this probability is accurately estimated by Equation 1 (see the Materials and methods section for the derivation and limitations of this estimate):

$$P \cong \frac{r}{4^n} \cdot \frac{(l-n+m)!}{m!(l-n)!} \quad (1)$$

We define the modularity as the number of interacting segments that form the motif. For a stem–loop motif, such as the R17 coat protein binding site, $m = 1$; for an internal loop motif, such as the HIV-1 Rev-binding site, $m = 2$; and for a branched internal loop, such as the hammerhead ribozyme, $m = 3$ (Figure 1). The redundancy r specifies the number of different sequence possibilities of length n that meet the criteria of the motif. A requirement for base pairing that can only be satisfied by the four Watson–Crick pairs contributes a factor of four to the redundancy, as does every arbitrary base; bases that are restricted to purines or pyrimidines contribute a factor of two. The redundancy is further modified by symmetries of the structural elements based on the underlying modularity of the motif. For example, the two segments of an internal loop motif (Figure 1b) can be inserted into an RNA strand in two orientations; segments can be connected by a loop either on the 'right' or 'left' side of the motif [11,12]. The redundancy can be quite large and its estimation is highly dependent on the particular details of the motif. For example, the hammerhead ribozyme ($n = 43$) has nine arbitrary nucleotides one core residue plus the two 4-nucleotide loops), 11 determined nucleotides, one purine–pyrimidine Watson–Crick base pair, 10 other Watson–Crick base pairs, and a residue that cannot be a G (Figure 1c). Combining these factors with a cyclic permutation factor of three gives an estimate of $r = 4.95 \times 10^{12}$ [4^9 (arbitrary positions) × 2 (restricted Watson–Crick pair) × 4^{10} (Watson–Crick pairs) × 3 (non-G residue{s}) × 3 (possible segment permutations)].

Clearly, such estimates of redundancy are inaccurate, even in the case of the hammerhead motif, which has undergone

Figure 1

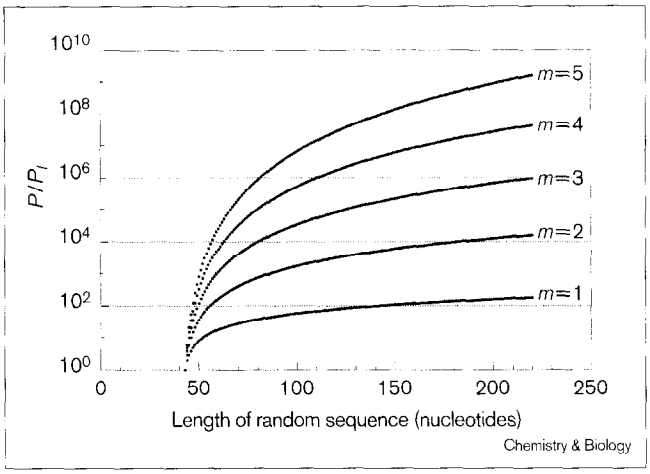


Three RNA motifs. The modular units of the motifs are indicated (solid lines). Base substitution experiments provide estimates for the redundancy of the motif: R, purine; Y, pyrimidine; W, A or U; S, C or G; D, A or G or U; X, A or C or U; N, arbitrary base; prime ('), a pairing requirement. **(a)** The binding site for bacteriophage R17 coat protein, $m = 1$ [16]. **(b)** The binding site of a peptide from HIV-1 Rev protein, an internal loop motif, $m = 2$ [17,18]. When inserted into an RNA strand two permutations are possible, depending on whether the segments are connected at the 'right' or 'left' side of the motif. **(c)** The hammerhead ribozyme, $m = 3$ [19]. Three cyclic permutations are possible because any two of the three stems can be connected by loops.

extensive structure–function analysis. The inaccuracies stem from unknown effects of changes at multiple positions and ignorance of sites where small insertions could be tolerated. Nevertheless, Equation 1 is useful for illustrating the effect of pool length on the probability of finding an RNA motif with a given size, modularity, and redundancy (Figure 2).

Fortunately, when calculating the advantage of one pool length over another for a particular motif, the redundancy need not be known. The advantage A of a pool with

Figure 2



The importance of modularity and pool length in finding a 43 nucleotide motif. Here, the intrinsic probability of finding the motif, $P_I = r/4^n$, has been factored out. The hammerhead ($n = 43$, $m = 3$) is represented by the middle curve. (The intrinsic probability of finding a hammerhead molecule is $P_I \cong 6.4 \times 10^{-14}$.)

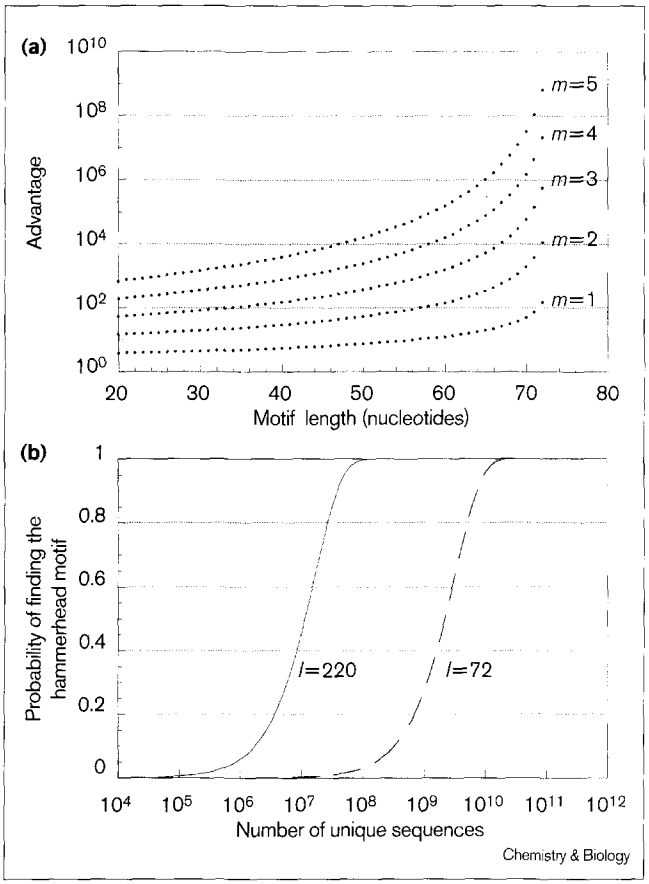
length l_1 over that of a pool with length l_2 is given by the ratio of the probabilities defined in Equation 1:

$$A \cong \frac{P(l_1)}{P(l_2)} \cong \frac{(l_1 - n + m)!\ (l_2 - n)!}{(l_2 - n + m)!\ (l_1 - n)!} \qquad (2)$$

In the experimental section of this paper we examine the inhibition of excess arbitrary sequence by inserting segments with 148 random residues within selected ribozyme sequences. Adding the 148 extra random residues is designed to simulate increasing the length of a random-sequence pool from 72 random positions to 220 random positions. The calculated advantage of such an increase in pool length is illustrated by solving Equation 2 for a range of motif sizes and modularities (Figure 3). For this broad range of possible motifs, going from the short pool to the longer pool significantly increases the chances of finding motifs. This increase is particularly dramatic for motifs with higher modularities or for motif sizes that approach the length of the shorter pool.

Increasing the length of random sequence from 72 to 220 nucleotides increases the probability of finding the hammerhead motif ($n = 43$, $m = 3$) by 200-fold (Figure 3). For a pool with complexity that can be achieved *in vivo* (e.g. 1.5×10^8 different molecules), this approximately three-fold increase in pool length increases the probability of the hammerhead appearing in the pool from unlikely (~0.05) to very likely (~0.9999) as demonstrated by Figure 3b. Because pools for *in vitro* selection typically contain $> 10^{13}$ different molecules, the hammerhead motif is expected in pools of either length. Nevertheless, for these more complex pools, the advantage of length would provide a large

Figure 3



The advantages of increasing pool length, determined by solving Equation 2. **(a)** The advantage a pool with 220 random positions has over a pool with 72 random positions. The advantage is calculated as a function of motif length and modularity (*m*). **(b)** The probability that a pool with a random-sequence length of 220 or 72 nucleotides contains a hammerhead motif; *l*, length.

practical benefit when accessing motifs significantly more rare than the hammerhead motif.

Empirical disadvantage of length

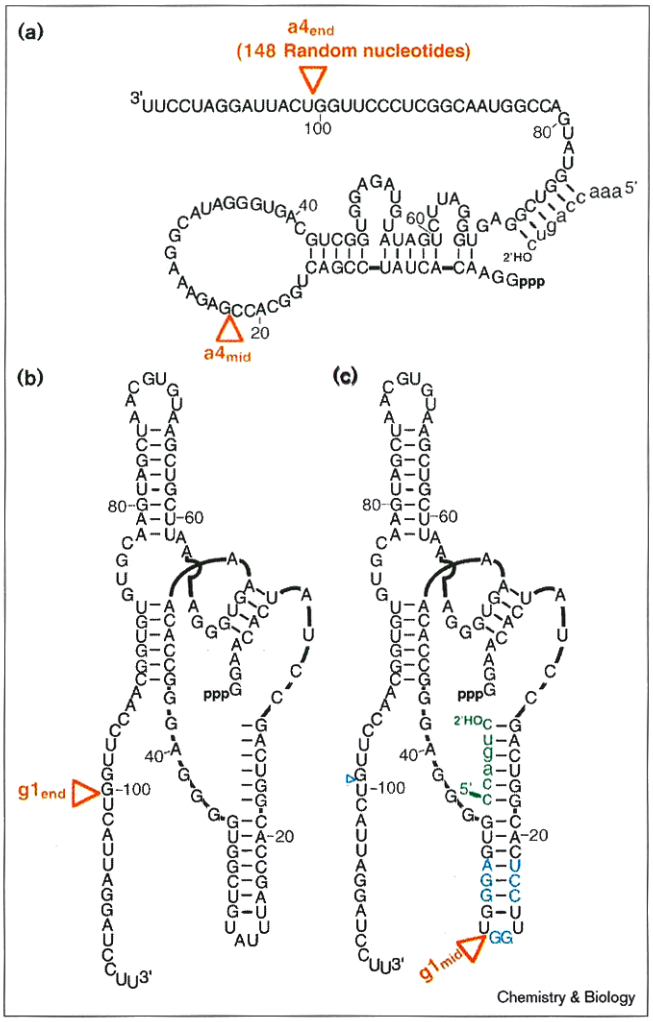
The degree to which functional motifs are more likely to be found in longer pools is clear from the probability arguments of the previous section. But the simple presence of a ribozyme or aptamer motif within an RNA sequence does not guarantee that the molecule will be active. A given RNA sequence can fold in many ways, and many longer RNA molecules have numerous metastable secondary structures. A sequence with the requisite motif can easily be locked in an inactive fold [13,14] and be lost during the first round of selection. Increasing the amount of superfluous sequence in a pool molecule will presumably increase the likelihood that the motif will misfold and be lost. In order to determine whether the greater abundance of motifs in longer pools truly imparts a practical advantage over shorter pools it is important to determine the degree to which this inhibitory effect of excess arbitrary sequence offsets the advantage of excess arbitrary sequence.

In an effort to examine the inhibitory effect of excess arbitrary sequence on RNA functional motifs, we have added excess random sequence to the class II and class III RNA ligase motifs (Figure 4). These two ribozyme motifs had been isolated multiple times in an *in vitro* selection experiment that started with a pool of about 10^{15} different RNAs [6,10]. The versions of the motifs used in this study are simple truncated forms of isolates from the original *in vitro* selection experiment ([6]; isolates a4 and g1, [10]). They promote a self-ligation reaction at 3.4×10^{-2} min^{-1} and 1.1×10^{-3} min^{-1}, respectively — rates typical of newly emergent ribozymes yet sufficiently rapid to permit detection of inhibition over several orders of magnitude. To assess whether location of the excess RNA influences its inhibitory effect, excess random sequence was added at two locations, either within the motif (libraries $a4_{mid}$ and $g1_{mid}$) or flanking the motif (libraries $a4_{end}$ and $g1_{end}$). The libraries were constructed by ligating restricted PCR fragments, followed by amplification and transcription of the ligation product. In order to look at the general effect of excess random sequences, without bias from a particular added sequence, libraries were constructed such that they contained many (> 10^{12}) different arbitrary segments. The extra arbitrary sequences had 148 random positions, constructed by linking together two smaller pools with 72 and 76 random positions.

When comparing catalytic activities of the four ribozyme libraries to those of the parents it was found that the random sequence was usually inhibitory and that the extent of the inhibition varied from no inhibition to 18-fold inhibition, depending on the motif and the location of the arbitrary sequences. For the class II ligase, extra random sequence was more inhibitory at the end of the motif than in the middle of the motif ($a4_{end}$, 5.5-fold inhibition; $a4_{mid}$, no detectable inhibition), whereas the opposite was true for the class III ligase ($g1_{end}$, 1.8-fold inhibition; $g1_{mid}$, 18-fold inhibition). The observed inhibition was primarily intramolecular rather than intermolecular; incubating the parent ribozymes with a ninefold excess of their respective libraries ($a4_{end}$ or $g1_{mid}$) resulted in insignificant (≤ twofold) inhibition of the parent molecules.

We explored whether certain incubation conditions can influence the inhibitory effect of excess random sequences. When selecting new ribozymes from random sequences the first rounds of the selection are often performed using permissive incubation conditions, including large concentrations of divalent cations. Accordingly, our initial assays of parent and pool activities were performed at large (60 mM) Mg^{2+} concentrations. Decreasing Mg^{2+} concentrations to 10 mM lowered the self-ligation rates of both the parent constructs and the libraries, but the effect was small

Figure 4



Ribozymes used to test the effect of additional arbitrary sequence. These ribozymes promote a self-ligation reaction in which the terminal 2′ hydroxyl of the substrate RNA (green) attacks the α phosphate of the ribozyme 5′ triphosphate. The two RNAs become joined by a 2′, 5′ linkage with concomitant release of pyrophosphate. **(a)** The a4 ribozyme. Two libraries were constructed based on the a4 parent ribozyme, a representative of the class II RNA ligase motif [10]. For the $a4_{mid}$ library, random-sequence segments were inserted within the ribozyme motif, between C22 and G23. For the $a4_{end}$ library, random-sequence segments were added at one end of the ribozyme motif, between G100 and U101. **(b)** The g1 ribozyme. To construct the $g1_{end}$ library, random sequence was inserted between G100 and U101 of the g1 parent ribozyme, a representative of the class III RNA ligase motif [10]. **(c)** The g1* ribozyme. To facilitate construction of the $g1_{mid}$ library, g1 was first altered (at positions indicated in blue) so that there would be a *StyI* restriction site at segment C23–G28 rather than at residues 95–100. For the $g1_{mid}$ library, random-sequence was inserted between G28 and U29.

(≤ twofold), with no consistent change in the pattern of inhibition. A denaturation-renaturation step, in which the ribozyme was heated in water to 80°C prior to the addition of salt, was found to be beneficial with or without the extra random sequence, although this effect was also small (≤ twofold). We did not find folding or incubation conditions that significantly altered the pattern or magnitude of inhibition by additional sequences.

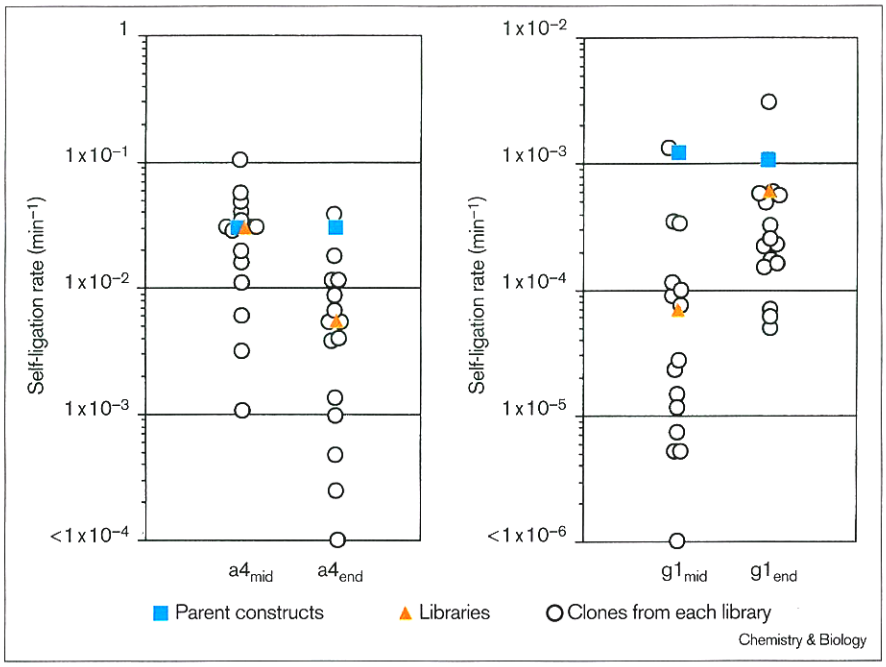
It is important to know how the inhibition by random sequences is distributed among the different members of the library. If some members of the library are completely inactivated by the extra arbitrary sequence, then such sequences would be lost in the first round of selection even though they contain the ribozyme motif. But if nearly all members of the pool retain at least marginal activity then, with very permissive conditions in the early rounds, most representatives of the more active motifs should be retained, despite the inhibitory effect of random sequences. Members of the four libraries were cloned and examined individually. All but two of the sixty clones examined retained some activity, although this activity ranged over two to three orders of magnitude (Figure 5). The median inhibition among the sixty clones was five-fold. For all four libraries, bulk activity of the library closely corresponded to the average activity of clones assayed in isolation, suggesting that the inhibitory effect of random sequence was not due to aggregation sometimes observed during incubation of complex pools of long random sequences [6].

Net advantage of longer pools

We have joined, in two contexts, random-sequence RNA to the class II and class III ligase ribozymes. The resulting four libraries of ligase–random-sequence chimeras were tested for ligation activity. For two pools, addition of random sequence led to moderate (5.5-fold and 18-fold) decreases in the ligation rate, whereas for the other two pools the effect was slight (< twofold). These minimal to modest effects were not altered by changes in incubation and pre-incubation conditions. When considering both the calculated advantage of pool length (Figure 3) and the empirical disadvantage of length (Figure 5) it is clear that rare motifs, particularly highly modular rare motifs, are much more easily accessed using longer pools.

The finding that most of the individuals of the libraries retained at least some activity (Figure 5) suggests that the net advantage of longer pools is even more pronounced for accessing rare ribozyme motifs than it is for accessing rare aptamer motifs. New ribozymes are usually selected based on their ability to modify themselves or a substrate attached to themselves [1]. Because this self-modification reaction need only occur once during the incubation, molecules that are slowed by misfolding can still be captured if they are incubated for a sufficient length of time. In contrast, in a binding selection (SELEX), molecules must have an affinity for the ligand during the brief window of time when bound and unbound RNAs are being separated; molecules that misfold during this time will be lost unless they are quickly exchanging between inactive and active folds.

Figure 5



Self-ligation rates of parent constructs, libraries, and clones from each library, for the a4 and g1 ribozymes.

Although the net advantage of a longer pool in accessing rare motifs is large, such a pool has little advantage when accessing common motifs. Short motifs and motifs of low modularity are not sufficiently more abundant in longer pools to counteract the inhibitory effect of the extra sequence (Figure 3a). Furthermore, some motifs are amply represented in short pools so their even greater abundance in longer pools is of little practical value (e.g. the hammerhead motif, Figure 3b). If such motifs would be acceptable leads for the desired activity, then a longer pool has negligible benefit. In this case, the intrinsic advantages of shorter pools dominate: shorter pools require less effort to construct; motifs within short isolates are easier to find and characterize; and shorter RNA molecules are less likely to be severed by hydrolysis, permitting longer incubation times for the initial round of ribozyme selections. Such considerations may also begin to limit the effectiveness of very long pools (i.e. pools with > 1000 nucleotides of random sequence), even when extremely rare motifs are being pursued.

We have examined the effects of excess arbitrary sequence on two ligase ribozymes. We anticipate that our results will be general to other motifs and activities, but this remains to be confirmed. The class II and class III ligase ribozyme motifs that we examined were isolated from a long pool and thus may have an unusual tolerance for excess arbitrary sequence. It would therefore be of particular interest to learn whether newly emergent ribozymes and aptamers from shorter pools would show the same degree of tolerance for excess random sequences. (Examining natural ribozymes would be less informative because in nature they must function in the context of long transcripts.)

Although it is impossible to know *a priori* the size of the motif that can satisfy a given selection criterion, it can be assumed that the more challenging the catalytic or binding task, the larger and more rare will be the structural solution. One approach for accessing large motifs is to build the motif stepwise, by first selecting for a more modest function or activity and then adding random sequence around the motif and selecting for improved or added function [15]. A second approach is to select the rare motif in a single step from a pool of long sequences and then to optimize by substituting residues within the motif and deleting residues outside the motif [9]. It is still too early to know which approach will be most productive. Likewise, it is unclear which approach dominated during the emergence of new functional motifs in early evolution. Our results indicate that inhibitory effects of excess arbitrary sequence do not seriously compromise the effectiveness of the second approach.

Significance

The technique of *in vitro* selection makes it possible to isolate molecules with interesting biochemical properties from large pools of random RNA or DNA sequences. With the development of this new method, exploration of the catalytic and binding abilities of RNA and DNA is no longer limited to functions found in contemporary biology. Instead, the starting points for such exploration are limited to functions that can be found in pools of

random sequences. Thus, it is important to learn how to design random-sequence pools that provide maximal access to rare activities. A critical, yet previously unexplored, aspect of pool design is the length of the random-sequence molecules. Our work indicates that increasing random-sequence length provides a large practical benefit in accessing rare functional motifs. For example, pools with 200 random positions are much better suited for accessing rare activities than are pools with 70 random positions. The insights gained from our work will be helpful for those attempting to extend further the basic knowledge of RNA's functional repertoire as well as those attempting to isolate RNAs with pharmaceutical or other beneficial uses.

Materials and methods

DNA constructs

DNA templates for a4 and g1 parental ribozymes (Figure 4) were constructed by deleting large internal segments of the original a4 and g1 isolates (GenBank accession numbers U26406 and U26412, respectively [10]). A 161 nucleotide *Ban*I fragment (residues 23–183 of the original isolate) was deleted to generate the a4 ribozyme template; a 160 nucleotide *Sty*I fragment (residues 101–260 of the original g1 isolate) was deleted to generate the g1 ribozyme template. Deletions were made by PCR, using the appropriate primers and plasmid templates [10]. The g1* ribozyme template was prepared in a similar manner except that longer PCR primers were employed which incorporated the desired base substitutions (Figure 4c) in addition to the 160 nucleotide deletion.

To prepare template for the $a4_{mid}$ library, the a4 PCR product was cut at the *Ban*I site (G17–C22, Figure 4a). The restriction fragment containing segment G23–U113 was purified on an agarose gel, then ligated to *Ban*I and *Sty*I restricted PCR fragments containing 72 and 76 random nucleotides, respectively. Fragments and ligation conditions were as described in Figure 2 in [6], except that ligation was done on a much smaller scale. This generated a library of $> 10^{12}$ different double-stranded DNA templates with the following sense strand: TTC*TAATACGACTCACTATA*GGAACACTATCCGACTGGCACC-N_{72}-**CCTTG**-**G**-N_{76}-**GGCACC**GAGAAAGGCATAGGGTGACGTCGGTGGAGATGTATAGTCTTAGGGTGAGGCTGGTATGACCGGTAACGGCTCCCTTGGTCATTAGGATCCTT (T7 promoter is italicized, *Ban*I and *Sty*I sites used in library construction are in bold, N indicates random residue position). A similar strategy produced a template for the $a4_{end}$ library (TTC*TAATACGACTCACTATA*GGAACACTATCCGACTGGCACCGAGAAAGGCATAGGGTGACGTCGGTGGAGATGTATAGTCTTAGGGTGAGGCTGGTATGACCGGTAACGGCTC**CCTTGG**-N_{76}-**GGCACC**-N_{72}-CCTTGGTCATTAGGATCCTT), a template for the $g1_{end}$ library (TTC*TAATACGACTCACTATA*GGAACACTATCCGACTGGCACCGATTTATGTCGGTGGGGAGGGCCACAAAGTGGGAAATTCGTCGAATGTGCAATCGATGAACGTGTGTGGCAA**CCTTGG**-N_{76}-**GGCACC**-N_{72}-CCTTGGTCATTAGGATCCTT) as well as a template for the $g1_{mid}$ library (TTC*TAATACGACTCACTATA*GGAACACTATCCGACTGGCACT**CCTTGG**-N_{76}-**GGCACC**-N_{72}-**CCTTGG**TGGGAGTGGGGAGGGCCACAAAGTGGGAAATTCGTCGAATGTGCAATCGATGAACGTGTGTGGCAACCTTGTCATTAGGATCCTT).

DNAs encoding parent ribozymes and libraries were amplified by PCR and the PCR product was transcribed *in vitro* by T7 RNA polymerase. A small portion of template for each library was cloned (T-Vector kit, Novagen). Plasmid inserts from 15 clones were amplified by PCR and then used as templates for *in vitro* transcription. Templates from clones with low activity in the self-ligation assays were sequenced to ensure that the low activity was not due to a construction artifact; no construction artifacts were found.

Self-ligation assays

Unless otherwise noted, self-ligation reactions were performed at 22°C in 30 mM Tris, pH 7.4, 200 mM KCl, 60 mM $MgCl_2$, 0.6 mM EDTA. Gel-purified ribozyme (1.0 μM, final concentration) was incubated in water at 80°C for 1 min, then allowed to cool at room temperature for 10 min prior to simultaneous addition of salt, buffer, and ^{32}P-labeled DNA–RNA substrate oligonucleotide (5′-dAdAdAdCrCrArGrUrC, 1.0 μM, final concentration). A DNA oligonucleotide 5′-AAGGATCCTAATGACCAAGG (1.0 μM, final concentration), which was complementary to the 3′ primer-binding segment, was also included to mask any inhibitory effect of the 3′ constant region [6,9,10]. Reactions were stopped by adding two volumes of 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 substrate converted to product divided by the duration of the incubation. As seen previously [6,9,10], the rate of self-ligation of both clones and libraries slowed at later time points; reported rates were based on the earliest time point in which the product signal could be accurately measured. When co-incubating the parent ribozymes with the libraries (to explore the possibility of intermolecular inhibition), parent and library ribozymes were mixed prior to the 80°C incubation (0.1 μM final concentration of parent, 0.9 μM final concentration of pool).

Derivation of motif abundance estimate

The calculation divides into two components. The first concerns the intrinsic probability of finding a motif of length n in random sequence n bases long:

$$P_I = \frac{r}{4^n} \tag{3}$$

The intrinsic probability is increased by considering all the ordered ways a motif with modularity m can be inserted into random sequence of length l. There are $(l - n + m)$ locations where the m motif segments can be positioned. When keeping the order of the motif segments fixed, there are:

$$\frac{(l-n+m)!}{m!((l-n+m)-m)!} \tag{4}$$

different ways that the segments can be distributed, resulting in the estimate given by Equation 1.

Although Equations 1 and 2 provide accurate estimates for most RNA motifs over a reasonable range of pool lengths, the equations cannot be applied to motifs with very short segments or to very long pools. This is because the combinatorial term in Equation 1 corresponds to keeping only the first term of a series, which should be written as:

$$P = \sum P(i) - \sum P(i \cap j) + \sum P(i \cap j \cap k) - \ldots \tag{5}$$

where the subtracted terms can be understood from the point of view of a Venn diagram in which intersecting regions of probability have been overcounted. The higher order corrections are negligible provided that:

$$\sum_{i=1}^{m} \frac{(l-n-n_i)}{4^{n_i}} << 1 \tag{6}$$

where n_i are the motif segment lengths. Therefore, when looking at pool molecules of length ≤ 300 nucleotides, Equation 1 is a good approximation if the shortest segment of the motif is ≥ 6 nucleotides.

Acknowledgements

We thank Scott Baskerville, Eric Ekland, Thomas Tuschl, and Kelly Williams for helpful discussions and Wendy Johnston for technical assistance. This research is supported by the NIH.

References

1. Williams, K.P. & Bartel, D.P. (1996). In vitro selection of catalytic RNA. In *Catalytic RNA*. (Eckstein, F. & Lilley, D.M.J., eds), pp. 367-381,

Springer-Verlag: Berlin Heidelberg.
2. Breaker, R.R. (1997). In vitro selection of catalytic polynucleotides. *Chem. Rev.* **97**, 371-390.
3. Gold, L., Polisky, B., Uhlenbeck, O. & Yarus, M. (1995). Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763-797.
4. Osborne, S.E. & Ellington, A.D. (1997). Nucleic acid selection and the challenge of combinatorial chemistry. *Chem. Rev.* **97**, 349-370.
5. Connell, G.J., Illangesekare, M. & Yarus, M. (1993). Three small ribooligonucleotides with specific arginine sites. *Biochemistry* **32**, 5497-5502.
6. Bartel, D.P. & Szostak, J.W. (1993). Isolation of new ribozymes from a large pool of random sequences. *Science* **261**, 1411-1418.
7. Crameri, A. & Stemmer, W.P. (1993). 10^{20}-fold aptamer library amplification without gel purification. *Nucleic Acids Res.* **21**, 4410.
8. Schultes, E., Hraber, P.T. & LaBean, T.H. (1997). Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence. *RNA* **3**, 792-806.
9. Ekland, E.H. & Bartel, D.P. (1995). The secondary structure and sequence optimization of an RNA ligase ribozyme. *Nucleic Acids Res.* **23**, 3231-3238.
10. Ekland, E.H., Szostak, J.W. & Bartel, D.P. (1995). Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**, 364-370.
11. Sassanfar, M. & Szostak, J.W. (1993). An RNA motif that binds ATP. *Nature* **364**, 550-553.
12. Famulok, M. (1994). Molecular recognition of amino acids by RNA-aptamers: an L-citrulline binding RNA motif and its evolution into an L-arginine binder. *J. Am. Chem. Soc.* **116**, 1698-1706.
13. Uhlenbeck, O.C. (1995). Keeping RNA happy. *RNA* 1, 4-6.
14. Herschlag, D. (1995). RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **270**, 20871-20874.
15. Lorsch, J.R. & Szostak, J.W. (1994). In vitro evolution of new ribozymes with polynucleotide kinase activity. *Nature* **371**, 31-36.
16. Schneider, D., Tuerk, C. & Gold, L. (1992). Selection of high affinity RNA ligands to the bacteriophage R17 coat protein. *J. Mol. Biol.* **228**, 862-869.
17. Bartel, D.P., Zapp, M.L., Green, M.R. & Szostak, J.W. (1991). HIV-1 Rev regulation involves recognition of non-Watson-Crick base pairs in viral RNA. *Cell* **67**, 529-536.
18. Battiste, J.L., Mao, H., Rao, N.S., Tan, R., Muhandiram, D.R., Kay, L.E., Frankel, A.D. & Williamson, J.R. (1996). α Helix-RNA major groove recognition in HIV-1 Rev peptide-RRE RNA complex. *Science* **273**, 1547-1551.
19. McKay, D.B. (1996). Three-Dimensional Structure of the Hammerhead Ribozyme. In *Catalytic RNA*. (Eckstein, F. & Lilley, D.M.J., eds), pp. 161-172, Springer-Verlag: Berlin Heidelberg.