Title
Improving the odds: Influence of starting pools on in vitro selection outcomes

Authors
Pobanz, Kelsey
Lupták, Andrej

Publication Date
2016-08-01

DOI
10.1016/j.ymeth.2016.04.021

Peer reviewed

Methods xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Improving the odds: Influence of starting pools on *in vitro* selection outcomes

Kelsey Pobanz [a], Andrej Lupták [a,b,c,*]

[a] Department of Chemistry, University of California, Irvine, CA 92697, USA
[b] Department of Pharmaceutical Sciences, University of California, Irvine, CA 92697, USA
[c] Department of Molecular Biology and Biochemistry, University of California, Irvine, CA 92697, USA

ARTICLE INFO

Article history:
Received 12 March 2016
Received in revised form 16 April 2016
Accepted 18 April 2016
Available online xxxx

Keywords:
SELEX
In vitro selection
Molecular evolution
Aptamer
Ribozyme
Genotype
Phenotype

ABSTRACT

As with any outcome of an evolutionary process, the success of *in vitro* selection experiments depends critically on the starting population. *In vitro* selections isolate functional nucleic acids that fold into specific structures and form unique binding and catalytic sites. The selection outcomes therefore depend on the molecular and structural diversity of the initial pools. In addition, the experiments are strongly influenced by the length of the starting pool. Longer randomized regions support the formation of more complex structures and presumably allow formation of more intricate tertiary interactions, but they also tend to misfold and aggregate, whereas shorter pools are sufficient to yield simpler motifs. Furthermore, introducing a sequence bias that promotes secondary structure formation appears to prejudice the population towards more functional macromolecules. We review the literature on the influence of the starting pools on the predicted and actual outcomes of laboratory evolution experiments.

Contents

1. Introduction

In vitro selection and evolution experiments have changed the way we think about the development of functional macromolecules in a number of fields, ranging from the Origin of Life to diagnostics and therapeutics. The experiments were conceptually anchored in the work of Spiegelman [1] and the discovery of catalytic RNAs [2,3], but became far more feasible through the development of chemical synthesis of DNA, PCR, and RT-PCR in the late 1980s. On a practical level, *in vitro* selections of novel functional DNAs and RNAs were not possible until DNA libraries could be chemically synthesized on a large scale using phosphoramidites and an automated process [4]. Synthesis of long, large-scale DNA pools with random sequences flanked by primer-binding regions meant that highly diverse sequence populations could be interrogated either directly as functional DNAs or

Abbreviations: SELEX, systematic evolution of ligands by exponential enrichment; PCR, polymerase chain reaction; RT, reverse transcription; ATP, adenosine triphosphate; nt, nucleotide; bp, base-pair.

* Corresponding author at: 2141 Natural Sciences 2, University of California, Irvine, Irvine, CA 92697, USA.
E-mail address: aluptak@uci.edu (A. Lupták).

http://dx.doi.org/10.1016/j.ymeth.2016.04.021
1046-2023/

indirectly through transcription to RNAs or further translation to peptides and proteins and other coded polymers in molecular evolution experiments. Once the technology took off, attention returned to the analysis of impact that the starting pools have on the outcomes of these genetic experiments.

Evolution requires the phenotype be linked to its genotype, and in the case of RNA and DNA, functionality (phenotype) and the encoding strand (genotype) are the same molecule; whereas, in proteins, non-amplifiable nucleic acids analogs, and other encoded molecules, they have to be physically linked or co-localized with their coding sequences [5,6]. Typically, an *in vitro* selection starts with a population of DNA sequences containing a random region flanked by fixed sequences that are required for amplification, (reverse-)transcription, and in the case of proteins, translation. RNA and DNA *in vitro* selections have been performed with random regions ranging from 20 to 220 nucleotides (nts) with up to $\sim 10^{16}$ starting diversity [7]. Selection occurs when a population of sequences is required to perform a function, such as catalysis or ligand binding, and the active sequences are physically separated from the inactive sequences. Once these selected sequences are amplified, a new generation of variants is available for continual rounds of selection, leading to an evolved RNA or DNA population that is more efficient and specific for the desired function.

In vitro selections for ligand-binding RNAs, or RNA aptamers, have been extensively used and include targets such as ATP and other adenosine derivatives, guanosine derivatives, amino acids, cofactors and antibiotics [5,8]. Other experiments have identified RNA aptamers for ions [9], small synthetic molecules [10], peptides [11], proteins [12], and even liposomes [13]. On the other hand, the first example of an *in vitro* selected catalytic RNA came in 1993 with the discovery of a ligase ribozymes [14], and one motif, the class I ligase, has been extensively studied and evolved into a polymerase capable of template-directed extension of RNA [15–17] and synthesis of another ribozyme [18]. Although there are no known naturally-occurring DNA enzymes, or deoxyribozymes, *in vitro* selection revealed a Pb^{2+}-dependent deoxyribozyme capable of cleaving an RNA phosphodiester bond in 1994 [19] and since then many other deoxyribozymes have been identified, including DNA and RNA ligases [20,21], DNA kinases [22], adenylases [23], depurinating enzymes [24], a Diels–Alderase [25], and an amidase [26], among others.

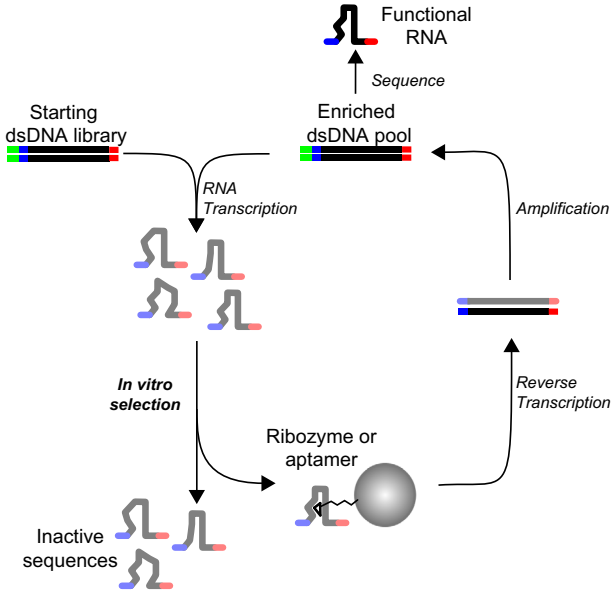

Fig. 1. Overview of an *in vitro* selection. The starting library represents the initial population from which a functional RNA is selected and can originate from entirely synthetic or genomic DNAs. Green represents the RNA polymerase promoter, blue and red are fixed regions for primer binding during amplification. The selection step separates the inactive sequences from active ones, which in this example remain attached to the beads, and are amplified by reverse transcription and PCR.

In general, an *in vitro* selection experiment requires a functional step (e.g. binding or catalysis), a selection/separation step (affinity purification, PAGE band shift, droplet sorting, etc.), and an amplification step (Fig. 1). To identify new functional nucleic acids, these experiments critically depend on a well-designed separation assays. Equally important, but often overlooked, parameter is the initial population pool, from which the functional RNA or DNA is isolated and evolved. Understanding what can reasonably be expected from a starting library and how the design will affect the outcome of a selection and its ability to isolate new and rare functional nucleic acids is an important aspect for the field of molecular evolution. We review the studies that have tackled this challenge, particularly in nucleic acids selections.

1.1. Effect of random region length on selection success

One parameter that is considered at the beginning of an *in vitro* selection is the length of the random region. Whereas shorter lengths will cover all or a large percentage of sequence space, longer regions are thought to allow for more complex structures that may be needed to fold into a functional RNA or DNA. For a random region of N nucleotides, the theoretical diversity is 4^N; that is, a random 28-mer has a theoretical diversity of $\sim 7 \times 10^{16} \approx 0.1$ μmol and a 50-mer could theoretically reach a diversity of $\sim 10^{30} = 2$ *million* mol $\approx 6 \times 10^5$ kg of ssDNA. Clearly, the diversity of a random pool is limited by the DNA synthesis and only a pool with a random region shorter than ~28 nts samples the theoretical diversity extensively. On the other hand, exhaustive sampling near a theoretical limit may not be necessary, because many sequences can fold into the same nucleic acid secondary structure (e.g. CGCGAT:ATCGCG and ACTGAC:GTCAGT both form a 6-bp double-helix, but their sequences are vastly different). For practical reasons, *in vitro* selections have been carried out with pools of up to $\sim 10^{16}$ members.

Intuitively, a large starting diversity may help contribute to the probability of a successful outcome, but the length of pool members is an important aspect that may also contribute to evolving a functional nucleic acid. Successful outcomes have been observed at both extremes, with the class I ligase originally evolved from a highly diverse pool of 220 nts [14], whereas the small isoleucine aptamer was evolved from a pool with a random region of 22 nts [27]. The question of optimal pool length has been addressed in several different ways, computationally and experimentally.

Several groups have approached the question of random region length computationally. Sabeti et al. derived an equation that estimated the probability of finding a motif in a random sequence, which considered the size, modularity and redundancy of the sequence [28]. The probability of finding the hammerhead ribozyme (length = 43 nt) in a 220-nt random region *versus* a 72-nt random region was estimated to increase 200 times. This increased the probability of finding the sequence in a small pool of 1.8×10^8 molecules from 0.05 to 0.999, which was significant if motifs rarer than the hammerhead motif were sought in an *in vitro* selection of 1×10^{15} molecules. Knight and Yarus built on this work by eliminating some of the approximations used to estimate the abundance of functional motifs in random pools and were able to consider active sequences and their probability of folding [29,30]. They used previously discovered functional RNAs, an isoleucine

Please cite this article in press as: K. Pobanz, A. Lupták, Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.04.021

aptamer with two modules and the common hammerhead ribozyme with three modules, and incorporated paired regions that had unspecified sequences but were required for formation of the active structure. In both examples, the probability of finding the motifs solely based on length ranged from 1.7×10^{-8} to 7.9×10^{-6} for the hammerhead motif and 3.5×10^{-9} to 9.0×10^{-8} for the isoleucine aptamer in random regions of 50- to 150-nts, respectively. However, poorer (predicted) folding contributed to lower probabilities for each motif at longer lengths, leading to maximal probabilities of 4.27×10^{-12} to 8.61×10^{-10} for the hammerhead motif and 1.88×10^{-10} to 1.06×10^{-9} for the isoleucine motif for pools ranging from 50- to 150-nt, indicating that the payoff for using a longer random region decreased due to misfolding. Nevertheless, at both lengths, the probabilities were within the realm of the starting diversity of a typical *in vitro* selection. This calculation was particularly relevant, because it incorporated not only sequence requirements for the example motifs but also structural (helical) requirements, which are not conserved in the primary sequence but exhibit strong sequence co-variation [31]. In an independent analysis, a structure-based search for hammerhead ribozymes in a random sequence of 2.2×10^{8} nucleotides revealed three putative ribozymes, none of which was predicted to fold into the correct secondary structure by RNA folding algorithms. For comparison, the same search through a microbial metagenomic dataset of identical size, clearly biased towards highly evolved functional sequences, revealed 13 active hammerhead ribozymes [32].

An experimental study of the effect of random region length on selection outcomes was first performed by ligating arbitrary PCR fragments to class II and class III RNA ligases [14,33] and creating four new libraries with 10^{12} members [28]. The catalytic activities of the new libraries were compared to the originating libraries and the median effect was a 5-fold decrease in activity, which was inconvenient for selecting smaller, simpler motifs, but worth the cost when accessing rarer, complex structures. Similarly, an experimental test of selection success for an RNA-mediated CoA-thioester synthesis using random regions of 30-, 60-, 100-, and 140-nts revealed only sequences from the 30- and 60-nt random region pools, indicating that the abundance of smaller and faster replicating sequences will outcompete longer sequences containing the same active core structure [34,35]. However, these results are weakened by the strong bias of the PCR reaction towards shorter amplicons, limiting the impact of the study. This result led to six parallel *in vitro* selections for the previously discovered isoleucine aptamer using random region lengths of 16-, 22-, 26-, 50-, 70- and 90-nts [36]. This aptamer selection was chosen because previous work indicated that there was a high probability of recurrence of functional motifs in all pool sizes [37–40]. Unexpectedly, the aptamer was 20- to 40-fold more abundant in the 50- and 70-nt sized pools, compared to all other lengths [36]. Since these selections were performed in parallel, the bias due to PCR (or other polymerase-based steps) was not present and the results likely represent true distribution of functional sequences in the starting pools.

In a different approach to mapping of sequence space for functional RNAs, a highly stringent selection for GTP aptamers yielded a number of motifs that ranged in their length, complexity, and target affinity and specificity (further discussed in Section 1.4) [41–43]. One interesting result of the selection was that the stronger-binding aptamers formed structures so long and complex that they had to incorporate the primer-binding sequences into their folds, thus extending the effective length of the pool beyond the random sequence.

In summary, small and simpler motifs isolated by *in vitro* selections generally do not benefit from longer random regions, which may cause misfolding or masking of the active sequence, and the abundance of these motifs may not increase significantly in longer random regions. The probability of isolating larger or more complex functional nucleic acids increases with increasing length, up to a point, and usually outweighs possible inhibitory effects.

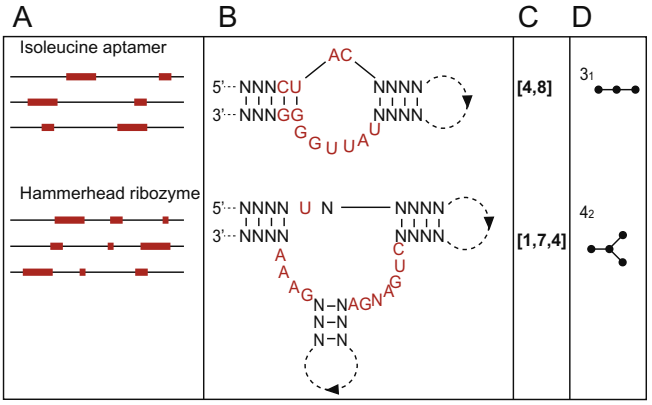

Fig. 2. Simple functional RNAs and their representations. Simplification of secondary structures of isoleucine aptamer and the hammerhead self-cleaving ribozyme led to (A) identification of sequences that performed the same function, (B) definition of the minimal structures for the function of the aptamer and ribozyme with modules (red) represented by (C) matrices [6,10] and [1,7,4], respectively, and (D) representation of the 2D RNA folds with tree graphs containing two to seven vertices with lines as stems and vertices as loops, bulges and junctions.

1.2. Structural diversity and abundance of RNA folds

Random region length of a pool is only one factor that affects the success of an *in vitro* selection. Since nucleic acids only contain four nucleotides, many secondary structure motifs, such as pseudoknots, tetraloops, and uridine turns, occur often and represent energetically efficient solutions when only a few key nucleotides are defined. For example, the GNRA tetraloop motif binds proteins or forms tertiary contacts within the large RNA structure of *Escherichia coli* 16S rRNA which contains 9 GNRA tetraloops [44]. As noted above, the majority of the primary sequence can be easily swapped, while still maintaining the same secondary structure. This degeneracy leads to many primary sequences that have the ability to fold into very similar structures that likely exhibit similar molecular functions and may be designed into an *in vitro* selection pool.

A simplistic approach to understanding the difference between a random primary sequence *versus* a random structure is to consider the probability of each case. In the case of a 200-nt random region, there are over 10^{120} possible sequences and if a starting pool contained 10^{15} molecules, the probability of finding any functional nucleic acid larger than 30 nts nucleotides would be negligible if a functional molecule originated from only a single sequence (i.e. if sequence space were about the same size as the structure space). However, many *in vitro* selections have successfully identified larger ribozymes or aptamers, pointing to the need for only a few conserved bases and structural support. For example, the hammerhead ribozyme only needs approximately 14 defined nucleotides flanked by three helical elements, which leads to a higher probability of isolation than that of hepatitis delta virus ribozymes, which require only about eight specific nucleotides but a more complex secondary structure [32,45]. A simple example of this effect was outlined with 76-nt cloverleaf self-alkylating ribozyme, containing 16 conserved bases. The probability of finding this example, including helices, was estimated to be 1.4×10^{-20} in a 200-nt pool. This probability increased to

6×10^{-13} if extra nucleotides in the pool were allowed within the structure, resulting in approximately 600 molecules in a starting pool of 10^{15} that would match the required structure [46]. If the random region was reduced to 100 nucleotides, only one molecule matching the structural requirements would be found in a starting pool of 10^{15} molecules. The cloverleaf ribozyme was identified by a series of *in vitro* selections that started with a substrate (biotin) affinity selection from a RNA pool of 5×10^{14} members with a 72-nt random region, followed by mutagenesis and selection of self-biolinylating ribozymes [47]. The first selection was dominated by a single biotin-binding sequence, suggesting that the aptamer's frequency in a 72-nt random sequence was about 10^{-14}. This aptamer was mutagenized, a short random sequence was appended to it, and the resulting pool selected for self-alkylating activity. The experiment revealed that the secondary structures of the aptamer and the ribozyme are significantly different (pseudoknot *vs.* cloverleaf), but a critical short sequence in a connecting strand was largely conserved and in the aptamer this sequence contacts the biotin moiety [48]. The experiment thus did not directly reveal whether the self-alkylating cloverleaf was present in the original 72-mer pool, but the drastic change in secondary structure upon reselection suggested that it would likely not be in the starting pool and the biotin aptamer was an obligatory evolutionary intermediate.

Estimating the abundance of functional motifs in random pools relies heavily on considering the modularity of these motifs. Modularity (m) is defined as the number of interacting segments that form the motif [28]. For example, the isoleucine aptamer has a modularity of two, whereas the branched internal loop of a hammerhead ribozyme has a modularity of three (Fig. 2). To determine the probability of finding a sequence within a longer random region, the minimal structures of the isoleucine aptamer [27] and the hammerhead ribozyme [39] were utilized as test cases for a combinatorial analysis [29]. The minimal sequence would only contain the essential nucleotides for functional activity and, in this example, the isoleucine aptamer (m = 2) was represented by a 4- and 8-nt module [4,8] and the hammerhead ribozyme (m = 3) was represented by [1,7,4] (Fig. 2). However, each sequence motif could include additional nucleotides (fixed paired bases) that maintained function, so the upper limit for the isoleucine aptamer and hammerhead ribozyme were [7,11] and [10,14,13], respectively. When considering the probability of finding a sequence motif within a long random region (100 nt), it was calculated that a pool size between 3.3×10^3 and 4.5×10^8 was required to find the isoleucine aptamer – a surprisingly wide range – but once correct secondary structure requirements were considered, the pool size would increase to 4.1×10^9. Conversely, for the hammerhead ribozyme motif, the sequence-based pool size estimation was between 2.2×10^2 and 3.1×10^{17}, and narrowed to 1.6×10^{10} when correct folding was considered [29,49]. One conclusion of this work was that the secondary structure of a functional RNA has to be considered for these approximations to be useful; another one was that motifs with evenly divided, smaller modules were more abundant than asymmetric and larger ones. Working within a 100-nt random sequence, over 1 million unique sequences were calculated for a structure represented by [5,5,5,5] while only ~8000 unique sequences existed for a structure represented by [17,1,1,1] [29]. More broadly, the authors proposed that within an *in vitro* selection experiment with 10^{15} molecules it should be possible to find a motif containing up to 26 nts with m = 1 and a motif with a maximum of 34 nts and m = 4 in a 40-nt and a 100-nt random region, respectively. An example of modularity beyond probability calculations was the minimized nucleotide synthase ribozyme that was determined to have 5 helices and a required ~40 nucleotides for activity [50]. With a modularity of five, the probability of isolating the structure from 228-nt random pool increased from 1 in 10^{22} to 1 in 10^{15}.

Another approach to understand the structural abundance of sequence space utilized graph theory to analyze complete sets of RNA secondary structures for complexity in random pools of 25-, 40-, 60-, 80-, and 100-nts [51,52]. The studies used the Vienna RNA folding package [53] on libraries of 10^4 random RNA sequences and then converted the 2D topologies to tree graphs with bulges, loops, and junctions as the vertices and stems as the lines (Fig. 2D). This approach collapsed a diverse pool of sequences into easily categorized shapes without consideration for detailed base-pair information or stem and loop sizes, allowing analysis of an entire random pool instead of a smaller sample size. More than 90% of folded structures were found to be simple topologies such as stem-loop motifs. Structural complexity, as defined by vertex number, increased with length of the pool and a general relation was put forth revealing that a pool of length L would be most abundant with L/20 stems. For example, a 100-nt pool would have the highest frequency of 5-stemmed structures and their results showed that 40% of the 100-nt pool had 6 vertices. While the Vienna folding package is unable to predict more complicated tertiary structures such as pseudoknots, these results were insensitive to folding stability parameters or the pool size. An application of this method to the three-stem class V GTP aptamer suggested that the aptamer had the highest abundance in a 60-nt random RNA pool, the same size that had been used to discover the motif *in vitro* [41].

1.3. Influence of structures on selection outcomes

Computational efforts to understand the abundance of RNA folds rely on the assumption that the population of more rare and complex RNA secondary structures will exhibit higher or new activities. Experimental validation of this assumption has been presented using information theory to connect structural diversity to activity by studying a set of eleven RNA aptamers that bound GTP. Many aptamers contain a stem-loop that contributes to

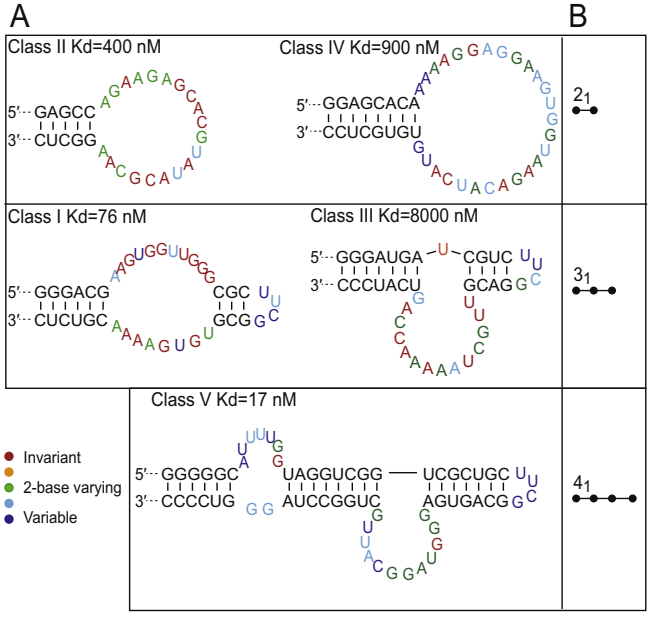

Fig. 3. GTP aptamer sequences and secondary structures. (A) Secondary structures of aptamers optimized for GTP binding and information content for each position. (B) Tree graph representations of the aptamers.

the stability of the structure [8,54–56]. For the GTP aptamer selection, a library was designed to contain a stem-tetraloop of specific sequence flanked by random regions and mixed with an unstructured (completely random) library in a 1:1 ratio [43]. High affinity GTP RNA aptamers were evolved with dissociation constants ranging from micro- to nanomolar and several unique aptamer families were isolated with three of the highest affinity GTP aptamers originating from the structured pool (Fig. 3). Out of 11 identified class I GTP aptamers, 7 originated from the structured pool (easily recognized by its sequence). In this case, it was shown that the advantage of introducing structural complexity in the form of a stem-loop to a random pool outweighed the possible disadvantage of eliminating possible registers within the random region where the GTP recognition loop could occur. The sequence of the introduced stem-loop did not seem to affect the outcome of the selection because the sequence of the loop apparently drifted, picking up mutations during the selection and re-selection, whereas the stem in most cases maintained non-biased base-pair covariation in the stem sequence [41].

The discovery of so many GTP aptamer families afforded an analysis of their structural and informational complexity with respect to the selection criterion, GTP affinity (actually off-rate), and starting pool abundance. An optimized sequence of each aptamer family was mutagenized up to 21% per position and reselected for GTP affinity to further optimize the motifs [41]. The minimal information to define the resulting sequences was used to compute their information complexity and the dissociation constants were determined by binding assays. Dissociation constants for the set of eleven RNA aptamers ranged from 8 μM to 9 nM and seven of the eleven RNA aptamers contained a designed stem-tetraloop from the original *in vitro* selection [43]. The relationship observed between activity and complexity showed that every 10-fold improvement in binding was ~1000× less frequent in a random pool. This result also held true in the case of two unique ligase ribozymes [33] and it was speculated that this relationship would hold as long as additional complexity contributed to stability of the overall RNA fold. As expected, the stronger binders were also structurally more complex, requiring more stems. Interestingly, the information content analysis revealed that the strongest binders utilized the primer binding regions designed into the starting pools to build the more complex structures (effectively extending the pool's length), and were statistically highly unlikely to be found in the starting pool population. This result suggested that if the selection pressure is strong enough, it may be possible to isolate highly unlikely functions from a given starting diversity but also that the motif is not likely to emerge from a second, independent trial (Fig. 4).

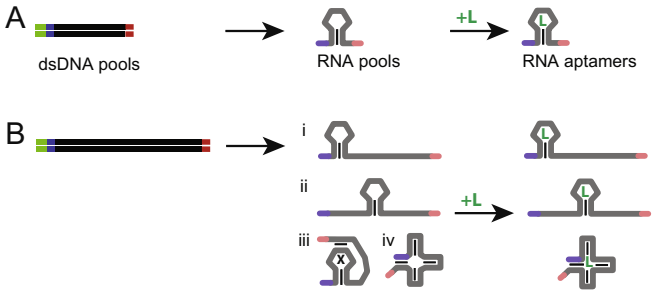

Fig. 4. Correlation of pool design with outcomes. The work described in this review is summarized for simple ligand (L)-binding motifs selected from (A) a short random pool or (B) longer pool that may conatain the simple motif at different positions (i and ii) along the sequence, but the same motif may be inhibited by the additional sequence present in the long RNA (iii; inhibited binding site is marked by X), preventing ligand binding. A longer pool can also form a more complex structure, perhaps incorporating the primer-binding sequence into the structure (iv; blue segment), with higher affinity for the ligand. Black lines within grey RNAs denote base-pairing interactions.

An *in vitro* selection of a biotin aptamer and self-alkylating ribozyme discussed previously provided an interesting example of a structural novelty that arises under selection pressure [47]. The biotin aptamer, selected from a random pool, formed a pseudoknot, whereas the ribozyme selected from its sequence was a cloverleaf. Another interesting example of evolution of the structures of functional RNAs involved the evolution of a three-way junction of an aminoacylating ribozyme [57] into a kinase ribozyme, switching the substrate from AMP-phenylalanine to GTP (γS), thus catalyzing reactions at different atoms (P *vs* C) and likely proceeding *via* different transition states (trigonal bipyramidal *vs* tetrahedral) [58]. The aminoacylating ribozyme was mutagenized to yield $\sim 4 \times 10^{14}$ sequences within about 12 substitutions of the parent sequence and a kinase selection was performed. Surprisingly, the new ribozymes used different sites for the covalent modification and seemed to fold into a wide variety of motifs. Two of the kinase ribozymes were analyzed further, revealing that they formed a pseudoknot and a four-way junction with a kissing-loop, representing both a simpler and more complex structures. The simpler motif did not preserve any base-pairs from the parent ribozyme secondary structure, whereas the more complex one retained 9 bps and created 23 new ones. The new ribozymes thus escaped the original fold and the implication for pool design is that a single starting fold is likely to be too constraining to promote formation of a variety of functional RNAs. On the other hand a partially structured pool may be enriched for active RNAs.

1.4. Evolving more complex structures

In vitro selections tend to find the simplest solutions, which have been supported repeatedly experimentally, for example for the hammerhead ribozymes, adenosine aptamers, and RNA-cleaving DNAzymes [26,39,59–65]. More complex structures may exhibit better activities, but if they are not magnitudes more efficient than the more abundant and simpler motif, they may not be discovered in the initial rounds of *in vitro* selection. In general, smaller motifs are observed in higher abundance even if more complex structures are more efficient and, consequently, selections can be further optimized by deletion, mutation, and recombination experiments. Other strategies, such as biasing the nucleotide composition of the starting pool, have been proposed to increase the abundance of rare secondary structures and may help to increase the chance of finding highly active functional nucleic acids.

Many discovered functional nucleic acids have exhibited compositional biases and tend to be purine-biased [66]. This idea was examined as a possibility to increase the initial structural complexity within a random pool computationally. For the case of the aforementioned isoleucine aptamer and hammerhead ribozyme, the composition that would lead to the optimal sequence abundance and folding, assuming a random region of 100 nucleotides, would be 15% A, 25% C, 35% G, and 25% U; and 35% A, 10% C, 25% G, and 30% U, respectively [49]. The probability of finding the two functional RNAs within a biased pool over an unbiased random region increased 3.5- and 2.3-fold, respectively. While the optimal composition for each example is vastly different, a composition that maximized the probability of finding both motifs was calculated to be 20% A, 15% C, 40% G, and 25% U, and required 6.23×10^9 molecules for 99% probability of occurrence of both motifs, a factor of 10 smaller than an unbiased pool. Although the sample size was limited, compositional changes in the initial pool of an *in vitro* selection may assist in favoring new and rare functional structures. A computational analysis using graph theory partially supported this result by showing that 20% A, U and 30% G, C increased the proportion of higher-vertex structures for a 40-nt

pool but structural abundances for a 100-nt pool were not altered significantly [52].

An experimental approach to measure the effect of compositional bias on *in vitro* selection outcomes confirmed that libraries patterned to increase secondary structure yield more functional RNAs [67]. Ruff et al. designed a pattern of purines and pyrimidines that increased the likelihood of forming stem-loop structures within otherwise random sequences. In three head-to-head *in vitro* selections with an unpatterned, random library of the same length, the patterned library won two times, yielding protein-binding aptamers of higher affinities. The two libraries performed equally well in the third selection. Interestingly, the structure-promoting patterned library required low-level doping with random nucleotides to yield one of the aptamers, suggesting that over-engineering the starting pools may be detrimental to success of the selection. Nonetheless, the structure-promoting patterned pool appears to be one advance that can be readily implemented in selections for rare (complex or energetically expensive) functions.

1.5. Modular pools

Beyond sequence bias, increased modularity in a random region has also been suggested to lead to higher abundances of more complex and rare structures. A modular evolution model has been introduced as a strategy to increase structural complexity [68]. To analyze the likelihood of the combination of smaller RNA modules to form a more complex functional nucleic acid, computational simulations of two independently evolving pools of short, random sequences (n = 35) revealed several advantages of modular evolution. These advantages included a finding that (1) the selection was equally efficient with 35-nt modules at double the mutation rates of those allowed for 70-nt sequences, (2) the time to select smaller motifs was significantly shorter, and (3) smaller pools sizes were required. Applied to an *in vitro* selection, small motifs could be quickly evolved using high mutation rates, exploring most of sequence space, and the resulting modules could be ligated and evolved further for more complex functions. Experimentally, modular evolution has been employed by (1) appending random region sequences to previously discovered ribozymes [16,18,69], (2) incorporating an aptamer motif for a substrate (ATP) within a random region for a kinase ribozyme selection [70], (3) fusing functional motifs to form allosteric ribozymes [71–73], (4) selecting a functional RNA that exhibits both ligation and RNA cleavage reactions [74], (5) including separate domains that allow ligand binding and subsequent cleavage [75], and (6) designing oligonucleotide base-pairing domain and selecting for ligand-induced conformation-switching aptamers [37,76].

Accessing rare secondary structures for new activities may be important for isolating new functional nucleic acids. Increasing the abundance of complex secondary structures in an initial random pool by installing a structured loop or changing the compositional nature are simple measures that could lead to new functional isolates. Combination of independently selected small modules could be another powerful tool for isolating complex and rare functional nucleic acids containing more than one active domain.

2. Conclusion

Design of a nucleic acid library is an important consideration for a successful *in vitro* selection, especially in search for new and highly active functional macromolecules. Increasing the starting diversity of a pool will increase the sampling of sequence space, but most practical considerations limit the number of library members to $10^{15} - 10^{16}$; therefore, other strategies must be employed to more readily access rare structures that have higher activity.

Generally, longer random regions are more prone to misfolding and aggregation, potentially inhibiting expected functional activities by masking the key structural motif or not allowing it to form. This effect is most pronounced for smaller, simpler motifs, whereas the probability of finding more complex and rare structures in long random regions increases significantly and balances the possible cost of misfolding (Fig. 4). What is needed, though, is a more quantitative and direct experimental testing of random and designed pools. Aggregation in particular is a behavior of random pools that has not been, to our knowledge, studied explicitly and it likely affects selection outcomes. Furthermore, methods sensitive enough to detect the activity of single molecules in highly diverse pools may reveal true distribution of functional molecules in these pools.

Since nucleic acids only contain four nucleotides, many secondary structure motifs such as pseudoknots, tetraloops, and uridine turns occur often. The majority of the primary sequence can be easily exchanged while still maintaining the same secondary structure. This degeneracy leads to many primary sequences that have the ability to fold into similar structures that may exhibit similar activities as well. Whereas sequence space is usually sparsely sampled, the structural diversity may be within the practical limits of an *in vitro* selection, at least in terms of basic secondary structure folds. Small, evenly sized modular motifs were shown to be the most abundant and increasing complexity of secondary structures correlated with increasing vertices in graph theory, pointing to longer random regions for unique structures that might reveal new activities.

In vitro selections tend to find the simplest solutions, which has been supported repeatedly experimentally. More complex structures may exhibit better activities but if they are not magnitudes more efficient than the more abundant and simpler motif, they may not be discovered in the initial rounds of *in vitro* selection. Strategies such as altering nucleotide composition of the random region, introducing stable structures into the initial random pool, and evolving small motifs that can be combined into more complex functional nucleic acids have been shown to increase the abundance of rare secondary structures that may be more efficient as well.

Ultimately, structural diversity affects the ability of the pool to evolve a new function more than sequence diversity. Biasing the starting population in an *in vitro* selection towards more structured sequences may be one strategy to increase the success rate of molecular evolution experiments. We expect that a combination of structurally biased pools, high-throughput analysis of early rounds of selections [77], and expansion of the chemical repertoire of the selected population will lead to the discovery of new and more active functional macromolecules.

Acknowledgements

Support of our research by the Pew Charitable Trusts, NIH EUREKA Program (R01GM094929), NSF (MCB-1330606), and the John Templeton Foundation/Foundation for Applied Molecular Evolution is gratefully acknowledged.

References

[1] G.F. Joyce, Forty years of in vitro evolution, Angew. Chem. Int. Ed. Engl. 46 (34) (2007) 6420–6436.
[2] K. Kruger et al., Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena, Cell 31 (1) (1982) 147–157.
[3] C. Guerrier-Takada et al., The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme, Cell 35 (1983) 849–857.

[4] L.J. McBride, M.H. Caruthers, Nucleotide chemistry. 10. An investigation of several deoxynucleoside phosphoramidites useful for synthesizing deoxyoligonucleotides, Tetrahedron Lett. 24 (3) (1983) 245–248.
[5] D.S. Wilson, J.W. Szostak, In vitro selection of functional nucleic acids, Ann. Rev. Biochem. 68 (1999) 611–647.
[6] J.J. Agresti et al., Selection of ribozymes that catalyse multiple-turnover Diels-Alder cycloadditions by using in vitro compartmentalization, Proc. Nat. Acad. Sci. U.S.A. 102 (45) (2005) 16170–16175.
[7] A. Peracchi, DNA catalysis: potential, limitations, open questions, ChemBioChem 6 (8) (2005) 1316–1322.
[8] M. Sassanfar, J.W. Szostak, An RNA motif that binds ATP, Nature 364 (6437) (1993) 550–553.
[9] J. Ciesiolka, M. Yarus, Small RNA-divalent domains, RNA 2 (8) (1996) 785–793.
[10] A.D. Ellington, J.W. Szostak, In vitro selection of RNA molecules that bind specific ligands, Nature 346 (6287) (1990) 818–822.
[11] D. Nieuwlandt, M. Wecker, L. Gold, In vitro selection of RNA ligands to substance P, Biochemistry 34 (16) (1995) 5651–5659.
[12] C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, Science 249 (4968) (1990) 505–510.
[13] A. Vlassov, A. Khvorova, M. Yarus, Binding and disruption of phospholipid bilayers by supramolecular RNA complexes, Proc. Nat. Acad. Sci. U.S.A. 98 (14) (2001) 7706–7711.
[14] D.P. Bartel, J.W. Szostak, Isolation of new ribozymes from a large pool of random sequences, Science 261 (5127) (1993) 1411–1418.
[15] E.H. Ekland, D.P. Bartel, RNA-catalysed RNA polymerization using nucleoside triphosphates, Nature 382 (6589) (1996) 373–376.
[16] W.K. Johnston et al., RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension, Science 292 (5520) (2001) 1319–1325.
[17] L.L. Martin, P.J. Unrau, U.F. Muller, RNA synthesis by in vitro selected ribozymes for recreating an RNA world, Life (Basel) 5 (1) (2015) 247–268.
[18] A. Wochner et al., Ribozyme-catalyzed transcription of an active ribozyme, Science 332 (6026) (2011) 209–212.
[19] R.R. Breaker, G.F. Joyce, A DNA enzyme that cleaves RNA, Chem. Biol. 1 (4) (1994) 223–229.
[20] W.E. Purtha et al., General deoxyribozyme-catalyzed synthesis of native 3'-5' RNA linkages, J. Am. Chem. Soc. 127 (38) (2005) 13124–13125.
[21] B. Cuenoud, J.W. Szostak, A DNA metalloenzyme with DNA ligase activity, Nature 375 (6532) (1995) 611–614.
[22] W. Wang, L.P. Billen, Y. Li, Sequence diversity, metal specificity, and catalytic proficiency of metal-dependent phosphorylating DNA enzymes, Chem. Biol. 9 (4) (2002) 507–517.
[23] Y. Li, Y. Liu, R.R. Breaker, Capping DNA with DNA, Biochemistry 39 (11) (2000) 3106–3114.
[24] T.L. Sheppard, P. Ordoukhanian, G.F. Joyce, A DNA enzyme with N-glycosylase activity, Proc. Nat. Acad. Sci. U.S.A. 97 (14) (2000). 7802-7.
[25] M. Chandra, S.K. Silverman, DNA and RNA can be equally efficient catalysts for carbon-carbon bond formation, J. Am. Chem. Soc. 130 (10) (2008) 2936–2937.
[26] C. Zhou et al., DNA-catalyzed amide hydrolysis, J. Am. Chem. Soc. 138 (7) (2016) 2106–2109.
[27] C. Lozupone, S. Changayil, I. Majerfeld, M. Yarus, Selection of the simplest RNA that binds isoleucine, RNA 9 (2003) 1315–1322.
[28] P.C. Sabeti, P.J. Unrau, D.P. Bartel, Accessing rare activities from random RNA sequences: the importance of the length of molecules in the starting pool, Chem. Biol. 4 (10) (1997) 767–774.
[29] R. Knight, M. Yarus, Finding specific RNA motifs: function in a zeptomole world?, RNA 9 (2) (2003) 218–230
[30] M. Yarus, R.D. Knight, The Genetic Code and Origin of Life, Landes Bioscience, Georgetown, TX, 2004.
[31] N.J. Riccitelli, A. Lupták, Computational discovery of folded RNA domains in genomes and in vitro selected libraries, Methods 52 (2) (2010) 133–140.
[32] R.M. Jimenez, E. Delwart, A. Lupták, Structure-based search reveals hammerhead ribozymes in the human microbiome, J. Biol. Chem. 286 (10) (2011) 7737–7743.
[33] E.H. Ekland, J.W. Szostak, D.P. Bartel, Structurally complex and highly active RNA ligase derived from random RNA sequences, Science 268 (1995) 364–370.
[34] Z.J. Huang et al., A simple and sensitive enzyme-mediated assay of biotin, Biotechniques 13 (4) (1992) 543–546.
[35] T.M. Coleman, F. Huang, RNA-catalyzed thioester synthesis, Chem. Biol. 9 (2002) 1227–1236.
[36] M. Legiewicz et al., Size, constant sequences, and optimal selection, RNA 11 (11) (2005) 1701–1709.
[37] R. Nutiu, Y. Li, In vitro selection of structure-switching signaling aptamers, Angew. Chem. Int. Ed. Engl. 44 (2005) 1061–1065.
[38] M.M. Hanczyc, R.L. Dorit, Replicability and recurrence in the experimental evolution of a group I ribozyme, Mol. Biol. Evol. 17 (2000) 1050–1060.
[39] K. Salehi-Ashtiani, J.W. Szostak, In vitro evolution suggests multiple origins for the hammerhead ribozyme, Nature 414 (2001) 82–84.
[40] R.P. Cruz, J.B. Withers, Y. Li, Dinucleotide junction cleavage versatility of 8–17 deoxyribozyme, Chem. Biol. 11 (2004) 57–67.
[41] J.M. Carothers et al., Informational complexity and functional activity of RNA structures, J. Am. Chem. Soc. 126 (16) (2004) 5130–5137.
[42] J.M. Carothers, S.C. Oestreich, J.W. Szostak, Aptamers selected for higher-affinity binding are not more specific for the target ligand, J. Am. Chem. Soc. 128 (24) (2006) 7929–7937.
[43] J.H. Davis, J.W. Szostak, Isolation of high-affinity GTP aptamers from partially structured RNA libraries, Proc. Nat. Acad. Sci. U.S.A. 99 (18) (2002) 11616–11621.
[44] C.R. Woese, S. Winker, R.R. Gutell, Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops", Proc. Nat. Acad. Sci. U.S.A. 87 (1990). 8476–71.
[45] D.J. Ruminski et al., Processing and translation initiation of non-long terminal repeat retrotransposons by hepatitis delta virus (HDV)-like self-cleaving ribozymes, J. Biol. Chem. 286 (48) (2011) 41286–41295.
[46] A.B. Wedel, Fishing the best pool for novel ribozymes, Trends Biotechnol. 14 (12) (1996) 459–465.
[47] C. Wilson, J.W. Szostak, In vitro evolution of a self-alkylating ribozyme, Nature 374 (6525) (1995) 777–782.
[48] J. Nix, D. Sussman, C. Wilson, The 1.3 A crystal structure of a biotin-binding pseudoknot and the basis for RNA molecular recognition, J. Mol. Biol. 296 (5) (2000) 1235–1244.
[49] R. Knight et al., Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids, Nucleic Acids Res. 33 (18) (2005) 5924–5935.
[50] K.E. Chapple, D.P. Bartel, P.J. Unrau, Combinatorial minimization and secondary structure determination of a nucleotide synthase ribozyme, RNA 9 (10) (2003) 1208–1220.
[51] W. Fontana et al., Statistics of RNA secondary structures, Biopolymers 33 (9) (1993) 1389–1404.
[52] J. Gevertz, H.H. Gan, T. Schilick, In vitro RNA random pools are not structurally diverse: a computational analysis, RNA 11 (6) (2005) 853–863.
[53] I.L. Hofacker, Vienna RNA secondary structure server, Nucleic Acids Res. 31 (13) (2003) 3429–3431.
[54] R.D. Jenison et al., High-resolution molecular discrimination by RNA, Science 263 (5152) (1994) 1425–1429.
[55] A. Geiger et al., RNA aptamers that bind L-arginine with sub-micromolar dissociation constants and high enantioselectivity, Nucleic Acids Res. 24 (6) (1996) 1029–1036.
[56] P. Fan et al., Molecular recognition in the FMN-RNA aptamer complex, J. Mol. Biol. 258 (3) (1996) 480–500.
[57] M. Illangasekare et al., Aminoacyl-RNA synthesis catalyzed by an RNA, Science 267 (5198) (1995) 643–647.
[58] E.A. Curtis, D.P. Bartel, New catalytic structures from an existing ribozyme, Nat. Struct. Mol. Biol. 12 (11) (2005) 994–1000.
[59] P. Burgstaller, M. Famulok, Isolation of RNA aptamers for biological cofactors by in-vitro selection, Angew. Chem.-Int. Ed. Engl. 33 (10) (1994) 1084–1087.
[60] D. Saran, J. Frank, D.H. Burke, The tyranny of adenosine recognition among RNA aptamers to coenzyme A, BMC Evol. Biol. 3 (2003) 26.
[61] M.M. Vu et al., Convergent evolution of adenosine aptamers spanning bacterial, human, and random sequences revealed by structure-based bioinformatics and genomic SELEX, Chem. Biol. 19 (10) (2012) 1247–1254.
[62] A. Peracchi, Preferential activation of the 8–17 deoxyribozyme by Ca(2+) ions. Evidence for the identity of 8–17 with the catalytic domain of the Mg5 deoxyribozyme, J. Biol. Chem. 275 (16) (2000) 11693–11697.
[63] K. Schlosser, Y. Li, Tracing sequence diversity change of RNA-cleaving deoxyribozymes under increasing selection pressure during in vitro selection, Biochemistry 43 (30) (2004) 9695–9707.
[64] S.W. Santoro, G.F. Joyce, A general purpose RNA-cleaving DNA enzyme, Proc. Nat. Acad. Sci. U.S.A. 94 (1997) 4266–4462.
[65] S.F. Torabi, Y. Lu, Identification of the same Na(+)-specific DNAzyme motif from two in vitro selections under different conditions, J. Mol. Evol. 81 (5–6) (2015) 225–234.
[66] E. Schultes, P.T. Hraber, T.H. LaBean, Global similarities in nucleotide base composition among disparate functional classes of single-stranded RNA imply adaptive evolutionary convergence, RNA 3 (7) (1997) 792–806.
[67] K.M. Ruff, T.M. Snyder, D.R. Liu, Enhanced functional potential of nucleic acid aptamer libraries patterned to increase secondary structure, J. Am. Chem. Soc. 132 (27) (2010) 9453–9464.
[68] S.C. Manrubia, C. Briones, Modular evolution and increase of functional complexity in replicating RNA molecules, RNA 13 (1) (2007) 97–107.
[69] L. Jaeger, M.C. Wright, G.F. Joyce, A complex ligase ribozyme evolved in vitro from a group I ribozyme domain, Proc. Nat. Acad. Sci. U.S.A. 96 (26) (1999) 14712–14717.
[70] J.R. Lorsch, J.W. Szostak, In vitro evolution of new ribozymes with polynucleotide kinase activity, Nature 371 (6492) (1994) 31–36.
[71] J. Tang, R.R. Breaker, Rational design of allosteric ribozymes, Chem. Biol. 4 (6) (1997) 453–459.
[72] Y. Komatsu et al., In vitro selection of hairpin ribozymes activated with short oligonucleotides, Biochemistry 41 (29) (2002) 9090–9098.
[73] J.M. Carothers et al., Model-driven engineering of RNA devices to quantitatively program gene expression, Science 334 (6063) (2011) 1716–1719.
[74] R.M. Kumar, G.F. Joyce, A modular, bifunctional RNA that integrates itself into a target RNA, Proc. Nat. Acad. Sci. 100 (17) (2003) 9738–9743.
[75] C. Romero-Lopez et al., Interfering with hepatitis C virus IRES activity using RNA molecules identified by a novel in vitro selection method, Biol. Chem. 386 (2) (2005) 183–190.
[76] K.A. Yang et al., Recognition and sensing of low-epitope targets via ternary complexes with oligonucleotides and synthetic receptors, Nat. Chem. 6 (11) (2014) 1003–1008.
[77] A. Long et al., Elucidating the molecular architecture of adaptation via evolve and resequence experiments, Nat. Rev. Genet. 16 (10) (2015) 567–582.

Please cite this article in press as: K. Pobanz, A. Lupták, Methods (2016), http://dx.doi.org/10.1016/j.ymeth.2016.04.021