*Published online 6 May 2010* *Nucleic Acids Research, 2010, Vol. 38, No. 13* ***e139***
*doi:10.1093/nar/gkq282*

# Computational generation and screening of RNA motifs in large nucleotide sequence pools

**Namhee Kim[1], Joseph A. Izzo[1], Shereef Elmetwaly[1], Hin Hark Gan[1] and Tamar Schlick[1,2,*]**

[1]Department of Chemistry, New York University, 100 Washington Square East, New York, NY 10003 and [2]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10021, USA

Received January 7, 2010; Revised April 2, 2010; Accepted April 4, 2010

## ABSTRACT

**Although identification of active motifs in large random sequence pools is central to RNA *in vitro* selection, no systematic computational equivalent of this process has yet been developed. We develop a computational approach that combines target pool generation, motif scanning and motif screening using secondary structure analysis for applications to $10^{12}$–$10^{14}$-sequence pools; large pool sizes are made possible using program redesign and supercomputing resources. We use the new protocol to search for aptamer and ribozyme motifs in pools up to experimental pool size ($10^{14}$ sequences). We show that motif scanning, structure matching and flanking sequence analysis, respectively, reduce the initial sequence pool by 6–8, 1–2 and 1 orders of magnitude, consistent with the rare occurrence of active motifs in random pools. The final yields match the theoretical yields from probability theory for simple motifs and overestimate experimental yields, which constitute lower bounds, for aptamers because screening analyses beyond secondary structure information are not considered systematically. We also show that designed pools using our nucleotide transition probability matrices can produce higher yields for RNA ligase motifs than random pools. Our methods for generating, analyzing and designing large pools can help improve RNA design via simulation of aspects of *in vitro* selection.**

## INTRODUCTION

RNA *in vitro* selection is a sensitive experimental technology for detecting rare active motifs in random pools of up to $10^{16}$ sequences (1–3). The versatility of the method has led to numerous nucleic acid molecules binding targets (aptamers) as diverse as organic molecules, antibiotics, proteins and whole viruses (3,4). Importantly, *in vitro* selection experiments have enabled discovery of new classes of RNA enzymes (ribozymes) and have ramifications for biomolecular engineering, including the design of allosteric ribozymes and aptamer-based biosensors (5–7), and aptamers capable of inhibiting protein function for functional genomics (8,9). Many aptamers and ribozymes have also been developed for therapeutic applications (10,11), such as aptamers inhibiting the TAR RNA element of HIV-1 (12) and the human vascular endothelial growth factor in cancer (13). See examples in Table 1.

*In vitro* selection of RNAs involves three essential steps: synthesize a large sequence pool, screen the sequence pool for aptamers or ribozymes and verify active RNA candidates using functional assays. Initially, a DNA-pool is chemically synthesized, amplified by PCR and then transcribed to generate the RNA pool. Ligand-binding RNAs are detected using, for example, column chromatography, where target ligands are bound. The ligand-bound RNAs are selected and then reverse-transcribed and amplified by PCR for further selection rounds (3). Ribozymes are selected using various strategies, including attaching chemical tags to RNAs (3). The entire pool generation and selection process can be laborious, and complications arise when searching for specific motifs: selection biases may also occur because detection strategies may favor some classes of active motifs; false positives may require further experimental tests (14).

These technical difficulties could be ameliorated by a systematic computational method for modeling the process of pool generation and selection of active motifs. More importantly, modeling could guide fruitful experimental efforts and discourage less productive search avenues through analysis and engineering of sequence pools for target motifs. Reliable simulation models could

*To whom correspondence should be addressed. Tel: +1 212 998 3116; Fax: +1 212 995 4152; Email: schlick@nyu.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

**Table 1.** *In vitro* selected RNAs, pool sequence length, pool size and motif yield[a]

| RNA | Sequence length (nt) | Pool size (seqs) | Yield (#/$10^9$) | Reference |
|---|---|---|---|---|
| GTP aptamer | 60–80 | $2.5 \times 10^{14}$ | 0.05 | (34) |
| ATP aptamer | 169 | $\sim 10^{14}$ | 1–10 | (31) |
| Neomycin B aptamer | 74 | $\sim 10^{15}$ | 0.001 | (30) |
| Chloramphenicol aptamer | 70–80 | $10^{14}$–$10^{15}$ | 0.074–0.0074 | (32) |
| Streptomycin aptamer | 74 | $10^{15}$ | 0.0043 | (33) |
| Macugen | 30 | $\sim 10^{14}$ | 0.0143 | (13) |
| DSL ligase | 30 | $1.4 \times 10^{14}$ | 0.0001 | (35) |
| T80 ligase | 35 | $\sim 10^{14}$ | 0.0002 | (36) |

[a] Motif yield is the number of reported active RNAs per $10^9$ sequences. This value can be biased by experimental details (e.g. RNA selection strategies, threshold values for binding constants and reaction rates).

also be used to corroborate experimental results and help to identify technical experimental problems. Ultimately, modeling and simulation could elucidate the physiochemical factors that dictate the presence of active RNAs in sequence pools and relate sequence to structure and function.

A major challenge in computational modeling of *in vitro* selection is the enormous size of sequence pools ($\sim 10^{15}$ molecules), roughly eight orders of magnitude larger than the human genome ($\sim 10^9$ nt) for 100-nt sequence pools. Modeling of pool generation and screening for active RNAs requires computation of RNA's primary, secondary and tertiary structures, as well as ligand interactions. Computations involving such large pool sizes demand the use of both novel approaches and large-scale computing resources.

Already, various mathematical approaches have been reported for modeling aspects of *in vitro* selection (15,16). Waterman and coworkers developed a mathematical model for *in vitro* selection and amplification by relating motif selection probabilities and protein binding constants (15). Levine and Nilsen-Hamilton (16) quantified the convergence of *in vitro* selection by providing upper and lower bounds on the number of rounds required to enrich the pool with a specified set of binding affinities by using an approach originally developed by Irvine *et al.* (17).

Knight *et al.* (18) combined approximate probabilistic analyses with a secondary folding algorithm which estimates motif probability; they used this approach to predict the frequencies of an isoleucine aptamer and hammerhead ribozyme in random pools by folding a large number of sequences using computing clusters. Their investigation showed that certain regions of the composition space are enriched with these motifs, and that their computed yields are consistent with reported experimental results. Recently, in an approach designed for RNA microarray applications (19), random pools of size $10^8$ sequences have also been screened for RNAs binding specific targets using a 3D folding algorithm and a docking program.

The distribution of RNA motifs in nucleotide sequences has also been investigated by the Cedergren (20) and Schlick (21) groups using motif scanning programs such as RNAMOT(22) and RNAMotif (23). These studies highlighted the over- and under-representation of specific RNA motifs in randomized sequences; our additional studies using RNA graphs also led to a similar conclusion (24). The Cedergren group identified motif hits without structure folding, whereas the Schlick group used folding and thermodynamic criteria to filter the candidates. The present work extends these tools and develops new methods to handle the voluminous data associated with large sequence pools.

Recently, we have developed a mathematical tool for generating pools by nucleotide transition probability matrices (or mixing matrices) and combined it with graph theory representations for analyzing RNA structure space and designing structured pools (25,26). Here, we employ the methods of nucleotide transition probability matrix and pool design and develop new tools to generate, screen and filter very large pools. Namely, we use the nucleotide transition probability matrix approach for pool generation and design, employ an efficient motif-scanning program RNAMotif for identifying known RNA motifs (23), develop tailored screening algorithms, and accelerate program performance by program restructuring to allow very large pools to be analyzed effectively. Our tailored screening algorithms involve screening of predicted secondary structures using a measure of structural similarity as well as flanking sequence analysis to assess the effect of random sequences flanking active motifs. These screening criteria help remove false positives.

The overall combination of pool generation and screening is packaged in a suite of tools to generate and screen efficiently large *in-vitro*-like RNA pools for specific active motifs. Using parallel computing resources (1000 processors of the IBM Blue Gene/L (power 440) or Intel 64 (2.33 GHz) linux clusters), we can generate and screen $10^9$ sequences of length 60 nt within 1.14 min on IBM Blue Gene and a pool size of $10^{14}$ in 162 h on the Intel cluster. To demonstrate the reliability of this approach, we show that computed yields for simple hypothetical motifs are in excellent agreement with theoretical motif yields calculated based on probability theory. We also demonstrate the feasibility of our method by searching for a set of known aptamers and ribozymes in random pools. These computed motif yields overestimate experimental data, as expected, for at least three reasons: sequences of different functions can map onto the same motif, only limited tertiary analysis is considered and no aptamer/ligand interactions have been computed; in addition, the experimental values represent lower bounds. We then utilize the pool screening tools and pool design concept to show that designed pools for RNA ligases can produce higher yields for targeted motifs than random pools.

The advantage of our combined approach is the generation and screening of RNA pools up to $10^{14}$ in size to simulate key aspects of the *in vitro* selection process. The approach can be utilized by RNA researchers to evaluate the productivity of random and non-random sequence pools for specific motifs. Our pool design approach also offers an avenue for developing biased pools for improving the yields of complex RNA motifs which are rarely found in random pools. Remaining future challenges include developing approaches for discovering novel motifs and for efficient screening of tertiary interactions.

## MATERIALS AND METHODS

### Algorithm for generating, searching and filtering RNA motifs in large pools

Our overall computational algorithm consists of three major steps (Figure 1):

(1) 'Generate' large sequence pools and scan sequences by the RNAMotif program (23) using a motif descriptor based on desired sequence and structural features, without any sequence folding.
(2) 'Filter or screen' sequence hits for folding into active conformations that are known experimentally using programs for RNA folding and structure comparison such as RNAfold and RNAdistance (27).
(3) 'Further filter' candidate sequences from Step 2 using flanking sequence analysis to account for the effects of random sequences around an active motif structure.

Step 1 combines sequence generation and scanning so that the large sequence pools need not be stored. In practice, a batch of 60 000 sequences is generated at a time, scanned, and then discarded, retaining only sequence hits; this procedure is repeated until the desired entire pool is generated and scanned. Further, we integrate pool generation and motif scanning so that the process takes place in memory with minimal overhead from file system read and write operations to provide highly efficient analysis of very large pools.

In the first step, we generate RNA pools of a given length and nucleotide composition using the high quality pseudorandom number generator named Mersenne Twister (28). On the IBM Blue Gene at the Rensselaer Polytechnic Institute (RPI) or the Intel cluster at the National Center for Supercomputing Applications (NCSA), our program for simultaneous generation and scanning pools of $10^9$ and $10^{14}$ sequences of length 60 nt requires 1.14 min and 162 h, respectively (see 'Program Implementation' section below for details).

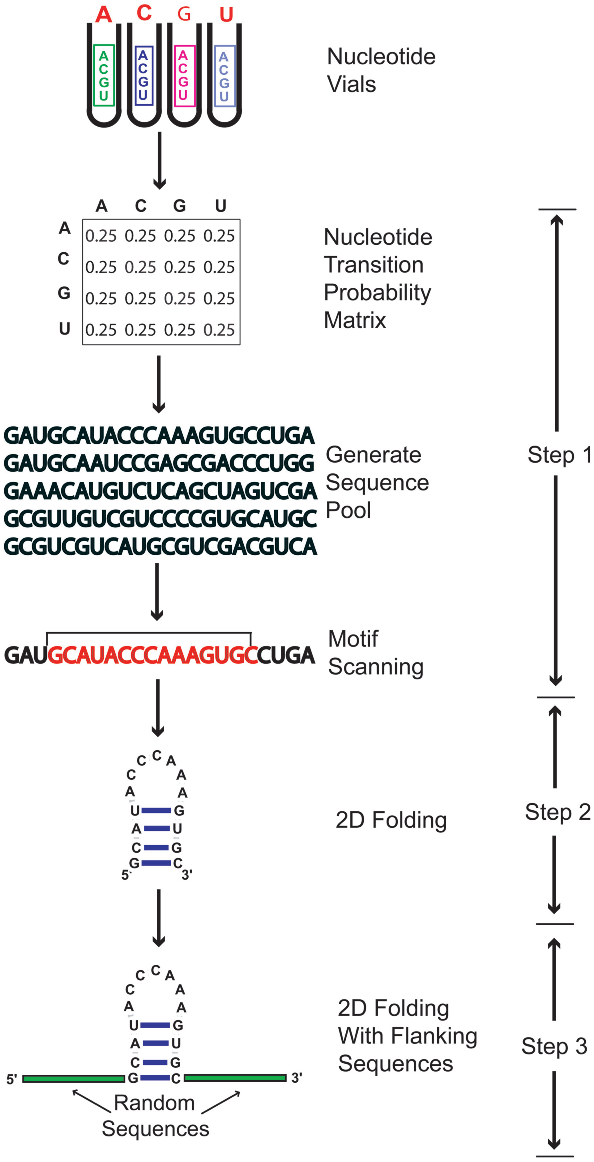

**Figure 1.** Schematic diagram of the three screening steps: (1) generate large sequence pools using nucleotide transition probability matrices and scan sequences by the RNAMotif program using a motif descriptor based on desired sequence and structural features, without any sequence folding. (2) Filter or screen sequence hits for folding into active conformations that are known experimentally using programs for RNA folding and structure comparison such as RNAfold and RNAdistance. (3) Further filter candidate sequences from Step 2 using flanking sequence analysis to account for the effects of random sequences around an active motif structure.

Also in this first step, we scan the pools using RNAMotif to simulate aspects of the experimental motif selection process. RNAMotif is a general pattern-search tool for target sequences and secondary structure motifs in nucleotide sequences (23). It can search for RNAs represented both as trees and as pseudoknots, of any degree of complexity. In particular, we formulate the motif descriptors based on the conserved sequence and structural features that are observed for multiple RNAs obtained from *in vitro* experiments (see the next section for descriptor details). The overall target secondary structure is defined by the number of paired and unpaired regions as well as the locations and sizes of stems, bulges, loops, junctions or pseudoknots. These features are estimated from sequence alone, without any folding. We allow sequence variability to the extent observed from experimental chemical probing or computational sequence alignment. The result of this scanning is a subpool of sequences that may match the target motif.

In Step 2, sequence matches from RNAMotif are further screened based on criteria such as correct target fold to remove false positives. Specifically, each of the candidate sequences identified by RNAMotif is folded using RNAfold (27), and the near-minimum free-energy structures (mfe) are determined. These structures are compared to the original descriptor for the target motif to determine if the specified structural features are present. To increase the efficiency of this screening step, we use a standard structure similarity measure (tree edit distance) to compare the structures of scanned motifs against the target motifs. The tree edit distance between two (full) tree secondary structures measures the minimum sum of the cost (insertion, deletion and replacement of nodes) along an edit path for converting one tree graph to another (29). We use this measure as implemented in RNAdistance of the Vienna RNA package (27). Only candidate sequences whose tree edit distances (obtained from their respective low free energy structures) are within a given threshold of a target structure are retained; others are discarded.

In Step 3, the motif candidates from Step 2 are further filtered using flanking sequence analysis to account for the effects of random sequences around an active motif structure. A random sequence is added at each end of the candidate active motif to the total experimental length for the motif. The length of the end random sequences is varied to reflect the random nature in which the active motifs are embedded in them. The structure 'Hamming distance' between target and folded candidate structure is calculated from aligned bracket symbols for the sequence region in consideration (27).

## RNA motif descriptors

How a motif is defined numerically is an important aspect of our algorithm. Below, the key features of descriptors for simple hypothetical motifs, aptamers and ligase ribozymes (Supplementary Figure S1, Figures 2 and 3) are summarized; descriptor files are provided in Supplementary Data.

*Simple hypothetical motifs.* We define two types of simple hypothetical motifs in RNA secondary structures, the stem-loop and stem-bulge–stem-loop. For each type, there is some variation of sequence conservation (Supplementary Figure S1). For example, structure 1 is 17 nt in length, and its overall form is a stem-loop, with a three-base-pair stem length. This structure is divided into three different descriptors: in descriptor 1, all regions are conserved; in descriptor 2, unpaired regions are conserved; and in descriptor 3, only stem-loop regions are conserved. Similarly, in structures 2 and 3, different regions are conserved in each of their descriptors. These descriptors of hypothetical motifs are designed to facilitate the comparison of theoretical and computational approaches.

*Aptamer descriptors.* We also define motifs for six aptamers obtained from *in vitro* selection (Figure 2): neomycin aptamer, ATP aptamer, chloramphenicol aptamer, macugen and GTP aptamer (see Table 1 for experimental data).

Many *in vitro* selected RNAs are members of families of similar sequences exhibiting the desired function. For antibiotics (neomycin, streptomycin, chloramphenicol) and ATP binding aptamers, we use motif descriptors defined by ref. (21). Using these descriptors, Laserson *et al.* (21) identified 37 candidate sequences in bacterial and archeal genomes that correspond to synthetic functional RNA motifs. For the macugen and GTP aptamers, we use structures selected from experimental results with consensus sequences for their family.

The neomycin aptamer has several key conserved regions in the helix with three G–U base pairings (30) (Figure 2a). With this definition in the descriptor, we maintain several residues in the loop region because they are involved in neomycin B binding. A common sequence feature is that the hairpin loop is composed of three G–U wobble pairs followed by a G–C pair just before the start of the loop. The loop exhibits a consensus where the first nucleotide is a guanine and the rest of the loop is adenine-rich; the loop is closed off by a non-canonical A–G pair.

All ATP-binding aptamers have a small, highly conserved consensus sequence embedded within a common secondary structure (31) (Figure 2b). This structure is composed of a simple hairpin motif with an asymmetric bulge containing the consensus sequence (GGAAG AAACUG), which mediates binding; the two helices surrounding the bulge are variable in length (we allow from 6 to 14 bp). The constructed descriptor precisely reflects this information.

Several of the selected chloramphenicol aptamer sequences exhibit sequence conservation and the capacity to fold into similar secondary structures (32) (Figure 2c). The structure is composed of a hairpin with two adenine-rich asymmetric bulges which are responsible for binding. The helix between the two bulges is only slightly variable in length (from 4 to 6 bp), while the two outer helices allow higher variability.

The structure of macugen (a therapeutic aptamer that binds to vascular endothelial growth factor) has a stem-bulge–stem-loop (13) (Figure 2d). The consensus sequences in the bulge region are conserved AAUCA on the 5′ side and A on the 3′ side. Two stem regions are variable in length (from 4 to 7 bp).

The streptomycin-binding aptamer has a consensus sequence in the bulge regions (33) (Figure 2e). The raw sequences are 75-nt long whereas the corresponding minimally active structure is only 40 nt. The minimal structure is composed of two asymmetric bulges in a hairpin structure, with the middle helix remaining roughly constant in length, and the other two helices are highly variable. We search for the sequence motif GNANNUG (where N is A, C, G or U) in the asymmetric bulge.

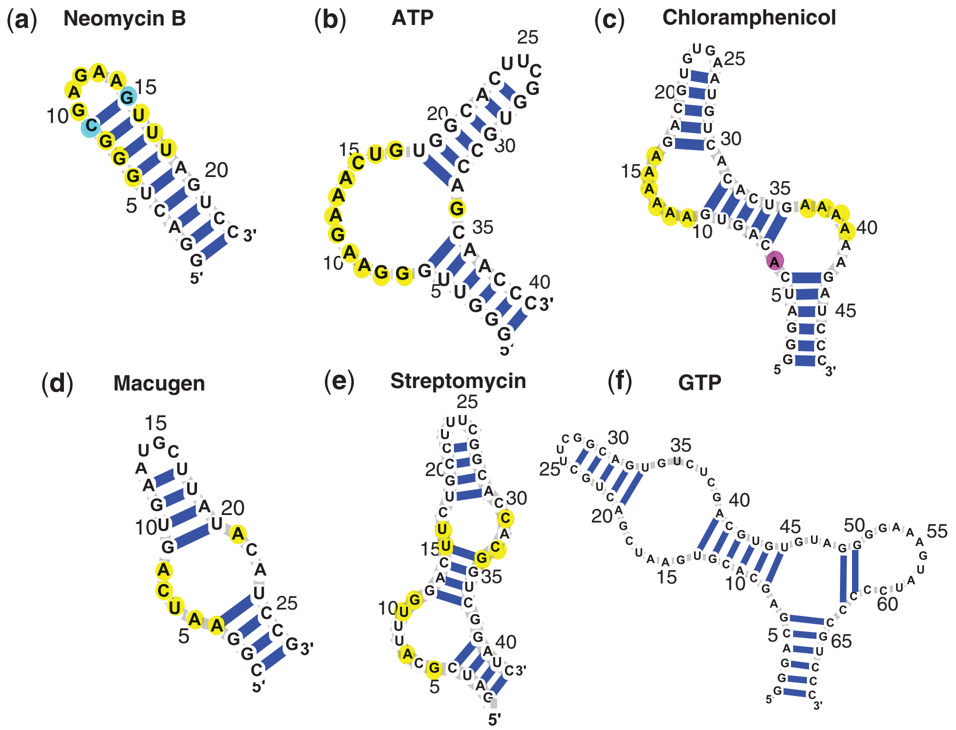

**Figure 2.** The aptamer motifs used to design motif descriptors and screen pools: (**a**) neomycin B, (**b**) ATP, (**c**) chloramphenicol, (**d**) macugen, (**e**) streptomycin, (**f**) GTP. Yellow bases are conserved, blue bases are C or G and pink bases are A, C or G. All displayed base pairs are required by the motifs. For (a–e), variations in length at both paired and unpaired regions are allowed. Details of conserved and variable motif elements are provided in motif descriptors in Supplementary Data.

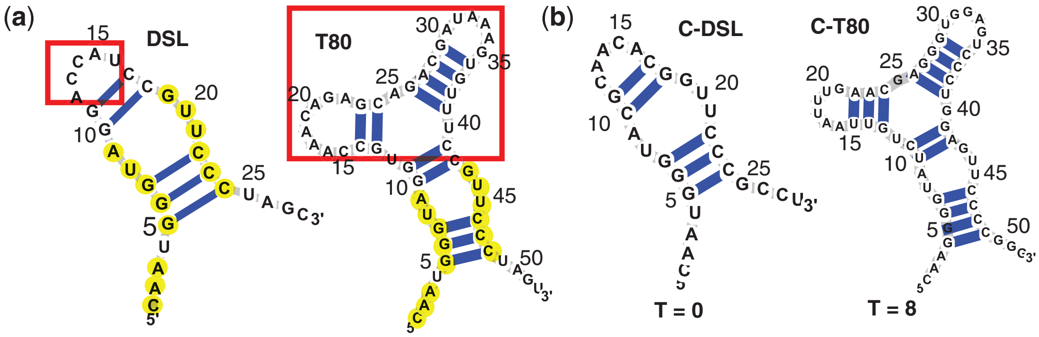

**Figure 3.** (a) Catalytic modules of the DSL and T80 ligases. Yellow bases are conserved. Red boxes represent the hairpin region in DSL and the modified random region in T80. (**b**) Candidates of the DSL and T80 screened/filtered from random pools (with size of $10^9$ and $10^{14}$ sequences, respectively) by our computational approaches.

In describing the GTP-binding aptamer, we choose the 9–4 class which has the highest binding affinity among seven GTP aptamer families (34). The 9–4 GTP aptamer has a sequence of 69 nt in length, and 4 stems as shown in Figure 2f.

*Ligase ribozyme descriptors.* We explore both the designed and selected ligase (DSL) (35) and T80 (36) ligases (Figure 3a and Table 1). Both catalyze the RNA-templated joining of an oligonucleotide 3′-hydroxyl and an oligonucleotide 5′-triphosphate, forming a 3′,5′-phosphodiester and releasing inorganic pyrophosphate (36). The DSL ligase was obtained by Inoue and his coworkers using a modular approach by attaching a 30 nt random sequence region to a structural scaffold derived from a naturally occurring ribozyme (35). Jaeger and coworkers also selected a DSL ligase ribozyme from a structured pool which had 30 nt random regions in the scaffold of a naturally occurring P4–P6 RNA with a known 3D structure (37,38). Its overall structure is defined by a stem-bulge–stem-loop of 28 nt. Based on the consensus sequences in ref. (35), we define our DSL catalytic module descriptor to have three conserved G–C base pairs in the stem near the end and several residues (yellow-colored) in the bulge and on the 5′ end side.

Recently, the Joyce group reported T80 ligase that exhibits enhanced activity. T80 was discovered by modifying the hairpin region of the catalytic module of the DSL ligase with a 35 nt random sequence (red boxes in Figure 3a represent the hairpin region in the DSL and a new random region in the T80) (36). The T80 ligase descriptor contains a three-way junction and a bulge-stem flanked by the four additional bases. Like the DSL ligase descriptor, sequences in the bulge region are conserved as UA and GUU on each side and three G–C base pairs in the stem.

### Theoretical motif yields in random pools

For random sequences, the probability of finding a given motif $M$ is calculated as follows:

$$\Pr(M) = \left(\frac{1}{4}\right)^{\text{cnt}} \times \left(\frac{1}{16}\right)^{\text{cbp}} \times \left(\frac{3}{8}\right)^{\text{ncbp}}, \tag{1}$$

where cnt, cbp and ncbp are the number of conserved nucleotides, conserved base pairs and non-conserved base pairs, respectively. This estimate can be compared with RNAMotif scanning results without secondary free energy filtering. For example, the probabilities of having DSL and T80 motifs are $5.24 \times 10^{-8}$ and $1.46 \times 10^{-12}$, respectively. When a motif has variable elements (range of stem or loop sizes), the probability of finding the motif is the sum of that in each distinct case ($M_1, M_2, \ldots, M_n$, where $n$ is the number of possible distinct motifs):

$$\Pr(M) = \sum_{i=1}^{n} \Pr(M_i)\,. \tag{2}$$

Motif yields in a pool $P$ of given length are calculated as follows:

$$\text{Yield}(M,P) = \Pr(M) \times S_p \times (l_p - l_m + 1), \tag{3}$$

where $S_p$, $l_p$, $l_m$ are the size of a pool, the length of sequences in the pool and the length of the given motif sequence. For example, in a $10^9$ sequence random pool having 60 nt in length, this corresponds to expected yields for DSL and T80 motifs of 17.29 and 0.01311, respectively.

### Designed sequence pools for ligase motifs

In addition to random pools, we consider non-random, designed pools to enhance the yield of selected motifs (e.g. RNA ligases). Our pool design approach is based on the idea that efficient sampling of the sequence space could lead to more productive searches for motifs. We use our nucleotide transition probability matrices (or mixing matrices) that can cover significant regions of sequence space (25,26); see also our web server, RAGPOOLS, available at http://rubin2.biomath.nyu.edu, for designing and analyzing structured pools for *in vitro* selection. Our set of 26 matrices covers vast regions of sequence space (Supplementary Figures S4 and S5). We apply these matrices to generate non-random pools for RNA ligases. To reduce the bias in our pool generation, we use starting sequences obtained by scanning random pools using the targeted motif descriptors rather than the known ligase sequences. The nucleotide transition probability matrices are then applied to selected candidate sequences from random pools to generate non-random pools, and motif yields are computed to identify the best designed pools.

### Program implementation

Our pool generation and scanning program consists of three integrated modules: module 1 is a pool generation routine utilizing one of two available pseudorandom number generators, the Scalable Parallel Random Number Generator and Meresenne Twister (all data here employ the latter); module 2 is a modified version of RNAMotif to run as a parallel subroutine for RNA motif scanning; module 3 is a statistical component for reporting and analysis of results. Our program has been implemented in the C language using the MPICH2 reference implementation of the message passing interface (MPI). The program has a linear complexity (Supplementary Table S1): $O(n)$ where $n$ is pool size times the length (nt) of each sequence. The combined program runs on multiple architectures including the IBM Blue Gene at RPI and Brookhaven National Laboratory (BNL), Intel cluster at NCSA, and a local SGI MIPS Origin 3200 cluster. The Blue Gene at RPI was used to produce most of the results in this article. On the Blue Gene, our program typically runs on 1000 processors for ~12 h to generate and scan a pool of $10^{12}$ sequences of 60-nt length (Supplementary Table S1) or ~1200 h for a pool size of $10^{14}$ sequences. On NCSA's Linux cluster with 1000 processors, the same analysis of the $10^{14}$ sequence pool required only 162 h. Previous motif pool screening efforts using secondary structure folding involved a Linux machine with 968 P3 compute processors (1 GHz) and required ~1000 processor hours for $10^8$–$10^9$ sequence pools (100 nt in length) (18). Although this computing approach has advantages in scalability and workstation utilization, our approach can be applied to larger pools due to an integration of pool generation and scanning modules.

### RNA tertiary folding algorithms

MC-Sym (39) predicts structures using structural knowledge about nucleotide conformations and statistical potentials. FARNA (40) uses fragment assembly similar to the Rosetta method for predicting protein structures (41). To aid 3D structure prediction, we use constraints on all base pairs with MC-Sym, and partial base pair constraints with FARNA, corresponding to 2D structures. We select the best 3D structures from MC-Sym using a probabilistic search over 30 min (neomycin B) and 12 h (ATP and streptomycin). For FARNA, we select the best 3D structure out of 25–50 models from simulations using 300 000 fragment insertions. Though highly approximate, such methods can help prune candidate folds of small RNAs (40 nt or less).

## RESULTS

Below we first compare computational yields for simple RNA motifs to theoretical expectations. Second, we analyze computational yields for aptamers in random pools using different screening tools to show that the candidate pools are reduced selectively and narrowed correctly, without losing too many good candidates. Third, we compute yields for ligases in random pools that serve to predict sequences for DSL and T80 ligase motifs. Fourth, we describe yields for ligases in random versus computationally designed pools to demonstrate the much higher yields in the latter.

### Theoretical versus computational yields of simple motifs in random pools

Comparing computational motif yields from RNAMotif to those from probability theory assuming motifs are found in one contiguous region [Equation (3) in 'Materials and Methods' section] in the limit of infinite pool size helps assess the pool size needed to reproduce theoretical yields. This comparison is performed without filtering with structure distance. Table 2 and Supplementary Figure S2 compare the computed yields for simple constructed motifs (Supplementary Figure S1) in 10 random pools containing $10^9$ unique 100 nt sequences. The computational yields are in excellent agreement with theoretical yields calculated by Equation (3), implying that this pool size is sufficient for the motifs studied; below, we use this pool size or larger for analysis of aptamer and ribozyme motifs. For example, the DSL ligase motif should have ~38.3 hits which is close to the mean of the computational yield, 36.7 hits (within the SD, ± 5.1). These results are consistent for all simple hypothetical motifs tested (Table 2). Of course, simple probability theory is inefficient for more complex structures, with variable lengths of paired and unpaired regions and mispairs/mismatches, which we treat next.

### Computational screening of aptamer motifs in random pools

Figure 4 shows the motif yields for a set of aptamer motifs (Figure 2) screened at different tree edit distances ($T$-values) of 6, 12 and unfiltered for $10^{12}$ sequences pools with given lengths 40 nt (neomycin B and macugen), 60 nt (ATP, chloramphenicol and streptomycin) and 100 nt (GTP); see also Supplementary Table S2. Our experience indicates that appropriate $T$-values are motif-length dependent: 30–40 nt motifs with $T \leq 6$ are sufficiently similar, whereas ~50 nt motifs require $T \leq 12$. The target aptamer motifs bind diverse targets including antibiotics, ATP, protein (macugen binds VEGF) and GTP. For each aptamer, we retain RNAMotif hits with motif energies (computed in RNAMotif based on motif matches) <50% of the minimum free energy (mfe) of the active aptamer sequence (Figure 2). The mfe values for macugen, neomycin B, ATP, chloramphenicol, streptomycin and GTP are −2.6, −7.10, −17.7, −14.8, −4.5 and −26.4 kcal/mol, respectively.

When T filtering is not applied (Supplementary Table S2), the hit numbers reported by RNAMotif are roughly correlated with the motif complexity, which can be described by the number of stems and conserved nucleotides in each aptamer motif: for neomycin B (one stem) 184 433 hits, ATP (two stems) 82 723 hits, macugen (two stems) 521 369 hits, chloramphenicol (three stems) 24 265 hits, streptomycin (three stems) 16 856 hits and GTP (four stems, one junction) 21 103 hits in $10^{14}$ random sequences. The pool fraction plot in Figure 4 shows that scanning of random pools for specific aptamer motifs results in a reduction of candidate pool size by six to eight orders of magnitude. Next, we perform *in silico* sequence folding and target structure matching on a much smaller set of motif candidates.

As we define the threshold tree edit distance to be smaller, the yields for the six aptamer motifs are reduced by one to two orders of magnitude, as shown in Figure 4 and Supplementary Table S2. For example, for the ATP

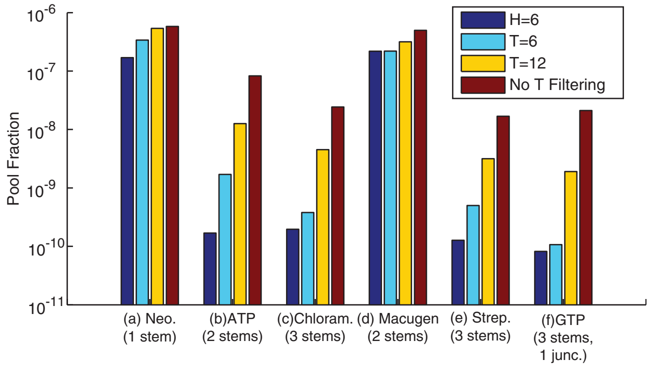

**Figure 4.** Frequencies of selected motifs for $10^{12}$ random-sequence pools for (**a**) neomycin B, (**b**) ATP, (**c**) chloramphenicol, (**d**) macugen, (**e**) streptomycin and (**f**) GTP. Experimental motifs are shown in Figure 2. Filtering analysis is done by the tree edit distance ($T$) of the minimal free energy structure (mfe) with respect to target motif and structure Hamming distance ($H$) from flanking sequences. See also Supplementary Table S2 for filtering of sequences with consideration of suboptimal states.

**Table 2.** Computational versus theoretical motif yields for simple motifs shown in Supplementary Figure S1

| Yield | Structure 1 | | | Structure 2 | | | Structure 3 | | | DSL Ligase |
|---|---|---|---|---|---|---|---|---|---|---|
| | S1-1 | S1-2 | S1-3 | S2-1 | S2-2 | S2-3 | S3-1 | S3-2 | S3-3 | |
| Mean (SD) | 4.9 (2.8) | 1064.2 (24.0) | 5000.2 (32.1) | 21 432.9 (178.6) | 772 577.1 (816.8) | 1 372 931 (1335.3) | 1.2 (1.2) | 1583.3 (49.8) | 75.1 (5.2) | 36.7 (5.1) |
| Theo. | 4.9 | 1056.1 | 5006.8 | 21 457.7 | 772 476.2 | 1 373 291 | 1.2 | 1565.3 | 77.2 | 38.3 |

All computational and theoretical yields are scaled to correspond to a $10^9$ random sequence pool. To estimate computational motif yields, 10 random pools of size $10^9$ each 100 nt are generated and their results are averaged (means and standard deviations provided). Note that structures 1−3 are simple hypothetical motifs in RNA secondary structures, the stem-loop and stem-bulge–stem-loop. For each type, there is some variation of sequence conservation (S1-1 to S3-3, see Supplementary Figure S1). For DSL ligase (designed and selected ligase), see Figure 3a.

aptamer, the motif yields are 1706, 12 634 and 32 546 as filtered by tree edit distance 6, 12 and 18, respectively, compared with 82 723 for no structure matching; for the GTP aptamer, the corresponding motif yields are 107, 1903 and 4939 versus 21 103. These results indicate that the vast majority (50–90%) of RNAMotif sequence hits do not fold to the target structure. Supplementary Figure S3 shows representative structures in different tree edit distance classes: it shows that structures with the smallest tree edit distance eight are close to the target GTP aptamer structure, with global topological features (four stems and one junction) preserved; large tree edit distances ($T > 12$) yield markedly different candidates.

To also show that the false positive rate is low, meaning that the remaining candidates after motif scanning are not likely to arise from randomized sequences, we randomly shuffle the actual aptamer sequences 1000 times and fold them using RNAfold (27). We confirmed that the number of shuffled sequences that fold correctly into the target motifs is low: none of the reshuffled sequences for ATP, chloramphenicol and streptomycin aptamers fold into the correct motifs with $T = 6$ and none for the GTP aptamer with $T = 12$, while 25 and 11 correspond to the small neomycin B and macugen aptamers with $T = 6$ (Supplementary Table S3). Likewise, the false negative rate due to incorrect folding of known aptamer structures is low since all the aptamers are folded to within $T = 6$. Of course, these tests are limited by the accuracy of 2D folding algorithms which can correctly predict ~75% of base pairs (42) but are much more accurate for short compared to long RNAs. Note that our folding simulations were applied with constraints on: two bases (G8, U16) for ATP, one base (A4) for macugen, two base pairs (G1–C42 and G12–C37) for streptomycin and four bases (G14, G51, G52, C61) for GTP because unconstrained RNAfold produces incorrect structures. With these constraints, the $T$-values for the known aptamers are all zero.

Because some suboptimal structures are similar to the target fold, we also filter the data with the minimal free energy structure and suboptimal states within 5–10% of the minimal free energy (Supplementary Table S2). The suboptimal states are calculated using the subopt function of the RNAfold package. The yield trends are similar with those filtered by minimum free energy structures. Within 5% of the mfe, the yields are up to 1.7 times larger compared to the mfe, and within 10% the yields are up to 3.67 times larger; the yield for candidate streptomycin structures decreases by 3-fold because we removed two base-pair constraint requirements, and this makes correct folding less likely.

### Flanking sequence analysis of screened aptamer motifs

The gap between experimental data and computational motif yield after structure matching is still considerable; for example, the pool of computational candidates is 10 times larger for the ATP aptamer and greater than five orders of magnitude for the neomycin B aptamer (Table 1). To further screen the sequence candidate pools, we developed the flanking sequence analysis test for filtered motifs. Random or non-essential nucleotide regions that flank the discovered functional motifs can destabilize the motif's fold due to entropic and energetic factors, thereby reducing the probability of observing active motifs. For the antibiotics and ATP binding aptamers under study (Figure 2), the sequence lengths range from 40 to 169 nt, while the actual functional regions of these molecules can be as short as 27 and as long as 50 nt (Table 1 and Supplementary Table S4). The entire sequence must fold in such a way that the flanking regions do not interfere with the formation of the active substructure.

To simulate this effect, we flank the candidate sequences that pass the folding requirements (threshold $T = 6$ and 12) with random sequences on either side such that the sum of candidate sequence plus the two flanking sequences equals the length $L$ of the RNAs in the respective experimental pools (Supplementary Table S4). For each candidate sequence, we perform 100 flanking sequence trials using variable (random) lengths and compute their structure Hamming distances ($H$). This measure of similarity between two RNA structures is calculated by alignment between the target structure and the substructure of candidate sequence excluding the random flanking regions (see 'Materials and Methods' section); the tree edit distance is not a suitable measure for this application since the substructure can interact with the flanking regions.

Figure 4 and Supplementary Table S4 show the results of filtering by flanking sequence analysis for candidate sequences with thresholds of $T = 6$ and 12. For $T = 6$ candidates of neomycin B aptamer, 64 241 sequences out of 87 435 pass the flanking test with $H = 6$, representing a reduction factor of 1.36. For the $T = 6$ macugen candidates, the reduction factor is small (216 470 from 219 692). However, many of the ATP, chloramphenicol and streptomycin aptamer candidates fail the flanking test (169, 196 and 127 from 1706, 380, and 499); the long flanking sequence length of ATP aptamer (~120 nt) relative to GTP aptamer (~20 nt) accounts for ATP aptamer's low yield. Thus, the flanking test reduction factor can be significant. Other studies with constant, primer flanking sequences for isoleucine motif suggest a reduction factor of ~10 (43). Our estimated neomycin B aptamer frequency of 64 241 in $10^{12}$ sequences is still higher than the yield of one suggested by experiments (30), and further screening by tertiary fold and ligand binding would be needed. Still, the candidate set is already a good starting point, considering also that the RNA selection experiments represent a lower bound and the initially much larger candidate pool size (Figure 4).

### Tertiary folds of candidate RNAs: a preliminary analysis

The functionality of candidate aptamers from random pools can be examined by predicting their tertiary structures. We use the folding algorithms MC-Sym and FARNA developed by Major (39) and Baker (40) groups, respectively. Though such methods are highly approximate, for small RNAs (~40 nt or shorter) they can be valuable for pruning candidate folds. First, we predict the

3D folds of the wild-type (WT) neomycin B, ATP and streptomycin aptamers whose tertiary structures have been solved experimentally. As shown in Supplementary Table S5, the RMSD values are 3.12, 5.86 and 9.0Å for WT neomycin B (23 nt), ATP (43 nt) and streptomycin (44 nt) aptamers, respectively. For the neomycin B aptamer, both folding algorithms predict similar RMSD values (3.12 Å from FARNA and 3.60 Å from MC-Sym). Because both programs perform poorly on the WT ATP and streptomycin aptamers (6 Å or more away from the native structure), better accuracy cannot be expected. Thus, we confine the fold prediction of candidate aptamers to neomycin B sequences. Figure 5 and Supplementary Table S5 show the superimposed structures and RMSD values for two candidate neomycin B aptamers, each with the respective predicted native structure. Figure 5 shows a candidate neomycin B structure (Candidate 1, $T = 6$) with 3.30 Å RMSD predicted by MC-Sym, and another candidate (Candidate 2, $T = 6$) has 2.33 Å RMSD predicted by FARNA. The predicted candidate structures display similar overall folds and reproduce the ligand binding regions fairly well; the small structural distortions could be caused by structural flexibility (e.g. induced fit of antibiotic ligands) as well as inaccuracies in folding algorithms.

Furthermore, as shown in Supplementary Table S5, the RMSD values predicted by FARNA of the binding pocket region between the neomycin B WT and our candidates (Candidates 1 and 2) are 4.17 and 2.40 Å for Candidates 1 and 2, respectively. Significantly, the candidate structures contain the conserved structural elements of the WT neomycin B-bound RNA, including the Watson–Crick stem segment, a continuous segment containing three consecutive G–U (6·18, 7·17 and 8·16) wobbles, a Watson–Crick G15–C9 pair, a GAGA hairpin loop closed by a sheared G10–A14 mispair and a looped-out A13 base that acts as a flap over the bound neomycin B. The G6–U18 mismatch and G15–C9 base pair represent the boundaries of the RNA-binding pocket with a prominent role for A13 in the encapsulation of a segment of the bound neomycin B (44). Similarly, the binding pocket of neomycin B aptamer Candidate 2 has three continuous G–U pairs, a Watson–Crick G–C pair and a GAUA hairpin loop closed by a sheared G–A mispair and a looped-out A13 base. The binding pocket of neomycin B aptamer Candidate 1 has the same helix region conformation with the neomycin B aptamer, but has a GUAU hairpin loop instead of a GAGA hairpin loop. This analysis of the neomycin B binding pockets of candidates suggests that promising active RNA molecules can be recovered from systematic screening of large pools, consistent with recent screening of smaller pools using folding and docking algorithms (19).

### Computational ligase motif yields in random pools from $10^9$ to $10^{14}$ sequences

RNA ligase ribozymes catalyze the formation of phosphodiester bonds (3′,5′ and 2′,5′ linkages). The DSL ligase was obtained by Inoue and coworkers using a modular approach by attaching a 30 nt random sequence region to a structural scaffold derived from a naturally occurring ribozyme (35). Jaeger and coworkers also selected a DSL ligase ribozyme from a structured pool which had 30 nt random regions in the scaffold of a naturally occurring P4–P6 RNA with a known 3D structure (37,38). The RNA scaffold acts as a stabilizing fold, and the ligase's catalytic module in the variable region is determined via *in vitro* selection and evolution. Here we explore computational screening of two RNA ligase ribozymes: the simple DSL ligase with a two-stem catalytic motif, and the more complex T80 ligase motif with four short stems and a three stem junction (Figure 3a).

We screen for these two ligase motifs in 60 nt random sequence pools ranging from size $10^9$ to experimental pool size of $10^{14}$ using RNAMotif and target structure matching. Table 3 shows the frequency of candidate

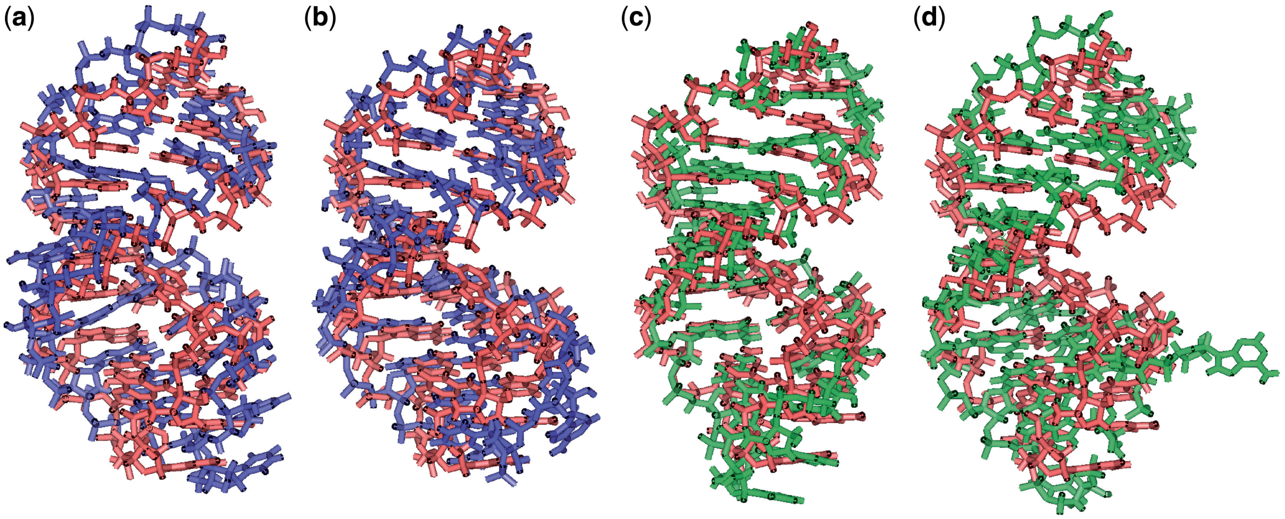

**Figure 5.** Tertiary structures of neomycin B aptamer: (**a**) Candidate 1 from FARNA superimposed with native structure (1NEM represented in red; RMSD = 4.61 Å), (**b**) Candidate 1 from MC-Sym superimposed with native structure (RMSD = 3.30 Å), (**c**) Candidate 2 from FARNA superimposed with native structure (RMSD = 2.33 Å), and (**d**) Candidate 2 from MC-Sym superimposed with native structure (RMSD = 4.12 Å).

**Table 3.** Computational motif yields of ligases in 60 nt random sequence pools ranging from $10^9$ to $10^{14}$ sequences

| Ligase | Computational motif yields in different pool sizes | | | | | | Experimental motif yields (in a $10^{14}$-sized pool) |
|---|---|---|---|---|---|---|---|
| | $10^9$ | $10^{10}$ | $10^{11}$ | $10^{12}$ | $10^{13}$ | $10^{14}$ | |
| DSL | 14 (1) | 174 (13) | 1739 (146) | 16 880 (1423) | 173 113 (14 286) | 1 729 810 (142 451) | 10 |
| T80[a] | 0 | 0 | 0 | 14 (0) | 138 (0) | 1295 (27) | 20 |

Both unfiltered and filtered results are presented; filtered results in parentheses are for tree edit distances 0 and 10 for DSL and T80, respectively.
[a]Added two base pair constraints.

motifs by RNAMotif and by filtering with target structure matching with a tree edit distance threshold of $T = 0$ for DSL and $T = 10$ for T80; for T80, we use constraints on two base pairs G26–U39 and G29–U36 to perform folding because unconstrained RNAfold produces a T80 structure without these base pairs. The motif frequency increases in proportion to the pool size. As expected, the RNAMotif hits or unfiltered results are in agreement with estimates from probability theory [Equation (3)]. Table 3 shows that candidate DSL motifs with the same target structure ($T = 0$) begin to emerge when the pool size is $10^9$ or larger, whereas candidate T80 motifs with $T \leq 10$ do not emerge until the pool size is $>10^{12}$. For DSL, 1 729 810 sequences are selected by RNAMotif from a $10^{14}$ sequence pool, and 142 451 of them fold into the DSL structure ($T = 0$), giving a reduction factor of 12 for structural matching screening. For T80, 1295 sequences are selected by RNAMotif from a $10^{14}$ sequence pool, and only 27-fold into a T80-like structure ($T = 10$), representing a reduction factor of 48. Experimentally, the yields for DSL and T80 motifs are 10 and 20 sequences, respectively, in a $10^{14}$ pool. For DSL, the higher computational yield is probably due to factors similar to those for neomycin B aptamer discussed above, namely 3D requirements and possible experimental omissions. For T80, the higher experimental yield is likely due to the structural enrichment of *in vitro* evolution rounds elaborated below.

Again, we confirm that the false positive rates for screening of the ligases are low. Using random shuffling of actual DSL and T80 sequences 1000 times and folding them, we find no cases with $T = 0$ from DSL and with $T = 10$ from T80 of target motif emergence (Supplementary Table S6).

These results clearly indicate that, even for the moderately complex T80 motif, simulations of very large pools approaching experimental pool size are needed to find good candidates. This explains the rarity of the T80 motif which was recovered experimentally from a pool of $10^{14}$ RNA molecules. The T80 candidates were subsequently subject to a number of continuous evolution transfers to improve catalytic rates (36).

Figure 3b shows a candidate DSL ligase motif ($T = 0$) produced from the $10^9$ pool and a candidate T80 motif ($T = 8$) from the $10^{14}$ pool. The former (C-DSL in Figure 3b) folds exactly into the target conformation ($T = 0$) corresponding to its lowest free energy structure. It has a short (28 nt) sequence and simple (stem-bulge–stem-loop) secondary structure which allows most of the bases to be paired; this candidate differs in 11 positions from the known DSL, even though the core region is conserved. Our T80 candidate (C-T80 in Figure 3b) is slightly different from the actual T80 ligase: C-T80 has a similar global structure (four stems with a junction) to the experimental structure, but it has one more base pair in each of the two stems.

### Enhancing the selection of ligase ribozyme motifs using designed pools

In addition to random pools, we now consider non-random, designed pools to enhance the yield of RNA ligases. Our pool design approach is based on the idea that efficient sampling of the sequence space could lead to more productive searches for motifs. Although random sequence pools can in theory sample a large diversity of sequences, this is rarely achieved due to incomplete sampling of sequence space for large pool lengths (>25 nt). As shown previously (25), our non-random matrices can achieve a greater coverage of sequence space than random sequences. In contrast, the sequence coverage for pools generated using base composition (nucleotide transition probability matrix class F in Supplementary Figure S5) is limited, as shown in the sequence space mapping in Supplementary Figure S4.

We use our nucleotide transition probability matrices (or mixing matrices) that can cover significant regions of sequence space (25,26); see also our web server, RAGPOOLS, available at http://rubin2.biomath.nyu.edu, for designing and analyzing structured pools for *in vitro* selection. Our set of 26 matrices covers vast regions of sequence space (Supplementary Figures S4 and S5). RAGPOOLS designs pools using novel structural motifs, represented as RNA graphs, and nucleotide transition probability matrices; RNA graphs can be equivalently represented as abstract shapes developed by Giegerich and collaborators (45,46).

We apply these matrices to generate non-random pools for RNA ligases. To reduce the bias in our pool generation, we use starting sequences obtained by scanning random pools (C-DSL and C-T80 in Figure 3b) using the targeted motif descriptors rather than the known ligase sequences or those from the limited number (30 sequences) of our RAGPOOLS web server. The mixing matrices (Supplementary Figure S5) are then applied to selected candidate sequences from random pools to generate non-random pools. These are scanned for target motifs and their yields are compared to identify the best designed pools.

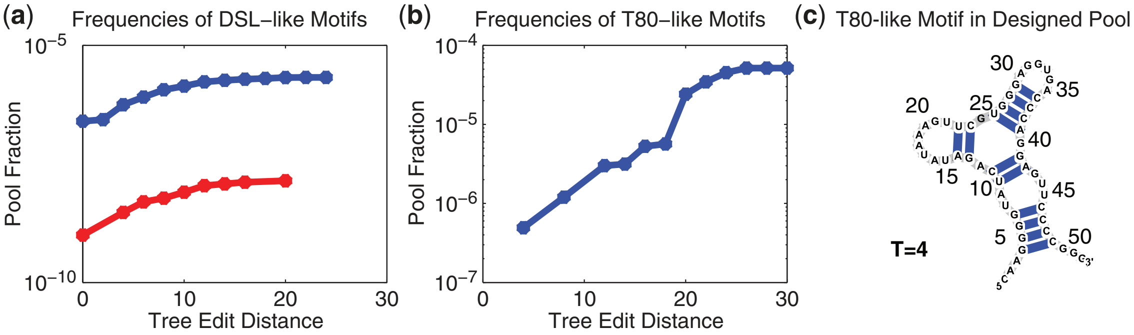

**Figure 6.** Frequencies of selected sequences as a function of tree edit distance for (**a**) DSL and (**b**) T80 in Figure 3a in $10^9$ sequence designed and random pools. Designed pools are generated from C-DSL and nucleotide transition probability matrix 14 for DSL and C-T80 and nucleotide transition probability matrix 9 for T80. (**c**) T80 candidate with $T = 4$ from a designed pool. No T80-like motif was found from a $10^9$ sequence random pool.

Figure 6 shows motif frequency as a function of the threshold tree edit distance for pools generated from C-DSL sequence with nucleotide transition probability matrix 14 and C-T80 sequence with matrix 9. These matrices consistently yield higher motif frequencies than others. As shown in Figure 6a for the DSL case, the motif frequency in the designed pool at all $T$-values is significantly higher than that in random pools. For example, at $T = 0$, the motif frequency of the designed pool (C-DSL and matrix 14) is 250 times higher relative to the random pool. The sequence diversity is also significantly increased relative to the random pool: for the random pool, candidate sequences with $T = 0$ differ in seven positions on average, whereas the designed pool yields 10.9, suggesting that the DSL designed pool enriches the target motif. Similarly, the designed pool by C-T80 and matrix 9 (Figure 6b) has higher yields for T80-like motifs for different $T$-values. Interestingly, the structure of T80 candidate with $T = 4$ (Figure 6c) selected from the designed pool is remarkably similar to the experimental motif (with sequence dissimilarity in 28 positions). Thus, the designed pools generated by these two matrices outperform random pools.

## DISCUSSION

We have proposed a set of tools, including pool generation, motif searching and motif screening techniques, to offer an automated approach for analyzing sequence pools to aid the discovery of RNA motifs. These include new tools for generating and designing large pools efficiently using mixing matrices and supercomputing resources, and development of a pipeline for motif screening using new and existing tools for flanking sequence analysis, motif search (23), 2D structure prediction/comparison (27) and 3D structure folding (39,40). Our approach is applicable to tree and pseudoknot structures in large random and designed pools. Currently, we can very quickly generate and scan pool sizes of $10^{12}$ but analyzing $10^{14}$ sized pools is also feasible, as demonstrated. Previous work by Knight *et al.* (18) screened $10^8$ size pools for sequences that satisfy the target 2D motifs (18). Recent screening of random pools using 3D folding and docking software for microarray applications also involved $10^8$ size pools (19). With further algorithmic advances in the near future, it may be possible to reach yet larger $10^{15}$ pool size. We have also developed techniques for filtering motif hit sequences: by utilizing secondary structure folding algorithms combined with the automated calculation of similarity between structures (tree edit and Hamming distance measures); and by flanking sequence analysis. The structural similarity screening aims to mimic aspects of experimental screening procedures (binding selection via chromatography) through which experimental pools are enriched with molecules possessing the desired function. Our consensus binding motifs here replace the affinity chromatography in *in vitro* experiments. Overall, our stepwise screening protocol leads to the following candidate pool size reductions: motif scanning (6–8 orders of magnitude); structure similarity (1–2 orders); and flanking sequence analysis (1 order). Moreover, we have shown that pools are appropriately narrowed by sequence shuffling experiments. Thus, we estimate that our screening methods can efficiently reduce candidate sequence pools by about nine orders of magnitude, underscoring the fact that most random sequences are non-functional.

The applications of our computational tools include assessment of sequence pools for specific motifs, optimization of pools with respect to sequence length and pool size, and the design of non-random sequence pools to improve motif yields. The success in finding specific motifs in sequence pools depends on motif complexity; selection experiments can fail if target motif or function cannot be found in the pool. Pool assessment for target motifs can be performed prior to a synthesis experiment to avoid performing experiments with low likelihood of finding the motifs. This strategy can be applied in the final stages of aptamer optimization after the target motifs are known. Motif yield is also dependent on sequence length and pool size. Our scanning and filtering tools can be used to determine the appropriate sequence length for a target motif and the smallest pool size to yield a sufficient number of good candidates. Finally, sequence pools can be enriched for target motifs by a combined application of pool assessment and design (25,26).

### Challenges in reproducing experimental motif yields computationally

Clearly, predicting novel active RNA sequences without laboratory experimentation is a challenge. Although the capability to scan all possible active sequences is an advantage of our approach, our computational yields are higher than the experimental yield. This discrepancy is due to a number of factors such as the limitation of pool size, design of motif descriptors and screening using motif and secondary structure information. Still, our screening methods allow analyses of large pools and drastically reduce the number of motif hits. Experimental error sources or selection inefficiencies (motifs missed by selection procedures) and under-sampling are also possible. However, overcoming the limitations enumerated below will help advance *in silico* RNA selection.

First, our computational approach is limited by the accuracy of the secondary folding algorithms used to predict structures and the lack of accurate tertiary structure folding. Currently, 3D folding algorithms such as MC-Sym (39) and FARNA (40) are available for RNA structures <40 nt in size. These could be incorporated into our design and screening protocol for the final pool candidates. Our preliminary analysis of neomycin B aptamers shows that 3D folding methods can provide valuable tertiary-structure information of small RNAs. Thus, such methods could be fully integrated into our design and screening protocol for the final pool candidates. Future work could incorporate more systematic 3D structure analysis based on recent experimental advances in 3D structure identification (47,48), in combination with ligand docking and molecular dynamics simulations.

Second, generation and screening of large pool sizes ($10^{15}$) within a reasonable length of time may be necessary. This is on the horizon with additional reprogramming strategies. Improvements are possible to reduce the run time and avoid file system bottlenecks.

Third, computer simulation of *in vitro* selection relies on the prediction of structure and sequence relationships to design an appropriate descriptor for a given function, while experimental searches can select functional molecules without any previous knowledge of the molecular features that confer this functionality. A better understanding of the relationship between sequence, structure and function of RNA molecules is a fundamental challenge in structural biophysics.

Our work offers experimentalists a tool for assessing pools and designing optimal pool parameters (sequence length, composition, motif type or complexity) before performing synthesis and selection experiments, through utilities available on our web server (http://rubin2.biomath.nyu.edu). The designed pools targeting two ligase ribozymes (DSL and T80) and promising candidates reported here (Figures 3b and 6) suggest that experimental yields might be improved and that new avenues for understanding sequence/structure/function relationships could be fruitful. We invite users to utilize our utilities and report to us their experiences.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank David Rooklin and James Potter for their preliminary flanking sequence and scanning analyses; and Dr Luc Jaeger for valuable comments. Computing time for this research was supported by resources of the CCNI and NYSTAR through group award, the National Science Foundation through TeraGrid resources provided by NCSA and LONI under (grant number TG-MCB080036N) and resources at the New York Center for Computational Sciences at Stony Brook University/Brookhaven National Laboratory which is supported by the US Department of Energy under Contract No. DE-AC02-98CH10886 and by the State of New York.

## FUNDING

National Science Foundation (DMS-0201160, CCF-0727001); Human Frontier Science Program, National Institutes of Health (R01GM081410); and NYU Dean's Undergraduate Research Fund (to J.A.I.). Funding for open access charge: National Institutes of Health (R01GM081410).

*Conflict of interest statement*. None declared.

## REFERENCES

1. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
2. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
3. Wilson,D.S. and Szostak,J.W. (1999) In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, **68**, 611–647.
4. Hermann,T. and Patel,D.J. (2000) Biochemistry - adaptive recognition by nucleic acid aptamers. *Science*, **287**, 820–825.
5. Soukup,G.A. and Breaker,R.R. (1999) Engineering precision RNA molecular switches. *Proc. Natl Acad. Sci. USA*, **96**, 3584–3589.
6. Soukup,G.A. and Breaker,R.R. (1999) Nucleic acid molecular switches. *Trends Biotechnol.*, **17**, 469–476.
7. Soukup,G.A. and Breaker,R.R. (2000) Allosteric nucleic acid catalysts. *Curr. Opin. Struct. Biol.*, **10**, 318–325.
8. Famulok,M. and Verma,S. (2002) In vivo-applied functional RNAs as tools in proteomics and genomics research. *Trends Biotechnol.*, **20**, 462–466.
9. Toulme,J.J., Di Primo,C. and Boucard,D. (2004) Regulating eukaryotic gene expression with aptamers. *FEBS Lett.*, **567**, 55–62.
10. Peracchi,A. (2004) Prospects for antiviral ribozymes and deoxyribozymes. *Rev. Med. Virol.*, **14**, 47–64.
11. Bagheri,S. and Kashani-Sabet,M. (2004) Ribozymes in the age of molecular therapeutics. *Curr. Mol. Med.*, **4**, 489–506.
12. Held,D.M., Kissel,J.D., Patterson,J.T., Nickens,D.G. and Burke,D.H. (2006) HIV-1 inactivation by nucleic acid aptamers. *Front Biosci.*, **11**, 89–112.
13. Lee,J.H., Canny,M.D., De Erkenez,A., Krilleke,D., Ng,Y.S., Shima,D.T., Pardi,A. and Jucker,F. (2005) A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of VEGF165. *Proc. Natl Acad. Sci. USA*, **102**, 18902–18907.

14. Joyce,G.F. (2004) Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.*, **73**, 791–836.
15. Sun,F.Z., Galas,D. and Waterman,M.S. (1996) A mathematical analysis of in vitro molecular selection-amplification. *J. Mol. Biol.*, **258**, 650–660.
16. Levine,H. and Nilsen-Hamilton,M. (2007) A mathematical analysis of SELEX. *Comput. Biol. Chem.*, **31**, 5924–5935.
17. Irvine,D., Tuerk,C. and Gold,L. (1991) Selexion - systematic evolution of ligands by exponential enrichment with integrated optimization by nonlinear-analysis. *J. Mol. Biol.*, **222**, 739–761.
18. Knight,R., De Sterck,H., Markel,R., Smit,S., Oshmyansky,A. and Yarus,M. (2005) Abundance of correctly folded RNA motifs in sequence space, calculated on computational grids. *Nucleic Acids Res.*, **33**, 5924–5935.
19. Chushak,Y. and Stone,M.O. (2009) In silico selection of RNA aptamers. *Nucleic Acids Res.*, **37**, e87.
20. Bourdeau,V., Ferbeyre,G., Pageau,M., Paquin,B. and Cedergren,R. (1999) The distribution of RNA motifs in natural sequences. *Nucleic Acids Res.*, **27**, 4457–4467.
21. Laserson,U., Gan,H.H. and Schlick,T. (2005) Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. *Nucleic Acids Res.*, **33**, 6057–6069.
22. Gautheret,D., Major,F. and Cedergren,R. (1990) Pattern searching alignment with RNA primary and secondary structures - an effective descriptor for transfer-RNA. *Comput. Appl. Biosci.*, **6**, 325–331.
23. Macke,T.J., Ecker,D.J., Gutell,R.R., Gautheret,D., Case,D.A. and Sampath,R. (2001) RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.*, **29**, 4724–4735.
24. Gevertz,J., Gan,H.H. and Schlick,T. (2005) In vitro RNA random pools are not structurally diverse: a computational analysis. *RNA*, **11**, 853–863.
25. Kim,N., Gan,H.H. and Schlick,T. (2007) A computational proposal for designing structured RNA pools for in vitro selection of RNAs. *RNA*, **13**, 478–492.
26. Kim,N., Shin,J.S., Elmetwaly,S., Gan,H.H. and Schlick,T. (2007) RAGPOOLS: RNA-As-Graph-Pools - a web server for assisting the design of structured RNA pools for in vitro selection. *Bioinformatics*, **23**, 2959–2960.
27. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
28. Matsumoto,M. and Nishimura,T. (1998) A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM TOMACS*, **8**, 3–30.
29. Shapiro,B.A. and Zhang,K.Z. (1990) Comparing multiple RNA secondary structures using tree comparisons. *Comput. Appl. Biosci.*, **6**, 309–318.
30. Wallis,M.G., Vonahsen,U., Schroeder,R. and Famulok,M. (1995) A novel RNA motif for neomycin recognition. *Chem. Biol.*, **2**, 543–552.
31. Sassanfar,M. and Szostak,J.W. (1993) An RNA motif that binds ATP. *Nature*, **364**, 550–553.
32. Burke,D.H., Hoffman,D.C., Brown,A., Hansen,M., Pardi.A. and Gold,L. (1997) RNA aptamers to the peptidyl transferase inhibitor chloramphenicol. *Chem. Biol.*, **4**, 833–843.
33. Wallace,S.T. and Schroeder,R. (1998) In vitro selection and characterization of streptomycin-binding RNAs: recognition discrimination between antibiotics. *RNA*, **4**, 112–123.
34. Carothers,J.M., Oestreich,S.C., Davis,J.H. and Szostak,J.W. (2004) Informational complexity and functional activity of RNA structures. *J. Am. Chem. Soc.*, **126**, 5130–5137.
35. Ikawa,Y., Tsuda,K., Matsumura,S. and Inoue,T. (2004) De novo synthesis and development of an RNA enzyme. *Proc. Natl Acad. Sci. USA*, **101**, 13750–13755.
36. Voytek,S.B. and Joyce,G.F. (2007) Emergence of a fast-reacting ribozyme that is capable of undergoing continuous evolution. *Proc. Natl Acad. Sci. USA*, **104**, 15288–15293.
37. Jaeger,L., Wright,M.C. and Joyce,G.F. (1999) A complex ligase ribozyme evolved in vitro from a group I ribozyme domain. *Proc. Natl Acad. Sci. USA*, **96**, 14712–14717.
38. Yoshioka,W., Ikawa,Y., Jaeger,L., Shiraishi,H. and Inoue,T. (2004) Generation of a catalytic module on a self-folding RNA. *RNA*, **10**, 1900–1906.
39. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51–55.
40. Das,R. and Baker,D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl Acad. Sci. USA*, **104**, 14664–14669.
41. Rohl,C.A., Strauss,C.E.M., Misura,K.M.S. and Baker,D. (2004) Protein structure prediction using rosetta. *Meth. Enzymol.*, **383**, 66–93.
42. Mathews,D.H. and Turner,D.H. (2006) Prediction of RNA secondary structure by free energy minimization. *Curr. Opin. Struct. Biol.*, **16**, 270–278.
43. Legiewicz,M., Lozupone,C., Knight,R. and Yarus,M. (2006) Size, constant sequences, and optimal selection. *RNA*, **11**, 1701–1709.
44. Jiang,L.C., Majumdar,A., Hu,W.D., Jaishree,T.J., Xu,W.K. and Patel,D.J. (1999) Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer. *Structure*, **7**, 817–827.
45. Steffen,P., Voss,B., Rehmsmeier,M., Reeder,J. and Giegerich,R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, **22**, 500–503.
46. Voss,B., Giegerich,R. and Rehmsmeier,M. (2006) Complete probabilistic analysis of RNA shapes. *BMC Biology*, **4**, 5–27.
47. Shechner,D.M., Grant,R.A., Bagby,S.C., Koldobskaya,Y., Piccirilli,J.A. and Bartel,D.P. (2009) Crystal structure of the catalytic core of an RNA-Polymerase ribozyme. *Science*, **326**, 1271–1275.
48. Ishikawa,J., Matsumura,S., Jaeger,L., Inoue,T., Furuta,H. and Ikawa,Y. (2009) Rational optimization of the DSL ligase ribozyme with GNRA/receptor interacting modules. *Arch. Biochem. Biophys.*, **490**, 163–170.