doi: 10.1093/nar/gky524

A pipeline for computational design of novel RNA-like topologies

Swati Jain[1], Alain Laederach[2], Silvia B. V. Ramos[3] and Tamar Schlick[1,4,5,*]

[1]Department of Chemistry, New York University, 1001 Silver, 100 Washington Square East, New York, NY 10003, USA, [2]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, [3]Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, [4]Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA and [5]NYU-ECNU Center for Computational Chemistry at New York University Shanghai, Room 340, Geography Building, North Zhongshan Road, 3663 Shanghai, China

Received October 24, 2017; Revised May 22, 2018; Editorial Decision May 23, 2018; Accepted May 24, 2018

ABSTRACT

Designing novel RNA topologies is a challenge, with important therapeutic and industrial applications. We describe a computational pipeline for design of novel RNA topologies based on our coarse-grained RNA-As-Graphs (RAG) framework. RAG represents RNA structures as tree graphs and describes RNA secondary (2D) structure topologies (currently up to 13 vertices, ≈260 nucleotides). We have previously identified novel graph topologies that are RNA-like among these. Here we describe a systematic design pipeline and illustrate design for six broad design problems using recently developed tools for graph-partitioning and fragment assembly (F-RAG). Following partitioning of the target graph, corresponding atomic fragments from our RAG-3D database are combined using F-RAG, and the candidate atomic models are scored using a knowledge-based potential developed for 3D structure prediction. The sequences of the top scoring models are screened further using available tools for 2D structure prediction. The results indicate that our modular approach based on RNA-like topologies rather than specific 2D structures allows for greater flexibility in the design process, and generates a large number of candidate sequences quickly. Experimental structure probing using SHAPE-MaP for two sequences agree with our predictions and suggest that our combined tools yield excellent candidates for further sequence and experimental screening.

INTRODUCTION

An understanding of the three-dimensional (3D) structure of macromolecules like RNA and proteins is crucial for deciphering critical cellular processes. Structural insights can be used to infer mechanisms as well as manipulate the functions of macromolecules for various therapeutic and industrial applications (1). As new 3D structures of macromolecules emerge from X-ray crystallography, Nuclear Magnetic Resonance (NMR), and cryo-EM (2–5), new opportunities for design applications arise. Modeling can play an important role in these design objectives.

Because of the importance of RNA molecules in cellular processes—from participating in transcription and translation of proteins (6) to catalysis (7–9) and gene regulation (10–13)—there has been a growing interest to determine and design structures of RNA molecules (14,15). Though RNA structure determination lags behind proteins, the number of solved RNA structures continues to grow, especially for structures of protein–RNA complexes (16). That new RNA molecules are continuously being discovered suggests that we have barely scratched the surface of RNA's rich repertoire of structures and possibly functions. Therefore, a systematic identification and design of new RNA structural topologies can help expand the pool of available RNA structures and better understand the fundamental forces that govern activity.

One of the most successful and common technique for designing novel RNA molecules is Systematic Evolution of Ligands by Exponential enrichment (SELEX) (17–19). This *in vitro* selection process involves multiple rounds of screening to select RNA molecules from a large pool of random or semi-random RNA molecules that bind a specific target or perform a specific function. SELEX has been successful for a variety of therapeutic applications (20–23). In addition, RNA molecules and their binding partners have been targeted for therapeutic interventions (24,25); RNA aptamers and ribozymes are being designed to bind specific targets (26,27); and drugs are being developed to target essential RNA molecules in disease causing organisms (28–30).

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

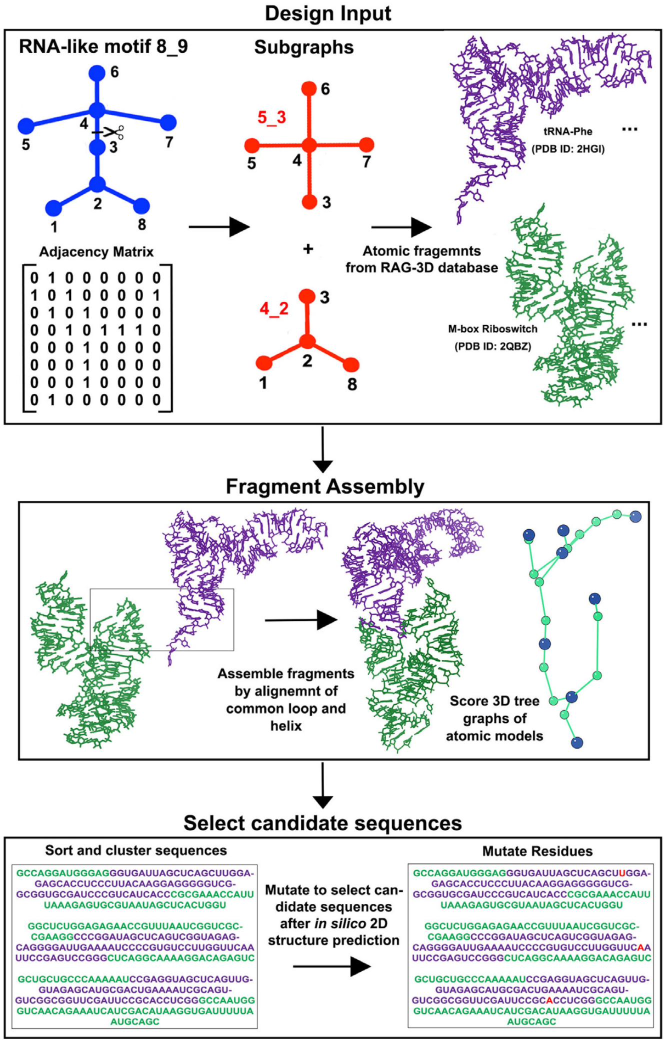
Various computational algorithms have also been developed to tackle the RNA inverse folding problem, i.e. design an RNA sequence that folds onto a target secondary (2D) structure. The pioneering RNAInverse program (ViennaRNA package) (31) randomly samples mutations and accepts one that bring the sequence closer to the target 2D structure. Other programs include RNA-SSD (32), which performs a hierarchical decomposition of the target structure followed by a local stochastic search; INFO-RNA (33), employing dynamic programming and probabilistic sampling of sequences; DSS-OPT (34), involving Newtonian dynamics in sequence space; NUPACK-Design (35,36), which calculates partition functions over the equilibrium ensemble, recently updated for multiple states; MODENA (37), which uses a multi-objective genetic algorithm to select sequences for both structure stability and similarity to the target structure; and design algorithms like RNAexinv, RNAfbinv, and IncaRNAfbinv (38–40) which use simulated annealing to design RNA shapes with additional physical constraints like thermodynamic stability, mutational robustness, and sequence. The EteRNA (41) open laboratory initiative was launched with the related goal to address issues on viability of 2D structures for design, to involve the broad community in RNA structural design, and provide feedback through laboratory experiments (42). In this paper, we present a computational pipeline for *in silico* design of novel RNA topologies using our RNA-As-Graphs (RAG) approach in combination with recently developed tools for graph-partitioning and fragment assembly.

RNA 2D structures have been described by graphs since the 1970s and 1980s by Waterman (43), Nussinov (44,45), Shapiro (46), and others (see reviews in (47–49)). Our RAG approach offers a systematic way to represent RNA 2D structures as planar, undirected tree and dual graphs (50). Such simplified representations reduce the conformational search space drastically, and allow us to study RNA structure using machinery in graph theory, like graph-isomorphism, partitioning, and enumeration (51). RAG has been successfully applied to computationally model the *in vitro* selection process of RNA molecules (52,53), develop a hierarchical graph-sampling methodology to predict RNA 3D graph topologies (RAG-TOP) (54–56), create a database of RNA structures and substructures (57) using graph-partitioning algorithms (58), and develop a fragment-assembly based approach called *F-RAG* to generate atomic models from candidate RNA 3D tree graphs (59).

The coarse-grained graph representations of RNA 2D structures facilitate the study of many possible RNA motifs and topologies. Importantly, we can enumerate the possible topologies and connectivities of a tree or a dual graph for a given number of vertices (60,61). Based on the features and characteristics of existing RNAs, we have classified (using clustering techniques) RNA tree graph topologies as 'existing', RNA-like, and non RNA-like (61,62). RNA-like motifs are 2D tree graph topologies more likely to correspond to RNA 2D structures that have not yet been discovered, and non RNA-like motifs are graph topologies that are less likely to be found in Nature. The RNA-like graphs are closely related (e.g. by an additional junction or loop) to existing topologies, while the non RNA-like are more different (e.g. 'asterisk' like graphs where junctions or loops emanate from a central point, see black motifs in Figure 3 later). Out of the 10 novel RNA-like topologies predicted in 2004 (61), at least five have since been solved (63). A more recent assessment (62) shows that our classification holds promise, since many more RNA-like compared to non RNA-like topologies have been solved since our last classification. Such RNA-like topologies are thus ideal candidates for RNA design. Such a modular approach of designing sequences corresponding to RNA-like topologies could allow for greater flexibility in the design process, and generate more viable sequences. It may potentially help generate novel RNA motifs systematically.

Here, we use our computational pipeline to design novel sequences and 3D folds corresponding to RNA-like graph topologies. As sketched in Figure 1, we start by partitioning the RNA-like target tree graph into subgraphs (using our graph partitioning (58)), and extracting the corresponding atomic fragments from our database of existing RNA substructures (RAG-3D database (57)). We then use our fragment assembly approach (similar to F-RAG developed for structure prediction (59)) to construct a complete 3D atomic model from the atomic substructures corresponding to the subgraphs. We score the generated models using our knowledge-based statistical potential developed for 3D structure prediction (54,56), and screen the sequences of the top scoring models further for their intended fold using the two 2D structure prediction programs RNAfold (available from the ViennaRNA package) (64) and NUPACK (65–67). We also subject representative sequences to manual mutations using EteRNA's puzzle maker interface to test for mutational robustness and improve our yield. Our intention here is to show how a systematic design could be pursued, and how mutations can be applied to assess or improve the results. We also incorporate specific structural motifs, like *k*-turns, to test the ability of our design software to incorporate user specified motifs into the design for additional stability or tailored functionality.

We apply this design methodology to derive sequences for six selected RNA-like topologies, namely 7_4, 8_4, 8_6, 8_7, 8_9 and 8_12 (see Figures 3 and 6 later), with and without an additional restriction of a *k*-turn motif. Results show promise for designing novel RNA topologies (or motifs) systematically and highlight the flexibility in the target 2D structure and sequence length. Furthermore, our fragment-assembly approach generates a large number of candidate sequences that can then be further screened. To test the accuracy of our predictions, we subject two representative designed sequences to experimental structure probing using SHAPE-MaP (68,69). Overall, the experimental data are in good agreement with our computational design predictions. We propose that such a pipeline could be applied to all RNA-like topologies to produce a library of novel RNA motifs. Designing RNA sequences that will fold onto RNA-like topologies can help increase our understanding of RNA structural features and explore potentially novel functions.



Adjacency Matrix

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Sort and cluster sequences

```
GCCAGGAUGGGAGGGUGAUUAGCUCAGCUUGGA-
GAGCACCUCCCUUACAAGGAGGGGGUCG-
GCGGUGCGAUCCCGUCAUCACCCGCGAAACCAUU
UAAAGAGUGCGUAAUAGCUCACUGGU

GGCUCUGGAGAGAACCGUUUAAUCGGUCGC-
CGAAGGCCCGGAUAGCUCAGUCGGUAGAG-
CAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCAA
UUCCGAGUCCGGGCUCAGGCAAAAGGACAGAGUC

GCUGCUGCCCAAAAAUCCGAGGUAGCUCAGUUG-
GUAGAGCAUGCGACUGAAAAUCGCAGU-
GUCGGCGGUUCGAUUCCGCACCUCGGGCCAAUGG
GUCAACAGAAAUCAUCGACAUAAGGUGAUUUUUA
AUGCAGC
```

Mutate Residues

```
GCCAGGAUGGGAGGGUGAUUAGCUCAGCUUGGA-
GAGCACCUCCCUUACAAGGAGGGGGUCG-
GCGGUGCGAUCCCGUCAUCACCCGCGAAACCAUU
UAAAGAGUGCGUAAUAGCUCACUGGU

GGCUCUGGAGAGAACCGUUUAAUCGGUCGC-
CGAAGGCCCGGAUAGCUCAGUCGGUAGAG-
CAGGGGAUUGAAAAUCCCCGUGUCCUUGGUUCAA
UUCCGAGUCCGGGCUCAGGCAAAAGGACAGAGUC

GCUGCUGCCCAAAAAUCCGAGGUAGCUCAGUUG-
GUAGAGCAUGCGACUGAAAAUCGCAGU-
GUCGGCGGUUCGAUUCCGCACCUCGGGCCAAUGG
GUCAACAGAAAUCAUCGACAUAAGGUGAUUUUUA
AUGCAGC
```

Figure 1. Flow chart of the pipeline to design sequences to fold onto novel RNA-like topologies. Design input: RNA-like topology, adjacency matrix, subgraphs, and corresponding atomic fragments from the RAG-3D database. The 8_9 graph topology can be partitioned into two subgraphs with RAG IDs 5_3 and 4_2, respectively. Atomic fragments corresponding to the two subgraphs are used to build the atomic model for the RNA-like graph topology 8_9 using fragment assembly. The atomic models are scored based on the statistical potential, and top sequences are selected based on *in silico* 2D structure prediction and mutational analysis.

MATERIALS AND METHODS

This section summarizes components of our computational pipeline used to design sequences that fold onto novel RNA-like topologies. Details are provided in Section S1 of the Supplementary data.

RAG tree graphs representation, classification, and partitioning

The RAG approach represents any 2D RNA structure as a planar, undirected, connected *tree* graph (50). Each vertex of the graph corresponds to a loop (single-stranded regions), and each edge corresponds to a helix (double-stranded stem) connecting the two loop vertices. This graph is a 2D representation of connectivity of the 2D structural elements of an RNA molecule (Figure 2B). To incorporate size and build 3D objects for our sampling approach RAGTOP (54), we convert this 2D graph into a 3D graph (Figure 2C) by adding vertices and scaling edges, as described in (54). A 3D tree graph can also be constructed from a given RNA 3D structure (Figure 2D), as specified in (54). Building 3D graphs from 3D structures allows us to score 3D atomic models using our statistical potential.

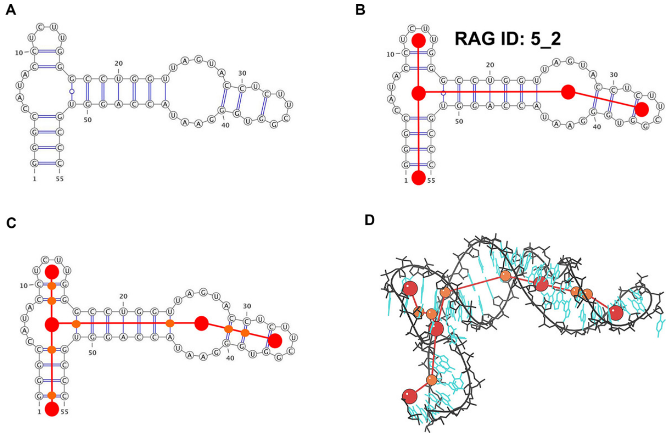


Figure 2. RAG tree graphs in 2D and 3D for a 5S rRNA fragment (PDB ID: 2HGH). (**A**) Secondary structure, (**B**) corresponding RAG 2D tree graph, (**C**) associated RAG 3D tree graph constructed from the 2D tree graph by adding additional vertices at helical ends, and respective edges, and (**D**) the 3D tree graph can also be constructed from the experimentally solved tertiary structure.

We use graph enumeration methods to generate possible tree graphs for a given number of vertices (60,61). The RAG database currently characterizes different tree graph topologies up to 13 vertices (62). Tree graph topologies are given unique RAG IDs. The graphs associated with known RNA structures are classified as 'existing RNA'. The remaining hypothetical graphs are classified as RNA-like, or non RNA-like as trained by known RNAs using the clustering algorithm PAM (Partitioning Around Medoids) (61). Figure 3 shows a sample of topologies from the RAG resource, classified as existing (red), RNA-like (blue) and non RNA-like (black).

To study the 2D structure submotifs of an RNA structure, we partition the 2D tree graph into topologically distinct subgraphs (58). Figure 4 shows the partitioning of the structure of the TPP riboswitch (PDB ID: 3D2G), and its various subgraphs and corresponding atomic fragments. Graph partitioning was applied to ≈1500 representative RNA structures to create a database of RNA structures and substructures called *RAG-3D* (57). The RAG-3D database catalogs atomic fragments associated with 51 different RAG IDs which we use in our fragment assembly procedure to design sequences for RNA-like topologies, as described next.

Fragment assembly procedure for design

Essentially, we partition the RNA-like graph topologies into subgraphs, and then obtain atomic fragments for each of these subgraphs from the RAG-3D database. Next, we

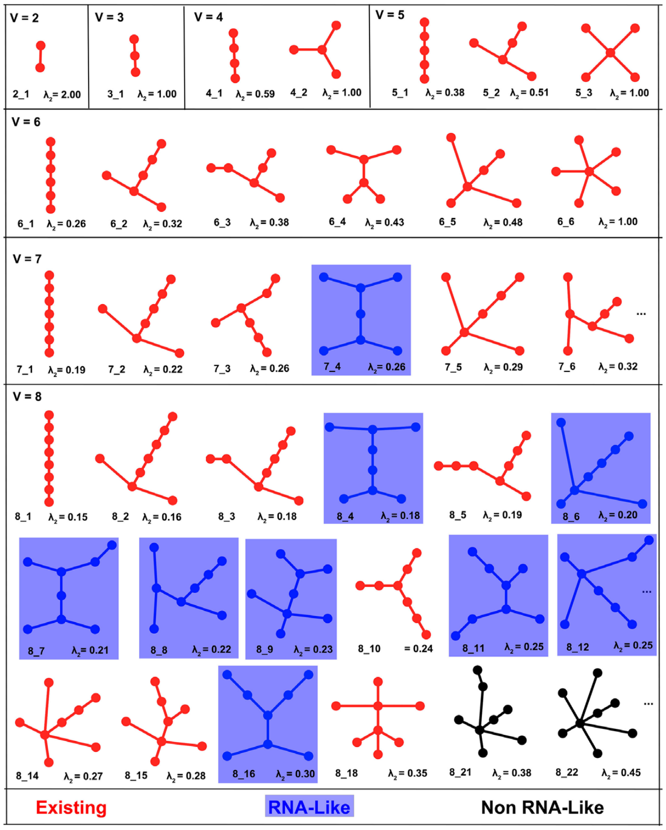
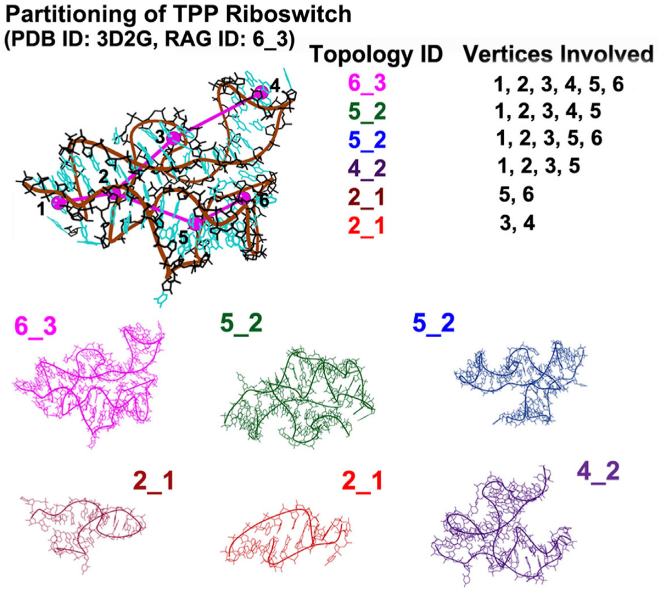


Figure 3. Sample subset of the RAG topologies catalogue (see full catalog in www.biomath.nyu.edu/?q=rag/tree_vertices.php). Topologies for a given number of vertices are classified as existing (red), RNA-like (blue) and non RNA-like (black), by a clustering approach (62). There are about ≈2300 such tree structures in the actual database up through 13 vertices. Among them, around 80 are existing motifs and ∼1600 of the remaining are hypothetical motifs classified as RNA-like.



Topology ID	Vertices Involved
6_3	1, 2, 3, 4, 5, 6
5_2	1, 2, 3, 4, 5
5_2	1, 2, 3, 5, 6
4_2	1, 2, 3, 5
2_1	5, 6
2_1	3, 4

Figure 4. Subgraphs of an RNA molecule. Subgraphs, their RAG IDs, and the corresponding atomic coordinates for the structure of the TPP riboswitch are shown as produced by RAG-3D (PDB ID: 3D2G).

piece together subgraphs by fragment assembly of corresponding atomic fragments to build 3D models and predict sequences to fold onto the given target. We have recently reported development of this fragment assembly based approach, *F-RAG*, to build atomic models for candidate RNA 3D tree graph topologies for RNA structure prediction (59). We use a modified F-RAG procedure in this paper for designing sequences for target RNA graph topologies.

Our automated procedure requires as input the following pieces (see Figure 1 for an example of various inputs for target graph 8_9): (i) target graph topology for design, with the order of the vertices specified in the 5′ to the 3′ direction; (ii) adjacency matrix A corresponding to the target graph; (iii) number of subgraphs of the target graph, along with their RAG IDs and vertices; (iv) list of RNA substructures corresponding to each of the target's subgraph RAG IDs (in the RAG-3D database), along with their 2D structure and atomic fragments and (v) loop number, along with a given 2D structure and atomic fragment of a specific motif, if an internal loop or a hairpin needs to be restricted to the given motif.

The design algorithm employs a recursive procedure for each subgraph of the target graph (starting from the 5′ direction) to generate atomic coordinates for that subgraph. For each subgraph, the number and connectivity of RNA loops from the 5′ to the 3′ direction is compared to each atomic fragment, and any mismatched fragments are eliminated. The remaining fragments for each subgraph are superimposed on the partially built atomic model from previous subgraphs (as part of the recursive procedure) using the residues in the common helix connecting the two subgraphs. For specified motif design, atomic fragment for the specified motif is used for that loop. The number and identity of the bases in the atomic fragments are left unchanged. Unpaired residues at the 5′ or 3′ ends of the sequence are removed. Once atomic coordinates are generated for all subgraphs, a 3D tree graph is calculated corresponding to the atomic model (as described in subsection S1.1 'RAG 2D and 3D tree graphs' in the Supplementary data). This 3D graph is scored using our knowledge-based statistical potential (initially developed for RNA structure prediction (56)). Atomic models with corresponding sequences, 2D structures, 3D tree graph, and scores are produced as output of the fragment assembly.

Selecting candidate sequences for target graph

The above fragment assembly procedure is applied for different orientations of the target graph. Each orientation represents a different 5′ to 3′ order of the RNA loops for the same 2D target graph topology. To sort through the large number of generated sequences and identify candidates quickly, we combine resulting sequences from every orientation of the target graph and order them by increasing score. Any model with large chain breaks is removed, and the top 200 models with unique sequences are retained for further analysis. These top 200 sequences are clustered based on the type of RNA origin of atomic fragments.

To further narrow the pool of candidate sequences, we subject the top 200 unique sequences selected above to RNAfold (available with the Vienna RNA Package

2.3.3) (64) and NUPACK (65–67) for *in silico* 2D structure prediction, with default parameters. That is, for each sequence, we identify the 2D graph topologies of the minimum energy and the centroid 2D structure produced by RNAfold, and the minimum energy structure produced by NUPACK. We consider a design successful if our sequence folds onto the same RAG topology with both RNAfold (either the minimum energy or the centroid structure) and NUPACK, regardless of whether this was the intended fold or not (since another fold might be produced). For all our top 200 candidate sequences, we classify the number of sequences that fold onto different RAG topologies, and select the sequences that are predicted to fold onto the target graph for further study.

As an additional step to test the robustness of the successfully designed sequences (by the criteria above) that fold onto the target graph, we subject them to manual mutations. Specifically, we use EteRNA's (41) puzzle-maker interface to perform mutations on our designed sequences. EteRNA uses the same software for 2D structure prediction (RNAfold and NUPACK) as done here, with a useful graphical interface and real time feedback. We select the top sequences for a target RNA-like topology based on their score, most productive clusters, and the above robustness test.

Experimental structure probing via SHAPE

As a proof of principle, we used SHAPE-MaP (Selective 2′-hydroxyl acylation analyzed by primer extension) with next-generation sequencing (69) to probe the structures of two candidate design sequences (namely 1b and 1d, as marked in Figure 6 later with stars). SHAPEMapper pipeline was used to evaluate mutations and obtain SHAPE data (70,71). The structures were folded based on normalized SHAPE reactivities using RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) (72) and RNAstructure (https://rna.urmc.rochester.edu/RNAstructure.html) (73), both with default parameters for the pseudo-free energy term, as described in (74). Structures were visualized in VARNA (75). See full details in subsection S1.7 of the Supplementary data.

RESULTS

Overview

Below we present the results of our computational pipeline to design sequences for six selected RNA-like topologies. The target topologies are shown in Figure 5. For the fragment-assembly procedure, each target graph is divided into two subgraphs, and two separate designs were performed for each target graph: an initial non-specific design, and a second with the specific requirement that the connecting internal loop between the two subgraphs be a *k*-turn motif. The latter showcases the ability of our design software to incorporate user given motifs in the design process for additional stability or functionality. Our design pipeline can work with any user given motif for internal loops and hairpins (as long as those loops are at ends of a subgraph). We choose the *k*-turn motif as a representative motif as it is a well studied motif with special features that we believed would be helpful in stabilizing a 2D structure.

For each design, the fragment-assembly procedure is run separately for each orientation illustrated in Supplementary Figure S1 (two different orientations for 7_4 and 8_4 topologies, four different orientations for 8_6, 8_7 and 8_12 topologies and five different orientations for the 8_9 topology). The different orientations of the target graphs shown in Supplementary Figure S1 were manually constructed by treating each terminal vertex of the graph in turn as vertex 1 (that corresponds to the 5′/3′ end of the sequence). Each orientation corresponds to a different 5′ to the 3′ order of loops for the same RAG topology. The 5′ to 3′ order of loops in the subgraphs of the target graph are also different for each orientation. It is possible that no atomic fragments in our RAG-3D database match the order of loops in the subgraphs of certain orientations. Therefore, our design algorithm does not produce models for such orientations.

We perform all designs on our local linux cluster (six nodes, each node with two Intel 2.67 GHz processors and 24GB RAM). For the six design targets, each run of the fragment assembly requires ≈15 min to 3.5 h depending on the number of atomic fragments and models generated. The number of models generated ranges from ≈20 (e.g. for orientation 1 of 8_7) to ≈8500 (e.g. for orientation 4 of 8_6).

For each target, our design pipeline can quickly generate hundreds of corresponding sequences because there are many existing RNA representatives for the various subgraph components. We score the atomic models using our statistical potential (as described in subsection titled 'Fragment-assembly procedure for design'), combine the results from orientations that produce models, and retain the top 200 unique candidates for each design for further analysis.

Analyzing and organizing the resulting candidates required us to develop further ways to examine them and select the most promising candidates. This is because our fragment assembly construction does not guarantee that the resulting fold is indeed the target fold, or even the major fold in the candidate pool. Thus, we cluster the top 200 candidates based on the RNA origins of the fragments used, and subject them to RNAfold and NUPACK (as described in subsection titled 'Selecting candidate sequences for target graph') and report the number of candidates for which both programs predict the same fold in Table 1, whether that was the intended fold or not. Table 1 also lists how many of these candidates are predicted to have the same fold as the target graph.

As we see, many candidates provide the target fold for five of the six target topologies (fewer for the sixth). The distributions of the major motifs obtained for each design are shown in Figure 5 (see Tables S1–S12 in Supplementary data for details of the sequence clusters and a full list of different motifs obtained from each design). Overall, we generally obtain a number of sequences for the target graphs, and requiring a *k*-turn connective loop (red histograms in the plots of Figure 5) can increase the overall yield. Interestingly, many of the other motifs generated (accidentally) in our design pipeline are existing (red) or RNA-like (blue), but very few non RNA-like (black). These results suggest an intriguing design protocol that demands further experi-

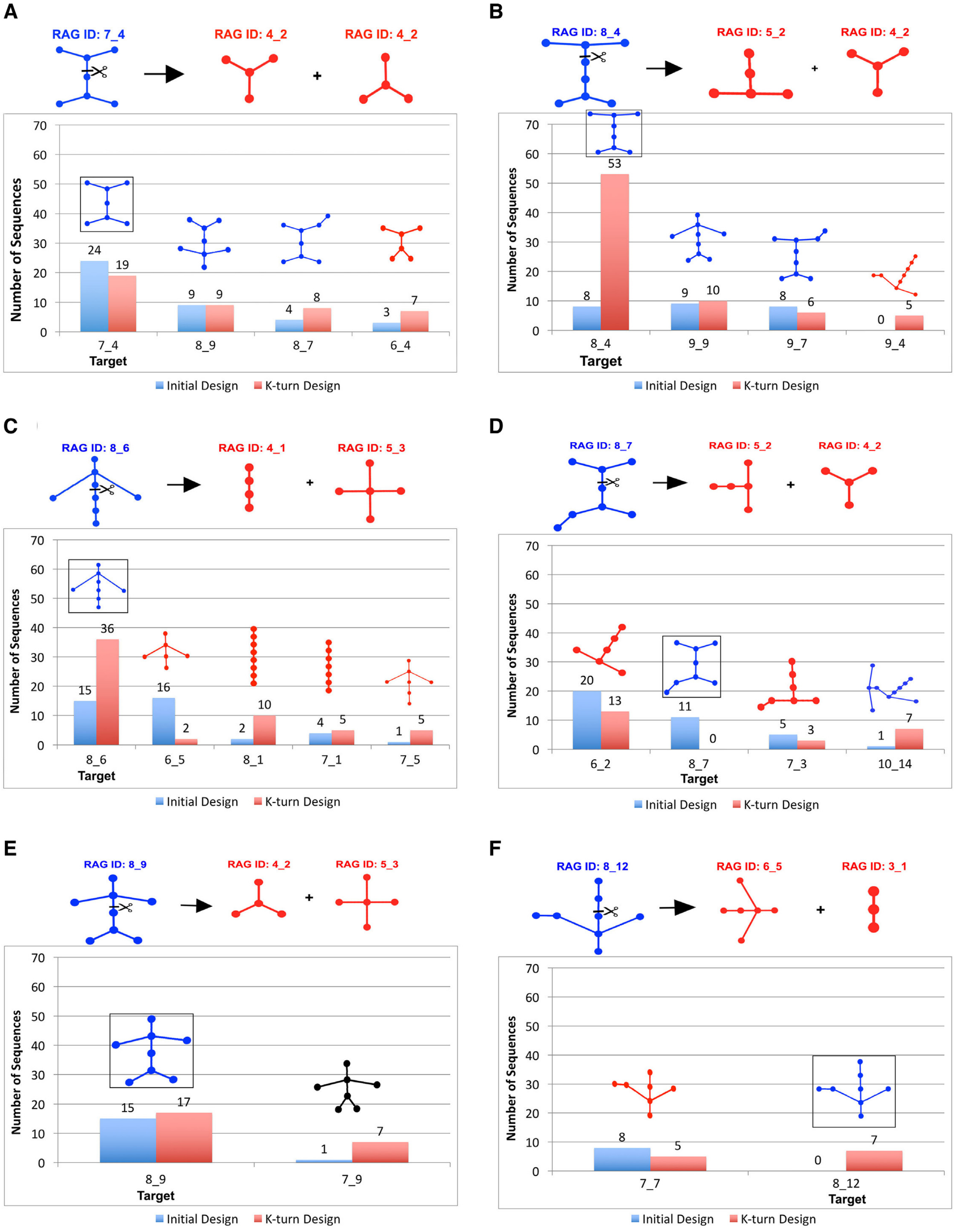


Figure 5. Distribution of sequences resulting from the design for the six RNA-Like motifs as predicted by RNAfold and NUPACK only when they coincide. The target is highlighted in the black box. Red, blue, and black colors are used for existing, RNA-like, and non RNA-like motifs respectively as classified by RNA clustering (62). Only topologies that have five or more sequences, in either the initial non-specific design or the specific *k*-turn design, are shown.

Table 1. Number of sequences out of the top 200 that are predicted to fold onto the same RAG topology ('Same fold') and the target RNA-like motif ('Target fold') with both RNAfold and NUPACK, for both the initial non-specific and specific *k*-turn design runs. The top RAG topologies obtained for the 'Same fold' group are shown in Figure 5 for each target. Refer to Tables S1-12 in the Supplementary data for obtained sequence clusters and the full distribution of RAG topologies

	Number of sequences			
	Initial design		*k*-turn design	
Target RNA-like motif	Same fold	Target fold	Same fold	Target fold
7_4	58	24	66	19
8_4	37	8	85	53
8_6	63	15	76	36
8_7	54	11	38	0
8_9	48	15	59	17
8_12	17	0	21	7

mental testing. Below we summarize results for each target including the most productive 'fragments' for each.

Obtained topologies and clusters

RNA-like motif 7_4 (Figure 5A, Supplementary Table S1). Of the top 200 unique sequences, 58 are predicted to fold onto the same RAG topology by both RNAfold and NUPACK. Around 41% of those (24) fold into the target 7_4 motif. Of the 24 sequences, 13 are from orientation 1 and 11 from orientation 2. The most productive design comes when atomic fragments from ribosomal RNAs (42% of the 24 sequences) and glycine riboswitches (42%) are used as the first fragment (RAG ID: 4_2), and from purine riboswitches (38%) and tetracycline riboswitches (25%) as the second fragment (RAG ID: 4_2).

Following the target 7_4 motif, the 8_9 RNA-like topology is produced by nine sequences. In contrast to the 7_4 sequences, the 8_9 sequences use exclusively tRNA structures for the second fragment. Topologically, the 7_4 and 8_9 tree graphs differ in the second subgraph: it is 4_2 (three-way junction) in 7_4, and 5_3 (four-way junction) in the 8_9 motif; and this difference is reflected in the preference for the second fragments.

RNA-like motif 8_4 (Figure 5B, Supplementary Table S2). Of the top 200 unique sequences, 37 are predicted to fold onto the same RAG topology by both RNAfold and NUPACK. Only eight of those fold onto the target 8_4 motif. All eight sequences are from orientation 1, and contain atomic fragments from glycine riboswitches for the first fragment (RAG ID: 5_2) for all sequences, and purine riboswitches and tetracycline riboswitches for the second fragment (RAG ID: 4_2) for three sequences each.

Apart from the target 8_4 motif, 9_9 and 9_7 RNA-like topologies result, with 9 and 8 sequences respectively. Both the 9_9 and 9_7 tree graphs differ from 8_4 in the second subgraph: it is 4_2 in 8_4, 5_3 in 9_9 and 5_2 in the 9_7 motif. However, these differences are reflected in the preferences for the second fragment only between 8_4 and 9_9 motifs. The 9_9 sequences prefer tRNA structures for the second fragment (Cluster 1), but the 9_7 sequences also work with purine riboswitches like 8_4. The 9_7 sequences prefer exclusively ribosomal RNAs for the first fragment.

RNA-like motif 8_6 (Figure 5C, Supplementary Table S3). Of the top 200 unique sequences, 63 are predicted to fold onto the same RAG topology by both RNAfold and NUPACK. However, only 15 of those fold onto the target 8_6 motif. All 15 sequences are from orientation 4. The target motif prefers atomic fragments from Myotonic Dystrophy RNAs (73% of the 15 sequences) as the first fragment (RAG ID: 4_1), and exclusively tRNA structures as the second fragment (RAG ID: 5_3).

Apart from the target 8_6 motif, the top RAG topology obtained is 6_5 graph topology with 16 sequences. Tree graphs 8_6 and 6_5 differ in the first subgraph (4_1 in 8_6 and 2_1 in the 6_5 motif), and this is reflected in the preference for structures for the first fragment between 8_6 and 6_5 sequences (most notable in Clusters 1 and 2), with 6_5 sequences preferring SAM I riboswitches.

RNA-like motif 8_7 (Figure 5D, Supplementary Table S4). Of the top 200 unique sequences, 54 are predicted to fold onto the same RAG topology by both RNAfold and NUPACK. Of those, 11 fold onto the target 8_7 motif. Of the 11 sequences, 10 are from orientation 3. The 8_7 sequences come mainly from Cluster 1 that uses fragments from TPP riboswitch structures for the first fragment (RAG ID: 5_2) and purine riboswitches for the second fragment (RAG ID: 4_2).

Apart from the target 8_7 motif, the top RAG topology obtained is 6_2 graph topology with 20 sequences. However, there are no differences in fragments preferences between 8_7 and 6_2 sequences, with both sequences coming from almost the same clusters.

RNA-like motif 8_9 (Figure 5E, Supplementary Table S5). Of the top 200 unique sequences, 48 are predicted to fold onto the same RAG topology by both RNAfold and NUPACK, but only 15 fold onto the target 8_9 motif, and there are no other significant topologies that emerge. Of the 15 sequences, 13 are from orientation 5. The most productive design comes when atomic fragments from ribosomal RNAs (60% of the 15 sequences) and M-box riboswitches (27%) are used as the first fragment (RAG ID: 4_2), and exclusively tRNA structures as the second fragment (RAG ID: 5_3).

RNA-like motif 8_12 (Figure 5F, Supplementary Table S6). The design of this RNA-like motif produced the least number of candidates, possibly due to limited number of atomic fragments available for its 6_5 subgraph. Only 17 of the top 200 unique sequences are predicted to fold onto the same RAG topology by both RNAfold and NUPACK, but none of those produce the target 8_12 motif. Most sequences fold onto the RAG topology 7_7. Atomic fragments from SAM-I riboswitches correspond to the first fragment (RAG ID: 6_5) for all sequences, but there is considerable variation for the second fragment (RAG ID: 3_1). This is not surprising as 3_1 is one of the most common subgraph in our RAG-3D database.

Incorporation of *k*-turn motif as a constraint

Our second design run where the connecting loop between each pair of subgraphs was restricted to be a *k*-turn motif demonstrates our ability to incorporate specific submotifs. The number of sequences that fold onto the target topology increases for four of the six RNA-like motifs (Table 1), decreases slightly for the 7_4 motif, and are completely eliminated for the 8_7 motif. Overall, the total number of sequences that fold onto the target topologies for all six RNA-like motifs almost doubles from 73 to 132. The major RAG topologies obtained in these designs for each RNA-like motif are shown as red histograms in Figure 5, with sequence clusters listed in Tables S1-S6 in the Supplementary data.

The *k*-turn design had the most significant impact on the design results of 8_4 and 8_6 RNA-like topologies. The number of sequences that fold onto the target 8_4 motif increases from 8 in the initial non-specific design to 53 in the *k*-turn design (Table 1), with the increase noted mainly in Clusters 1, 4 and 6 (Supplementary Table S2 in the Supplementary data), making results of the *k*-turn design more productive for the 8_4 motif. Of the 53 8_4 sequences, 44 are from orientation 1, and remaining 9 from orientation 2. The number of sequences that fold onto the unintended 9_7 and 9_9 byproduct however remain almost the same (Figure 5B).

Similarly for the 8_6 RNA-like motif, the number of sequences that fold onto this target more than doubles to 36, and the alternate 6_5 RAG topology is almost eliminated by the *k*-turn design (Figure 5C). However, the sequences for 8_1 and 7_5 topologies increase in the *k*-turn design. Orientation 1 now also contributes 13 of the 36 sequences, with the remaining 23 coming from orientation 4. Four new clusters now emerge for the 8_6 constrained design (Cluster 5, 6, 7 and 8 in Supplementary Table S3 in the Supplementary data). There is more variety in the preference for structures of the first fragment: the contribution of Myotonic Dystrophy RNAs (Cluster 1) is reduced, and SRP structures (Cluster 7) and ribosomal RNAs (Cluster 4) emerge (Supplementary Table S3 in the Supplementary data).

For the 8_12 motif, the *k*-turn design results in seven sequences that fold onto the target motif (as opposed to none with the initial non-specific design). All seven sequences are from orientation 1. The *k*-turn design also leads to decrease in the number of unintended 7_7 sequences (Figure 5F).

The number of sequences that fold onto the target 8_9 topology increases slightly to 17 with the *k*-turn design, and are now more concentrated in Cluster 1 (Supplementary Table S5 in the Supplementary data). The orientation preference remains the same, with 15 of the 17 sequences coming from orientation 5. The *k*-turn design also leads to increase in the number of unintended 7_9 sequences.

For the 7_4 motif, the number of sequences that fold onto the target decreases from 24 to 19, though the preference for the first fragment changes. The first fragment now comes mainly from ribosomal RNAs, while the most productive clusters with the glycine riboswitch disappear (Supplementary Table S1 in the Supplementary data). The preference for the orientation also changes, with 14 of the 19 sequences now coming from orientation 2. The *k*-turn design also results in increase in the sequences for 8_7 and 6_4 topologies, while the sequences for the 8_9 topology remain the same.

The *k*-turn design negatively impacts the results of the 8_7 motif: the number of sequences that fold onto the target decreases from 11 to none, and there is a significant increase in the sequences folding onto the 10_14 RAG topology (Figure 5D).

Mutational analysis of candidate sequences

Overall, a large number of candidate sequences were obtained quickly, indicating that our computational pipeline has the potential for success in designing candidate sequences that fold onto the target RNA-like topologies.

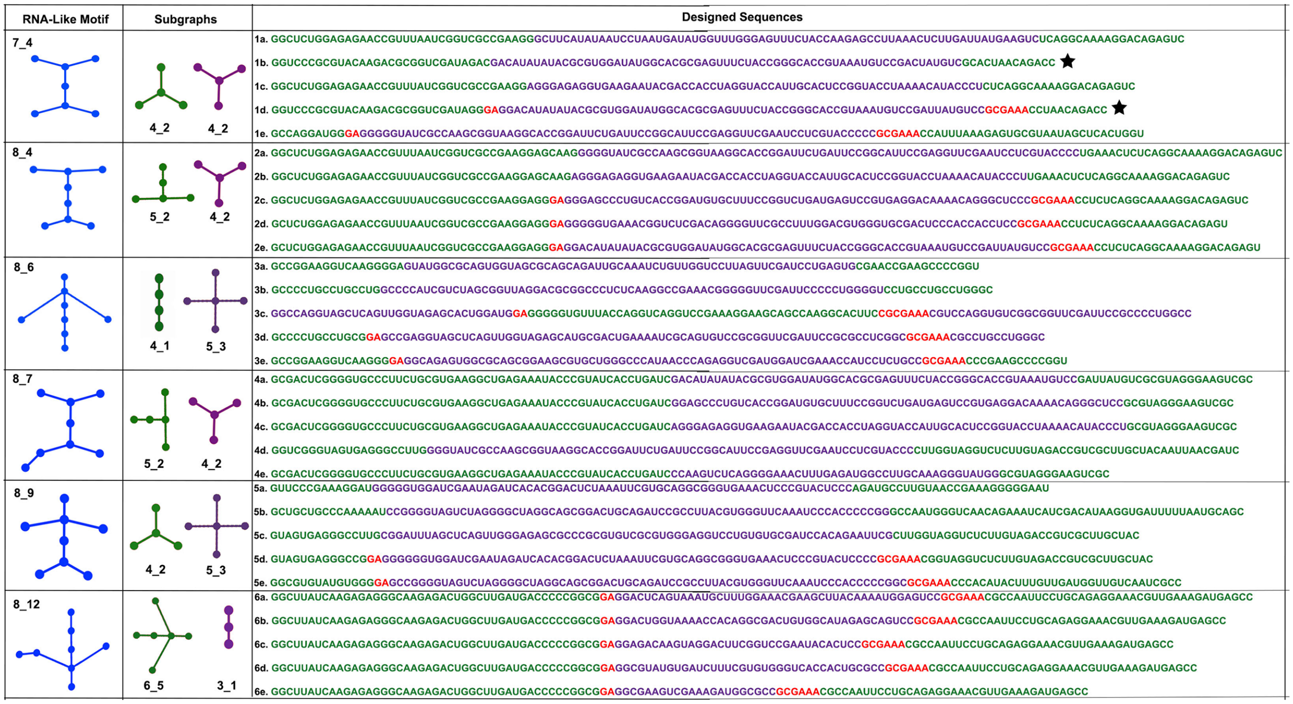
As an additional step to test the robustness of the designed sequences '*in silico*', we subject the top sequences that fold onto the target topology to manual mutations and compensatory mutations with EteRNA (41) (as described in subsection S1.6 'Mutational analysis for robustness and increased yield' in the Supplementary data). For the six RNA-like topologies, with both the initial and the *k*-turn specific design, we select top sequences from clusters with sequences that fold onto respective target topologies for manual analysis. Based on the most productive clusters and our robustness analysis, we identify five sequences most likely to fold onto each of our six target RNA-like topologies in Figure 6. See Supplementary data (Supplementary Figures S3–S8, and Tables S13–S22) for sequence identities, nucleotide distributions, specificity, and sensitivity values of the top sequences.

In addition, we also subjected top sequences that did not fold onto the target topologies to mutations to fold onto the target topologies. We selected top sequences from the most productive clusters that fold onto the target topologies with only one of RNAfold or NUPACK (hence are not considered 'successful') and mutated them with EteRNA (as described in subsection S1.6 'Mutational analysis for robustness and increased yield' in the Supplementary data). For *k*-turn design sequences, mutations were also performed to ensure that the *k*-turn motif is preserved. For the total of 12 such sequences selected, seven sequences required one mutation, three sequences required two mutations and remaining two sequences required four mutations. We report these as additional sequences that fold onto the target topologies. See Supplementary Table S23 in the Supplementary data for a description of these sequences.

Experimental probing of representative sequences

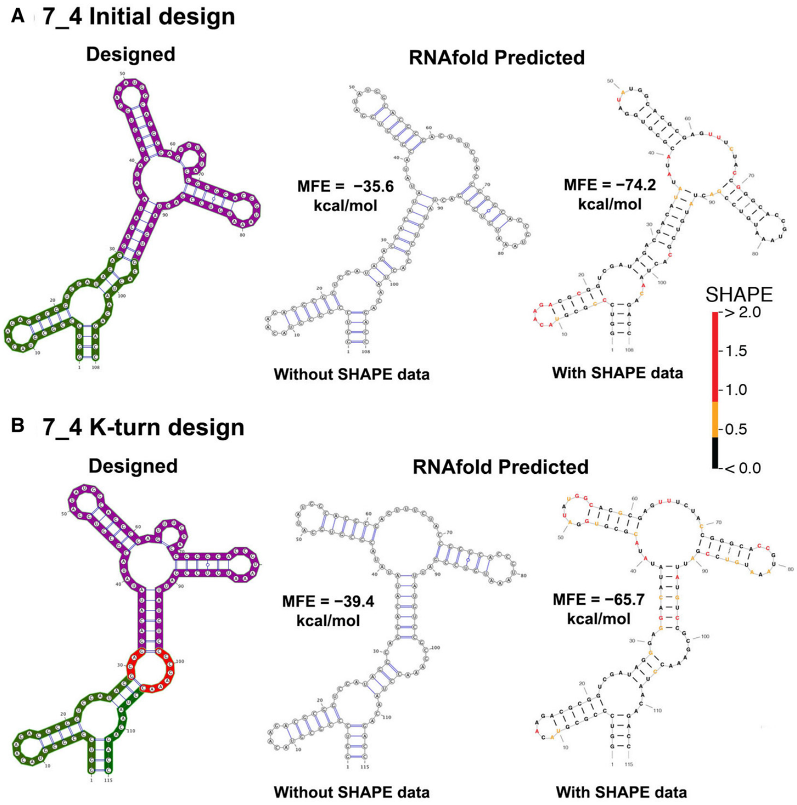
To explore the potential application of our pipeline, we also pursued experimental probing by SHAPE of two specific design sequences, 1b and 1d in Figure 6 (black stars) corresponding to the 7_4 RNA-like topology. Our design candidates were chosen based on score and manual check for mutational robustness. Both sequences were selected as the top scoring initial (1b) and *k*-turn (1d) design sequence from Cluster 1 (Supplementary Table S1 in the Supplementary data). SHAPE reactivities for both sequences were obtained as detailed in subsection S1.7 'Experimental structure probing via SHAPE' in the Supplementary data.

The 2D structures corresponding to the sequences as predicted using RNAfold with and without SHAPE experimental reactivities are shown in Figure 7. In both cases,



RNA-Like Motif	Subgraphs	Designed Sequences
7_4	4_2 4_2	1a. GGCUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGGCUUCAUAUAAUCCUAAUGAUAUGGUUUGGGAGUUUCUACCAAGAGCCUUAAACUCUUGAUUAUGAAGUCUCAGGCAAAAGGACAGAGUC 1b. GGUCCCGCGUACAAGACGCGGUCGAUAGACGACAUAUAUACGCGUGGAUAUGGCACGCGAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGUCGCACUAACAGACC ★ 1c. GGCUCUGGAGAGAACCGUUUAUCGGUCGCCGAAGGAGGGAGAGGUGAAGAAUACGACCACCUAGGUACCAUUGCACUCCGGUACCUAAAACAUACCCUCUCAGGCAAAAGGACAGAGUC 1d. GGUCCCGCGUACAAGACGCGGUCGAUAGGGAGGACAUAUAUACGCGUGGAUAUGGCACGCGAGUUUCUACCGGGCACCGUAAAUGUCCGAUUAUGUCCGCGAAACCUAACAGACC ★ 1e. GCCAGGAUGGGAGGGGGUAUCGCCAAGCGGUAAGGCACCGGAUUCUGAUUCCGGCAUUCCGAGGUUCGAAUCCUCGUACCCCGCGAAACCAUUUAAAGAGUGCGUAAUAGCUCACUGGU
8_4	5_2 4_2	2a. GGCUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGAGCAAGGGGGUAUCGCCAAGCGGUAAGGCACCGGAUUCUGAUUCCGGCAUUCCGAGGUUCGAAUCCUCGUACCCCUGAAACUCUCAGGCAAAAGGACAGAGUC 2b. GGCUCUGGAGAGAACCGUUUAUCGGUCGCCGAAGGAGCAAGAGGGAGAGGUGAAGAAUACGACCACCUAGGUACCAUUGCACUCCGGUACCUAAAACAUACCCUUGAAACUCUCAGGCAAAAGGACAGAGUC 2c. GGCUCUGGAGAGAACCGUUUAUCGGUCGCCGAAGGAGGGAGGGAGCCCUGUCACCGGAUGUGCUUUCCGGUCUGAUGAGUCCGUGAGGACAAAACAGGGCUCCCGCGAAACCUCUCAGGCAAAAGGACAGAGUC 2d. GCUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGAGGGAGGGGGUGAAACGGUCUCGACAGGGGUUCGCCUUUGGACGUGGGUGCGACUCCCACCACCUCCGCGAAACCUCUCAGGCAAAAGGACAGAGU 2e. GCUCUGGAGAGAACCGUUUAAUCGGUCGCCGAAGGAGGGAGGACAUAUAUACGCGUGGAUAUGGCACGCGAGUUUCUACCGGGCACCGUAAAUGUCCGAUUAUGUCCGCGAAACCUCUCAGGCAAAAGGACAGAGU
8_6	4_1 5_3	3a. GCCGGAAGGUCAAGGGGAGUAUGGCGCAGUGGUAGCGCAGCAGAUUGCAAAUCUGUUGGUCCUUAGUUCGAUCCUGAGUGCGAACCGAAGCCCCGGU 3b. GCCCCUGCCUGCCUGGCCCCAUCGUCUAGCGGUUAGGACGCGGCCCUCUCAAGGCCGAAACGGGGGUUCGAUUCCCCCUGGGGUCCUGCCUGCCUGGGC 3c. GGCCAGGUAGCUCAGUUGGUAGAGCACUGGAUGGAGGGGGUGUUUACCAGGUCAGGUCCGAAAGGAAGCAGCCAAGGCACUUCCGCGAAACGUCCAGGUGUCGGCGGUUCGAUUCCGCCCCUGGCC 3d. GCCCCUGCCUGCGGAGCCGAGGUAGCUCAGUUGGUAGAGCAUGCGACUGAAAAUCGCAGUGUCCGCGGUUCGAUUCCGCGCCUCGGCGCGAAACGCCUGCCUGGGC 3e. GCCGGAAGGUCAAGGGGAGGCAGAGUGGCGCAGCGGAAGCGUGCUGGGCCCAUAACCCAGAGGUCGAUGGAUCGAAACCAUCCUCUGCCGCGAAACCCGAAGCCCCGGU
8_7	5_2 4_2	4a. GCGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGAUCGACAUAUAUACGCGUGGAUAUGGCACGCGAGUUUCUACCGGGCACCGUAAAUGUCCGAUUAUGUCGCGUAGGGAAGUCGC 4b. GCGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGAUCGGAGCCCUGUCACCGGAUGUGCUUUCCGGUCUGAUGAGUCCGUGAGGACAAAACAGGGCUCCGCGUAGGGAAGUCGC 4c. GCGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGAUCAGGGAGAGGUGAAGAAUACGACCACCUAGGUACCAUUGCACUCCGGUACCUAAAACAUACCCUGCGUAGGGAAGUCGC 4d. GGUCGGGUAGUGAGGGCCUUGGGGUAUCGCCAAGCGGUAAGGCACCGGAUUCUGAUUCCGGCAUUCCGAGGUUCGAAUCCUCGUACCCCUUGGUAGGUCUCUUGUAGACCGUCGCUUGCUACAAUUAACGAUC 4e. GCGACUCGGGGUGCCCUUCUGCGUGAAGGCUGAGAAAUACCCGUAUCACCUGAUCCCAAGUCUCAGGGGAAACUUUGAGAUGGCCUUGCAAAGGGUAUGGGCGUAGGGAAGUCGC
8_9	4_2 5_3	5a. GUUCCCGAAAGGAUGGGGGUGGAUCGAAUAGAUCACACGGACUCUAAAUUCGUGCAGGCGGGUGAAACUCCCGUACUCCCAGAUGCCUUGUAACCGAAAGGGGGAAU 5b. GCUGCUGCCCAAAAAUCCGGGGUAGUCUAGGGGCUAGGCAGCGGACUGCAGAUCCGCCUUACGUGGGUUCAAAUCCCACCCCCGGGCCAAUGGGUCAACAGAAAUCAUCGACAUAAGGUGAUUUUUAAUGCAGC 5c. GUAGUGAGGGCCUUGCGGAUUUAGCUCAGUUGGGAGAGCGCCCGCGUGUCGCGUGGGAGGUCCUGUGUGCGAUCCACAGAAUUCGCUUGGUAGGUCUCUUGUAGACCGUCGCUUGCUAC 5d. GUAGUGAGGGCCGGAGGGGGGUGGAUCGAAUAGAUCACACGGACUCUAAAUUCGUGCAGGCGGGUGAAACUCCCGUACUCCCCGCGAAACGGUAGGUCUCUUGUAGACCGUCGCUUGCUAC 5e. GGCGUGUAUGUGGGGAGCCGGGGUAGUCUAGGGGCUAGGCAGCGGACUGCAGAUCCGCCUUACGUGGGUUCAAAUCCCACCCCCGGCGCGAAACCCACAUACUUUGUUGAUGGUUGUCAAUCGCC
8_12	6_5 3_1	6a. GGCUUAUCAAGAGAGGGCAAGAGACUGGCUUGAUGACCCCCGGCGGAGGACUCAGUAAAUGCUUUGGAAACGAAGCUUACAAAAUGGAGUCCGCGAAACGCCAAUUCCUGCAGAGGAAACGUUGAAAGAUGAGCC 6b. GGCUUAUCAAGAGAGGGCAAGAGACUGGCUUGAUGACCCCCGGCGGAGGACUGGUAAAACCACAGGCGACUGUGGCAUAGAGCAGUCCGCGAAACGCCAAUUCCUGCAGAGGAAACGUUGAAAGAUGAGCC 6c. GGCUUAUCAAGAGAGGGCAAGAGACUGGCUUGAUGACCCCCGGCGGAGGAGACAAGUAGGACUUCGGUCCGAAUACACUCCGCGAAACGCCAAUUCCUGCAGAGGAAACGUUGAAAGAUGAGCC 6d. GGCUUAUCAAGAGAGGGCAAGAGACUGGCUUGAUGACCCCCGGCGGAGGCGUAUGUGAUCUUUCGUGUGGGUCACCACUGCGCCGCGAAACGCCAAUUCCUGCAGAGGAAACGUUGAAAGAUGAGCC 6e. GGCUUAUCAAGAGAGGGCAAGAGACUGGCUUGAUGACCCCCGGCGGAGGCGAAGUCGAAAGAUGGCGCCGCGAAACGCCAAUUCCUGCAGAGGAAACGUUGAAAGAUGAGCC

Figure 6. Top five selected sequences that fold onto the target for the 6 RNA-like topologies. The initial design sequences are listed first, followed by the *k*-turn design sequences. The listing order within each category is in increasing score (lower scores are better). The nucleotides in the sequences are colored green and purple according to the subgraphs they belong to shown at left. Red nucleotides are part of the *k*-turn motif. The sequences indicated with the black star were selected for experimental probing with SHAPE.



A 7_4 Initial design

Designed

RNAfold Predicted

B 7_4 K-turn design

Designed

RNAfold Predicted

Figure 7. Analysis of 2D structures of two 7_4 candidate sequences using SHAPE data. For the 7_4 initial and *k*-turn design sequences (1b and 1d in Figure 6), we show in (**A**) and (**B**) the designed 2D structure and the predicted 2D structures with and without SHAPE data as determined by RNAfold, with default parameters. See Supplementary data for details. Note that the structures generated based on experimental SHAPE data are the same as the predicted 2D structure, except with even lower free energy; this supports the structural prediction pipeline.

the 2D structure predicted using SHAPE reactivities is the same as the one predicted without it. Furthermore, for both sequences, the SHAPE data produce structures with lower free energy (by ≈30 kcal/mol) while not altering the predicted structure. For both sequences, the predicted structures differ by only two base pairs (one missing and one extra) from the designed structures; these differences do not affect the overall RAG topology since the tree graph (design target) is the same. This suggests both strong agreement of the SHAPE data with the predicted structure and that the RNAs fold in solution. Of course, alternative folds based on SHAPE data cannot be ruled out. Interestingly, 2D structures predicted with decoys in the form of shuffled variations of the real SHAPE data differ from the 2D structure predicted without the SHAPE data for majority of the 100 shuffles considered (95% for 1b and 82% for 1d, see subsection S1.7 and Supplementary Figures S9 and S10 in the Supplementary data). None of the structures produced with the shuffled data have lower free energy than that with the real SHAPE data for sequence 1b, and only 14% have lower energies for sequence 1d. Note that the shuffled SHAPE data do not represent a random reactivity vector, so some false positives can be expected. Further experimental testing is limited by the high cost of the SHAPE-MaP next generation sequencing approach used here. However, these experimental tests and further computational experiments suggest strongly that our predicted sequences fold onto the intended topology and 2D structure.

DISCUSSION

Our developed computational pipeline to design sequences for candidate topologies that emerged as RNA-like from graph enumeration is based on our coarse-grained RNA-As-Graphs (RAG) approach. The pipeline uses tools for graph-partitioning (58), substructure matching (RAG-3D database (57)), and a fragment-assembly approach (similar to F-RAG (59)), followed by screening using *in silico* 2D structure prediction and targeted mutations. Our design strategy can also incorporate specific structural motifs (like *k*-turns) as desired. Our application to six selected RNA-like topologies, with and without a constraint for a *k*-turn, succeeds '*in silico*' to generate a number of candidate sequences for five of the six RNA-like topologies (fewer for one topology). Although the usage of highly valuable but naturally imperfect 2D structure prediction tools cannot demonstrate actual success *in vivo*, the consensus sequences that yield the same fold by more than one program appear quite promising. The SHAPE-MaP probing data (obtained for two candidate sequences) produce structures with lower free energy while not altering the predicted 2D structure; together with the fact that shuffled variations of the SHAPE data largely produce different 2D structures, these results indicate strong agreement with the our predicted structure. Of course, we cannot rule out alternate folds based on SHAPE data. Overall, our results reveals that our design pipeline has the potential for generating viable sequences that fold onto intended folds.

The advantage of our design pipeline is that our designed sequences target an RNA-like RAG topology, rather than a specific 2D structure. They can easily employ various RNA shapes or other constraints into the design through specified segments. Targeting an RNA-like topology can produce more viable sequences. Thus, our graph enumeration and clustering provide an automatic way to identify RNA-like motifs for design, and our graph partitioning, database of RNA substructures, knowledge-based potential, and fragment-assembly approach provide the needed tools to automate the inverse folding procedure. In addition, the user can choose to incorporate specific structural motifs for hairpins and internal loops (as we illustrate for *k*-turn motifs), and select specific atomic fragments for subgraphs for added stability or functionality. The resulting approach generates a large number of candidate sequences, can create quickly a library of different sequence and 2D structures for previously unexplored RNA topologies, and can produce plausible atomic models by fragment assembly of the candidate sequences.

Various improvements can be envisioned to enhance yield and accuracy. Our current code assembles atomic models by all combinations of matching atomic fragments for each subgraph. While this exhaustive approach generates many candidate sequences quickly, the computational time will increase significantly with the number of subgraphs. A more selective strategy for combining fragments may be needed for more than two subgraphs. For example, we divided the 8_12 RNA-like topology into two subgraphs: 6_5 and 3_1. While the RAG-3D database contains ∼900 representative fragments for RAG ID 3_1, the number for 6_5 fragments is only ∼60. Dividing the 6_5 subgraph further into 5_3 and 2_1 will lead to more input fragments and better candidate sequences. This will also require improvements to our scoring function to select for best hairpins and junction loops, and possibly a measure for 2D structure viability. Because our tree graphs cannot represent pseudoknots explicitly, pseudoknots are not considered in our analysis (although they may exist in our 3D atomic fragments and hence the plausible 3D model). Extending the different components of our pipeline to dual graphs will enable us to perform designs with more complex motifs; such work in underway (Jain, Bayrak, Petingi, and Schlick, 'Dual graph partitioning highlights a small group of pseudoknot-containing RNA submotifs', Submitted).

Our mutational robustness analysis can also be automated in the future. The results of our targeted mutation exercise with EteRNA are also encouraging. We successfully identified target residues for mutation based on comparison between the target and predicted RAG topologies. An automated approach to perform these mutations can thus be envisioned in the future. Such systematic mutations can be applied to increase the yield of our design pipeline. Finally, it may be interesting to try to design topologies we identified as non RNA-like to identify whether such targets would present greater challenges for inverse folding and if so, why. In this way, we might better understand fundamental relationships between sequence and structure, and structure and function.

DATA AVAILABILITY

The executable files for our software are available for download from our website http://www.biomath.nyu.edu/?q=software/RAGTOP.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Cigdem S. Bayrak, Dr Gavin Bascom and Dr Ognjen Perišić for helpful discussions and comments, Charlotte Davison and Maggie Pan for their work as part of the Girls' Science, Technology, Engineering and Mathematics (GSTEM) summer research program, and Shereef Elmetwaly for technical assistance.

FUNDING

National Institute of General Medical Sciences, National Institutes of Health (NIH) [GM100469, R35GM122562 to T.S., GM101237, HG008133 to A.L.]; National Institute of Child Health and Human Development, NIH [K08 HD069597 to S.B.V.R.]. The funding bodies listed above did not play any role in the design or conclusions of this study. Funding for open access charge: National Institute of General Medical Sciences, National Institutes of Health (NIH) [GM100469, R35GM122562 to T.S.].

Conflict of interest statement. None declared.

REFERENCES

1. Dirks,R.M., Lin,M., Winfree,E. and Pierce,N.A. (2004) Paradigms for computational nucleic acid design. *Nucleic Acids Res.*, **32**, 1392–1403.
2. Garman,E.F. (2014) Developments in X-ray crystallographic structure determination of biological macromolecules. *Science*, **343**, 1102–1108.
3. Gong,Z., Schwieters,C.D. and Tang,C. (2015) Conjoined use of EM and NMR in RNA structure refinement. *PLoS ONE*, **10**, 1–9.
4. Marchanka,A., Simon,B., Althoff-Ospelt,G. and Carlomagno,T. (2015) RNA structure determination by solid-state NMR spectroscopy. *Nat. Commun.*, **6**, 7024.
5. Earl,L.A., Falconieri,V., Milne,J.L. and Subramaniam,S. (2017) Cryo-EM: beyond the microscope. *Curr. Opin. Struct. Biol.*, **46**, 71–78.
6. Crick,F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.
7. Zaug,A.J. and Cech,T.R. (1986) The intervening sequence RNA of Tetrahymena is an enzyme. *Science*, **231**, 470–475.
8. Lilley,D.M.J. (2011) Mechanisms of RNA catalysis. *Philos. Trans. R Soc. B: Biol. Sci.*, **366**, 2910–2917.
9. Wilson,T.J., Liu,Y. and Lilley,D.M.J. (2016) Ribozymes and the mechanisms that underlie RNA catalysis. *Front. Chem. Sci. Eng.*, **10**, 178–185.
10. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
11. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9**, 1043–1049.
12. Kaikkonen,M.U., Lam,M.T. and Glass,C.K. (2011) Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc. Res.*, **90**, 430–440.
13. Patil,V.S., Zhou,R. and Rana,T.M. (2014) Gene regulation by non-coding RNAs. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 16–32.
14. Schlick,T. and Pyle,A.M. (2017) Opportunities and challenges in RNA structural modeling and design. *Biophys. J.*, **113**, 225–234.
15. Pyle,A.M. and Schlick,T. (2016) Challenges in RNA structural modeling and design. *J. Mol. Biol.*, **428**, 733–735.
16. Berman,H.M., Narayanan,B.C., Costanzo,L.D., Dutta,S., Ghosh,S., Hudson,B.P., Lawson,C.L., Peisach,E., Prlić,A., Rose,P.W. *et al.* (2013) Trendspotting in the Protein Data Bank. *FEBS Lett.*, **587**, 1036–1045.
17. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
18. Wilson,D.S. and Szostak,J.W. (1999) In vitro selection of functional nucleic acids. *Annu. Rev. Biochem.*, **68**, 611–647.
19. Stoltenburg,R., Reinemann,C. and Strehlitz,B. (2007) SELEX–a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Eng.*, **24**, 381–403.
20. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
21. Soukup,G.A. and Breaker,R.R. (1999) Engineering precision RNA molecular switches. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 3584–3589.
22. Que-Gewirth,N.S. and Sullenger,B.A. (2007) Gene therapy progress and prospects: RNA aptamers. *Gene Ther.*, **14**, 283–291.
23. Prakash,J.S. and Rajamanickam,K. (2015) Aptamers and their significant role in cancer therapy and diagnosis. *Biomedicines*, **3**, 248–269.
24. Hermann,T. and Westhof,E. (2000) Rational drug design and high-throughput techniques for RNA targets. *Comb. Chem. High T Scr.*, **3**, 219–234.
25. Sullenger,B.A. and Gilboa,E. (2002) Emerging clinical applications of RNA. *Nature*, **418**, 252–258.
26. Thiel,K.W. and Giangrande,P.H. (2009) Therapeutic applications of DNA and RNA aptamers. *Oligonucleotides*, **19**, 209–222.
27. Meyer,C., Hahn,U. and Torda,A.E. (2013) RNA aptamer design. In: *De novo Molecular Design*. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 519–542.
28. Hermann,T. (2000) Strategies for the design of drugs targeting RNA and RNA-protein complexes. *Angew. Chem. Int. Ed.*, **39**, 1890–1905.
29. Hong,W., Zeng,J. and Xie,J. (2014) Antibiotic drugs targeting bacterial RNAs. *Acta Pharmaceut. Sin. B*, **4**, 258–265.
30. Gallego,J. and Gabriele,Varani (2001) Targeting RNA with small-molecule drugs: a therapeutic promise and chemical challenges. *Acc. Chem. Res.*, **34**, 836–843.
31. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. Chem. Mon.*, **125**, 167–188.
32. Andronescu,M., Fejes,A.P., Hutter,F., Hoos,H.H. and Condon,A. (2004) A new algorithm for RNA secondary structure design. *J. Mol. Biol.*, **336**, 607–624.
33. Busch,A. and Backofen,R. (2006) INFO-RNA – a fast approach to inverse RNA folding. *Bioinformatics*, **22**, 1823–1831.
34. Matthies,M.C., Bienert,S. and Torda,A.E. (2012) Dynamics in sequence space for RNA secondary structure design. *J. Chem. Theory Comput.*, **8**, 3663–3670.
35. Zadeh,J.N., Steenberg,C.D., Bois,J.S., Wolfe,B.R., Pierce,M.B., Khan,A.R., Dirks,R.M. and Pierce,N.A. (2011) NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.*, **32**, 170–173.
36. Wolfe,B.R., Porubsky,N.J., Zadeh,J.N., Dirks,R.M. and Pierce,N.A. (2017) Constrained multistate sequence design for nucleic acid reaction pathway engineering. *J. Am. Chem. Soc.*, **139**, 3134–3144.
37. Taneda,A. (2011) MODENA: a multi-objective RNA inverse folding. *Adv. Appl. Bioinforma Chem.*, **4**, 1–12.
38. N,D., Avihoo,A. and Barash,D. (2008) Reconstruction of natural RNA sequences from RNA shape, thermodynamic stability, mutational robustness, and linguistic complexity by evolutionary computation. *J. Biomol. Struct. Dyn.*, **26**, 147–162.
39. Avihoo,A., Churkin,A. and Barash,D. (2011) RNAexinv: An extended inverse RNA folding from shape and physical attributes to sequences. *BMC Bioinformatics*, **12**, 319.
40. Weinbrand,L., Avihoo,A. and Barash,D. (2013) RNAfbinv: an interactive Java application for fragment-based design of RNA sequences. *Bioinformatics*, **29**, 2938–2940.
41. Lee,J., Kladwang,W., Lee,M., Cantu,D., Azizyan,M., Kim,H., Limpaecher,A., Gaikwad,S., Yoon,S., Treuille,A. *et al.* (2014) RNA

design rules from a massive open laboratory. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 2122–2127.

42. Anderson-Lee,J., Fisker,E., Kosaraju,V., Wu,M., Kong,J., Lee,J., Lee,M., Zada,M., Treuille,A. and Das,R. (2016) Principles for predicting RNA secondary structure design difficulty. *J. Mol. Biol.*, **428**, 748–757.
43. Waterman,M. (1978) Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Stud.*, **1**, 167–212.
44. Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **77**, 6309–6313.
45. Le,S., Nussinov,R. and Maizel,J. (1989) Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, **22**, 461–473.
46. Shapiro,B.A. and Zhang,K. (1990) Comparing multiple RNA secondary structures using tree comparisons. *Bioinformatics*, **6**, 309–318.
47. Laing,C. and Schlick,T. (2011) Computational approaches to RNA structure prediction, analysis, and design. *Curr. Opin. Struct. Biol.*, **21**, 306–318.
48. Kim,N., Fuhr,K.N. and Schlick,T. (2013) Graph applications to RNA structure and function. In: Russell,R (ed). *Biophysics of RNA Folding*. Springer, NY, 23–51.
49. Schlick,T. (2018) Adventures with RNA graphs. *Methods*, doi: 10.1016/j.ymeth.2018.03.009.
50. Fera,D., Kim,N., Shiffeldrim,N., Zorn,J., Laserson,U., Gan,H.H. and Schlick,T. (2004) RAG: RNA-As-Graphs web resource. *BMC Bioinformatics*, **5**, 1.
51. Kim,N., Petingi,L. and Schlick,T. (2013) Network theory tools for RNA modeling. *WSEAS Trans. Math.*, **9**, 941–955.
52. 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.
53. 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.
54. Kim,N., Laing,C., Elmetwaly,S., Jung,S., Curuksu,J. and Schlick,T. (2014) Graph-based sampling for approximating global helical topologies of RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 4079–4084.
55. Kim,N., Zahran,M. and Schlick,T. (2015) Chapter 5 - Computational prediction of riboswitch tertiary structures including pseudoknots by RAGTOP: a hierarchical graph sampling approach. In: Chen,S-J and Burke-Aguero,DH (eds). *Computational Methods for Understanding Riboswitches, Vol. 553 of Methods in Enzymology*. Academic Press, Waltham, MA, pp. 115–135.
56. Bayrak,C.S., Kim,N. and Schlick,T. (2017) Using sequence signatures and kink-turn motifs in knowledge-based statistical potentials for RNA structure prediction. *Nucleic Acids Res.*, **45**, 5414–5422.
57. Zahran,M., Bayrak,C.S., Elmetwaly,S. and Schlick,T. (2015) RAG-3D: a search tool for RNA 3D substructures. *Nucleic Acids Res.*, **43**, 9474–9488.
58. Kim,N., Zheng,Z., Elmetwaly,S. and Schlick,T. (2014) RNA graph partitioning for the discovery of RNA modularity: a novel application of graph partition algorithm to biology. *PLoS ONE*, **9**, e106074.
59. Jain,S. and Schlick,T. (2017) F-RAG: Generating atomic models from RNA graphs using fragment assembly. *J. Mol. Biol.*, **429**, 3587–3605.
60. Gan,H.H., Pasquali,S. and Schlick,T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic Acids Res.*, **31**, 2926–2943.
61. Kim,N., Shiffeldrim,N., Gan,H.H. and Schlick,T. (2004) Candidates for novel RNA topologies. *J. Mol. Biol.*, **341**, 1129–1144.
62. Baba,N., Elmetwaly,S., Kim,N. and Schlick,T. (2016) Predicting large RNA-Like topologies by a knowledge-based clustering approach. *J. Mol. Biol.*, **428**, 811–821.
63. Izzo,J.A., Kim,N., Elmetwaly,S. and Schlick,T. (2011) RAG: an update to the RNA-As-Graphs resource. *BMC Bioinformatics*, **12**, 219.
64. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorith. Mol. Biol.*, **6**, 26.
65. Dirks,R.M. and Pierce,N.A. (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comp. Chem.*, **24**, 1664–1677.
66. Dirks,R.M. and Pierce,N.A. (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comp. Chem.*, **25**, 1295–1304.
67. Dirks,R.M., Bois,J.S., Schaeffer,J.M., Winfree,E. and Pierce,N.A. (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.*, **49**, 65–88.
68. Wilkinson,K.A., Merino,E.J. and Weeks,K.M. (2006) Selective 2′-hydroxyl acylation analyzed by primer extension SHAPE: quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*, **1**, 1610–1616.
69. Siegfried,N.A., Busan,S., Rice,G.M., Nelson,J. A.E. and Weeks,K.M. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods*, **11**, 959–965.
70. Smola,M.J., Rice,G.M., Busan,S., Siegfried,N.A. and Weeks,K.M. (2015) Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc.*, **10**, 1643–1669.
71. Busan,S. and Weeks,K.M. (2018) Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA*, **24**, 143–148.
72. Lorenz,R., Luntzer,D., Hofacker,I.L., Stadler,P.F. and Wolfinger,M.T. (2016) SHAPE directed RNA folding. *Bioinformatics*, **32**, 145–147.
73. Xu,Z.Z. and Mathews,D.H. (2016) Experiment-assisted secondary structure prediction with RNA structure. In: Turner,DH and Mathews,DH (eds). *RNA Structure Determination: Methods and Protocols*. Springer, NY, pp. 163–176.
74. Deigan,K.E., Li,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 97–102.
75. Darty,K., Denise,A. and Ponty,Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.