Bioinformatics Advance Access published September 13, 2007

RAGPOOLS: RNA-As-Graph-Pools – A Web Server for Assisting the Design of Structured RNA Pools for *In Vitro* Selection

Namhee Kim[1], Jin Sup Shin[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 10012

Associate Editor: Prof. Thomas Lengauer

ABSTRACT

Summary: Our RNA-As-Graph-Pools (RAGPOOLS) web server offers a theoretical companion tool for RNA *in vitro* selection and related problems. Specifically, it suggests how to construct RNA sequence/structure pools with user-specified properties and assists in analyzing resulting distributions. This utility follows our recently developed approach for engineering sequence pools that links RNA sequence space regions with corresponding structural distributions via a "mixing matrix" approach combined with a graph theory analysis of RNA secondary-structure space (Kim *et al.*, 2007); the mixing matrix specifies nucleotide transition rates, and graph theory links sequences to simple graphical objects representing RNA motifs. The companion RAGPOOLS web server ("Designer" component) provides optimized starting sequences, mixing matrices, and associated weights in response to a user-specified target pool structure distribution. In addition, RAGPOOLS ("Analyzer" component) analyzes the motif distribution of pools generated from user-specified starting sequences and mixing matrices. Thus, RAGPOOLS serves as a guide to researchers who aim to synthesize RNA pools with desired properties and/or experiment *in silico* with various designs by our approach.

Availability: The web server is accessible on the web at http://rubin2.biomath.nyu.edu

Contact: schlick@nyu.edu

1 INTRODUCTION

RNA *in vitro* selection is a versatile experimental approach for screening large random RNA sequence libraries (10^{15}) for specific functions, such as binding or catalysis. Numerous novel aptamers and ribozymes have been discovered via RNA *in vitro* selection (Wilson and Szostak, 1999). Enhancing the scope of *in vitro* selection experiments via pool design could widen the range of structures and functions found in RNA pools and, in turn, expand upon associated applications in technology and bioengineering.

Many RNAs identified from random pools have simple structural motifs (e.g., stem-loop, stem-bulge-stem-loop). For example, our graph-based analysis of random pools demonstrated that the generated RNA secondary topologies are far from uniformly distributed and, in fact, favor simple motifs (Gevertz *et al.*, 2005). Thus, designed RNA pools that favor complex structures could enhance the discovery of novel RNAs.

We have recently developed a computational approach for designing structured RNA pools by modeling pool synthesis using graph theory for analyzing RNA structure space and mixing matrices for generating designed pools (Kim *et al.*, 2007). To make the design approach available to experimentalists and other RNA researchers, we have developed a companion web server, RAGPOOLS, for designing and analyzing structured pools for *in vitro* selection. RAGPOOLS aims to: help design structured RNA pools with target motif distribution; analyze structural distributions of RNA pools produced by our approach; and stimulate discoveries of novel RNAs via combined experimental and theoretical pool design.

2 METHODS

Full details of our targeted design approach are provided in Kim *et al*. (2007) and the web server tutorial (rubin2.biomath.nyu.edu/tutorials.html). Essentially, we design structured RNA pools using both random and biased sequence mutations around a specific sequence. The mixing matrices (MM) have elements that specify mixing ratios in the four phosphoramidite (A, C, G, and U (or T)) vials (i.e., synthesis ports); applying these matrices to starting sequences leads to designed sequence pools. Using such matrices to represent pool generation allows computational analysis of pool properties. In its "Designer" component, RAGPOOLS optimizes the set of mixing matrices, starting sequences, and associated weights for a given user-specified structural distribution in the pool.

2.1 Mixing matrices and starting sequences

For pool synthesis using four vials, the mixing matrix M is a 4 by 4 matrix; M_{ij} denotes the molar fraction of base j in vial "for base i".

Mixing matrices with symmetric elements, $M_{AU} = M_{UA}$, $M_{CG} = M_{GC}$, tend to preserve base pairs. Such matrices cover the sequence subspace approximating covariance mutations (e.g., AU to UA, CG to GC). Alternatively, to disrupt stems and generate new structures, we consider asymmetric matrices without the property of covariance mutations; non-covariance mutations, including random mutations, are commonly used for *in vitro* selection applications. Based on these biologically-motivated mutations, we construct six representative matrix classes for a total of 34 mixing matrices (see Kim *et al*., 2007). This number of matrices will increase in future versions of our program.

As suggested by RNA graph theory (Gevertz *et al.*, 2005), we use starting sequences/structures to represent distinct RNA topologies in structure space and to allow exploration of their structural neighbors via mutations. We use 30 starting sequences classified by shape, length, and function. For example, the starting sequences with distinct RNA tree structures are: tRNA (81-nt), hammerhead ribozyme (49-nt), GTP-binding aptamer (69-nt), and modified GTP-binding aptamer (54-nt). RAGPOOLS has pre-calculated results for all secondary motif distributions (as determined by Vienna RNAfold) corresponding to all mixing matrix/starting sequence

*To whom correspondence should be addressed.

combinations. These data for 5000 total sequences serve as reference for the pool optimization algorithm.

2.2 An algorithm for designing structured pools

The algorithm is based on analyses of sequence and structure spaces to enrich pools for specific structures. The algorithm exploits reference data that relate mixing matrices and starting sequences to pool motif distributions. Here, a motif is defined as a 2D RNA tree topology or shape. By knowing the structural distributions of mixing matrix/starting sequence pairs, we optimize the choice of starting sequences, mixing matrices, and associated weights (pool fractions) to approximate the target structured pool.

Recall that reference data are available for motif distributions corresponding to all starting sequence and mixing matrix combinations. The user specifies three items: (a) a target distribution of RNA tree topologies (see RAG, http://monod.biomath.nyu.edu/rna, for enumerated topologies), (b) number of mixing matrices, and (c) starting sequences to be used for approximating the target distribution. By the optimization procedure described in Kim *et al.* (2007) (see also the tutorial at rubin2.biomath.nyu.edu/algorithm.html), RAGPOOLS then determines an optimal combination of starting sequences, mixing matrices, and associated weights for the target RNA motif distribution. Essentially, the algorithm involves calculations of associated weights for all possible cases, estimation of topology distribution and error from target distribution, and minimization of errors. See Table 1 for examples of input and output.

Table 1. Examples of structured RNA pools designed by RAGPOOLS; see tutorials for definitions of mixing matrices (MM) and starting sequences (SS).

Input			Output
Target distributions	Number of MM	Starting sequences	Optimized weights, MM, and SS
4_1, 4_2: 30%, 30%	2 (Class B)	All	78%, MM13, mod. GTP aptamer 22%, MM12, Hammerhead ribo.
5_1, 6_1: 30%, 30%	2	80-100 nt	38.5%, MM3, tRNA 61.5%, MMT8, let-7 ncRNA

2.3 Implementation

The server's architecture consists of three components: Web interface, Engine, and Back End. The web interface is made of html pages and java scripts. The engine consists of 4 perl scripts which validate user input and call the c programs for predicting RNA secondary structure, converting secondary structures to tree graphs, and optimizing mixing matrices. The back end of the server contains reference data and databases (e.g., pre-calculated motif distributions, tree graphs in RAG) used to process calculations and analyses. We use an SGI 1450 computing system with 4 Intel Pentium III 700MHz processors and 2GB memory.

3 FEATURES OF RAGPOOLS

RAGPOOLS contains two parts: RNA pool designer and RNA pool analyzer. Figure 1 shows the organization of RAGPOOLS web server. It also contains the tutorial pages to define key concepts and methods, including *in vitro* selection, mixing matrix, starting sequence, optimization algorithm, and examples of designed pools.

3.1 RNA pool designer

As detailed above, the RNA pool designer computes the optimal designed pool parameters corresponding to the user input. For example, if the user requests to use two matrices with the conservation of C and G and all sequences to achieve 30% of 4_1 and 30% of 4_2 tree motifs, the optimization specifies 78% of matrix 13 with modified GTP aptamer and 22% of matrix 12 with the hammerhead ribozyme. This combination yields the desired structural distribution. The user specified input variables (Table 1) are limited to the numbers available in the web server (e.g., currently 34 MM and 30 SS); we have found in practice that the error depends on the target distribution and is large when complex topologies with high frequencies are sought (e.g., 5_3: 100%, error = 49%). Of course, all the design and analysis described here depend on the accuracy of 2D folding algorithms. However, for the generally short sequences we used here (< 200 nt), prediction should be quite accurate. In our experience, optimization generally requires several seconds.

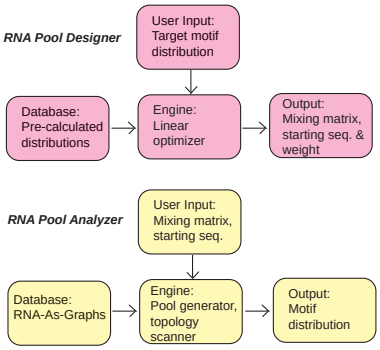

Fig. 1. Organization of the RAGPOOLS.

3.2 RNA pool analyzer

This part of the server analyzes the structural distribution of a pool generated by user-specified starting sequence and matrix, as shown in Figure 1. The resulting motif distribution is sent by email to users. For a sequence <100 nt, analysis requires around 30 min.

4 CONCLUSIONS

RAGPOOLS offers a general tool for designing and analyzing structured RNA pools with specified target motif distributions. We plan to expand the set of starting sequences and mixing matrices and provide further analyses of structural properties. We invite the users to explore RAGPOOLS and provide us feedback on usage and application by contacting us at: ragpools@biomath.nyu.edu

ACKNOWLEDGEMENTS

This work was supported by Human Frontier Science Program (HFSP) and by a joint NSF/NIGMS Initiative in Mathematical Biology (DMS-0201160).

Conflict of Interest: none declared.

REFERENCES

Gevertz, J., Gan,H.H., Schlick,T. (2005). *In vitro* RNA random pools are not structurally diverse: A computational analysis. *RNA*, **11**, 853-863.

Kim, N., Gan, H.H., Schlick, T. (2007). A computational proposal for designing structured RNA pools for *in vitro* selection of RNAs. *RNA*, **13**, 478-492.

Wilson, D.S., Szostak, J.W. (1999). *In vitro* selection of functional nucleic acids. *Annu.Rev.Biochem*., **68**, 611-647.