Invited Talk

Searching for 2D RNA Geometries in Bacterial Genomes

Uri Laserson Hin Hark Gan Tamar Schlick*
Department of Chemistry and Courant Institute of Mathematical Sciences
New York University, 251 Mercer Street, New York, New York 10012
schlick@nyu.edu

Categories and Subject Descriptors: J.3 [Life and medical sciences]: Biology and genetics

1. INTRODUCTION

The central dogma of biology, that DNA makes RNA makes protein, was again shown to be outdated with the recent discovery of RNAs that have essential regulatory functions (e.g., metabolite-binding RNAs, transcription regulation) [2, 15]. These finding have stimulated a large effort to search for small, functional RNA motifs (either embedded inside larger messenger RNA molecules or as separate molecules in the cell). For example, it is known that cells make use of a variety of small non-coding RNAs (such as microRNAs) as a mechanism for gene regulation. The very existence of these small motifs in Nature suggests that the functional, artificial RNA molecules developed through (experimental) *in vitro* selection technology[1] may shed some light on the scope and functional diversity of these small RNA molecules *in vivo*.

The binding and catalytic properties of nucleic acid molecules are conferred by specific sequence and structural motifs. Indeed, recent discoveries show that metabolite-induced RNA conformational changes constitute another form of bacterial gene regulation [12, 20, 21]. Since *in vitro* selection is a process that simulates evolution, it is reasonable that novel nucleic acid motifs discovered through *in vitro* selection experiments may have also evolved in the cell, especially since such motifs often target molecules (e.g., ATP, cAMP, antibiotics, etc.) that are prevalent in Nature.

The structure of RNA is hierarchical in nature. The primary (1D) structure of an RNA molecule is the oriented, linear ordering of the nucleotides A, C, G, and U. The secondary (2D) structure of the molecule is described by the pairs of nucleotides that "base-pair" with each other by forming hydrogen bonds according to the traditional Watson-Crick base pairing rules (i.e., A with U and G with C), plus other common pairing schemes, like the non-canonical G-U base pair. Finally, the tertiary (3D) structure of RNA refers the the arrangement of the 2D geometries in space. Though it is ultimately the 3D structure that confers function onto the molecule, it is known (thermodynamically) that RNA structure is hierarchical, so it is possible to assign functions to specific 2D geometries [17].

We present a method of extracting relevant structural details from experimental studies (mainly *in vitro* selection experiments) and applying several computational tools to search the genomes of various organisms for sequences that may potentially fold into similar structures, should they be transcribed in the cell. We also develop several tools for analyzing the significance of our results.

After applying our techniques to several aptamer motifs,[2] we report several promising candidate sequences in the genomes of various bacterial organisms that may exhibit the desired functional characteristics.

2. APTAMER SEARCH ALGORITHM

Our method for searching for the experimental aptamer structures in various genomes is comprised of three major steps:

1. Motif Descriptor Construction. The motif descriptor is written manually by extracting the critical sequence and structural motifs from the experimental 2D aptamer structure that mediate the specific physical/chemical properties of the molecule.[3]
2. Genome Search. The genomes of selected organisms are searched for the structural motif defined in the descriptor.
3. Analysis of Candidate Sequences. The results of the searches are subjected to several analyses to assess their quality, and filter out false positive signals.

The first step involves the analysis of the structure of the motif as elucidated by the *in vitro* selection experiment (usually at the level of secondary structure) and, if available,

*To whom correspondence should be addressed

[1] *In vitro* selection is a process that mimics evolution in which a large pool of small, random-sequence RNA molecules is subjected to an iterative process that selects for a specific physical or chemical property; see [19] for a review.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SCG'04, June 9–11, 2004, Brooklyn, New York, USA.
Copyright 2004 ACM 1-58113-885-7/04/0006 ...$5.00.

[2] Aptamers are small, artificial nucleic acid molecules that exhibit binding affinity to a specific molecule.

[3] See our articles on RNA geometries for introductions into RNA structural motifs and modeling: [5, 4] and `http://monod.biomath.nyu.edu/rna` . The Appendix of this article reviews work in our group on using graph theory to study RNA structures.

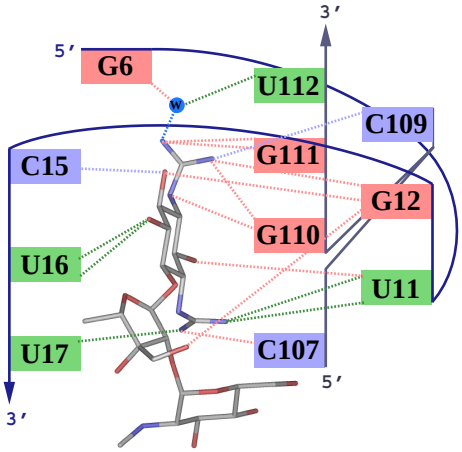

Figure 1. The hydrogen-bonding scheme of the streptomycin aptamer developed by the Schroeder group [18] and crystallized by the Patel group [16]. The dotted lines represent hydrogen bonds between the streptomycin molecule (center) and the nucleotides in the binding pocket. This type of information is ideal for pinpointing the specific nucleotides that account for the binding specificity. (Adapted from [16].)

also any detailed 3D structural characterizations of the motif (e.g., X-ray crystallography or NMR studies). This allows us to pinpoint the important structural qualities of the motif such as the qualitative structure (e.g., helices, bulges, hairpins, etc.) and the specific nucleotide sequences that are critical in the motif's physical/chemical activity.

This process is illustrated with the streptomycin-binding aptamer in Figure 1. Since 3D crystallographic data is available, the hydrogen-bonding scheme is clearly discernible, and it is a simple matter to pinpoint the specific nucleotides involved in conferring the binding specificity of the RNA molecule. Additionally, since the secondary structure of the molecule is also known (from the original *in vitro* selection study; see top of Figure 2), it is simple to extract the overall structural motif (which acts as a structural scaffold).

This information is subsequently compiled in a motif "descriptor" (a computer language designed as part of the RNAMotif software [10] used to define nucleic acid secondary structures), and the motif is searched for in the genomes of various organisms using the RNAMotif tool. This tool produces a set of genomic sequences that have the *potential* to fold into the specified motif based on Watson-Crick base-pairing rules (i.e., A with U and G with C). We subsequently filter the initial pool of candidates by using standard tools to predict the secondary structures of the candidates and discard those RNA sequences that fold into structures that are not similar to the original experimental form.

Finally, we subject the remaining pool of candidates to further tests to evaluate which candidates are most likely biological sequences, and to hopefully filter out additional false-positives. These include energetic calculations (to evaluate the stability of the candidate sequences) as well as statistical tests (to deduce whether there is a statistical over- or under-representation of the specific motif). These methods are summarized below, though full details will be given in a separate work.

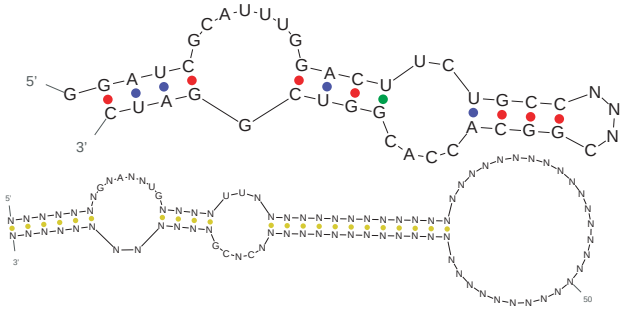

Figure 2. The secondary structure of the streptomycin aptamer (top) with the corresponding motif descriptor (bottom). Regions with many 'N's are variable in length. The nucleic acid symbols are: $N = \{A, C, G, U\}$, $V = \{A, C, G\}$, and $S = \{C, G\}$. We search several genomes for sequences fits into the descriptor consensus structure (allowing for slight mismatches). The descriptor structure shows what information is considered critical for the binding specificity. The aptamer experimental study is [18, 16].

2.1 Energetic Analyses

After initially filtering the pool of candidates by predicting whether they adopt the correct geometry, the remaining candidates are subjected to two energetic analyses to evaluate their stability, an analysis of heat capacity curves ("melting curves") and the generation of a conformational energy landscape (a "folding funnel").

The melting curve of a molecule plots its specific heat versus temperature. Higher and sharper peaks at high temperatures correspond to higher structural stability. Therefore, we compute the melting curve of our candidate sequence (using the Vienna RNA package [8]) as well as melting curves for a large number of random permutations of the sequence. We can then compare the melting temperature of the candidate with the melting temperatures of the shuffled sequences to see whether the candidate stands out.

Conformational energy landscapes are obtained by calculating the suboptimal structures of the candidate sequence (also with the Vienna RNA package), and plotting the energy of each such structure against its "distance" from the minimum energy structure. The distance metric can be defined in many ways. Since the secondary structure of an RNA molecule can be defined as a set of pairs of nucleotides, we use a simple distance metric that defines the distance between two structures (on the same sequence), as the symmetric difference set between the two sets of base-pairs. Structures that are very stable should have steep "funnels" where the greater the distance from the minimum energy structure, the greater the energy of the corresponding suboptimal structure.

2.2 Statistical Analysis

Since the motifs can be described as "words" in the 4-letter nucleotide alphabet, it is possible to analytically compute the theoretical frequency of observing each motif in a random sequence of nucleotides [13]. This makes it possible to evaluate whether a specific genome has a statistical over- or under-representation of the given motif. However, this computation can become extremely complex. Therefore, a Monte Carlo estimation method is far more appropriate.

Specifically, we implement a simple Sample-Mean method [14] that, for each iteration, generates a 1 Mbp uniformly distributed sequence of nucleotides (using the Mersenne Twister algorithm [11]). We then search the sequence for the specified motif using RNAMotif and average the number of matches over one million iterations.

2.3 Computational Performance

Creating the actual motif descriptors is done manually and requires biological intuition and (most importantly) experimentation. Once the descriptor is defined, the genome searches using RNAMotif are very fast (for bacterial-sized genomes, < 10 Mbp). The search conducted on the streptomycin-binding aptamer in Figures 1 and 2 was completed in less than one minute for the largest genome searched (*Streptomyces avermitilis*, ~ 9 Mbp). In general, the computational speed depends on both the genome size as well the the complexity of the descriptor.

The Vienna RNA package, which we use to predict the secondary structures of the candidates (i.e., "fold" them), can fold a sequence less than 100 nucleotides long in a matter of seconds. (The predicted secondary structure corresponds to the minimum potential energy conformation; therefore, the folding algorithm is essentially an optimization algorithm.) However, computing suboptimal structures for a given sequence (i.e., local minima that are higher in energy than the global minimum) is relatively lengthy and variable. The problem is that certain sequences, despite being of similar length, have extremely different numbers of suboptimal structures (for example, some sequences have about 150 suboptimal structures while other sequences of similar length have about 200,000). However, the generation of each suboptimal structure is fast and even the case with over 200,000 structure is completed within 2 minutes. (The generation of suboptimal structures is used for the construction of conformational energy landscapes, discussed below.)

The screening analyses (step 3) we perform are more computationally intensive. Generating heat capacity curves (also using the Vienna package) can take several minutes. Since we generate many curves for each sequence, the process is relatively computationally expensive (up to several hours of CPU time). Additionally, we implement a Monte Carlo method for computing the theoretical expected frequency of each motif. We perform on the order of one million iterations of the Sample-Mean method since our sample-space has very high dimensionality. The combined calculations take on the order of weeks (with the MC accounting for the largest amount of computation).

For more detailed analyses of the performance of the tools used, we refer the reader to the literature describing the tools in detail. Our computations were performed on an SGI 300 MHz MIPS R12000 IP27 processor with 4 GB of memory.

3. RESULTS

We tested our method on several aptamers that bind to antibiotics (chloramphenicol [1], streptomycin [18, 16], and neomycin B [9]). We searched the genomes of various bacterial organisms for these motifs to see if the algorithm is successful in producing potential candidate sequences that may bind the molecule in question.[4] The results are summarized in Tables 1 and 2 for statistics of matches and sequences, respectively.

[4]The genomes are available for download at the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov`).

Table 1. Number of initial sequence matches for the three antibiotic-binding aptamers in selected bacterial genomes.

Genome	Size (Mbp)	Chlor phen	Str myc	Neom B	**Total**
Strep. avermitilis	9.2	0	2	6	8
Strep. coelicolor	8.8	0	1	1	2
E coli K12	4.7	7	1	11	19
Ec O157:H7	5.6	7	1	17	25
Ec O157:H7 EDL933	5.6	7	1	19	27
Ec CFT073	5.3	7	0	17	24
Neis. men. MC58	2.3	1	0	9	10
Neis. men. Z2498	2.2	5	0	9	14
Sinorh. meliloti	3.7	1	1	3	5
Chlamydia trach.	1.1	4	0	1	5
	Total	39	7	93	139

Chloramphenicol Candidate Sequences				
`1:TCAGAGCTGAAAAACTGGCCCCGGGTGCAGCTAAAAACTGA`				
`2:TCAGAGCTGAAAAACTGGCCCCGAGTGCAGCTAAAAACTGA`				
`3:TCAGAGCTGAAAAACTGGCCCCGAGTGCAGCTAAAAACTGA`				
`4:GGGAAACGCAAAAATCAGGTACAGGCAGACGACGGTGGCGTAAAACTCCC`				
Sequence:	1	2	3	4
Genome:	*E coli*	*E coli*	*Ec O157:H7*	*N. men.*
Genome:	*K12*	*O157:H7*	*EDL933*	*Z2491*
EFFE:	−11.9	−10.58	−10.58	−9.17
Melt T:	78.0	72.6	72.6	54.4
Melt Curve:	Accept	Accept	Acceptable	Poor
CE Landscape:	Good	Fair	Fair	Fair

Table 2. Distilled candidate pool for chloramphenicol (after filtering by looking at predicted 2D structures). Ensemble Folding Free Energy (EFFE) is measured in Kcal/mol and Melting Temperature is measured in degC. The folding free energies are computed using the Vienna RNA package. The corresponding RNA sequences have `T` replaced with `U`. Note that the second and third candidates have identical sequences.

Several trends are immediately discernible from the results in Table 1. For example, the average number of matches to neomycin B in the *Streptomyces* genomes is 4, while the average number for *E. coli* genomes is four times as many. However, the *Streptomyces* genome is about twice the size of the *E. coli* genome, so based on probability alone, one would expect that *Streptomyces* would have about twice the number of matches compared with *E. coli*, not one-fourth, as we observe.

Another significant trend is that the neomycin B pool clearly has many more matches than either the chloramphenicol or streptomycin pools. This is due to the fact the the neomycin B aptamer is simpler (or more rigorously, the neomycin B aptamer has less Shannon entropy [3]). Therefore, it is important to estimate the expected number of matches in a random sequence to be able to significantly conclude whether or not there is a statistical over- or under-representation of a specific motif. These estimates are obtained by implementing a Monte Carlo scheme as described previously. (These calculations are still underway.)

Following the initial search, we filter the pool of candidate sequences by folding each sequence and discarding those that fold into structures that are not similar to the target experimental motif geometry. We then subject the sequences to

two energetic analyses, an analysis of heat capacity curves (melting curves) and the generation of a conformational energy landscape.

Since functional, biological RNA molecules are expected to form stable structures, we can use both of these analyses to determine exactly which candidate sequences have the highest potential to actually be functional *in vivo*. The final batch of candidates for the chloramphenicol motif are shown in Table 2, along with descriptions of the results of their energetic analyses.

4. DISCUSSION AND DIRECTIONS

We have developed a method to search for specific RNA geometries in the genomes of various organisms. Our study is motivated by the deduction that since these RNA molecules were developed by a method that simulates evolution (*in vitro* selection), similar structures are likely present in the cell. After applying this method to several aptamers, we found many potential candidate sequences that correspond to the experimental structures.

Clearly, much work remains ahead. We are in the process of automating the analysis so that all genomes available at NCBI can be searched at once. We are also developing better ways to eliminate false-positives from our search results. Most importantly, we plan to conduct an experimental collaboration to verify whether our candidate sequences both exhibit the desired physical or chemical property and are expressed *in vivo*. Significantly, our search methodology may ultimately be used to apply *in vitro* selection technology as a tool for identifying novel RNA genes or RNA-based regulatory mechanisms.

Acknowledgements

We thank Marco Avellaneda and Mike Waterman for many suggestions regarding the methods used in this work. We thank Tom Macke and Ivo Hofacker for technical assistance with RNAMotif and the Vienna RNA Package, respectively, and Yanli Wang for preparing Figure 1. We also thank Dave Scicchitano for supporting this study through the Department of Biology. UL thanks the Howard Hughes Medical Institute for an undergraduate research fellowship through the Honors Summer Institute at NYU. We also gratefully acknowledge the support of the NSF, NIH, and Human Frontier Science Program for supporting this research.

References

[1] D. H. Burke, D. Hoffman, A. Brown, M. Hansen, A. Pardi, and L. Gold. RNA aptamers to the peptidyl transferase inhibitor chloramphenicol. *Chem. Biol.*, 4:833–843, 1997.

[2] J. A. Doudna and T. R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418:222–228, 2002.

[3] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[4] H. H. Gan, D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick. RAG: RNA-As-Graphs database – concepts, analysis, and features. *Bioinformatics*, 20:1285–1291, 2004.

[5] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucl. Acids Res.*, 31:2926–2943, 2003.

[6] F. Harary. The number of homeomorphically irreducible trees and other species. *Acta Math.*, 101:141–162, 1959.

[7] F. Harary. *Graph Theory*. Addison-Wesley, 1969.

[8] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[9] L. Jiang, A. Majumdar, W. Hu, T. J. Jaishree, W. Xu, and D. J. Patel. Saccharide-RNA recognition in a complex formed between neomycin B and an RNA aptamer. *Structure Fold Des.*, 7:817–827, 1999.

[10] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucl. Acids Res.*, 29:4724–4735, 2001.

[11] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. on Modeling and Comp. Simulation*, 8:3–30, 1998.

[12] A. Nahvi, N. Sudarsan, M. S. Ebert, X. Zou, K. L. Brown, and R. R. Breaker. Genetic control by a metabolite binding mRNA. *Chem. Biol.*, 9:1043–1049, 2002.

[13] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.*, 7:1–46, 2000.

[14] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley, New York, 1981.

[15] G. Storz. An expanding universe of noncoding RNAs. *Science*, 296:1260–1263, 2002.

[16] V. Tereshko, E. Skripkin, and D. J. Patel. Encapsulating streptomycin within a small 40-mer RNA. *Chem. Biol.*, 10:175–187, 2003.

[17] I. Tinoco, Jr. and C. Bustamante. How RNA folds. *J. Mol. Biol.*, 293:271–281, 1999.

[18] S. T. Wallace and R. Schroeder. In vitro selection and characterization of streptomycin-binding RNAs: Recognition discrimination between antibiotics. *RNA*, 4:112–123, 1998.

[19] D. S. Wilson and J. W. Szostak. In vitro selection of functional nucleic acids. *Ann. Rev. Biochem.*, 68:611–647, 1999.

[20] W. Winkler, A. Nahvi, and R. R. Breaker. Thiamine derivatives bind messenger RNAs directly to regulate bacterial expression. *Nature*, 419:952–956, 2002.

[21] W. C. Winkler, S. Cohen-Chalamish, and R. R. Breaker. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. USA*, 99:15908–15913, 2002.

APPENDIX

A.1 RNA genomics and graph theory

An alternate approach to RNA genomics developed in our group is the use of graph theory (Gan and Schlick, in progress, see also [5, 4]). Graph theory analysis of genomes is promising because all RNA structures can be schematically represented as two-dimensional graphs. Thus, novel graph topologies from graphical enumeration can be used to drive discovery of novel RNA motifs in genomes via methods and analyses similar to those described above. Below, we outline the essentials and advantages of graph theory for describing, cataloguing, and predicting RNA structures.

A.2 RNA structural motifs & graphs

RNA molecules are hierarchical in nature since their secondary structures are known to be stable independently of their tertiary structures [17]. RNA secondary motifs have a network-like topology with stems linking loops, bulges, and junctions (Figure 3). Such a topological RNA representation allows exploration of RNA topologies using graph theory.

Figure 3 shows three RNA secondary (tree) motifs represented as tree graphs: the vertices (•) are RNA loops, bulges or junctions, and the edges (lines, —) are RNA stems (precise rules are detailed in [5]). Thus, the schematic tree graphs represent the connectivity between the RNA secondary elements (e.g., stems,

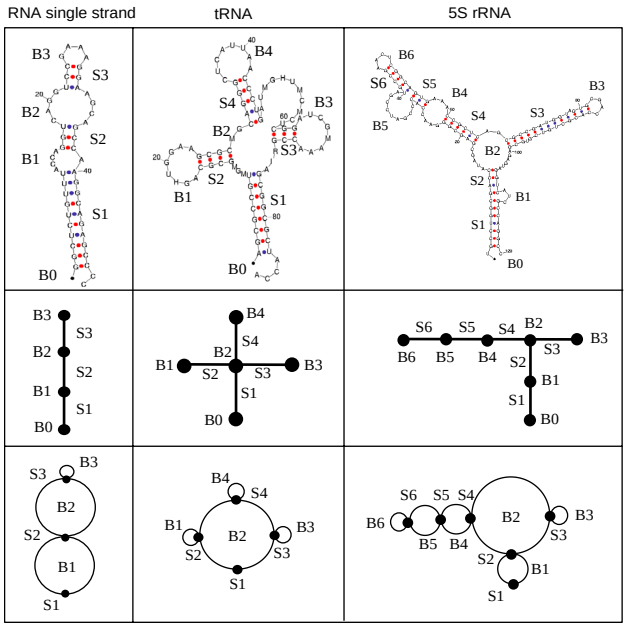

Figure 3. Graphical representations of RNA secondary structures (top) as tree (middle) and dual (bottom) graphs.

loops, bulges, junctions). The tree graphs provide intuitive representations of RNA structures, but they cannot represent other important RNA types, such as pseudoknots. For completeness, we developed another class of RNA graphs called *dual graphs* (third row of Figure 3; [5]); dual graphs can represent all RNA trees and pseudoknots and can be generalized to represent unusual RNA structures with triple, quadruple, and higher-order helices.

Since the "RNA graphs" are discrete, they allow us to enumerate all possible 2D RNA motifs using enumeration methods of graph theory. Graphical enumeration of RNA topologies can be performed analytically or computationally depending on the complexity of the structures. For example, for unlabeled trees, the number of possible graphs with i vertices is obtained from the coefficients c_i associated with the x^i term of the counting polynomial derived by Harary and Prins [7].

These sets of distinct graphs represent libraries of theoretically possible RNA topologies, which include naturally occurring, candidate, and hypothetical RNA motifs (see Schlick lab's RNA-As-Graphs (RAG) Database at `http://monod.biomath.nyu.edu/rna/` and [4]). Known RNAs in public databases (NDB and others) can thus be matched to the topologies we describe (see Figure 3). Significantly, because we found that the known 2D RNA motifs represent only a small subset of all possible topologies, we hypothesize that some of the missing motifs may represent undiscovered *naturally occurring* RNAs while others may be designed and then synthesized in the laboratory.

A.3 Sequence space vs. topology space

Current theoretical and experimental approaches to RNA structure explore RNA's sequence space. Experimental *in vitro* selection techniques exploit random sequence pools for comprehensive searches for novel RNAs. In the search for RNA genes in genomes, scanning algorithms require sequence and structural motifs as input. In contrast, our RNA analysis focuses on structural motifs rather than sequences per se. A critical advantage of RNA graph analysis is that the space of topologically distinct structures is vastly smaller than the nucleotide sequence space. In fact, we estimate, based on Harary-Prins enumeration formula for tree graphs (above) [7, 6], that the number of distinct RNA tree topologies can be parameterized as $\sim 2.5^{(N/20)-3}$ for $N > 60$ compared with 4^N for the nucleotide sequence space! The markedly smaller RNA topology space implies great potential for the search for novel RNA structures. Once a novel target topology/motif is identified, the corresponding RNA sequences can be found in two ways: for natural RNAs, the selected motif can be found by scanning the genomes; and, for synthetic RNAs, they can be designed using modular assembly of existing RNA fragments (i.e., using a library of sequence/motif building blocks and application of 2D folding algorithms). Both of these research directions are currently being pursued in our laboratory.

A.4 RAG: RNA-As-Graphs Database

Our RNA graphical representations present an opportunity for cataloguing of RNA structures based on their topological properties (Figure 4). Cataloguing RNA's structural diversity, including hypothetical motifs, is vital for identifying novel RNA structures and for pursuing RNA genomics initiatives. Our RNA-As-Graphs (RAG; `http://monod.biomath.nyu.edu/rna`) database catalogues and ranks all mathematically possible (including existing and candidate) RNA secondary motifs on the basis of graphical enumeration results. We archive RNA tree motifs as "tree graphs" and other RNAs, including pseudoknots, as general "dual graphs." All RNA motifs are catalogued by graph vertex number (a measure of sequence length) and ranked by topological complexity (second smallest eigenvalue (λ_2) corresponding to the graph's *Laplacian* matrix). RAG's inventory immediately suggests candidates for novel RNA motifs, either naturally occurring or synthetic. Through RAG, we hope to pursue and further stimulate efforts to predict and design novel RNA motifs and thereby contribute to RNA genomics initiatives.

V	λ_2	Tree Graph	Secondary Structure
4	0.5858		RNA single strand (PR0021)
	1.0000		tRNA (PR0019)
5	0.3820		70S (F) (RR0003)
	0.5188		tRNA (TR0001)
	1.0000		tRNA (TRNA12)

Figure 4. RNA tree motif libraries for $V = 4, 5$. V is vertex number and λ_2 is the second smallest eigenvalue of the Laplacian matrix.