

Candidates for Novel RNA Topologies

Namhee Kim^{1†}, Nahum Shiffeldrim^{1†}, Hin Hark Gan¹ and Tamar Schlick^{1,2*}

¹Department of Chemistry
New York University, 100
Washington Square East
Room 1001, New York, NY
10003, USA

²Courant Institute of
Mathematical Sciences, New
York University, 251 Mercer
Street, New York, NY 10012
USA

Because the functional repertoire of RNA molecules, like proteins, is closely linked to the diversity of their shapes, uncovering RNA's structural repertoire is vital for identifying novel RNAs, especially in genomic sequences. To help expand the limited number of known RNA families, we use graphical representation and clustering analysis of RNA secondary structures to predict novel RNA topologies and their abundance as a function of size. Representing the essential topological properties of RNA secondary structures as graphs enables enumeration, generation, and prediction of novel RNA motifs. We apply a probabilistic graph-growing method to construct the RNA structure space encompassing the topologies of existing and hypothetical RNAs and cluster all RNA topologies into two groups using topological descriptors and a standard clustering algorithm. Significantly, we find that nearly all existing RNAs fall into one group, which we refer to as "RNA-like"; we consider the other group "non-RNA-like". Our method predicts many candidates for novel RNA secondary topologies, some of which are remarkably similar to existing structures; interestingly, the centroid of the RNA-like group is the tmRNA fold, a pseudoknot having both tRNA-like and mRNA-like functions. Additionally, our approach allows estimation of the relative abundance of pseudoknot and other (e.g. tree) motifs using the "edge-cut" property of RNA graphs. This analysis suggests that pseudoknots dominate the RNA structure universe, representing more than 90% when the sequence length exceeds 120 nt; the predicted trend for <100 nt agrees with data for existing RNAs. Together with our predictions for novel "RNA-like" topologies, our analysis can help direct the design of functional RNAs and identification of novel RNA folds in genomes through an efficient topology-directed search, which grows much more slowly in complexity with RNA size compared to the traditional sequence-based search.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: RNA secondary structure; novel RNA; pseudoknot; graph theory; clustering algorithm

*Corresponding author

Introduction

The notion that biological function follows structure holds for protein and RNA molecules. Like proteins, RNAs perform important cellular functions, and exhibit a repertoire that is expanding rapidly.^{1–4} Uncovering the range of RNA's structural repertoire is a key to understanding the functional diversity of the RNA universe. Indeed, protein structural genomics projects are well

underway to map protein structures with the aim of functionally characterizing protein sequences in databases;^{5,6} the goal of protein structural genomics is to systematically determine novel protein fold classes and estimate the number of such classes.⁷ In contrast to proteins, such important developments have not been matched for RNA molecules despite the growing recognition of their functional importance. A parallel effort to uncover the structural repertoire of RNA molecules will lead to a more comprehensive mapping of functional molecules in the cell.⁸ To help close the gap of what we know about protein and RNA structure worlds, we describe applications of a novel theoretical approach to predict RNA-like motifs and estimate RNA's structural repertoire based on analyses of

† N.K. and N.S. contributed equally to this work.
Abbreviations used: PAM, partitioning around medoids; tRNA, transfer RNA; RAG, RNA-As-Graphs.
E-mail address of the corresponding author:
schlick@nyu.edu

RNA topological properties and information about existing RNAs.

RNA molecules are hierarchical, maintaining independently stable, and conserved, secondary folds and tertiary structures.⁹ This property implies that RNA's function is strongly correlated with its secondary fold or topology. Analysis of the secondary structure, though certainly incomplete, is less complex than the tertiary structure and provides an excellent starting point for investigating RNA structures. This advantage was exploited in our recent articles exploring the theoretical RNA secondary structure repertoire and their classification, where we employed graphical representations to enumerate, construct, and analyze two-dimensional (2D) RNA secondary topologies.^{10,11} Indeed, various 2D graphical representations have already shown to be useful for comparing and analyzing RNA secondary structures.^{12–14} RNA graphs specify the connectivity of the secondary elements such as stems, loops, bulges, and junctions. Although the level of representation is coarser than the atomic-level description, RNA graphs capture the essential topological properties defining known RNA families.

A key advantage of this topological approach is that both existing and hypothetical RNA structural motifs can be systematically generated and analyzed, thereby opening a new avenue for predicting novel RNA motifs and estimating their abundance in the RNA structure universe. This strategy is similar to combinatorial chemistry methods for generating structural diversity and for discovering novel compounds.¹⁵ Thus, predicting novel 2D RNA topologies by specifying the motif connectivity patterns may be considered as the equivalent problem of determining the number of protein fold classes. The more difficult problem of determining all 3D RNA structure classes will require both theoretical and experimental efforts.

In a previous work, we have heuristically generated small RNA topologies and described existing and “missing” motifs.¹⁰ Here, we report candidate novel RNA-like motifs, or topologies possessing properties similar to existing RNAs, using graph theory and clustering methods; these candidates may prove useful for experimental work in the growing field of RNA design.^{16–19} Specifically, we develop and apply graph theory methods to generate libraries of theoretical RNA topologies of different sizes, quantitatively describe topological properties (i.e., topological descriptors) of RNA motifs, and distinguish different topological types (pseudoknots and non-pseudoknots). Our topological descriptors are derived from the eigenvalues of the (Laplacian) matrix specifying the patterns of connectivity in the topology.¹¹ Since not all theoretical RNA topologies are physically meaningful, or RNA-like, we partition libraries of topologies into two groups to identify the RNA-like group using topological descriptor variables and a clustering algorithm Partition Around Medoids (PAM).²⁰

Identification of the RNA-like group in a given

set of theoretical topologies is central to our prediction scheme. Significantly, we find that most existing or natural topologies for 60–80 nt RNAs fall in the RNA-like group; we consider the remaining motifs less RNA-like. In the RNA-like group, the yet unfound topologies constitute viable candidates for novel RNA motifs. Specifically, we found ten novel motifs in the range of 60–80 nt containing natural RNA substructures, half of which are pseudoknots. We have also proposed candidate sequences that might fold into these motifs. Larger (>80 nt) candidate RNA motifs are similarly predicted using our clustering procedure, leading to sets of RNA-like motifs of different sizes (Figures 7–10, below). We find that the number of novel motifs increases with RNA size. Interestingly, existing and candidate RNA-like motifs have similar topological (Z-score) profiles, whereas the profile for non-RNA motifs is distinct, implying that natural motifs have specific topological properties. Generally, members of the RNA-like group are compact pseudoknots and branched trees, whereas non-RNA topologies are unbranched trees and pseudoknots with domains joined by single strands.

Another major prediction of our analysis is the abundance of pseudoknots, trees, and other motifs in the RNA secondary structure universe. We find that the proportion of pseudoknots rises rapidly with RNA size, exceeding 90% when length is >120 nt, implying that the universe of RNA motifs is dominated by pseudoknots rather than trees. Significantly, our predicted trends for pseudoknots and non-pseudoknots in the range <100 nt agree with available data for natural motifs. These predictions were made possible by our development of a graph theory algorithm for characterizing fundamental motif types (e.g. pseudoknot, tree).

Information about RNA's structural repertoire, even for secondary topologies, will likely benefit both theoretical and experimental search for novel functional RNA molecules. For example, novel motifs can be designed theoretically¹⁷ and their functional properties determined experimentally. Indeed, modular design of functional RNA molecules *via* the RNA *in vitro* selection technology has been exploited for possible applications in biotechnology, chemistry and medicine.^{16,21} Our predicted motifs provide a rich source of RNA topologies for such design efforts. Another emerging application of novel RNA topologies is in the computational search for novel RNA genes. Current RNA gene searches are based on sequence conservation and information about specific sequence motifs.^{22–24} Combining this approach with the knowledge of novel RNA topologies can lead to a more effective and comprehensive search for RNA genes in genome sequences.

The remainder of this article is organized as follows. The Methods section presents various techniques and algorithms for representing, describing, characterizing, and clustering RNA secondary structures. The Results section presents

a clustering analysis of sets of RNA topologies into RNA-like and non-RNA groups and estimation of the proportion of pseudoknots and trees in RNA structure space. The Discussion section compares topological and sequence approaches to finding novel RNAs, explains factors determining RNA motif abundance, and explores similarity/dissimilarity of protein and RNA structural repertoires. In the Conclusion section, we briefly discuss cataloguing of our novel motifs and future directions.

Methods

The methods and algorithms presented below include: graphical representation of RNA; the Laplacian eigenvalue spectrum and associated transformations as topological descriptors; a probabilistic graph-growing algorithm to generate RNA graphs; a clustering method called PAM for grouping possible RNA graphs; an algorithm for discriminating pseudoknot and non-pseudoknot motif types; and a Z-score for profiling the significance RNA topologies.

Graphical representation of RNA

Discrete graphical representations of RNA structures have the advantage that topologically distinct motifs can be theoretically enumerated and analyzed using available graph theory methods. The rules for representing RNA pseudoknot and non-pseudoknot motifs as dual graphs are detailed in our previous work.¹⁰ RNA tree graphs have been widely used for comparing RNA structures, but they are limited to tree structures.^{12,13} Thus, dual graphs are more general than tree graphs. Here, we use only dual graphs in our analysis. We map an RNA secondary structure onto a dual graph by the following rules (see Figure 1(a), inset). (1) A vertex (●) represents a double-stranded helical stem with ≥ 2 complementary bp. (2) An edge (–) represents a single strand that may occur in segments connecting the secondary elements (e.g. bulges, loops, junctions, and stems), where a bulge has more than one unmatched nucleotide or non-complementary base-pair; we consider AU, GC and GU as complementary base-pairs. Essentially, RNA graphs represent the topological properties of the connectivity pattern among RNA's secondary structural elements, such as loops, bulges, stems, and junctions; for example, the star-shaped transfer RNA (tRNA) is topologically distinct from the branched structure of 5S ribosomal RNA. Each vertex in a dual graph represents about 20 nt. The complete sets of RNA dual graph from two to four vertices, together with examples of larger topologies, are documented on our RNA-As-Graphs (RAG) database†; RAG also separately catalogues RNA tree graphs with up to ten vertices.

The double helical stems of RNAs imply specific rules for the construction of RNA dual graphs. An interior stem is connected to other stems by four strands (two incoming and two outgoing), whereas end stems (at the 3' and 5' ends) can have fewer connectors (see Figure 1(a), inset). A stem with four emanating strands is represented by a vertex with four radiating (incident) edges (see nodes labeled with 4 in Figure 1(a), inset). If the ends occur on the same stem, then it is represented by a vertex with two incident edges (see node labeled with 2). The case where the ends terminate on two different stems, the two stems are represented as vertices with three incident edges each (see node labeled with 3). Thus, the two allowed vertex patterns for RNA dual graphs are $(4, \dots, 4, 2)$ and $(4, \dots, 4, 3, 3)$, which are also called degree sequences in graph theory.²⁵

Topological descriptors of RNA graphs: spectrum of the Laplacian matrix and associated transformations

An RNA secondary topology can be intuitively characterized by the number of hairpin loops, junctions, bulges, and stems, as well as their connectivity properties. These parameters are examples of RNA topological descriptors. Here, we develop topological descriptors based on spectral representations of the connectivity of RNA topologies. In graph theory, the connectivity of an RNA graph is quantified using an adjacency matrix (A). The adjacency matrix and associated eigenvalues also aid characterization of isomorphism (similarity) between graphs (S. Pasquali, H.H.G. & T.S., unpublished results). Specifically, we use the V by V Laplacian matrix (L) representation of V -vertex graphs constructed from the symmetric adjacency (A) and degree (D) matrices. Namely, the square Laplacian matrix $L(G)$ of a graph G with vertices $1, 2, \dots, V$ is defined as $L = D - A$, where each element A_{ij} specifies the number of links or edges connecting i and j vertices, and D_{ii} specifies the degree of connectivity of vertex i . A V -vertex graph is characterized by the ordered eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \lambda_V$ of $L(G)$, called the spectrum of G , which is independent of the labeling of graph vertices. If the Laplacian spectra of two graphs are different, the graphs are not isomorphic; the converse, however, is not true because identical spectra can be associated with different topologies.²⁶

The pattern of a graph's connectivity is related to its eigenvalue spectrum. For example, the number of zeros in the spectrum represents the number of disconnected components of the graph. The second eigenvalue λ_2 measures compactness: a linear chain has a smaller second eigenvalue than a branched structure. The Laplacian eigenvalues are examples of topological descriptors; many other molecular structure descriptors have also been used.^{13,27,28}

To capture essential topological features of an RNA dual graph, we reduce the number of descriptors from the set of positive Laplacian

† <http://monod.biomath.nyu.edu/rna>

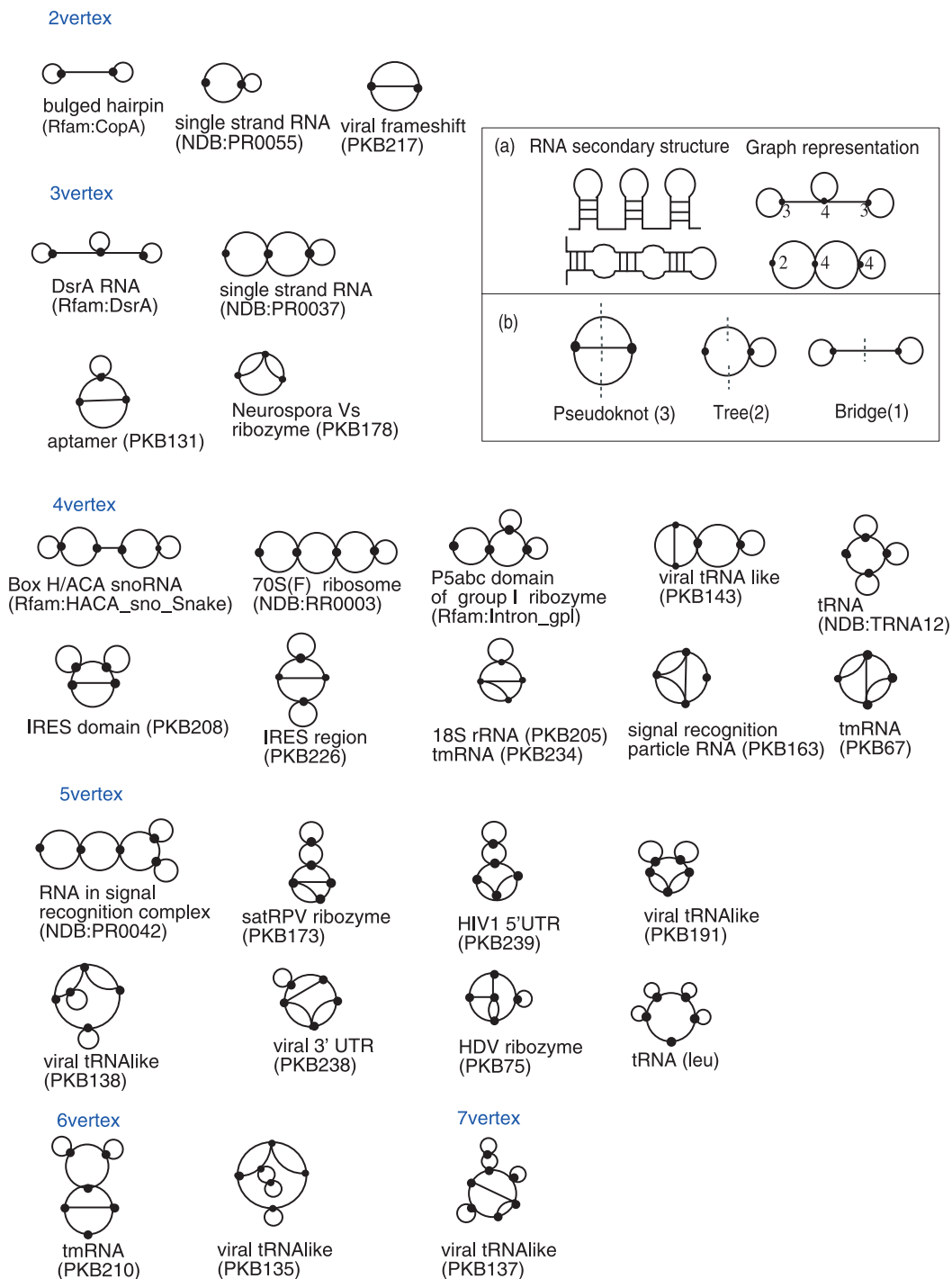


Figure 1. Existing RNA topologies for two to seven vertices and (inset) RNA graph representations. The 28 existing RNA topologies include both entire RNAs and RNA domains. The RNAs are from Nucleic Acids Database (NDB), RNA families database (Rfam), and Pseudobase (PKB). In the inset box, the RNA secondary structures and their dual graph representations, with labeled vertex degrees (incident edges) are shown in (a), and the minimal edge-cut numbers for three RNA motif types: pseudoknot, tree, and bridge, are shown in (b).

eigenvalues $\lambda_2, \dots, \lambda_V$ to two variables α and β : the positive slope β and the intercept α are calculated using the least-squares method applied to $\lambda_2, \dots, \lambda_V$. Thus, β measures the average spacing between positive eigenvalues, and the intercept α represents the second largest eigenvalue calibrated by β . This variable reduction is commonly used in clustering analysis such as performed in drug design, known

as quantitative structure–activity relationships (QSAR).²⁹

We found that β decreases with V , and therefore, it is reasonable to consider $V^*\beta$ as a quantity independent of V . We thus characterize an RNA dual graph using α and $V^*\beta$. That is, we represent each graph by $(V^*\beta, \alpha)$ and use these quantities to perform clustering of RNA-like and non-RNA-like

Table 1. Number of theoretical, existing, and candidate (RNA-like) RNA topologies with two to nine vertices; each vertex represents about 20 nt

| | V , vertex no. | | | | | | | |
|-------------|------------------|---|----|-----|-----|------|--------|--------|
| | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Theoretical | 3 | 8 | 30 | 108 | 494 | 2388 | 12,184 | 38,595 |
| Existing | 3 | 4 | 10 | 8 | 2 | 1 | | |
| RNA-like | | 2 | 8 | 64 | 307 | 1604 | 8777 | 25,810 |

motifs. For example, in Figure 1(a) (inset), a linear chain graph with three vertices and three loops is characterized by $\alpha = -1$ and $V^*\beta = 6$ ($\lambda_2 = 1$ and $\lambda_3 = 3$). In contrast, the more compact graph with three vertices and no loop (*Neurospora* VS topology in Figure 1) yields $\alpha = 2$ and $V^*\beta = 6$ ($\lambda_2 = 4$ and $\lambda_3 = 6$).

Further information for graph characterization can be gained by considering powers of the Laplacian spectrum. By definition, L satisfies: $L^n x_i = \lambda_i^n x_i$, where $n = 1, 2, \dots, N$, $i = 1, \dots, V$ and x_i is an eigenvector corresponding to λ_i^n . Thus, by variable reduction we can associate the spectrum $\{\lambda_i^n\}$ of L^n to $(V^*\beta_n, \alpha_n)$, where β_n and α_n are the spectrum's slope and intercept, respectively. For an N order Laplacian, the descriptor variable set is $\{(V^*\beta_1, \alpha_1), \dots, (V^*\beta_N, \alpha_N)\}$; N will be determined by clustering analysis in Results.

Computational generation of RNA graphs using a graph-growing algorithm

We can theoretically generate the space of all RNA secondary topologies using the rules for constructing dual graphs, i.e. each graph must satisfy the vertex-degree sequence $(4, \dots, 4, 2)$ or $(4, \dots, 4, 3, 3)$. Specifically, we use a probabilistic graph-growing method to generate each connected V -vertex RNA graph. In this stepwise algorithm, two initial vertices are randomly selected from a set of V vertices and connected either by one, two, or three edges. The number of edges connecting any two vertices is chosen based on a random uniform distribution.³⁰ To proceed, the next vertex is randomly picked and connected to the previous vertex until all V vertices are connected. The graphs generated by this method are automatically connected, since the previous vertex is selected from a connected subgraph. The above process is reiterated so as to generate an ensemble of RNA dual graphs after many graph-growing cycles.

Our algorithm can lead to isomorphic graphs (i.e. equivalent topologies). We thus trim the ensemble by removing isomorphic graphs to ensure that no Laplacian spectra are repeated. This method is imperfect, since some non-isomorphic graphs have the same Laplacian spectrum, but the error for dual graphs is generally a few percent.^{11,31} Indeed, no existing graph spectra (invariants) are capable of discriminating all non-isomorphic graphs. We use the Laplacian spectra to determine the convergence of our graph growing cycles by ensuring that the number of non-isomorphic graphs in the ensemble

of dual graphs is not increasing. We verified that our graph-growing algorithm and isomorphism test can successfully generate $V = 2, 3, 4$ dual graphs, which were previously determined heuristically.¹⁰ For the cases of $V = 5, 6, 7$ and 8 , our algorithm generates 108, 494, 2388 and 12,184 distinct RNA dual graphs, respectively, as listed in Table 1.

Clustering of RNA graphs into RNA-like and non-RNA-like groups using partitioning around medoids (PAM)

To distinguish RNA-like from non-RNA-like graphs, we employ the well-known clustering technique called PAM (pam function implemented in the cluster library package of R, an environment for statistical computing and graphics† to group enumerated graphs.²⁰ Essentially, to cluster k most diverse groups, PAM chooses k centering points of the distribution and determines which point is included in which group. This algorithm computes k representative objects, or medoids, and assigns members in each group by minimizing the medoid's dissimilarity to all the objects in the cluster. This procedure is repeated until the total Euclidean distance between the medoids and the objects in the group converges. Recently, PAM was used to analyze protein structure similarity.³²

Quantitatively, to analyze similarity and diversity of RNA graphs, we use the Euclidean distance $\delta_{ij} = f(G_i, G_j)$ for two graphs G_i and G_j based on the topological descriptors $(V^*\beta_n, \alpha_n)$ for $n = 1, \dots, N$ of an RNA dual graph. We then construct the symmetric matrix $D = \{\delta_{ij}\}$ for $i, j = 1, \dots, K$, where K is the number of graphs considered.²⁹ To find the similarity within a group, the PAM algorithm minimizes the total Euclidean distance between members in the group, i.e. the intragroup distance. To search the diversity, PAM selects the medoids to maximize the intergroup distance for k groups. PAM combines the search for similarity and diversity in the set with K graphs by iteratively optimizing selection of k representatives, maximizing intergroup distance, and minimizing intragroup distance.^{33,34}

To visualize the PAM clustering, we project 2N dimensions of the descriptors $(V^*\beta_n, \alpha_n)$, $n = 1, \dots, N$ of an RNA dual graph into m (usually two or three) dimensions, using the multi-dimensional scaling (MDS) method.^{29,33} The MDS projection preserves the original distance matrix $D = \{\delta_{ij}\}$ as possible. The

† <http://www.r-project.org/>

m -dimensional vector for an RNA dual graph G_i is determined by the singular value decomposition (SVD) method in numerical linear algebra.³⁵ We use MDS as implemented in the `cmdscale` function from the multivariable analysis library package of R[†].

Characterization of RNA pseudoknot and non-pseudoknot graphs

Our dual graph generation algorithm produces large numbers of RNA graphs belonging to pseudoknot and non-pseudoknot motif types; Table 1 shows that there are over 12,000 eight-vertex (~ 160 nt) graphs. The RNA secondary structure universe can be partitioned into tree, pseudoknot and bridge motif types (see Figure 1(b), inset).¹⁰ An RNA bridge refers to a topology with substructures that are connected by a single strand (a more precise definition is given below). Bridge motifs are biologically meaningful because they can suggest modular RNA motifs. The abundance of various RNA types in Nature can be predicted by characterizing or determining the RNA type of each hypothetical graph. We develop an algorithm based on graph properties to automate the process of differentiating pseudoknot, tree and bridge motifs. Our algorithm is based on the concept of edge cut, or equivalently, Eulerian tour.³⁰ Here, edge-cut is defined as the minimal number of edges whose removal makes the graph disconnected.

An RNA secondary structure is a pseudoknot if its graphical representation contains a non-Eulerian subgraph. An Eulerian tour is a walk defined as an alternating sequence of vertices (v_i) and edges (e_i), $v_0, e_0, v_1, e_1, \dots, v_V$, ending with the starting vertex, $v_0 = v_V$, where all e_i are distinct.³⁰ An Eulerian graph is a graph that has an Eulerian tour. A well-known property of an Eulerian graph is that it should have an even number of edge cuts.^{25,36} RNA trees, pseudoknots and bridges can be categorized as distinct topological types by their edge-cut property. As shown in Figure 1(b) (inset), RNA pseudoknots are characterized by at least a minimal edge-cut with three edges, RNA trees by a minimal edge-cut with two edges, and RNA bridges by a minimal edge-cut with one edge. A bridge graph can also be a pseudoknot if it contains a pseudoknot subgraph. To be consistent with the biological literature, we classify a graph as a pseudoknot if it has a pseudoknot subgraph. (Note that in our previous usage a bridge graph with a pseudoknot subgraph is also a bridge graph.¹⁰) Likewise, we classify a graph as a bridge even though it has a tree subgraph with a bridge structure. Based on these conditions, we define the RNA pseudoknot, bridge and tree topological types as follows.

Let G be any RNA graph, then: (1) G is a pseudoknot if and only if there exists three or four edge-cuts from the same vertex of G such that G becomes an $(M + 1)$ -component graph, where M is

the number of disconnected components of G after all bridges are removed. (2) G is a bridge if and only if there exists one edge cut of G that creates two components. (3) G is a tree if and only if it is neither a pseudoknot nor a bridge.

By using these definitions for RNA motif types, we construct an algorithm to determine the character of any RNA graph. We note that self-loop (or hairpin loop) structures do not affect the graph character, because it is determined only by the connectivity between different stems or vertices. Consequently, self-loops are removed in our calculations. Our algorithm is as follows. First, we remove all edges having one edge-cut property. We can count the number of components by the number of zeros of the spectrum of Laplacian.³⁷ Second, we check which of the following cases is true.

Case 1. If there is a component of G with three or four edge-cut property, we characterize the graph as a pseudoknot.

Case 2. If the number of components of G after removing all edges with one-edge-cut property is greater than two and G has no pseudoknot subgraph (Case 1), we characterize G as a bridge.

Case 3. If G is neither case 1 nor case 2, we characterize G as a tree.

Our algorithm can characterize the topological types of all enumerated RNA graphs.

Z-score for profiling the significance of RNA motifs

We consider a Z-score to profile sets of RNA graphs based on the topological descriptors $\{V^* \beta_1, \alpha_1, V^* \beta_2, \alpha_2, \dots, V^* \beta_N, \alpha_N\}$. For t_{iG} , descriptor i of graph G , we define the normalized Z-score as:

$$Z_{iG} = (t_{iG} - \langle t_{iG} \rangle) / \text{std}(t_{iG})$$

where $\langle t_{iG} \rangle$ and $\text{std}(t_{iG})$ are the mean and standard deviation of each descriptor in a set of graphs. The significance profile of descriptor i of graph G is:

$$SP_{iG} = Z_{iG} / \sqrt{\sum_{i=1}^{2N} Z_{iG}^2} \quad (1)$$

It is a normalized vector of Z-scores. Profiling a set of graphs using SP_{iG} allows comparison of subgroup properties (e.g. RNA-like or non-RNA-like) within the set.

Results

We examine several aspects emerging from our RNA graph analysis: clustering of RNA motifs into RNA-like and non-RNA-like groups; significance profiles of RNA motifs; candidate novel RNA motifs; and relative abundance of pseudoknot, tree and other motifs in RNA universe. Our analysis relies on existing RNA topologies compiled in Figure 1 from various sources in the literature and

† <http://www.r-project.org/>

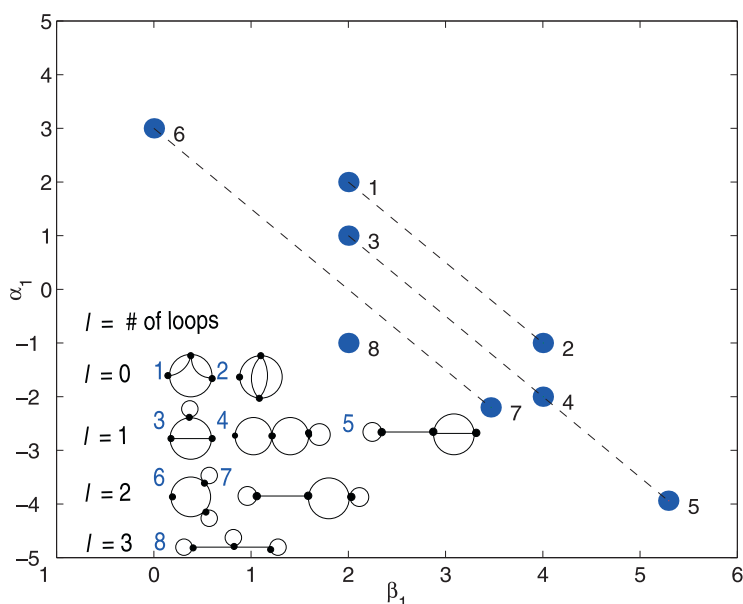


Figure 2. Mapping of eight three-vertex RNA dual graphs using topological descriptors α_1 and β_1 .

databases.^{38,39} The 28 existing RNAs range from two to seven vertices, or ~ 40 nt to ~ 140 nt, representing both whole RNAs and RNA domains. The set covers all topological (tree, pseudoknot and bridge) types. They are fewer large (>140 nt) RNA topologies for systematic analysis.

Topological descriptors map RNA motif classes and compactness

We examine the RNA motif features displayed by topological descriptors α_1 and β_1 by plotting in Figure 2 the intercept α_1 versus slope β_1 corresponding to the spectrum of L for eight three-vertex RNA

graphs. Interestingly, the α_1, β_1 plot maps RNA topologies according to characteristics such as compactness and the number of hairpin loops. Figure 2 has four linear lines representing possible zero, one, two, three hairpin loops for three-vertex graphs or motifs. For example, the three one-loop motifs lie on a straight line; there is only one three-loop motif. Similarly, for the 30 four-vertex motifs, there are five lines representing zero to four possible hairpin loops (data not shown). Generally, a linear topology with one or more “bridge” strands has a large β_1 and a small α_1 , whereas a compact graph has a small β_1 and a large α_1 . Previously, we showed that the second eigenvalue λ_2 increases

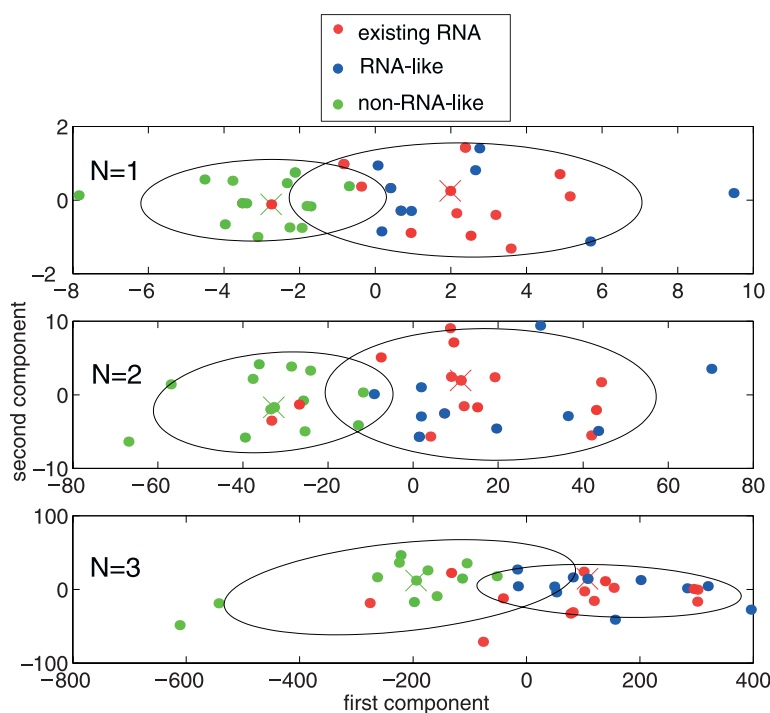


Figure 3. 2D clustering plots of the MDS transformation for 38 (three- and four-vertex) RNA dual graphs using Laplacian orders $N=1, 2, 3$. Existing RNA, RNA-like, and non-RNA-like topologies are color coded as red, blue, and green, respectively. Each ellipse encloses at least 85% of the RNA-like or non-RNA-like group members.

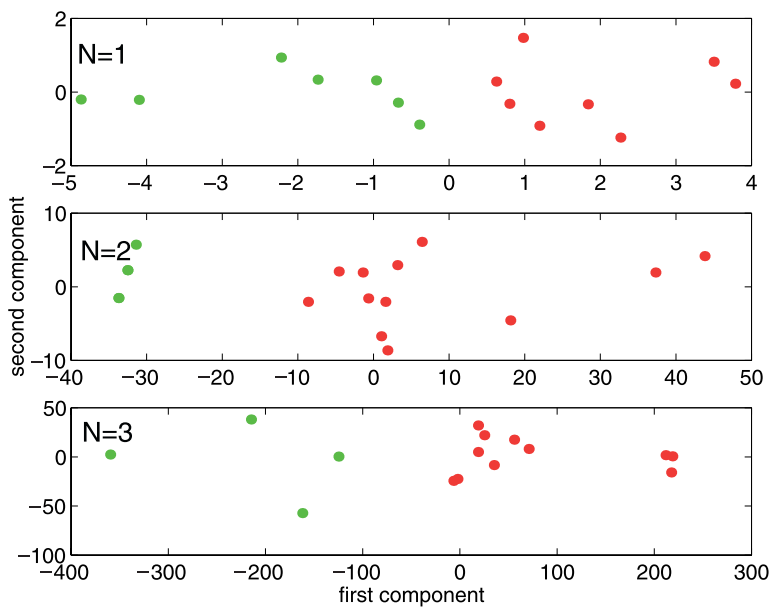


Figure 4. 2D clustering plots of the MDS transformation for 15 existing three and four-vertex RNA dual graphs using Laplacian orders $N=1, 2, 3$.

with motif compactness.¹¹ Thus, the α_1, β_1 map organizes possible RNA motifs into hairpin-loop classes which are arranged by compactness. Note that the four existing three-vertex RNAs with zero, two, three hairpin loops display no discernible clustering pattern different from hypothetical RNA motifs.

Clustering of topologies into RNA-like and non-RNA-like groups

Figure 3 shows 2D clustering of 38 (eight three-vertex and 30 four-vertex) RNA topologies into two groups using the PAM algorithm. The smaller RNA topologies are used since there are 15 existing motifs. To aid visualization, we have also drawn an ellipse to enclose 85% of topologies in each group. As shown, we applied PAM on the same 38 graphs using Laplacian orders ($N=1, 2$ and 3); $N=1$ denotes clustering with $(V^*\beta_1, \alpha_1)$ descriptor variables, $N=2$ with both $(V^*\beta_1, \alpha_1)$ and $(V^*\beta_2, \alpha_2)$ descriptor variables, and so on. Significantly, in all $N=1, 2, 3$ clusterings, at least 12 out of 15 existing RNA motifs are clustered in one group, meaning that the natural RNAs are topological neighbors. The numbers for topologies clustered in the non-RNA-like group are 12, 15 and 15 for $N=1, 2, 3$, respectively; the number of misclassified motifs (existing RNAs in the group) is less than three. Increasing the Laplacian order from 2 to 3 yields similar results, implying that $N=2$ is near optimal for clustering RNA topologies. Moreover, 3D clustering yields results similar to those in 2D projection.

We call the group with the majority of existing RNAs the “RNA-like” class and the other “non-RNA-like.” Interestingly, the RNA-like group contains 9 to 11 motifs ($\sim 30\%$ of total) not yet found in Nature. We call these RNA topologies the candidate novel RNA motifs. Our 2D clustering analysis

suggests that these predicted motifs are topologically similar to existing RNAs. To ascertain that our clustering procedures are not fortuitous, we also applied the PAM clustering to the 15 existing RNA motifs belonging to three to four vertex graphs. Figure 4 shows that for Laplacian orders 2 and 3 there is a disproportionate number (11 or 12) of graphs in one group. Thus, forcing existing RNAs to cluster into two groups does not yield a random partition of motifs. For $N=2$ case (middle panel, Figure 4), the three members of the smaller group (green dots) are DsrA RNA (Rfam:DsrA), tRNA (NDB:TRNA12), and IRES RNA (PKB:226).

Another prediction from our clustering analysis includes the medoids of the RNA-like and non-RNA-like groups. A group’s medoid is its “center of gravity” exhibiting the common topological features of the group. In Figure 3, the medoids for $N=1, 2$ and 3 are four-vertex motifs of viral 5'-UTR (PKB209), tmRNA (PKB234) and P5abc domain of group I intron ribozyme. The viral 5'-UTR and tmRNA are pseudoknots, whereas the P5abc domain is a tree structure. In the optimal $N=2$ clustering, the occurrence of tmRNA as the center of the RNA-like group is interesting because most topologies are pseudoknots and there are two distinct four-vertex tmRNA motifs (PKB234 and PKB67). The tmRNA has both tRNA-like and mRNA-like functions. It participates in a translation process where the unfinished protein is tagged for degradation and release from the stalled ribosome due to a defective mRNA.⁴⁰

Significance profiles of RNA motifs

Significance profiling of RNA networks provides another perspective on the clustering of RNA-like and non-RNA-like groups. This approach was recently applied to complex biological, technological, and sociological networks to compare and

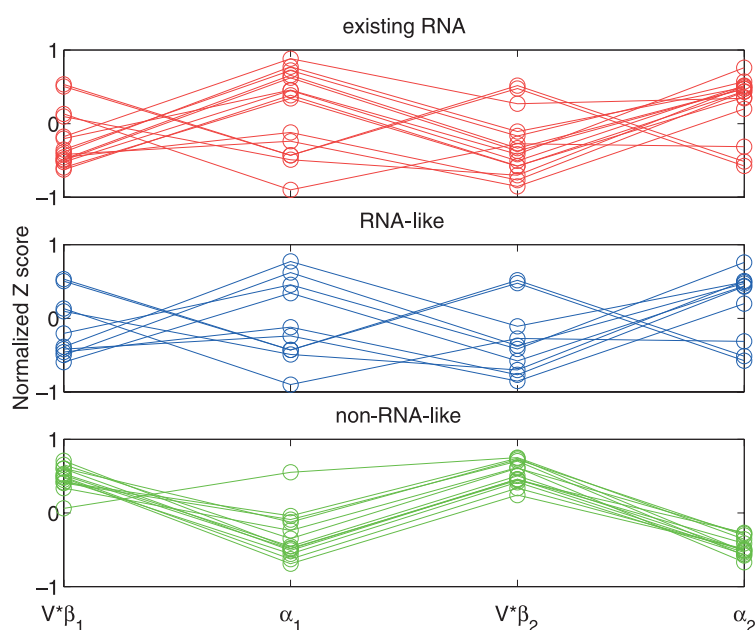


Figure 5. Significance profiles (Z scores) of existing RNA, RNA-like and non-RNA-like topologies with respect to four topological descriptors.

identify superfamilies of network structures; biological networks in protein signaling, developmental genetic networks, and neuronal wiring belong to the same superfamily.⁴¹ Figure 5 plots the normalized Z-score versus the four topological descriptors (Laplacian order 2) for 38 (three and four-vertex) existing RNA, RNA-like and non-RNA-like motifs deduced from the clustering analysis above. As defined in equation (1), the normalized Z-score is a measure of the deviation of

a descriptor variable t_{iG} from the mean value $\langle t_{iG} \rangle$ for a set of graphs considered. The figure shows that the 15 existing RNA topologies display a pattern distinct from the 13 motifs clustered as non-RNA-like graphs. Remarkably, our ten predicted novel RNA-like motifs (Figure 7) have a significance profile rather similar to that for existing RNAs. Still, some existing and candidate RNA motifs display patterns deviating from the general profile. Thus, the RNA-like topologies may be considered

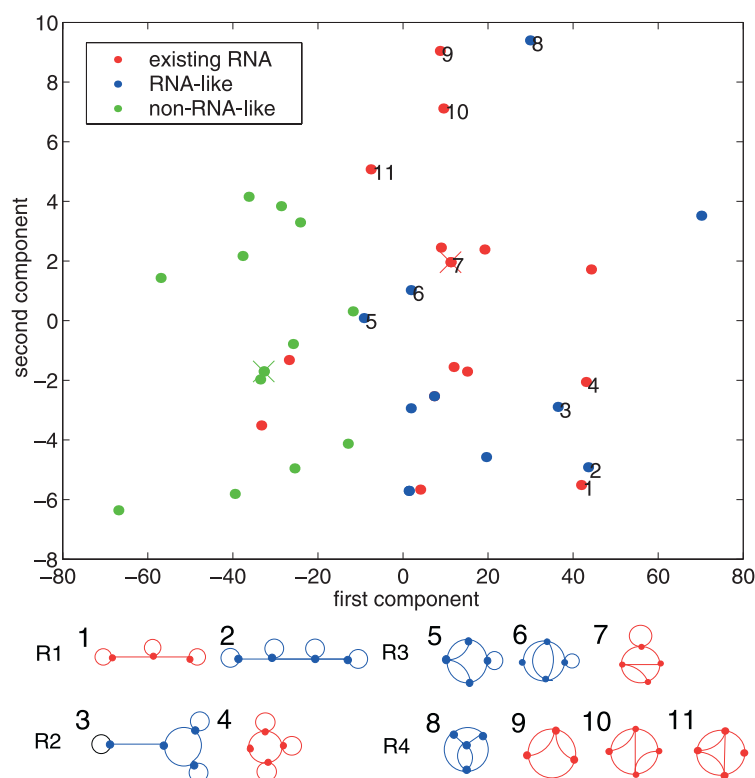


Figure 6. 2D visualization of three and four-vertex RNA graphs for Laplacian order 2 showing subgroups (R1,...,R4) with topologically similar RNAs. Existing RNA, RNA-like, and non-RNA-like motifs are color coded as red, blue and green, respectively. Red and green crosses (X) are medoids (centroids) of RNA-like and non-RNA-like groups, respectively.

as belonging to the same network family. Overall, significance profiling essentially confirms the validity of clustering possible RNA graphs into RNA-like and non-RNA groups.

Topologically similar subgroups revealed by motif clustering

Figure 6 illustrates four subgroups (R1, R2, R3 and R4) within the RNA-like cluster with very similar topologies or connectivity patterns. R1 subgroup contains an existing linear, three-stem-loop structure and a predicted linear four-stem-loop motif. The tRNA fold with four-stem junction is in the R2 subgroup, which also contains a predicted RNA-like bridge motif with a three-stem junction. The R3 and R4 subgroups contain only pseudoknot motifs. R3 contains the tmRNA motif (R3-7), the group's medoid, and two other predicted pseudoknots, one (R3-5) of which only differs from the tmRNA topology by the location of a stem-loop. The R4 subgroup has three similar existing motifs (*Neurospora* VS, signal recognition particle RNA, tmRNA) with no hairpin loops and a

predicted topology (R4-8) similar to the HDV ribozyme (PKB75) except for the absence of a stem-loop. Analyzing topologically similar subgroups may allow inference of the functional properties of the predicted motifs.

Similarity of predicted novel topologies to structures of functional RNAs

Figure 7 displays the main results of this work: ten candidate novel RNA secondary motifs (C1,...,C10) predicted by our PAM clustering, including existing RNAs that are structurally similar to them. The expected size of the candidate motifs varies from 60 nt to 80 nt. Two of them are tree motifs (C2,C7), four are bridge motifs (C1,C3,C4,C5), and four are pseudoknots (C6,C8,C9,C10). A comparison of the candidates with existing motifs in Figure 1 suggests that at least five of the predicted RNA-like topologies are similar to those for functional RNAs. For example, snoRNA, a molecule involved in post-transcriptional modification of other RNAs, is topologically similar to candidate motif C1 but with a missing stem-loop element. The three-stem junction motif

| Graph representation with natural submotif | RNA secondary structure with natural submotif | Similar existing topology |
|--|---|--|
| C1 | single strand RNA (NDB:PR0055) | Box H/ACA snoRNA (Rfam:HACA_sno_Snake) |
| C2 | bulged hairpin (Rfam:CopA) | tRNA (NDB:TRNA12) |
| C3 | DsrA RNA (Rfam:DsrA) | |
| C4 | bulged hairpin (Rfam:CopA) single strand RNA (NDB:PR0055) | |
| C5 | bulged hairpin (Rfam:CopA) | |
| C6 | DsrA RNA (Rfam:DsrA) | |
| C7 | single strand RNA (NDB:PR0055) | P5abc domain of group I ribozyme (Rfam:Intron_gpl) |
| C8 | single strand RNA (NDB:PR0037) | |
| C9 | DsrA RNA (Rfam:DsrA) | Neurospora Vs ribozyme (PKB178) |
| C10 | single strand RNA (NDB:PR0037) | HDV ribozyme (PKB75) |

Figure 7. Ten predicted candidate novel RNA motifs with about 60 nt to 80 nt deduced from the clustering in Figure 3, which are RNA-like topologies in the case of Laplacian order 2.

C2 may be viewed as a substructure of the four-stem junction tRNA. The linear four-stem-loop structure C3 is similar to the linear three-stem-loop motif of DsrA RNA. We have found no apparent structural similarity between existing RNAs and C4, C5 and C6 topologies. The four-vertex tree motif C7 is similar to the topology of P5abc, a domain of group I intron ribozyme, except for the location of a hairpin loop.

Interestingly, the complex topology of candidate pseudoknot C9 appears to be a variation of several existing pseudoknots, including viral frameshifting (PKB178), HIV-1 5'-UTR (PKB239), and virial tRNA-like (PKB191). The common motif of these pseudoknots is the core double-pseudoknot (PKB178) with addition of one or two hairpin loops to generate topological diversity. This theme holds also for predicted motif C10 and HDV ribozyme (PKB75), which has an extra stem-loop. We have identified topological similarities between the predicted novel motifs and existing RNAs to indicate their potential to be functional molecules, but not necessarily functionally related to the existing structures compared.

Many functional RNA topologies (Figure 1) are compact pseudoknots and moderately branched

trees (i.e. low order junctions¹⁰), with no pseudoknots possessing bridge substructures. The candidate topologies in Figure 7 have similar features; there is only one pseudoknot (C6) with a bridge substructure. For comparison, the topologies in the non-RNA-like group generally consist of pseudoknots with one or more single strands connecting the substructures of a motif, as well as several complex pseudoknots (see our RAG database).

It is also instructive to identify natural submotifs of our candidate RNA topologies. Since the candidate RNAs are small, we are restricted to consideration of tiny RNA structures such as bulged hairpin (Rfam:CopA), DsrA RNA (Rfam:DsrA), and single strands (NDB:PR0055, PR0037). The folds of single strands are found as submotifs of C1, C4 and C7. The linear stem-loop structure of DsrA RNA is found in C3, C6 and C9. The C2 and C5 topologies contain the bulged hairpin fold.

Identifying sequences folding into candidate topologies

To deduce possible sequences that will fold into the new candidate motifs, we suggest employing a

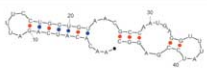

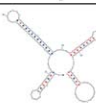
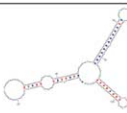
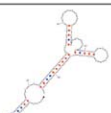
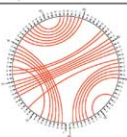

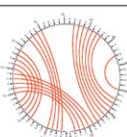
| ID | Designed sequence | Novel RNA structure |
|-----|---|---|
| C1 | AACACAUCAGAUUUCCUGGUGUAA CGCCAAUGAGGUUUUAUCCGAGGC |  |
| C2 | AGCGCCGUGGCAGGGCUCUAACC CUGAUGUCCUCGGAUCGAAACCGA GCGGCGCUACCA |  |
| C3 | AACACUCAGAUUUCCUGGUGUAA GAAUUUUUAAGUGCUUCUUGCUU AAGCAAGUUUCUACCCGACCCCU CAGGGUCGGGAUUUUGGACCUCCA UGACGUUAUGGUCC |  |
| C4 | AACACUCAGAUUGGACCUCAUGAC GUUAUGGUCCUCCUGGUGUAACG AAUUUUUAAGUGCUUCUUGCUUA AGCAAGUUUCUACCCGACCCCU AGGGUCGGGAUUU |  |
| C5 | CCUGGUAUUGCAGUACCUCCAGGU AGCGCCGUGGCAGGGCUCUAACC CUGAUGUCCUCGGAUCGAAACCG AGCGGCGCUACCA |  |
| C6 | AGACCGUCAAACACAGACUAAAUGU CGGUCGGGAAGAUGUAUUCUUCU CAUAAGAUUAAGUCGGCCUGGUU UGCAGUACCUCCAGGU |  |
| C7 | GGCAGUACCAAGUCGCGAAAGCGA UGAUGGUAAGCCUUGCAAAGGGUU AAGCUGCC |  |
| C8 | not yet found | |
| C9 | CUUCUUAUAUGAUUAGGUUGUCAU UUAGAAUAAGAAAACCUUGUAUUG CAGUACCUCCAGGUUAACCG |  |
| C10 | not yet found | |

Figure 8. Ten RNA sequences that fold into candidate novel RNA motifs in Figure 7 as constructed by a build-up procedure are shown.

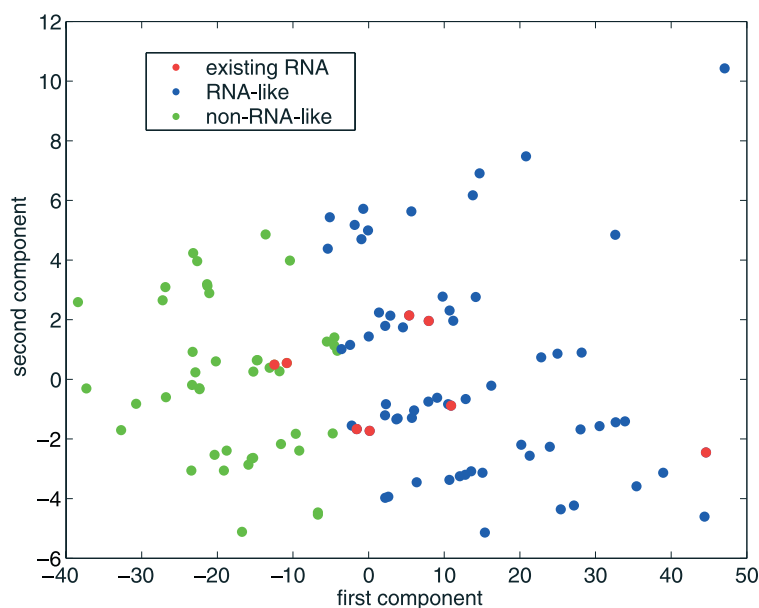


Figure 9. 2D clustering plot of the MDS transformation for five-vertex RNA dual graphs using Laplacian order 2.

modular design protocol, to be followed by experimental testing.¹⁰ The idea is to construct candidate topologies using small existing RNA structures and their associated known sequences. For example, we can propose sequences for the C1, C3, C4, C6 and C9 motifs using the following approach. For motif C1, a bridge structure with three stems, we combine a tree structure with two stems (NDB:PR0055) with a tree structure with one stem; the latter is a fragment of a bridge structure with three stems (DsrA RNA). We then “fold” this built-up structure using Mfold (Figure 8). In this case, Mfold indeed produces motif C1 as the optimal fold ($dG = -11.4$ kcal/mol). For C3, a bridge structure with four hairpins, we combine two existing structures, a bridge with three hairpins (DsrA RNA) and a hairpin structure (NDB:PTR016). For C4, we insert a fragment of a hairpin structure (PTR016) into a bridge structure with three hairpins (DsrA RNA). For C6–pseudoknot structure with a bridge, we add a tiny hairpin structure (PTR 016) to the end of an existing pseudoknot RNA, PKB77. For the pseudoknot structure C9, we insert a hairpin structure (PTR 016) into the middle of a pseudoknot (PKB216). These sequences fold into their candidate topologies using the PKNOTS program available from the Eddy group†. In Figure 8, we provide the results of sequences and structures using the *sir_graph* and Mfold programs from the Zuker group.

We can also consider cutting out or moving a fragment from an existing RNA structure into another. For example, we identified candidate sequences for C2 and C7 using this backward approach. For C2, we used tRNA structure and cut out a stem-loop. For C7, we moved a stem-loop from the P5abc domain of group I intron of *Tetrahymena thermophila*. These preliminary experiments already suggest several sequences that can be

examined in the laboratory and indicate that the above ideas are worth exploiting systematically.

Prediction of the relative abundance of pseudoknot, tree and other motifs

Our analysis and prediction of small (60–80 nt) RNA motifs can be extended to larger topologies where fewer or no existing RNAs are available. We overcome the paucity of larger existing RNAs by using the medoids determined from the clustering of smaller three and four-vertex topologies. We assume that their medoids for both RNA-like and non-RNA-like groups can be used to similarly cluster five, six, and higher-vertex RNA graphs into two groups. More precisely, we use fixed medoids for clustering all higher-vertex graphs. This procedure is expected to yield less accurate results with increasing vertex number. Figure 9 shows the PAM clustering of 108 five-vertex topologies where six out of eight functional topologies are clustered in the RNA-like group. Table 1 tabulates the number of topologies in the RNA-like group as a function of vertex number.

Figure 10(a) shows the proportion of pseudoknot motifs in the RNA universe as a function of vertex number, or size (~ 20 nt/vertex). The percentage of pseudoknots from our clustering analysis rises rapidly from 30% for three-vertex graphs to >90% for seven- and higher-vertex graphs (i.e. >140 nt). Note that pseudoknot topologies are determined using the edge-cut or Eulerian tour property. The trend for the existing RNAs up to five-vertex motifs is similar to our predicted pseudoknot abundance curve, with deviations of $\sim 10\%$. Quantitative agreement with data is not expected because RNA structures in databases are most likely not representative of the RNA structure universe. Another problem is that very few RNA structures are known, unlike protein structures. The pseudoknot

† <http://www.genetics.wustl.edu/eddy>

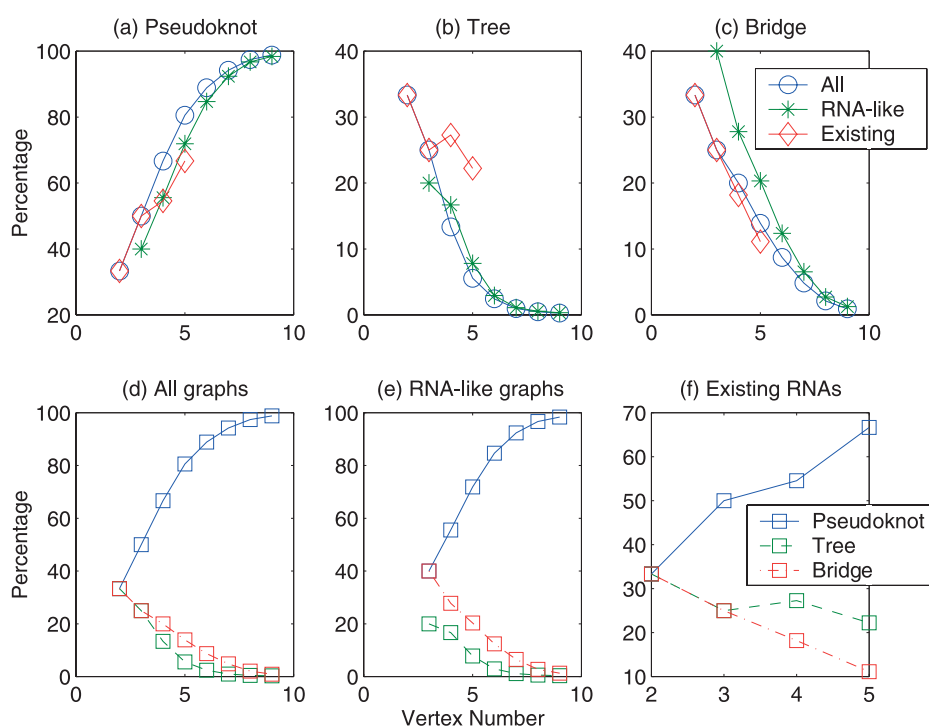


Figure 10. Size dependence (vertex number) of the relative proportion of pseudoknot (a), bridge (b), and tree (c) motifs in the RNA universe, and their distributions in all possible graphs (d), RNA-like (e) and non-RNA-like (f) groups.

abundance with clustering analysis is consistently lower by 10% than without (hypothetical) clustering analysis, except for seven and higher-vertex numbers where the results converge. The abundance of non-pseudoknots is computed by subtracting 100% from pseudoknot abundance.

The proportion of tree motifs decreases with RNA size (Figure 10(b)), from a high of 33% for two-vertex case to <10% for five and higher-vertex motifs. The same trend is observed with or without clustering analysis. The proportion of existing RNA trees also decreases with size, but the data do not quantitatively agree with prediction for four and five-vertex cases. A similar trend is seen for bridge motifs (Figure 10(c)) in agreement with data.

We also analyze the abundance trends of pseudoknots, trees and bridges in entire RNA, RNA-like and existing RNA groups (Figure 10(d)–(f)). The trends are the same in all groups considered: pseudoknot abundance increases with size whereas non-pseudoknot abundance (trees and bridges) decreases with size. For all sizes, our calculations show that tree topologies are less abundant than bridge motifs, which in turn is less abundant than pseudoknots. We thus conclude that the RNA universe is dominated by pseudoknots, especially for larger RNAs. It is somewhat surprising that the tree motif type is the least abundant considering the importance of RNA tree structures, such as tRNA and 5 S and 23 S ribosomal RNAs, in the development of the field.

Discussion

Advantages and disadvantages of the sequence and topological approaches for exploring RNA structure space

Conventional approaches to RNA genomics and the search for functional RNAs have focused on identifying sequences that fold to novel structures or corresponding to desired functions. For example, the analysis of genomes for novel non-coding RNAs relies on sequence conservation and specific sequence motifs.^{22,23,42} A number of candidate and novel RNAs have been identified in this way. Another active area is the identification synthetic functional RNAs from random sequence pools using *in vitro* selection techniques,^{43–46} which essentially explore the sequence space. Recent attempts to design novel RNAs by fitting sequences to structures highlight the complexity of the sequence space.^{17,47} Sequence-based approaches are principally limited by the astronomical size of the sequence space and by the lack of sequence conservation among many functionally and structurally related RNA molecules (e.g. tRNA).

In contrast to sequence-based approaches, our topological approach has the advantage that the RNA structure space can be comprehensively and efficiently explored. Of course, the skeletal graphical structures are starting points and lack information about sequence details. Still, our

combination of topology construction, clustering and characterization enables prediction of specific candidate RNA topologies (Figure 7) and the estimation of the abundance of various RNA types as a function of size in the RNA universe (Figure 10). To the best of our knowledge, this is the first attempt to assess the size of RNA structure space based on information of existing and hypothetical RNA motifs. Our topological approach focuses on global features of the RNA space, such as RNA motif connectivity and their properties, whereas conventional sequence-based approaches search for specific motifs (e.g. GNRA tetraloops) in functional RNAs. Thus, topological and sequence approaches are complementary. Combining their respective advantages in future analysis of genomes and design of functional RNAs is likely to lead to a more productive search from novel RNAs.

Topological descriptors and functional RNAs

Our clustering of existing and hypothetical RNA topologies into RNA-like and non-RNA groups is based solely on topological descriptors such as the scaled slope ($V^*\beta_i$) and intercept (α_i) of the Laplacian eigenvalue spectrum. Despite the absence of information about specific sequence motifs in our procedures, we find that most existing RNAs are clustered in the same group. Our analysis of the significance profiles of existing, candidate and hypothetical RNAs confirms that our topological descriptors can successfully discriminate RNA-like from non-RNA topologies. Furthermore, there are intriguing similarities between candidate and functional RNA topologies. Thus, our clustering results and analyses suggest that significant features of functional RNA structures are encoded in their topologies. For example, our previous survey showed that functional RNA tree topologies are moderately branched.¹⁰ Clearly, our level of description does not include detailed features of functional motifs like UNCG and GNRA tetraloop motifs, bulge-G, bulge-helix-bulge, U-turn, biloop and triloop, and A-stack, although their presence is implicit in our topological descriptors. Of course, 2D RNA structures and their graphical representations do not contain information about tertiary interactions, including interactions with metal ions which are integral to many RNA structures.^{48,49} Ultimately, knowledge of 3D structures is required to understand functional properties in detail.

Relative trends of RNA motif types are dictated by mathematical possibilities

Pseudoknots, trees and bridges are fundamental motif types of RNA secondary structures. Although numerous examples of each motif type are archived in various databases,^{38,39} their relative distributions have not been estimated before. Proportions are easier to estimate than the absolute motif number because not many existing RNAs are available for accurate prediction. This analysis was made

possible by our topology characterization algorithm, allowing automated distinction of motif types regardless of motif complexity. Our main prediction is that the RNA structure universe is dominated by pseudoknots instead of tree and bridge motif types. This reflects the many more mathematically possible ways to form pseudoknot than tree topologies, although tree motifs are better known through examples such as tRNA and 5 S and 23 S ribosomal RNAs. Of course, Nature also selects functional topologies based on energetic and folding criteria. The qualitative agreement between predicted and existing relative abundance of pseudoknots strongly suggests that Nature avails herself to numerous mathematically possible pseudoknot structures. Future estimates of motif abundance should consider energetic factors.

Comparing the RNA and protein structure universes

Current predictions of the size of protein structure universe rely considerably on extrapolation of the statistics of existing protein folds.^{7,50,51} Geometric and thermodynamic factors are known to influence the distribution of fold classes.⁵² A recent estimate suggests that the protein structure universe has ~16,000 structures of all sizes.⁷ In contrast, our estimate of the RNA structure universe is size dependent and based on topological constraints of RNA secondary structures. We predict that the number of candidate RNA motifs grows with RNA size. RNA size and constraints, including motif types, can be precisely specified in graphs, allowing exact enumeration and clustering analysis of the RNA structure space. However, no similar analysis is available for protein folds due to essential differences between RNA and protein molecules. Still, the four protein fold classes⁵³ (i.e. α , β , $\alpha + \beta$, and α/β), which are a basis of SCOP database classification,⁵⁴ may be regarded as equivalent to RNA pseudoknot, tree and bridge motif types. This comparison is reasonable because α and β are protein secondary structural elements. The distribution of RNA motif types is uneven (Figure 10), whereas protein fold classes are more homogeneous: the proportions of α , β , $\alpha + \beta$, and α/β classes are 25%, 20%, 30% and 25%, respectively.⁵¹

Conclusion

Existing, candidate and non-RNA topologies are systematically catalogued in our RAG database†. The topologies are organized by vertex number and by complexity as measured according to the Laplacian spectrum. In addition to the results for dual graphs discussed here, RAG catalogues results for tree motifs up to ten vertices, or a maximum RNA size of ~180 nt. We hope that RAG's

† <http://monod.biomath.nyu.edu/rna/rna.html>

cataloguing of existing and hypothetical motifs will help organize the universe of RNA motifs and stimulate the search for novel RNAs. One possibility, under development, is to systematically identify sequences that fold into the desired, candidate topologies by a modular build-up procedure similar to our preliminary folding experiments.^{10,55}

Acknowledgements

We thank Uri Laserson for the suggestion on Eulerian tour, Samuela Pasquali for discussions on graph-growing algorithms, and Daniela Fera for valuable assistance. This work was supported by Human Frontier Science Program (HFSP) and by a Joint NSF/NIGMS Initiative in Mathematical Biology (DMS-0201160).

References

- Doudna, J. A. & Cech, T. R. (2002). The chemical repertoire of natural ribozymes. *Nature*, **418**, 222–228.
- Gesteland, R. F., Cech, T. & Atkins, J. F. (1999). *The RNA World*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Storz, G. (2002). An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
- Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature Rev. Genet.* **2**, 919–929.
- Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Swaminathan, S. *et al.* (1999). Structural genomics: beyond the Human Genome Project. *Nature Genet.* **23**, 151–157.
- Chance, M. R., Bresnick, A. R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A. *et al.* (2002). Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.* **11**, 723–738.
- Vitkup, D., Melamud, E., Moulton, J. & Sander, C. (2001). Completeness in structural genomics. *Nature Struct. Biol.* **8**, 559–566.
- Doudna, J. A. (2000). Structural genomics of RNA. *Nature Struct. Biol.* **7**, 954–956.
- Tinoco, I. & Bustamante, C. (1999). How RNA folds. *J. Mol. Biol.* **293**, 271–281.
- Gan, H. H., Pasquali, S. & Schlick, T. (2003). Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucl. Acids Res.* **31**, 2926–2943.
- Gan, H. H., Fera, D., Zorn, J., Shiffeldrim, N., Tang, M., Laserson, U., Kim, N. & Schlick, T. (2004). RAG: RNA-As-Graphs Database—concepts, analysis, and features. *Bioinformatics*, **20**, 1285–1291.
- Le, S. Y., Nussinov, R. & Maizel, J. V. (1989). Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.* **22**, 461–473.
- Benedetti, G. & Morosetti, S. (1996). A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. *Biophys. Chem.* **59**, 179–184.
- Fontana, W., Konings, D. A. M., Stadler, P. F. & Schuster, P. (1993). Statistics of RNA secondary structures. *Biopolymers*, **33**, 1389–1404.
- Burke, M. D., Berger, E. M. & Schreiber, S. L. (2003). Generating diverse skeletons of small molecules combinatorially. *Science*, **302**, 613–618.
- Soukup, G. A. & Breaker, R. R. (1999). Engineering precision RNA molecular switches. *Proc. Natl Acad. Sci. USA*, **96**, 3584–3589.
- Andronescu, M., Fejes, A. P., Hutter, F., Hoos, H. H. & Condon, A. (2004). A new algorithm for RNA secondary structure design. *J. Mol. Biol.* **336**, 607–624.
- Davis, J. H. & Szostak, J. W. (2002). Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proc. Natl Acad. Sci. USA*, **99**, 11616–11621.
- Schultes, E. A. & Bartel, D. P. (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, **289**, 448–452.
- Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.
- Soukup, G. A. & Breaker, R. R. (2000). Allosteric nucleic acid catalysts. *Curr. Opin. Struct. Biol.* **10**, 318–325.
- Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**, 1369–1373.
- Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* **15**, 1637–1651.
- Carter, R. J., Dubchak, I. & Holbrook, S. R. (2001). A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.* **29**, 3928–3938.
- Read, R. C. (1959). The enumeration of locally restricted graph(I). *J. London Math. Soc.* **34**, 417–436.
- Cvetkovic, D. M., Doob, M. & Sachs, H. (1995). *Spectra of Graphs*. Johann Ambrosius Barth, Heidelberg.
- Randic, M. & Zupan, J. (2001). On interpretation of well-known topological indices. *J. Chem. Inf. Comput. Sci.* **41**, 550–560.
- Randic, M. (2001). Novel shape descriptors for molecular graphs. *J. Chem. Inf. Comput. Sci.* **41**, 607–613.
- Schlick, T. (2002). *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer, New York.
- Gross, J. L. & Yellen, J. (1999). *Graph Theory and its Applications*. CRC Press, Boca Raton, FL.
- van Dam, E. R. & Haemers, W. H. (2003). Which graphs are determined by their spectrum? *Linear Algebra Appl.* **373**, 241–272.
- Choi, I. G., Kwon, J. & Kim, S. H. (2004). Local feature frequency profile: a method to measure structural similarity in proteins. *Proc. Natl Acad. Sci. USA*, **101**, 3797–3802.
- Xie, D. X., Tropsha, A. & Schlick, T. (2000). An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-Newton minimization. *J. Chem. Inf. Comput. Sci.* **40**, 167–177.
- Xie, D. X., Singh, S. B., Fluder, E. M. & Schlick, T. (2003). Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases. *Math. Program.* **95**, 161–185.
- Trefethen, L. N. & Bau, D. (1997). *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia.

36. Fleischner, H. (2004). *Eulerian Graphs and Related Topics Annals of Discrete Mathematics*, vol. 45, chapt. 4. North-Holland, Amsterdam.
37. Mohar, B. (1991). The Laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications* (Alavi, Y., Chartrand, G., Oellermann, O. R. & Schwenk, J. A., eds), pp. 871–899, Wiley, New York.
38. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T. *et al.* (1992). The Nucleic-Acid Database—a comprehensive relational database of 3-dimensional structures of nucleic-acids. *Biophys. J.* **63**, 751–759.
39. van Batenburg, F. H. D., Gulyaev, A. P., Pleij, C. W. A., Ng, J. & Oliehoek, J. (2000). PseudoBase: a database with RNA pseudoknots. *Nucl. Acids Res.* **28**, 201–204.
40. Valle, M., Gillet, R., Kaur, S., Henne, A., Ramakrishnan, V. & Frank, J. (2003). Visualizing tmRNA entry into a stalled ribosome. *Science*, **300**, 127–130.
41. Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S. & Ayzenshtat, I. (2004). Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.
42. Eddy, S. R. (2002). Computational genomics of noncoding RNA genes. *Cell*, **109**, 137–140.
43. Ellington, A. D. & Szostak, J. W. (1990). *In vitro* selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
44. Wilson, D. S. & Szostak, J. W. (1999). *In vitro* selection of functional nucleic acids. *Annu. Rev. Biochem.* **68**, 611–647.
45. Hermann, T. & Patel, D. J. (2000). Biochemistry—adaptive recognition by nucleic acid aptamers. *Science*, **287**, 820–825.
46. Jaschke, A. (2001). Artificial ribozymes and deoxy-ribozymes. *Curr. Opin. Struct. Biol.* **11**, 321–326.
47. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**, 167–188.
48. Hermann, T. & Patel, D. J. (1999). Stitching together RNA tertiary architectures. *J. Mol. Biol.* **294**, 829–849.
49. Draper, D. E. (2004). A guide to ions and RNA structure. *RNA*, **10**, 335–343.
50. Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* **5**, 1029–1032.
51. Chothia, C., Hubbard, T., Brenner, S., Barns, H. & Murzin, A. (1997). Protein folds in the all-beta and all-alpha classes. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 597–627.
52. Finkelstein, A. V. & Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* **50**, 171–190.
53. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, **261**, 552–558.
54. Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G. & Chothia, C. (1999). SCOP: a Structural Classification of Proteins database. *Nucl. Acids Res.* **27**, 254–256.
55. Breaker, R. R. (2002). Engineered allosteric ribozymes as biosensor components. *Curr. Opin. Biotechnol.* **13**, 31–39.

Edited by J. Doudna

(Received 20 April 2004; received in revised form 10 June 2004; accepted 21 June 2004)