



## RAG: RNA-As-Graphs database—concepts, analysis, and features

Hin Hark Gan<sup>1,2</sup>, Daniela Fera<sup>1</sup>, Julie Zorn<sup>1</sup>, Nahum Shiffeldrim<sup>1</sup>, Michael Tang<sup>1</sup>, Uri Laserson<sup>1,3</sup>, Namhee Kim<sup>1</sup> and Tamar Schlick<sup>1,2,3,\*</sup>

<sup>1</sup>Department of Chemistry, New York University, New York, NY 10003, USA, <sup>2</sup>Howard Hughes Medical Institute, 100, Washington Square East and <sup>3</sup>Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012, USA

Received on May 30, 2003; revised on October 27, 2003; accepted on November 11, 2003  
Advance Access publication February 12, 2004

### ABSTRACT

**Motivation:** Understanding RNA's structural diversity is vital for identifying novel RNA structures and pursuing RNA genomics initiatives. By classifying RNA secondary motifs based on correlations between conserved RNA secondary structures and functional properties, we offer an avenue for predicting novel motifs. Although several RNA databases exist, no comprehensive schemes are available for cataloguing the range and diversity of RNA's structural repertoire.

**Results:** Our RNA-As-Graphs (RAG) database describes and ranks all mathematically possible (including existing and candidate) RNA secondary motifs on the basis of graphical enumeration techniques. We represent RNA secondary structures as two-dimensional graphs (networks), specifying the connectivity between RNA secondary structural elements, such as loops, bulges, stems and junctions. 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. The RAG inventory immediately suggests candidates for novel RNA motifs, either naturally occurring or synthetic, and thereby might stimulate the prediction and design of novel RNA motifs.

**Availability:** The database is accessible on the web at <http://monod.biomath.nyu.edu/rna>

**Contact:** [schlick@nyu.edu](mailto:schlick@nyu.edu)

### INTRODUCTION

The broad range of RNA's functional roles in the cell is now being unravelled as never before (Gibbs, 2003). Indeed, with the recent sequencing and analysis of the mouse genome, the abundance of RNA-coding genes and RNA's unanticipated functional roles have been suggested (Waterston *et al.*, 2002; Okazaki *et al.*, 2002; Rivas *et al.*, 2001; Eddy, 2001; Storz,

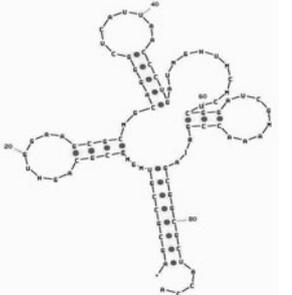
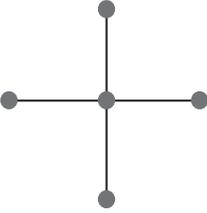
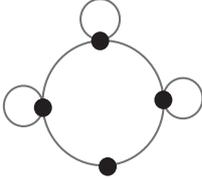
2002). Functional RNA molecules possess well-defined secondary and tertiary structures, which are conserved for each functional class (e.g. tRNA, ribosomal RNA, group I intron). Just as for protein structural genomics projects (Burley and Bonanno, 2002), an understanding of the range of RNA's structural repertoire is crucial in the identification of novel functional RNAs. Here, we present a new RNA motif database RNA-As-Graphs (RAG) that catalogues both existing and other possible RNA secondary motifs.

Present RNA databases include the Nucleic Acids Database (NDB; Berman *et al.*, 2003, <http://ndbserver.rutgers.edu/NDB>) and the RNA Structure Database (RNABase, <http://www.rnabase.org/>), which archive three-dimensional (3D) RNA structures; the Structural Classification of RNA (SCOR; Klosterman *et al.*, 2002) provides hierarchical classification of RNA motifs; Gutell's database catalogues the secondary ribosomal RNA structures (Cannone *et al.*, 2002; <http://www.rna.icmb.utexas.edu>); PseudoBase archives existing pseudoknots (Batenburg *et al.*, 2001, [www.bio.leidenuniv.nl/~Batenburg/PKB.html](http://www.bio.leidenuniv.nl/~Batenburg/PKB.html)); Rfam catalogues conserved RNA families derived from multiple sequence alignments and covariance models (Griffiths-Jones *et al.*, 2003); NCIR, a database of non-canonical interactions in RNAs (Nagaswamy *et al.*, 2002); and other specialized RNA databases (see the RNA World; <http://www.imb-jena.de/RNA.html>).

In contrast to existing databases, which provide information about known RNA primary, secondary and tertiary structures, RAG provides a quantitative method for cataloguing and classifying all RNA structures based on the topological properties of their secondary motifs using graph theory results, such as graphical enumeration (Harary, 1969) and generation, and spectral analysis (Fiedler, 1989; Cvetkovic *et al.*, 1995).

Topological properties refer to the connectivity pattern among RNA's secondary structural elements, such as loops, bulges, stems and junctions. For example, the star-shaped

\*To whom correspondence should be addressed.

Structure ID	Secondary Structure	Tree Graph Representation	Dual Graph Representation
tRNA (NDB: TRNA12)			

**Fig. 1.** Tree graph and dual graph representation for the corresponding tRNA (NDB: TRNA12) secondary structure generated by Mfold.

transfer RNA (tRNA) motif is topologically distinct from the branched structure of 5S ribosomal RNA. Here, ‘motif’ refers to topological properties rather than a sequence or tertiary patterns; our topological level of description is coarser than that of the traditional RNA motifs, such as base pairing, base stacking or kissing hairpins. A coarse description is required for exploring the range and diversity of RNA’s structural repertoire.

Since RNA secondary topologies of different functional classes are remarkably well conserved, their topological characteristics provide a basis for organizing RNA secondary structures broadly (Griffiths-Jones *et al.*, 2003, <http://rfam.wustl.edu/>); however, some degree of topological diversity within the same functional RNA class (e.g. RNase P, tRNA) is expected. RAG has immediate applications for exploring RNA’s repertoire—both existing and potential motifs (Gan *et al.*, 2003).

An RNA graph is a formal construct composed of lines (edges) linking nodes (vertices) representing an RNA secondary structure (Fig. 1). It is characterized by the number of vertices ( $V$ ) and the associated connectivity. Significantly, our discrete tree and dual graphical representations of RNA secondary structures allow all distinct RNA motifs to be enumerated. We use both heuristic and exact graphical enumeration results to build RNA motif libraries in the RAG database. Enumeration of RNA motifs provides unparalleled opportunity for investigating natural and hypothetical RNA motifs, including trees and pseudoknots (Gan *et al.*, 2003). Already, graph enumeration techniques have been exploited to generate chemical libraries for applications in pharmacology (Wieland *et al.*, 1996).

RAG catalogues RNA motifs (graphs) according to the number of vertices ( $V$ ) and ranks the associated  $\lambda_2$ , second smallest nonzero eigenvalue of the graph’s (Laplacian) matrix (Cvetkovic *et al.*, 1995; Mohar, 1991) that characterizes the complexity of the RNA topology. To help stimulate the search for RNAs with novel structures, we annotate all motifs as natural or hypothetical. RAG’s inclusion of both these motif types

is unique. The hypothetical motifs with features resembling natural RNAs may be candidates for novel RNAs. Indeed, we show that existing RNAs represent only a small subset of all possible motifs. RAG also offers a tool for aiding structural and functional characterization of RNA sequences based on their secondary structures, since secondary structures are related to tertiary and functional properties of RNAs.

## METHODS

RNA trees and pseudoknots are two major types of 2D RNA secondary structures, distinguished by the topology of their base pairing patterns. We employ ‘tree graphs’ for archiving RNA tree motifs and ‘dual graphs’ for general RNA motifs, including pseudoknots (Fig. 1). The tree representation is simpler but cannot capture pseudoknots; the latter is more general but less intuitive to create. The advantage of our discrete RNA representations is that they allow enumeration and quantitative characterization of all existing and hypothetical RNA topologies.

### RNA tree graphs

For tree graphs (Le *et al.*, 1989), we use the following rules to assign edges and vertices to RNAs (Gan *et al.*, 2003). (1) A nucleotide bulge, hairpin loop or internal loop is a vertex ( $\bullet$ ) when there is more than one unmatched nucleotide or non-complementary base pair. The special case of the GU wobble base pair is regarded as a complementary base pair. (2) The 3’ and 5’ ends of a helical stem are considered a vertex. (3) An RNA stem is considered an edge ( $—$ ); an RNA stem must have two or more complementary base pairs. (4) An RNA junction is a vertex; a junction is the location where three or more stems meet. Thus, rule 1 defines a bulge/hairpin/internal loop to have two or more unpaired bases; a junction may have any number of unpaired bases. For tree graphs, an edge represents roughly 20 nt.

Naturally, our simplified tree representation of RNAs cannot resolve the type of secondary structure (loop/bulge/junction/3’, 5’ ends), size/length of secondary structures (stems/loops),

chain polarity (direction of strands), angle between secondary structural elements, non-canonical base pairing and sequence level information. Consequently, two distinct functional RNAs may map onto the same graph. This problem can be overcome by using labelled trees and more advanced graph constructs. Although more complex graphical representations have obvious advantages, they make the problem of enumerating distinct RNA topologies intractable. Our schematic RNA models are designed to allow assessment of the range of RNA's structural repertoire.

### RNA dual graphs

Our dual graphs (Gan *et al.*, 2003) can represent all RNA trees as well as pseudoknots and can also be generalized to represent RNA structures with triple, quadruple and higher order helices. We construct dual graphs using the following rules. (1) A vertex ( $\bullet$ ) represents a double-stranded helical stem with  $\geq 2$  complementary base pairs. (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, as in the tree-graph rules. Thus, in contrast to the situation in tree graphs, a vertex now represents a stem instead of a bulge, a loop or a junction; an edge represents a strand in the bulge/loop/junction instead of a stem; see the dual graph of tRNA in Figure 1. As for sequence length  $l$  to  $V$  relation, we find that it is  $l = 20V$  for dual graphs.

### RNA motif libraries from graph enumeration methods

Graphical enumeration can be performed analytically or computationally depending on the complexity of the structures. For unlabelled trees, the number of possible graphs is obtained from the coefficients  $\{c_i\}$  associated with the  $x^i$  term of the counting polynomial derived by Harary and Prins (Harary, 1969):

$$t = \sum_i c_i x^i \\ = x + x^2 + x^3 + 2x^4 + 3x^5 + 6x^6 + 11x^7 + 23x^8 \\ + 47x^9 + 106x^{10} + 235x^{11} + 551x^{12} + \dots$$

For example, there is only one distinct graph each for  $V = 1, 2, 3$  vertices (since  $c_1 = c_2 = c_3 = 1$ ) and two distinct 4-vertex graphs ( $c_4 = 2$ ), three distinct 5-vertex graphs ( $c_5 = 3$ ) and so on. These sets of distinct graphs represent libraries of theoretically possible RNA topologies, which include naturally occurring candidate and hypothetical RNA motifs, with different RNA sequence lengths.

To enumerate and construct graphs, we use probabilistic graph-growing techniques (Gross and Yellen, 1999). For dual graphs with  $V = 2, 3, 4, 5, 6$  and  $7$  we have identified 3, 8, 30, 180, 494 and 2388 distinct, non-isomorphic graphs,

**Table 1.** Number of existing and hypothetical RNA tree and dual graph motifs

$V$	Tree motifs	Existing tree motifs	Dual graph motifs	Existing dual graph motifs
2	1	1	3	3
3	1	1	8	4
4	2	1	30	12
5	3	3	180*	6
6	6	2	494*	2
7	11	1	2388*	1
8	23		NA	
9	47		NA	
10	106		NA	

\*Estimated by a probabilistic graph-generation method.

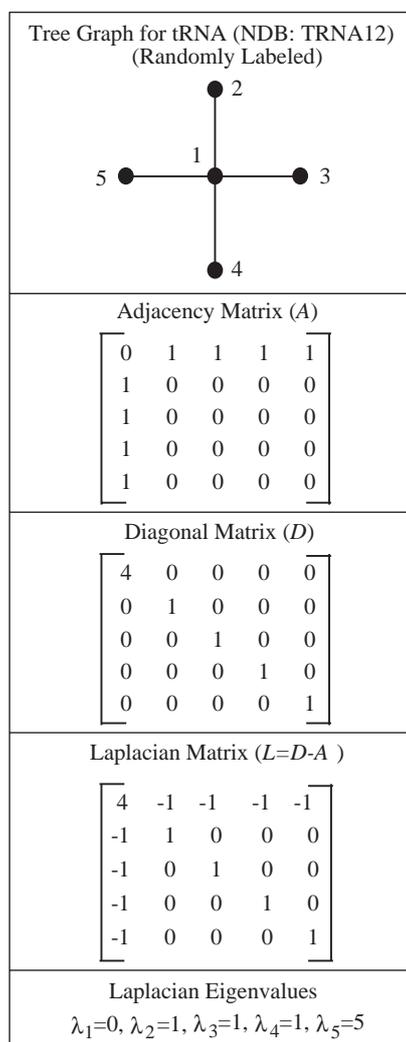
respectively (Table 1). For larger  $V$ , the number of dual graphs increases rapidly, and only subsets can be generated by exhaustive computation.

### Spectral analysis of RNA graphs: Laplacian matrix and eigenvalues

RNA tree and dual graphs can be quantitatively characterized using spectral techniques in graph theory (Cvetkovic *et al.*, 1995). A graph's matrix specifies the degree of connectivity between the vertices of the graph (Fig. 2), and the matrix's eigenvalue spectrum is a tool to quantify connectivity as well as characterize graph similarity.

Specifically, we use the Laplacian ( $V$  by  $V$ ) matrix ( $L$ ) representation for a graph of  $V$  vertices, useful for modelling physical systems such as vibrating strings (with beads) and membranes (Van Dam and Haemers, 2002).  $L$  is constructed from the matrices  $D$  and  $A$  that define the graph:  $L(G)$  of graph  $G$  with vertices  $1, 2, \dots, V$  is defined as  $L = D - A$ , where  $A$  and  $D$  are the adjacency and degree matrices of the graph, respectively. The elements ( $a_{ij}$ ) of the  $V \times V$  symmetric matrix  $A$  specify the number of links or edges connecting  $i$  and  $j$  vertices (a self-loop at vertex  $i$  contributes 2 to the diagonal element  $a_{ii}$ ).  $D$  is a square diagonal matrix whose elements ( $d_{ii}$ ) specify the valency or the degree of connectivity of vertex  $i$ ; e.g. a vertex with four (incident) edges emanating from it has a degree of 4. Figure 2 shows the 5-vertex tree graph for tRNA with its corresponding adjacency ( $A$ ), diagonal ( $D$ ) and Laplacian ( $L$ ) matrices.

A  $V$ -vertex graph is thus characterized by the ordered eigenvalues  $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_V$  of  $L(G)$ , called the spectrum of  $G$ , which is independent of the labelling of graph vertices and a measure of connectivity. The spectrum of the tRNA example (Fig. 2) is  $\lambda_1 = 0$ ,  $\lambda_2 = \lambda_3 = \lambda_4 = 1$  and  $\lambda_5 = 5$ , indicating a 4-fold symmetry and a 4-stem junction. In general, all connected (RNA) graphs have the property that  $\lambda_1 = 0$  and  $\lambda_2 > 0$ , and graphs with similar  $\lambda_2$  values have similar topologies. Thus, the second



**Fig. 2.** A 5-vertex (tRNA) tree graph and its matrix representations (the vertices are randomly labelled). The adjacency matrix ( $A$ ) is a connectivity matrix whose entries specify the number of edges connecting pairs of graph vertices ( $i, j$ ); e.g. the pairs connected by an edge have a matrix element of 1 and the non-connected pairs have a matrix element of 0. The elements of the diagonal matrix ( $D$ ) specify the degree of connectivity of each vertex. Also shown is the eigenvalue spectrum of the Laplacian matrix ( $L$ ), defined as  $L = D - A$ .

eigenvalue  $\lambda_2$  reflects the overall pattern of connectivity of a graph (Fiedler, 1989); a linear chain has a smaller second eigenvalue than a branched structure (since some of the latter graph's vertices have degree  $d_{ii} > 2$ ). However, the Laplacian spectrum does not uniquely determine the topology of a graph, because of the existence of non-isomorphic co-spectral graphs (Van Dam and Haemers, 2002, <http://netec.mcc.ac.uk/WoPEc/data/Papers/dgrkubcen200266.html>). This problem is not severe for the tree and dual graphs of interest to us (e.g. 3% co-spectral graphs for dual graphs with  $V = 4$ ).

## RNA-As-Graphs FEATURES

### RNA motif libraries and quantitative characterization of motifs

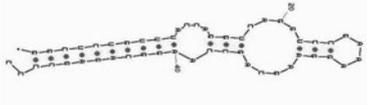
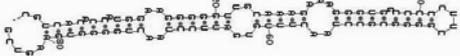
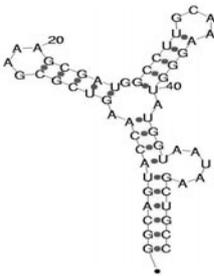
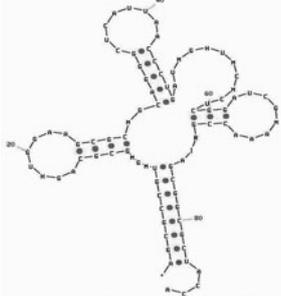
The RNA motif libraries for RAG are derived from the counting polynomial of Harary–Prins for (unlabelled) tree topologies and from computational enumeration techniques for dual graph topologies. For a given vertex number  $V$ , a library of possible RNA motifs is generated, with size depending on  $V$  and the motif type (tree or pseudoknot).

We rank the tree and dual graphs in each motif library ( $V = 2, 3, 4, \dots$ ) using  $\lambda_2$  to allow automatic cataloguing of RNA graphs (and search of existing and hypothetical RNA motifs) but provide the complete eigenvalue spectrum ( $\lambda_1, \lambda_2, \dots, \lambda_V$ ) of each topology. Thus, all existing or hypothetical motifs are referenced as  $(V, \lambda_2)$ . For easy reference, we also index graphs as  $(V, n)$  where  $n$  are integer numbers ordered according to the  $\lambda_2$  ranking; distinct graphs with the same eigenvalue are assigned different  $n$  values. This scheme allows RNA motifs of varying degrees of topological complexity to be distinguished. Since  $\lambda_2$  depends on  $V$ , it is only meaningful to compare the  $\lambda_2$  values of motifs in the same library  $V$ . Similarly, eigenvalues of the tree and dual graphs are not directly comparable. Since tree graphs can also be represented as dual graphs, we link between the two libraries where relevant. Each library catalogues both hypothetical and existing RNA motifs, with links to other RNA sequences, structures (2D and 3D) and function databases (e.g. NDB, PseudoBase, 5S). Figure 3 shows  $V = 4$  and 5 tree motif libraries with corresponding RNA secondary folds predicted by Mfold (Zuker *et al.*, 1999).

RAG's systematic ordering of all RNA motifs according to  $\lambda_2$  constitutes an RNA motif dictionary in which the neighbouring motifs do not necessarily have a structural or functional relationship, although the sequence lengths of the various motifs are similar. Such a scheme allows easy access to and retrieval of RNA motifs.

### Libraries of RNA tree topologies from tree graphs

Table 1 summarizes the size of the RNA tree motif libraries for 2, 3, ..., 10-vertex graphs and the number of distinct natural and 'missing' RNA topologies in each motif library. The natural RNA tree motifs found include U2 snRNA ( $V = 2, \lambda_2 = 2$ ), single-strand RNAs ( $V = 3, \lambda_2 = 1$  and  $V = 4, \lambda_2 = 0.5858$ ), tRNA ( $V = 5, \lambda_2 = 1$ ), 70S ( $V = 5, \lambda_2 = 0.3820$ ), 5S rRNA ( $V = 7, \lambda_2 = 0.2254$ ) and signal recognition particle RNA ( $V = 9, \lambda_2 = 0.2311$ ). While some graphs correspond to natural RNA topologies, many other tree graphs have not yet been found in nature. For example, the numbers for missing tree motifs for  $V = 4, 5, 6$  and 7 are 1, 0, 4 and 10, respectively (Table 1). Though the majority of the missing motifs may not correspond to natural RNAs or potential RNAs with novel properties, some

$V$	$\lambda_2$	Tree Graph	Secondary Structure	nt.
4	0.5858		RNA single strand (PR0021) 	51
	1.0000		missing motif	
5	0.3820		70S (F) (RR0003) 	85
	0.5188		P5abc 	56
	1.0000		tRNA (TRNA12) 	87

**Fig. 3.** The motifs of 4- and 5-vertex RNA tree libraries and their corresponding second Laplacian eigenvalues ( $\lambda_2$ ); each tree motif is accompanied by an existing (or missing) RNA secondary structure (generated by Mfold) conforming to that motif.

motifs may be identified in natural systems in the future, while others may be designed in the laboratory.

### Libraries of RNA tree and pseudoknot topologies from dual graphs

Complete libraries (Table 1) of small RNA dual graphs ( $V < 7$ ) can be readily enumerated using computational approaches such as exhaustive search and probabilistic graph-growing methods (Gross and Yellen, 1999). Each library of

dual graph motifs contains RNA tree, pseudoknot and bridge types; an RNA bridge is a motif that becomes disconnected upon removal of an edge. Naturally, dual graph libraries are larger than tree libraries.

We identified a total of 22 distinct pseudoknot topologies in the literature and the PseudoBase database (Batenburg *et al.*, 2001), as follows: 9 for  $V = 2, 3, 4$ ; 12 for  $V = 5-18$ , and 22; and one for 16S rRNA pseudoknot ( $V = 87$ ). We have also identified six examples of naturally occurring RNA

bridge motifs with 4–22 vertices. These RNAs include the hepatitis C virus ( $V = 18, \lambda_2 = 0.1317$ ), group I intron ( $V = 22, \lambda_2 = 0.0848$ ; also has a pseudoknot submotif) and box H/ACA snoRNA ( $V = 4, \lambda_2 = 0.7639$ ).

### RNA structural and functional identification: RNA Matrix Program

We use a graph comparison procedure (RNA Matrix Program) to aid the search for structurally/functionally isomorphic RNAs. Our scheme is advantageous over sequence comparison because RNAs in the same functional class (e.g. tRNA, 5S rRNA, group I intron) have similar or conserved secondary and tertiary structures, whereas the conservation of RNA sequences is less apparent. Basically, we search for an annotated RNA motif in the database that has the same vertex number  $V$  and eigenvalue spectrum as the secondary structure submitted by the user. A positive match can lead to a functional interpretation of the queried structure.

Specifically, the RNA Matrix Program accepts a base pairing file (or file known as 'ct' generated by Zuker's folding algorithm) for a secondary structure submitted by the user, converts the file into a graphical RNA representation using the tree graph rules, and then computes the Laplacian eigenvalue spectrum of the corresponding RNA graph. The output of the program is then used to search for motifs in the database with the same topological characteristics.

Our approach of combining the RNA motif database with a structural identification program has several limitations. First, not many distinct RNA topologies are currently known, but the situation will probably change in the near future as more functional RNAs are identified (Storz, 2002; Eddy, 2001; Rivas *et al.*, 2001). Second, our schematic graphical representations suppress many features such as conservation of sequence segments and biologically active sites. Still, RAG's approach can serve as a preliminary search tool to assist refined structural/functional annotation methods, such as multiple sequence alignment and comparative analysis (Cannone *et al.*, 2002). Some of the current limitations may also be removed by extensions of graph theory (e.g. labelled and weighted graphs).

### Summary and conclusion

Our RAG database for RNA secondary motifs, including trees and pseudoknots, exploits the conservation of RNA secondary structures and their intimate connection to tertiary structure and function. By representing RNAs as graphs, we introduce a systematic and quantitative approach to the organization of RNA motifs. RAG describes RNA secondary motifs according to the number of vertices  $V$  (equivalently, sequence length) and the topological complexity (e.g. second smallest Laplacian eigenvalue of the RNA graph). It contains RNA tree libraries, for tree graphs, and the more general (but less intuitive) dual graph libraries, for tree and pseudoknot motifs. RAG can thus catalogue all existing and hypothetical

RNA motifs. It immediately suggests missing RNAs. Among the missing motifs, we can pinpoint RNA-like topological properties (e.g. those found in the vicinity of existing RNA motifs).

Many avenues can be pursued to improve RAG. Labelled and directed graphs can allow more information about RNA secondary motifs by differentiating specific loops/bulges/junctions and specifying strand directions. A substructure search utility tool for analysing RNA secondary structures (Gan *et al.*, 2003), including large ribosomal RNAs, can help identify structural/functional relationships; we plan to incorporate information about existing large RNAs, such as large ribosomal RNAs. We also plan to expand the number of dual graphs available by exploiting database technologies for storage and retrieval, work on linking available 3D RNA structures (e.g. NDB, RNABase) to RAG topologies, and classify existing RNA topologies into functional categories to complement our mathematical cataloguing scheme.

A great challenge emerging from our work on RAG is to define a 'mapping' of sequences onto the most promising new motif candidates. New RNA motifs can be exploited for biotechnological applications. We are currently exploring our idea of using a modular assembly of existing RNA fragments (Gan *et al.*, 2003) to form a sequence/structure motif. Another intriguing application of RAG under investigation is to use the catalogued missing motifs to direct the search for novel RNA motifs/functions in genomes (Zorn *et al.*, 2004; Laserson *et al.*, 2004). A coupling of theoretical to experimental efforts is necessary in this quest. We invite users to explore RAG and send us their comments to RAG@biomath.nyu.edu.

### ACKNOWLEDGEMENTS

We thank the referees for their constructive suggestions for improving RAG and Danny Banash for comments related to the use of the Laplacian second Eigen value. This work was supported by a Joint NSF/NIGMS Initiative in Mathematical Biology (DMS-0201160) as well as Human Frontier Science Program (HFSP). D.F. acknowledges support from the Dean's Undergraduate Research Fund and a summer fellowship from the Department of Chemistry.

### REFERENCES

- Batenburg, F.H.D., van Gultyaev, A.P. and Pleij, C.W.A. (2001) PseudoBase: structural information on RNA pseudoknots. *Nucleic Acids Res.*, **29**, 194–195.
- Berman, H.M., Westbrook, J., Feng, Z., Iype, L., Schneider, B. and Zardocki, C. (2003) The Nucleic Acid Database: a repository of three-dimensional structural information about nucleic acids. *Structural Bioinformatics*, John Wiley and Sons, NY, pp. 199–216.
- Burley, S.K. and Bonanno, J.B. (2002) Structuring the universe of proteins. *Annu. Rev. Genomics Hum. Genet.*, **3**, 243–262.
- Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., Feng, B., Lin, N., Madabusi, L.V.,

- Muller, K.M. *et al.* (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.
- Cvetkovic, D.M., Doob, M. and Sachs, H. (1995) *Spectra of Graphs*. Johann Ambrosius Barth Verlag, Heidelberg.
- Eddy, S.R. (2001) Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, **2**, 919–929.
- Fiedler, M. (1989) *Laplacian of Graphs and Algebraic Connectivity, Combinatorics and Graph Theory*. Banach Center Publications, Vol. 25, Warsaw, pp. 57–70.
- Gan, H.H., Pasquali, S. and Schlick, T. (2003) Exploring the repertoire of RNA secondary motifs using graph theory; with implications for RNA design. *Nucleic Acids Res.*, **31**, 2926–2943.
- Gibbs, W.W. (2003) The Unseen Genome: gems among the junk. *Sci. Am.*, 46–53.
- Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
- Gross, J. and Yellen, J. (1999) *Graph Theory and Its Applications*. CRC Press, Boca Raton.
- Harary, F. (1969) *Graph Theory*. Addison-Wesley, Reading, Mass.
- Klosterman, P.S., Tamura, M., Holbrook, S.R. and Brenner, S.E. (2002) SCOR: a structural classification of RNA database. *Nucleic Acids Res.*, **30**, 392–394.
- Laserson, U., Gan, H.H. and Schlick, T. (2004) Exploring the connection between synthetic and natural RNAs in genomes via a novel computational approach. In S. Yancopoulos (ed.), *Pattern Discovery in Biosequences Data: Tools, Techniques and Applications*: Wang, J.T.L., Shapiro, B.A. and Shasha, D. (series eds.), Oxford University Press, NY.
- Le, S.Y., Nussinov, R. and Maizel, J.V. (1989) Tree graphs of RNA secondary structures and their comparisons. *Comput. Biomed. Res.*, **22**, 461–473.
- Mohar, B. (1991) The laplacian spectrum of graphs. In Alavi, Y., Chartrand, G., Ollermann, O. and Schwenk, A. (eds), *Graph Theory, Combinatorics, and Applications*. John Wiley and Sons, Inc., New York, pp. 871–898.
- Nagaswamy, U., Larios-Sanz, M., Hury, J., Collins, S., Zhang, Z., Zhao, Q. and Fox, G.E. (2002) NCIR: a database of non-canonical interactions in known RNA structures. *Nucleic Acids Res.*, **30**, 395–397.
- Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
- Rivas, E., Klein, R.J., Jones, T.A. and Eddy, S.R. (2001) Computational identification of noncoding RNAs in *E.coli* by comparative genomics. *Curr. Biol.*, **11**, 1369–1373.
- Storz, G. (2002) An expanding universe of noncoding RNAs. *Science*, **296**, 1260–1263.
- Van Dam, E.R. and Haemers, W.H. (2002) Which graphs are determined by their spectrum? *Discussion Paper*, No. 2002–66.
- Waterston, R.H., Lindblad-Toh, K., Birneg, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexanderson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
- Wieland, T., Kerber, A. and Laue, R. (1996) Principles of the generation of constitutional and configurational isomers. *J. Chem. Inf. Comput. Sci.*, **36**, 413–419.
- Zorn, J., Shiffeldrim, N., Gan, H.H. and Schlick, T. (2004) Structural motifs in ribosomal RNAs: implications for RNA design and genomics. *Biopolymers*, **70**, 343–347.
- Zuker, M., Mathews, D.H. and Turner, D.H. (1999) Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In Barciszewski, J. and Clark, B.F.C. (eds) *RNA Biochemistry and Biotechnology*, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, pp. 11–43.