Network Theory Tools for RNA Modeling

Namhee Kim,
New York University Department of Chemistry Courant Institute of Mathematical Sciences 251 Mercer Street New York, NY 10012, USA

Louis Petingi, and
College of Staten Island City University of New York Department of Computer Science 2800 Victory Boulevard Staten Island, NY 10314, USA louis.petingi@csi.cuny.edu

Tamar Schlick
New York University Department of Chemistry Courant Institute of Mathematical Sciences 251 Mercer Street New York, NY 10012, USA schlick@nyu.edu

Abstract

An introduction into the usage of graph or network theory tools for the study of RNA molecules is presented. By using vertices and edges to define RNA secondary structures as tree and dual graphs, we can enumerate, predict, and design RNA topologies. Graph connectivity and associated Laplacian eigenvalues relate to biological properties of RNA and help understand RNA motifs as well as build, by computational design, various RNA target structures. Importantly, graph theoretical representations of RNAs reduce drastically the conformational space size and therefore simplify modeling and prediction tasks. Ongoing challenges remain regarding general RNA design, representation of RNA pseudoknots, and tertiary structure prediction. Thus, developments in network theory may help advance RNA biology.

Keywords

Network Theory; RNA-As-Graphs; RNA Prediction; *In Vitro* Selection

> "To know what to leave out and what to put in; just where and just how, ah, *that* is to have been educated in knowledge of simplicity."
>
> – Frank Lloyd Wright.

1 Introduction

Modeling biological systems involves not only choices of what to approximate and how, as well as what can be neglected, but also selection of appropriate tools, both existing and new, to design and apply to complex problems. Many successes in science can be credited to borrowing tools from seemingly disparate fields and applying them in new ways. Examples include quantum computing using DNA [1], analyzing DNA recombination products using

Corresponding author: nk401@nyu.edu.

knot theory [91], storing digital information within DNA [15], modeling economic scenarios using game theory [70], and applications of Mandelbrot's fractal geometry to architecture, financial markets and turbulence [63, 64, 65].

Network or graph theory is a well-established field of mathematics that has been used extensively in a variety of economic, social, biological, and medical contexts. Essentially, foundations of graph the ory can be used to enumerate and analyze combinatorial properties of various systems, such as communication, chemical, and biological networks. Specific examples include chemical structures (e.g., hydrocarbons, drug compounds) [38, 62], genetic and cellular relationship [11, 40, 67, 30, 58], transportation arrays [5], Internet linkages [88], and social media communications [35]. In mathematical terms, an undirected graph $G = (V, E)$ is a discrete object described by a finite set of vertices V and a set E of unordered pair of vertices called edges, where each edge represents a connection between two vertices. If E is composed instead of ordered pairs of vertices (each representing direction of an edge), G is called a directed graph (digraph). In this work, we describe how graphs can be used to model and study RNA structure. See also our recent review [47].

2 RNA Background

RNA has come to the forefront of science with recent discoveries of its regulatory roles [19, 85]. Beyond protein synthesis, RNA can regulate gene expression and modify the genetic message by gene silencing [33], chemical modifications of ribosomal RNAs [2], control of protein stability [90], and changing conformations of ligand-binding sites of messenger RNA [71, 78, 79, 12].

These new discoveries have also opened new avenues for thinking about therapeutic biotechnology applications of RNA, because RNA's editing, silencing, and other regulatory capabilities can be manipulated to shut off and turn on genes, deliver drugs, diagnose gene activity, and design new materials [80, 31, 13].

Information on RNA structure comes from experimental information (X-ray crystallography, NMR, chemical probing) as well as modeling (e.g., as reviewed in [52]). These structures and their analyses have established RNA's hierarchical structure, in which building blocks (motifs) combine stepwise to form complex active shapes of RNA [52, 47]: The primary (1D) structure or sequence leads to the seconday (2D) structure – pattern of hydrogen bonding arrangements – which in turn triggers tertiary (3D) structure formation, created by all interactions including long-range contacts between 2D substructures (see, Figure 1). RNA is a single-stranded polymer whose sugar-phosphate backbone contains four standard bases, Adenine (A) and Guanine (G), Uracil (U) and Cytosine (C), and their modified bases in various order. This single-stranded polymer folds upon itself, to form GC, AU, or GU ("wobble") base pairs which define double-helical regions ("stems"), imperfect with single-stranded regions named "hairpin loops", "internal loops", and "junctions", which have one, two, or more adjacent helical arms, respectively (see 2D structure elements in Figure 2a and Figure 3). When two single-stranded regions flanked by a stem are base-paired, an interwined RNA structure called *pseudoknot* forms (see Figure 1; top middle, and

top right, for the interwining green and red stems). Long-range interactions between secondary structural motifs form complex tertiary network-like structures.

These hierarchical, modular, and network-like features of RNA structures invite mathematical and computational approaches to model, analyze, and predict RNA structures (see recent reviews [47, 52]). For example, programs for predicting RNA 2D structures by free energy minimization [95, 82], folding 3D structures of small RNAs [72, 18], and annotating motifs in 2D/3D structures [66, 76] are widely used. Figure 1 shows the abstraction of 2D structures as a tree graph and a knot-like graph (called "dual graph").

3 Graph Theory Approaches to RNA Structure Modeling

3.1 Early Graph Approaches by Waterman, Nussinov and Shapiro

Pioneering modeling of graphs for RNA began in the late 1970s. Waterman developed graphical representations of RNA in 1978 with the aim of analyzing the secondary structure of tRNA (Figure 2b) [89]. Specifically, he offered the first graph-theoretic definition of secondary structure and classified graphs of RNA secondary structures with the goal of finding stable 2D structures. These planar graphs were analyzed for base pairing using an adjacency matrix. His method allows comparison of two different RNA 2D structures by approximating the free energy based on the adjacency matrix representations. In 1989, Le et al. also developed the ordered labeled tree representation to compare 2D structures of RNA (Figure 2c) [57].

In 1990, Shapiro and coworkers used a tree representation of RNA 2D structures to measure structural similarities (Figure 2d) [83]. They developed an algorithm for analyzing multiple RNA 2D structures by multiple string alignment. In particular, they defined the tree edit distance between two (full) tree 2D structures to quantify the minimum cost (insertion, deletion, and replacement of nodes) along an edit path for converting one tree into another. Morosetti further studied similarities in tree graph representations by using topology connectivity indices known as the Randi index [6]. That is, for a given graph $G = (V, E)$, and $d(u)$ representing the number of edges incident at vertex u of G (i.e., the degree of vertex u), the Randi index, $R(G)$, is defined in term of the degrees of the vertices as $R(G) = \sum_{(u,v) \in E} (d(u) \cdot d(v))^{1/2}$. This index was originally introduced to assess the connectivity in chemical graphs as this measure is sensitive to the shape of a chemical cluster.

3.2 RAG Tree and Dual Graphs

In 2003, Schlick and coworkers introduced tree and dual graphical representations of RNA 2D motifs in a framework called RAG (RNA-As-Graphs) [22, 23, 20, 37] (see Figures 2e, 3 and 4).

In the RAG tree graphs, RNA 2D structural elements – stems, loops, bulges, and junctions – are converted into 2D graphical objects with the following rules: (1) an edge (−) represents a double-stranded helical stem with more than one base pair; (2) a vertex (●) represents a single strand that occurs in segments connecting secondary structural elements such as bulges, loops, and junctions. Here, a bulge motif is considered to be an internal symmetric or asymmetric loop with more than one unmatched nucleotide or one unstable base pair.

Although RAG tree graphs offer an intuitive description of RNA 2D topologies, they cannot represent pseudoknots. The dual graph representation translates RNA 2D structure into a more abstract graphical model by interchanging the vertices and edges for opposite motifs from the translation rules of RAG tree graphs [22, 20, 23, 37]: (1) a vertex (●) represents a double-stranded helical stem with more than one base pair; (2) an edge (−) represents a single strand that occurs in segments connecting 2D structural elements such as bulges, loops, and junctions where a bulge motif is the same as the translation rules of RAG tree graphs (i.e., asymmetric loop with more than one unmatched nucleotide or one unstable base pair). See Figure 3 for example.

The above definitions can be modified as needed to construct more detailed 2D network models [22, 32, 41]. For example, by adding the flow direction on edges (i.e., labeled vertices), RNA structures with different starting and ending vertices can be differentiated [22, 32, 41]. Currently, in the goal of RNA tertiary structure prediction, Schlick and coworkers are developing 3D tree graph representations for parallel and anti-parallel helical arrangements with the following additions and modification of vertices and edges to RAG tree graphs: (1) additional vertices (●) are set at helix ends; (2) additional edges (−) are set to connect RAG tree vertices representing loops to additional proximal vertices; (3) additional vertices (●) are set to represent bulges for asymmetric loop with one unmatched nucleotide (see Figures 3 and 4). When the geometrical features such as helix lengths and angles are added to 3D tree graphs, we can also represent RNA 3D global geometries intuitively. See end of Section 4.4 for an overview.

3.3 Advantages of the RAG Representation

The graphical representation of RNA secondary structures makes it possible to apply graph theory methods to quantitatively describe topological properties (i.e., topological descriptors) of RNA motifs [20, 43].

Let $G = (V, E)$ be an undirected graph and $n = |V|$; we define the associated Laplace matrix (called the Laplacian) $M = (m_{ij})$ as the $n \times n$ matrix where

$$m_{ij} = \begin{cases} d(i) & : & i=j \\ -w_{ij} & : & (i,j) \in E \\ 0 & : & (i,j) \notin E, \end{cases} \quad (1)$$

and the value w_{ij} is the number of edges between vertices i and j. The Laplacian thus decribes the connectivity of a graph (see Figures 2e and 3 and Ref. [20, 43]). The Laplacian eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are non-negative and real because M is symmetric, and the smallest eigenvalue $\lambda_1 = 0$ [69]; the Laplacian eigenvalues have been extensively studied because they describe combinatorial properties of a topology. As an example, a spanning tree of a graph G is a minimally connected subgraph (i.e., connected and acyclic) containing all the vertices of G. Each RAG tree graph has only one spanning tree, itself, since any removal of an edge makes a graph disconnected. For dual graphs, however, this is not the case; dual graphs can have several spanning trees. An arbitrary spanning tree can be found by an application of Depth First Search [29] under the assumption that G is connected. The number of spanning trees of a graph G on n vertices, $t(G)$, can be evaluated as a function of

the Laplacian eigenvalues as $t(G) = \left(\prod_{i=2}^{n} \lambda_i\right)/n$ [7]. An open problem in extremal graph theory, originally proposed in the 1970s [42], is to characterize graphs with maximum number of spanning trees (called t-optimal graphs) among all graphs with e edges and n vertices (see [42, 14, 74, 75]); t-optimal graphs play an important role in the design of reliable communication networks [8, 9] in reliability theory as well as in optimization design theory in statistics [14].

The second smallest eigenvalue, λ_2, and its corresponding eigenvector, μ_2, provide information about topological compactness (i.e., algebraic connectivity) and partition properties of graphs, respectively. The eigenvalue λ_2, also known as the *Fiedler* eigenvalue [21], increases with the compactness of the structure. For example, λ_2 is smaller in a linear chain than in a branched structure for a family of closely related graphs (see examples in Figure 4). Unlike traditional topological connectivity, algebraic connectivity or compactness depends on the number of vertices and how they are connected.

The eigenvector $\mu_2 = (v_1, v_2, \ldots, v_n)$ provides information on how to subdivide a large RNA into smaller fragments by using spectral-graph partitioning methods [77] that minimize dependencies between fragments. A cut $C(A, A')$ of a graph $G = (V, E)$ is a set of edges that, when removed from G, partitions the set of vertices V into two sets A and A'. The cut-ratio $\Omega(A, A')$ of the cut $C(A, A')$ is defined as $|C(A, A')|/min(|A|, |A'|)$. The idea of spectral partitioning is to bisect (i.e., split a set into two sets whose cardinalities differ by at most in one) the vertex set V into $\{i \in V : v_i > s\}$ and $\{i \in V : v_i \quad s\}$ for given value s. Spielman et al. showed that there exists a value s that can be determined in term of the *Fiedler* eigenvalue, for which the bisection of the *Fiedler* eigenvector yields the best possible cut-ratio in G [86]. For example, μ_2 corresponding to RAG tree graph 5_2 in Figure 5 is (−0.20, 0.34, −0.42, −0.42, 0.70). When s is the median of the eigenvector components, −0.20, the graph is partitioned into two sets of vertices, $\{5, 2\}$ and $\{1, 3, 4\}$, representing an internal loop and a three-way junction, respectively.

Another important graph-theoretical descriptor is the *diameter* of a graph $G = (V, E)$ which is defined as the maximum distance (length of the shortest path) between every pair of vertices $u, v \in V$. Intuitively, the diameter represents the maximum possible communication delay between a pair of vertices of a graph. Even though the *Fiedler* eigenvalue and the diameter assess different structural properties of graphs, the algebraic connectivity is related to the reciprocal of the average distance of a graph. Also $4/(n \cdot \lambda_2)$ represents a lower bound for the diameter of a graph G on n vertices [68].

Laplacian eigenvalues [20, 43], graph diameter [25], and spanning trees are examples of topological descriptors widely used to determine properties of graphs. These quantitative measures can be used to compare graphs, subdivide graphs into recurring motifs, and identify subgraphs [43, 73].

In communication applications, graphs with maximum number of spanning trees are useful for reliable network design [74, 75]. Consider a communication network $G = (V, E)$, in which edges fail independently (vertices are perfectly reliable) and each edge x fails with probability ρ_x. The *all-terminal reliability*, $R_V(G)$, gives the probability that each pair of

vertices u, $v \in V$ will remain connected via an operational path after deletion of the failed edges. It is not difficult to show that if a graph G_0 has maximum number of spanning trees (i.e., G_0 is t-optimal) among all graphs in a class with equal number of edges $e = |E|$ and vertices $n = |V|$, Ωn, e), then when edges are failing (i.e., for each edge x the probability of failure ρ_x approaches 1), G_0 is also the most reliable graph in $\Omega(n, e)$. Thus t-optimal graphs play a significant role when characterizing most reliable graphs given a class of graphs $\Omega(n, e)$.

Independently from this result in graph theory, Cheng [14] discovered in 1982 that the optimization problem of characterizing t-optimal graphs in a class of graphs $\Omega(n, e)$ is equivalent to the optimization problem of characterizing the so-called balanced incomplete-block designs in statistics (i.e., design of network with specific requirements such as the number of objects in each block). We first consider block designs with t treatments to be tested on b blocks of size k with $k < t$. Each statistical experiment can be mapped to a graph in $\Omega(n, e)$, where $n = t + b$ and $e = b * k$, and the optimal block design is the experiment represented by a t-optimal graph in $\omega(n, e)$.

A key advantage of this topological approach is that it reduces the size of RNA space enormously. The RAG framework make it possible to enumerate and generate RNA 2D topologies, opening a new avenue for predicting novel RNA motifs [22, 43]. For RAG unlabeled tree graphs, the number of possible graphs is obtained using the counting polynomial derived by Harary and Prins [22, 34]. To enumerate and construct RAG dual graphs, probabilistic graph-growing techniques are used [43]. These sets of distinct graphs represent libraries of theoretically possible RNA topologies, which include existing and hypothetical ("missing") RNA motifs.

Since not all hypothetical RNA topologies are physically meaningful, Kim et al. predicted motifs, named as "RNA-like" motifs, which are more likely to occur in nature than others by a learning approach based on known RNAs [43]. They clustered RNA graphs into two groups based on topological variables via the transformation of Laplacian eigenvalues and predicted novel candidate RNA-like topologies that possess topological properties similar to existing RNAs [43]. Based on these predictions, RAG catalogued all RNA graphs into existing and RNA-like, and non-RNA-like motifs (http://www.biomath.nyu.edu/rna): see Figures 5–7 for RAG segments. Among the ten RNA-like dual graphs with 3 or 4 vertices predicted in 2004, five motifs (see C1, C2, C3, C4, and C7 in Figure 6b) have been identified in nature with regulatory or catalytic functions [37].

RAG's systematic catalogue offers a comprehensive search ability and has been used in analyzing, designing, and predicting RNA structures [73, 56, 49]. For example, Pasquiali et al. analyzed modular pseudoknot architectures by dual subgraph-isomorphism searches in large ribosomal RNA [73]. Laserson et al. identified motifs corresponding to antibiotics-binding motifs in genomes by searching tree graphs [56]. Knisley and coworkers applied RAG to predict larger RNA-like structures by merging two small graphs and applying neural network analysis [49]. See also [47] for the impact of RAG on the RNA field.

4 Graph Applications to RNA Structure Prediction

4.1 Motif Analysis

Network analysis applied to RNA structures has identified modular RNA structure which are composed of repetitive motifs, in which patterns appear hierarchically, from 2D to 3D structure levels. Essentially, motifs are recurring structural elements in RNA molecules.

Statistical analysis of 2D networks helped reveal motif distributions of known RNAs and provided improved structural and energetic parameters for 2D structure prediction [94, 24]. Zorn et al. annotated paired and unpaired bases in stems, junctions, hairpin loops, bulges, and internal loops in ribosomal RNAs and discovered that the paired and unpaired bases in structural motifs have characteristic distribution shapes and ranges: the frequency distribution of paired bases in stems declines linearly with the number of bases, whereas, for unpaired bases in junctions, it has a pronounced peak [94]. Gardner et al. correlated structural statistics of hairpins and internal loops to energetic parameters by analyzing sequence alignments and 2D structure data sets to improve the accuracy of RNA 2D prediction [24].

The RNA 3D network plays an important role in RNA 3D folding and provides models of the principles of organization of complex RNA structures [59, 87, 93]. Leontis and Westhof developed the network representation rules to represent 3D structure in 2D (including base pairing, base-stacking, and basebackbone interactions), which facilitates the discrete models and classification of 3D motifs [59]. Using this 3D network representation, they classified base pair interactions into 12 geometric families in terms of base pairing and base-stacking [59], and catalogued by similar base pair interactions that can be substituted by compensatory mutation [87]. Zirbel et al. analyzed and classified 10 base-backbone geometrical families based on their hydrogen bond interaction patterns [93].

The Westhof/Leontis 3D network representation and classifications have been used to define recurrent 3D motifs in known RNA 3D structures [92, 28, 10]. By manual inspection by experts in the field of RNA structure, several 3D motifs have been identified (e.g., coaxial helix, A-minor, ribose zipper, and kissing hairpin, see Ref. [92] for the definition of motifs) and more continue to be discovered. Recently, the Jaeger and Steinberg groups found new 3D motifs found common right-angles, twist-joints and double twist-joints motifs in ribosomal RNAs and other large RNAs, which establish specific helical arrangements [28, 10].

The increased repertoire of RNA motifs has stimulated the development and application of statistical and computational analysis in RNA structures to better understand the network interaction properties of RNA [92, 36]. Xin et al. comprehensively analyzed a data set of 54 high-resolution RNA crystal structures for motif occurrence and correlations. Specifically, they searched seven RNA 3D motifs by various computer programs for motif occurrence and correlations (see the resulting frequencies of motifs at http://www.biomath.nyu.edu/motifs). Pyle and coworkers represented RNA backbone structures using a simplified virtual bond system and reported the pseudo-torsional angles focusing on local RNA backbone

geometry [36]. Statistics of RNA 3D motifs provide the groundwork for RNA 3D structure prediction via networks of restraints.

4.2 Junctions Classification and Prediction

RNA junctions are important structural elements that form when two or more helices come together in space in RNA 3D structures associated with 2D structures. The Al-Hashimi, Westhof and Schlick groups performed geometrical analysis and classification of 2-way (or internal loops), 3-way, and 4-way (and higher-order) junctions, respectively [3, 4, 60, 55, 51, 50].

Al-Hashimi and coworkers showed that 2D structure features (e.g., loop size) encode topological constraints on the 3D helical orientations of 2-way junctions and additional long-range contacts serve to stabilize specific helical conformations [3, 4].

Westhof and coworkers classified 3-way junctions to 3 families (called A, B, and C) based on coaxial-stacking and orientation of helices, parallel, diagonal, or anti-parallel [60]. They also developed an algorithm for automatically predicting the topological family of any RNA 3-way junction, from given only 2D structure information, based on the parameters derived from a data set of 3-way junctions taken from known RNAs [55].

Schlick and coworkers focused on 4-way and higher-order junctions [51, 50, 53]. Laing et al. classified 4-way junctions into nine families according to coaxial stacking patterns and helical configurations [51]. They further developed a computational tool, called Junction-Explorer, to predict helical arrangements in 3- and 4-way junctions by a data mining approach, known as random forests, which relies on a set of decision trees trained using length, sequence, and free energy specified for any given junction [53]. The web server Junction-Explorer is freely available at http://bioinformatics.njit.edu/junction. Junction-Explorer can predict the junction family with an accuracy of 85% for 3-way junctions and 74% for 4-way junctions. These studies of RNA junctions help understand the global interaction network of RNAs and thus have great potential for guiding the challenging task of RNA 3D structure prediction.

4.3 *In Vitro* Selection Modeling *In Silico*

In the laboratory, new RNAs are designed by an experimental procedure called *in vitro* selection which involves the generation of large random-sequence pools ($\sim 10^{15}$) and screening the random pool for molecules that can perform a specific function such as binding to specific targets.

Theoretical analysis of random pools by Gevertz et al. [25] has shown that the random pools are biased towards simple topologies (e.g., see aptamer motifs in Figure 8).

Kim et al. [44, 46] developed a computational framework for modeling the *in vitro* selection process with specific targets in mind and provided computational tools for analyzing and improving RNA *in vitro* selection in three ways: by designing structured pools, by generating large sequence pools, and by screening target motifs in large pools. The concept of the "nucleotide transition probability matrix" was introduced to represent the mixing

ratios of nucleotides in synthesis ports. This makes it possible to design sequence pools targeting RNA-like topologies that are more likely to be active when produced experimentally.

To assist experimentalists, they made available design tools through a web-server, RAGPOOLS: RNA-As-Graph-Pools at http://rubin2.biomath.nyu.edu [45]. This work represents a first-generation effort to simulate *in vitro* selection processes *in silico* and exponentially increases the number of quality samples available to experimentalists. Efficient modeling has extended the approach to pools of up to 10^{14} sequences.

Other efforts in this direction include work by Chushak and Stone on generation and screening of 10^8 random-sequence pools for RNA aptamers for binding specific targets using a 3D folding and docking algorithms [16]. Luo et al. developed computational approaches to generate and evolve DNA pools for five-way junctions [61].

One of main challenges *in vitro* selection modeling *in silico* remains 3D structural characterization of the products to complement the 2D motif analysis. Advances in the prediction of RNA 3D structure require more conceptual breakthroughs to address the global positioning of RNA's 2D structure elements.

4.4 RNA Structure Prediction

RNA's modular, hierarchical, and network-like properties invite the application of graph approaches to RNA structure prediction. The prediction tools of RNA 2D structures from a given sequence (e.g., Mfold [95]) using energy minimization are already widely used. An alternative approach is the topology-based method, which generates an exhaustive list of all possible topologies, and also uses clustering analysis and shape parameters to narrow down the promising topologies [43]. Once these topologies are identified, the next step is to generate sequences to fold into them. Build-up approaches have been developed for this purpose [43].

The prediction of atomic 3D structure based on a given sequence is an ultimate goal of RNA structure prediction. Several programs predicting all-atom structures for small RNAs have been developed using fragment assembly and energy minimization (FARNA [18]) and modular approach using 4-6 nt building blocks called graph cycles (MC-SYM [72]). NAST [39] and iFold [84] use coarse-grained approach with one-bead or three-beads model, respectively. However, such programs are generally poor at predicting long-range interactions and, for higher prediction accuracy, manual manipulation and expert intuition are required; see recent reviews [52, 17].

The challenge remains for prediction of larger and more complex RNAs. An intermediate step toward 3D atomic structure prediction of any size and shape has been achieved using graphical representation to simplify 3D structures and reduce sampling space [26, 27]. Gillespie et al. represented RNA backbones as polygonal curves on the 3D triangular lattices and simulated the folding of RNA pseudoknots, by considering only the 3D conformations that can realize pseudoknot structures in the 3D space given the base pair restrictions [26]. Gopal et al. predicted and visualized the average structure of 2D/3D topology ensembles of

long viral RNAs (with thousands nucleotides) using enumeration of RAG tree graphs and coarse-grained molecular dynamics [27]. Ongoing efforts using graph approaches are aimed at predicting helical arrangements given 1D and 2D information via 3D motif constraints.

Currently, Schlick and coworkers are developing a hierarchical RNA folding approach using RAG 3D tree graphs and the geometrical analyses of RNA junctions and 3D motifs [54, 48]. The approach exploits the advantage of 3D tree graphs to reduce the conformational space for RNA and accelerate global sampling of candidate RNA 3D topologies. From given RNA 2D structures, initial 3D graphs are constructed by a junction family prediction program called Junction-Explorer [54]: Junction-Explorer is based on a data mining study that uses random trees to predict junction topology based on input feature vectors. The algorithm is trained on known RNA junctions. Following initial junction prediction, the Monte Carlo graph sampling is performed where each candidate graph is scored by a statistical potential energy function that guides preferred local and global conformations [48]. This graph-based hierarchical model is being tested on a set of known 3D structures, and shows promise for predicting large RNA structures. The translation of graph predictions to atomic models is also under current development.

5 Concluding Remarks and Open Problems

Network representations of complex RNA structures have been shown to be useful for representing, analyzing, predicting, and designing RNAs. From Waterman's graph approaches [89] to RAG tree and dual graphs [22], these various mathematical objects offer an efficient way of visualizing, classifying, and predicting RNA structures.

RAG's tree and dual graphs exploit the network-like nature of RNA 2D topologies [22, 23, 20, 37]. The application of graph theory to RAG representations has been used to measure overall compactness and enumerate all possible 2D topologies [43, 20]. The cataloguing, classification, and clustering of RAG graphs has expanded our understanding of RNAs structural repertoire, and allowed us to predict RNA-like topologies [43, 37, 49]. Graph enumeration and clustering algorithms have been used to explore RNA-like topologies [43, 37], and the graph merge method has been developed to predict larger RNA topologies [49]. The RAG database has been used to perform topology-based approaches which led to prediction of non-coding RNA searches in genomes [56] and submotif searches in larger RNAs [73].

Graph representations in combination with network and geometrical analysis have been applied to RNA structure prediction in multiple ways. By exploring network properties on 2D and 3D levels, recurrent motifs have been searched for and identified in known RNAs, which lays the groundwork for 3D structure prediction [94, 59, 92]. In particular, the 3D geometrical analysis of junction networks are led to topological classification of junction family types [3, 4, 60, 55, 51, 50]. To discover new RNAs with RNA-like topologies, the RAG catalog has been used to analyze, generate, and design targeted RNA structures to ultimately enhance *in vitro* selection experimental procedures [44, 46, 45]. The substantial reduction in conformational space size can enhance RNA structure prediction [26, 27, 54, 48].

Graph theory is expected to continue to be a useful tool for representing, analyzing, and predicting RNAs; such applications also provide exciting research opportunities in biology for mathematical scientists. We end with three open questions that require collaboration between mathematicians and biologists:

First, can we develop an alternative approach to clustering approaches for predicting RNA-like topologies based on the spectrum of a graph? The clustering analysis used previously by [43, 37] is based on supervised approaches trained on the whole set of all existing RNA topologies. This method has a considerable number of false positives [37]. An alternative approach is to identify in the RAG database those graphs that are non-isomorphic but have the same eigenvalues, determine which are known in nature, and then explore whether a certain spectral pattern lead to a fruitful way to predict novel RNAs. For example, if there are three specific non-isomorphic trees with the same spectrum and only one has been found experimentally, the other two may be RNA-like and good candidates for design. This approach relies on the well-known result in graph theory that two different non-isomorphic trees (for the same number of vertices) can have the same eigenvalues; that is, the vertex number and the eigenvalues alone cannot uniquely identify a topology.

Second, can we extend the RAG database for design of RNAs by build up by systematic identification of RNA submotifs for a target structure and links of submotifs to 1D/3D structures from the PDB? This problem involves rigorously identifying submotifs for a target structure, which can be an automated search of 3D subgraphs based on subgraph isomorphism. Established graph theory properties regarding subgraph isomorphism and maximum number of spanning trees could be appropriate here. Ongoing activity on determining graphs with maximum number of spanning trees (among graphs with the same number of vertices and edges) using the Laplacian eigenvalues and other optimization techniques could be valuable here.

Third, can we expand the repertoire of 3D motifs by applying spectral graph partitioning and the algebraic connectivity? So far, only a limited number of known 3D motifs has been identified and analyzed by manual inspection and computational searches from solved RNAs. The application of spectral graph partitioning could be useful to enumerate all existing 3D motifs by dividing 3D graphs translated from RNA 3D structures. The topological properties of 3D motifs could also be described by the *F iedler* eigenvalue and related to the diameter of the whole structure.

Unlike Frank Lloyd Wright's architecture, biological macromolecules cannot be made arbitrarily simple, but perhaps simplified network representations could lead to more innovative and successful approaches for addressing difficult real-life problems in RNA structure prediction and design.

Acknowledgments

References

[1]. Amos, M. Natural Computing Series. Springer; New York, NY: 2005. Theoretical and Experimental DNA Computation.

[2]. Bachellerie JP, et al. The expanding snoRNA world. Biochimie. 2002; 84(8):775–790. [PubMed: 12457565]

[3]. Bailor MH, et al. 3D maps of RNA interhelical junctions. Nat Protoc. 2011; 6(10):1536–1545. [PubMed: 21959236]

[4]. Bailor MH, Sun X, Al-Hashimi HM. Topology links RNA secondary structure with global conformation, dynamics, and adaptation. Science. 2010; 327(5962):202–206. [PubMed: 20056889]

[5]. Barabasi AL, Bonabeau E. Scale-free networks. Sci. Am. 2003; 288(5):60–69. [PubMed: 12701331]

[6]. Benedetti G, Morosetti S. A graph-topological approach to recognition of pattern and similarity in RNA secondary structures. Biophys. Chem. 1996; 59:179–184. [PubMed: 8867337]

[7]. Biggs, N. Algebraic graph theory. Cambridge University Press; 1993.

[8]. Boesch F, Lee X, Suffel C. On the existence of uniformly most reliable graphs. Networks. 1991; 2(21):181–194.

[9]. Boesch F, Suffel C, Satyanarayana A. A survey of some network reliability analysis and synthesis results. Networks. 2009; 2(54):99–107.

[10]. Boutorine YI, Steinberg SV. Twist-joints and double twist-joints in RNA structure. RNA. 2012; 18(12):2287–2298. [PubMed: 23060425]

[11]. Bray D. Molecular networks: the top-down view. Science. 2003; 301(5641):1864–1865. [PubMed: 14512614]

[12]. Breaker RR. Riboswitches and the RNA World. Cold Spring Harb Perspect Biol. 2012 doi:pii: a003566.

[13]. Burnett JC, Rossi JJ. RNA-based therapeutics: current progress and future prospects. Chem Biol. 2012; 19(1):60–71. [PubMed: 22284355]

[14]. Cheng C. Maximizing the total number of spanning trees in a graph: two related problems in graph theory and optimization design theory. Journal of Combinatory Theory (B). 1981; 31:240–248.

[15]. Church GM, Gao Y, Kosuri S. Next-generation digital information storage in DNA. Science. 2012; 337(6102):1628. [PubMed: 22903519]

[16]. Chushak Y, Stone M. In silico selection of RNA aptamers. Nucleic Acids Res. 2009; 37(12):e87. [PubMed: 19465396]

[17]. Cruz JA, et al. RNA-Puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction. RNA. 2012; 18(4):610–625. [PubMed: 22361291]

[18]. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. Proc. Natl. Acad. Sci. U. S. A. 2007; 104(37):14664–14669. [PubMed: 17726102]

[19]. R S. Eddy, Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. 2001; 2(12):919–929. [PubMed: 11733745]

[20]. Fera D, et al. RAG: RNA-As-Graphs web resource. BMC Bioinformatics. 2004; 5:88. [PubMed: 15238163]

[21]. Fiedler M. Algebraic connectivity of graphs. Czechoslovak Mathematical Journal. 1973; 23:298.

[22]. Gan HH, et al. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nucleic Acids Res. 2003; 31(11):2926–2943. [PubMed: 12771219]

[23]. Gan HH, et al. RAG: RNA-As-Graphs database–concepts, analysis, and features. Bioinformatics. 2004; 20(8):1285–1291. [PubMed: 14962931]

[24]. Gardner DP, et al. Statistical potentials for hairpin and internal loops improve the accuracy of the predicted RNA structure. J. Mol. Biol. 2011; 413(2):473–483. [PubMed: 21889515]

[25]. Gevertz J, et al. In vitro RNA random pools are not structurally diverse: a computational analysis. RNA. 2005; 11(6):853–863. [PubMed: 15923372]

[26]. Gillespie J, Mayne M, Jiang M. RNA folding on the 3D triangular lattice. BMC Bioinformatics. 2009; 10:369. [PubMed: 19891777]

[27]. Gopal A, et al. Visualizing large RNA molecules in solution. RNA. 2012; 18(2):284–299. [PubMed: 22190747]

[28]. Grabow WW, et al. The Right Angle (RA) Motif: A Prevalent Ribosomal RNA Structural Pattern Found in Group I Introns. J. Mol. Biol. 2012; 424(1-2):54–67. [PubMed: 22999957]

[29]. Gross, J.; Yellen, J. Graph Theory and Its Applications. CRC Press; 1996.

[30]. Gunsalus KC, et al. Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. Nature. 2005; 436(7052):861–865. [PubMed: 16094371]

[31]. Guo P. The emerging field of RNA nanotechnology. Nat Nanotechnol. 2010:833–842. [PubMed: 21102465]

[32]. Hamada M, et al. Mining frequent stem patterns from unaligned RNA sequences. Bioinformatics. 2006; 22(20):2480–2487. [PubMed: 16908501]

[33]. Hamilton AJ, Baulcombe DC. A species of small antisense RNA in posttranscriptional gene silencing in plants. Science. 2009; 286(5441):950–952. [PubMed: 10542148]

[34]. Harary, F. Graph Theory. Addison-Wesley; Mass: 1969.

[35]. Hotz RL. Decoding our chatter. The Wall Street Journal. Oct 1.2011 :C1–C2.

[36]. Humphris-Narayanan E, Pyle AM. Discrete RNA libraries from pseudo-torsional space. J. Mol. Biol. 2012; 421(1):6–26. [PubMed: 22425640]

[37]. Izzo JA, et al. RAG: an update to the RNA-As-Graphs resource. BMC Bioinformatics. 2011; 12:219. [PubMed: 21627789]

[38]. Johnson M. Structure-activity maps for visualizing the graph variables arising in drug design. J. Biopharm. Stat. 1993; 3(2):203–236. [PubMed: 8220404]

[39]. Jonikas MA, et al. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA. 2009; 15(2):189–199. [PubMed: 19144906]

[40]. Kalir S, Alon U. Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. Cell. 2004; 117(6):713–720. [PubMed: 15186773]

[41]. Karklin Y, et al. Classification of non-coding RNA using graph representations of secondary structure. Pac. Symp. Biocomput. 2005; 4:15.

[42]. Kelmans A, Chelnokov V. A certain polynomial of a graph and graphs with maximum number of spanning trees. Journal of Combinatory Theory (B). 1974; 16:197–214.

[43]. Kim N, et al. Candidates for novel RNA topologies. J. Mol. Biol. 2004; 341(5):1129–1144. [PubMed: 15321711]

[44]. Kim N, et al. A computational proposal for designing structured RNA pools for in vitro selection of RNAs. RNA. 2007; 13(4):478–492. [PubMed: 17322501]

[45]. Kim N, et al. RAGPOOLS: RNA-As-Graph-Pools–a web server for assisting the design of structured RNA pools for in vitro selection. Bioinformatics. 2007; 23(21):2959–2960. [PubMed: 17855416]

[46]. Kim N, et al. Computational generation and screening of RNA motifs in large nucleotide sequence pools. Nucleic Acids Res. 2010; 38(13):e139. [PubMed: 20448026]

[47]. Kim, N.; Fuhr, KN.; Schlick, T. Graph Applications to RNA Structure and Function. In: Russell, Rick, editor. Biophysics of RNA Folding. Springer; 2013. Chapter 3

[48]. Kim N, et al. Graph-based sampling approach for approximating global helical topologies of RNA. 2013 In preparation.

[49]. Koessler DR, et al. A predictive model for secondary RNA structure using graph theory and a neural network. BMC Bioinformatics. 2010; 11(Suppl 6):S21. [PubMed: 20946605]

[50]. Laing C, Schlick T. Analysis of four-way junctions in RNA structures. J. Mol. Biol. 2009; 390(3):547–559. [PubMed: 19445952]

[51]. Laing C, et al. Tertiary motifs revealed in analyses of higher-order RNA junctions. J. Mol. Biol. 2009; 393(1):67–82. [PubMed: 19660472]

[52]. Laing C, Schlick T. Computational approaches to 3D modeling of RNA. J Phys. Condens Matter. 2010; 22(28):283101–283118. [PubMed: 21399271]

[53]. Laing C, et al. Predicting coaxial helical stacking in RNA junctions. Nucleic Acids Res. 2012; 40(2):487–498. [PubMed: 21917853]

[54]. Laing C, et al. Predicting helical topologies in RNA junctions as tree graphs. 2013 Submitted.

[55]. Lamiable A, et al. Automated prediction of three-way junction topological families in RNA secondary structures. Comput. Biol. Chem. 2012; 37:1–5. [PubMed: 22326420]

[56]. Laserson U, et al. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs. Nucleic Acids Res. 2005; 33(18):6057–6069. [PubMed: 16254081]

[57]. Le S, et al. Tree graphs of RNA secondary structures and their comparisons. Comput. Biomed. Res. 1989; 22:461–473. [PubMed: 2776449]

[58]. Lee DS, et al. The implications of human metabolic network topology for disease comorbidity. Proc. Natl. Acad. Sci. U. S. A. 2008; 105(29):9880–9885. [PubMed: 18599447]

[59]. Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. RNA. 2001; 7(4):499–512. [PubMed: 11345429]

[60]. Lescoute A, Westhof E. Topology of three-way junctions in folded RNAs. RNA. 2006; 12(1):83–93. [PubMed: 16373494]

[61]. Luo X, et al. Computational approaches toward the design of pools for the in vitro selection of complex aptamers. RNA. 2010; 16(11):2252–2262. [PubMed: 20870801]

[62]. Mandado M, et al. Chemical graph theory and n-center electron delocalization indices: a study on polycyclic aromatic hydrocarbons. J. Comput. Chem. 2007; 28(10):1625–1633. [PubMed: 17342712]

[63]. Mandelbrot, B. Fractals: Form, Chance and Dimension. W H Freeman and Co; 1977.

[64]. Mandelbrot, B. The Fractal Geometry of Nature. W H Freeman and Co; 1982.

[65]. Mandelbrot, B. A Multifractal Walk down Wall Street. Scientific American; Feb. 1999

[66]. Mäcke TJ, et al. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. 2001; 29(22):4724–4735. [PubMed: 11713323]

[67]. Milo R. Superfamilies of evolved and designed networks. Science. 2004; 303(5663):1538–1542. [PubMed: 15001784]

[68]. Mohar, B. The Laplacian spectrum of graphs. In: Alavi; Chartrand; Qellermann; Schwenk, editors. Graph theory, combinatorics, and applications. Wiley; 1991. p. 817-898.

[69]. Mohar, B. Graph Laplacians. In: Beineke; Wilson, editors. Topics in algebraic graph theory. Cambridge University Press; 2004. Chapter 4

[70]. Nobel Foundation. Nobel Prize in Economics 2012: Alvin E. Roth and Lloyd S. Shapley "for the theory of stable allocations and the practice of market design". Oct 15.2012

[71]. Nudler E. Flipping riboswitches. Cell. 2006; 126(1):19–22. [PubMed: 16839869]

[72]. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008; 452(7183):51–55. [PubMed: 18322526]

[73]. Pasquali S, Gan HH, Schlick T. Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs. Nucleic Acids Res. 2005; 33(4):1384–1398. [PubMed: 15745998]

[74]. Petingi L, Boesch F, Suffel C. On the Characterization of Graphs with Maximum Number of Spanning Trees. Discrete Mathematics. 1998; 1-3(179):155–166.

[75]. Petingi L, Rodriguez J. A New Technique for the Characterization of Graphs with Maximum Number of Spanning Trees. Discrete Mathematics. 2002; 1-3(244):351–373.

[76]. Petrov AI, et al. WebFR3D–a server for finding, aligning and analyzing recurrent RNA 3D motifs. Nucleic Acids Res. 2011; 39(Web Server issue):W50. [PubMed: 21515634]

[77]. Pothen A, et al. Partition space matrices with eigenvectors in graphs. SIAM J. on Matrix Analysis and Applications. 1990; 11:430.

[78]. Quarta G, et al. Analysis of riboswitch structure and function by an energy landscape framework. J. Mol. Biol. 2009; 393(4):993–1003. [PubMed: 19733179]

[79]. Quarta G, Sin K, Schlick T. Dynamic energy landscapes of riboswitches help interpret conformational rearrangements and function. PLoS Comput Biol. 2012; 8(2):e1002368. [PubMed: 22359488]

[80]. Que-Gewirth NS, Sullenger BA. Gene therapy progress and prospects: RNA aptamers. Gene Ther. 2007; 14(4):283–291. [PubMed: 17279100]

[81]. Ren A, Rajashankar KR, Patel DJ. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. Nature. 2012; 486(7401):85–89. [PubMed: 22678284]

[82]. Rivas E, Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 1999; 285(5):2053–2068. [PubMed: 9925784]

[83]. Shapiro B, Zhang K. Comparing multiple RNA secondary structures using tree comparisons. Comput. Appl. Biosci. 1990; 6(5):309–318. [PubMed: 1701685]

[84]. Sharma S. iFold: a platform for interactive folding simulations of proteins. Bioinformatics. 2006; 22(21):2693–2694. [PubMed: 16940324]

[85]. Sharp PA. The centrality of RNA. Cell. 2009; 136(4):577–580. [PubMed: 19239877]

[86]. Spielman DA, Teng SH. Spectral partitioning works: Planar graphs and finite element meshes. Linear Algebra and Its Applications. 2007; 421:284–305.

[87]. Stombaugh J, et al. Frequency and isostericity of RNA base pairs. Nucleic Acids Res. 2009; 37(7):2294–2312. [PubMed: 19240142]

[88]. Yook SH, et al. Modeling the Internet's large-scale topology. Proc. Natl. Acad. Sci. U. S. A. 2003; 99(21):13382–13386. [PubMed: 12368484]

[89]. Waterman MS. Secondary structure of single-stranded nucleic acids. Adv. Math. Suppl. Stud. 1978; 1:167–212.

[90]. Williams KP. The tmRNA Website: invasion by an intron. Nucleic Acids Res. 2002; 30(1):179–182. [PubMed: 11752287]

[91]. Witz G, Dietler G, Stasiak A. Tightening of DNA knots by supercoiling facilitates their unknotting by type II DNA topoisomerases. Proc. Natl. Acad. Sci. U.S.A. 2011; 108(9):3608–3611. [PubMed: 21321228]

[92]. Xin Y, et al. Annotation of tertiary interactions in RNA structures reveals variations and correlations. RNA. 2008; 14(12):2465–2477. [PubMed: 18957492]

[93]. Zirbel CL, et al. Classification and energetics of the base-phosphate interactions in RNA. Nucleic Acids Res. 2009; 37(15):4898–4918. [PubMed: 19528080]

[94]. Zorn J, et al. Structural motifs in ribosomal RNAs: implications for RNA design and genomics. Biopolymers. 2004; 73(3):340–347. [PubMed: 14755570]

[95]. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003; 31(13):3406–3415. [PubMed: 12824337]
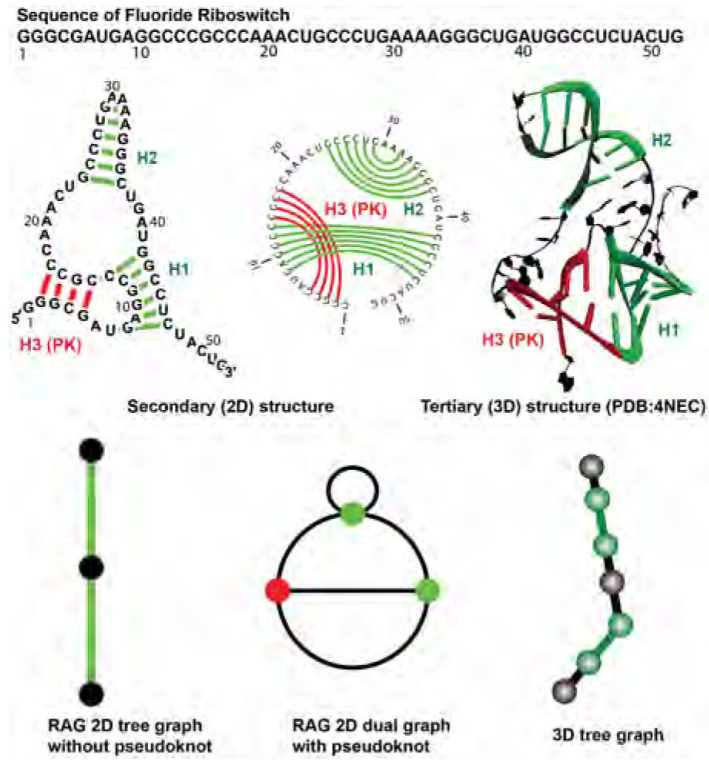
Figure 1.
RNA's hierarchical network-like structures from 1D to 3D. The primary (1D), secondary (2D), and tertiary (3D) structures associated with fluoride riboswitch (PDB:4NEC [81]) are shown. The RAG tree and dual graphs representing its 2D topology without and with pseudoknot, respectively, are also shown. The double-stranded or stem regions are colored green for tree topologies and red for pseudoknot formation.
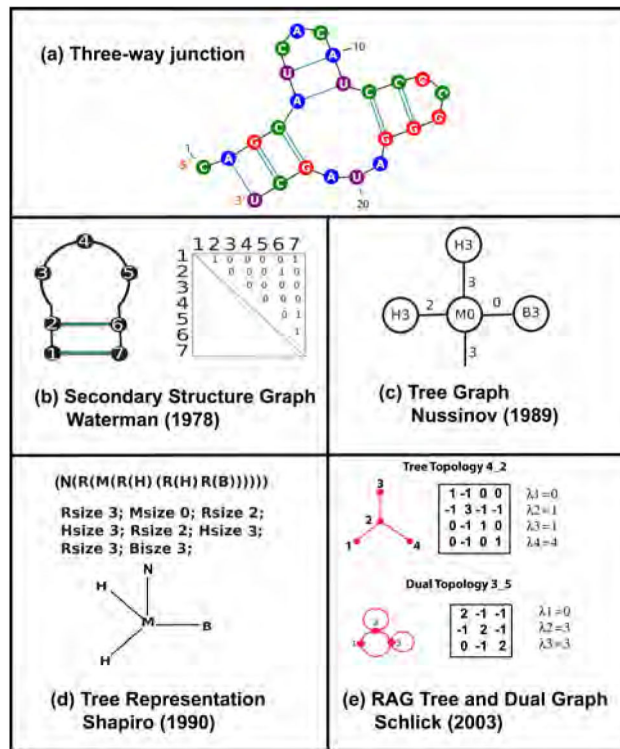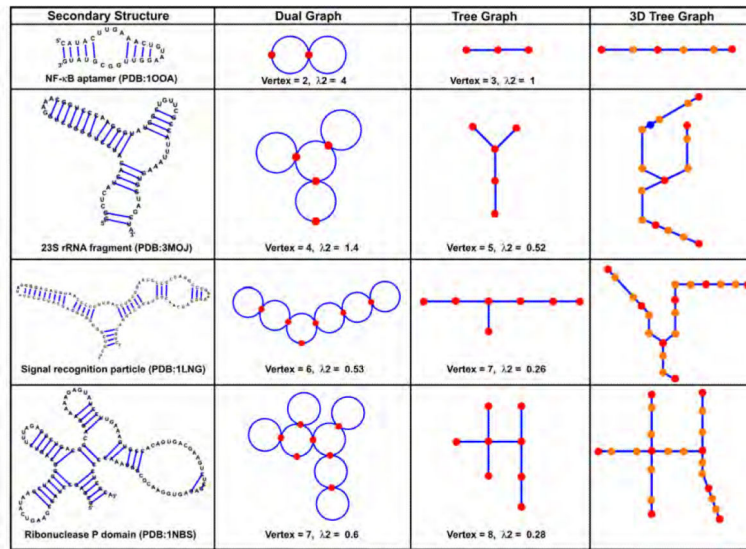
Figure 2.
Graph theoretical models of an RNA 2D structure for a three-way junction (a) by Waterman, Nussinov, Shapiro, and Schlick. (b) Waterman (1978): graphs at base-pair level; (c) Nussinov (1989): ordered labeled tree graphs B, I, H, M, and S represent bulge, internal loop, bifurcation loop, and single-stranded region, respectively; (d) Shapiro (1990): simplified tree graphs where R represent a paired region; (e) Schlick (2003): RAG tree and dual graphs.

Figure 3.
The RAG translation rules of RNA 2D structures as dual, tree, and 3D tree graphs (as developed in our recent work, see text). For each 2D structural element, the associated dual, tree, and 3D tree graphs are shown along with Laplacian of the tree graph. RNAs with pseudoknot structures can only be presented by a dual graph.
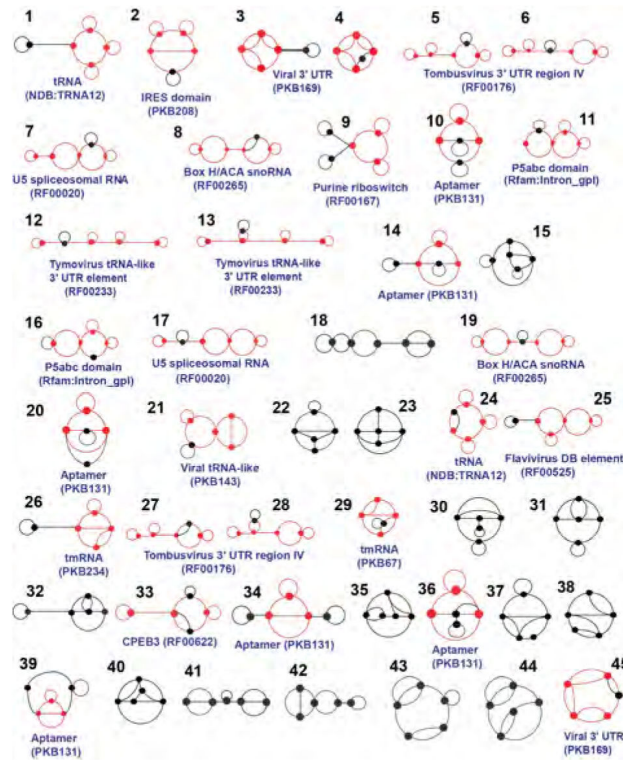
Figure 4.
Examples of known RNAs with their RAG represenations: dual, tree, and 3D tree graphs (as developed in our recent work, see text). Each RAG graph is also described by the vertex number and the second smallest Laplacian eigenvalue.
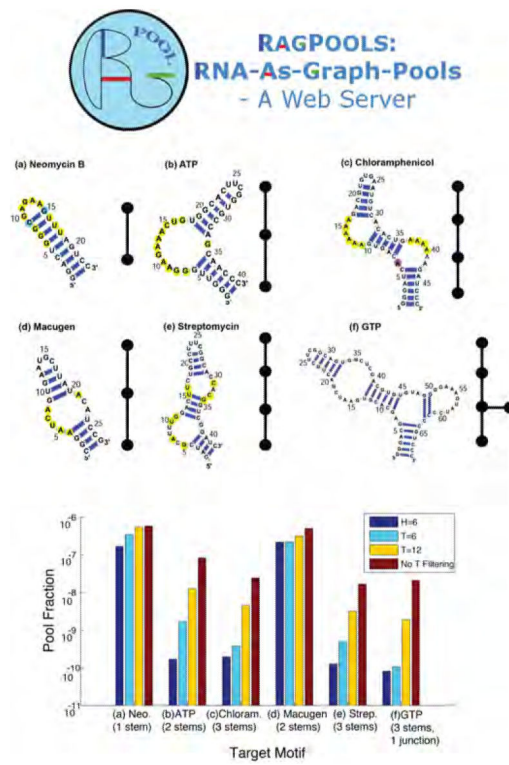
Figure 5.
RNA tree graph library segment. Tree graphs are associated with existing RNAs (from the PDB) and classified as RNA-like and non-RNA-like. The eigenvalues of RNA topologies are a measure of structural compactness.

Figure 6.
RAG tree and dual graphs. (a) The library of topologies of RAG tree graphs between 2 and 10 vertices, with the second smallest Laplacian eigenvalues (λ_2) listed. (b) RAG dual graphs between 2 and 4 vertices, with topological classification (tree, bridge, and pseudoknot). The status of existing, RNA-like, and non-RNA-like topologies is represented by red, blue, and black (dashed), respectively (see text). The motifs in the yellow box indicate the five candidate motifs (C1, C2, C3, C4, and C7) as of 2004 [43] which have been identified as CPEB3, purine riboswitch, tRNA-like, Tombusvirus 3′ UTR region, and Flavivirus, respectively, in the updated RAG as of 2011 [37].

Figure 7.
45 RNA-like topologies with 5 vertices predicted by supervised clustering analysis based on existing RNAs [37]. The submotifs found in nature are colored red.

Figure 8.
Pool screening and filtering analysis for 10^{12} random-sequence pools for aptamers.