

Contents lists available at [ScienceDirect](#)

Methods

journal homepage: www.elsevier.com/locate/ymeth

Adventures with RNA graphs

Tamar Schlick ^{a,b,c,*}^a Department of Chemistry, 100 Washington Square East, Silver Building, New York University, New York, NY 10003, USA^b Courant Institute of Mathematical Sciences, New York University, 251 Mercer St., New York, NY 10012, USA^c New York University ECNU - Center for Computational Chemistry at NYU Shanghai, 3663 North Zhongshan Road, Shanghai, 200062, China

ARTICLE INFO

Article history:

Received 11 January 2018

Received in revised form 7 March 2018

Accepted 26 March 2018

Available online xxxxx

Keywords:

RNA structure

Graphs

Coarse-grained modeling

RNA secondary structure

Mathematical biology

RNA design

ABSTRACT

The structure of RNA has been a natural subject for mathematical modeling, inviting many innovative computational frameworks. This single-stranded polynucleotide chain can fold upon itself in numerous ways to form hydrogen-bonded segments, imperfect with single-stranded loops. Illustrating these paired and non-paired interaction networks, known as RNA's secondary (2D) structure, using mathematical graph objects has been illuminating for RNA structure analysis. Building upon such seminal work from the 1970s and 1980s, graph models are now used to study not only RNA structure but also describe RNA's recurring modular units, sample the conformational space accessible to RNAs, predict RNA's three-dimensional folds, and apply the combined aspects to novel RNA design. In this article, we outline the development of the RNA-As-Graphs (or RAG) approach and highlight current applications to RNA structure prediction and design.

© 2018 Elsevier Inc. All rights reserved.

Contents

1. Introduction	00
1.1. Mathematical biology modeling	00
1.2. Examples of graphs or networks	00
1.3. Utility of RNA representations as networks	00
2. RNA 2D structures as graphs	00
2.1. RNA structure	00
2.2. History of RNA graphs	00
3. RAG-RNA-As graphs framework – 2D and 3D graphs, linear algebra machinery	00
3.1. Coarse-grained models	00
3.2. RAG overview	00
3.3. Linear algebra machinery	00
3.4. Software and application overview	00
4. RAG motif enumeration and design	00
4.1. Clustering into RNA-like and non-RNA-like motifs	00
4.2. RAG-3D for graph partitioning and motif search	00
4.3. Automated fragment assembly for design	00
5. RAGTOP for 3D structure prediction	00
5.1. RNA structure prediction difficulties	00
5.2. RAGTOP protocol	00
6. Conclusions	00
6.1. Biomolecular simulations as a field on its own right	00
6.2. Future challenges with RAG	00
Acknowledgment	00
References	00

* Address: Department of Chemistry, 100 Washington Square East, Silver Building, New York University, New York, NY 10003, USA.

E-mail address: schlick@nyu.edu

1. Introduction

1.1. Mathematical biology modeling

The complexity and beauty of the biological world has long been interpreted through a mathematical lens. Mathematical analysis was crucial to the interpretation of heredity by Gregor Mendel and to the theory of evolution by Thomas Darwin. The language of calculus, laid by Isaac Newton and Gottfried Leibniz, is fundamental for studying many types of motion, from planetary orbits to the dynamics of biomolecules fundamental for life's basic processes. Mathematical and statistical frameworks have been essential for studying genetic diseases, the spread of epidemics, or the behavior of cellular ensembles. As biology has advanced in dazzling speed over the past decades since the elucidation of DNA structure, opportunities for employing various areas of mathematics to study biological systems have exploded. Now, mathematics and statistics are key aspects of biomolecular modeling, genomics, proteomics, brain research, environmental science, and many other modern biology and scientific subfields. There is no wonder that mathematical and computational frameworks are considered to be invaluable for the future health, wealth, and security of our modern technological society.

In turn, over the past hundreds of years, biological systems and problems have also motivated theoretical developments in many fields of mathematics and computation, notably geometry and topology, algebra, analysis, and computer science theory. In particular, graph, or network, theory and topology are fields of mathematics that have received much attention as they are applicable to networks found in numerous fields, from social networks and cellular connections to economic, and transportation, and security networks.

1.2. Examples of graphs or networks

A graph (or network) is a discrete object $G = (V, E)$ with vertices V and edges E ; graphs can be *directed* or *undirected* and *weighted* or *unweighted* by their content (e.g., number of bases); simple rules are needed to translate elements of the network into edges and vertices.

Examples of graphs or networks are the London tube map (see <http://content.tfl.gov.uk/standard-tube-map.pdf>), or a genetic network of budding yeast regulating cell cycle and sporulation (e.g., Fig. 1 of [1]).

The algorithm by which Google returns an ordered list of links/websites to the user's query keyword (PageRank scheme) is a network: The Internet is one giant graph; each webpage is a node; and two pages are joined by an edge if there is a link from one page to the other. PageRank works by the principle that the more links to a page, the more important it is perceived. Thus, knowing how this algorithm works can be exploited to get one's site near the top of the list.

The well-known "traveling salesman problem" [2] leads to a graph: Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city? This is well-recognized as "Non-deterministic polynomial time" hard problem in combinatorial optimization. It can be modeled as an undirected weighted graph, such that cities are the graph's vertices, paths are the graph's edges, and a path's distance is the edge's length. Today, such problems are easily solved on modern computers for thousands of cities. However, in 1962, a contest by Proctor & Gambler offered \$10,000 cash prizes (valued to be almost ten times as much today) to solutions of such a problem with 33 cities. See poster in <http://www.math.uwaterloo.ca/tsp/history/pictorial/car54.html>. A

detailed solution is provided in [3]; see also solution illustration in p. 15 of [2].

More recently, a shareability network model was developed to estimate how car pooling might reduce cost and pollution in New York City [4]. Traditionally, such "dynamic pickup and delivery" problems – where a number of goods or customers must be picked up and delivered efficiently at specific destination within time windows (think Uber taxis or FreshDirect grocery deliveries) – are solved by linear programming. Linear programming involves a system of linear equations for a set of variables subject to constraints. Instead of solving this cab sharing problem by linear programming, the researchers in [4] defined a shareability network model. Each trip in the network is represented by a vertex, and each shared trip is represented by an edge. Application of the method to a dataset of 150 million taxi trips in New York City suggested that considerable savings in cost and traffic can be realized while still keeping the travel time low [4]. Most large cities around the world today will need to adapt such traffic-cutting measures to solve current traffic woes.

1.3. Utility of RNA representations as networks

RNA structure has been a natural subject for mathematical analysis by graph theory. Though similar in chemical composition to molecular biology's superstar, DNA, RNA's single-strandedness with its 2' hydroxyl group allows this polymer chain to fold upon itself and form networks of hydrogen bonds, imperfect with single-stranded loops (see Fig. 1). This hydrogen-bonding network immediately suggests networks or graphs.

Why study RNA? RNA has always been an important biological subject due its key role in the central dogma of biology. However, interest in RNA structure, function, and design has exploded over the recent two decades with the discovery of noncoding regulatory RNAs [5,6], including micro RNAs (miRNAs) and long noncoding RNAs (lncRNAs) that control gene expression through diverse cellular pathways.

RNA's structural versatility translates to functional wizardry. RNA can replicate itself, act as an enzyme, and serve as a template for protein synthesis. Moreover, the crucial relation of RNAs to many human diseases is also becoming apparent [6–10], opening new opportunities for disease detection and therapy [11]. For example, CRISPR RNA technology for gene editing has found numerous applications in research and medicine, having the potential to treat genetic diseases and offer novel targeted drug therapies for human diseases using nucleic acid targets rather than proteins and other compounds. Besides genomic RNAs, synthetic RNAs by *in vitro* selection [12,13] have significantly expanded RNA's repertoire and created many opportunities in nanotechnology and biomedicine [14–16].

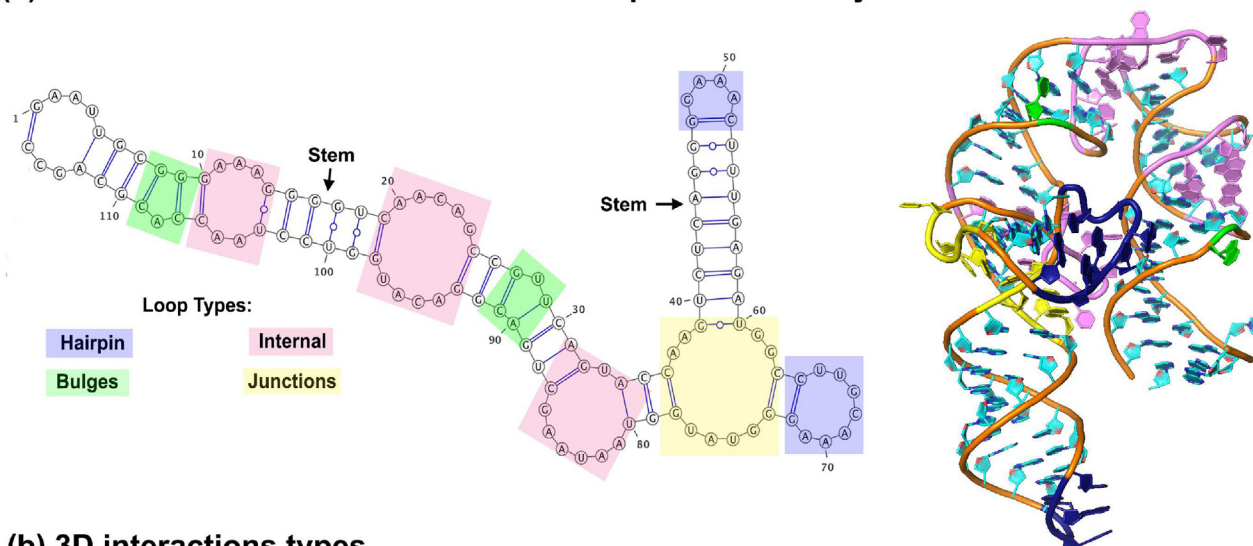
As with protein structural genomics, a primary goal of ribonomics is to catalog all distinct RNA folds across functional RNA classes to determine sequence/motif/functional relationships [17,18]. Mathematical modeling has thus played an active role in advancing this exciting field of RNA science. In the remainder of this article, we introduce graph representations of RNA and describe various applications to RNA structure analysis and design using our group's RAG (RNA-As-Graphs) framework.

2. RNA 2D structures as graphs

2.1. RNA structure

RNA structure can be described by its *primary* (nucleotide sequence), *secondary* or *2D* (hydrogen-bond pairing arrangements that define double-stranded or stem regions and single-stranded

(a) 2D and 3D Structures of the P4-P6 Group I Intron Ribozyme Domain



(b) 3D interactions types

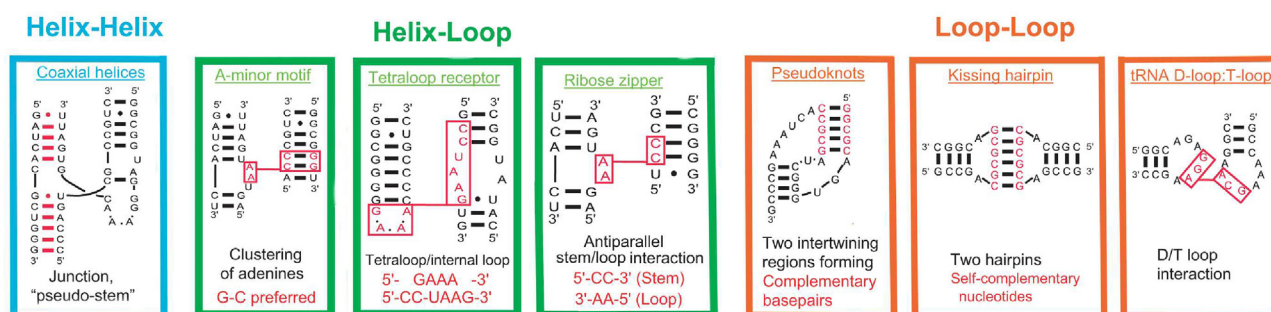


Fig. 1. RNA secondary (2D) and tertiary (3D) interactions. (a) The 2D and 3D structures of the P4-P6 Group I Ribozyme Domain are sketched. (b) Various types of possible 3D interactions are illustrated.

loops), and *tertiary or 3D* interactions (see Fig. 1). RNA pseudoknots are *super secondary* interactions defined when hydrogen-bonding interactions dictate intertwined, or knotted-like regions (Fig. 2), leading to complex tertiary folds. Though RNA structure organization is generally thought to be hierarchical, with its structural organization occurring in discrete states or transitions from 2D to 3D interactions [19,20], recent work on ribozymes suggests a more intricate coupling between 2D and 3D elements, with protein and/or ions guiding the folding process and dictating the folding pathways [21,22].

Nonetheless, the well-defined geometric aspects of RNA 2D structures are good starting points for analyzing and modeling RNAs, and the network of hydrogen bonds in RNA 2D structures explains why graphs and graph-like objects have long been natural objects to represent RNAs (see Fig. 3).

2.2. History of RNA graphs

In 1971, Tinoco and co-workers have introduced what we call the *Tinoco plot* (see Fig. 3a). These plots indicate 2D structures of RNA, as deduced by energy minimization using nearest-neighbor thermodynamic parameters for the different base-pair terms and solved by efficient dynamic programming (DP) algorithms [23]. Today, such dynamic programming algorithms and programs for predicting RNA 2D structures, like RNAfold [24] and NUPACK [25], though imperfect, are excellent starting points from a given nucleotide sequence [26].

In 1978, Nussinov and co-workers introduced convenient circular and linked graphs views to easily visualize the base-pairing in RNAs by arcs: see Fig. 3b [27]. The circular variant is more economical in usage of space. Both views allow easy detection of pseudoknots by the crossing of base-pair connections.

In the same year, Waterman and co-workers pioneered graphical representations of RNA with the aim of analyzing 2D structures of tRNA [28]. Waterman offered the first graph-theoretic definition of 2D structures, planar graph, along with the adjacency matrix. This matrix describes which edges of the graph are connected to one another: see Fig. 3c.

An interesting mountain plot was introduced in 1984 by Hoggeweg & Hesper [29]: the height $m(k)$ is the cumulative number of paired bases at position k . This mountain plot allows straightforward comparison of structures and inspired a convenient algorithm for comparison of 2D structures (see Fig. 3d).

Soon after, Nussinov developed the ordered labeled tree graph to compare 2D structures of RNA [30] (see Fig. 3e).

Tree graphs were also used by Shapiro and collaborators to measure 2D-structural similarities. In particular, they defined the *tree edit distance* between two (full) 2D tree structures to quantify the minimum cost (by insertion, deletion, and replacement of nodes) along an edit path for converting one tree graph into another [31,32] (see Fig. 3f).

In 2003, our group has contributed to these exciting advances in the RNA field by developing a graph-theoretical framework and web server called RAG (RNA-As-Graphs) to describe and catalog 2D structures of RNA (<http://www.biomath.nyu.edu/rag/home>),

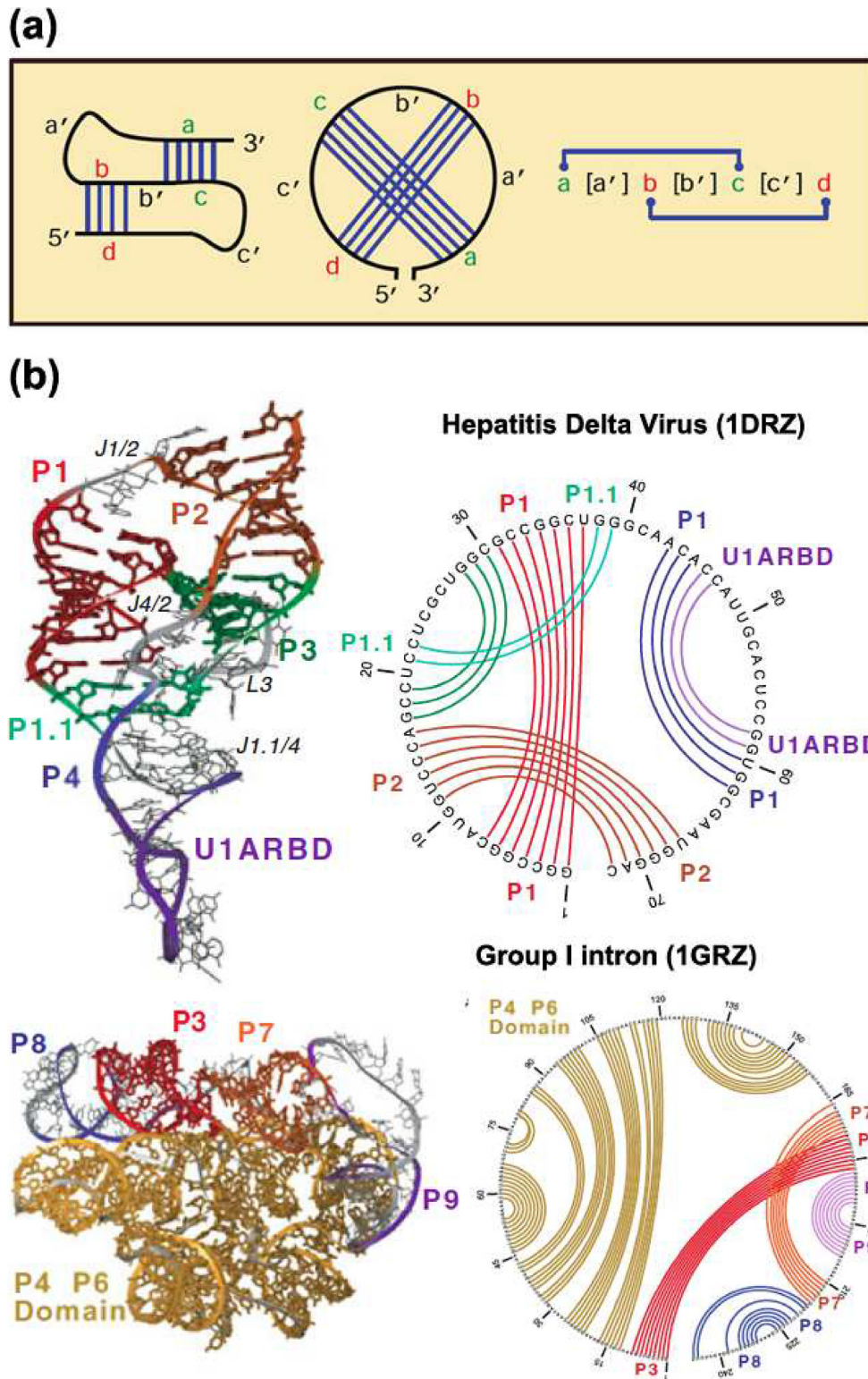


Fig. 2. RNA pseudoknots. (a) RNA pseudoknots are defined by an intertwined form of base pairing, which leads to crossing of base pairs in the circular representation. (b) Examples of pseudoknotted RNAs are also shown.

both as planar tree graphs and dual graphs [33] (see Fig. 3g and Fig. 4). In tree graphs, the 2D elements bulges, junctions, and loops are represented as vertices, and stems are edges. In dual graphs, the rules used above to formulate tree graphs are reversed: bulges, junctions, and loops are edges, and stems are vertices. Dual graphs are less intuitive to work with but they have the important advantage

of being able to represent pseudoknots, a motif frequent in many RNAs. Later, we extended our RAG tree representation into three-dimensional (3D) space by including additional vertices at helix ends and junction centers, and scaling edges to represent helix sizes (see Fig. 5) [34]. Such 3D graph representations may or may not correspond to the folded RNA structure. That is, our

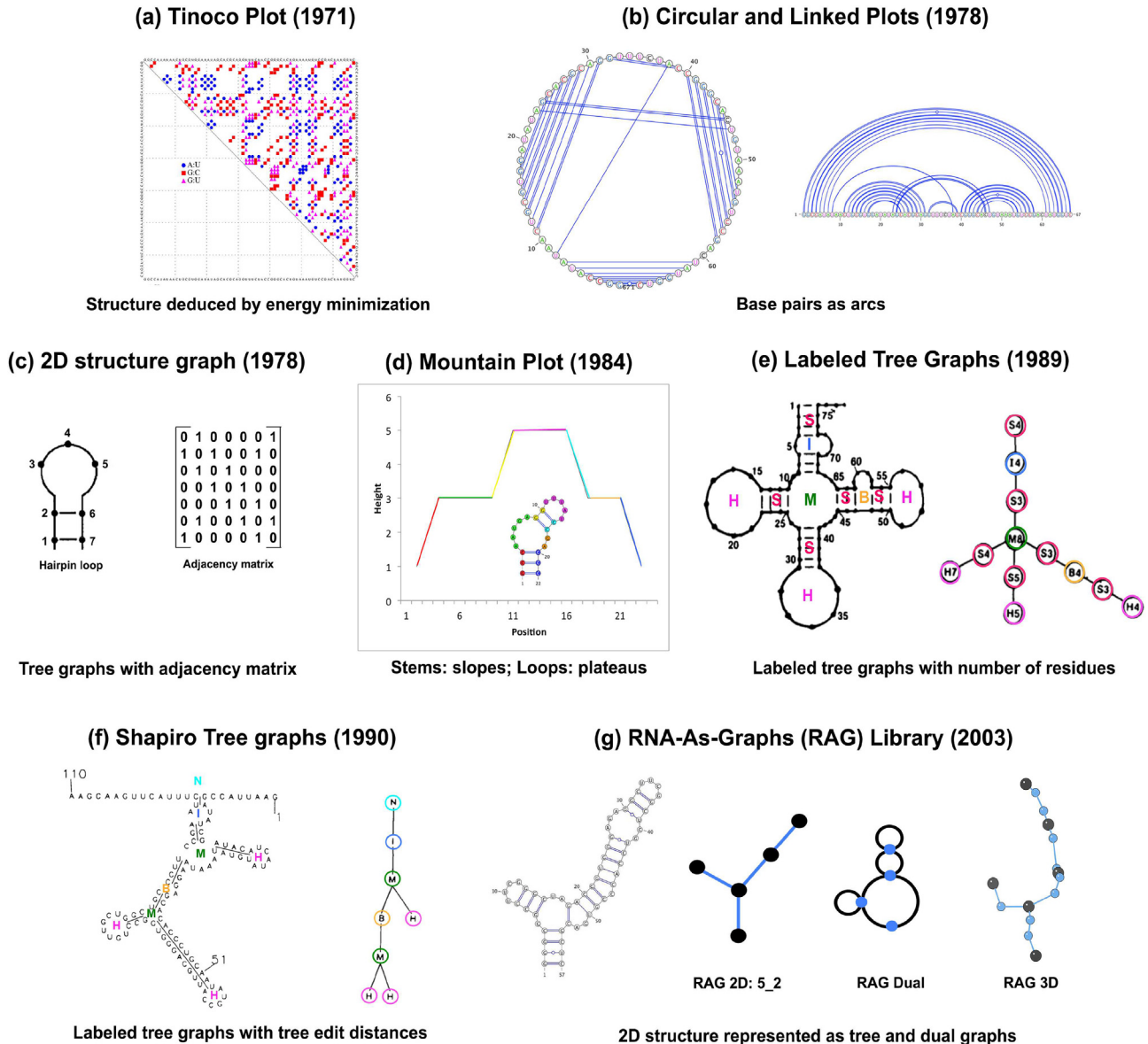


Fig. 3. History of RNA Graphs. See text for details.

3D RAG graph may be constructed from the solved structure, as shown in Fig. 5, or predicted by our RAGTOP sampling procedure, as shown in Fig. 14 (Candidate graph), described below.

The motivation for these RAG representations was not only to establish a convenient framework for enumerating the universe of RNA motifs but also for stimulating the prediction of structures and design of new RNA motifs on modern computers. When motifs are enumerated on the sequence level, there are numerous possibilities; some examples in the classes of double-stranded (coaxial helices), double/single strands, and single/single stranded motifs are sketched in Fig. 1.

3. RAG-RNA-As graphs framework – 2D and 3D graphs, linear algebra machinery

3.1. Coarse-grained models

The graph models of RNAs described above, developed as early as the 1970s, by Waterman, Nussinov, Shapiro, and others, as

reviewed recently [35,36], provide valuable representations for analysis. The graph theoretical approaches developed in our lab aim to extend the graph constructs to RNA simulation, exploiting the drastic reduction in conformational space by graph representations.

While it is possible and valuable to simulate RNA at atomic resolution, for example to gain insights into ribosome motion [37] and catalysis [38], such simulations require enormous computational resources and expert knowledge in the modeling details. Expertise is required to treat the solvation and ionic atmosphere of RNA and to handle force-field issues for atomic RNA models to guarantee stability and reliability of the results [39,40].

Complementary to these atomic-level simulations and modeling studies, various coarse-grained representations of RNA (e.g., [41–45]) have shown to be effective in many applications, including configurational sampling, structure prediction, and RNA design. Simplified representations can capture essential features of biomolecules while making computations accessible for a variety of applications due to a drastic reduction in the number of degrees of freedom. Historically, united-atom representations for proteins

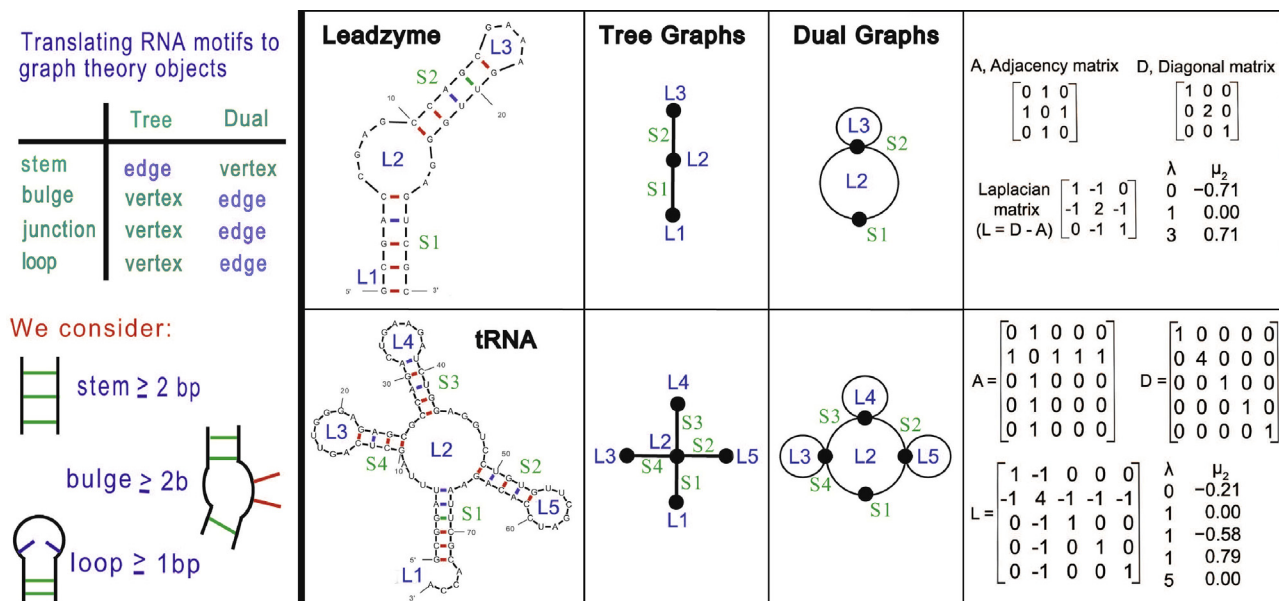


Fig. 4. RAG representations of RNA 2D structures as tree and dual graphs. The associated matrices (adjacency, diagonal, and Laplacian) are also shown, along with the eigenvalues $\{\lambda_i\}$ and the second eigenvector, μ_2 .

were used to simulate their dynamics (e.g., [46]), and recent coarse grained models of chromatin have provided insights into chromatin architecture [47].

For RNA, the graph framework offers a natural way to coarse grain the system and employ a large body of graph theory and topology machinery for structure analysis, partitioning, and simulation.

3.2. RAG overview

In our RAG representation, we translate RNA 2D structures into tree and dual graph objects with the following definitions: In tree graphs, stems are edges, and junctions, bulges, and loops are vertices. We reverse those definitions to define dual graphs [48] (see examples in Figs. 4 and 5).

For 3D RNA sampling and structure prediction, our extension of RAG 2D tree graphs [49] considers the size of helices and loops as well as the parallel/anti-parallel helical arrangements. In our revised graphs, we add vertices at helix ends to represent parallel and anti-parallel helical arrangements. Thus, in 3D graphs, a helix is represented as an edge and two vertices; additional edges connect new helix vertices to a loop vertex. We also add weights to each edge to represent helix lengths and size of unpaired regions.

Two examples of 3D graphs constructed from the known 3D structures are shown in Fig. 5, superimposed onto the 3D structures. We also show the corresponding 2D structures, and the planar tree and dual graphs. In the 3D graphs, the lengths of the edges depend on the sizes of the helix, loop, bulge, and junction elements. Optionally, we can add edges in such 3D graphs to represent pseudoknot interactions by connecting specific loops/junctions to other loops/junctions (see Section 5).

In Fig. 5, the RNase P (PDB 1NBS) has 120 nucleotides and 2583 atoms; the corresponding graph has 28 vertices. The 23S ribosomal RNA (PDB 1S72) of *H. marismortu* from archaea has 2922 nucleotides and 59021 atoms; our 3D graph of this rRNA has 469 vertices. Such 3D graph representations provide an efficient computational framework to model RNA 3D geometry and sample RNA 3D space by simplified, coarse-grained biomolecular models.

3.3. Linear algebra machinery

Besides the pictorial view, the associated linear algebra machinery is useful to describe, compare, and analyze graph objects. Specifically, we associate the n by n Laplacian matrix L with each graph, where n = number of vertices, to be the difference between the diagonal matrix D (which describes the connections from each vertex) and A , which indicates whether each i, j pair is connected or not. See Fig. 4 for illustrations of these matrices.

The Laplacian has non-negative eigenvalues ($0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$), and the second eigenvalue λ_2 is a measure of overall compactness. We use λ_2 to order our RNA graphs in the RAG catalog. We also use the second eigenvector μ_2 with n components $\{\mu_{2,1}, \dots, \mu_{2,n}\}$ for partitioning graphs [50]. Specifically, we use the *gap cut* method to partition tree graphs into subgraphs [50]: this method divides the graph between the two indices k, m whose respective components of μ_2 generate the largest numerical difference (i.e., $|\mu_{2,k} - \mu_{2,m}| = \max |\mu_{2,i} - \mu_{2,j}|$ for $i, j \in 1, \dots, n$). This partitioning method leaves junctions intact [50] (Fig. 6). With computer scientist Louis Petingi, a partitioning method for dual graphs has also been developed, which divides a dual graph into pseudoknot and pseudoknot-free regions [51] (Fig. 6).

3.4. Software and application overview

The main advantage of graphical representations of RNA secondary structures is that all possible motifs can be described explicitly by graph enumeration methods [48]. It also makes systematic (exhaustive) studies possible because the graph motif space is much smaller than sequence space. We exploit this aspect, for example, in sampling the conformational space of tree graphs by Monte Carlo/Simulated Annealing in hierarchical program RAG-TOP for predicting the global topologies of RNAs [49,34,52–54]. To date, RAG has been used (see reviews in [55,35,36]) to classify [33,56–59], catalog [56,57,60], and design RNA motifs [61–64]; partition RNAs into building blocks [60,63,65,51]; and to predict global RNA topologies [49,34,52,53]. See next two sections for more details.

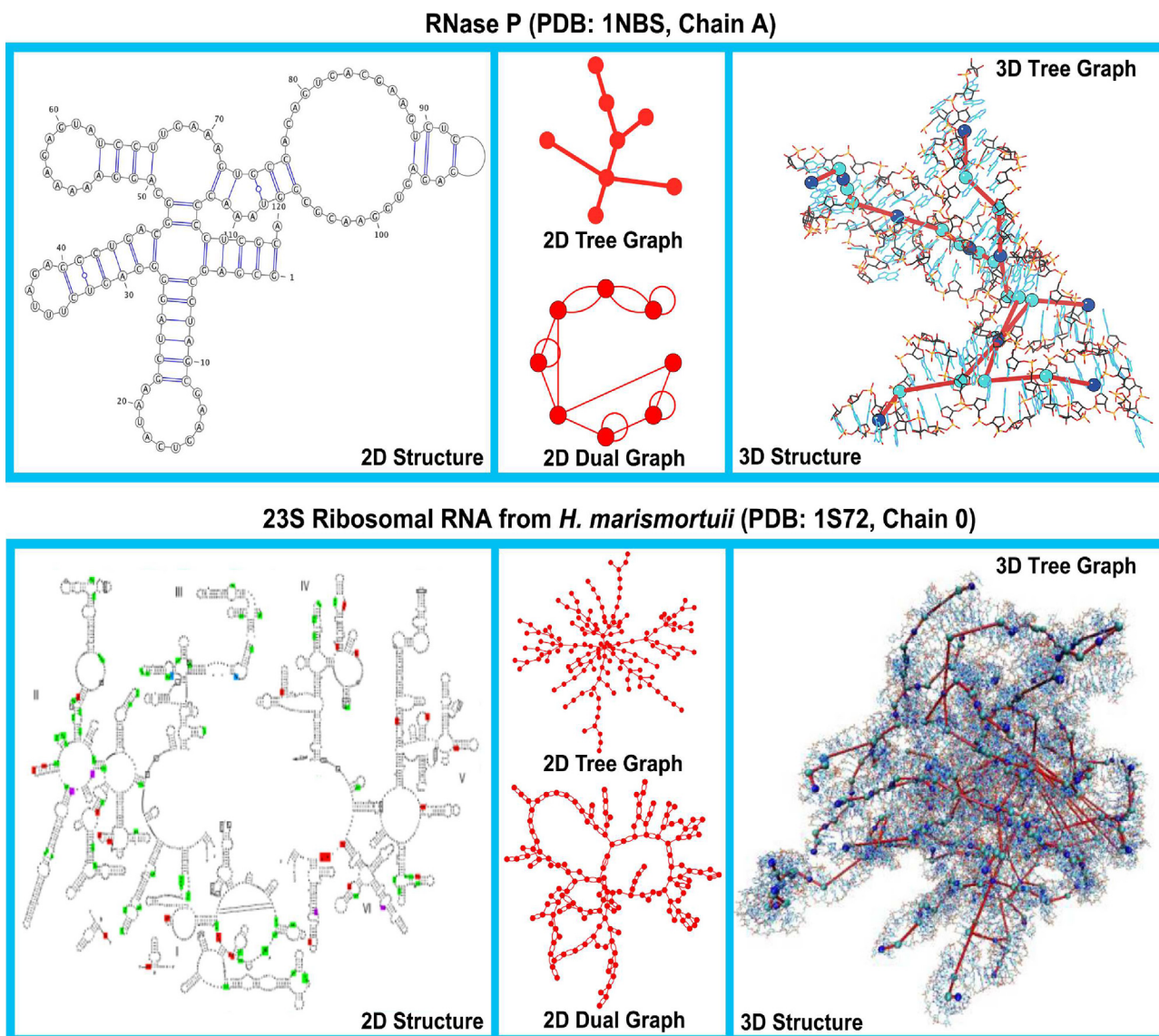


Fig. 5. Examples of RAG 2D and 3D graphs: For RNase P (top) and rRNA (bottom), shown are the experimentally determined 2D (left) and 3D (right) structures. The corresponding 2D RAG tree and dual graphs are also shown (middle), as well as the 3D tree graph superimposed on the experimental 3D structure.

Software and web servers using RAG have been made available to the community. The RAG (<http://www.biomath.nyu.edu/rag/home>) catalog is available for identifying solved RNAs with graphs and exploring the universe of RNA motifs. The RAG-3D database and server are available for RNA motif search and graph partitioning (<http://www.biomath.nyu.edu/RAG3D/>) [65], both useful for RNA design and analysis.

Our web server RAGPOOLS [61,62] utilizes RAG2D to analyze and design RNA pools with targeted topological distributions (i.e., enrichment of motifs) to help discover new synthetic RNAs by computational simulation of the *in vitro* selection procedure (see Fig. 7, link from the main RAG resource, <http://www.biomath.nyu.edu/rag/home>, and program notes).

Our programs JunctionExplorer [59,49] and CHSalign [66] (<http://nature.njit.edu/biosoft/Junction-Explorer/>) offer a data-mining tool to predict and compare junction structures. We developed a data-mining protocol to predict junction topologies using a decision tree based on “features” of the system (loop size, adenine content, free energy) [59]. Our methods predict coaxial stacking arrangements and junction family arrangements in 3- and 4-way junctions with accuracy of 75% and better.

More recently, RAGTOP exploits RNA 3D graphs for 3D structure prediction by graph sampling [49,34,52,53]. Here we utilize a hierarchical approach starting from 2D structure to 3D graphs to all-atom models using our junction prediction [59], Monte Carlo sampling, and fragment assembly to translate graph candidates into all-atom models [54,53].

The next two sections provide details on selection of novel RNA motifs and their design, and tertiary structure prediction, respectively.

4. RAG motif enumeration and design

One advantage of graphical representations of RNA secondary structures is that all possible motifs can be described explicitly by graph enumeration methods [48].

We have recently enumerated all RNA tree graphs in the RAG framework up to 13 vertices using graph enumeration techniques, and clustered them to suggest which hypothetical RNAs, or those not yet found in Nature, may be RNA-like, that is, good design candidates [56]. Thus, for example, all existing motifs are associated

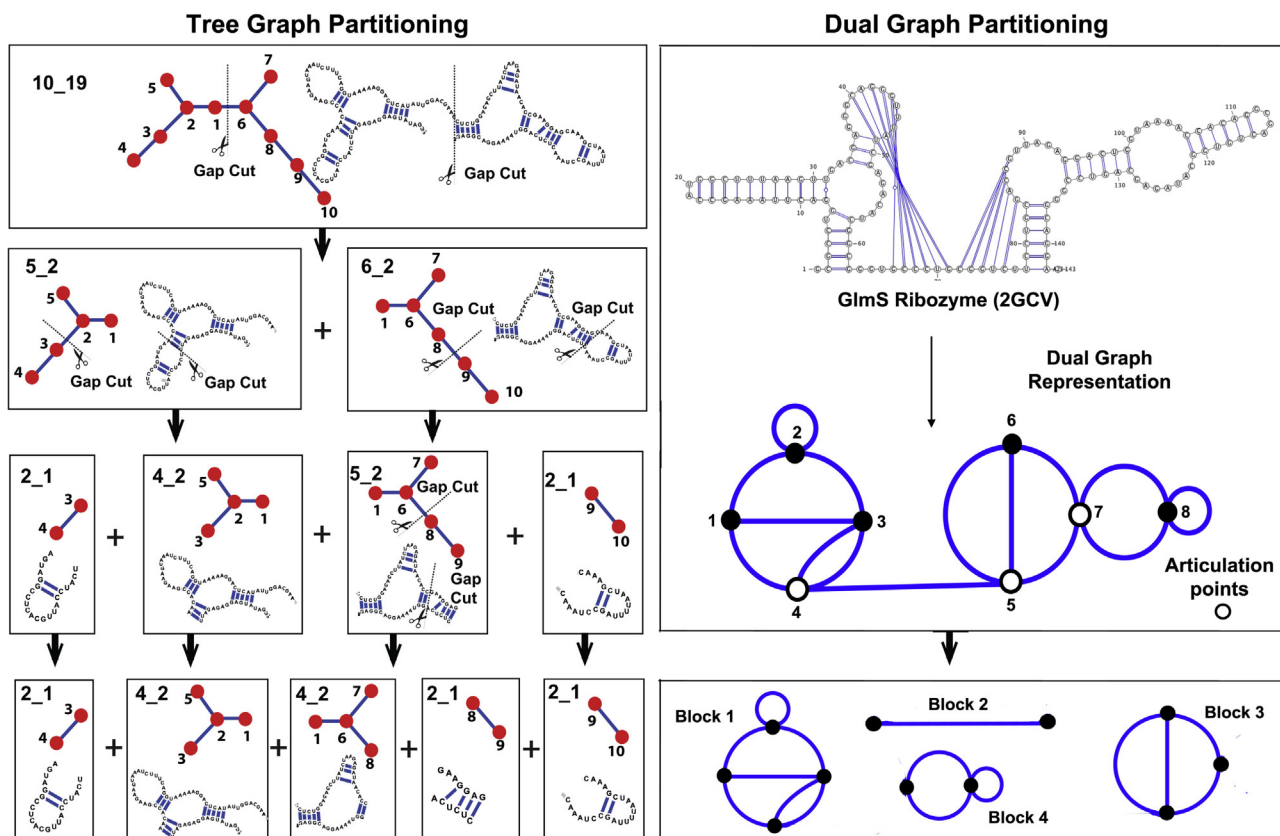


Fig. 6. RNA partitioning. (a) Partitioning of tree graphs by the gap cut method [50], and (b) partitioning of dual graphs by articulation points [51].

with a representative PDB structure in our database. Each RAG tree is associated with unique label, by vertex number and branch complexity, the latter described by λ_2 .

Thus in Fig. 8, which shows segments of our RAG motif atlas, red graphs indicate existing graphs and black graphs indicate hypothetical graphs.

4.1. Clustering into RNA-like and non-RNA-like motifs

To further classify the hypothetical motifs into RNA-like or non-RNA-like motifs, we employ graph clustering based on features of known RNAs. We use, as topological descriptors, transformations of Laplacian eigenvalues ($0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$) and vertex number (n) and a standard clustering algorithm called *partitioning around medoids* (PAM) to predict RNA-like motifs based on training set of known RNAs [56]. Essentially, PAM minimizes the sum of distances between two members within a group and maximizes the sum of distances between the two groups. The blue graphs in Fig. 8 are those determined to be RNA-like, while the black colored graphs are those considered to be non-RNA-like. Similar procedures can be used for dual graph clustering and classification [64].

Our recent enumeration and classification of hypothetical graphs into RNA-like and non-RNA-like has shown that RNA-like structures provide good candidates for RNA design, better than those classified as “non-RNA-like” [56]. Specifically, 10% of our RNA-like versus only 4% of non-RNA-like graphs have been experimentally determined since 2011.

4.2. RAG-3D for graph partitioning and motif search

How could these hypothetical RNA-like motifs be designed? That is, what sequences could be created that will fold onto these desired motifs?

We have developed tools for this purpose using our database and web server “RAG-3D” [65]. RAG-3D extends the RAG catalog to 3D graphs (<http://www.biomath.nyu.edu/RAG3D/>) and links solved PDB structures to these 3D graphs. Moreover, in response to a query RNA structure or PDB file, RAG-3D searches for similar blocks in the database and partitions any solved RNA, represented as graphs, into building blocks, or modular units.

For example, the signal recognition particle 75.S SRP (PDB ID 1LNG) corresponds to the 7_3 graph. RAG-3D can partition this RNA graph into 9 subgraphs that contain various 6_3, 6_2, 5_2, 4_2, 3_1, and 2_1 graphs (Fig. 9). Our partitioning requires that all junctions remain intact in the subgraphs. Similarly, RAG-3D’s search tool can identify and rank matching structures to query subgraph (as PDB structures and associated coordinates).

These tools immediately suggest how to design RNAs with desired novel motifs: sequences and 3D structures corresponding to the different subgraph fragments can be assembled together.

Since each subgraph has a corresponding known sequence [65], the idea is to piece them together to build the candidate atomic model. In our previous work on this concept, we used manual piecing together of building blocks to design 10 RNA-like motifs containing pseudoknots, described as dual graphs (see Fig. 10). For each design candidate, we partitioned the graph into two subgraphs, both of which have been solved. Then we

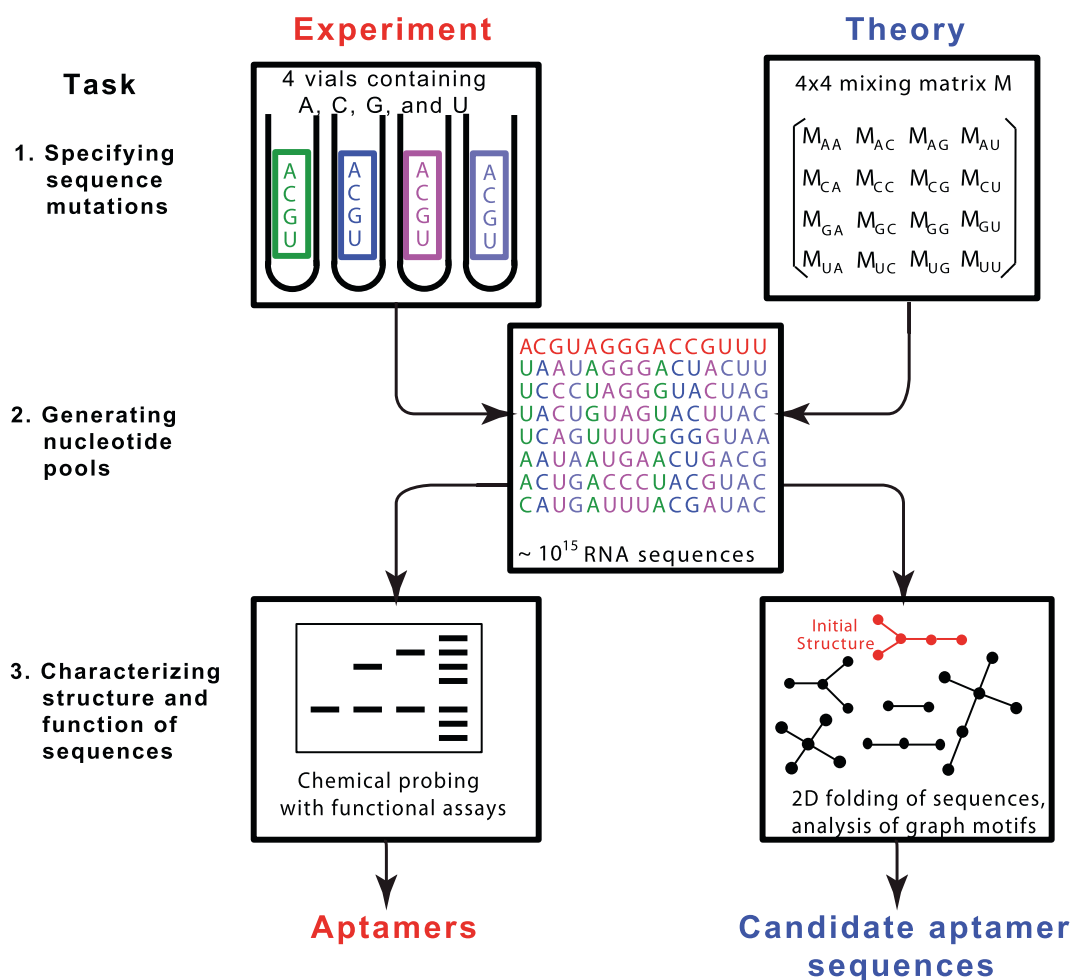


Fig. 7. The *in silico* RAGPOOLS approach to simulate the experimental *in vitro* selection process for novel RNA motifs. Using ‘mixing matrices’ [61,62] we can specify which graph motifs to blend into the selection pools to help obtain desired RNA motifs. The resulting sequences are ‘folded’ using available 2D algorithms and filtered to obtain the desired products.

pieced together the sequences for the two subgraphs, and tested the resulting combined sequence’s ability to fold onto the target fold by applying 2D prediction algorithms [63]. The design exercise proved its utility: Since 2014, at least half of the candidate models were solved experimentally, as shown in the last column of Fig. 10, with significant correlations in the designed sequences [35].

4.3. Automated fragment assembly for design

Very recently, we have developed an automated “fragment assembly” protocol to piece these subgraphs (see Fig. 11) [54].

The fragment assembly approach, based on empirical potentials derived from available experimental structures, has proven successful for predicting tertiary structures of proteins from sequence using the Rosetta program [67]. Significantly, in recent studies even side chains can be predicted in atomic resolution. As evident from the superb performance in the RNA-Puzzles initiative [68], Rosetta’s adaptation for RNA by Das and co-workers has benefited from the inclusion of chemical mapping information tailored for RNA [69].

Starting from the RNA-like graph motif, we first extract 2D sub-motifs of existing graph IDs. Second, we extract all-atom fragments catalogued with the same graph ID as the existing graph ID from the RAG-3D database (see Fig. 11). We select the fragments that

have the required number of internal loops and hairpins. Third, the two fragments are assembled by overlapping the base pairs that flank the common loop between the two fragments. Fourth, the 3D tree graph is constructed from the final all-atom model and scored based on our RAGTOP statistical potential (described below) [53].

To illustrate its application, we summarize in Fig. 12 results for 6 candidate RNA-like motifs. For each target RNA-like motif, the top 400 (top 200 each from two different design runs) scoring candidate models/sequences (after removing models with chain breaks) were subjected to *in silico* folding by programs RNAfold [24] and NUPACK [25]. The yield in the figure indicates the number of sequences out of the top 400 that fold onto the desired fold as determined by both RNAfold and NUPACK. Preferences for the desired fold are indicated by denoting the fragments that produce the highest number of sequences with the desired fold. The number in the parenthesis indicates the number of sequences that fold onto the desired fold with that fragment.

Clearly, this design protocol provides a large number of candidates for further analysis. These preliminary results indicate promise for our design protocol, as recently described for six target sequences, including preliminary experimental testing [64]. Though the main ingredients are now in place, further steps need to be developed to optimize and select the most promising candidates and subject them to experimental testing.

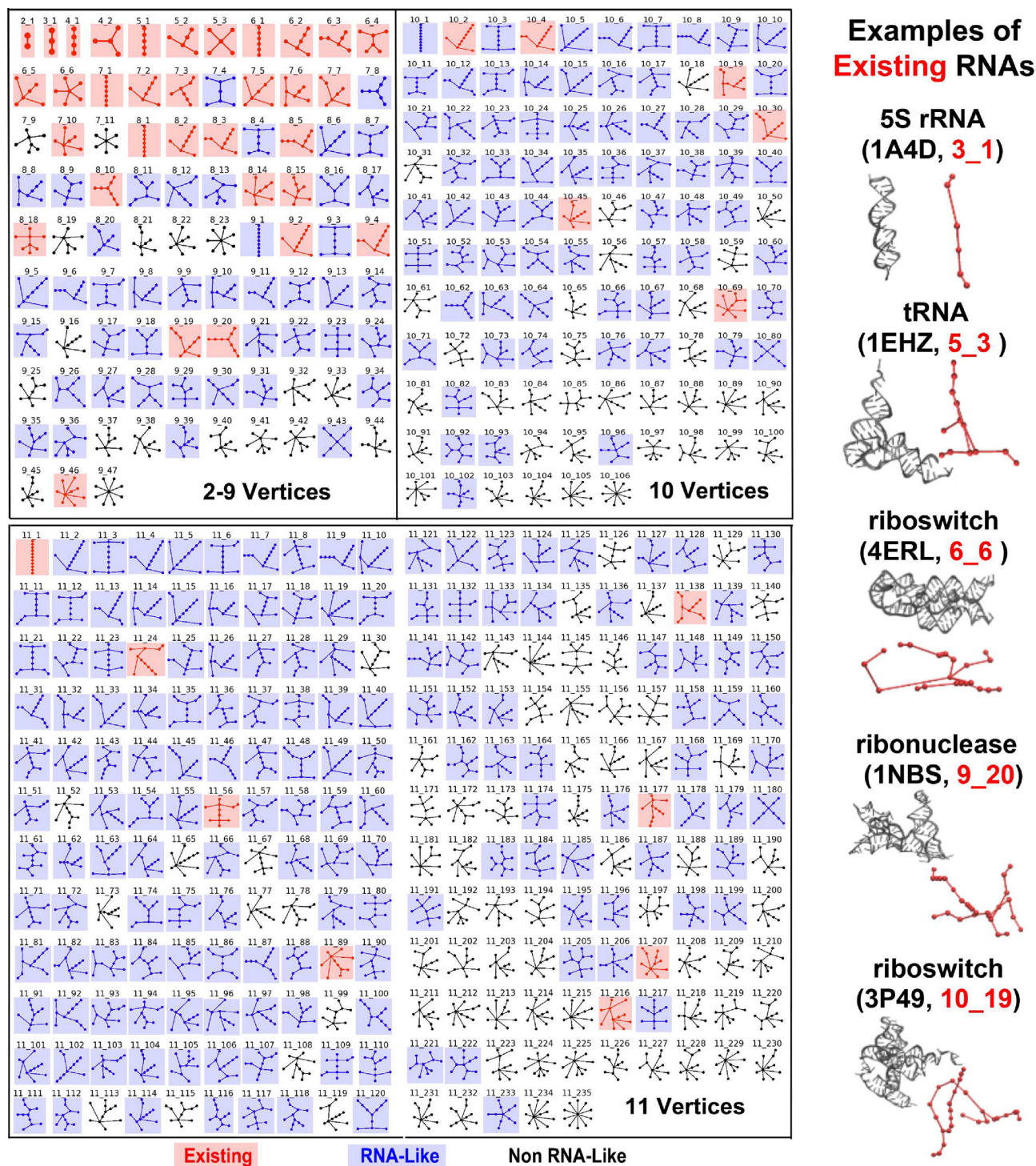


Fig. 8. Segments of RAG's motif atlas, with classifications into existing, RNA-like and non-RNA-like motifs as determined by clustering [56].

5. RAGTOP for 3D structure prediction

5.1. RNA structure prediction difficulties

RNA structure prediction is challenged by a combination of factors, including: limited structural information (compared to proteins), enormous structural repertoire, RNA's high flexibility and structure dependence on bound ions and proteins, and the difficulty of predicting global interactions a priori.

An RNA building block has 7 degrees of freedom corresponding to the nucleic-acid backbone torsions plus 2 degrees of freedom for the base torsions. In molecular dynamics simulations, the sensitive behavior of RNA base-pair stability and RNA geometry to torsional angle parameters has been noted, and structural distortions are common [40]. Another difficulty involves protein-binding effects. These can induce significant distortions, such as helical bending of loops. Together with the multiple conformations that RNAs can adopt, for example in riboswitches, these combined problems

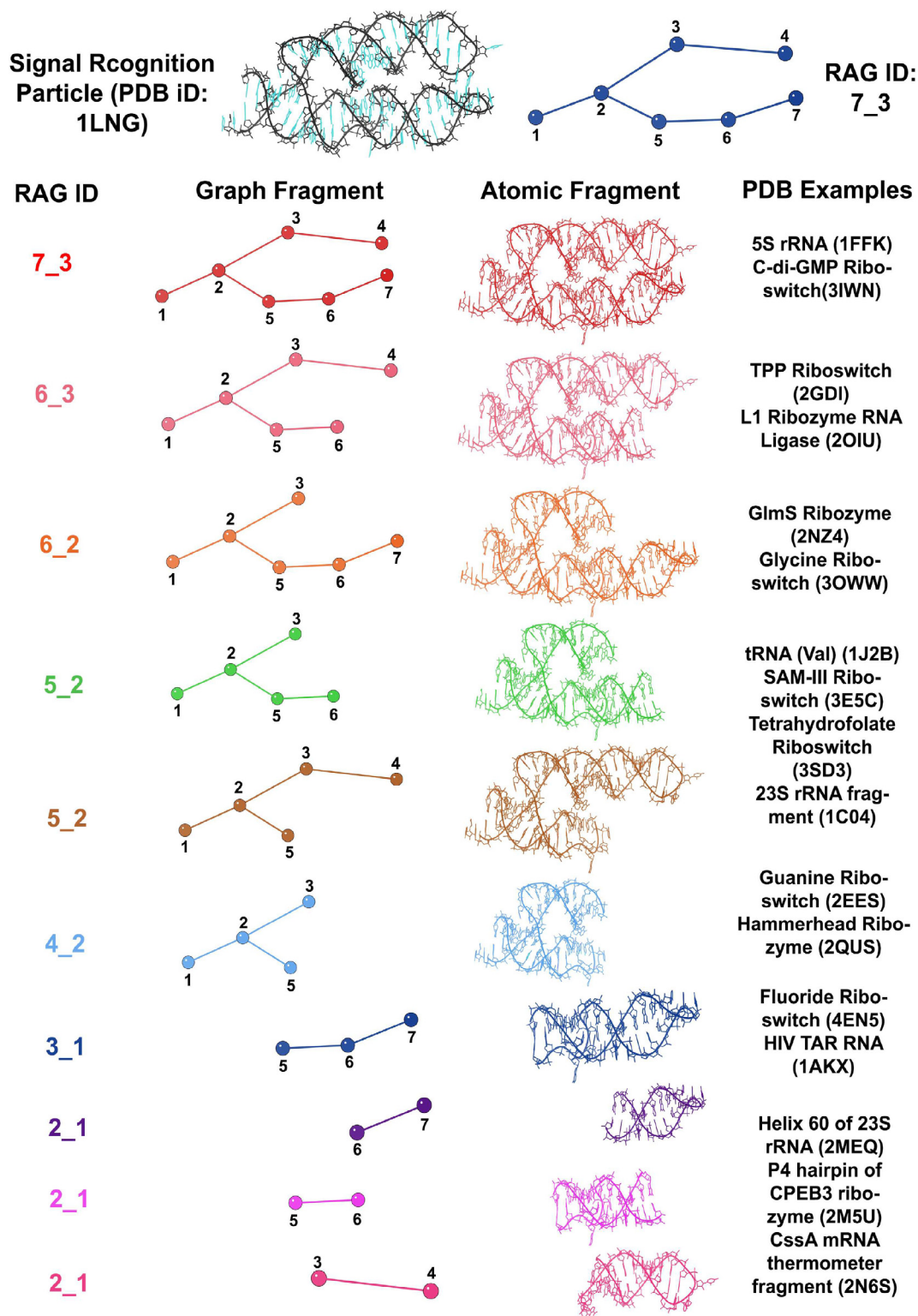


Fig. 9. RAG partitioning of the signal recognition particle into subgraphs.

challenge the modeling and structure prediction of RNA structure and its interactions.

Though lagging behind protein prediction, RNA 3D structure prediction has improved in recent years due to a combination of experimental and computational advances. Programs for predicting RNA 2D structures, though imperfect, are excellent starting points from a given nucleotide sequence [26]. But predicting how

these 2D elements fold in 3D is challenging. As we have recently found in our comparative assessment of available 3D folding programs for RNA [70,55], automatic 3D prediction methods remains difficult for large RNAs. In particular, breakthroughs are needed to arrange helical elements and determine long-range contacts.

Current approaches for RNA 3D structure prediction include those that have worked for proteins, including Rosetta-based sam-

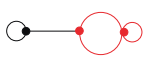
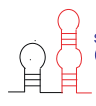
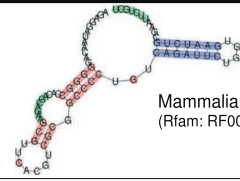

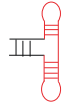
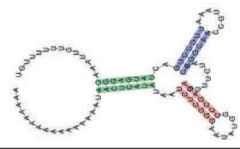
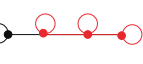

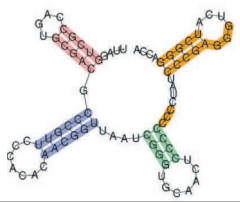


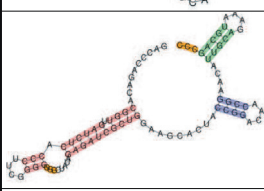
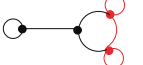
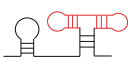


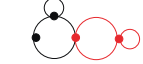
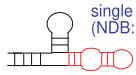
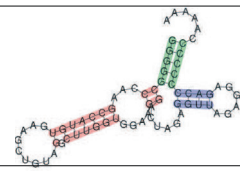

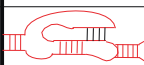
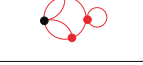
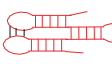


Graph Representation With Natural Submotif	RNA Secondary Structure With Natural Submotif	Candidates Discovered After 2004
C1 	 single strand RNA (NDB:PR0055)	 Mammalian CPEB3 Ribozyme (Rfam: RF00622)
C2 	 bulged hairpin (Rfam:CopA)	 Purine Riboswitch (Rfam: RF00167)
C3 	 DsrA RNA (Rfam:DsrA)	 Tymovirus tRNA-like 3' UTR element (Rfam: RF00233)
C4 	 bulged hairpin (Rfam:CopA) single strand RNA (NDB:PR0055)	 Tombuvirus 3' UTR region IV (Rfam: RF00176)
C5 	 bulged hairpin (Rfam:CopA)	
C6 	 DsrA RNA (Rfam:DsrA)	
C7 	 single strand RNA (NDB:PR0055)	 Flavivirus DB element (Rfam:RF00525)
C8 	 single strand RNA (NDB:PR0037)	
C9 	 DsrA RNA (Rfam:DsrA)	
C10 		

Fig. 10. Ten designed dual graphs [63].

pling, which assembles short fragments from existing RNA structures, like FARNA [71–73]; comparative modeling approaches based on RNA homology [74]; and coarse-grained approaches like MC-Sym [42], NAST [75,43], or others (e.g., [44,45]).

To stimulate advances in RNA modeling, a new exercise was founded by Eric Westhof called *RNA-puzzles*. Already, *RNA-puzzles* has demonstrated a fundamental difficulty in our ability to handle long-range RNA interactions, which are very difficult to predict *a priori* [76,68].

5.2. RAGTOP protocol

The assembly of global interactions is the focus of our recently introduced hierarchical RNA graph sampling approach for topology prediction called RAGTOP [49,34,52] (see sketch in Fig. 13). RAG exploits coarse-grained RNA graphs for efficient sampling of the associated conformational space.

Junction family prediction. Our program RAGTOP begins with a 2D RNA structure, which can be provided experimentally or pre-

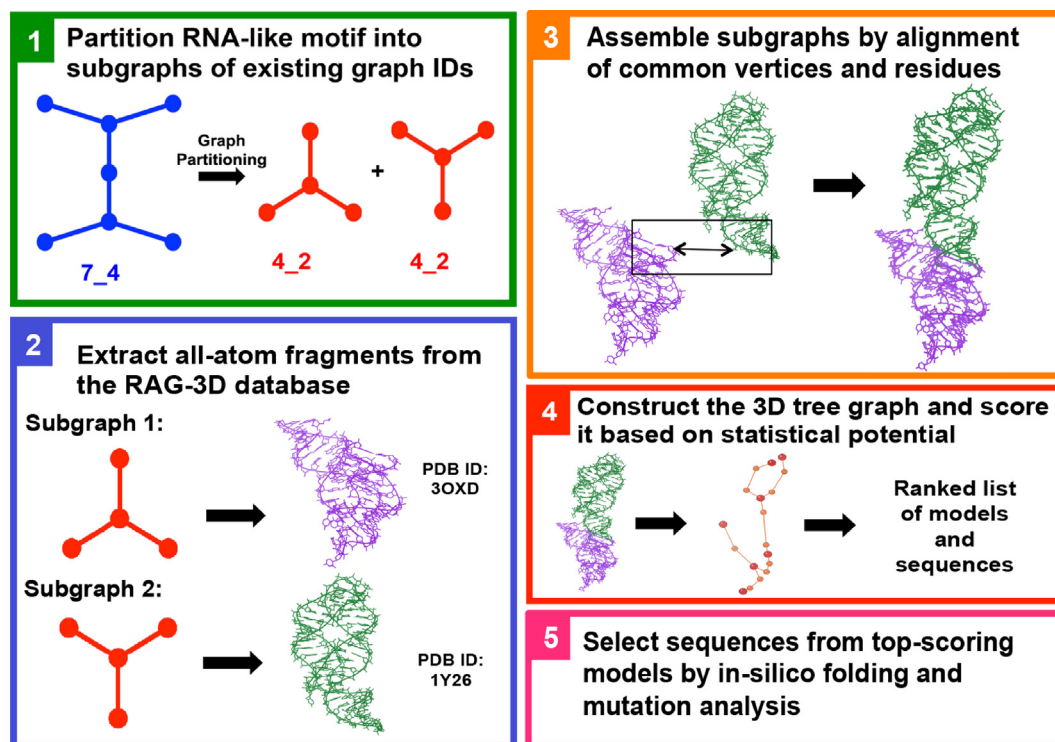


Fig. 11. Sketch of the fragment assembly approach.

dicted using various *in silico* programs. To provide a good initial RNA graph for sampling, we apply our bioinformatics/data-mining tool called JunctionExplorer to predict initial junction arrangements [59].

This is made possible by analyzing and classifying all existing junctions. Building upon Westhof's initial classification of 3-way junctions [77], we have analyzed 207 RNA junctions in the NDB that ranged from 3-way to 10-way junctions [78] (see Fig. 13 later, quadrant 1).

Then, our random forest data-mining protocol [59] predicts junction topologies using a decision tree based on "features" or vectors which we define and train on known RNAs. These feature vectors involve loop size, adenine content, and free energy of stacking criteria. These features work because small loop size enhances the probability of coaxial stacking; adenines in loops tend to form A-minor motifs; and free-energy parameters account for stacking forces between base pairs at the end of helices. The trained random-forest decision tree protocol [59,49] is then applied to each initial 2D graph. The resulting 2D graph is then extended into 3D directed graphs to incorporate sequence lengths of the loops and junctions.

MC/SA with statistical scoring function. This 3D graph is then subjected to a Metropolis Monte Carlo (MC)/Simulated Annealing (SA) sampling procedure guided by a statistical scoring function. The Metropolis algorithm generates a series of conformations to sample the *canonical* ensemble of a system (constant number of particles, temperature T , and volume) so that the sequence of states tends to a lower energy region (see details in [79]). Essentially, each state depends only on the prior state, and these are related to one another by a specified perturbation. A new state X' that leads to lower energy than the prior state X is always accepted, while a state that increases the energy (i.e., $\Delta E = X' - X > 0$) is accepted with probability $p = \exp(-\beta\Delta E)$, where $\beta = (1/k_B T)$ and k_B is Boltzmann's constant. In practice, this probability is achieved by comparing p to a uniformly-generated number ran between

zero and 1. If $p > ran$, we accept X' as the new configuration in the sequence; otherwise, we generate another trial configuration and repeat the process. In our context, we perturb each graph by small changes in bend and twist angles. The SA component helps accelerate convergence further by decreasing the effective temperature T with the MC iteration with a well chosen function that works in practice [34,53].

Such an MC procedure can be applied to any system where the potential energy or score of each configuration can be evaluated, for example by empirical functions using a standard all-atom or coarse-grained force field (see [79]), or by a statistical scoring potential. The latter involves collecting a set of solved structures, analyzing their properties of interest as a function of internal geometric variables, and then deriving probability distribution functions that describe the likelihood of any given state.

In our case, we have collected a database of high-resolution non-redundant RNA structures, and divide all the loops in the database into families L, R , where $L \leq R$ describe the number of bases in each single-stranded region of the loop (see Fig. 13, quadrant 2). Many L/R families are possible, but based on observed RNA structures, we use 27 families $L \leq R$: $0/1, 0/2, 0/6+, 1/1, \dots, 6/6+$, where the plus sign indicates that the number of nucleotides can be equal or greater than specified. We have found that the bending and twisting angles about these loops depend on L and R .

We analyze such information at a certain bin size (e.g., 36 bins, with bin size of 10 degrees), and smooth the probability distribution if needed. Then the probability that a certain bending angle θ (or torsion angle τ) exists is simply the number of loop entries in that bin divided by the total number of loops in the sample: $\Pr(\theta) = N_i/N$ where N_i is the number of internal loops with θ in that bin, and N is the total internal loops in the dataset. This structure-derived probability $\Pr(\theta)$ can be compared to the probability that an angle will be in that bin, which is, for example, $P_{\text{random}} = 1/36$ for 10-degree bins. Applying Boltzmann statistics

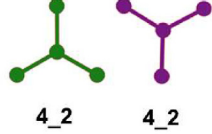
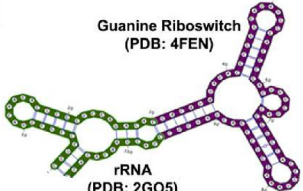

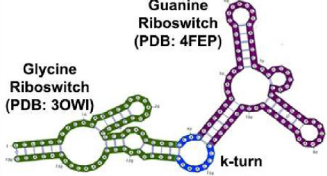
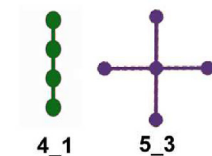
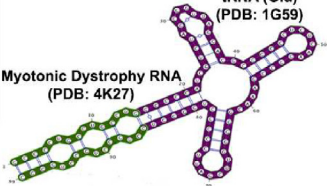
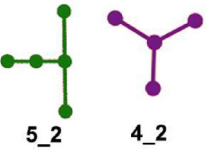
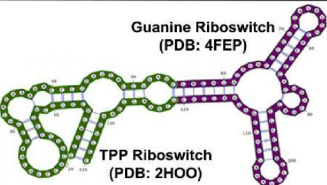
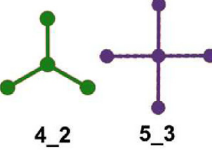
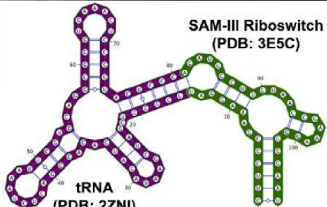
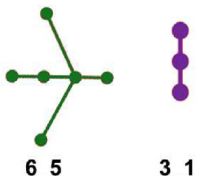
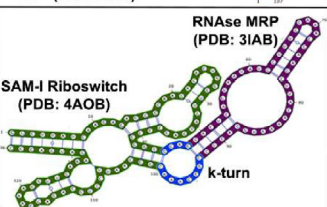
RNA-Like Motif	Subgraphs	Designed Sequence	Yield	Preference for Desired fold
7_4	 4_2 4_2	 Guanine Riboswitch (PDB: 4FEN) rRNA (PDB: 2GO5)	43	One fragment of ribosomal RNA (4_2) (23)
8_4	 5_2 4_2	 Guanine Riboswitch (PDB: 4FEP) Glycine Riboswitch (PDB: 3OWI) k-turn	61	One fragment of the Glycine Riboswitch (5_2) (52)
8_6	 4_1 5_3	 tRNA (Glu) (PDB: 1G59) Myotonic Dystrophy RNA (PDB: 4K27)	50	One fragment of the transfer RNA (5_3) (50)
8_7	 5_2 4_2	 Guanine Riboswitch (PDB: 4FEP) TPP Riboswitch (PDB: 2HOO)	11	One fragment of the thi-box Riboswitch (5_2) (10)
8_9	 4_2 5_3	 SAM-III Riboswitch (PDB: 3E5C) tRNA (PDB: 2ZNI)	32	One fragment of the transfer RNA (5_3) (32)
8_12	 6_5 3_1	 RNase MRP (PDB: 3IAB) SAM-I Riboswitch (PDB: 4AOB) k-turn	7	One fragment of the SAM-I Riboswitch (6_5) (7)

Fig. 12. Illustration of results from design of six RNA-like motifs by fragment assembly [64]. For each RNA-like motif, shown are the two fragments we piece together based on known RNAs, the yield of the intended motifs as determined by two *in silico* programs, and the trends we identified for obtaining that fold.

to internal loop angles, we derive the free-energy score for internal loop angles to be:

$$\Delta G(\theta) = -k_B T \ln(\text{Pr}(\theta)/P_{\text{random}}).$$

This is the free energy term for bending that is used in our RAGTOP MC/SA process.

An identical form is used for the torsion angles (τ) for the same dataset and bins, resulting in our statistical scoring potential:

$$\Delta G_{\text{internal}} = \sum_i \Delta G(\theta_i) + \Delta G(\tau_i).$$

In addition to these two internal potential terms for bending and twisting terms about internal loops, our scoring function ΔG also includes a radius of gyration term for the whole RNA (ΔG_{R_g}), and an optional pseudoknot term to restrain the 3D structure to a length corresponding to a pseudoknot edge (ΔG_{pk}). The R_g term

helps drive overall compactness of the RNA, and the pseudoknot term helps maintain in 3D a pseudoknot interaction known from the 2D structure.

Our analysis has shown that radii of gyration of RNA 3D graphs increase logarithmically with the RNA sequence length L and decrease logarithmically with the vertex number V :

$$R_g(L, V) = C_1 \ln(L) + C_2 \ln(V) + C_3.$$

Here, C_1 , C_2 , and C_3 are constants fitted to the RNA data to define the target R_g , namely \bar{R}_g , for a given RNA graph with length L and V vertices.

Then, we define the resulting potential term that drives the system with corresponding R_g to the target value \bar{R}_g as:

$$\Delta G_{R_g} = |R_g - \bar{R}_g|.$$

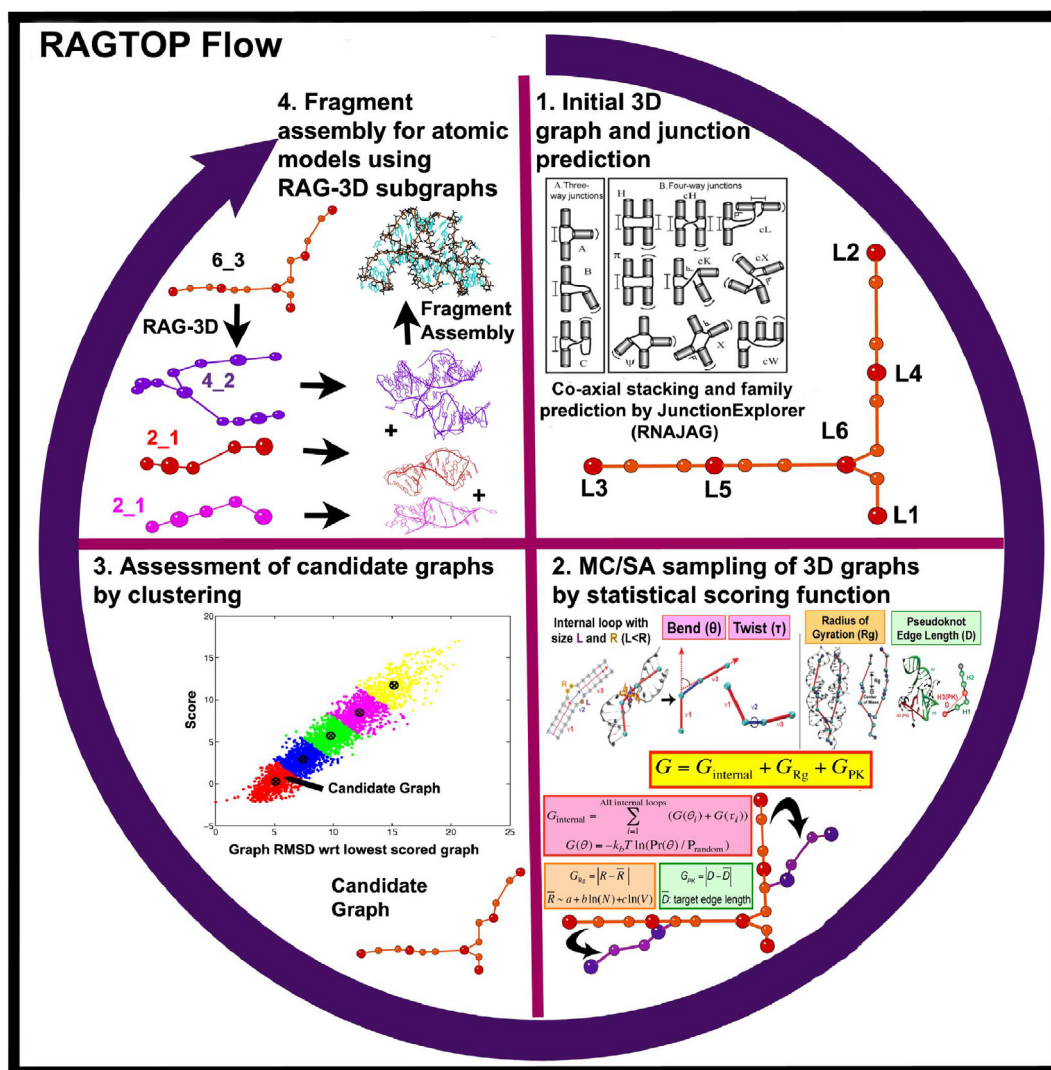


Fig. 13. Sketch of the RAGTOP hierarchical sampling approach [49,34,50,53]. 1. Initial junction topology prediction; 2. MC sampling of 3D graphs scored by a statistical scoring function with components for bend, twist, radius-of-gyration, and pseudoknot terms; 3. clustering of generated graphs to identify candidate graph; and 4. determination of atomic models from the candidate graph by our fragment assembly algorithm using RAG-3D subgraph partitioning.

Optionally, to model pseudoknots, we add a similar term for the pseudoknot edge, restraining it to some reasonable value, \bar{D} .

A recent addition involves dividing the bending term further into k-turn-like bends and non-k-turn-like bends [53]. Such a division is reasonable because in the long term we would like to define the geometry of an RNA tertiary structure in terms of its constituent sequence/structure motifs. Our tailored kink-turn motif term recognizes these motifs by a consensus sequence [53]. Kink-turns are a widely distributed motif in large RNA/protein assemblies like the ribosome and in other RNAs like riboswitches, snoRNAs, and more [80,81]. They direct and sculpt the trajectory of helical segments within the RNA by forming unusually large angles of around 50° between the two axes of the helical arms that are stabilized by two cross-strand A-minor interactions. Thus, the sequence signature of a standard k-turn involves a duplex RNA with a short bulge followed by G-A and A-G base pairs. The associated A-minor motifs involve nearby residues.

The Monte Carlo moves are implemented by local rotations around loops, while keeping the predicted junction family intact. We have considered various MC protocols, in which angular displacements are randomly sampled from the full 2π range, or alternatively restricted, so that the range is lowered with the MC

iteration, to help convergence. We have found both to work and have recently employed a random move protocol combined with Simulated Annealing to guide convergence [53,54]. The 3D predictions by RAGTOP depend on the quality of the initial 2D structure. The better the 2D structure is (from experiment or from computational predictions), the better the final result. In general, our results are competitive with other approaches and work especially well for RNAs with junctions [34,52–54]. This is because our junction assembly component helps bring essential tertiary elements in space in the right orientation; the tailored kink-turn potential for loops helps obtain reasonable bend orientations [53]; the pseudoknot-like term guides the structure assembly of RNAs with pseudoknots [52], and automated fragment assembly generates reasonable atomic models, especially for RNAs with junctions [54]. See Fig. 14 for examples of RAGTOP results from graph models to full atomic models [54].

Naturally, many improvements can be envisioned, concerning the statistical potential and the MC moves. For example, other sequence-dependent features could be incorporated into the MC scoring function, like the k-turn bend potential, and different stems within junctions should be moved with respect to one another to model RNA's natural flexibility. The handling of pseudoknots could

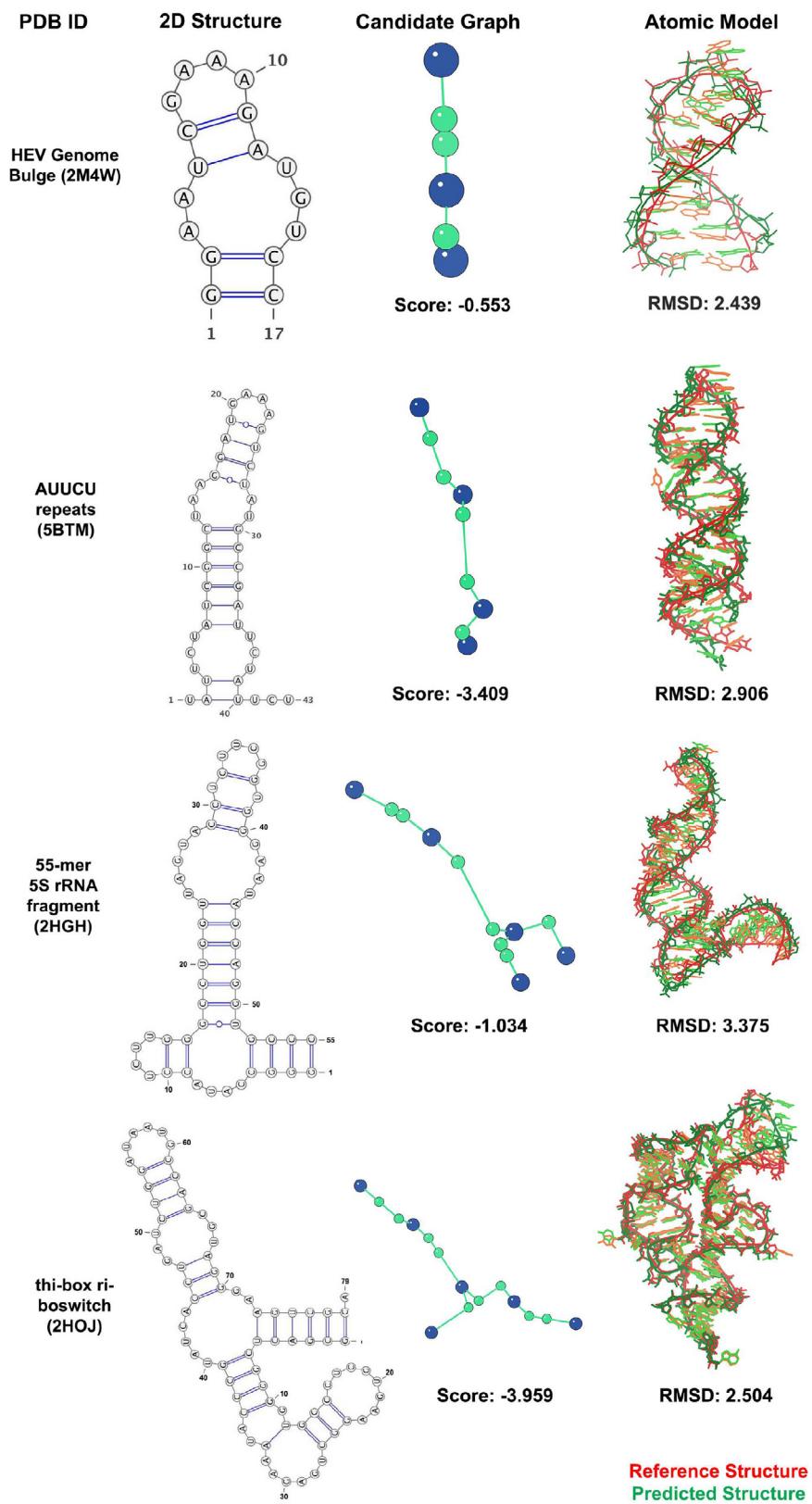


Fig. 14. Examples of RAGTOP results from graph models to full atomic models.

also be made more rigorous using dual graphs, but the latter requires new machinery for graph scoring and sampling altogether.

6. Conclusions

6.1. Biomolecular simulations as a field on its own right

Impressive developments in both computational technology and simulation algorithms have led to rapid advances in the biomolecular modeling field [82]. These enhancements have rendered computational biology a valuable field on its own right rather than a mere accessory to experimental structure determination and analysis [83].

Because of emerging discoveries of RNA's structural and functional versatility, including exciting RNA-based CRISPR applications in biology, engineering, and medicine, RNA is rising to a superstar status. The related problems and applications offer many incentives and opportunities for RNA research, including computational approaches. *In silico* strategies, in particular, have the potential to offer systematic solutions to challenging structure prediction and design problems [84,85].

The typical atomic-level modeling of biomolecular with standard force fields [79] can address many problems, but the computational complexity of nucleotide/nucleotide contacts increases quadratically with the sequence size. The difficulties in modeling floppy RNAs with solvent and salt are also well appreciated [40]. Thus, as emphasized in this chapter, reduced representations, such as offered by coarse graining using graphs to model RNA 2D and 3D structures, can help define and solve problems with a fresh perspective. The successful applications described with the RAG approach involve structure annotation, motif enumeration and classification, RNA partitioning, structure prediction, and RNA design. All these aspects exploit the power of the graph representations to enumerate RNA motifs; employ associated linear algebra tools for graph partitioning, clustering, and classification; sample RNA configuration space efficiently to predict tertiary topologies of RNAs from the 2D structure; and combine all these tools to design novel RNA motifs. Of course, a combination of methods and approaches could always be fruitful, including the incorporation of experimental data (such as chemical reactivity) in the models [85], and iterative experimental testing and modeling of the designed RNA. It is particularly exciting that our recent *in silico* design predictions were confirmed by experimental testing by chemical reactivity data [64].

6.2. Future challenges with RAG

Ongoing work with the RAG approach involves applying the dual-graph partitioning algorithm [51] to the full library of dual graphs representing RNAs up to 9 vertices (98 different graph topologies for 1785 RNA secondary structures) [86]. Using the Hopcroft and Tarjan algorithm for identifying non-separable graph components in a connected graph, our algorithm uses the adjacency matrix of the graph as input and determines articulation points. Such vertex points v are defined if $G - v$ (where G is the dual graph of the RNA secondary structure) lead to a disconnected graph. In this way, we define all subgraphs, or building blocks, of the dual graph library that contain no such articulation points (non-separable units). Interestingly, preliminary work suggests that several dual subgraphs are common to the entire set of dual graphs. Work is continuing to analyze the biological features of these subgraphs.

Another interesting application involves evolutionary analysis of organism complexity using graphs. Analysis of ribosome struc-

tures from different species, from archaea to human, has been used by Williams and coworkers to detect evolutionary relationships [87]. Ribosomal RNAs catalyze the synthesis of proteins essential to life. Ribosomal RNA structures, both secondary and tertiary, are thus highly conserved, and a common structural/functional core has been identified. This core evolves in complexity as the organism's complexity increases from archaea to human. However, detecting such relationships using detailed 2D networks is far from easy. It is possible that graph analysis using partitioning of both the small and large ribosomal RNA subunits for several organisms will help identify core structural features as well as branching patterns with evolution. These subgraph patterns could then be connected to other evolutionary techniques to make biological inferences.

Exciting applications and extensions of graphs to many areas of biology can be envisioned in the near future. Mathematical and computer scientists, in particular, may find intriguing opportunities to contribute to the exciting field of RNA structure and design. Together, the experimental and computational communities will undoubtedly continue to work together to analyze, interpret, predict, and design RNA molecules and their complexes and to pursue important biomedical and engineering applications.

Acknowledgment

Funding from NIH NIGMS R01GM100469 and R35GM122562 Awards to T. Schlick is gratefully acknowledged. The author is grateful to Swati Jain for preparing many of the figures.

References

- [1] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.* 16 (7) (2001).
- [2] D.L. Applegate, R.E. Bixby, V.V. Chvátal, W.J. Cook, *The Traveling Salesman Problem*, Princeton University Press, Princeton, NJ, 2007.
- [3] R.L. Karg, J.L. Thompson, A heuristic approach to solving traveling salesman problems, *Manage. Sci.* 10 (2) (1964) 225–248.
- [4] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S.H. Strogatz, C. Ratti, Quantifying the benefits of vehicle pooling with shareability networks, *Proc. Natl. Acad. Sci. USA* 111 (37) (2014) 13290–13294.
- [5] S.R. Eddy, Non-coding RNA genes and the modern RNA world, *Nat. Rev. Genet.* 2 (2001) 919–929.
- [6] P.D. Zamore, B. Haley, Ribo-gnome: the big world of small RNAs, *Science* 309 (2005) 1519–1524.
- [7] M. Esteller, Non-coding RNAs in human disease, *Nat. Rev. Genet.* 12 (2011) 861–874.
- [8] T. Huang, A. Alvarez, B. Hu, S.-Y. Cheng, Non-coding RNAs in cancer and stem cells, *Chin. J. Cancer* 32 (2013) 582–593.
- [9] S.W. Cheetham, F. Gruhl, J.S. Mattick, M.E. Dinger, Long noncoding RNAs and the genetics of cancer, *Brit. J. Cancer* 108 (2013) 2419–2425.
- [10] M.F. Mehler, J.S. Mattick, Non-coding RNAs in the nervous system, *J. Physiol.* 572 (2006) 333–3411.
- [11] S.B. Baylin, P.A. Jones, A decade of exploring the cancer epigenome – biological and translational implications, *Nat. Rev. Cancer* 11 (2011) 726–734.
- [12] D.S. Wilson, J.W. Szostak, *In Vitro* selection of functional nucleic acids, *Ann. Rev. Biochem.* 68 (1999) 611–647.
- [13] G.F. Joyce, *In vitro* evolution of nucleic acids, *Curr. Opin. Struct. Biol.* 4 (1994) 331–336.
- [14] P. Guo, The emerging field of RNA nanotechnology, *Nature Nanotechnol.* 5 (2010) 833–842.
- [15] P. Ceres, J.J. Trausch, R.T. Batey, Engineering modular 'ON' RNA switches using biological components, *Nucl. Acids Res.* 41 (2014) 10449–10461.
- [16] A.B. Kennedy, J.V. Vowles, L. d'Espaux, C.D. Smolke, Protein-responsive ribozyme switches in eukaryotic cells, *Nucl. Acids Res.* 29 (2014) 12306–12321.
- [17] S.K. Burley et al., Structural genomics: beyond the human genome project, *Nature Genet.* 23 (1999) 151–157.
- [18] J. Chandonia, S. Brenner, The impact of structural genomics: expectations and outcomes, *Science* 311 (2006) 347–351.
- [19] P. Brion, E. Westhof, Hierarchy and dynamics of RNA folding, *Ann. Rev. Biophys. Biomol. Struct.* 26 (1997) 113–137.
- [20] I. Tinoco Jr., C. Bustamante, How RNA folds, *J. Mol. Biol.* 293 (1999) 271–281.
- [21] H.T. Lee, D. Kilburn, R. Behrouzi, R.M. Briber, S.A. Woodson, Molecular crowding overcomes the destabilizing effects of mutations in a bacterial ribozyme, *Nucl. Acids Res.* 43 (2015) 1170–1176.
- [22] S.C. Abeyirigunawardena, S.A. Woodson, Differential effects of ribosomal proteins and Mg²⁺ ions on a conformational switch during 30S ribosome 5'-domain assembly, *RNA* 21 (2015) 1859–1865.

- [23] I. Tinoco Jr., O.C. Uhlenbeck, M.D. Levine, Estimation of secondary structure in ribonucleic acid, *Nature* 230 (1971) 362–367.
- [24] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, I.L. Hofacker, ViennaRNA package 2.0., *Alg. Mol. Biol.* 6 (1) (2011) 26.
- [25] R.M. Dirks, J.S. Bois, J.M. Schaeffer, E. Winfree, N.A. Pierce, Thermodynamic analysis of interacting nucleic acid strands, *SIAM Rev.* 49 (1) (2007) 65–88.
- [26] P.H. Higgs, RNA secondary structure: physical and computational aspects, *Quart. Rev. Biophys.* 33 (2001) 199–253.
- [27] R. Nussinov, G. Pieczenik, J.G. Griggs, D.J. Kleitman, Algorithms for loop matchings, *SIAM J. App. Math.* 35 (1) (1978) 68–82.
- [28] M.S. Waterman, T.F. Smith, RNA secondary structure: a complete mathematical analysis, *Math. Biosci.* 42 (1978) 257–266.
- [29] P. Hogeweg, B. Hesper, Energy directed folding of RNA sequences, *Nucl. Acids Res.* 12 (1) (1984) 67–74.
- [30] S.-Y. Le, J. Owens, R. Nussinov, J.-H. Chen, B. Shapiro, J.V. Maizel, RNA secondary structures: comparison and determination of frequently recurring substructures by consensus, *Comput. Appl. Biosci.: CABIOS* 5 (3) (1989) 205–210.
- [31] B.A. Shapiro, An algorithm for comparing multiple RNA secondary structures, *Comput. Appl. Biosci.: CABIOS* 4 (3) (1988) 387–393.
- [32] B.A. Shapiro, K. Zhang, Comparing RNA secondary structures using tree comparisons, *Comput. Appl. Biosci.: CABIOS* 6 (4) (1990) 309–318.
- [33] H.H. Gan, D. Fera, J. Zorn, M. Tang, N. Shiffeldrim, U. Laserson, N. Kim, T. Schlick, RAG: RNA-As-Graphs database – concepts, analysis, and features, *Bioinformatics* 20 (2004) 1285–1291.
- [34] N. Kim, C. Laing, S. Elmetwaly, S. Jung, J. Curuksu, T. Schlick, Graph-based sampling approach for approximating global helical topologies of RNA, *Proc. Natl. Acad. Sci. USA* 111 (2014) 4079–4084.
- [35] N. Kim, N. Fuhr, T. Schlick, Graph applications to RNA structure and function, in: R. Russell (Ed.), *Biophysics of RNA Folding, Biophysics for the Life Sciences*, chapter 3, Springer-Verlag, New York, 2013, pp. 23–51.
- [36] N. Kim, L. Petingi, T. Schlick, Network theory tools for RNA modeling, *WSEAS Trans. Math.* 12 (2013) 941–955.
- [37] P.C. Whitford, S.C. Blanchard, J.H.D. Cate, K.Y. Sanbonmatsu, Connecting the kinetics and energy landscape of tRNA translocation on the ribosome, *PLoS Comp. Biol.* 9 (2013). e1003003.
- [38] C.S. Gaines, D.M. York, Ribozyme catalysis with a twist: active states of twister ribozyme in solution predicted from molecular simulation, *J. Amer. Chem. Soc.* 138 (2016) 3058–3065.
- [39] J. Šponer, P. Banáš, P. Jurečka, M. Zgarbová, P. Kührová, M. Havrila, M. Krepl, P. Stadlbauer, M. Otyepka, Molecular dynamics simulations of nucleic acids, from tetranucleotides to the ribosome, *Chem. Phys. Lett.* 5 (2014) 1771–1782.
- [40] P. Kührová, R.B. Best, S. Bottaro, G. Bussi, J. Šponer, M. Otyepka, P. Banáš, Computer folding of RNA tetraloops: identification of key force field deficiencies, *J. Chem. Theor. Comput.* 12 (2016) 4534–4548.
- [41] S.-Y. Le, R. Nussinov, J. Maizel, Tree graphs of RNA secondary structures and their comparisons, *Comput. Biomed. Res.* 22 (1989) 461–473.
- [42] M. Parisien, F. Major, The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data, *Nature* 452 (2008) 51–55.
- [43] M.A. Jonikas, R.J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, R.B. Altman, Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters, *RNA* 15 (2009) 189–199.
- [44] X. Xu, S.-J. Chen, Physics-based RNA structure prediction, *Biophys. Rep.* 1 (2015) 2–13.
- [45] M.J. Boniecki, G. Lach, W.K. Dawson, K. Tomala, P. Lukasz, T. Soltysinski, K.M. Rother, J.M. Bujnicki, SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction, *Nucl. Acids Res.* (2015).
- [46] J.A. McCammon, B.R. Gelin, M. Karplus, Dynamics of folded proteins, *Nature* 267 (1977) 585–590.
- [47] G. Ozer, A. Luque, T. Schlick, The chromatin fiber: multiscale problems and approaches, *Curr. Opin. Struct. Biol.* 31 (2015) 124–139.
- [48] H.H. Gan, S. Pasquali, T. Schlick, Exploring the repertoire of RNA secondary motifs using u graph theory: implications for RNA design, *Nucl. Acids Res.* 31 (2003) 2926–2943.
- [49] C. Laing, S. Jung, N. Kim, S. Elmetwaly, M. Zahran, T. Schlick, Predicting helical topologies in RNA junctions, *PLoS One* 8 (2013). e71947.
- [50] N. Kim, Z. Zheng, S. Elmetwaly, T. Schlick, RNA graph partitioning for the discovery of RNA modularity: a novel application of graph partition algorithm to biology, *PLoS One* 9 (2014). e106074.
- [51] L. Petingi, T. Schlick, Partitioning and classification of RNA secondary structures into pseudonotated and pseudoknot-free regions using a graph-theoretical approach, *Int. Assoc. Eng. Intl. J. Comp. Sci.* 44 (2) (2017).
- [52] N. Kim, M. Zahran, T. Schlick, Computational prediction of riboswitch tertiary structures including pseudoknots by RAGTOP: a hierarchical graph sampling approach, *Meth. Enzym.* 553 (2015) 115–135.
- [53] C.S. Bayrak, N. Kim, T. Schlick, RNA structure prediction with knowledge-based statistical potentials for kink-turn motifs, *Nucl. Acids Res.* 45 (9) (2017) 5414–5422.
- [54] S. Jain, T. Schlick, F-RAG: generating atomic coordinates from RNA graphs by fragment assembly, *J. Mol. Biol.* 429 (2017) 3587–3605.
- [55] C. Laing, T. Schlick, Computational approaches to RNA structure prediction, analysis, and design, *Curr. Opin. Struct. Biol.* 21 (2011) 306–318.
- [56] N. Baba, S. Elmetwaly, N. Kim, T. Schlick, Predicting large RNA-like topologies by a knowledge-based clustering approach, *J. Mol. Biol.* 428 (2016) 811–823.
- [57] J. Izzo, N. Kim, S. Elmetwaly, T. Schlick, RAG: an update to the RNA-As-Graphs resource, *BMC Bioinf.* 12 (2011) 219.
- [58] N. Kim, Exploring RNA Structure Space Using Multidisciplinary Approaches with Applications for Novel RNA Design (Ph.D. thesis), New York University, Department of Chemistry (Program in Computational Biology), New York, NY, May 2009.
- [59] C. Laing, D. Wen, J.T.L. Wang, T. Schlick, Predicting coaxial helical stacking in RNA junctions, *Nucl. Acids Res.* 40 (2011) 487–498.
- [60] H.H. Gan, S. Pasquali, T. Schlick, A survey of existing RNAs using graph theory with implications to RNA analysis and design, *Nucl. Acids Res.* 31 (2003) 2926–2943.
- [61] N. Kim, H.H. Gan, T. Schlick, Designing structured RNA pools for *in vitro* selection of RNAs, *RNA* 13 (2007) 478–492.
- [62] N. Kim, J. Sup Shin, S. Elmetwaly, H.H. Gan, T. Schlick, RAGPOOLS: RNA-As-Graph-Pools – A web server for assisting the design of structured RNA pools for *in vitro* selection, *Bioinformatics* 23 (2007) 2959–2960.
- [63] N. Kim, N. Shiffeldrim, H.H. Gan, T. Schlick, Candidates for novel RNA topologies, *J. Mol. Biol.* 341 (2004) 1129–1144.
- [64] S. Jain, S.B. Ramos, A. Laederach, T. Schlick, A pipeline for computational design of novel RNA-like topologies, *In Revision*, 2018.
- [65] M. Zahran, C. Bayrak, S. Elmetwaly, T. Schlick, RAG-3D: a search tool for RNA 3D substructures, *Nucl. Acids Res.* 43 (2015) 9474–9488.
- [66] L. Hua, Y. Song, N. Kim, C. Laing, J.T.L. Wang, T. Schlick, CHSalign: a web server that builds upon JunctionExplorer and RNAJAG for pairwise alignment of RNA secondary structures with coaxial helical stacking, *PLoS One* 11 (2016). e0147097.
- [67] S. Ovchinnikov, D.E. Kim, R.Y.R. Wang, Y. Liu, F. DiMaio, D. Baker, Improved de novo structure prediction in CASP11 by incorporating co-evolution information into Rosetta, *Prot.: Struct. Fun. Bioinf.* 84 (2016) 67–75.
- [68] Z. Miao et al., RNA-puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures, *RNA* 21 (2015) 1–19.
- [69] S. Tian, R. Das, RNA structure through multidimensional chemical mapping, *Quart. Rev. Biophys.* 49 (e7) (2016) 1–30.
- [70] C. Laing, T. Schlick, Computational approaches to RNA 3D modeling, *J. Phys. Cond. Matter* 22 (2010) 283101.
- [71] R. Das, D. Baker, Automated de novo prediction of native-like RNA tertiary structures, *Proc. Natl. Acad. Sci. USA* 104 (2007) 14664–14669.
- [72] R. Das, J. Karanicolas, D. Baker, Atomic accuracy in predicting and designing noncanonical RNA structure, *Nature Meth.* 7 (2010) 291–294.
- [73] C.Y. Cheng, F.-C. Chou, R. Das, Modeling complex RNA tertiary folds with Rosetta, *Meth. Enzym.* 553 (2015) 35–64.
- [74] M. Rother, K. Rother, T. Puton, J.M. Bujnicki, RNA tertiary structure prediction with ModeRNA, *Brief. Bioinf.* 12 (2011) 601–613.
- [75] M.A. Jonikas, R.J. Radmer, R.B. Altman, Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models, *Bioinformatics* 25 (2009) 3259–3266.
- [76] J.A. Cruz et al., RNA-puzzles: a CASP-like evaluation of RNA three-dimensional structure prediction, *RNA* 18 (2012) 610–625.
- [77] C. Laing, T. Schlick, Analysis of four-way junctions in RNA structures, *J. Mol. Biol.* 390 (2009) 547–559.
- [78] C. Laing, S. Jung, A. Iqbal, T. Schlick, Tertiary motifs revealed in analyses of higher-order RNA junctions, *J. Mol. Biol.* 393 (2009) 67–82.
- [79] T. Schlick, *Molecular Modeling: An Interdisciplinary Guide*, second ed., Springer-Verlag, New York, NY, 2010.
- [80] J. Wang, P. Daldrop, L. Huand, D.M. Lilley, The k-junction motif in RNA structure, *Nucl. Acids Res.* 42 (2014) 5322–5331.
- [81] L. Huang, J. Wang, D.M. Lilley, A critical base pair in k-turns determines the conformational class adopted, and correlates with biological function, *Nucl. Acids Res.* 44 (2016) 5390–5398.
- [82] T. Schlick, R. Collepardo-Guevara, L.A. Halvorsen, S. Jung, X. Xiao, Biomolecular modeling and simulation: a field coming of age, *Quart. Rev. Biophys.* 44 (2011) 191–228.
- [83] T. Schlick, The 2013 nobel prize in chemistry celebrates computations in chemistry and biology, *SIAM News* 46 (2013) 1–4.
- [84] F. Liu, S. Somarowthu, A.M. Pyle, Visualizing the secondary and tertiary architectural domains of IncRNA RepA, *Nat. Chem. Biol.* 13 (2017) 282–289.
- [85] T. Schlick, A.M. Pyle, Opportunities and challenges in RNA structural modeling and design, *Biophys. J.* 113 (2017) 225–234.
- [86] S. Jain, C.S. Bayrak, L. Petingi, T. Schlick, Dual graph partitioning highlights a small group of pseudoknot-containing submotifs in RNAs, in preparation, 2018.
- [87] A.S. Petrov, B. Gulen, A.M. Norris, N.A. Kovacs, C.R. Bernier, K.A. Lanier, G.E. Fox, S.C. Harvey, R.M. Wartell, N.V. Hud, L.D. Williams, History of the ribosome and the origin of translation, *Proc. Natl. Acad. Sci. USA* 112 (50) (2015) 15396–15401.