Computational Prediction of Riboswitch Tertiary Structures Including Pseudoknots by RAGTOP: A Hierarchical Graph Sampling Approach

Namhee Kim, Mai Zahran, Tamar Schlick

Department of Chemistry and Courant Institute of Mathematical Sciences, New York University, New York, USA

1Corresponding author: e-mail address: schlick@nyu.edu

Abstract

The modular organization of RNA structure has been exploited in various computational and theoretical approaches to identify RNA tertiary (3D) motifs and assemble RNA structures. Riboswitches exemplify this modularity in terms of both structural and functional adaptability of RNA components. Here, we extend our computational approach based on tree graph sampling to the prediction of riboswitch topologies by defining additional edges to mimic pseudoknots. Starting from a secondary (2D) structure, we construct an initial graph deduced from predicted junction topologies by our data-mining algorithm RNAJAG trained on known RNAs; we sample these graphs in 3D space guided by knowledge-based statistical potentials derived from bending and torsion measures.
of internal loops as well as radii of gyration for known RNAs. We present graph sampling results for 10 representative riboswitches, 6 of them with pseudoknots, and compare our predictions to solved structures based on global and local RMSD measures. Our results indicate that the helical arrangements in riboswitches can be approximated using our combination of modified 3D tree graph representations for pseudoknots, junction prediction, graph moves, and scoring functions. Future challenges in the field of riboswitch prediction and design are also discussed.

1. INTRODUCTION

1.1. Riboswitch structure and function

Many noncoding RNAs have important regulatory and catalytic roles in various cells and viruses. Riboswitches represent a common type of noncoding RNA that is present in the 5' UTRs of certain mRNAs (Serganov & Nudler, 2013). They offer important specialized components involved in the regulation of cellular function and operate through a conformational switch upon binding to a ligand (Barrick & Breaker, 2007; Breaker, 2012; Serganov & Patel, 2007). The regulatory mechanisms involved include, for example, formation or deletion of transcription terminator (Peselis & Serganov, 2012; Proshkin, Mironov, & Nudler, 2014), sequestration of ribosome-binding sites (Winkler & Breaker, 2005), and emergence of alternative cleavage sites (Cheah, Wachter, Sudarsan, & Breaker, 2007). The ligand can be a small molecule or ion, and the binding interaction triggers a conformational change in the RNA and subsequent altered expression of the open reading frame located within its mRNA. Typically, each riboswitch consists of an aptamer domain that forms the binding pocket for the target metabolite or ion and an expression platform, which overlaps with the aptamer region of the riboswitch and exerts genetic control by one of several possible mechanisms (Breaker, 2012).

Riboswitches exhibit a diverse range of secondary (2D) and tertiary (3D) structures (Montange & Batey, 2008; Peselis & Serganov, 2012), but they generally contain a natural ligand binding or “aptamer domain” (Gold, Polisky, Uhlenbeck, & Yarus, 1995) and an “expression platform” (Winkler & Breaker, 2003; see Fig. 1). The thiamine pyrophosphate (TPP) riboswitch is a classic example. The TPP riboswitch can adopt two structural conformations upon binding to TPP. TPP can only bind upon formation of a thi-box domain between the aptamer domain and the expression platform. The binding triggers a switch in the entire RNA structure
from an active “on” state to an inactive “off” state that causes the formation of a terminator hairpin (Edwards, Klein, & Ferre-D’Amare, 2007; Serganov, Polonskaia, Phan, Breaker, & Patel, 2006).

Based on solved 3D structures, riboswitches have been classified into two types—Type I and Type II (Montange & Batey, 2008). Type I riboswitches are characterized by a single binding pocket supported by a largely pre-established global fold. This arrangement limits ligand-induced conformational changes in the RNA to a small region. The purine riboswitches, the glmS riboswitch, and the S-adenosylmethionine (SAM)-II riboswitch are of this type. Type II riboswitches contain binding pockets split into at least two spatially distinct sites. As a result, binding induces both local changes to the binding pocket as well as global rearrangements to the architecture of the RNA. This latter class includes the TPP riboswitch, SAM-I riboswitch, and M-box magnesium riboswitch (Montange & Batey, 2008). Similar features are found in other noncoding RNAs, making it possible to begin to build a hierarchical classification of RNA structure based on the spatial organization of their active sites and associated 2D structural elements.

1.2. Riboswitch motifs

Riboswitch selectivity is encoded within their conserved sensing domains. These domains can vary in the size and complexity of their 2D and 3D structures. All major riboswitch classes are determined at high resolution in complex with their ligands (see Table 1 for examples). Even though they adopt different conformations, most riboswitch structures can be classified depending on the motif they contain: junctions or pseudoknots. Other motifs
such as kink-turn (k-turn) motifs (Lilley, 2014; Wang, Daldrop, Huang, & Lilley, 2014) recur in many riboswitches as well.

Junctions include single-stranded regions with three or more helical arms. These junctions form long-distance 3D interactions stabilizing the overall conformation. For example, junctions in the Mg\(^{2+}\) class II riboswitches (Serganov & Nudler, 2013) are positioned far from the regulatory helix, but one of the helices of the junction folds back toward the regulatory helix and stabilizes it through long-range 3D interactions. Ligands can bind to the RNA in the junction region or close to the regulatory helix, thus stabilizing the global conformation and 3D interactions.

Some riboswitch architectures, including the SAM-I and fluoride riboswitches, are governed by small pseudoknots, which are formed when two single-stranded regions flanked by a stem are base paired. Some junction riboswitches like \(\text{glmS}\) riboswitch and SAM-I riboswitch also contain pseudoknots that are particularly crucial in the formation of ligand-binding pockets and long-distance 3D contacts.
Many riboswitches contain recurrent structural motifs, which are present in other natural and artificial RNAs, such as k-turn motifs. A k-turn motif is a bulge, which generates a kink between two helices with an angle of \(\sim 50^\circ\) (Lilley, 2014). For example, the cyclic-diGMP, cobalamine, and T-box riboswitches have k-turn motifs. As this motif introduces a tight bend into the axis of the duplex RNA, k-turn motifs serve as key architectural elements that help generate specific ligand-binding pockets (Lilley, 2014). Like other functional RNAs, riboswitches employ these motifs as basic building blocks in their complex spatial conformations.

### 1.3. Advances in computational approaches for RNA structure prediction

Understanding the mechanisms behind RNA functions requires RNA 3D structural knowledge. Multidisciplinary approaches in biology have been commonly used in the past decades and are particularly valuable in the study of RNA molecules. Indeed, RNA’s regulatory roles combined with its modular architecture makes it a suitable subject for systematic computational approaches (see our recent reviews, Kim, Fuhr, & Schlick, 2013; Kim, Petingi, & Schlick, 2013; Laing & Schlick, 2011). Theoretical contributions to the prediction of RNA structure have been made from the prediction of 2D structure to the prediction of 3D folds. For example, programs for predicting RNA 2D structures such as Mfold (Zuker, 2003), RNAfold (Hofacker, 2003), ContextFold (Dowell & Eddy, 2006), and PKNOT (Rivas & Eddy, 1999) for pseudoknot folding, are widely used. Programs to fold 3D structures of small RNAs, such as NAST (Jonikas et al., 2009), FARNA (Das & Baker, 2007; Das, Karanicolas, & Baker, 2010), Vfold (Cao & Chen, 2011; Xu, Zhao, & Chen, 2014), and MC-Sym (Parisien & Major, 2008), have also been developed using coarse-grained models, free energy minimization, and fragment assembly approaches, respectively. These programs can predict 3D structures of small RNAs up to \(\sim 40\) nucleotides within \(\sim 6\) Å root-mean-square-deviation (RMSD) of atomic positions from native structures (Laing & Schlick, 2011). Other programs to annotate motifs in 2D and 3D structures such as FR3D (Petrov, Zirbel, & Leontis, 2011, 2013), MC-Annotation (Gendron, Lemieux, & Major, 2001), and RNAVIEW (Yang et al., 2003) provide useful tools to extract 2D information from 3D structures (Antczak et al., 2014; Kim, Laing, et al., 2014; Laing et al., 2013) and expand our knowledge of recurrent 3D interactions in RNA structures (Kim et al., 2013).
We have contributed to the field of RNA structure modeling by developing RNA-As-Graphs (RAG), a resource for RNAs modeled as planar tree and dual graphs to assist the cataloging, analyzing, and designing of RNA structures (see Fig. 2A; Fera et al., 2004; Gan et al., 2004; Izzo, Kim, Elmetwaly, & Schlick, 2011). The simplification and abstraction of RNA
structures as graphs drastically reduces the conformational space and allows enumeration and classification of RNA structures according to essential but simplified topological aspects, such as helical arrangements and loop connectivity. RAG has been applied to the prediction of RNA-like topologies (Izzo et al., 2011; Koessler, Knisley, Knisley, & Haynes, 2010), *in silico* modeling of *in vitro* selection (Kim, Izzo, Elmetwaly, Gan, & Schlick, 2010), analysis of large viral RNA (Gopal, Zhou, Knobler, & Gelbart, 2012), analysis and design of riboswitches (Quarta, Kim, Izzo, & Schlick, 2009), prediction of RNA junction topology (Laing et al., 2013), and graph partitioning for the discovery of RNA modularity (Kim, Zheng, Elmetwaly, & Schlick, 2014).

Recently, we have extended RNA graph representations from 2D to 3D graphs and developed a hierarchical sampling approach (which we term “RAGTOP” here for RNA-As-Graph-Topologies) to predict global 3D topologies compatible with a given RNA 2D structure (see Fig. 2; Kim, Laing, et al., 2014). RNA 3D graphs represent both topological connectivity of 2D structures and geometrical aspects of helical arrangements in 3D. Our overall approach exploits graph representations to accelerate conformational sampling and generate a graph approximation to an RNA 3D structure. We utilize the modular and hierarchical features of RNA structures in two steps. First, we predict junction topologies based on our data-mining program called RNAJAG (RNA-Junction-As-Graphs; Laing et al., 2013). Second, we sample 3D graphs guided by knowledge-based statistical potentials derived from bending and torsion measures of internal loops as well as radii of gyration for known RNAs (Kim, Laing, et al., 2014). This graph sampling approach RAGTOP has demonstrated significant improvements over current approaches for characterizing 3D global helical arrangements in large RNAs from a given 2D structure (Kim, Laing, et al., 2014).

Here, we modify our 3D tree graph representation to represent riboswitches with pseudoknots and apply our hierarchical graph sampling tool for the prediction of representative riboswitch structures. In Section 2, we characterize how to represent RNA 2D and 3D structures including both pseudoknot-free and pseudoknot structures by extended tree graphs. We also describe our hierarchical graph folding approach for the prediction of riboswitch structures. Section 3 presents our graph sampling and prediction results for 10 different riboswitches, 6 with pseudoknots, and compares them to solved structures based on global and local RMSD measures. We conclude in Section 4 with summary and future challenges.
2. HIERARCHICAL GRAPH FOLDING APPROACH

Here, we modify RAGTOP described as in Kim, Laing, et al. (2014) to handle pseudoknots by altering the 3D graph representations by additional edges for pseudoknot interactions, adding a term for pseudoknots in our scoring function, and treating MC pivot moves at pseudoknots differently (see below).

2.1. RNA 2D and 3D graph representation

RNA graph representations provide the basis of our hierarchical folding approach RAGTOP (see Fig. 2A). Recently, we developed 3D tree graph representations which preserve the rules for 2D tree graph representation and can further represent parallel and antiparallel helical arrangements in 3D space: (1) unpaired RNA 2D building blocks (hairpin loops, internal loops, junctions) and the helix ends are translated to vertices; (2) helices in pseudoknot-free structures are translated to edges; (3) edges are also set to represent the connection between vertices of unpaired regions and helix ends; (4) 3D coordinates are assigned for each vertex at the centers of helices and loops (see Kim, Laing, et al., 2014 for full equations). Here, to represent pseudoknot structures, we modify our 3D graph representation by additional edges for pseudoknot interactions: (5) the pseudoknot interactions are translated into edges. The pseudoknot edge is formed by the connection of two loop vertices, which interact via pseudoknot base pairing. Figure 2A shows an example of a modified 3D graph for a pseudoknot in the fluoride riboswitch (PDB entry 4ENC), formed by base pairs connecting one internal loop and the other dangling end, represented by a pseudoknot edge.

2.2. Junction prediction by RNAJAG

RNAJAG predicts helical arrangements of RNA junction structures as tree graphs from a given 2D structure. RNAJAG can indicate the family type and stacking orientation of a given junction by a data-mining approach based on the random forests (decision tree) procedure trained using loop length, adenine base content, and free energy estimates of two base pairs in junction helix ends (Laing et al., 2013). The 3- and 4-way junctions are classified into three families (called A, B, and C; Lescoute & Westhof, 2006) and nine families (H, cH, cL, cK, π, cW, Ψ, cX, and X), respectively, according to coaxial stacking and helical configuration (Laing & Schlick, 2009). RNAJAG is used in RAGTOP to generate initial graph structures with specific helical
arrangements (Laing et al., 2013). Four recurrent junction family types (A, B, C, and cL) in riboswitch structures are shown in Fig. 2B.

### 2.3. Monte Carlo simulated annealing (MC/SA) graph sampling

Starting from initial graph setup by size measures and junction prediction, we perform MC/SA sampling of RNA 3D graphs at the flexible vertices in internal loops (Kim, Laing, et al., 2014). We use two types of moves—restricted pivot moves, by reciprocally decreasing angle ranges from 360° to 10° along MC steps, and random pivot moves (Kim, Laing, et al., 2014). Here, in addition to internal loops, we allow pivot moves for hairpin loop vertices if they are involved in a pseudoknot interaction. To score our graphs, we developed knowledge-based statistical potentials from statistical analyses of geometrical features of solved RNA structures, including bending and torsion angles between two helices of internal loops, and radii of gyration of the entire RNAs (Kim, Laing, et al., 2014). Here, to model pseudoknots, we modify the scoring function by adding a term for the pseudoknot edge \( G_{pk} = D - \overline{D} \) where \( D \) is the length of an pseudoknot edge of each sampled graph and \( \overline{D} \) is the “equilibrium” length of a pseudoknot edge observed from known pseudoknots (between 10 and 15 Å). Thus, the scoring function for a pseudoknot structure \( G \) is the sum of scores for pseudoknot-free tree graph \( G_{internal} + G_{Rg} \) (see Kim, Laing, et al., 2014 for details) plus \( G_{pk} \). The scores guide the conformational sampling: if the score for a new conformation is lower than that of the old conformation, the new conformation is accepted. If the new score is higher, the SA sampling proceeds: some moves with higher score at each step are accepted with decreasing probability along the MC steps (for details, see Kim, Laing, et al., 2014).

### 2.4. Assessment of sampled graphs

After MC/SA, the candidate graphs are compared to the 3D graphs translated from solved RNAs by three procedures (P1–P3; Kim, Laing, et al., 2014). P1 directly compares the candidate graph with the lowest RMSD score among the final pool of accepted graphs to the reference graph translated from the solved structure. P2 compares our lowest-scored graph among accepted graphs to the reference graph. For random moves, conformational space is more globally sampled compared with restricted moves, and additional clustering is required to select a representative graph from among five clusters (P3). Thus, for P3 we compare the cluster representatives to the reference graph. We compare resulting graphs to reference
graphs translated from solved structures by the average RMSDs. Graph-based RMSD provides a valid measure to compare global topological similarities with positive correlation to atomic RMSD (Kim, Laing, et al., 2014). However, as elaborated in Parisien, Cruz, Westhof, and Major (2009), RMSD spreads the structural dissimilarities and does not specify local errors such as base interactions and local helical arrangements. As we start with given 2D structures, base interactions are same for all sampled graphs. Thus, in addition to graph RMSDs, we compare resulting graphs in terms of vertex-to-vertex distances, which account for the specificity of local helical arrangements.

2.5. All-atom building by RAG-3D

For the all-atom model building, we use a threading-like procedure based on a search for graph similarities with the 3D graphs classified in our database RAG-3D (Zahran, Elmetwaly, & Schlick, 2014), an extension of the original RAG database containing 2D planar graphs (Fera et al., 2004; Izzo et al., 2011). RAG-3D contains 3D atomic models extracted from RNA structures present in the PDB database, and linked to corresponding 3D graphs. In RAG-3D, the 3D graphs are classified based on the original RAG motif IDs, which reflect topological properties of 2D structural elements. All-atom models are constructed in three steps: (i) identifying the motif ID of the target graph; (ii) comparing the target graph to all 3D graphs catalogued with the same motif ID in RAG-3D based on a standard RMSD calculation; and (iii) selecting the lowest-score RMSD graph, extracting its all-atom 3D coordinates, and adjusting base content to match that of the target sequence, while keeping the backbone intact.

3. APPLICATION TO RIBOSWITCH STRUCTURE PREDICTION

We apply the RAGTOP procedure as modified here for pseudoknots to a set of 10 representative riboswitches in Table 1. These 10 riboswitches have diverse structural features in terms of sequence length and diverse combinations of structural elements, namely internal loops, junctions with 3- and 4-way junctions, and pseudoknots. In particular, among these 10 riboswitches, 7 structures have junctions and 6 have pseudoknots.

Table 2 and Fig. 3 provide our graph results for junction prediction and graphs after Monte Carlo sampling for these 10 representative riboswitches.
## Table 2 Graph results for 10 riboswitches

<table>
<thead>
<tr>
<th>PDB</th>
<th>Class</th>
<th>L</th>
<th>J</th>
<th>Native Coaxial</th>
<th>Family Coaxial</th>
<th>RMSD after MC/SA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Coaxial Family</td>
<td>Coaxial Family</td>
<td>P1</td>
</tr>
<tr>
<td>4ENB</td>
<td>Fluoride</td>
<td>52</td>
<td>No</td>
<td>N/A</td>
<td>N/A</td>
<td>2.45</td>
</tr>
<tr>
<td>2KZL</td>
<td>T-box</td>
<td>55</td>
<td>No</td>
<td>N/A</td>
<td>N/A</td>
<td>6.06</td>
</tr>
<tr>
<td>2G9C</td>
<td>Purine</td>
<td>67</td>
<td>3WJ</td>
<td>H1H3</td>
<td>C</td>
<td>3.88</td>
</tr>
<tr>
<td>3RKF</td>
<td>Guanine</td>
<td>67</td>
<td>3WJ</td>
<td>H1H3</td>
<td>C</td>
<td>3.67</td>
</tr>
<tr>
<td>3Q3Z</td>
<td>c-di-GMP-II</td>
<td>77</td>
<td>No</td>
<td>N/A</td>
<td>N/A</td>
<td>2.22</td>
</tr>
<tr>
<td>3D2G</td>
<td>TPP</td>
<td>77</td>
<td>3WJ</td>
<td>H1H2</td>
<td>A</td>
<td>3.62</td>
</tr>
<tr>
<td>2GDI</td>
<td>TPP</td>
<td>80</td>
<td>3WJ</td>
<td>H1H2</td>
<td>A</td>
<td>4.25</td>
</tr>
<tr>
<td>2HOJ</td>
<td>TPP</td>
<td>83</td>
<td>3WJ</td>
<td>H1H2</td>
<td>A</td>
<td>4.70</td>
</tr>
<tr>
<td>2GIS</td>
<td>SAM</td>
<td>94</td>
<td>4WJ</td>
<td>H1H4 and H2H3</td>
<td>cL</td>
<td>11.41</td>
</tr>
<tr>
<td>4B5R</td>
<td>SAM</td>
<td>95</td>
<td>4WJ</td>
<td>H1H4 and H2H3</td>
<td>cL</td>
<td>10.77</td>
</tr>
</tbody>
</table>

For riboswitches with junctions, each junction is listed with its junction family and coaxial-stacking arrangement from native structures and predicted structures by RNAJAG. N/A indicates that the structure does not include junction. After MC/SA sampling, graph RMSDs between reference graphs from solved structures and our sampled graphs by MC/SA—the lowest (P1, random moves), the lowest score (P2, restricted moves), and the lowest cluster representative (P3, random moves) after MC/SA—are shown. See Figs. 3 and 4 for vertex-to-vertex distance measures between corresponding vertices of reference and sampled graphs (P1–P3).
Figure 3 See legend on opposite page.
We assess candidate graphs with respect to predicted graphs (translated from solved structures) rather than predicted atomic models versus solved atomic models, by junction types, coaxial stacking, and graph RMSD. Our analysis has shown that graph and all-atom RMSDs are positively correlated with positive correlation coefficient (\( \sim 0.89 \)); thus, overall similarity between structures can be captured by graph RMSD (Kim, Laing, et al., 2014). Figure 4 shows graph results by vertex-to-vertex distance specifying local errors by helical arrangements.

For RNAs containing junctions (7 of 10 riboswitches in Table 1), junction families and coaxial stacking are perfectly predicted by RNAJAG based on a collective training set of 244 junctions. The purine and guanine riboswitches (PDB entries 2G9C and 3RKF) have 3-way junctions with Family C having coaxial stacking of H1 and H3 and a parallel helix H2, which is predicted correctly by RNAJAG. Similarly, for TPP riboswitches (PDB entries 3D2G, 2HOJ, and 2GDI), the 3-way junctions are classified and predicted as Family A with coaxial stacking of H1 and H2 and a perpendicular helix H3. For SAM riboswitches (PDB entries 2GIS and 4B5R), the 4-way junctions are classified and predicted as cL, which is an expanded version of Family A 3-way junction, with coaxial stacking of H1 and H4 and H2 and H3. Thus, RNAJAG can predict junction classes and coaxial stacking for riboswitches very well, with prediction accuracy much higher than overall prediction accuracy for all RNAs: 95%/92% in 3-/4-way junctions for coaxial tacking and 94%/87% for family type in 3-/4-way junctions. The abundance of junction structures in riboswitches increases the prediction accuracy for junction classifications.

Starting from an initial graph constructed from the RNAJAG junction prediction, we sample riboswitch graphs using 10^4 steps for restricted pivot moves, which converge to one region of conformational space, as well as random pivot moves, which explore multiple regions of space and thus require clustering analysis. Graph sampling improves orientation of loops, pseudoknots, and overall 3D topology geometries. For RMSDs relative to reference graphs (P1 in Table 2), the lowest values range from 2.22 Å.
Figure 4  Local geometry analysis by vertex-to-vertex distance measures from reference graphs to the lowest RMSD (P1, green line (light gray in the print version)), the lowest-scored (P2, blue line (dark gray in the print version)), and the lowest cluster representative (P3, red line (gray in the print version)) graphs for 10 riboswitches in Tables 1 and 2. Vertex ID matches with graphs in Fig. 3.
(c-di-GMP-II riboswitch, PDB entry 3Q3Z) to 11.41 Å (SAM riboswitch, PDB entry 2GIS) using random moves. For most riboswitches except the SAM riboswitch (PDB entries 2GIS and 4B5R), the RMSD with respect to known structure (P1) is less than or close to 6 Å. For SAM riboswitches with 4-way junctions, the RMSD is high (11.41 Å for PDB entry 2GIS and 10.77 Å for PDB entry 4B5R). This is because the initial junction geometries are held rigid during MC sampling. Although the 4-way junction class for the initial junction geometry is predicted correctly (family cL and coaxial stacking of H1H4 and H2H3), related distances are imperfect. For example, the distance between helices H1 and H3 should be much closer than that between H2 and H4, but the generated graph has the longer distance between H2 and H4.

When the reference graph is not known, we consider both the lowest-scored graph based on restricted moves (P2) and the lowest-scored graph representatives among five clusters based on random moves (P3). For P2, graph RMSEs range from 2.76 Å (fluoride riboswitch, PDB entry 4ENB) to 17.98 Å (TPP riboswitch, PDB entry 2GDI). For random graphs, representative graphs from five clusters sorted by score from low- to high-offer candidate 3D topologies in the absence of solutions (Fig. 3). Representative graphs from cluster 1 have RMSEs ranging from 3.08 Å (fluoride riboswitch, PDB entry 4ENB) to 18.34 Å (TPP riboswitch, PDB entry 3D2G; P3 in Table 1), similar to the lowest-scored graphs (P2).

The modified 3D tree graphs with additional edges handle pseudoknots effectively. For fluoride riboswitch (PDB entry 4ENB), an edge connecting the internal loop vertex ID 5 and the dangling end vertex 1 represent its pseudoknot (see Fig. 3 for vertex ID). For the purine and guanine riboswitches (PDB entries 2G9C and 3RKF), two distant hairpin loops are interconnected to form a pseudoknot as represented by an edge connecting vertices 5 and 3. For the c-di-GMP-II riboswitch (PDB entry 3Q3Z), a pseudoknot is formed by the connection between one hairpin (vertex ID 3) and the other internal loop (vertex ID 5). For the SAM riboswitches (PDB entries 2GIS and 4B5R), a pseudoknot edge is formed by the connection between a junction (vertex ID 15) and a hairpin loop (vertex ID 3). The sampled graphs approximate global interactions due to pseudoknots by closer distances between the two interconnected strands estimated by the size of pseudoknot edges. After MC/SA, the size of pseudoknot edge in the resulting graphs (P1–P3) is similar with that in reference riboswitches, which ranges from 10 Å (pseudoknot edge connecting vertices 3 and 5 in PDB entries 2G9C and 3RKF) to 20 Å (pseudoknot edge connecting vertices
5 and 15). All RMSD measures between resulting graphs (P1–P3) and reference graphs are relatively small for these pseudoknot structures. Even for the SAM riboswitches (PDB entries 2GIS and 4B5R) whose initial graph geometries are poor, the overall graph RMSDs are reduced by around 2 Å compared to graph results without pseudoknot edges (from 17.43 Å (Kim, Laing, et al., 2014) to 15.64 Å as shown in P3 in Table 2). For P3, RMSD ranges from 3.08 Å (PDB entry 4ENB) to 15.64 Å (PDB entry 2GIS).

As an alternative to the global RMSD measures, we also analyze local geometrical features. Figure 4 shows the results of local distances between corresponding vertices in aligned resulting graphs and reference graphs, which is developed in a similar spirit in local base interactions whose dissimilarities are spread in RMSD measures (Parisien et al., 2009). These vertex-to-vertex distance measures indicate the local similarities/dissimilarities. We see that the distance for hairpins is larger than that for junctions and internal loops (for example, see hairpin vertices ID 3 and 5 in TPP riboswitch, PDB entry 3D2G). This is because we locate the hairpin edge based on only size estimated by the sequence length of a hairpin sequence, without consideration of the angles about a hairpin edge.

Even though our graph prediction provides the information about the overall helical arrangements of riboswitches, atomic models are ultimately required. We applied our build-up approach to the 3-way junction structure of guanine riboswitch (PDB entry 3RKF; Laing et al., 2013). This 3-way junction guanine riboswitch RNA contains 53 nucleotides. RNAJAG was able to correctly predict both the junction family type and the coaxial stacking and predicted a graph with RMSD value of 4.32 Å with respect to the graph of its native structure translated from the PDB structure. We superimposed the predicted graph against all the graphs of the same motif ID family (namely, 4_2) available in the RAG-3D database, and ranked them based on their RMSDs with respect to the target graph. We extracted the all-atom coordinates of the lowest-RMSD graph found (4.41 Å), and created a model by mutating the bases to match the query sequence. We obtained an RMSD value of 5.09 Å for the all-atom model junction region compared to its native structure, as shown in figure 10 in Laing et al. (2013).

4. FUTURE CHALLENGES AND PERSPECTIVES

We have extended and applied our hierarchical computational approach (named RAGTOP here) to predict riboswitch tertiary structure with pseudoknots by combining our coarse-grained graph sampling
approach (Kim, Laing, et al., 2014), which utilizes RNAJAG for initial junction prediction and knowledge-based scoring functions for MC/SA sampling, with modified features for pseudoknots. Three added features for pseudoknots include modified 3D graphs with pseudoknot edges, graph moves for hairpin loop vertices involving a pseudoknot formation, and updated scoring function with an additional term for pseudoknot edge lengths. Our sampling based on geometric statistical potentials produces graphs whose 3D shapes resemble native structures, and the lowest-scored graphs are also reasonably selected without knowledge of reference graphs both in our restricted (P2) and random (P3) move protocols. The application of our approach to 10 riboswitches with internal loops, 3- and 4-way junctions, and pseudoknots shows that graph-based sampling can reasonably predict the junction structures of riboswitches, and provide a good approximation for global helical arrangements. Graph-based structure prediction is expected to continue to a useful tool to predict and design riboswitch structures.

Our current graph sampling approach is primarily applicable to the prediction of riboswitch structures with 3- and 4-way junctions. However, riboswitches can have higher-order junctions (e.g., lysine riboswitch, 5-way junctions). RNAJAG can potentially be extended to predict higher-order junctions. For example, 5-way junctions can be partitioned into 3- and 4-way subjunctions. Our prediction of RNA junctions could also be extended from the current three discrete models (parallel, perpendicular, and diagonal helical arrangements) to model continuous helical orientations.

Our 3D graph representations might also be modified and expanded to represent long-range interactions. Here, we added an edge to represent a pseudoknot formed by intertwined and long-range base pair interactions between two hairpin loops and between a hairpin loop and a junction/inter- nal loop. This representation provides a reasonable framework for pseudoknot topology sampling. The corresponding pivot moves for hairpin loops involving pseudoknots and the modified scoring function term to target a pseudoknot edge length guide our graph sampling towards native-like pseudoknot topologies. In the future, dual graphs could also be explored to model pseudoknot structures rigorously.

Our scoring function does not account for k-turn motifs, hairpin angles, and ligand-binding cases. However, since k-turn motifs have highly conserved sequence contents and 2D structures, it is possible to identify them based on primary and 2D structures. Our scoring function could discern
internal loops where k-turn can potentially occur and score them differently. In addition, the shape of hairpin loops could also be considered, so as to locate nonplanar hairpin edges. A separation of self-folding RNA parameters from those for substrate-binding RNAs could also be envisioned.

The application of RAGTOP to riboswitch design appears promising. Combined with other bioinformatics searches of sequences and accurate 2D folding algorithms, our approach could be applied to design new riboswitch structures. Furthermore, since riboswitches are potential targets for the construction of artificial genetic circuits that could be controlled by nonnatural compounds, understanding the regulatory and structural principles exploited by natural riboswitches could help in the development of synthetic riboswitches that respond to a particular ligand. Computational structure prediction of riboswitch candidates with unknown 3D structures could ultimately help identify and design new riboswitch classes.

ACKNOWLEDGMENTS
This work is supported by the National Institutes of Health (GM100469, GM081410). We also gratefully acknowledge the Telluride Science Research Center (TSRC) and the NIH conference grant (GM112216-01) which brought together RNA scientists and stimulated this work.

REFERENCES


