Topological Constraints and Their Conformational Entropic Penalties on RNA Folds

Chi H. Mak[a,b,c,1] and Ethan N. H. Phan[a]

[a]Department of Chemistry, [b]Center of Applied Mathematical Sciences, [c]Department of Biological Sciences, University of Southern California, Los Angeles, California 90089, USA

[1]To whom correspondence may be addressed:

Chi H. Mak, Department of Chemistry, University of Southern California, Los Angeles, CA 90089
Tel: 213-740-4101, Fax: 213-740-3972, E-mail: cmak@usc.edu

Key words:

RNA structure, conformational entropy, nucleic acid backbone conformations, folding free energy

Running title:

Topological Constraints on RNA Folds

ABSTRACT

Functional RNAs can fold into intricate structures using a number of different secondary and tertiary structural motifs. Many factors contribute to the overall free energy of the target fold. This study aims at quantifying the entropic costs coming from the loss of conformational freedom when the sugar-phosphate backbone is subjected to constraints imposed by secondary and tertiary contacts. Motivated by insights from topology theory, we design a diagrammatic scheme to represent different types of RNA structures so that constraints associated with a folded structure may be segregated into mutually independent subsets, enabling the total conformational entropy loss to be easily calculated as a sum of independent terms. We used high-throughput Monte Carlo simulations to simulate large ensembles of single-stranded RNA sequences in solution to validate the assumptions behind our diagrammatic scheme, examining the entropic costs for hairpin initiation and formation of many multiway junctions. Our diagrammatic scheme aides in the factorization of secondary/tertiary constraints into distinct topological classes and facilitates the discovery of interrelationships among multiple constraints on RNA folds. This novel perspective leads to useful insights into the inner workings of some functional RNA sequences, demonstrating how they might operate by transforming their structures among different topological classes.

INTRODUCTION

RNA sequences are predominantly found single-stranded in the cell, but they can assemble into specific higher-order structures by utilizing secondary and tertiary structural building blocks. The free energy change starting from an open unfolded chain going to the final folded conformation, ΔG_{fold}, determines the stability of the fold. A number of molecular factors control this folding free energy, including chain conformational fluctuations, base stacking, base complementarity interactions, as well as other solvent-induced forces such as counterion-mediated intrachain attractions (1-6). For the fold to be thermodynamically stable, the overall ΔG_{fold} from these various factors must add to produce a downhill driving force, i.e. a net negative ΔG_{fold}. Of all the factors that make up ΔG_{fold}, there is only one term that is guaranteed to be positive, and this is $-T\Delta S_{\mathrm{b}}$, where ΔS_{b} is the change in conformational entropy of the RNA backbone upon folding.

Formation of secondary and tertiary contacts on the RNA sequence introduces constraints into the conformation of the chains. The conformational contribution to the free energy $-T\Delta S_{\mathrm{b}}$ must therefore be uphill. On the secondary structural level, base pairing requires two nucleobases from different positions on the RNA sequence to adopt a specific relative geometry, while base stacking constrains two adjacent bases to a different relative geometry putting one base on top of the other. On the tertiary level, contacts such as kissing hairpins or loop-receptor type interactions place other kinds of constraints on the conformation of the chain. A thermodynamic ensemble of free chains has none of these constraints, and the variational statement of the second law of thermodynamics states that the introduction of internal constraints into the ensemble must raise the free energy, or at the minimum leaves it unchanged (7, 8). Therefore, the conformational entropy of the RNA backbone is necessarily suppressed when constraints are imposed. Another way to view this is to consider a chain that has been compacted by internal constraints. Upon the removal of these constraints, it will unfurl if no force other than chain conformational entropy is present. Therefore, folding must suppress ΔS_{b} producing a thermodynamically uphill penalty against the folded conformation.

The fact that the chain conformational entropy ΔS_{b} upon folding is always less than zero has important consequences. First, if we denote all terms in ΔG_{fold} due to factors other than backbone entropy -- base complementarity interactions, base stacking interactions, counterion-mediated electrostatic interactions, and excluded volume interactions -- by $\Delta G'$, the thermodynamic requirement that $\Delta G_{\mathrm{fold}} = \Delta G' - T\Delta S_{\mathrm{b}} < 0$ for a stably folded RNA demands that $\Delta G'$ must be more negative than $T\Delta S_{\mathrm{b}}$. The magnitude of $T\Delta S_{\mathrm{b}}$ therefore places a rigorous lower bound on the strengths of all the other thermodynamic forces that make the fold overall stable. Second, the backbone conformational entropy can help answer the question of how different RNAs assemble their folds. If folding proceeds predominantly via the formation of local domains, the $(T\Delta S_{\mathrm{b}})_i$ penalty for each domain i (a domain is defined here as any segment on the RNA sequence that forms a local higher-order structure) must be offset by the free energy gain $\Delta G'$ within the same local domain such that $(\Delta G' - T\Delta S_{\mathrm{b}})_i < 0$ for all domains i before the global fold is assembled. If

on the other hand folding is non-hierarchical and corporative, as seen in existing studies of RNA folding mechanism (9, 10), then $(\Delta G^{'} - T\Delta S_{\mathrm{b}})_i$ for some domains might be negative while others are positive, but it is only through the sum of them that $\sum_i(\Delta G' - T\Delta S_{\mathrm{b}})_i$ becomes net negative. Therefore, being able to compute the conformational entropy within different folding domains is also important.

In this paper, our goal is to develop the theoretical basis for calculating ΔS_{b} as a function of the constraints on the RNA backbone imposed by known secondary or tertiary structures. The first question is a technical one. Is there an efficient computational methodology to accurately quantify backbone conformational entropy? The second question is a conceptual one. How do we define these constraints, and more importantly, how do we decide whether a set of constraints are independent or correlated? This paper addresses these two questions by formulating a topological view of RNA folds.

MATERIAL AND METHODS

Relationship between Constraints and Backbone Conformational Entropy

Examples of the kind of constraints that defines the secondary and tertiary structures of a RNA may be base pairs, stacked bases or other tertiary interactions. We denote each constraint symbolically by c_j and in a folded RNA there could be N of these. In a thermal ensemble of free RNA chains in solution, the entropy cost ΔS_{b} of imposing these constraints $\{c_1, c_2, c_3, \ldots\ c_N\}$ on the chains can be calculated from the probability of observing chains that meet these conditions (11, 12):

$$P(c_1, c_2, c_3, \ldots\ c_N) = e^{\Delta S_{\mathrm{b}}/R}, \qquad (1)$$

where R is the gas constant. For even a short chain with any appreciable secondary or tertiary structure, the number of base pairs, stacked bases and other tertiary contacts is usually quite large. The joint probability of all these constraints occurring on the same chain is consequently small, and ΔS_{b} is usually large and very negative. While Eq. 1 is a possible way to compute ΔS_{b}, the number of chain conformations that must be sampled is impractically and prohibitively large.

A reduction of the joint probability is possible if the constraints can be divided into subsets which are independent of each other. If this is the case, Eq. 1 can be simplified. For instance, if there are six constraints and they can be factored into three independent subsets $\{c_1, c_2\}$, $\{c_3\}$ and $\{c_4, c_5, c_6\}$, then $e^{\Delta S_{\mathrm{b}}/R} = P(c_1, c_2, c_3, \ldots\ c_6) = P(c_1, c_2)P(c_3)P(c_4, c_5, c_6)$, and the entropy becomes a sum of three independent terms, one for satisfying each of these three independent sets of constraints. If this is the case, the entropy can be more easily evaluated because each of the joint probabilities that has to be computed requires many fewer conditions to be jointly satisfied. In the next section, we will devise a topological representation of these constraints to help us better understand how to factor them into independent subsets.

Topological Representation of Secondary and Tertiary Structural Constraints

In this section, we describe a useful topological representation for some of the common secondary and tertiary constraints found in typical RNA folds. The use of graphs in the study of RNA structure is a well-documented practice that has allowed the tools and results of graph theory to be put to bear on problems such as secondary structure enumeration and comparison (13-15). Early uses of graph theory in RNA studies heavily relied on so called tree graphs of RNA structure which represented junctions and loops in secondary structures as vertices (points) of a graph and helices as the edges connecting the vertices of the graph. Though useful in allowing graph theoretic results to be applied to analyzing RNA structure, tree graphs can only show structure that contains helices and loops. This issue was eventually addressed by the introduction of dual graphs by Schlick and coworkers (16-19). In the dual graph representation, helices are represented by vertices of the graph, while the unpaired segments are represented by the edges connecting the vertices. This results in a graph that is not visually relatable to the 2D secondary structure but allows for pseudoknot and structures such as quadruplex and triple helices to be shown explicitly.

Figure 1 shows several examples of the secondary structural motifs seen in many RNA folds and their corresponding graph representation. Figure 1(a) depicts a three-way junction with two hairpins in the interior of the sequence and a helix between the 5' and 3' terminal residues, with three intervening single-stranded loop segments. In this case, the constraints associated with the secondary structure are the base pairing and stacking forces that hold the helices together. If these forces are removed, the chain will unfurl. The backbone conformational entropy is the logarithm of the joint probability of observing all these constraints being satisfied on one chain. In the middle row of Fig. 1(a), we group all the constraints that come from the same stem into one set. There are three stems in this structure and hence three subsets of constraints. The reason why we choose to view each stem as one subset is because the multiple constraints in each set (i.e. base pairs and base stacks) are clustered. Unless there are additional tertiary contacts between these stems, they should be largely unaware of the existence of the constraints in the other sets.

While the division of constraints into the three subsets depicted in the second row of Fig. 1(a) seems reasonable, we have omitted the central fact that the three helices are connected by single-stranded segments that make up the rest of the three-way function. The connectivity among the helices, while not explicitly given in our list of constraints, is implicit due to the backbone continuity of the RNA. In the topological representation, the segments labelled ***a*** through ***c*** in the second row of Fig. 1(a) remind us that these strands, as well as those in the hairpins ***d*** and ***e***, must be counted as implicit constraints for this construct.

The third row of Fig. 1(a) shows our topological representation of all the constraints inherent in the structure 1(a), including both explicit as well as implicit ones. All the constraints due to a single stem (base pairing and base stacking forces) are represented by one solid circle. Following standard terminology in topology, each circle is a "vertex". The loops labelled ***a*** through ***e*** are called "arcs", or edges of the graph, and they make manifest the implicit constraints coming from the backbone

connectedness. Notice that four arcs passes through every vertex. This corresponds to the physical observation that each helix can have at most 2 strand coming from either end of the helix. The half circle at the lower right is actually two arcs, denoting the 5' and 3' free termini of the chain. Free ends on the 5' and 3' termini of a chain do not cost any entropy, hence ΔS_{b} for a structure with or without free ends would have been the same. This topological reduction of the secondary structure in Fig. 1(a) delineates the key constraints that define the fold as well as the relationships among them. Notice that while all helices are represented by just dots, the intrinsic entropy of each stem depends on the size of each helix measured in nucleotide (nt) units, which must be specified in order for its entropy to be evaluated correctly.

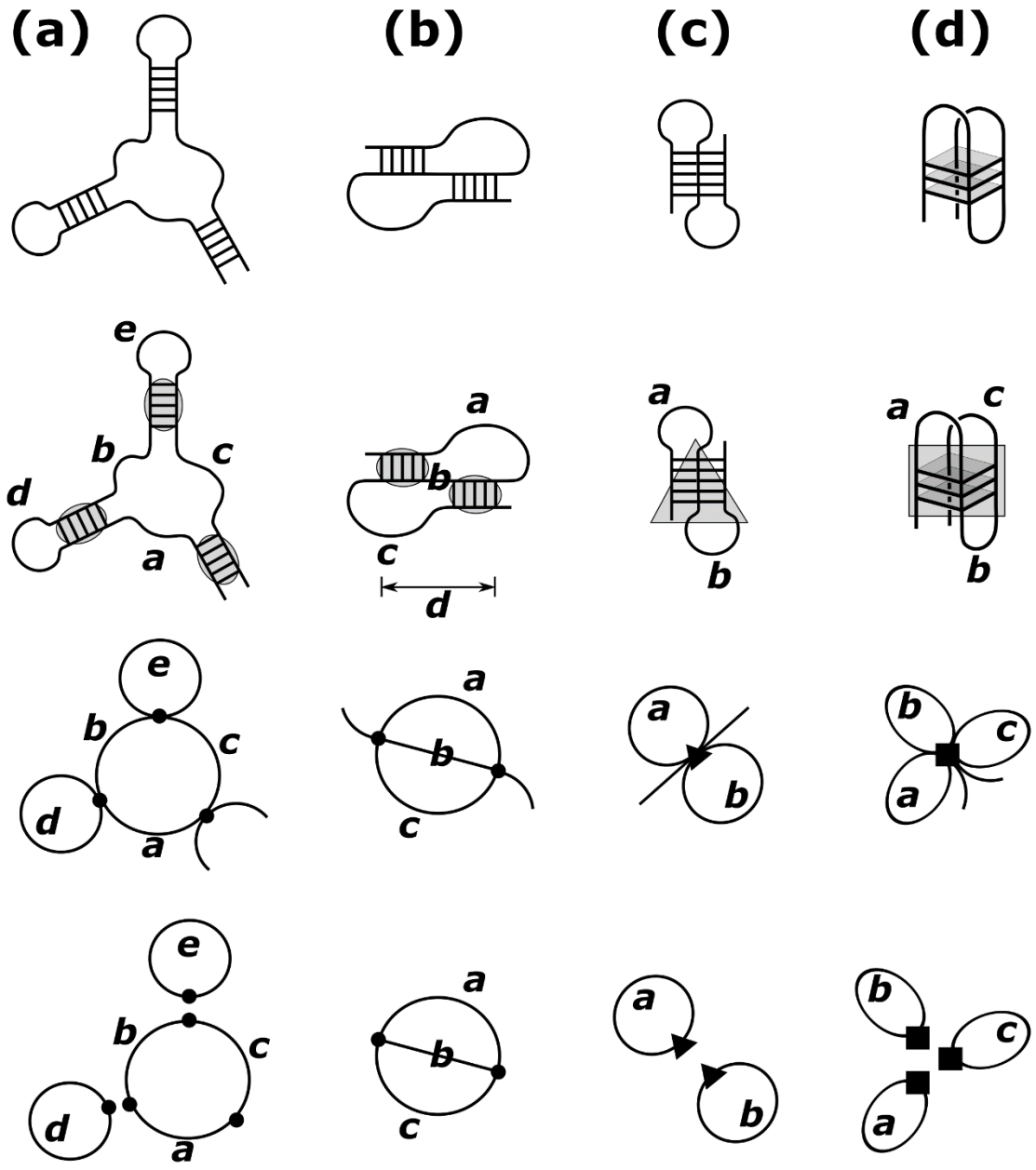

FIGURE 1.

Various secondary structures, the total enumeration of the constraints that define them, and their conversion into a diagrammatic topological representation followed by factorization. (a) A three-way junction is defined by five single-stranded lengths and 3 helices. It is factored into 3 independent subsets which can be treated separately. (b) A pseudoknot is defined by three single-stranded length and two helices. Due to backbone connectivity, the diagram is not factorizable. (c) A triple helix is defined by two single-stranded loops and one triple helix structure. The factorization suggests the two loops are approximately independently of each other. (d) A quadruplex is defined by several loops threaded through the quadruplex core. The factorization shown here suggests that the three loops, after topological reduction, should become approximately independent of each other.

Figure 1(b) shows a schematic structure of a pseudoknot, which helps illustrates additional features of our topological representation. The second row of Fig. 1(b) suggests that the constraints coming from each stem can be grouped together into one subset. The three single-stranded regions internal to the pseudoknot are labelled ***a*** through ***c***. These segments constitute the implicit constraints originating from the connectedness of the backbone. The third row of Fig. 1(b) shows the topological representation of all these constraints in reduced form. The arcs labelled ***a***, ***b*** and ***c*** correspond to the loops depicted in the second row. As in the three-way junction, four arcs go through every vertex. Though not explicitly shown in the topological representation, the number of nucleotides between the entry point into the pseudoknot and the exit point, labelled ***d*** in the second row of Fig. 1(b), needs to be specified for the entropy to be evaluated properly. Again, the free 5' and 3' ends are indicated by open arcs, but as described above they do not cost additional entropy.

Figure 1(c) shows a schematic drawing of a triple helix, and 1(d) shows a quadruplex. The same topological reduction procedures described above lead to the diagrams on the third row of Figs. 1(c) and (d). For the triple helix in Fig. 1(c), its topological representation has only one vertex, but six arcs go through it. To differentiate this from the vertices in Fig. 1(a) and (b), the vertex in Fig. 1(c) is shown as a solid triangle. The two relevant loops are labelled ***a*** and ***b***. Again, the size of the triple helix in nt units must be specified for the entropy to be computed properly. The quadruplex structure in the first row of Fig. 1(d) reduces to the diagram on the third row. There are three loops labelled ***a***, ***b*** and ***c***. This vertex, which has eight arcs going through it, is shown as a solid square. The size of the quadruplex stack in nt units must be specified for the total entropy to be calculated properly.

Factoring Diagrams into Approximately Independent Pieces

While the topological reductions introduced in the last section transform the constraints that define the secondary and/or tertiary structure of a RNA fold into diagrammatic elements, the fact that the vertices and arcs in the topological representation remain connected suggests that they are still correlated with each other. However, there exists an implicit assumption within the literature for RNA secondary structure modeling that loops can be factored into independent components. Examples of this assumption being used include the nearest-neighbor model of Turner and Mathews (20-23), web servers that utilizes the nearest-neighbor model to calculate free energy of RNA structures such as MFold (20, 24-27) or NUPack (28), and discrete chain models in which loops are formed as part of a random walk (29, 30). In the following, we develop a rigorous factorization scheme to divide each diagram into approximately independent pieces in a way that is consistent with the existing literature.

A possible factorization scheme is illustrated on the last row of Fig. 1(a) for the three-way junction. First, as discussed earlier, the free segments on the 5' and 3' ends of the chain do not incur any entropic costs. In the factored diagram, the two open arcs representing these two termini have been eliminated. Second, the loops labelled ***d*** and ***e*** have been factored out from the composite arc ***a***-***b***-***c***. This factorization scheme is motivated by the fact that the hairpin loop on one end of each stem is largely isolated from the loops on the opposite end of the stem, except in the case where they make direct

contact with each other, such as in a pseudoknot. Otherwise, loops on opposite ends of a helix are largely agnostic of each other except for the fact that they are both on the same stem, so factoring the loops on the opposite ends of a stem into approximately independent parts seems to be justified, as long as there are no explicit constraints between them. In this sense, every vertex "insulates" a pair of arcs on one side of the vertex from another pair of arcs on the other side, facilitating this factorization. We note that this postulated independence is not exact but only approximate. The validity of this conjecture will be demonstrated by the simulation studies presented below, and the data will show this postulated independence is quite accurate.

While the factorization shown in the last row of Fig. 1(a) suggests the two hairpin loops ***d*** and ***e*** are largely independent of the three loops ***a***, ***b***, and ***c*** forming the three-way junction, the composite ***a***-***b***-***c*** loop cannot be factorized further. The reason is that each vertex only insulates a pair of arcs from another pair, and the ***a***-***b***-***c*** loop must be treated as interdependent.

Before going on to demonstrate how to factorize the other diagrams in Fig. 1, we turn to the theory of topology to try to show why vertices with four arcs going through them can be factorized, but those with only two cannot. For planar networks such as the ones shown in the third row of Fig. 1, a basic definition in topology for Eulerian circuits guarantees that the entire network of arcs connected only by even vertices (i.e. those with an even number of arcs going through them) could be traversed by a continuous closed path that traces over each arc once and only once. Conversely, if a close path can traverse a network over each arc once and only once, the vertices must all be even (31, 32). When expressed in the context of RNAs, this theorem simply expresses the obvious fact that a RNA, having a continuous backbone, must be able to traverse all the constraints on its folded structure; therefore, all vertices representing such constraints are necessarily even. Furthermore, if we factor the diagram in the third row of Fig. 1(a) into that on the last row, the requirement of backbone continuity remains intact because every even vertex ensures that there is a closed path on both sides of the vertex after it has been factored. Conversely, if we factor a diagram and find that one or more of the elements in the resulting diagram can no longer be traversed by a closed path, then chain connectivity has been violated and such factorization is illegitimate. Thus, the fewest number of edges that must be connected to a vertex to ensure that each subgraph maintains backbone continuity is two, and vertices with only two arcs cannot be factored further as this is equivalent to splitting the helix along its length. By this, we see that further factoring the ***a***-***b***-***c*** loop in the last row of Fig. 1(a) is impossible because that would necessarily break one or more implicit constraints imposed by the continuity requirement of the RNA backbone. With this, it is easy to see that any part of a diagram that begins and ends on the same vertex can be factored out if and only if there is a close path that traverses all the arcs inside this part of the diagram once and only once. This is commonly referred to in graph theory as a circuit decomposition. Because of this, all self-contained peripheral loops, like those in the last row of Fig. 1(a), are factorizable from the rest of the diagram. Therefore, to facilitate the factorization of diagrams, it is convenient to introduce another topological feature called an "articulation point". An articulation point is any vertex which when removed separates the diagram into

two disjoint parts, each of which can be traversed by a closed path. The three vertices in Fig. 1(a) all represent articulation points.

Now going to the example of the pseudoknot in Fig. 1(b), we can first remove the two free ends producing the diagram in the last row of Fig. 1(b). But further factorization of this diagram is impossible because the two vertices are now both odd (i.e. having an odd number of arcs going through them). A theorem in topology states that for a network that has exactly two odd vertices, it can be traversed by exactly one path that begins on one of the vertex and ends on the other one. Further factorizing the diagram would violate the continuity requirement of the chain because neither of the two vertices are articulation points. Finally, for the triple helix in Fig. 1(c) and the quadruplex in Fig. 1(d), factorization leads to the diagrams on the last row. The results of these factorizations are analogous to the three-way junction in Fig. 1(a), producing multiple disjoint closed loops. Though the diagrammatic factorization would suggest that triple helices and quadruplexes have mostly independent loops, there is currently no data to support the factorization for Fig. 1(c) or 1(d). Thus, the factorizations suggested for Fig. 1(b), 1(c), and 1(d) are only conjectures. This work will focus on validating the factorization for multi-way junctions which all share the same topology as Fig 1(a). This will provide theoretical support to the long-standing assumption of factorizability for loops in secondary structure and serve as a lead into future studies that focus on the factorization of the more complex structures.

It should be noted that this separation of constraints into independent subsets and the subsequence factorization to be introduced is valid for the backbone conformational term, ΔS_b. There are terms in $\Delta G'$, particularly the electrostatics and the excluded volume interactions, that are not expected to factor due to the long-range nature of these forces. However, the intrinsic factorizability of the backbone conformational entropy term, ΔS_b, is unaltered. In future work, we will show how these other terms in $\Delta G'$ could be layered onto the backbone entropy term by interpolating between the graphical representation described in this paper and a fully 3D atomistic model.

Monte Carlo Simulation Studies

The factorization schemes introduced above for dividing constraints inherent from known secondary/tertiary structures of a RNA into approximately independent subsets were tested against large-scale Monte Carlo simulations. We simulated large ensembles of poly-U sequences with or without constraints to ascertain the interdependencies of different constraints corresponding to the ones that define hairpins with various loop lengths, as well as two-way, three-way and four-way junctions of different sizes.

The Monte Carlo (MC) simulations were carried out using our in-house Nucleic MC program based on the computational method described previously (33). The Nucleic program enables high-throughput atomistic MC simulations to be carried out for RNAs or DNAs by using a mixed numerical/analytical method to treat the sugar-phosphate backbone. Given positions and orientations of the bases, Nucleic uses a chain closure algorithm to sum over all possible backbone conformations arising from the torsional

degrees of freedom of the sugar-phosphate backbone for all nucleotide units on the chain (33-37). In the process, the summation takes into account steric interactions within all parts of the chain: between atoms in the sugar-phosphate backbone, between all bases in the side chains, and between the backbone and nucleobase side chain. Unlike molecular dynamics, Nucleic MC can sum over a massive number of backbone conformations with numerical efficiencies orders of magnitude faster, enabling a diverse ensemble of chain conformations to be generated rapidly. To further cut down on CPU requirements, Nucleic also uses high-level theoretical models (38-43) to represent the solvent's and the counterions' influences on the nucleic acid implicitly without the need of explicitly including solvent molecules and/or counterions in the simulation. Using our in-house parallel computing resources, a thermal ensemble consisting of several million uncorrelated chain conformations for RNA and DNA sequences up to a hundred nucleotides could be simulated in several days. The accuracy of Nucleic MC in terms of the chain structures that it produces has been fully validated in several studies (33-35).

For the present study, we simulated polyU chains of different lengths, with or without constraints. To focus our investigation exclusively on backbone entropic effects, we turned off all base stacking and base complementarity interactions except those explicitly dictated by the constraints during the simulations. The steric interactions, in keeping with our focus on entropic effects, is represented by the Weeks-Chandler-Andersen (WCA) potential (38). The WCA potential captures the repulsive branch of common two-body potentials such as Lennard-Jones and reflects lack of stabilization associated with base pairing and base stacking. Counterion-mediated forces are necessary to accurately mimic physiological ionic conditions, and we calibrated these interactions in our simulations to match the ambient ionic strength of an approximately 0.1 M NaCl solution (33, 39, 44).

Several series of simulations were carried out. These consisted of: (1) polyU chains with no constraints, to assess the entropic costs of hairpin loop initiations, (2) polyU chains with one internal constraint corresponding to a pre-formed hairpin loop in the interior of the sequence, to assess the entropic costs of initiating a second base-pair contact anywhere else along the chain, seeding the formation of either a two-way junction or a second hairpin loop, (3) polyU chains with two internal constraints corresponding to two pre-formed hairpin loops separated by a variable-length loop between them, to assess the entropic costs of initiating different three-way junctions of various sizes, and (4) polyU chains with three internal constraints corresponding to three pre-formed hairpin loops separated by two fixed-length loops, to assess the entropic costs of initiating a four-way junction. Entropic costs were evaluated by conducting a counting experiment on all MC frames produced by Nucleic MC. The number of times that a given pair of nucleotides—labeled as i and j—satisfies the base pairing constraints (vide infra) is collected and normalized by the total number of MC frames analyzed. This provides a probability of observing the nucleotides I and j in a configuration that satisfies the base pairing constraint, $P(i,j)$, within the thermal ensemble. The associated entropy cost is then calculated as

$$\Delta G = -k_B T \ln[P(i,j)] \quad (2)$$

All entropic costs in this work were calculated at 310K. While these simulations were designed to test the conjectures made above regarding the interdependencies of various constraints, the full thermodynamic data set presented below will also enable any researcher to easily calculate the backbone entropy costs of any known RNA fold. Care should be taken when using or referencing the reported values as they pertain only to the backbone entropy cost. Thus, the values should not be compared directly to experimental entropy values which have contributions from all parts of the system—the solvent for example. The reported entropy costs should ideally serve as a guide to determine trends in dependence and extrapolation into larger loop sizes at which point the backbone entropy tends to be the dominant contributor to the free energy. Alternatively, these values can also serve as a validity check towards studies of enthalpy as the sum of all non-entropy parts must offset, at a minimum, the backbone entropy costs reported within this work. Figure 2 shows a sample snapshot of a chain conformation from the MC simulations.

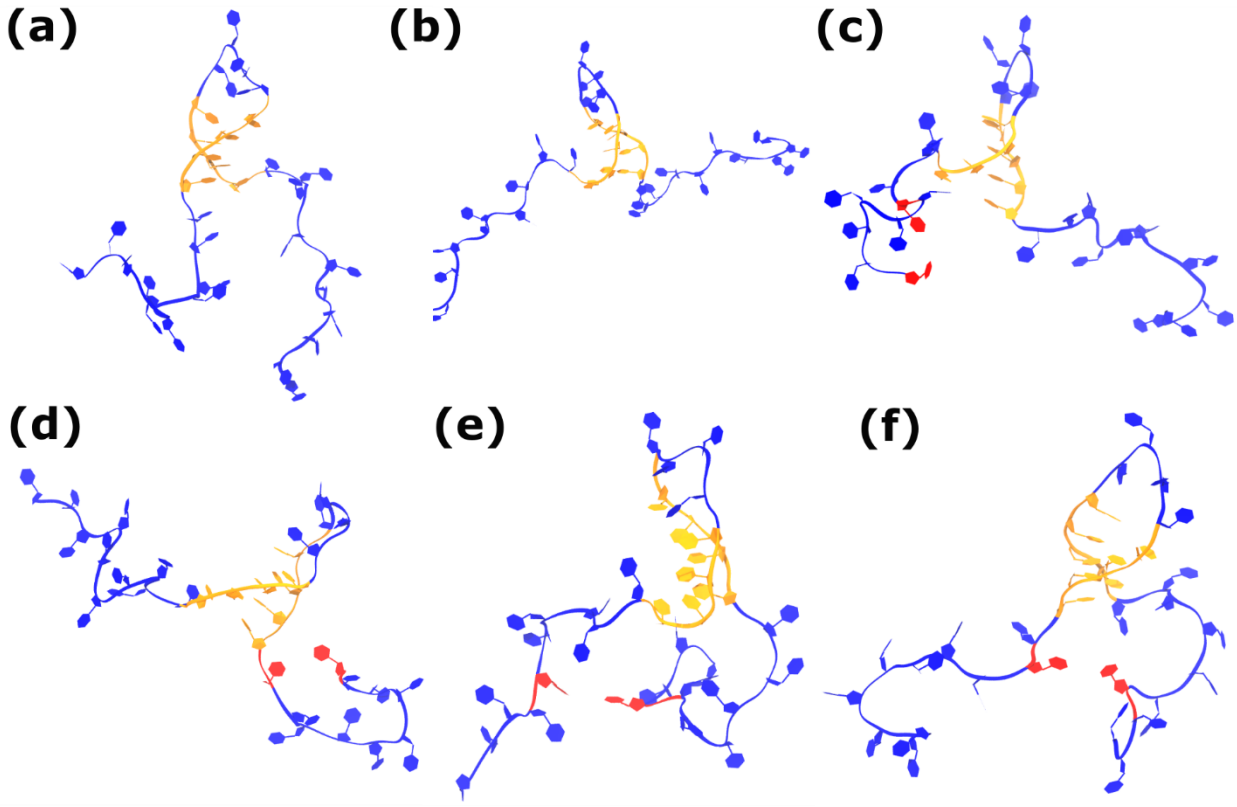

FIGURE 2.

Sample conformations obtained from the same starting constraints (in yellow) for a 34 nt poly-U chain. The newly formed base pair is in red. Conformations (a) and (b) show no newly formed base pairs. Conformations (c) and (d) show newly formed base pair initiating loops in the head and tail respectively. Conformations (e) and (f) show newly formed base pair creating internal junctions

RESULTS

Hairpin Loops

While U does not form canonical base pairs with itself, the entropic penalty necessary to put the sugar-phosphate backbone into a conformation ready to facilitate base pairing between them can be easily computed by counting the number of chain conformations that meet the conditions shown in the inset of Fig. 3 over the entire ensemble. This combination of N_b -N_b distance (9.0 ± 0.5 Å), virtual bond angles ($125 \pm 20°$) and virtual torsion angle ($0 \pm 40°$) between the two C1'-N_b glycosidic bonds of the two bases to be paired selects out base configurations which are in position to form an "ideal" complementary pair (http://ndbserver.rutgers.edu/) (45, 46). It should be noted that the choice of accepted values for the four base pairing criteria can be tightened or relaxed to match experimental geometries. As this determines the phase space volume that is associated with the constraints, the calculated entropy cost to form a structure will decrease as the range of accepted values for the criteria is increased and vice versa, so the entropy will have a constant offset depending on how the constraints are precisely defined. For an example, see Figure S1 in the supplemental information. Figure 3 shows the free energy $\Delta G = -T\Delta S_b$ at T = 310K for the spontaneous initiation of a hairpin loop of different lengths a anywhere along the sequence of a $(U)_{22}$ strand as the open circles. The loop initiation free energy increases smoothly from about 4.7 kcal/mol for a 3-nt hairpin loop to 6.6 kcal/mol for a 10-nt loop. The free energy for loop initiation starting at specific locations on the sequence are shown for several positions in Fig. 3, red (toward the 5' end) to violet (toward the 3' end). Experimental values are included in green. Loop initiation free energies seem to be slightly lower on the chain ends as they are expected to have more freedom, but only by a very small amount. Interior loops farther from the chain ends appear to be formed with roughly uniform probability along the entire sequence. Both the magnitude and loop-length dependence of these data compare well with the thermodynamic data reported by Turner and Mathew in green (https://rna.urmc.rochester.edu/NNDB/index.html) (20-23, 47) based on RNA melting experiments; most of the deviations are within 0.6kcal/mol (1k_B at 310K). The observed trends and deviations from experimental values collected in the work of Turner and Mathew match those obtained by prior simulation studies (29, 30, 48). As we are only investigating the parts of the free energy that comes from the backbone, the differences are expected and result from the experimental values capturing contribution from other terms besides the backbone. They are also consistent with previous MC data from our group using slightly different backbone closure parameters (34, 35).

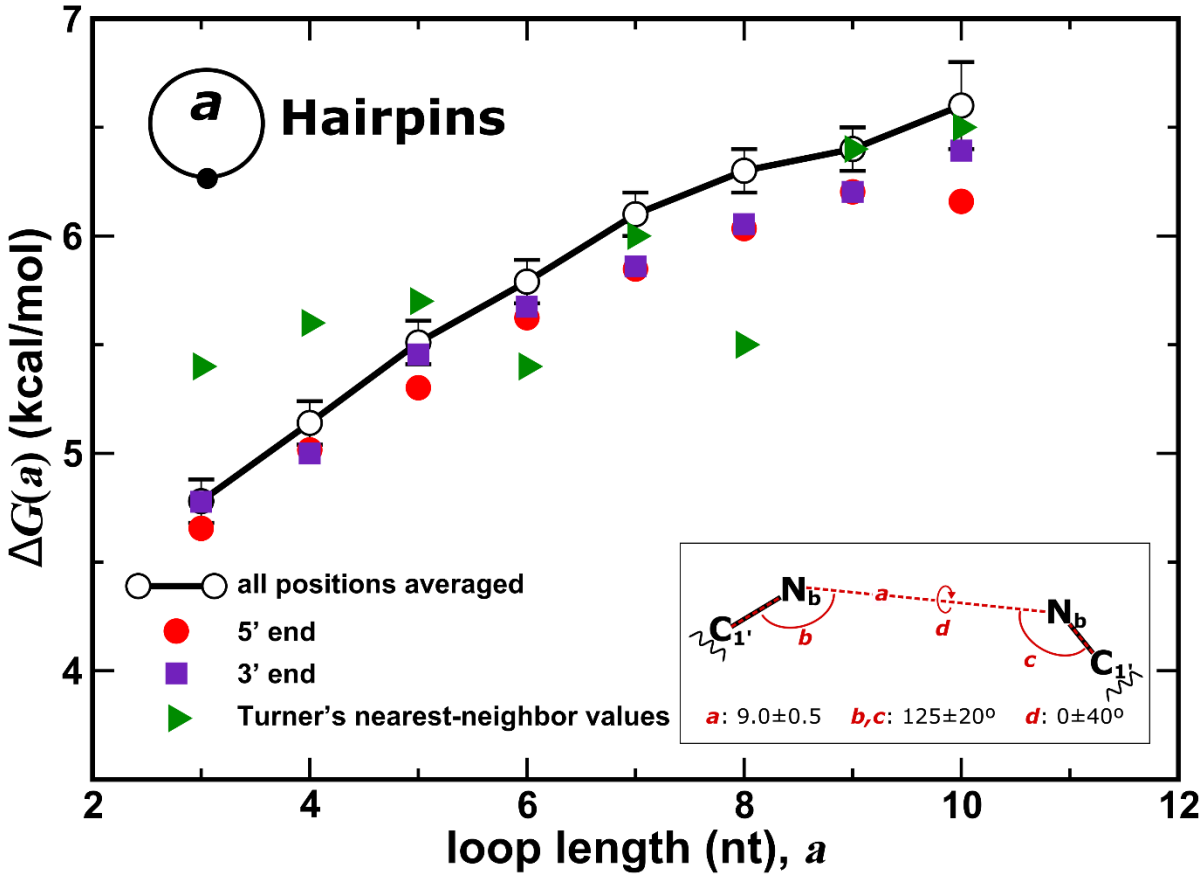

FIGURE 3.

Free energy cost due to conformational entropy loss at 310K for loop initiation in an unconstrained chain. The cost increases smoothly as a function of loop size (nt) with no significant position dependence along the sequence other than at the chain's ends where the cost decreases slightly. Experimental data for hairpin initiation obtained from melting experiments and aggregated in the nearest-neighbor model's database (21) have been included for comparison purposes. Error bars have been included for all points in the average value series. (Inset) the backbone geometric criteria used to define a base pair in the MC simulation. All parameters are chosen to put the C1'-N_b glycosidic bonds in the correct geometry to form a Watson-Crick pair.

Once a loop has been initiated, the helix can propagate by stacking more paired bases onto the first one. MC data show that the free energy cost due to backbone conformational entropy required for propagating the stem is 5.22 ± 0.03 kcal/mol per rung, in agreement with previous results (33). This value is independent of the length of the existing helix.

Initiation of a Second Hairpin

The formation of a second hairpin on a RNA strand that already contains one provides the first test for assessing whether the constraints associated with two side-by-side hairpins are independent. Figure 4

shows initiation free energy for the second hairpin as a function of its loop length. The open circles are loop initiation free energies for the first hairpin taken from Fig. 3. The green markers are initiation free energies for a second hairpin formed on the strand in which the first loop has a minimal stem length of 1 and the spacer length c is variable. The grayscale markers are initiation costs for a fix length of c and variable length in the stem of loop a. Fig. 4 shows that, to within statistical errors, the initiation of the second hairpin costs as much entropy as the first one of the same loop length b. This proves that the constraints associated with two side-by-side hairpins are indeed largely independent.

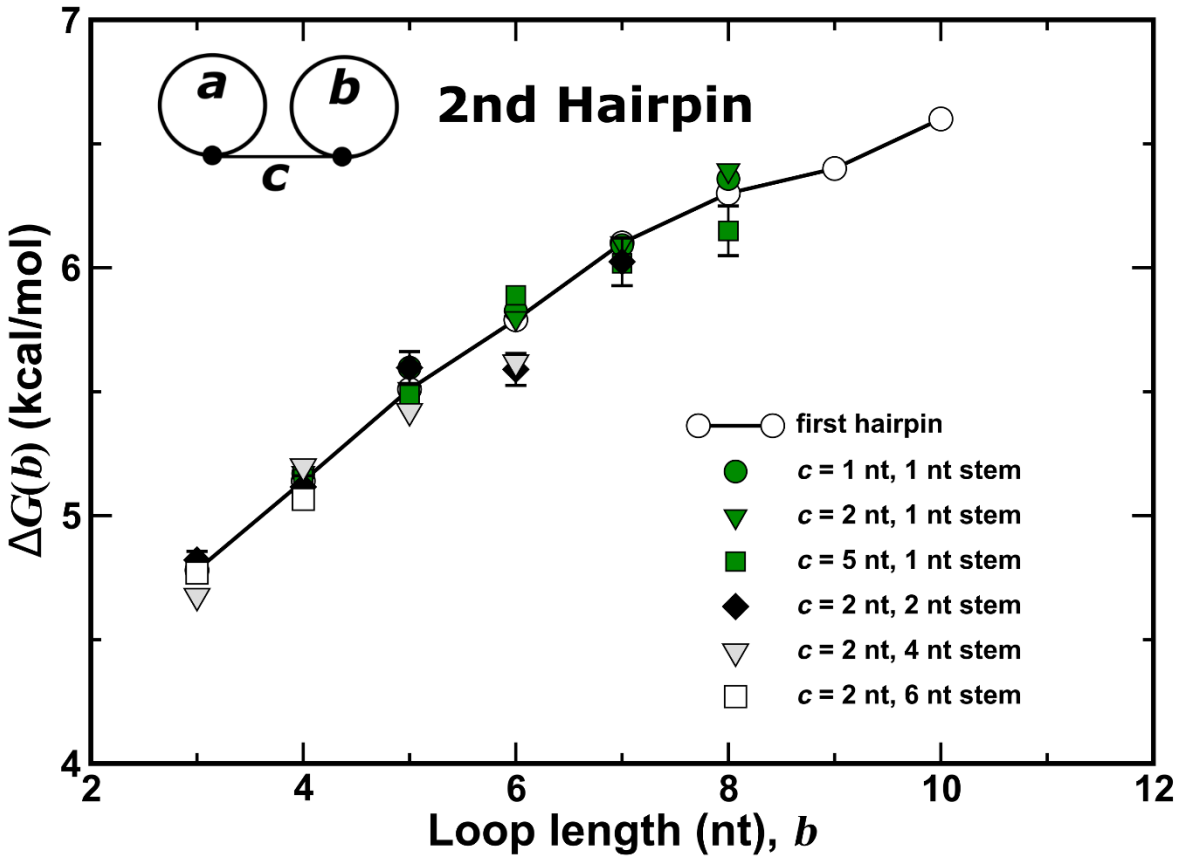

FIGURE 4.

Free energy cost at 310K to initiate a second loop of length b in a chain already containing a loop. The green symbols show the cost of the second loop b is independent of the spacer length c between it and the first loop a, which has a minimal stem length of 1. The grayscale symbols show the cost of loop b is independent of the stem length of loop a for a spacer length c = 2 nt. Other data showing similar independence for different spacer lengths c as well as the stem length on loop a are not presented. Note that error bars were included even though some of them are too small to be observed.

Two-way Junctions

The free energies for forming two-way junctions are shown in Fig. 5. In the topological representation of a two-way junction, depicted on the top left of Fig. 5, there are three relevant loop lengths: ***a*** is the length of the hairpin loop, ***b*** is the length of the junction on the 5' side, and ***c*** is the other junction on the 3' side. The dangling free ends of the chain are omitted as usual because they do not cost free energy. Figure 5 shows the additional free energy needed to initiate a two-way junction after the hairpin loop ***a*** is in place, as a function of the two junction lengths b and c in nt. Figure 5 illustrates that the free energy $\Delta G(b,c)$ is approximately the same when b and c are swapped, indicating that the initiation costs of a two-way junction is roughly symmetric with respect to the 5' and 3' junction lengths. The numerical values for $\Delta G(b,c)$ are tabulated in Table 1, with error estimates given in parentheses. While a precise comparison between the numerical values obtained from experiments versus simulations is difficult due to the fact that the simulations only accounted for the backbone entropy, the trend observed in our data are nevertheless similar to those from experiments used in constructing the nearest neighbor model. The entropic cost in general increases as the size of the loop $(b+c)$ grows and exhibits asymptotic behavior for sufficiently large loop size (21, 22, 48).

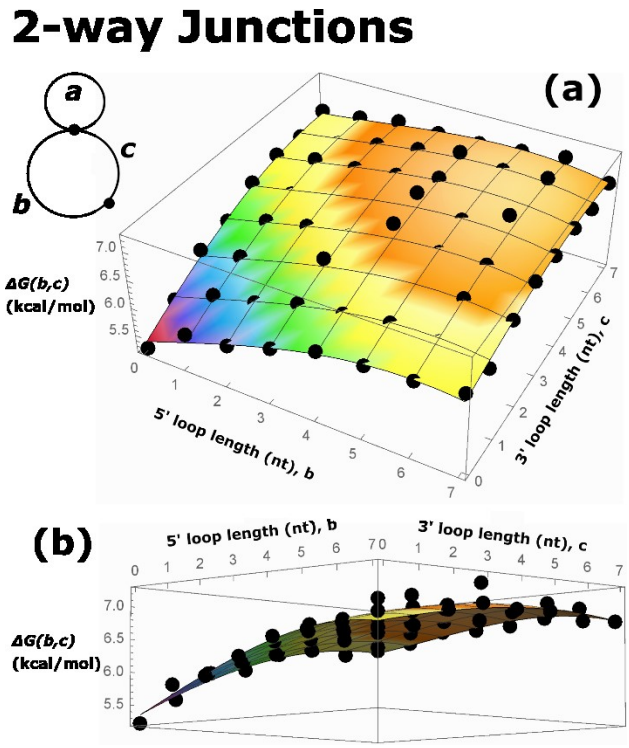

FIGURE 5.

The free energy costs of forming a two-way junction with 5' and 3' junction length b and c respectively given that a loop a is already in place. (a) Top view. (b) Side view. In general, the free energy cost grows as the junction size increases and is roughly symmetric when the 5' and 3' lengths are swapped.

5' loop length (nt), b	3' loop length (nt), c								
	0	1	2	3	4	5	6	7
0	5.22 (0.03)	5.62 (0.04)	6.07 (0.06)	6.16 (0.07)	6.45 (0.09)	6.57 (0.10)	6.53 (0.09)	6.65 (0.11)
1	5.77 (0.05)	5.98 (0.06)	6.20 (0.07)	6.40 (0.09)	6.47 (0.09)	6.70 (0.11)	6.61 (0.10)	6.72 (0.11)
2	5.86 (0.05)	6.12 (0.07)	6.27 (0.08)	6.55 (0.10)	6.58 (0.10)	6.75 (0.12)	6.79 (0.12)	6.85 (0.13)
3	6.11 (0.07)	6.38 (0.08)	6.68 (0.11)	6.62 (0.10)	6.73 (0.11)	6.85 (0.13)	6.92 (0.13)	6.83 (0.12)
4	6.36 (0.08)	6.48 (0.09)	6.55 (0.10)	7.08 (0.16)	7.18 (0.17)	7.02 (0.15)	7.08 (0.16)	6.85 (0.13)
5	6.42 (0.09)	6.61 (0.10)	6.55 (0.10)	6.83 (0.12)	6.94 (0.14)	6.79 (0.12)	6.97 (0.14)	6.92 (0.13)
6	6.47 (0.09)	6.58 (0.10)	6.79 (0.12)	6.87 (0.13)	7.26 (0.18)	6.89 (0.13)	7.05 (0.15)	7.02 (0.15)
7	6.58 (0.10)	6.42 (0.09)	6.70 (0.11)	6.85 (0.13)	6.81 (0.12)	6.83 (0.12)	6.73 (0.11)	6.77 (0.12)

TABLE 1.

Table of free energy cost of forming a two-way junction in kcal/mol as a function of the 5' and 3' junction lengths in nt, b and c, respectively. Error estimates from the simulation are given in parentheses.

Figure 6 shows how the two-way junction free energy depends on the loop length of the hairpin on the other side of the helix and the length of the stem itself. The conjecture that motivates our topological reduction scheme argues that they should be largely independent. Figure 6 plots the free energy of initiating a symmetric two-way junction (i.e. $b = c$) as a function of the junction size for a 4 nt hairpin loop with three different stem lengths (1 nt, 4 nt, and 6 nt), as well as a 6 nt hairpin loop with a 1 nt stem, and a 7 nt loop with a 1 nt stem. Clearly, the entropic costs for junction formation is independent of the hairpin on the other side of the constraint as well as the helix length. Note that the variation in costs for larger loop sizes is a natural result of the counting experiment. A higher entropic cost corresponds to a smaller number of recorded occurrences which is more heavily impacted by counting uncertainty. While not shown explicitly here, results for all two-way junctions, symmetric or asymmetric, demonstrate similar independence. Error bars are shown explicitly for a few data points to illustrate the size of the typical uncertainties.

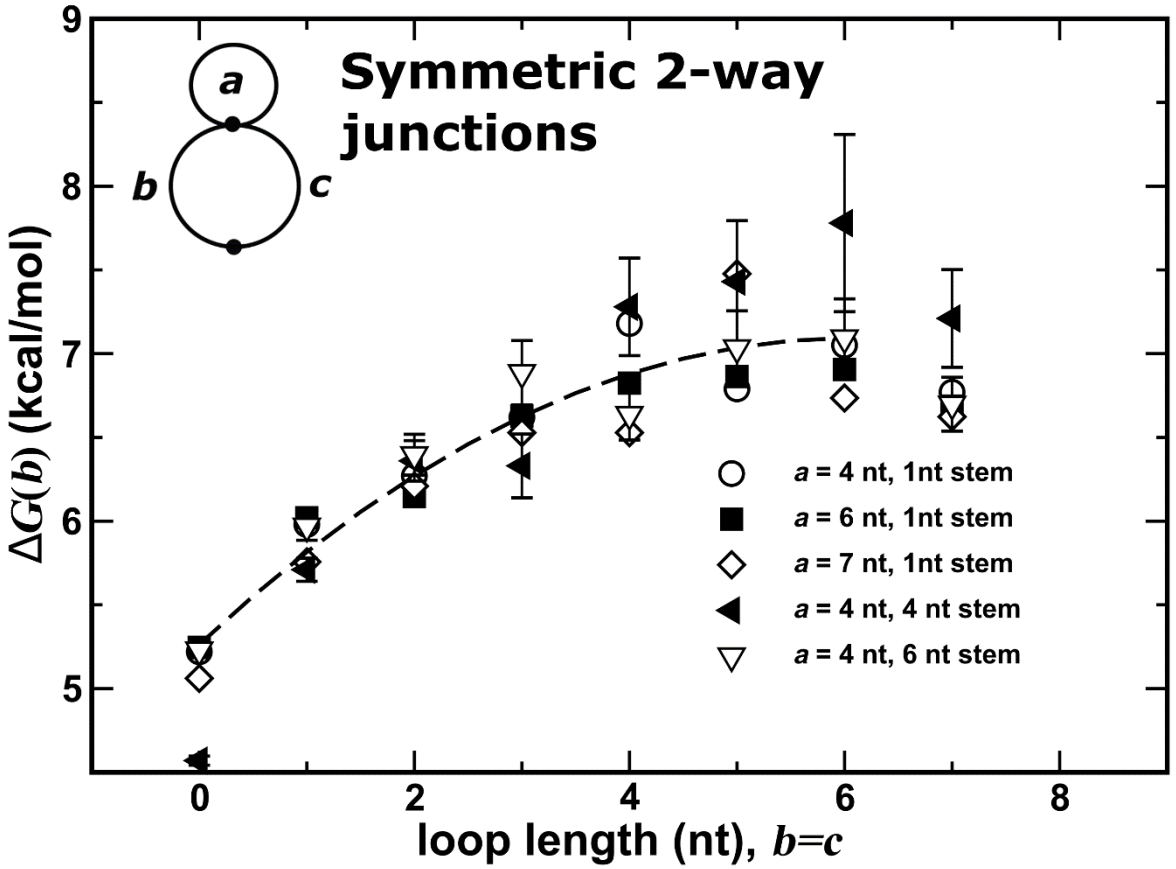

FIGURE 6.

The free energy cost of forming symmetric two-way junctions plotted for chains with different sizes of the first loop, a, and for different lengths of the stem separating a from the 2-way junction (b, c). Over the set of three values used for a, the free energy cost to close the junction are consistent with each other. This indicates that the two-way junction is dependent on only the two junction lengths b and c, but not the loop on the opposite side, a. Over the three different stem lengths, the cost to close the symmetric junction shows no discernible dependence on the length of the stem. Typical error bars for selected data points are included. The error for larger loop sizes can be attributed to errors in the counting experiment. Dash line is a guide to the eye.

Three-way Junctions

Three-way junctions are characterized by three different junction lengths as shown in Fig. 7. As in the case of two-way junctions, the free energy cost of initiating a three-way junction is largely independent of the hairpins on the opposite side of all three constraints. In Table 2, we tabulate the values of $\Delta G(a, b, c)$, where a is the length of the 5' junction, c is the length of the 3' junction, and b is the length of the junction in the middle. Only one value for b are shown in Table 2; data tables for all other values of b studied are included in the supplementary information. Not surprisingly, closing a three-way junction costs more free

energy than two-way junctions, but this additional cost is only marginal. Comparison of our data against experimental results shows some deviations; this is expected as the introduction of larger loops and more branching helices yields larger contribution to the experimental results from sources that are not included in our simulations such as sequence-dependent stabilization and coaxial stacking of helices. In terms of comparing against existing simulation results, we observed the same dependence on loop size and number of branching helices as Aalberts & Nandagopal (48). As the loop size increases the free energy cost increases. Additionally as the number of branching helices increases, there is an overall destabilizing effect that increases the cost for all loop sizes (48). This can be seen in the decreased range spanned by the entropy cost as we move from two-way junction to three- and four-way junctions. The trends are also similar to results obtained in other studies (29, 30), though our predicted entropic costs are somewhat higher. This difference most likely originates from the way in which each simulation handles the torsional motion of the backbone with the other studies using highly discretized models—diamond lattice for Cao & Chen (29) and discrete states configuration space for Zhang et al. (30)

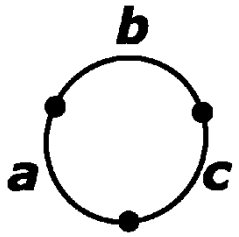

FIGURE 7.

Reduced topological representation of the set constraints defining a three-way junction. For the purposes of this study. For Table 2 and all sub-tables of Table S1, the value for b is fixed while a and c changing to give rise to the different sizes of three-way junctions.

center loop length b = 1 nt									
5'loop length (nt), a	3' loop length (nt), c								
	0	1	2	3	4	5	6	7
0	6.54 (0.10)	6.97 (0.15)	7.21 (0.18)	6.84 (0.13)	7.03 (0.16)	7.09 (0.17)	7.34 (0.21)	7.00 (0.15)
1	6.94 (0.14)	7.06 (0.16)	7.25 (0.19)	7.06 (0.16)	7.25 (0.19)	7.59 (0.27)	7.09 (0.17)	7.30 (0.20)
2	6.87 (0.13)	6.82 (0.13)	7.03 (0.16)	7.30 (0.20)	7.46 (0.23)	7.30 (0.20)	7.17 (0.18)	7.40 (0.22)
3	7.25 (0.19)	7.21 (0.18)	7.25 (0.19)	7.34 (0.21)	7.25 (0.19)	7.46 (0.23)	7.40 (0.22)	7.13 (0.17)
4	7.30 (0.20)	7.30 (0.20)	6.97 (0.15)	7.88 (0.37)	7.46 (0.23)	7.46 (0.23)	7.46 (0.23)	7.59 (0.27)
5	7.09 (0.17)	7.30 (0.20)	7.59 (0.27)	7.52 (0.25)	7.46 (0.23)	7.34 (0.21)	7.25 (0.19)	7.00 (0.15)
6	7.34 (0.21)	7.40 (0.22)	7.25 (0.19)	7.40 (0.22)	7.34 (0.21)	7.34 (0.21)	7.30 (0.20)	7.21 (0.18)
7	6.97 (0.15)	7.09 (0.17)	6.94 (0.14)	7.17 (0.18)	7.06 (0.16)	7.25 (0.19)	7.46 (0.23)	7.13 (0.17)

TABLE 2.

Table of free energy costs of forming a three-way junction in kcal/mol as a function of the 5' and 3' junction length in nucleotide (a and c respectively) with the centre junction length (b) as a parameter. For $b = 0$ and $b \geq 2$, see Table S1 in the supplemental material. Error estimates from the simulation are given in parentheses. Entries which have "inf" errors were too infrequently observed during the simulation for errors to be accurately calculated.

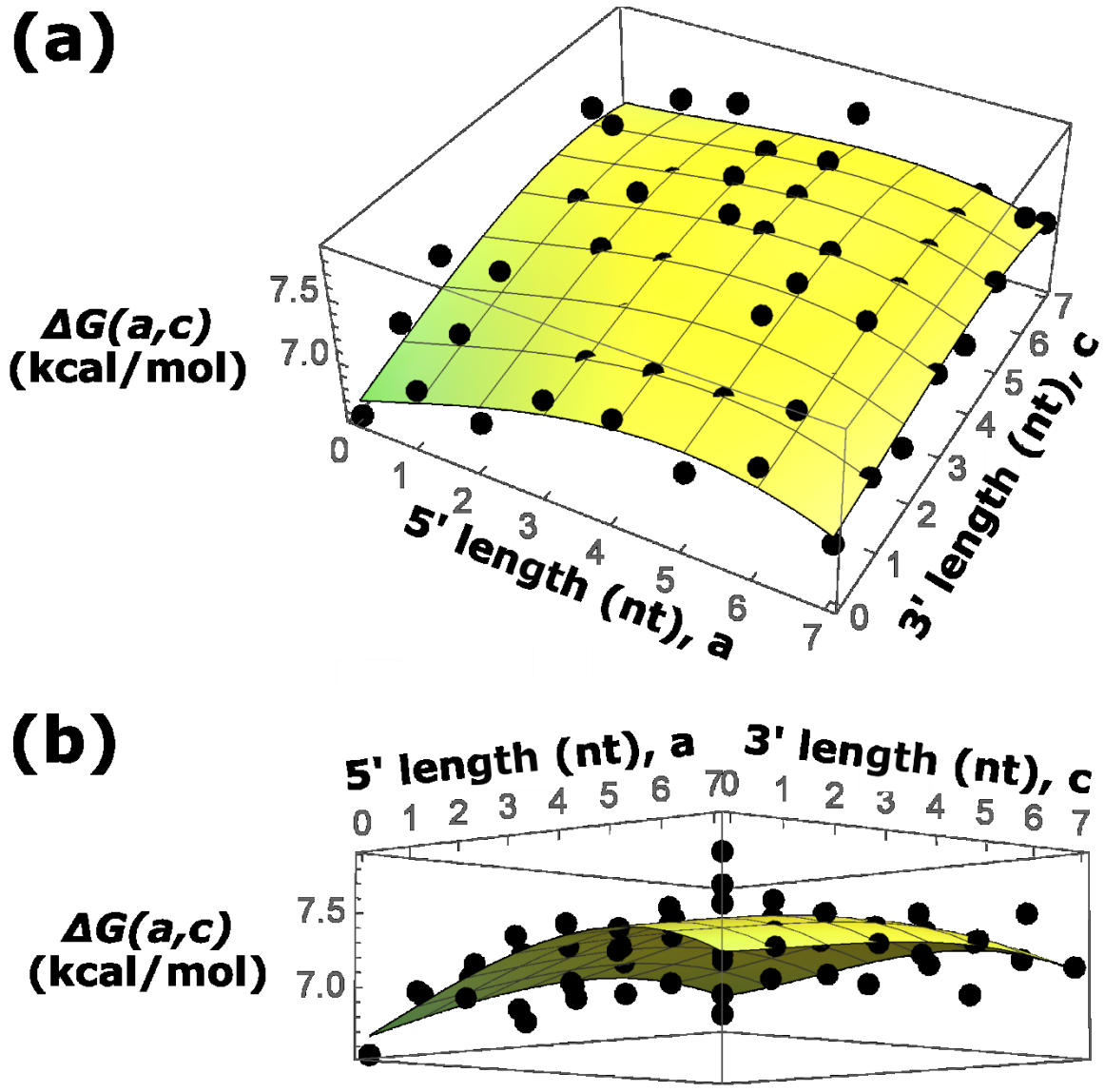

FIGURE 8.

The free energy costs of forming a three-way junction with 5' and 3' junction length a and c respectively given that junction length b is fixed at 1 nt; this surface corresponds to the data given in table 2 above. (a) Top view. (b) Side view.

Initiation of a Third Hairpin

Figure 9 shows the free energy for initiating a third hairpin ***c*** after two others (***a*** and ***b***) have been formed, as a function of loop length c in nt. The open circles are initiation free energy for the first hairpin taken from Fig. 3. Red circles show hairpin initiation on the 5' side of loop ***a***. Violet squares show hairpin initiation on the 3' side of loop ***b***, and green diamonds show hairpin initiation on the strand between ***a*** and ***b***. Analogous to the results for the initiation of a second hairpin shown in Fig. 3, the third hairpin is largely

independent of the first two. The segment length between any two hairpins in this set of data varies from 0 to 4 nt.

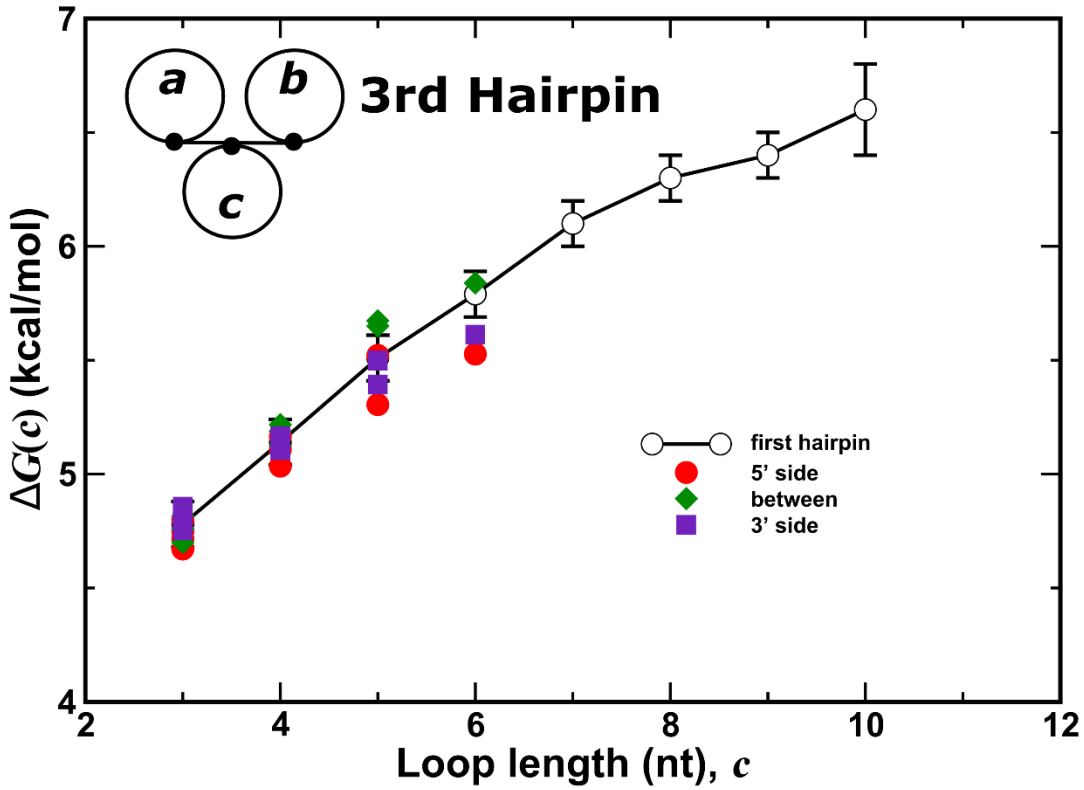

FIGURE 9.

The free energy cost of initiating a third hairpin of length c in the presence of two existing loops (a and b). When compared against the cost of initiating a hairpin loop on the free chain, the cost of the third loop is comparable and shows no dependence on the location of the new loop relative to the existing loops. This suggests that independence of hairpin loops can be extended to any number of loops within a chain. Note that error bars were included for the average cost of the first hairpin like in Fig.3; some of them are not visible due to their size.

Four-way Junctions

Figure 10 shows the reduced topological representation of a four-way junction, with the loop on the other side of every hairpin having been factored out. The free energy of formation of a four-way junction is a function of the four junction lengths a, b, c and d. Initiation free energies for an example of a four-way junction are tabulated in Table 3, for one particular combination of junction lengths b = c = 4 nt. Data shown are the additional free energy cost for the fourth constraint to be met after the first three constraints are in place. To obtain this data set, an ensemble of 2 million MC simulated conformations of $(U)_{42}$ chains was used. Free energies in Table 3 show that closing a four-way junction generally costs more entropy than a three-way junction (see Table 2), which in turn costs more entropy than two-way

junctions. Again, error estimates are given in parentheses. The error bars are a little larger than for the two- and three-way junctions because the probability of observing a four-way junction was quite low. In Table 3, cells that are blank indicate combinations that failed to show up in the 2-million-member MC simulated ensemble.

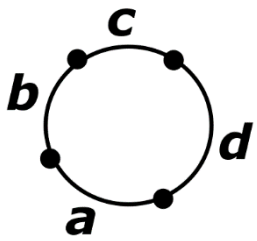

FIGURE 10.

Reduced topological representation of the set constraints defining a four-way junction. For the purposes of this study, the two of the lengths were constrained to be equal and fixed in value $(b = c = 4$ nt) while the other lengths (a and d) can vary.

Four-way junctions with center loops with lengths b = c = 4 nt									
5' loop length (nt), a		3' loop length (nt), d							
		0	1	2	3	4	5	6	7
	0	7.85 (0.43)	7.85 (0.43)	7.60 (0.32)	8.03 (0.53)	8.27 (0.76)	8.03 (0.53)	---	7.42 (0.27)
	1	7.71 (0.37)	8.03 (0.53)	8.27 (0.76)	8.03 (0.53)	8.27 (0.76)	8.03 (0.53)	---	7.85 (0.43)
	2	8.03 (0.53)	8.03 (0.53)	7.85 (0.43)	8.27 (0.76)	7.71 (0.37)	8.27 (0.76)	7.71 (0.37)	7.60 (0.32)
	3	8.03 (0.53)	7.85 (0.43)	7.85 (0.43)	7.85 (0.43)	---	8.03 (0.53)	7.71 (0.37)	7.85 (0.43)
	4	7.85 (0.43)	7.71 (0.37)	7.85 (0.43)	8.27 (0.76)	7.85 (0.43)	8.27 (0.76)	8.03 (0.53)	7.85 (0.43)
	5	7.71 (0.37)	8.03 (0.53)	7.85 (0.43)	---	7.71 (0.37)	8.03 (0.53)	7.71 (0.37)	7.42 (0.27)
	6	7.71 (0.37)	7.71 (0.37)	8.27 (0.76)	---	7.71 (0.37)	7.60 (0.32)	7.50 (0.29)	7.60 (0.32)
	7	7.50 (0.29)	7.85 (0.43)	7.60 (0.32)	7.42 (0.27)	7.71 (0.37)	8.03 (0.53)	7.85 (0.43)	7.85 (0.43)

TABLE 3.

Table of free energy cost of forming a four-way junction in kcal/mol as a function of the 5' and 3' junction length in nucleotide (a and d respectively) with the middle junction lengths fixed $(b = c = 4$ nt$)$. Error estimates from the simulation are given in parentheses. Blank entries correspond to events that were not observed during the simulation despite the large size of the ensemble generated.

DISCUSSION

The topological representation we have developed above has been used to aide in the factorization of the joint constraints imposed by typical RNA secondary structure motifs into approximately independent subsets. Here, we discuss the broader application of this scheme.

First, using the topological reduction scheme and data presented above, calculating the total free energy cost arising from backbone conformational constraints associated with any structure is simple. Using the three-way junction from Fig. 1(a) as an example, we will illustrate this procedure for junction lengths a = 6, b = 4, c = 5, d = 3, e = 6 nt, with one of the two stems having f base pair steps and the other having g. From Fig. 3, the free energy for seeding hairpin loops d = 3 nt and e = 6 nt are 4.8 and 6.6 kcal/mol, respectively. The cost for propagating a seeded hairpin is 5.2 kcal/mol/base-pair-step, so the free energy associated with the two stems combined is $5.2 \times (f + g)$ kcal/mol. From Table S1(d), the free energy for a 6-4-5 three-way junction is 7.4 kcal/mol. The total is therefore $18.8 + 5.2(f + g)$ kcal/mol.

Topological reduction can also be used to analyze the interdependence of more complex constraints coming from tertiary contacts. An example is shown in Fig. 8. Many riboswitches, such as the guanine-responsive riboswitch from *the xpt-pbuX* operon of *B. subtilis* (49) and TPP-specific riboswitch of *Arabidopsis thaliana* (50), make use of a three-way junction architecture to form their aptamer domain. When the aptamer binds its target ligand, additional constraints arising from the reconfiguration of the binding pocket either destabilize existing tertiary interactions or stabilize addition tertiary contacts, leading to a rearrangement of the folded structure causing an upstream or downstream switching sequence to rehybridize and produces a global shape transformation in the riboswitch RNA (51-53). Figure 10 shows how some of these interactions renormalize the topology of a three-way junction.

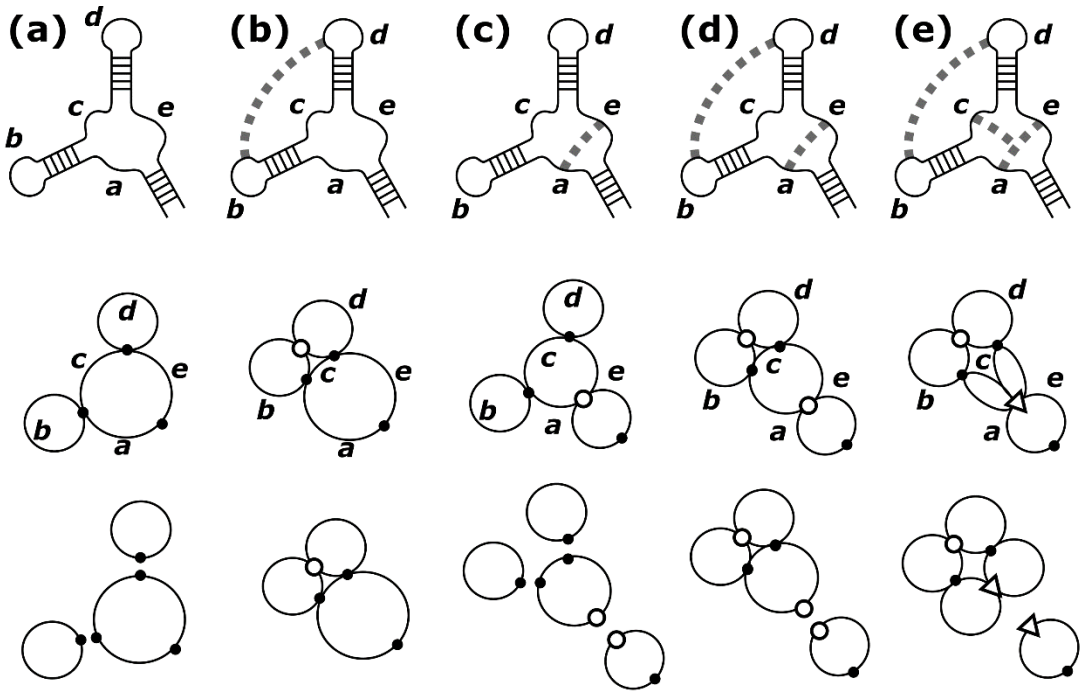

FIGURE 11.

Diagrammatic representation of the topology of a three-way junction and how it can be altered by introduction of new tertiary interactions. (a) An unmodified three-way junction like the one shown in Fig. 1(a). (b) Representation of kissing loops. The new constraint represented by the thick dash line in the top row of (b) results in a change in connectivity that no longer allows the two loops $\boldsymbol{b}$ and $\boldsymbol{d}$ to be factored. (c) Representation of ligand-mediated base-base contact in the three-way junction. The new constraint closes a portion of the three-way junction into a loop, giving rise to a diagram that is factorizable into 4 independent subsets corresponding to two hairpins, one two-way junction, and one three-way junction. (d) The kissing loop and ligand-mediated base-base interaction are combined. The effect changes the connectivity to yield a factorizable diagram consisting of a two-way junction and the structure previously seen in (b). (e) The kissing loop interaction is now combined with a triple base interaction. This yields a new structure factorizable into a two-way junction and a new multiply-connected loop structure.

The top row of Fig. 11(a) shows the same three-way junction architecture from Fig. 1(a) without tertiary contacts. The second row in Fig. 11(a) shows its topological representation and the third row shows the final factorized diagram from Fig. 1(a). As described above, without tertiary contacts the two hairpin loops and the junctions are largely independent, and from this we derive three disjoint sets of constraints. Now consider the addition of a kissing-loop interaction, denoted in Fig. 11(b) by a thick dashed line, between hairpins ***b*** and ***d***. The topological representation of this structure is shown in the second row of Fig. 11(b), where the constraint imposed by the kissing-loop interaction is represented by a white circle. Due to this extra constraint, this structure is no longer factorizable because it contains no articulation points. Therefore, the kissing-loop interaction modifies the topological structure of the diagram

fundamentally. In the language of topology, this diagram now belongs to a different “class” than the diagram in Fig. 11(a). This new nonfactorizable topological class is shown on the bottom row of Fig. 11(b).

In Fig. 11(c), a different tertiary interaction is introduced into the three-way junction. The dashed line in the top row of Fig. 11(c) denotes a new base-base contact between two of the junctions mediated by a ligand upon binding. The topological representation of this structure is shown in the second row of Fig. 11(c), and complete factorization leads to the diagram on the bottom row of Fig. 11(c). In this case, the two loops ***b*** and ***d*** corresponding to the hairpins remain factorizable, but the new interaction between loops ***a*** and ***e*** renormalizes the diagram into a different topological class. The final factorized representation, shown on the bottom row of Fig. 11(c), is topologically equivalent to two hairpin loops, one two-way junction, and one three-way junction.

The structure in Fig. 11(d) combines a kissing-loop tertiary contact between ***b*** and ***d*** with a base-base tertiary interaction between ***a*** and ***e***. The final factorized diagram is shown on the bottom row of Fig. 11(d), consisting of one two-way junction, plus three multiply-connected loops, which happens to belong to the same topological class as the structure in Fig. 11(b).

Finally, Fig. 11(e) introduces a new type of tertiary interaction. The thick three-way dashed line in Fig. 11(e) denotes a triple base interaction, such as the one observed in the crystallographic structure of the G-box riboswitch when a guanine is bound into the aptamer domain. The ligand forms contacts simultaneously with three bases, leading to a triplet interaction. Figure 11(e) considers the topological renormalization that is produced by mixing a kissing-loop interaction between hairpins ***b*** and ***d*** with a base-triple interaction among junctions ***a***, ***c*** and ***e***. The final factorized diagram is shown in the bottom row of Fig. 11(e). This diagram suggests that the structure in Fig. 11(e) is topologically equivalent to one two-way junction, plus four mutually connected loops. This results also explains how riboswitches based on a three-way junction motif might utilize tertiary interactions coming from ligand binding to induce loop-loop interactions in distal regions of its RNA sequence.

We conclude by mentioning one useful property of factorizable diagrams. After complete factorization, each disjoint piece consists of a self-contained substructure that traces out a close circuit beginning with an initial vertex and ending on the same vertex, traversing every arc inside the substructure once and only once. For each of these substructures, a basic theorem in topology states that the choice of the initial vertex is arbitrary, and the choice of the first arc to follow to start the circuit is also arbitrary. This means that when calculating the entropy of a substructure, the answer does not depend on which constraint (i.e. vertex) to start and end with. On the other hand, for substructures that do not begin and end on the same vertex, such as the one in Fig. 1(b), they must have exactly two odd vertices. There is only one way to traverse the entire path through such structures, which is by starting on one of the odd vertices and ending on the other one.

These examples in this and the last sections show how our proposed topological perspective of RNA structures could lead to new insights into the interplay among multiple constraints inherent in the

secondary and tertiary structures of folded RNAs. Work is currently in progress to generate data for the entropic penalties of a library of tertiary contacts as well as for pseudoknots.

By extending our study to more complex secondary structures such as those in Fig. 1(c) and 1(d), we should be able to examine the validity of the factorization hypothesis on more complex elements and evaluate their entropic costs from simulation. This can then be used to study more complex tertiary folds by mapping the 3D structure to the corresponding 2D graphs which we can separate into the independent subsets to calculate their entropic costs. Using our atomistic simulations, we can also reconstitute 3D structures from 2D graphs with defined constraints, and this would open new ways to study RNA structures within the space of all possible 2D graphs, providing a rigorous strategy to interpolate between 2D and 3D structures, forming the foundation for a large-scale Monte Carlo simulation algorithm for RNA tertiary structures.

CONCLUSION

In this paper, we take a fresh look at how to interpret the various types of secondary and tertiary structural motifs encountered in typical RNA folds from the point of view of graph theory. We have proposed a diagrammatic scheme to quantify the entropic penalty imposed on the sugar-phosphate backbone of a folded RNA coming from constraints imposed by the secondary and/or tertiary contacts needed to stabilize the fold. Among the various terms in the folding free energy, the free energy coming from entropy depression due to the loss of backbone conformational freedom is the only term that is guaranteed to be always uphill, and as such, it provides a rigorous lower bound on the magnitudes of all the other free energy contributors that must act to stabilize the fold. Whether folding occurs locally via domains or cooperatively can also be resolved by examining the free energy balance within each domain against backbone entropic costs.

A simple diagrammatic device is designed to help factor the many secondary- and tertiary-constraints typically seen in folded RNAs into approximately independent sets, in order to separate the backbone entropy into additive parts. This new approach generates an interesting and intuitive topological view of RNA structures. We further show how topological reduction can be carried out for typical secondary and tertiary structure motifs, and comparing the results of the reduction against large-scale Monte Carlo simulations of equilibrium ensembles of different RNA constructs in solution, we demonstrate the accuracy and the usefulness of the topological perspective. Extensive data sets and simple recipes are provided in the paper to enable any RNA scientist to easily estimate the magnitude of backbone entropy depression due to common RNA secondary motifs such as hairpin loops and multi-way junctions. Studies quantifying the conformational entropic penalties arising from pseudoknots as well as longer-range tertiary interactions are underway.

AUTHOR CONTRIBUTIONS

CHM designed the study. CHM and ENHP carried out the work. CHM wrote the manuscript.

ACKNOWLEDGEMENTS

This material is based in part upon work supported by the National Science Foundation under Grant Numbers CHE-0713981 and CHE-1664801.

REFERENCES

1. Draper, D. E., D. Grilley, and A. M. Soto. 2005. Ions and RNA Folding. Annu. Rev. Biophys. Biomol. Struct. 34:221-243.
2. Wong, G. C. L., and L. Pollack. 2010. Electrostatics of Strongly Charged Biological Polymers: Ion-Mediated Interactions and Self-Organization in Nucleic Acids and Proteins. Annu. Rev. Phys. Chem. 61:171-189.
3. Chen, S.-J. 2008. RNA Folding: Conformational Statistics, Folding Kinetics, and Ion Electrostatics. Annu. Rev. Biophys. 37:197-214.
4. Liu, L., and S.-J. Chen. 2010. Computing the conformational entropy for RNA folds. J. Chem. Phys. 132:235104.
5. Woodson, S. A. 2010. Compact intermediates in RNA folding. Annu. Rev. Biophys. 39:61-77.
6. Turner, D. H. 1996. Thermodynamics of base pairing. Curr. Opin. Struct. Biol. 6:299-304.
7. Chandler, D. 1987. Introduction to Modern Statistical Mechanics. Oxford University Press, New York, NY.
8. Hill, T. L. 2013. Statistical mechanics: principles and selected applications. McGraw-Hill, New York, NY.
9. Ding, F., S. Sharma, P. Chalasani, V. V. Demidov, N. E. Broude, and N. V. Dokholyan. 2008. Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. RNA 14:1164-1173.
10. Ding, F., C. A. Lavender, K. M. Weeks, and N. V. Dokholyan. 2012. Three-dimensional RNA structure refinement by hydroxyl radical probing. Nat. Methods 9:603.
11. De Gennes, P.-G. 1979. Scaling concepts in polymer physics. Cornell University Press, Ithaca, NY.
12. Flory, P., and M. Volkenstein. 1969. Statistical Mechanics of Chain Molecules. Interscience Publishers, New York, NY.
13. Le, S.-Y., R. Nussinov, and J. V. Maizel. 1989. Tree graphs of RNA secondary structures and their comparisons. Comput. Biomed. Res. 22:461-473.
14. Schmitt, W. R., and M. S. Waterman. 1994. Linear trees and RNA secondary structure. Discrete Apl. Math. 51:317-323.
15. Shapiro, B. A., and K. Zhang. 1990. Comparing multiple RNA secondary structures using tree comparisons. Bioinformatics 6:309-318.
16. Laing, C., and T. Schlick. 2011. Computational approaches to RNA structure prediction, analysis, and design. Curr. Opin. Struct. Biol. 21:306-318.
17. Kim, N., C. Laing, S. Elmetwaly, S. Jung, J. Curuksu, and T. Schlick. 2014. Graph-based sampling for approximating global helical topologies of RNA. Proc. Natl. Acad. Sci. USA 111:4079-4084.
18. Gan, H. H., D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick. 1987. RAG: RNA-As-Graphs database—concepts, analysis, and features. Nutr. Health 5:1285-1291.
19. Fera, D., N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. H. Gan, and T. Schlick. 2004. RAG: RNA-As-Graphs web resource. BMC Bioinf. 5:88.
20. Mathews, D. H., J. Sabina, M. Zuker, and D. H. Turner. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288:911-940.
21. Turner, D. H., and D. H. Mathews. 2009. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res. 38:D280-D282.
22. Mathews, D. H., and D. H. Turner. 2006. Prediction of RNA secondary structure by free energy minimization. Curr. Opin. Struct. Biol. 16:270-278.

23. Diamond, J. M., D. H. Turner, and D. H. Mathews. 2001. Thermodynamics of three-way multibranch loops in RNA. Biochemistry 40:6971-6981.
24. Zuker, M. 1989. [20] Computer prediction of RNA structure. In Methods Enzymol. Academic Press. 262-288.
25. Zuker, M., D. H. Mathews, and D. H. Turner. 1999. Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide. In RNA Biochemistry and Biotechnology. J. Barciszewski, and B. F. C. Clark, editors. Springer Netherlands, Dordrecht, Netherlands. 11-43.
26. Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406-3415.
27. Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. Science 244:48-52.
28. Zadeh, J. N., C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce. 2011. NUPACK: Analysis and design of nucleic acid systems. J. Comput. Chem. 32:170-173.
29. Cao, S., and S.-J. Chen. 2005. Predicting RNA folding thermodynamics with a reduced chain representation model. RNA 11:1884-1897.
30. Zhang, J., M. Lin, R. Chen, W. Wang, and J. Liang. 2008. Discrete state model and accurate estimation of loop entropy of RNA secondary structures. J. Chem. Phys. 128:03B624.
31. Arnold, B. H. 2011. Intuitive concepts in elementary topology. Dover Publications, Mineola, N.Y.
32. Balakrishnan, R., and K. Ranganathan. 2012. A textbook of graph theory. Springer Science & Business Media, New York, NY.
33. Mak, C. H. 2015. Atomistic Free Energy Model for Nucleic Acids: Simulations of Single-Stranded DNA and the Entropy Landscape of RNA Stem–Loop Structures. J. Phys. Chem. B 119:14840-14856.
34. Mak, C. H., L. L. Sani, and A. N. Villa. 2015. Residual Conformational Entropies on the Sugar–Phosphate Backbone of Nucleic Acids: An Analysis of the Nucleosome Core DNA and the Ribosome. J. Phys. Chem. B 119:10434-10447.
35. Mak, C. H., T. Matossian, and W.-Y. Chung. 2014. Conformational entropy of the RNA phosphate backbone and its contribution to the folding free energy. Biophys. J. 106:1497-1507.
36. Mak, C., W.-Y. Chung, and N. D. Markovskiy. 2011. RNA conformational sampling: II. Arbitrary length multinucleotide loop closure. J. Chem. Theory Comput. 7:1198-1207.
37. Mak, C. 2008. RNA conformational sampling. I. Single‐nucleotide loop closure. J. Comput. Chem. 29:926-933.
38. Weeks, J. D., D. Chandler, and H. C. Andersen. 1971. Role of repulsive forces in determining the equilibrium structure of simple liquids. J. Chem. Phys. 54:5237-5247.
39. Henke, P. S., and C. H. Mak. 2014. Free energy of RNA-counterion interactions in a tight-binding model computed by a discrete space mapping. J. Chem. Phys. 141:08B612_611.
40. Mak, C., and P. S. Henke. 2012. Ions and RNAs: free energies of counterion-mediated RNA fold stabilities. J. Chem. Theory Comput. 9:621-639.
41. Mak, C. H. 2016. Unraveling Base Stacking Driving Forces in DNA. J. Phys. Chem. B 120:6010-6020.
42. Rury, A. S., C. Ferry, J. R. Hunt, M. Lee, D. Mondal, S. M. O. O'Connell, E. N. H. Phan, Z. Peng, P. Pokhilko, D. Sylvinson, Y. Zhou, and C. H. Mak. 2016. Solvent Thermodynamic Driving Force Controls Stacking Interactions between Polyaromatics. J. Phys. Chem. C 120:23858-23869.
43. Hummer, G., S. Garde, A. E. Garcia, A. Pohorille, and L. R. Pratt. 1996. An information theory model of hydrophobic interactions. Proc. Natl. Acad. Sci. USA 93:8951-8955.
44. Henke, P. S., and C. H. Mak. 2016. An implicit divalent counterion force field for RNA molecular dynamics. J. Chem. Phys. 144.

45. Coimbatore Narayanan, B., J. Westbrook, S. Ghosh, A. I. Petrov, B. Sweeney, C. L. Zirbel, N. B. Leontis, and H. M. Berman. 2013. The Nucleic Acid Database: new features and capabilities. Nucleic Acids Res. 42:D114-D122.
46. Berman, H. M., W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. Srinivasan, and B. Schneider. 1992. The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys. J. 63:751-759.
47. Serra, M. J., and D. H. Turner. 1995. [11] Predicting thermodynamic properties of RNA. Methods Enzymol. 259:242-261.
48. Aalberts, D. P., and N. Nandagopal. 2010. A two-length-scale polymer theory for RNA loop free energies and helix stacking. RNA 16:1350-1355.
49. Batey, R. T., S. D. Gilbert, and R. K. Montange. 2004. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. Nature 432:411-415.
50. Thore, S., M. Leibundgut, and N. Ban. 2006. Structure of the eukaryotic thiamine pyrophosphate riboswitch with its regulatory ligand. Science 312:1208-1211.
51. Manzourolajdad, A., and J. Arnold. 2015. Secondary structural entropy in RNA switch (Riboswitch) identification. BMC Bioinf. 16:133.
52. Roth, A., and R. R. Breaker. 2009. The structural and functional diversity of metabolite-binding riboswitches. Annu. Rev. Biochem. 78:305-334.
53. Montange, R. K., and R. T. Batey. 2008. Riboswitches: emerging themes in RNA structure and function. Annu. Rev. Biophys. 37:117-133.

TABLE AND FIGURES LEGENDS

TABLE 1.

Table of free energy cost of forming a two-way junction in kcal/mol as a function of the 5' and 3' junction lengths in nt, b and c, respectively. Error estimates from the simulation are given in parentheses.

TABLE 2.

Table of free energy costs of forming a three-way junction in kcal/mol as a function of the 5' and 3' junction length in nucleotide (a and c respectively) with the centre junction length (b) as a parameter. For $b = 0$ and $b \geq 2$, see Table S1 in the supplemental material. Error estimates from the simulation are given in parentheses. Entries which have "inf" errors were too infrequently observed during the simulation for errors to be accurately calculated.

FIGURE 1.

Various secondary structures, the total enumeration of the constraints that define them, and their conversion into a diagrammatic topological representation followed by factorization. (a) A three-way junction is defined by five single-stranded lengths and 3 helices. It is factored into 3 independent subsets which can be treated separately. (b) A pseudoknot is defined by three single-stranded length and two helices. Due to backbone connectivity, the diagram is not factorizable. (c) A triple helix is defined by two single-stranded loops and one triple helix structure. The factorization suggests the two loops are approximately independently of each other. (d) A quadruplex is defined by several loops threaded through the quadruplex core. The factorization shown here suggests that the three loops, after topological reduction, should become approximately independent of each other.

FIGURE 2.

Sample conformations obtained from the same starting constraints (in yellow) for a 34 nt poly-U chain. The newly formed base pair is in red. Conformations (a) and (b) show no newly formed base pairs. Conformations (c) and (d) show newly formed base pair initiating loops in the head and tail respectively. Conformations (e) and (f) show newly formed base pair creating internal junctions.

FIGURE 3.

Free energy cost due to conformational entropy loss at 310K for loop initiation in an unconstrained chain. The cost increases smoothly as a function of loop size (nt) with no significant position dependence along the sequence other than at the chain's ends where the cost decreases slightly. Experimental data for hairpin initiation obtained from melting experiments and aggregated in the nearest-neighbor model's database (21) have been included for comparison purposes. Error bars have been included for all points in the average value series. (Inset) the backbone geometric criteria used to define a base pair in the MC

simulation. All parameters are chosen to put the C1'-N_b glycosidic bonds in the correct geometry to form a Watson-Crick pair.

FIGURE 4.

Free energy cost at 310K to initiate a second loop of length b in a chain already containing a loop. The green symbols show the cost of the second loop b is independent of the spacer length c between it and the first loop a, which has a minimal stem length of 1. The grayscale symbols show the cost of loop b is independent of the stem length of loop a for a spacer length c = 2 nt. Other data showing similar independence for different spacer lengths c as well as the stem length on loop a are not presented. Note that error bars were included even though some of them are too small to be observed.

FIGURE 5.

The free energy costs of forming a two-way junction with 5' and 3' junction length b and c respectively given that a loop a is already in place. (a) Top view. (b) Side view. In general, the free energy cost grows as the junction size increases and is roughly symmetric when the 5' and 3' lengths are swapped.

FIGURE 6.

The free energy cost of forming symmetric two-way junctions plotted for chains with different sizes of the first loop, a, and for different lengths of the stem separating a from the 2-way junction (b, c). Over the set of three values used for a, the free energy cost to close the junction are consistent with each other. This indicates that the two-way junction is dependent on only the two junction lengths b and c, but not the loop on the opposite side, a. Over the three different stem lengths, the cost to close the symmetric junction shows no discernible dependence on the length of the stem. Typical error bars for selected data points are included. The error for larger loop sizes can be attributed to errors in the counting experiment. Dash line is a guide to the eye.

FIGURE 7.

Reduced topological representation of the set constraints defining a three-way junction. For the purposes of this study. For Table 2 and all sub-tables of Table S1, the value for b is fixed while a and c changing to give rise to the different sizes of three-way junctions.

FIGURE 8.

The free energy costs of forming a three-way junction with 5' and 3' junction length a and c respectively given that junction length b is fixed at 1 nt; this surface corresponds to the data given in table 2 above. (a) Top view. (b) Side view.

FIGURE 9.

The free energy cost of initiating a third hairpin of length c in the presence of two existing loops (a and b). When compared against the cost of initiating a hairpin loop on the free chain, the cost of the third loop is comparable and shows no dependence on the location of the new loop relative to the existing loops. This suggests that independence of hairpin loops can be extended to any number of loops within a chain. Note that error bars were included for the average cost of the first hairpin like in Fig.3; some of them are not visible due to their size.

FIGURE 10.

Reduced topological representation of the set constraints defining a four-way junction. For the purposes of this study, the two of the lengths were constrained to be equal and fixed in value ($b = c = 4$ nt) while the other lengths (a and d) can vary.

FIGURE 11.

Diagrammatic representation of the topology of a three-way junction and how it can be altered by introduction of new tertiary interactions. (a) An unmodified three-way junction like the one shown in Fig. 1(a). (b) Representation of kissing loops. The new constraint represented by the thick dash line in the top row of (b) results in a change in connectivity that no longer allows the two loops $\boldsymbol{b}$ and $\boldsymbol{d}$ to be factored. (c) Representation of ligand-mediated base-base contact in the three-way junction. The new constraint closes a portion of the three-way junction into a loop, giving rise to a diagram that is factorizable into 4 independent subsets corresponding to two hairpins, one two-way junction, and one three-way junction. (d) The kissing loop and ligand-mediated base-base interaction are combined. The effect changes the connectivity to yield a factorizable diagram consisting of a two-way junction and the structure previously seen in (b). (e) The kissing loop interaction is now combined with a triple base interaction. This yields a new structure factorizable into a two-way junction and a new multiply-connected loop structure.