JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 21, Number 9, 2014
© Mary Ann Liebert, Inc.
Pp. 649–664
DOI: 10.1089/cmb.2014.0107

Shapes of Interacting RNA Complexes

BENJAMIN M.M. FU and CHRISTIAN M. REIDYS

ABSTRACT

Shapes of interacting RNA complexes are studied using a filtration via their topological genus. A shape of an RNA complex is obtained by (iteratively) collapsing stacks and eliminating hairpin loops. This shape projection preserves the topological core of the RNA complex, and for fixed topological genus there are only finitely many such shapes. Our main result is a new bijection that relates the shapes of RNA complexes with shapes of RNA structures. This allows for computing the shape polynomial of RNA complexes via the shape polynomial of RNA structures. We furthermore present a linear time uniform sampling algorithm for shapes of RNA complexes of fixed topological genus.

Key words: bijection, interacting RNA complexes, shape polynomials, topological genus, uniform generation

1. INTRODUCTION

In this article we study shapes of RNA complexes, which constitute one of the fundamental mechanisms of cellular regulation. We find such interactions in a variety of contexts, such as small RNAs binding a larger (m)RNA target, including the regulation of translation in both prokaryotes (Narberhaus and Vogel, 2007) and eukaryotes (McManus and Sharp, 2002; Banerjee and Slack, 2002; the targeting of chemical modifications (Bachellerie et al., 2002); insertion editing (Benne, 1989); and transcriptional control (Kugel and Goodrich, 2007). RNA–RNA interactions are far more complex than simple sense–antisense interactions. This is observed for a vast variety of RNA classes including miRNAs, siRNAs, snRNAs, gRNAs, and snoRNAs.

An RNA molecule is a linearly oriented sequence of four types of nucleotides, namely, **A**, **U**, **C**, and **G**. This sequence is endowed with a well-defined orientation from the $5'$- to the $3'$-end and referred to as the backbone. Each nucleotide can form a base pair by interacting with at most one other nucleotide by establishing hydrogen bonds. Here we restrict ourselves to Watson-Crick base pairs **GC** and **AU** as well as the wobble base pairs **GU**. In the following, base triples as well as other types of more complex interactions are neglected.

RNA structures can be presented as diagrams by drawing the backbone horizontally and all base pairs as arcs in the upper half-plane (Fig. 1). This set of arcs provides our coarse-grained RNA structure, ignoring any spatial embedding or geometry of the molecule beyond its base pairs.

As a result, specific classes of base pairs translate into distinct structure categories, the most prominent of which being secondary structures (Kleitman, 1970; Nussinov et al., 1978; Waterman, 1978a,b).

Department of Mathematics and Computer Science, University of Southern Denmark, Odense M, Denmark.

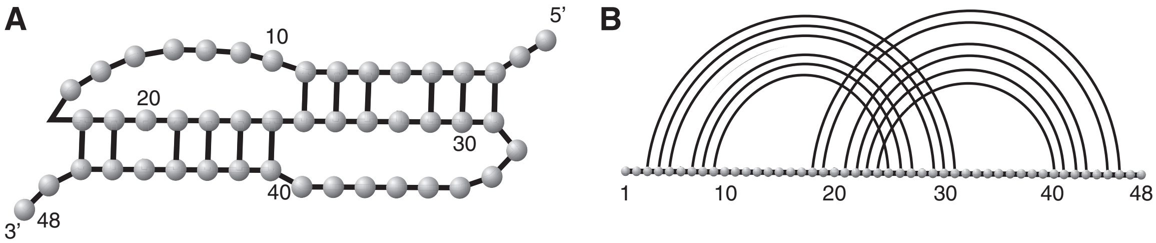


FIG. 1. **(A)** An RNA secondary structure and **(B)** its diagram representation.

Represented as diagrams, secondary structures have only non-crossing base pairs (arcs). Beyond RNA secondary structures, we find RNA pseudoknot structures. These exhibit cross serial interactions (Rivas and Eddy, 1999). Once such cross-serial interactions are considered, the question of a meaningful filtration arises, since the folding of unconstrained pseudoknot structures is NP-hard (Lyngsø and Pedersen, 2000).

It turns out that topological genus is one such meaningful observable. The genus of pseudoknotted, single stranded RNA has been studied in Vernizzi and Orland (2005), Vernizzi et al. (2005), Bon et al. (2008), and Andersen et al. (2011), and there are several alternative filtrations of cross-serial interactions (Orland and Zee, 2002; Reidys et al., 2010, 2011).

The objects studied here are derived from RNA complexes, which are diagrams over two backbones. Distinguishing internal and external arcs, the former being arcs within one backbone and the latter connecting the backbones, RNA complexes can be represented by drawing the two backbones on top of each other (Fig. 2).

We shall study shapes of RNA complexes, which are obtained by recursively removing all arcs of length one and collapsing all parallel arcs (Fig. 3).

Shapes are tailored to preserve the topological information of the molecule. The particular topologization is obtained via the notion of fat graphs, which date back to Heffter (1891). The classification and expansion of pseudoknotted RNA structures in terms of topological genus of a fat graph or double line graph were first proposed by Orland and Zee (2002) and Bon et al. (2008). In the context of RNA secondary structures, fat graphs were employed even earlier in Penner and Waterman (1993) and Penner (2004). The results of Orland and Zee (2002) are based on the matrix models and are conceptually independent. Genus, as well as other topological invariants of fat graphs, were introduced and studied as descriptors of proteins in Penner et al. (2010).

The approach undertaken here is combinatorial and follows Andersen et al. (2012). Starting with the diagram representation, we inflate each edge, including backbone edges, into ribbons. As each ribbon has two sides, and by specifying a counter-clockwise rotation around each vertex, we obtain so-called boundary cycles with a unique orientation. It is clear that we have thus constructed a surface, and its topological genus provides the desired filtration. Naturally, there are many such ribbon graphs that produce the same topological surface (by gluing the two ‘‘complementary’’ sides of each ribbon); this is how we obtain the desired equivalence (complexity) classes of structures.

It is easy to see that transforming an interaction structure into its shape preserves topological genus, and in Lemma 3.1, we shall see that for fixed genus g there exist only finitely many such shapes of RNA complexes. This means that for a fixed genus, there are only finitely many topologically distinct configurations, and important information is captured in the generating polynomial. In Theorem 4.5, we shall compute this polynomial and relate its coefficients to shapes of RNA structures by means of bijections relating one and two backbone shapes.

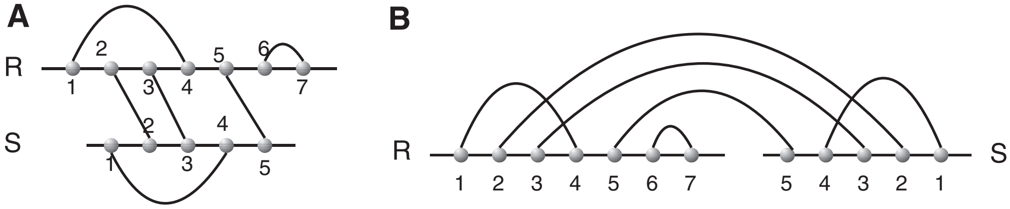


FIG. 2. Diagram representation of an RNA complex.

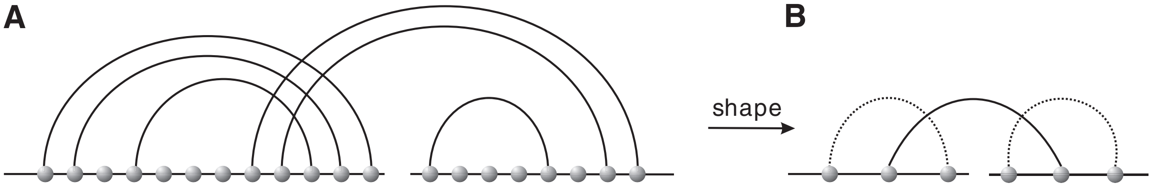


FIG. 3. From a 2-backbone diagram to its shape. The dashed arcs represent the rainbows (plants) of the shape.

In Huang and Reidys (2014), a linear time algorithm for uniformly generating shapes of RNA structures of fixed topological genus was given. By means of the bijection of Theorem 4.2 relating one and two backbone shapes, we can use this algorithm to generate, uniformly, shapes of RNA complexes.

The article is organized as follows: In Section 2, we introduce diagrams and the basic framework in which we formulate our results. We discuss fat graphs and the topological filtration, namely, as drawing these diagrams on orientable surfaces of higher topological genus. In Section 3, we develop the concept of shapes and establish basic properties. We recall some key results on shapes of RNA structures, in particular the two-term recursion for computing their coefficients. In Section 4, we analyze shapes of RNA complexes and relate them to shapes of RNA structures. Several constructions show how to derive one from the other by specific ''shape-surgery.'' Here we also present the uniform generation algorithm of shapes of RNA complexes of fixed topological genus. In Section 5, we discuss specific RNA complexes, which all have a fixed shape, and in Section 6, we integrate and discuss our results.

2. SOME BASIC FACTS

Definition 2.1. *A diagram is a labeled graph over the vertex set* $[n]=\{1, 2, \dots, n\}$ *represented by drawing the vertices* $1, 2, \dots, n$ *on a horizontal line in the natural order and the arcs* (i, j)*, where* $i < j$*, in the upper half-plane. The backbone of a diagram is the sequence of consecutive integers* $(1, \dots, n)$ *together with the edges* $\{\{i, i+1\} \mid 1 \le i \le n-1\}$*. A diagram over b backbones is a diagram together with a partition of* $[n]$ *into b backbones* (*Fig. 4*).

We shall distinguish backbone edges $\{i, i+1\}$ from arcs $(i, i+1)$, which we refer to as 1-arcs. Two arcs (i, j), (r, s), where $i < r$, are crossing if $i < r < j < s$ holds. Parallel arcs of the form $\{(i,j), (i+1, j-1), \cdots, (i+\ell-1, j-\ell+1)\}$ are called a stack, and ℓ is called the length of the stack. A stack on $[i, j]$ of length k naturally induces $(k-1)$ pairs of intervals of the form $([i+l, i+l+1], [j-l-1, j-l])$, where $0 \le l \le k-2$. Any of these $2(k-1)$ intervals is referred to as a *P-interval*. An interval $[i, i+1]$ is called a *gap* if there exists a pair of subsequent backbones B_1 and B_2 such that $i(i+1)$ is the rightmost(leftmost) vertex of $B_1(B_2)$. The vertex i is referred to as *cut vertex*. Any interval other than a gap or P-interval is called a *σ-interval*. Clearly, a diagram over $[n]$ contains $(n-1)$ intervals of length 1, and we distinguish three types: gap intervals, P-intervals, and σ-intervals (Fig. 5).

Vertices and arcs of a diagram correspond to nucleotides and base pairs, respectively. For a diagram over b backbones, the leftmost vertex of each backbone denotes the $5'$ end of the RNA sequence, while the rightmost vertex denotes the $3'$ end. The particular case $b = 2$ is referred to as RNA interaction structures or RNA complexes. RNA complexes are oftentimes represented alternatively by drawing the two backbones on top of each other, (Fig. 6).

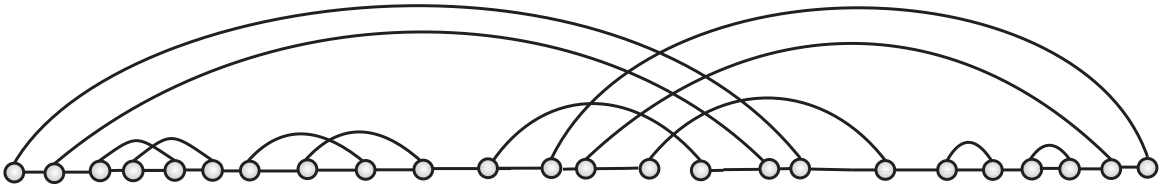


FIG. 4. A two-backbone diagram with 24 vertices and 12 arcs.

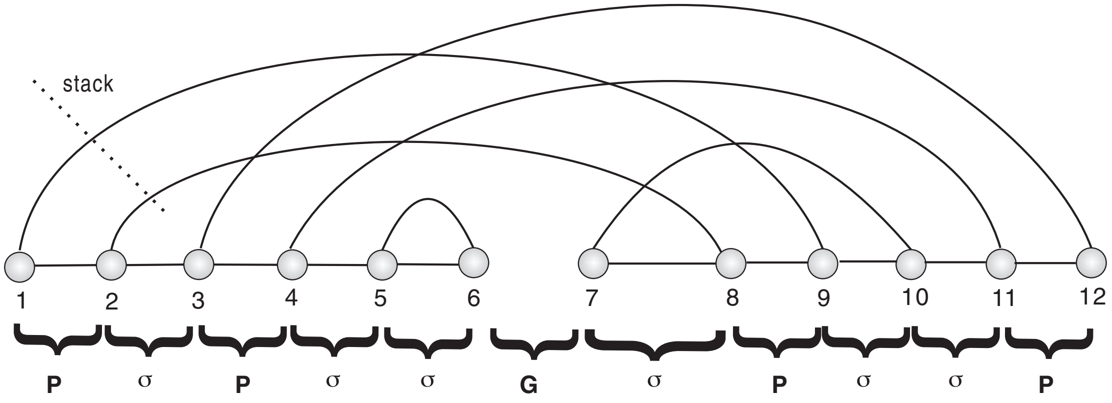


FIG. 5. Stacks and intervals: gap intervals, σ-intervals, and P-intervals labeled by G, σ, and P, respectively. There are four stacks: {(1, 9), (2, 8)}, {(3, 12), (4, 11)}, {(5, 6)}, and {(7, 10)}.

We will add an additional ''rainbow-arc'' over each respective backbone and refer to these diagrams as *planted diagrams* (Fig. 7).

A fat graph is a graph enriched by a cyclic ordering of the incident half-edges at each vertex and consists of the following data: a set of half-edges, H; cycles of half-edges as vertices; and pairs of half-edges as edges. The idea of half-edges stems from the observation that untwisted ribbons have two sides and are traversed in complementary directions. It is then a matter of convention to denote the terminal half of these sides as half-edge.

The specific drawing of a diagram G in the plane determines a cyclic ordering on the half edges of the underlying graph incident on each vertex, thus defining a corresponding fat graph $\mathbb{G}$. The collection of cyclic orderings is called fattening, one such ordering on the half-edges incident on each vertex (Fig. 8).

A fat graph $\mathbb{G}$ can be embedded in a compact orientable surface $F(\mathbb{G})$, such that its complement is a disjoint union of simply connected domains (called the faces or boundary components) and considered up to oriented homeomorphism. We can define the genus g of the fat graph by the genus of the surface. Clearly, $F(\mathbb{G})$ contains G as a deformation retract, and each $\mathbb{G}$ represents a cell-complex (Massey, 1967) over $F(\mathbb{G})$ (Fig. 9).

A diagram G hence determines a unique surface $F(\mathbb{G})$. Equivalence of simplicial and singular homology implies that Euler characteristic χ and genus g of $F(\mathbb{G})$ are independent of the choice of the cell-complex $\mathbb{G}$ and given by $\chi = v - e + r$ and $g = 1 - \frac{1}{2}\chi$, where v, e, r are the number of discs, ribbons, and boundary components in $\mathbb{G}$, respectively.

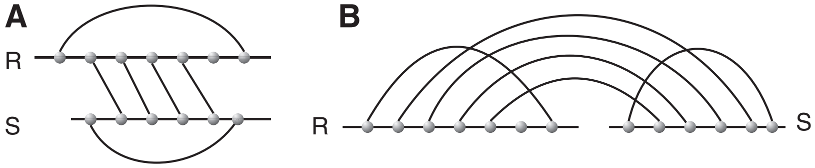
Without affecting topological type of the surface, one may collapse each backbone to a single vertex with the induced fattening called the polygonal model of the RNA (Fig. 10).

This backbone collapse preserves orientation, Euler characteristic, and genus. It is reversible by inflating each vertex to form a backbone. Using the collapsed fat graph representation, we see that for a connected diagram over b backbones, the genus g of the surface is determined by the number n of arcs and the number r of boundary components, namely, $2 - 2g - r = v - e = b - n$.

Boundary components are in the following, oftentimes referred to as loops. We distinguish the following loop-types:

- *hairpin loops*, which are boundary components of length one,
- *interior loops*, which are boundary components of length two,
- *multi-loops*, which are boundary components of length $2 \geq 3$.

FIG. 6. **(A)** An RNA complex presented by drawing the two backbones on top of each other. **(B)** The corresponding diagram over two backbones.



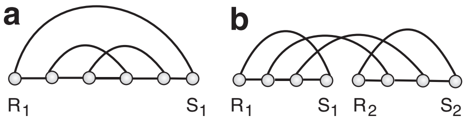


FIG. 7. **(a)** A planted one-backbone diagram with the plant arc (R_1, S_1); **(b)** a planted two-backbone diagram with the plant arc $\{(R_1, S_1), (R_2, S_2)\}$.

We furthermore distinguish within multiloops *pseudoknot loops*, which are multiloops containing some crossing arcs in the diagram representation. In interaction structures, we shall distinguish α*-loops* and β*-loops*, and α *stacks* and β *stacks*, depending on whether or not they contain only arcs whose endpoints are on one backbone.

3. SHAPES

A diagram is called a preshape if it contains neither 1-arcs [the arcs have the form $(i, i+1)$] nor stacks (parallel arcs), and isolated vertices (the vertices not paired). A preshape without a rainbow is called pure. A shape is then obtained from a pure preshape by adding a rainbow for every backbone (Fig. 11). We can obtain the shape of a planted diagram by iterating the following two steps: First collapse each stack into an arc; secondly remove all the 1-arcs and isolated vertices. Iteration generates an unique diagram without stacks, 1-arcs, and isolated vertices (Fig. 12).

For fixed genus g, there exist only finitely many shapes over one backbone (two backbones) (Andersen et al., 2012; Reidys et al., 2011).

Lemma 3.1. *Given a one-backbone shape of genus g with n edges, we have* $2g + 1 \leq n \leq 6g - 1$. *Therefore, for fixed genus g, there exist only finitely many shapes.*

Proof. First note that if there is more than one boundary component, then there must be an arc with different boundary components on its two sides, and removing this arc decreases r by exactly one while preserving g since the number of arcs is given by $n = 2g + r - 1$. Furthermore, if there are v_l boundary components of length l in the polygonal model, then $2n = \sum_l lv_l$ since each side of each arc is traversed once by the boundary (including the plant). For a shape, $v_1 = 1$, because the plant gives the only boundary component of length 1; $v_2 = 0$ by the definition of shapes. It therefore follows that $2n = \sum_l lv_l \geq 3(r-1)+1$, so $2n = 4g + 2r - 2 \geq 3r - 2$, that is, $4g \geq r$. Thus, we have $n = (2g + 4g - 1) = 6g - 1$, that is, any shape can contain at most $6g - 1$ arcs. The lower bound $2g + 1$ follows directly from $n = 2g + r - 1$ since $r \geq 2$.

For fixed genus g, the number of arcs in the shape is at most $6g - 1$, and the second assertion follows. ∎

Lemma 3.1 implies that the generating function for one-backbone shapes of genus g is a polynomial. For example, for the shapes over one backbone with genus 1 to 3, we have

$$\begin{aligned}
S_1(z) &= z^3 + 2z^4 + z^5,\\
S_2(z) &= 21z^5 + 189z^6 + 651z^7 + 1134z^8 + 1071z^9 + 525z^{10} + 105z^{11},\\
S_3(z) &= 1485z^7 + 26928z^8 + 198451z^9 + 808478z^{10} + 2054305z^{11} + 3442340z^{12}\\
&\quad + 3883363z^{13} + 2928926z^{14} + 1419418z^{15} + 400400z^{16} + 50050z^{17}
\end{aligned}$$

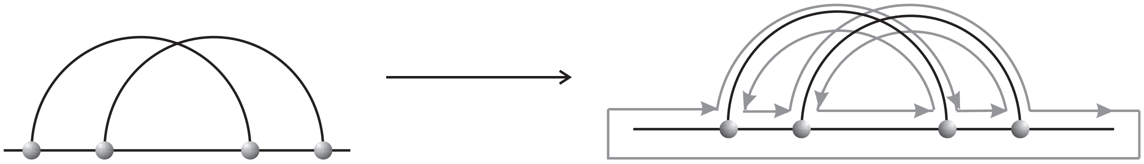


FIG. 8. The fattening.

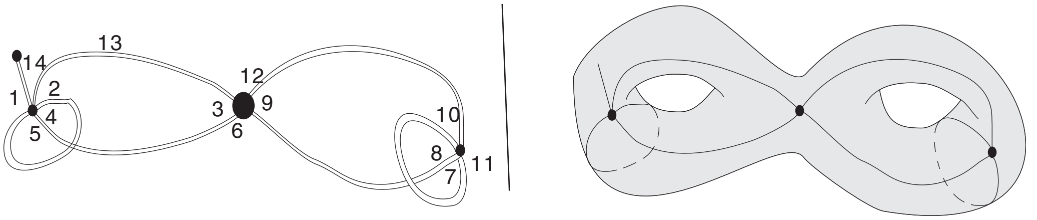


FIG. 9. A fatgraph and its embedding.

Explicit formulas for the coefficients of the shape polynomial of arbitrary fixed genus have been given in Huang and Reidys (2014). There the *Poincaré dual* of shapes, a unicellular map, was constructed, and a construction of Chapuy (2011) is refined to slice such a map into a tree with certain labeled vertices. The latter represent the blueprint to rebuild the original unicellular map and the shape, respectively.

Theorem 3.2. *(Huang and Reidys, 2014) The shape polynomial of genus g is given by*

$$S_g(z) = \sum_{t=1}^{g} \kappa_t^{(g)} z^{2g+t-1}(1+z)^{2g+t-1}, \tag{1}$$

where $\kappa_t^{(g)} = a_{t-1}^{(g)}\mathrm{Cat}(2g+t)$ *and*

$$a_t^{(g)} = \sum_{\substack{0=g_0<g_1<\cdots<g_r=g \\ 0=t_0=t_1\leq t_2\leq\cdots\leq t_r=r-t}} \prod_{i=1}^{r} \frac{1}{2g_i}\binom{2g+t-(2g_{i-1}+(i-1))+t_i}{2(g_i-g_{i-1})+1}. \tag{2}$$

Huang and Reidys (2014) furthermore derives from the underlying bijections a uniform generation algorithm, *UniformShape*, for shapes of a fixed genus g, which has linear time complexity.

Li and Reidys (personal communication, 2014) study the sequence $(\kappa_t^{(g)})_{t=1}^{g}$ (Table 1), which emerged originally in the computation of the virtual Euler characteristic of a curve (Harer and Zagier, 1986). Li and Reidys (personal communication, 2014) shows that $(\kappa_t^{(g)})_{t=1}^{g}$ is log-concave and hence unimodal and derives

$$\kappa_t^{(g)} = \frac{(2(2g+t-1))!}{2^{2g}(2g+t-1)!\sum_{\gamma\vdash g}\prod_i m_i!(2i+1)^{m_i}}.$$

Furthermore,

Proposition 3.3. *(Li and Reidys, personal communication, 2014)* $\kappa_t^{(g)}$ *satisfies*

$$(2g+t)\kappa_t^{(g)} = (2(2g+t)-3)(2(2g+t)-5)\left((2g+t-2)\kappa_t^{(g-1)} + 2(2(2g+t)-7)\kappa_{t-1}^{(g-1)}\right),$$

where $\kappa_1^{(1)} = 1$, $\kappa_t^{(g)} = 0$, *if* $t < 1$ *or* $t > g$.

The above recursion has also been derived by Chekhov (1997) using matrix models.

4. SHAPES OVER TWO BACKBONES

In this section, we study shapes over two backbones. Our main observation is that shapes over two backbones correspond to particular shapes over one backbone with topological genus increased by one.

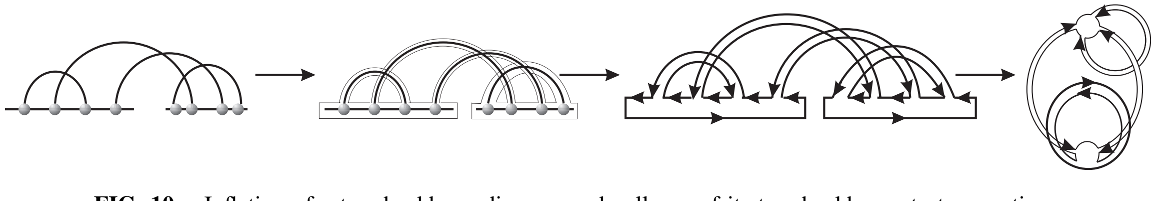


FIG. 10. Inflation of a two-backbone diagram and collapse of its two backbones to two vertices.

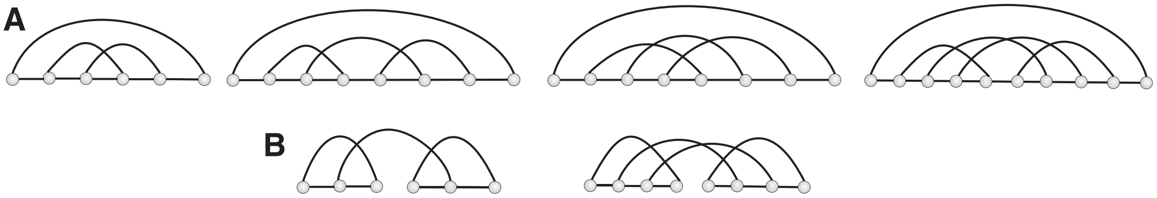


FIG. 11. **(A)** The four shapes of genus 1 over one backbone. **(B)** The two shapes of genus 0 over two backbones.

We denote a shape over one backbone by (B, α), where

$$B := [R_1, 1, 2, \cdots 2n, S_1]$$

is the sequence of vertices along the backbone and α is a fixed-point free involution, which contains (R_1, S_1) as one cycle (rainbow); α-cycles represent edges, and (R_1, S_1) is the plant.

We shall now distinguish two types of shapes. A shape is an A-shape if the vertex following $\alpha(1)$ is paired with the last vertex before S_1 and a B-shape otherwise (Fig. 13). Let the set of A- and B-shapes having n edges and genus g be denoted by $\mathcal{A}_g(n)$ and $\mathcal{B}_g(n)$, respectively. Furthermore, let $\mathcal{A}_g = \bigcup_n \mathcal{A}_g(n)$ and $\mathcal{B}_g = \bigcup_n \mathcal{B}_g(n)$, $\mathcal{S}_g(n) = \mathcal{A}_g(n) \bigcup \mathcal{B}_g(n)$.

Lemma 4.1. *We have a bijection:*

$$\theta : \mathcal{A}_g(n+2) \longrightarrow \mathcal{B}_g(n+1),$$

that is, there exists a pairing $(x, \theta(x))$ *associated to each A-shape and its unique B-shape. In particular,*

$$\mathcal{S}_g = \mathcal{A}_g \dot{\cup} \mathcal{B}_g$$

and $|\mathcal{S}_g|/2 = |\mathcal{A}_g|$.

Proof. Let $\Gamma = ([R_1, 1, 2, \cdots 2n+1, 2n+2, S_1], \alpha)$ be an A-shape having n + 2 arcs, containing the arc $(\alpha(1) + 1, 2n + 2)$. Since Γ is a shape, there are no nested arcs or 1-arcs, whence removal of $(\alpha(1) + 1, 2n + 2)$ maps an A-shape into a B-shape.

Furthermore, as an A-shape, Γ has a boundary component of size three, γ_3, traversing the sides of the rainbow, $(1, \alpha(1))$ and $(\alpha(1) + 1, 2n + 2)$. Let θ be the mapping defined by removing the arc $(\alpha(1) + 1, 2n + 2)$ together with its incident vertices and subsequent relabeling of the remaining vertices. Then θ decreases both: the number of boundary components, r, as well as the number of arcs $n + 2$ by 1. To see this, we note that $(\alpha(1) + 1, 2n + 2)$ is traversed by two distinct boundary components, γ, γ_3. Removing $(\alpha(1) + 1, 2n + 2)$ consequently merges γ and γ_3, whence the number of boundary components decreases by one. Euler's characteristic equation, $2 - 2g - r = 1 - (n + 2)$, shows that θ preserves g (Fig. 14).

We next specify θ^{-1}. Given a B-shape having $n + 1$ edges and genus g, we insert an arc with endpoints between $[\alpha(1), \alpha(1) + 1]$ and $[2n, S_1]$ and subsequently relabel the diagram. This insertion maps any B-shape into an A-shape. Namely, by construction, it creates neither nested arcs nor 1-arcs (the latter would imply that the rainbow has a nested arc). After relabeling, the inserted arc is incident to $(\alpha(1) + 1, 2n + 2)$ and creates a new boundary component, γ_3, as specified above. Euler's characteristic equation then shows that θ^{-1} does preserve genus (Fig. 14). ■

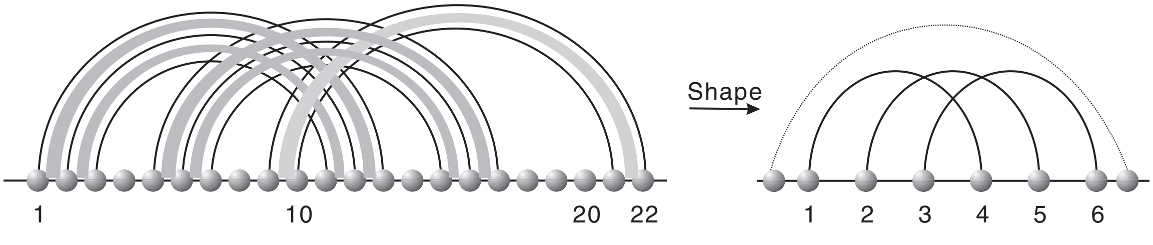


FIG. 12. From a diagram to a shape by removing all 1-arc and parallel arcs. The dashed arc is a rainbow, displayed together with a nested preshape.

TABLE 1. THE COEFFICIENTS $\kappa_t^{(g)}$

	g = 1	2	3	4	5
t = 0	1	21	1485	225225	59520825
1		105	18018	4660227	1804142340
2			50050	29099070	18472089636
3				56581525	78082504500
4					117123756750

Let $\mathcal{Q}_g$ denote the set of shapes over two backbones of genus g, and $\mathcal{S}_g^2$ denote the set of pairs of disconnected one-backbone shapes whose sum of genera equals g. Let $\mathcal{Q}'_g = \mathcal{Q}_g \cup \mathcal{S}_g^2$.

Theorem 4.2. *We have the following commutative diagram of bijections:*

$$\begin{tikzcd}
\mathcal{Q}'_g \arrow[r, "\eta"] \arrow[d, dash] & \mathcal{A}_{g+1} \arrow[d, dash] \\
\mathcal{Q}'_g(n+2) \arrow[r, "\eta_n"] & \mathcal{A}_{g+1}(n+3)
\end{tikzcd}$$

Proof. Since any $\mathcal{Q}'_g$-diagram has a unique number of arcs, it suffices to specify the bijections η_n. An $\mathcal{Q}'_g(n+2)$-element can be denoted by

$$x = ([[R_1, 1, 2, \cdots, m, S_1], [R_2, m+1, \cdots 2n, S_2]], \alpha),$$

having the rainbows (R_1, S_1), (R_2, S_2).

We define the mapping η_n as follows:

- first we glue the two backbones into

$$[R_1, 1, 2, \cdots, m, S_1, R_2, m+1, \cdots 2n, S_2],$$

- secondly we add a new rainbow,
- thirdly we relabel the vertices.

This produces a unique backbone

$$[R_1, 1, 2, \cdots, 2n-1, 2n, 2n+1, 2n+2, S_1]$$

and transforms the two rainbows into the new arcs

$$(R_1, S_1) \mapsto (1, \alpha(1)) \quad \text{and} \quad (R_2, S_2) \mapsto (\alpha(1)+1, 2n),$$

respectively. Accordingly, $\eta_n(x)$ is an A-shape having $(n + 3)$ edges (Fig. 15).

The mapping η_n eliminates one backbone, that is, $b' = b - 1$; generates a γ_3-boundary component merging the two original rainbow-boundaries and adds a new rainbow boundary, that is, $r' = r$; and adds one edge, that is, $n' = n + 3$. In view of $2 - 2g - r = 2 - (n + 2)$ we obtain

$$2g' = 2 - r - (2-1) + (n+3) = 2(g+1),$$

which proves that $\mathcal{A}_{g+1}(n+3)$.

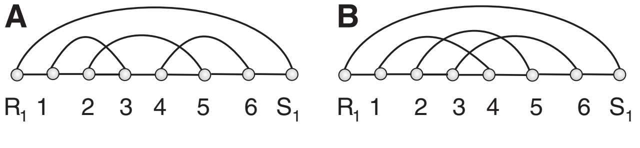


FIG. 13. A-shapes [$\alpha(1) + 1 = 4$ is paired with 6] and B-shapes [$\alpha(1) + 1 = 5$ is not paired with 6].

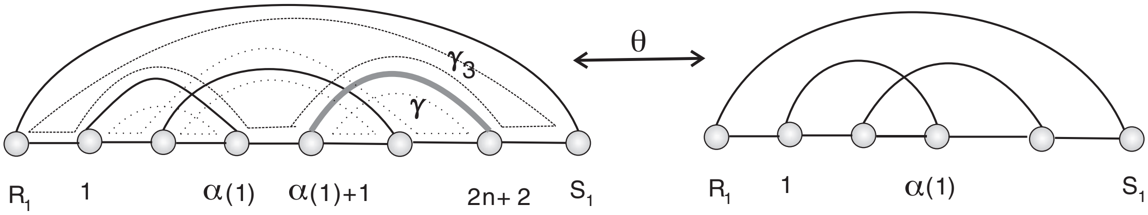


FIG. 14. θ: removal of ($\alpha(1)$ + 1, 2n + 2) creates a B-shape.

We next construct η_n^{-1} as follows: Consider an A-shape y $\in \mathcal{A}_{g+1}(n+3)$, then

- remove the rainbow,
- cut the backbone between $\alpha(1)$ and $\alpha(1)$ + 1, and
- relabel the two respective backbones.

By construction, the edges (1, $\alpha(1)$), ($\alpha(1)$ + 1, 2n) become the rainbows of the new backbones. The mapping η_n^{-1} reverses η_n, and our above accounting of backbones, boundary components, and edges applies here. Thus η_n^{-1}(y) is a two-backbone diagram of genus g having n + 2 edges (Fig. 15). ■

Corollary 4.3. *Let $x \in \mathcal{Q}'_g(n+2)$ be a shape over two backbones containing ℓ-multiloops, then $\eta_n(x) \in \mathcal{A}_{g+1}(n+3)$ is an A-shape over one backbone having ℓ + 1 multiloops.*

Proof. The map η_n merges two rainbow-boundary components of x and the new rainbow into a multiloop of length 3 (Fig. 15). ■

Algorithm 1: Uniform generation of shapes over two backbones

```
UniformBi-shape (TargetGenus)
while 1 do
    $\mathfrak{s}_1 \leftarrow$ UnifromShape(TargetGenus + 1)
    if $\mathfrak{s}_1$ is type A then
        $\mathfrak{s}_2 \leftarrow \eta^{-1}(\mathfrak{s}_1)$
    else
        $\mathfrak{s}_2 \leftarrow \eta^{-1}\theta^{-1}(\mathfrak{s}_1)$
    end if
    if Connection ($\mathfrak{s}_2$) then
        return $\mathfrak{s}_2$
    end if
end while
```

A first application of Theorem 4.2 is a uniform generation algorithm for shapes over two backbones of fixed topological genus g. We show the pseudocode in Algorithm 1.

Corollary 4.4. *Algorithm 1 generates two-backbone shapes of genus g uniformly.*

Proof. *UniformShape* (Huang and Reidys, 2014) generates one-backbone shape uniformly and any two-backbone shape corresponds to either an A-shape via η or a B-shape via $\theta \circ \eta$. Since A- and B-shapes are generated uniformly, any two-backbone shape is generated uniformly with multiplicity two. ■

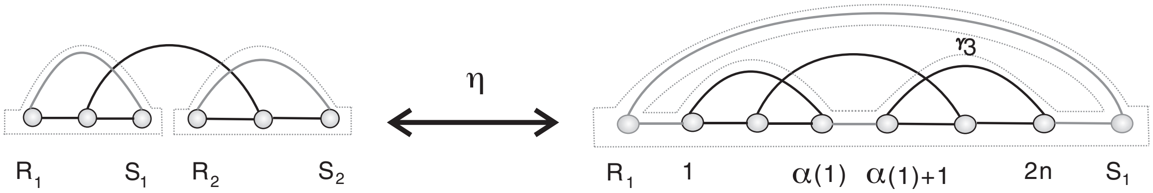


FIG. 15. The mapping η.

Let $\mathcal{S}^2$ denote the set of pairs of disconnected shapes whose sum of genera equals g and let $s_g^2(n)$ denote the number of these shapes having n arcs. Then $S_g^2(z) = \sum_n s_g^2(n)z^n$ satisfies $S_g^2(z) = \sum_1^g S_i(z)S_{g+1-i}(z)$.

Theorem 4.5. *The polynomial of shapes of genus g over two backbones, $Q_g(z) = \sum_l q_g(l)z^l$, is given by $Q_g(z) = Q'_g(z) - \sum_1^g S_i(z)S_{g+1-i}(z)$, where*

$$Q'_g(z) = \frac{S_{g+1}(z)}{(1+z)} = \sum_{t=1}^{g+1} \kappa_t^{(g+1)} z^{2g+t+1}(1+z)^{2g+t}.$$

Proof. Each $\mathcal{Q}'_g$-diagram is a $\mathcal{Q}'_g(n)$-diagram for a unique n. As such we have

$$\begin{tikzcd}
\mathcal{Q}'_g(n+2) \arrow[r, "\eta_n"] \arrow[dr, "\theta\circ\eta_n"'] & \mathcal{A}_{g+1}(n+3) \arrow[d, "\theta"] \\
& \mathcal{B}_{g+1}(n+2)
\end{tikzcd}$$

Suppose the generating function of A- and B-shapes is $A_g(z) = \sum_n a_g(n)z^n$ and $B_g(z) = \sum_n b_g(n)z^n$, respectively. From the bijection $\theta : \mathcal{A}_{g+1}(n+3) \leftrightarrow \mathcal{B}_{g+1}(n+2)$, we obtain $b_{g+1}(n+2) = a_{g+1}(n+3)$. Then $S_{g+1}(z) = A_{g+1}(z) + B_{g+1}(z)$ implies $S_{g+1}(z) = (1 + 1/z)A_{g+1}(z)$, or equivalently, $A_{g+1}(z) = \frac{S_{g+1}(z)}{1+1/z}$. By the bijection η, the generalized two-backbone shape $\mathfrak{s} \in \mathcal{Q}'_g$ has one arc less than $\eta(\mathfrak{s})$, which implies

$$Q'_g(z) = \frac{A_{g+1}(z)}{z} = \frac{S_{g+1}(z)}{(1+z)}.$$

Subtracting the set of disconnected two-backbone shapes, $S_g^2(z)$, the result follows. ■

For genus $g = 0, 1, 2$, we accordingly have

$$\begin{aligned}
\mathbf{Q}_0(z) &= z^3 + z^4 \\
\mathbf{Q}_1(z) &= 21z^5 + 167z^6 + 479z^7 + 645z^8 + 416z^9 + 104z^{10} \\
\mathbf{Q}_2(z) &= 1485z^7 + 25401z^8 + 172546z^9 + 633370z^{10} + 1413585z^{11} + 2015525z^{12} + 1852256z^{13} \\
&\quad + 1064616z^{14} + 348880z^{15} + 49840z^{16}
\end{aligned}$$

5. FIBERS

In the previous section, we computed the shape polynomials of shapes over two backbones of fixed topological genus. Their coefficients can be recursively determined and are directly related to the coefficients of polynomials of shapes over one backbone.

Furthermore, Theorem 4.2 implies a linear time sampling algorithm for such two-backbone shapes of genus g. By means of their preimages, shapes induce a natural partition of RNA complexes, and here we shall study the sets of RNA complexes having a fixed shape, $\mathfrak{s}$, to which we refer to as the fiber of $\mathfrak{s}$.

Given a two-backbone shape having l arcs and genus g, $\mathfrak{s}_{g,l}$, let $q_{\mathfrak{s}l,g}(n)$ be the number of two-backbone matchings of genus g having the shape $\mathfrak{s}_{l,g}$.

Theorem 5.1. *The generating function of matchings of genus g having shape $\mathfrak{s}_{l,g}$ is given by*

$$Q_{\mathfrak{s}_{l,g}}(z) = \sum_n q_{\mathfrak{s}_{l,g}}(n)z^n = C_0(z)^{2l+2}\frac{z^{l+2}}{(1-zC_0(z)^2)^{l+2}},$$

where $C_0(z) = \frac{1-\sqrt{1-4z}}{2z}$. In particular, the number of two-backbone structures of length n having genus g and shape $\mathfrak{s}_{l,g}$ depends only on l and

$$q_{\mathfrak{s}_{l,g}}(n) \sim \frac{k}{(l+1)!} n^{l+1} 4^{n-l-2},$$

where k is some positive constant.

Proof. By the following steps, we can inflate an RNA-complex from a shape (Fig. 16).

Step 1: We inflate each arc in $\mathfrak{s}_{l,g}$ into a sequence of induced arcs; an induced arc $\mathcal{N}$ is an exterior arc together with at least one nontrivial genus 0 matching in either one or both P-intervals. Clearly, we have $N(z) = z(2(C_0(z) - 1) + (C_0(z) - 1)^2) = z(C_0(z)^2 - 1)$. Furthermore, we inflate the arc into a sequence $\mathcal{M}$ of induced arcs $M(z) = \frac{1}{1-z(C_0(z)^2-1)}$. Inflating all $l + 2$ arcs (including the two rainbows) into a sequence of induced arcs leads to

$$z^{l+2} M(z)^{l+2} = z^{l+2} \left(\frac{1}{1 - z(C_0(z)^2 - 1)} \right)^{l+2}.$$

Denote the matching after this step by x_1.
Step 2: We inflate each arc in x_1 into a stack. The corresponding generating function is

$$\left(\frac{\frac{z}{1-z}}{1 - \frac{z}{1-z}(C_0(z)^2 - 1)} \right)^{l+2} = \frac{z^{l+2}}{(1 - zC_0(z)^2)^{l+2}}. \quad (3)$$

Step 3: We insert a $\mathcal{C}_0$ matching into the respective $(2l + 2)$ σ-intervals of $\mathfrak{s}_{l,g}$. The corresponding generating function is $C_0(z)^{2l+2}$.

Combining the above three steps, we derive

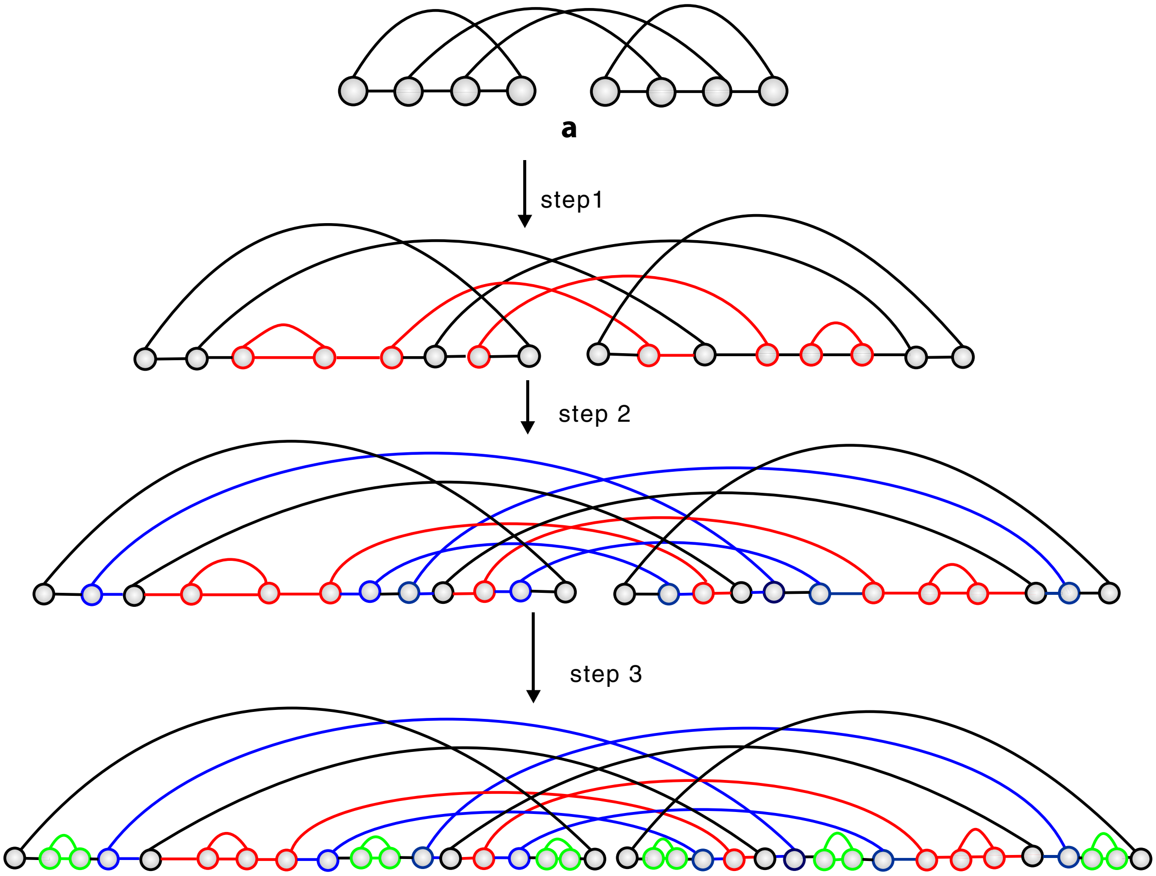


FIG. 16. **(a)** A shape of genus 1 with 4 arcs; step 1: inflate each arc to a sequence of induced arcs (red); step 2: inflate each exterior arc to a stack (blue); step 3: insert a $\mathcal{C}_0$-matching into the σ-intervals (green).

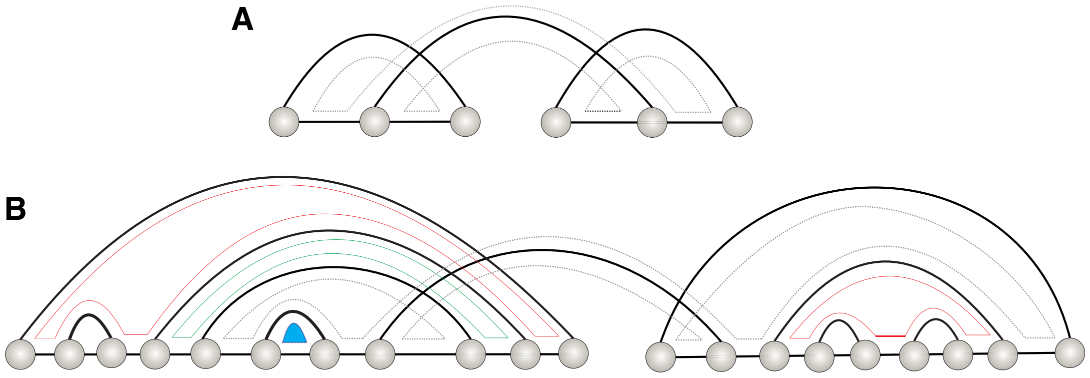


FIG. 17. A shape with a distinguished loop **(A)**. Inflation generates hairpin loops (blue), interior loops (green), and two types of nonshape multiloops (red) **(B)**. The length of the distinguished shape-loop increased by two.

$$Q_{\mathfrak{s}_l,\,g}(z)=\sum_n q_{\mathfrak{s}_l,\,g}(n)z^n=C_0(z)^{2l+2}\frac{z^{l+2}}{(1-zC_0(z)^2)^{l+2}},$$

where $q_{\mathfrak{s}_l,\,g}(n)$ denotes the number of genus g matchings generated from $\mathfrak{s}_{l,\,g}$.

The generating function has a unique, dominant singularity $\rho = 1/4$ with multiplicity $l+2$. Standard singularity analysis (Flajolet and Sedgewick, 2009) implies

$$q_{\mathfrak{s}_l,\,g}(n)\sim\frac{k}{(l+1)!}n^{l+1}4^{n-l-2}.$$

■

Corollary 5.2. *The generating function $W_g(z)$ of two-backbone matchings of genus g is given by*

$$W_g(z)=\sum_l q_g(l)Q_{\mathfrak{s}_l,\,g}(z)=\sum_l q_g(l)C_0(z)^{2l+2}\frac{z^{l+2}}{(1-zC_0(z)^2)^{l+2}}.$$

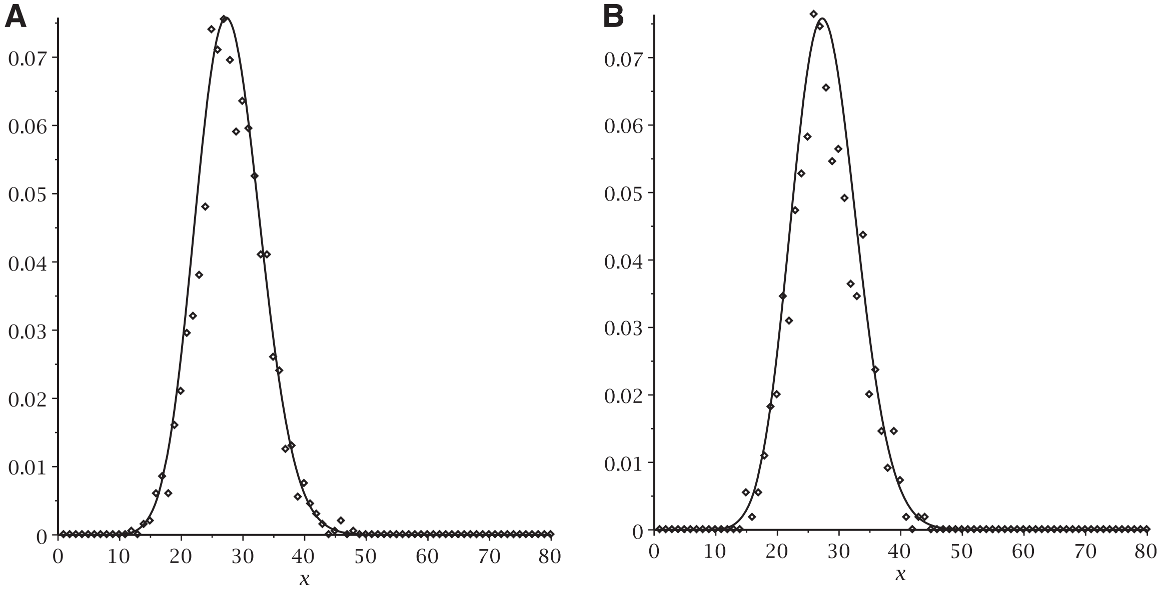


FIG. 18. Global and local sampling of shapes of RNA complexes of fixed topological genus: $N = 5 \times 10^5$ shapes of genus 1 were generated, and we display their multiplicities (dots) together with the binomial coefficients that are observed from uniform sampling **(A)**. Local sampling: we generate $N = 5 \times 10^5$ shapes of genus 1 with Seven arcs **(B)**.

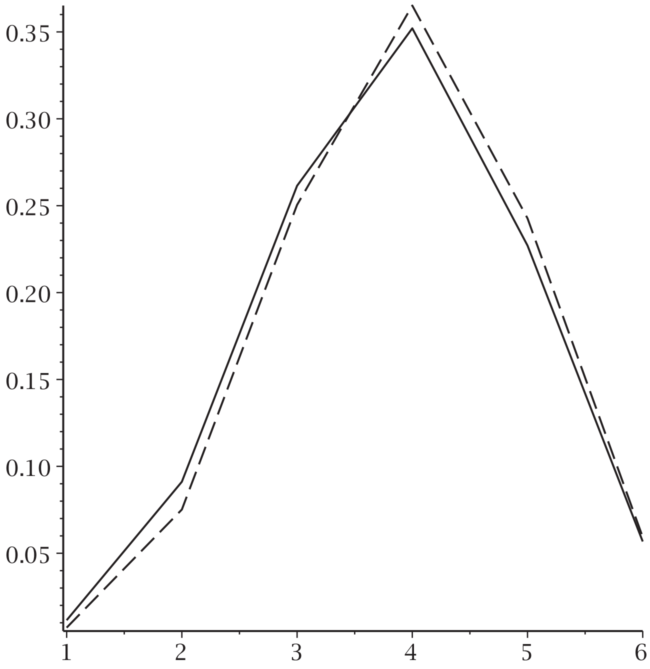


FIG. 19. Uniform sampling of RNA complexes of genus 1 with length 40, 80, 100, 150, 200, and (5×10^5). The solid curve displays the distribution induced by the coefficients of the shape polynomial, while the dashed curve displays distribution obtained from the sampling. Displayed is the average of the coefficients obtained from sampling the above different lengths.

In particular, we have $W_0(z) = \frac{z^3}{(1-4z)^2}$,

$$W_1(z) = \frac{(20z+21)z^5}{(1-4z)^5}, \quad W_2(z) = \frac{(1696z^2+6096z+1485)z^7}{(1-4z)^8}.$$

We conclude this section by discussing loops in shape-fibers. By construction, there are only multiloops and pseudoknot-loops in a shape. We observe that the lengths of the original shape-loops increase in structures of the shape-fiber. Structures of the shape-fiber exhibit, in addition, hairpin loops, interior loops, and two types of multiloops (Fig. 17).

6. DISCUSSION

In this article we study shapes of RNA complexes. We show that these shapes are directly related to shapes of RNA structures of increased topological genus. More precisely, we show in Lemma 4.1 that there is a bipartition of RNA-shapes into *A*-shapes and *B*-shapes. Furthermore, *A*- and *B*-shapes are in one-to-one correspondence. We establish in Theorem 4.2 that each respective type is in one-to-one correspondence to shapes of RNA complexes. These relations have various implications.

First, Lemma 3.1 guarantees that there are only finitely many such shapes. This leads to the shape polynomials for shapes of fixed topological genus *g*. The above correspondences reduce the computation of the coefficients of these polynomials for shapes of RNA complexes to those of shapes of RNA structures. For the latter, Proposition 3.3 gives a simple two-term recursion, which allows us to obtain any such polynomials for shapes of structures and complexes of fixed topological genus in constant time.

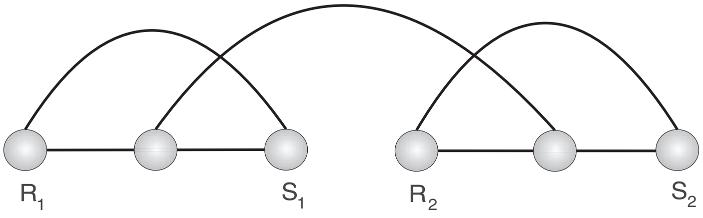


FIG. 20. The shape extracted from the biological RNA complexes (Richter and Backofen, 2012).

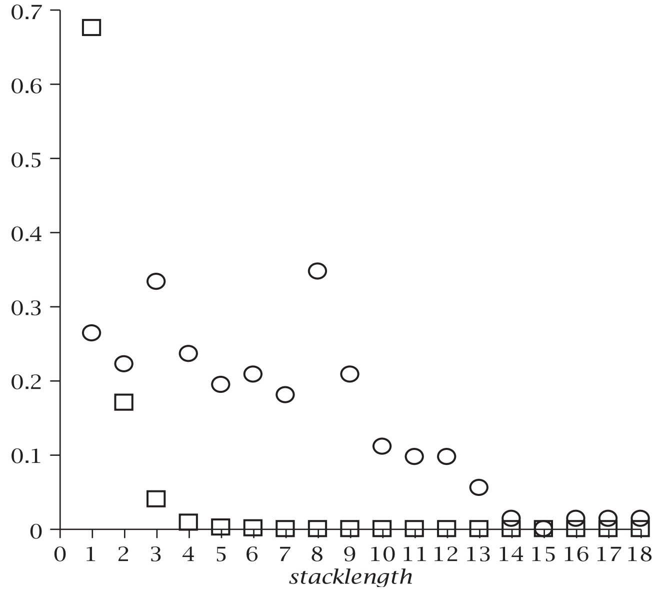


FIG. 21. The distribution of the lengths of exterior stacks in uniformly sampled structures having the shape in Figure 20 (box); the distribution of the length of exterior stacks in the biological RNA complexes obtained from Richter and Backofen (2012) (circle).

Secondly we obtain a sampling algorithm, Algorithm 1, for shapes of RNA complexes that have linear time complexity. Algorithm 1 and the sampling algorithm of RNA shapes are freely available online. This algorithm provides us with a plethora of statistics for shapes of RNA complexes of fixed topological genus. To illustrate local and global uniformity, we display in Figure 18 the multiplicities of shapes of genus 1. Here by local uniformity we mean that we can uniformly sample shapes of RNA complexes with a fixed number of arcs.

Lemma 3.1 shows that there are only finitely many shapes of RNA complexes. Hence the shape polynomial determines their numbers filtered by the number of arcs. This means that we can extract a finite observable from interaction structures that captures their topological core.

Let us calibrate this information by inspecting what happens when we sample uniformly RNA complexes of fixed topological genus (Fu et al., 2013). We uniformly sample RNA complexes having genus 1 and record the frequencies of their associated shapes. We observe that the distribution of shapes of different lengths equals the distribution obtained by normalizing the coefficients of the shape polynomial (Fig. 19).

Accordingly, the shape polynomial represents precisely the uniform case. As a result we can now compute the shapes of databases of RNA complexes and derive empirical coefficients (distributions) and

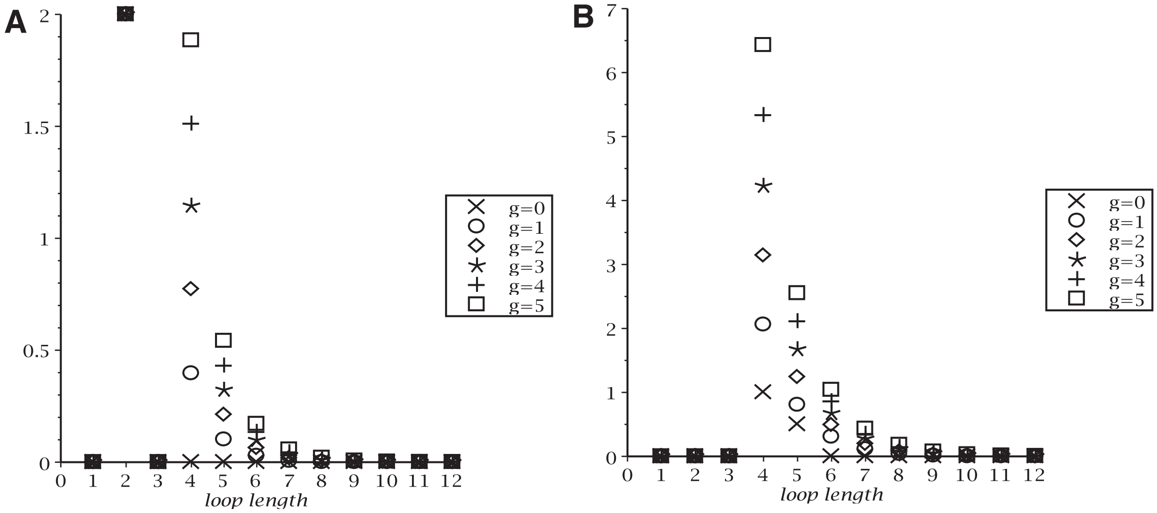


FIG. 22. The distribution of the average number of loops in the shapes of different genus: **(A)** the distribution of the α-loops (loops contained in one backbone) and **(B)** the distribution of the β-loops (loops over two backbones).

hence extract finite information from databases reflecting the topological properties of the biological complexes.

Along these lines we study the shapes of biological RNA complexes obtained from (Richter and Backofen, 2012). Because the data set contained only exterior arcs, we derived only one shape of genus zero (Fig. 20).

We accordingly compare the distribution of the exterior stack lengths of biological with that of uniformly sampled RNA complexes (Fig. 21).

We finally study loops in shapes of RNA complexes. By construction, such loops are multiloops, except for the two rainbow loops. We uniformly generate 5×10^5 shapes of RNA complexes from genus 0 to 5 and display the average number of loops (Fig. 22). The data suggest a central limit theorem for the average number of loops since their mean scales linearly with topological genus.

ACKNOWLEDGMENTS

We wish to thank Fenix W.D. Huang and Thomas J.X. Li for discussions. This work is funded by the Future and Emerging Technologies (FET) programme of the European Commission within the Seventh Framework Programme (FP7), under the FET-Proactive grant agreement TOPDRIM, FP7-ICT-318121.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Andersen, J.E., Huang, F.W.D., Penner, R.C., and Reidys, C.M. 2012. Topology of RNA-RNA interaction structures. *J. Comp. Biol.* 19, 928–943.

Andersen, J.E., Penner, R.C., Reidys, C.M., and Waterman, M.S. 2011. Topological classification and enumeration of RNA structures by genus. *J. Math. Biol.* 1–18.

Bachellerie, J.-P., Cavaillé, J., and Hüttenhofer, A. 2002. The expanding snoRNA world. *Biochimie* 84, 775–790.

Banerjee, D., and Slack, F. 2002. Control of developmental timing by small temporal RNAs: a paradigm for RNA–mediated regulation of gene expression. *Bioessays* 24, 119–129.

Benne, R. 1989. RNA–editing in trypanosome mitochondria. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression* 1007, 131–139.

Bon, M., Vernizzi, G., Orland, H., and Zee, A. 2008. Topological classification of RNA structures. *J. Mol. Biol.* 379, 900–911.

Chapuy, G. 2011. A new combinatorial identity for unicellular maps, via a direct bijective approach. *Adv. Appl. Math.* 47, 874–893.

Chekhov, L. 1997. Matrix model tools and geometry of moduli spaces. *Acta Applicandae Mathematica* 48, 33–90.

Flajolet, P., and Sedgewick, R. 2009. *Analytic Combinatorics*. Cambridge University Press, Cambridge, MA.

Fu, B.M.M., Han, H.S.W., and Reidys, C.M. 2013. On the RNA-RNA interaction structures of fixed topological genus. *arXiv:1311.0684v2*.

Harer, J., and Zagier, D. 1986. The euler characteristic of the moduli space of curves. *Invent. Math.* 85, 457–486.

Heffter, L. 1891. Über das Problem der Nachbagebiete. *Math. Ann.* 38, 477–508.

Huang, F.W.D., and Reidys, C.M. 2014. Shapes of topological RNA structures. *arXiv:1403.2908*.

Kleitman, D. 1970. Proportions of irreducible diagrams. *Studies in Appl. Math.* 49, 297–299.

Kugel, J.F., and Goodrich, J.A. 2007. An RNA transcriptional regulator templates its own regulatory RNA. *Nature Chemical Biology* 3, 89–90.

Lyngsø, R.B., and Pedersen, C.N. 2000. Pseudoknots in RNA secondary structures. *In* Proceedings of the fourth annual international conference on Computational Molecular Biology. *ACM*, pp. 201–209.

Massey, W.S. 1967. *Algebraic Topology: An Introduction*. Springer-Verlag, New York.

McManus, M.T., and Sharp, PA. 2002. Gene silencing in mammals by small interfering RNAs. *Nature Reviews Genetics* 3, 737–747.

Narberhaus, F., and Vogel, J. 2007. Sensory and regulatory RNAs in prokaryotes: A new german research focus. *RNA Biology* 4, 160–164.

Nussinov, R., Pieczenik, G., Griggs, J.R., and Kleitman, D.J. 1978. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics* 35, 68–82.

Orland, H., and Zee, A. 2002. RNA folding and large N matrix theory. *Nuclear Physics B* 620, 456–476.

Penner, R.C. 2004. Cell decomposition and compactification of Riemann's moduli space in decorated Teichmüller theory, 263–301. *In* Tongring, N., and Penner, R.C., ed. *Woods Hole Mathematics-Perspectives in Math and Physics*. World Scientific, Singapore.

Penner, R.C., Knudsen, M., Wiuf, C., and Andersen, J.E. 2010. Fatgraph models of proteins. *Comm. Pure Appl. Math.* 63, 1249–1297.

Penner, R.C., and Waterman, M.S. 1993. Spaces of RNA secondary structures. *Advances in Mathematics* 101, 31–49.

Reidys, C.M., Huang, F., Andersen, J.E. et al. 2011. Topology and prediction of RNA pseudoknots. *Bioinformatics* 27, 1076–1085.

Reidys, C.M., Wang, R.R., and Zhao, A.Y.Y. 2010. Modular, k-noncrossing diagrams. *The Electronic Journal of Combinatorics* 17, 1.

Richter, A.S., and Backofen, R. 2012. Accessibility and conservation: General features of bacterial small RNA–mRNA interactions? *RNA Biology* 9, 954–965.

Rivas, E., and Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.* 285, 2053–2068.

Vernizzi, G., and Orland, H. 2005. Large–N random matrices for RNA folding. *Acta PhysicA PolonicA Series B* 36, 2821.

Vernizzi, G., Orland, H., and Zee, A. 2005. Enumeration of RNA structures by matrix models. *Physical Review Letters* 94, 168103.

Waterman, M.S. 1978a. Combinatorics of RNA hairpins and cloverleaves. *Studies Appl. Math* 60, 91–96.

Waterman, M.S. 1978b. Secondary structure of single–stranded nucleic acids. *Adv. Math. Suppl. Studies* 1, 167–212.

Address correspondence to:
Christian M. Reidys
Department of Mathematics and Computer Science
University of Southern Denmark
Campusvej 55
DK-5230 Odense M
Denmark

E-mail: duck@santafe.edu