Generation of RNA pseudoknot structures with topological genus filtration

Fenix W.D. Huang[a], Markus E. Nebel[a,b], Christian M. Reidys[a,*]

[a] *Department of Mathematic and Computer science, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark*
[b] *Department of Computer Science, University of Kaiserslautern, Germany*

Abstract

In this paper we present a sampling framework for RNA structures of fixed topological genus. We introduce a novel, linear time, uniform sampling algorithm for RNA structures of fixed topological genus g, for arbitrary $g > 0$. Furthermore we develop a linear time sampling algorithm for RNA structures of fixed topological genus g that are weighted by a simplified, loop-based energy functional. For this process the partition function of the energy functional has to be computed once, which has $O(n^2)$ time complexity.

Keywords: RNA secondary structure, RNA pseudoknot structure, diagram, topological surface, topological genus, partition function, sampling

1. Introduction

Pseudoknots have long been known as important structural elements in RNA [1]. These cross-serial interactions between RNA nucleotides are functionally important in tRNAs, RNaseP [2], telomerase RNA [3], and ribosomal RNAs [4]. Pseudoknots in plant virus RNAs mimic tRNA structures, and *in vitro* selection experiments have produced pseudoknotted RNA families that bind to the HIV-1 reverse transcriptase [5]. Import general mechanisms, such as ribosomal frame shifting, are dependent upon pseudoknots [6].

*Corresponding author
Email addresses: fenixprotoss@gmail.com (Fenix W.D. Huang), nebel@informatik.uni-kl.de (Markus E. Nebel), duck@santafe.edu (Christian M. Reidys)

Lyngsø *et al.* [7] have shown that the prediction of general RNA pseudoknot structures is NP-complete. Thus, in order to provide prediction tools of feasible time complexity one frequently sticks to subtle subclasses of pseudoknots suitable for the dynamic programming paradigm [8, 9]. Alternative approaches to the prediction of RNA secondary structure (with or without pseudoknots) build on random sampling of foldings compatible to a given sequence. Here both, the underlying probability model and the efficiency of the sampling algorithm are crucial for being successful.

In this paper we propose a linear time uniform random sampler for pseudoknotted RNA structures of given topological genus which might be considered a promising starting point for the design of efficient solutions to the structure prediction problem. Topological genus means the number of handles attached to a sphere and measures topological complexity of its associated surface [10]. In fact any topological surface is fully characterized by its orientability and its genus [10].

Our approach is based on the observation that pseudoknotted RNAs are in a natural way related to topological surfaces. In fact pseudoknotted RNA structures can be viewed as drawings on orientable surfaces of genus g, that is by means of the classical classification theorem either on the sphere (secondary structures) or connected sums of tori (pseudoknotted structures). Our approach is a natural evolution from Waterman *et al.* pioneering work [11, 12, 13] on secondary structures.

Secondary structures are coarse grained RNA contact structures, see Figure 1 (A). They can be represented as diagrams, i.e. labeled graphs over the vertex set $[n] = \{1, \ldots, n\}$ with vertex degrees ≤ 3, represented by drawing its vertices on a horizontal line and its arcs (i, j) $(i < j)$, in the upper half-plane, see Figure 1. We assume the vertices to be connected by the edges $\{i, i+1\}$, $1 \leq i < n$, which are not considered arcs (but contribute to a nodes's degree). Furthermore, vertices and arcs correspond to the nucleotides **A**, **G**, **U** and **C** and Watson-Crick base pairs (**A-U**, **G-C**) or wobble base pairs (**U-G**), respectively.

Considering only the Watson-Crick and wobble base pair RNA structures, we set the restriction that one vertex can only pair with at most another vertex. Let $i < r$, we call arcs (i, j) and (r, s) crossing if $i < r < j < s$ holds. In this representation a pseudoknot-free secondary structure is a diagram without crossing arcs. Otherwise, i.e. diagrams with crossings represent pseudoknot structures, see Figure 1 (B).

In this paper, we present a framework for generating diagrams with cross-

ings, filtered by topological genus. The topological filtration of RNA structures has first been proposed by Penner and Waterman in [14] and later, as an application of the Matrix model in [15] and [16]. In [17, 18] a representation theoretic ansatz is employed that traces back to Zagier [19]. [18] connects the naturally appearing RNA shapes of fixed topological genus with Riemann's moduli space. The work presented here, however, is based on the combinatorial work of Chapuy [20].

As for the sampling of structures, let us start with pseudoknot-free secondary structures. The Moztzkin path interpretation of the latter already implies a linear time uniform sampler. This sampler can by construction not be extended to pseudoknotted structures. McCaskill [21] introduced the partition function computation for RNA secondary structures without pseudoknot, which implied a Boltzmann sampler, subsequently improved by [22].

The partition function of a restricted class of pseudoknotted structures has first been presented by Rivas [8]. [17] computed the partition function and Boltzmann sampler of topological RNA structures using a fully loop-based energy model. In [23] a heuristic method has been employed, constructing a pseudoknotted structure via selecting certain, compatible helices employing a Monte Carlo algorithm.

Uniform generation however has not been covered by the partition function frameworks. This is mainly due to the fact that run times are prohibitively high. Namely, $O(n^4)$ for [24, 17] and in case of [23] no complexity analysis is given.

The work in spirit closest to this paper deals with an entirely different, combinatorial class of RNA structures [25], the k-noncrossing structures. These are pseudoknotted structures, that have the property to not contain any k, mutually crossing arcs. In [26] a linear time sampler, based on a Markov chain of representations of the symmetric group, is presented. The construction uses the so called reflection principle, which is the key to compute the transition probabilities. While a uniform sampling with linear time complexity is obtained, the work is methodologically completely disjoint to this paper.

In order to understand how topology enters the picture for RNA molecules we need to pass from diagrams or contact-graphs to that of topological surfaces. Only the associated surface carries the key invariants leading to a meaningful filtration of RNA structures. The mental picture here is to "thicken" the edges into (untwisted) bands and to expand each vertex to a disk as shown in Figure 2. This inflation of edges leads to a fatgraph $\mathbb{D}$

[27, 28].

A fatgraph, sometimes also called a "map", is a graph equipped with a cyclic ordering of the incident half-edges at each vertex. Thus, $\mathbb{D}$ refines its underlying graph D insofar as it encodes the ordering of the ribbons incident on its disks. In fact a fatgraph constitutes to a cell-complex structure – combinatorial data in a sense – that have a topological surface as geometric realization [10].

Our sampling process consists of two steps: first we generate a diagram without crossing arcs and second we lift the topological genus to some fixed g. The process has linear time and is thereby very efficient.

The paper is organized as follows: we first introduce the topological filtration of diagrams. Then we introduce a genus induction process and finally, we describe and analyze the sampling processes.

2. Some basic facts

2.1. Diagrams

A diagram is a labeled graph over the vertex set $[n] = \{1, \ldots, n\}$ in which each vertex has degree ≤ 3, represented by drawing its vertices in a horizontal line. The backbone of a diagram is the sequence of consecutive integers $(1, \ldots, n)$ together with the edges $\{\{i, i+1\} \mid 1 \leq i \leq n-1\}$. The arcs of a diagram, (i, j), where $i < j$, are drawn in the upper half-plane. We shall distinguish backbone edges $\{i, i+1\}$ from arcs $(i, i+1)$, which we refer to as a 1-arc. Two arcs (i, j), (r, s), where $i < r$ are crossing if $i < r < j < s$ holds. The arc $(1, n)$ is called rainbow, see Figure 3.

2.2. Fatgraphs and unicellular maps

In this section, we discuss the filtration of diagrams by topological genus. In order to extract topological properties of diagrams those need to be enriched to fatgraphs. The latter are tantamount to a cell-complex structures over topological surfaces. Formally, we make this transition [18] by "thickening" the edges of the diagram into (untwisted) bands or ribbons. Furthermore each vertex is inflated into a disc as shown in Figure 2 (B). This inflation of edges and vertices means to replace a set of incident edges by a sequence of half-edges. This constitutes the fatgraph $\mathbb{D}$ [27, 28].

A fatgraph is thus 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. Consequently, we have the following definition:

Definition 1. A fatgraph is a triple (H, σ, α), where σ is the vertex-permutation and α a fixed-point free involution.

In the following we will deal with orientable fatgraphs[1]. Each ribbon has two boundaries. The first one in counterclockwise order shall be labeled by an arrowhead, see Figure 2 (C).

A fatgraph $\mathbb{D}$ exhibits a phenomenon, not present in its underlying graph D. Namely, one can follow the (directed) sides of the ribbons rotating counterclockwise around the vertices. This gives rise to $\mathbb{D}$-cycles or boundary components, constructed by following these directed boundaries from disc to disc. Algebraically, this amounts to form the permutation $\gamma = \alpha \circ \sigma$.

In the following we consider only diagrams with rainbow. As we shall see, the rainbow arc provides a canonical first boundary component, which travels on top of the rainbow arc and around the backbone of the diagram, see Figure 4.

A fatgraph, $\mathbb{D}$, can be viewed as a "drawing" on a certain topological surface. $\mathbb{D}$ is a 2-dimensional cell-complex over its geometric realization, i.e. a surface without boundary, $X_{\mathbb{D}}$, realized by identifying all pairs of edges [10]. Key invariants of the latter, like Euler characteristic [10]

$$\chi(X_{\mathbb{D}}) = v - e + r, \tag{1}$$

$$g(X_{\mathbb{D}}) = 1 - \frac{1}{2}\chi(X_{\mathbb{D}}), \tag{2}$$

where v, e, r denotes the number of discs, ribbons and boundary components in $\mathbb{D}$ [10] are defined combinatorially. However, equivalence of simplicial and singular homology [29] implies that these combinatorial invariants are in fact invariants of $X_{\mathbb{D}}$ and thus topological. This means the surface $X_{\mathbb{D}}$ provides a topological filtration of fatgraphs.

Since, adding a rainbow or collapsing the backbone of a diagram does not change the Euler characteristic, the relation between genus and number of boundary components is solely determined by the number of arcs in the

[1]Here ribbons may also be allowed to twist giving rise to possibly non-orientable surfaces [10].

upper half-plane:

$$2 - 2g - r = 1 - n, \tag{3}$$

where n is number of arcs and r the number of boundary components. The latter can be computed easily and allows us therefore to obtain the genus of the diagram.

Definition 2. A unicellular map $\mathfrak{m}$ of size n is a fatgraph $\mathfrak{m}(n) = (H, \alpha, \sigma)$ in which the permutation $\alpha \circ \sigma$ is a cycle of length $2n$.

While unicellular maps are simply particular fatgraphs, they naturally arise in the context of diagrams, by two observations. First in the diagram one may collapse the backbone into a single vertex. Second the mapping

$$\pi \colon (H, \sigma, \alpha) \ \mapsto \ (H, \alpha \circ \sigma, \alpha),$$

is evidently a bijection between fatgraphs having one vertex and unicellular maps, see Figure 5. The mapping is called the *Poincaré dual* and interchanges boundary components by vertices, preserving topological genus. In the following, we use π to denote the Poincaré dual.

Given a unicellular map the permutation σ and γ induces two linear orders of half-edges

$$r <_\gamma \gamma(r) <_\gamma \cdots <_\gamma \gamma^{2n-1}(r), \quad r <_\sigma \sigma(r) <_\sigma \cdots <_\sigma \sigma^k(r).$$

Let a_1 and a_2 be two distinct half-edges in $\mathfrak{m}$. Then $a_1 <_\gamma a_2$ expresses the fact that a_1 appears before a_2 in the boundary component $\gamma = \alpha \circ \sigma$. Suppose two half-edges a_1 and a_2 belong to the same vertex v. Note that v is effectively a cycle which we assume to originate with the first half-edge along which one enters v traveling γ. Then $a_1 <_\sigma a_2$ expresses the fact that a_1 appears (counterclockwise) before a_2.

The Poincare-dual maps the rainbow into a distinguished vertex of degree one and provides thereby a natural origin for the cycle γ. We call this vertex the *plant*, see Figure 5. Given a unicellular map we call a half-edge the minimum half-edge of a vertex v if it is the first half-edge via which γ visits v.

2.3. Genus induction

In this section we present a construction of [20], which plays a key role for our main result. It consists of two processes: a slicing-map Ξ and a

gluing-map Λ, which, when restricted to the proper classes, are inverse to each other.

The slicing process splits a vertex into $(2g+1)$ vertices and thereby reduces the genus of the map by g. Gluing is effectively inverse to slicing, namely: gluing any $(2g+1)$ vertices in a unicellular map increases the genus of the map by g. Slicing and gluing preserve unicellularity.

Definition 3. A half-edge h is an *up-step* if $h <_\gamma \sigma(h)$, and a *down-step* if $\sigma(h) \leq_\gamma h$. h is called a *trisection* if h is a down-step and $\sigma(h)$ is not the minimum half-edge of its respective vertex.

The number of trisections in a unicellular map of genus g is given by the following lemma:

Lemma 1. *[20] Let $\mathfrak{m}$ be a unicellular map of genus g. Then $\mathfrak{m}$ has exactly $2g$ trisections.*

Given a unicellular map $\overline{\mathfrak{m}}$ and a vertex $\overline{v}$ together with a trisection τ contained in $\overline{v}$. Let a_1 be the minimum half-edge of $\overline{v}$. Then set $a_3 = \overline{\sigma}(\tau)$ and a_2 to be the smallest half-edge between a_1 and a_3 (with respect to the order $<_{\overline{\sigma}}$) such that $a_2 >_{\overline{\gamma}} a_3$.

Since τ is a trisection such an a_2 exists. Then we refer to the replacement of

$$\overline{v} = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1})$$

by the three vertices where $v_i = (a_i, h_i,^1, \ldots, h_i^{m_i})$, $i = 1, 2, 3$, see Figure 6, as slicing. Slicing produces the unicellular fatgraph $\overline{\mathfrak{m}} = (H, \overline{\sigma}, \alpha)$.

Conversely, let $\mathfrak{m}$ be a unicellular map and let a_1, a_2 and a_3 be three half-edges belonging to three distinct vertices, $v_i = (a_i, h_i,^1, \ldots, h_i^{m_i})$ for some $m_i \geq 0$ and $i = 1, 2, 3$. Furthermore suppose $a_1 <_\gamma a_2 <_\gamma a_3$.

Then, replacing the cycles v_1, v_2 and v_3 by the cycle

$$\overline{v} = (a_1, h_2^1, \ldots, h_2^{m_2}, a_2, h_3^1, \ldots, h_3^{m_3}, a_3, h_1^1, \ldots, h_1^{m_1}),$$

is referred to as gluing. Gluing produces the unicellular fatgraph $\overline{\mathfrak{m}} = (H, \overline{\sigma}, \alpha)$, see Figure 6, in which the half-edge $\overline{\sigma}^{-1}(a_3)$ is, by construction, a trisection.

Lemma 2. *[20] Slicing maps a unicellular map together with a trisection into a unicellular map together with three labeled vertices. Gluing maps a unicellular map together with three labeled vertices into a unicellular maps with a trisection.*

Suppose we slice $(\overline{\mathfrak{m}}, \tau)$ into $(\mathfrak{m}, v_1, v_2, v_3)$, where in $\overline{\mathfrak{m}}$ holds $a_1 <_{\overline{\gamma}} a_3 <_{\overline{\gamma}} a_2$. Then we observe that in $\mathfrak{m}$ a_1 remains minimum in its new vertex and so does a_2, because a_2 is by definition the minimum half-edge where $a_3 <_{\overline{\gamma}} a_2$. However, a_3 becomes either the minimum half-edge, or remains a half-edge following a trisection. This gives rise to *two* types of trisections:

Definition 4. Let $\overline{\mathfrak{m}}$ be a unicellular map and $\overline{v}$ a vertex containing a trisection τ. Slicing $(\overline{\mathfrak{m}}, \tau)$ we obtain $(\mathfrak{m}, v_1, v_2, v_3)$. If the minimum half-edge of v_3, denoted by a_3 is the half-edge $\overline{\sigma}(\tau)$ in $\overline{\mathfrak{m}}$, we call the trisection τ to be of *type I* and *type II*, otherwise.

Proposition 1. *Let $\mathfrak{m}_g$ denote a unicellular map of genus g having n edges. Let furthermore τ^I denote a trisection of type I and τ^{II} denote a trisection of type II. Then we have the mappings Φ and Ψ:*

$$\Phi(\mathfrak{m}_g, v_1, v_2, v_3) = (\mathfrak{m}_{g+1}, \tau^I), \quad \Psi(\mathfrak{m}_g, v_1, v_2, \tau) = (\mathfrak{m}_{g+1}, \tau^{II})$$

are bijections, where v_1, v_2 and v_3 denote three distinct vertices in $\mathfrak{m}_g$ and $\mathfrak{m}_{g+1}$ is a unicellular map of genus $g+1$ having n edges.

Here Φ generates the trisection τ^I in a unicellular map of genus $g+1$ and the trisection τ^{II} persists when applying the mapping Ψ.

Gluing can be described as follow:
Given a unicellular map of $\mathfrak{m}_{g-k}$, together with a sequence of vertices $V = \{v_1, \dots v_{2k+1}\}$, where $v_i <_\gamma v_{i+1}$, $\forall 1 \leq i < 2k+1$. Then:
I. we glue the last three vertices v_{2k-1}, v_{2k} and v_{2k+1} via Φ, thereby obtaining the unicellular map $\mathfrak{m}_{g-k+1}$ together with trisection τ^I.
II. we apply $\Psi(\mathfrak{m}_{g-k+i}, v_{2k-2i-1}, v_{2k-2i}, \tau^I)$ $k-1$ times for $i = 1$ to $i = k-1$. This produces the unicellular map $\mathfrak{m}_g(n)$, together with a trisection τ^{II}. The process defines a mapping

$$\Lambda(\mathfrak{m}_{g-k}, v_1, \dots, v_{2k+1}) = (\mathfrak{m}_g, \tau),$$

where we do not label τ by type since in general we do not know whether Ψ has been applied. The order of the vertices in V is given by the partial order determined by γ. Thus V can be considered as a set of vertices in $\mathfrak{m}_{g-k}$, ordered by $<_\gamma$. Λ merges vertices from right to left by first applying Φ once then applying Ψ several times.

Λ is reversed as follows: given a unicellular map $\mathfrak{m}_g$ of genus g and $i = 0$:
1. if τ is type II trisection in $\mathfrak{m}_{g-i}$, then let $(\mathfrak{m}_{g-i-1}, v_{2i+1}, v_{2i+2}, \tau) =$

$\Psi^{-1}(\mathfrak{m}_{g-i}, \tau)$. We increase i to $i+1$ and repeat step **1**.
2. if τ has type I, let $(\mathfrak{m}_{g-i}, v_{2i+1}, v_{2i+2}, v_{2i+3}) = \Phi^{-1}(\mathfrak{m}_{g-i-1}, \tau)$.
Then we return

$$\Xi(\mathfrak{m}_g, \tau) = (\mathfrak{m}_{g-i}, V_\tau).$$

By construction, Λ and Ξ are inverse to each other.

Theorem 1. *[20] Let U_g^t denote the set of tuples $(\mathfrak{m}_g, v_1, \ldots, v_t)$, where $v_1, \ldots, v_t$ is a sequence of vertices in $\mathfrak{m}_g$. Furthermore, let D_g denote the set of tuples $(\mathfrak{m}_g, \tau)$, where τ is a trisection of $\mathfrak{m}_g$. Then*

$$\Lambda \colon \dot{\bigcup_{k=0}^{g-1}} U_k^{2g-2k+1} \to D_g, \quad \Xi \colon D_g \to \dot{\bigcup_{k=0}^{g-1}} U_k^{2g-2k+1}$$

are bijections and $\Lambda \circ \Xi = \mathrm{id}$ and $\Xi \circ \Lambda = \mathrm{id}$.

Let $\epsilon_g(n)$ denote the number of unicellular map of genus g having n edges. Then we have the following enumerative corollary

Corollary 1.

$$2g \cdot \epsilon_g(n) = \binom{n+1-2(g-1)}{3}\epsilon_{g-1}(n) + \cdots + \binom{n+1}{2g+1}\epsilon_0(n). \tag{4}$$

Here the $2g$-factor on left hand side counts the number of trisections in $\mathfrak{m}_g$ and the binomial coefficients on the right hand side counts the number of distinct selections of subsets of $(2k+1)$ vertices from a unicellular map $\mathfrak{m}_{g-k}$.

Iterating Ξ, we obtain

$$\epsilon_g(n) = \sum_{0=g_0<g_1<\cdots<g_r=g} \prod_{i=1}^{r} \frac{1}{2g_i}\binom{n+1-2g_{i-1}}{2(g_i-g_{i-1})+1} \cdot \epsilon_0(n), \tag{5}$$

where $\epsilon_0(n)$ is the number of planar trees having n edges, i.e. the Catalan number $\frac{1}{n+1}\binom{2n}{n}$.

3. Uniform generation of matchings

In this section, we show how to generate a matching of a given genus g over $2n$ vertices with uniform probability.

Any unicellular map $\mathfrak{m}_g$ together with one of its $2g$ trisections is mapped via Ξ into a unicellular map of lower genus. Note that the genus decreases at least by one. Therefore, by iterating the process finitely many times (at most g), we arrive at a unicellular map of genus 0, i.e a planar tree.

For our construction it is important to keep track of the particular slicing process. Accordingly, we introduce *slice/glue paths* as follows.

Definition 5. Suppose $\mathfrak{m}_g$ is a unicellular map of genus g having n edges. Then a sequence of unicellular maps

$$(\mathfrak{m}^0 = \mathfrak{m}_{g_0=0}, \mathfrak{m}^1 = \mathfrak{m}_{g_1}, \ldots, \mathfrak{m}^r = \mathfrak{m}_{g_r=g})$$

is called a slice path from $\mathfrak{m}_g$ to $\mathfrak{m}_0$ and a glue path when considered from $\mathfrak{m}_0$ to $\mathfrak{m}_g$, where $\Xi(\mathfrak{m}_{g_i}, \tau_i) = (\mathfrak{m}_{g_{i-1}}, V_{g_{i-1}})$ holds for some τ_i in $\mathfrak{m}_{g_i}$, $0 < i \leq r$.

We next consider $P_g(\mathfrak{m}^0)$, the set of distinct glue paths from a given $\mathfrak{m}^0 = \mathfrak{m}_0$ to some unicellular maps of fixed genus g.

Lemma 3. *The cardinality of $P_g(\mathfrak{m}^0)$ is given by*

$$\sum_{0=g_0<g_1<\cdots<g_r=g} \prod_{i=1}^{r} \frac{1}{2g_i} \binom{n+1-2g_{i-1}}{2(g_i-g_{i-1})+1}.$$

Proof. In order to construct $\mathfrak{m}_{g_i}$ from $\mathfrak{m}_{g_{i-1}}$, $0 < i \leq r$, we need to select $2(g_i - g_{i-1}) + 1$ vertices from $\mathfrak{m}_{g_{i-1}}$.

Euler characteristic shows that there are $(n + 1 - 2g_{i-1})$ distinct vertices in $\mathfrak{m}_{g_{i-1}}$, whence there are $\binom{n+1-2g_{i-1}}{2(g_i-g_{i-1})+1}$ ways to select a subset of vertices $V_{g_{i-1}}$.

On the other hand, the mapping Λ produces $\mathfrak{m}_{g_i}$ with a labeled trisection τ_i, i.e., the same $\mathfrak{m}_{g_i}$ will be produced exactly $2g_i$ times. Accordingly, we need to normalize the production by a factor $1/2g_i$ for each application of Λ.

As a result the total number of glue paths in $P_g(\mathfrak{m}^0)$ is

$$\sum_{0=g_0<g_1<\cdots<g_r=g} \prod_{i=1}^{r} \frac{1}{2g_i} \binom{n+1-2g_{i-1}}{2(g_i-g_{i-1})+1},$$

which is exactly the coefficient of $\epsilon_0(n)$ in eq. (5). □

The problem of generating a unicellular map of genus g having n edges with uniform probability thus splits into two parts: we first generate a planar

tree $\mathfrak{m}_0$ with n edges with uniform probability. Second we generate a glue path from $P_g(\mathfrak{m}^0)$ with uniform probability. It is well-known how to implement the first step by a linear time [2] (rejection) sampler [31] and it thus remains to present an algorithm for the second step.

We construct a glue path inductively. Suppose we are at step i and we have constructed a unicellular map $\mathfrak{m}^i$ of genus g_i. Then the next genus g_{i+1} is suggested by the process `NextGenus`. This process considers the sequence of genus $g_0, \ldots, g_i$ and the target genus g as input, and returns the genus g_{i+1}. Let $\mathbb{P}(g_{i+1} = t \mid g_0, \ldots, g_i, g)$ denote the probability of g_{i+1} equals t under the condition that $g_0, \ldots, g_i$ are the genus of the previous steps and g is the target genus. Then

$$\mathbb{P}(g_{i+1} = t \mid g_0, \ldots, g_i, g) = \frac{\sum_{t_0=g_0,\ldots,t_i=g_i,g_{i+1}=t<t_{i+1}<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}}{\sum_{t_0=g_0,\ldots,t_i=g_i<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}}. \tag{6}$$

Next we select the sequence of vertices from $\mathfrak{m}^i$ by process `SelectVertex`. This process chooses vertices in $2(g_{i+1} - g_i) + 1$ independent steps. The probability of a vertex being selected is given by $1/(n + 2 - 2g_i - k)$, where $(n + 2 - 2g_i - k)$ is the number of remaining non-selected vertices in the kth step, $1 \leq k \leq 2(g_{i+1} - g_i) + 1$. Since the selected vertices are ordered automatically by $<_{\gamma_{\mathfrak{m}^i}}$, the same set is generated with multiplicity $(2(g_{i+1} - g_i) + 1)!$. Normalizing the resulting term by the factor $1/(2(g_{i+1} - g_i) + 1)!$, the probability of the set V_i, $0 \leq i < r$ is given by

$$\mathbb{P}_{\texttt{Select}}(V_i) = \frac{1}{(2(g_{i+1} - g_i) + 1)!} \cdot \frac{1}{n + 1 - 2g_i} \cdots \frac{1}{n - 2g_{i+1}} = \frac{1}{\binom{n+1-2g_i}{2(g_{i+1}-g_i)+1}}. \tag{7}$$

After the sequence of vertices V_i is selected, a unicellular map $\mathfrak{m}^{i+1}$ is constructed by the process `Glue`, applying mapping Λ. We present the pseudocode of the procedures in Algorithm 1.

[2] Please note that we refer to an average-case linear time approximate-size sampler here, which for given n guarantees a size in $[n(1-\epsilon), n(1+\epsilon)]$ for arbitrarily small but fixed $\epsilon > 0$. An exact-size rejection sampler might have expected quadratic time and we may switch to a boustrophedonic algorithm as presented in [30] with a worst-case runtime in $O(n \log(n))$.

Algorithm 1

```
1: UniformMatching (𝔪^0, TargetGenus)
2: i ← 0
3: while g_i ≤ TargetGenus do
4:    g_{i+1} ← NextGenus (g_0, ..., g_i, TargetGenus)
5:    V_i ← SelectVertex (𝔪^i, 2(g_{i+1} − g_i) + 1)
6:    𝔪^{i+1} ← Glue (𝔪^i, V_i)
7:    i ← i + 1
8: end while
9: return 𝔪^i
```

Assuming the target genus to be constant and taking into account that during our construction the genus is strictly increasing, the while-loop of Algorithm 1 is executed only a constant number of times. Using appropriate memorization techniques, `NextGenus` and `Glue` can be implemented in constant time and `SelectVertex` in linear time. Thus, combined with a linear time sampler for planar trees, our approach allows for the uniform generation of random matchings in time $O(n)$.

Lemma 4. *Given a planar tree $\mathfrak{m}^0$ with n edges and a genus g, the probability of a glue path p_g generated by Algorithm 1 is $\epsilon_0(n)/\epsilon_g(n)$.*

Proof. Assume a glue path

$$p_g = \{\mathfrak{m}^0 = \mathfrak{m}_{g_0=0}, \mathfrak{m}^1 = \mathfrak{m}_{g_1}, \ldots, \mathfrak{m}^r = \mathfrak{m}_{g_r=g}\}$$

is generated by Algorithm 1. Since for each step, the process of choosing the genus for the next step and selecting labeled vertices is independent, the probability of P_g is given by

$$\mathbb{P}(P_g) = \prod_{i=0}^{r-1} \mathbb{P}(g_{i+1} = t_{i+1} | g_0, \ldots, g_i, g) \cdot \mathbb{P}(V_i).$$

We substitute eq. (6) and eq. (7) and obtain

$$\begin{aligned}
\mathbb{P}(P_g) &= \prod_{i=0}^{r-1} \frac{\sum_{t_0=g_0,\ldots,t_i=g_i,g_{i+1}=t<t_{i+1}<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}}{\sum_{t_0=g_0,\ldots,t_i=g_i<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}} \cdot \frac{1}{\binom{n+1-2g_i}{2(g_{i+1}-g_i)+1}} \\
&= \frac{\prod_{t_1=g_1,\ldots,t_r=g_r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}}{\sum_{0=t_0<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}} \cdot \prod_{i=0}^{r-1} \frac{1}{\binom{n+1-2g_i}{2(g_{i+1}-g_i)+1}} \\
&= \frac{1}{\sum_{0=t_0<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}} \\
&= \frac{\epsilon_0(n)}{\epsilon_0(n) \cdot \sum_{0=t_0<\cdots<t_r=g} \prod_{i=1}^{r} \frac{1}{2t_i}\binom{n+1-2t_{i-1}}{2(t_i-t_{i-1})+1}} \\
&= \frac{\epsilon_0(n)}{\epsilon_g(n)},
\end{aligned}$$

whence the lemma. □

Corollary 2. *Suppose a planar tree $\mathfrak{m}^0$ is uniformly generated, i.e., with probability $1/\epsilon_0(n)$. Then a unicellular map $\mathfrak{m}^r = \mathfrak{m}_g$ is uniformly generated by Algorithm 1 with probability $1/\epsilon_g(n)$.*

4. Uniform generation of diagrams

In this section, we extend our result of Section 3 in order to generate diagrams of genus g with uniform probability. The idea is to uniformly generate first a matching of genus g with n arcs. In a second step we choose $(\ell - 2n)$ unpaired vertices and insert them into the matching.

Let $\mathbb{P}_d(t = n|\ell, g)$ denote the probability of the diagram having exactly n arcs, $0 \leq n \leq \lfloor \ell/2 \rfloor$. In the following we compute $\mathbb{P}_d(t = n|\ell, g)$.

Let $\delta_g(\ell)$ denote the number of diagrams of genus g over ℓ vertices. Furthermore, let $\delta_g(\ell, n)$ denote the number of diagrams of genus g over ℓ vertices having exactly n arcs, $2n \leq \ell$. Then $\ell - 2n$ vertices are unpaired and

$$\delta_g(\ell, n) = \binom{\ell}{\ell - 2n} \epsilon_g(n).$$

Furthermore

$$\delta_g(\ell) = \sum_{n=0}^{\lfloor l/2 \rfloor} \delta_g(\ell, n) = \sum_{n=0}^{\lfloor l/2 \rfloor} \binom{\ell}{\ell - 2n} \epsilon_g(n). \tag{8}$$

In order to generate a diagram of genus g over ℓ arcs uniformly, we need to solve

$$\frac{1}{\delta_g(\ell)} = \mathbb{P}_d(t = n|\ell, g) \cdot \frac{1}{\epsilon_g(n)} \cdot \frac{1}{\binom{\ell}{\ell-2n}},$$

whence $\mathbb{P}_d(t = n|\ell, g) = \delta_g(\ell, n)/\delta_g(\ell)$.

We present the pseudocode of `UniformDiagram` as Algorithm 2. The subroutine `NumberofArcs` returns n with probability $\mathbb{P}_d(t = n|\ell, g)$, which determines the number of arcs in diagram $D_g(\ell)$. `UnifomTree` is a standard process uniformly generating a matching of genus 0 with n arcs. Finally, the process `InsertUnpairedVertices` first chooses $(\ell - 2n)$ vertices from ℓ vertices as unpaired. It leaves $2n$ vertices not selected, which are considered to be paired. Then the process maps the $2n$ vertices of the matching generated by `UniformMatching` and keeps the arcs in the upper half-plane. Accordingly, a diagram of genus g over ℓ vertices with exactly n arcs is generated. The result of some experiments conducted in connection with the generation of random matchings and diagrams using our algorithms is shown in Figure 7.

Algorithm 2

1: `UniformDiagram` $(\ell, TargetGenus)$
2: $n \leftarrow$ `NumberofArcs`(ℓ, g)
3: $\mathfrak{m}_0 \leftarrow$ `UnifomTree`(n)
4: $\mathfrak{m}_g \leftarrow$ `UniformMatching`$(\mathfrak{m}, TargetGenus)$
5: $D_g \leftarrow$ `InsertUnpairedVertices`$(\mathfrak{m}_g, \ell)$
6: **return** D_g

5. Non-uniform sampling

RNA structures can be represented as diagrams and are, due to the biophysical context subject to certain constraints with respect to their free energy [32]. The latter energy is oftentimes modeled as a function of the loops of the underlying RNA structure [32], S. These loops are in fact equal to the boundary components of the fatgraph constructed from the molecule. In the following we shall discuss, $\eta(S)$, a simplified version of the actual bio-physical loop-energy of a structure S.

Let us start with RNA secondary structures without pseudoknot, that correspond to diagrams of genus 0. For a pseudoknot-free secondary structure S_0, we denote its corresponding (see Section 2, duality mapping π) unicellular map by $\mathfrak{m}_0 = \pi(S_0)$. The bonds or arcs of the structure then correspond to edges of the unicellular map $\mathfrak{m}_0$ and loops or boundary components to vertices. Three types of loops are distinguished: hairpin loops, interior loops (including helices and bludge loops) and multi-loops. Accordingly, the duality maps hairpin loops into vertices of degree one, interior loops into vertices of degree two, and multi-loop into vertices of degree greater than two, see Figure 8. $\eta(S)$ extends these types in order to deal with structures having arbitrary genus $g \geq 0$ as follows.

Let S_g denote an RNA structure having length ℓ, n arcs and genus g, and $\mathfrak{m}_g = \pi(S_g)$ its corresponding unicellular map. Then $\eta(S_g)$ is given by

$$\eta(S_g) = n \cdot b + \sum_{v \in \mathbb{V}} T(v) + L_g^{pk}. \tag{9}$$

Here b represents an energy contribution of arcs, $\mathbb{V}$ the set of all vertices $\mathfrak{m}_g$, $T(v)$ is function given by

$$T(v) = \begin{cases} L^{hp} & \text{if } d(v) = 1, \\ L^{int} & \text{if } d(v) = 2, \\ L^{mul} & \text{if } d(v) > 2 \text{ and } v \text{ has no trisection}, \\ 0 & \text{if } v \text{ is attached to the root}, \end{cases}$$

where $d(v)$ is the degree of vertex v, and L^X is the contribution of a loop of type X, where $X = \{hp, int, mul\}$. Finally, L_g^{pk} represents a contribution that stems from novel loop-types emerging for genus $g > 0$. In this model, we do not take contributions from unpaired vertices into account.

In case of $g = 1$, there are four different types of pseudoknots [33], see Figure 9. This is analogous for any genus: there are always only finitely many corresponding shadows [34, 33], see Figure 9. Here, a shadow is a diagram without unpaired vertices in which all stacks (parallel arcs) have size one. Formally, a shadow of a structure can be obtained by first removing all its unpaired vertices, second removing all noncrossing arcs (together with their vertices) and then replacing a set of parallel arcs (and the incident vertices) of the form $\{(i, j), (i+1, j-1), \ldots, (i+\ell-1, j-\ell+1)\}$ by a single arc (and

two vertices).

Let us have a closer look at the boundary components of these shadows in case of genus 1, as shown in Figure 9. (H) is inspected to have one boundary component, whence $\eta(S_{\mathrm{H}}) = 2b + L^{mul} + L_1^{pk}$. (K) and (L) have two boundary components and accordingly $\eta(S_{\mathrm{K}}) = \eta(S_{\mathrm{L}}) = 3b + 2L^{mul} + L_1^{pk}$. Finally, for (M) we have $\eta(S_{\mathrm{M}}) = 4b + 3L^{mul} + L_1^{pk}$.

Consider a matching S_1 of genus 1 having n arcs and $\mathfrak{m}_1 = \pi(S_1)$ the unicellular map given by the duality. By selecting a trisection τ in $\mathfrak{m}_1$ and applying the mapping Ξ, we obtain $\Xi(\mathfrak{m}_1, \tau) = (\mathfrak{m}_0, v_1, v_2, v_3)$ and three labeled vertices. Here we write $\mathfrak{m}_0^{(3)} = (\mathfrak{m}_0, v_1, v_2, v_3)$ for short. Let $S_0^{(3)}$ denote a pseudoknot-free secondary structure with three labeled boundary component where $\pi(S_0^{(3)}) = \mathfrak{m}_0^{(3)}$. Let further $\mathbb{S}_{0,n}^{(3)}$ and $\mathbb{S}_{1,n}$ denote the set of S_1 and $S_0^{(3)}$ respectively. By Lemma 1, there are two trisections in $\mathfrak{m}_1$. Therefore, by selecting the same S_1 and different trisection τ, Ξ results in different $S_0^{(3)}$. Thus we have the cardinality $2|\mathbb{S}_{1,n}| = |\mathbb{S}_{0,n}^{(3)}|$. Figure 9 shows this for the four shadows of genus 1 and their pseudoknot-free secondary structure with three labeled boundary components.

We next formulate an "energy" for structures $S_0^{(3)}$, $\eta(S_0^{(3)})$, that matches the energy η for their corresponding counterpart of genus one after gluing. Note that this allows us to reduce everything to pseudoknot-free secondary structures with three labeled boundary components. To this end, let v_1, v_2 and v_3 be three labeled vertices in $\mathfrak{m}_0^{(3)}$, where $\mathfrak{m}_0^{(3)} = \pi(S_0^{(3)})$. Setting $T(v_1) = T(v_2) = T(v_3) = (L_1^{pk} + L^{mul})/3$ we observe

Proposition 2. *We have* $\eta(S_1) = \eta(S_0^{(3)})$.

Proof. The mapping Ξ is a bijection and $\Lambda(\mathfrak{m}_0, v_1, v_2, v_3) = (\mathfrak{m}_1, \tau)$. The three labeled vertices in $\mathfrak{m}_0$ are glued as $\overline{v}$, where $d(\overline{v}) \geq 3$. Hence $T(\overline{v}) = L^{mul} + L_1^{pk} = T(v_1) + T(v_2) + T(v_3)$, because $T(v_1) = T(v_2) = T(v_3) = (L_1^{pk} + L^{mul})/3$. The other vertices in $\mathfrak{m}_0$ maintain hence their scores are not changed. □

Given $\eta(S)$ we proceed along the lines of [21] and construct a probability space of structures by computing the partition function of a given sequence. Let $\theta(n) = \sum_{S \in \mathbb{S}_n} e^{\eta(S)}$ denote the total energy of all structures. A structure, S, is sampled with probability $e^{\eta(S)}/\theta(n)$. In case of pseudoknot-free secondary structures, loop-based and arc-based energy models are compatible to the standard recursion of pseudoknot-free secondary structure [35] and

$\theta_0(n)$ can be computed by the recursion

$$\theta(n) = \sum_{i=1}^{n-2} \theta(i)\theta(n-i-1) + e^{L^{hp}+b} \cdot \theta(1) + e^{L^{int}+b} \cdot \theta(n-1) + e^{L^{mul}+b} \cdot \sum_{i=1}^{n-3} \theta(i)\theta(n-i-2),$$

where η is the energy functional, discussed above. As there is only one summation in the above recursion, $\theta(n)$ is computed in $O(n^2)$ time.

We proceed by showing that the new functional, $\eta(S_0^{(3)})$, is also compatible with the pseudoknot-free secondary structure recursions.

Lemma 5. *Let* $\theta_1(n) = \sum_{S \in \mathbb{S}_{1,n}} e^{\eta(S)}$ *and* $\theta_0^{(3)}(n) = \sum_{S \in \mathbb{S}_{0,n}^{(3)}} e^{\eta(S)}$. *Then* $\theta_1(n) = \theta_0^{(3)}(n)/2$ *can be computed in* $O(n^2)$ *time. Once* $\theta_1(n)$ *is computed, a structure of genus one,* S_1*, is sampled with probability* $e^{\eta(S_1)}/\theta_1(n)$ *in* $O(n)$ *time.*

Proof. We have $\eta(S_1) = \eta(S_0^{(3)})$ for all $S_1 \in \mathbb{S}_{1,n}$ and $S_0^{(3)} \in \mathbb{S}_{0,n}^{(3)}$, and $2|\mathbb{S}_{1,n}| = |\mathbb{S}_{0,n}^{(3)}|$. Therefore,

$$\theta_1(n) = \sum_{S \in \mathbb{S}_{1,n}} e^{\eta(S)} = \frac{1}{2} \sum_{S \in \mathbb{S}_{0,n}^{(3)}} e^{\eta(S)} = \frac{1}{2}\theta_0^{(3)}(n).$$

We next show that $\theta_0^{(3)}(n)$ can be computed in $O(n^2)$ time. Let $\theta_0^{(2)} = \sum_{S \in \mathbb{S}_{0,n}^{(2)}} e^{\eta(S)}$ and $\theta_0^{(1)} = \sum_{S \in \mathbb{S}_{0,n}^{(1)}} e^{\eta(S)}$, where $\mathbb{S}_{0,n}^{(2)}$ and $\mathbb{S}_{0,n}^{(1)}$ denote the sets of pseudoknot-free secondary structures with two and one labeled boundary components. The functionals of these labeled boundary components are computed exactly as in the case of $S_0^{(3)}$.

Then we have, see also Figure 10:

$$\begin{aligned}
\theta_0^{(3)}(n) &= 2\sum_{i=1}^{n-2}\theta_0^{(3)}(i)\theta_0(n-i-1) + 2\sum_{i=1}^{n-2}\theta_0^{(2)}(i)\theta_0^{(1)}(n-i-1) \\
&+ e^{(L^{mul}+L_1^{pk})/3+b}\cdot\left(2\sum_{i=1}^{n-2}\theta_0^{(2)}(i)\theta_0(n-i-1) + \sum_{i=1}^{n-2}\theta_0^{(1)}(i)\theta_0^{(1)}(n-i-1)\right) \\
&+ 2e^{L^{mul}+b}\cdot\left(\sum_{i=1}^{n-3}\theta_0^{(3)}(i)\theta_0(n-i-2) + \sum_{i=1}^{n-3}\theta_0^{(2)}(i)\theta_0^{(1)}(n-i-2)\right) \\
&+ e^{(L^{mul}+L_1^{pk})/3+b}\cdot\left(2\sum_{i=1}^{n-3}\theta_0^{(2)}(i)\theta_0(n-i-2) + \sum_{i=1}^{n-3}\theta_0^{(1)}(i)\theta_0^{(1)}(n-i-2)\right) \\
&+ e^{L^{int}}\cdot\theta_0^{(3)}(n-1) + e^{(L_1^{pk}+L^{mul})/3+b}\cdot\theta_0^{(2)}(n-1).
\end{aligned}$$

Analogously, we have recursions for $\theta_0^{(2)}(n)$ and $\theta_0^{(1)}(n)$, which can be computed in $O(n^2)$ time. Therefore, $\theta_1(n) = \theta_0^{(3)}(n)$ can be computed in $O(n^2)$ time.

In order to sample a diagram of genus 1 over ℓ vertices, D_1, we need first to determine its number of arcs. As in the case of uniform sampling, we have $\vartheta_1(\ell) = \sum_{n=0}^{\lfloor \ell/2 \rfloor}\vartheta_1(\ell,n)$, where $\vartheta_1(\ell,n) = \binom{\ell}{\ell-2n}\theta_1(n)$. Replacing in the formulae for uniform sampling $\epsilon_1(n)$ by $\theta_1(n)$ and $\delta_1(n)$ by $\vartheta_1(n)$, we find that the probability of sampling a diagram with n arcs is given by $\vartheta_1(\ell,n)/\vartheta_1(\ell)$. It remains to sample a matching $S_0^{(3)}$ with n arcs and to subsequently glue the three labeled vertices in $\mathfrak{m}_0^{(3)} = \pi(S_0^{(3)})$. This generates a unicellular map of genus one, $\mathfrak{m}_1$, which is associated to $S_1 = \pi^{-1}(\mathfrak{m}_1)$ by duality. Note that choosing different slice-paths for S_1 generates two different $S_0^{(3)}$, see eq. (10).

$$\begin{array}{ccc}
S_0^{(3)} & \longrightarrow & S_1 \\
\downarrow & & \uparrow \\
(\mathfrak{m}_0, v_1, v_2, v_3) & \xrightarrow{\Lambda} & (\mathfrak{m}_1, \tau)
\end{array} \tag{10}$$

The probability of a structure of genus one, S_1, is then given by

$$\frac{2e^{\eta(S_0^{(3)})}}{\theta_0^{(3)}(n)} = \frac{2e^{\eta(S_1)}}{2\theta_1(n)} = \frac{e^{\eta(S_1)}}{\theta_1(n)}.$$

Finally, we insert the unpaired vertices into S_1 and obtain D_1 with the probability

$$\frac{\vartheta_1(\ell,n)}{\vartheta_1(\ell)} \cdot \frac{\binom{\ell}{\ell-2n} e^{\eta(D_1)}}{\theta_1(n)} = \frac{e^{\eta(D_1)}}{\vartheta_1(\ell)}.$$

□

6. Conclusion

In this paper we have proposed an original and highly efficient (linear time) approach to sample random RNA pseudoknotted structures in the uniform and a non-uniform model. The later builds on a simplified concept of free energy, favoring foldings of a native appearance. This is a first step towards efficient prediction algorithms for pseudoknotted RNA since structure predictions of good quality can easily be derived from suitable (high quality) random samples (see [36] and the references given there). To this end, our algorithms need to be extended towards two directions:

1. The probability model needs to be improved further, and
2. the RNA sequence needs to be taken into account.

Our sampler represents a paradigm shift from the original uniform sampler for pseudoknot-free secondary structures to topological structures of fixed topological genus in the following sense: the former employs an induction on the length of the path while the former employs an induction on the topological genus.

By construction, the uniform sampler is different from the Boltzmann sampler derived from the partition function. However, when we sample non-uniformly, we reduce essentially to an "enhanced" Boltzmann sampler as follows: the key point is that we translate topological genus into labeling schemes and then show compatibility of the Boltzmann sampler with these labellings. This allows us to sample non-uniformly w.r.t. the simplified energy model, introduced in Section 5, in $O(n^2)$ time.

An immediate application of the uniform sampler is the distributions of loops in structures of genus g. We have shown that the loops in structures are translated into vertices of their associated unicellular maps. In particular, a hairpin loop corresponds to a vertex of degree one, an interior loop to a vertex of degree two and a multi-loop is to some vertex having degree greater than two without a trisection. Finally a pseudoknot loop corresponds to a

vertex having degree greater than two containing a trisection. In Fig. 11 we present the respective data, filtered by genus.

It is well known in context of pseudoknot-free secondary structures how to use either a sophisticated model for the free energy or stochastic concepts like the maximum likelihood approach to obtain realistic probability models applicable to random sampling. Our approach seems to be suitable to apply the latter and it is a topic for future research to work out the details. Incorporating the sequence is a more complicated task but again results for classic RNA secondary structures prove it feasible with only small losses in efficiency [37].

Thus we assume our findings of this paper an important contribution towards the development of efficient structure prediction tools for pseudoknotted RNA structures. Those are also in need for state of the art tools addressing the inverse folding problem with similar efficiency of those for pseudoknot-free secondary structures. The latter employ search heuristics based on genetic algorithms [38, 39] in order to process the space of possible sequences using structure prediction tools to judge the quality (similarity to input) of current solutions. For the large number of calls, the efficiency of the prediction algorithm is crucial for the applicability of the entire approach. Today's established algorithms for the prediction of pseudoknotted RNA with run times at least $O(n^4)$ [9, 39] are not suited for this task.

7. Acknowledgments.

We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme (FP7) for Research of the European Commission, under the FET-Proactive grant agreement TOPDRIM, number FP7-ICT-318121.

References

[1] E. Westhof, L. Jaeger, RNA pseudoknots, Curr. Opin. Struct. Biol. 2 (1992) 327–333.

[2] A. Loria, T. Pan, Domain structure of the ribozyme from eubacterial ribonuclease, RNA 2 (1996) 551–563.

[3] D. W. Staple, S. E. Butcher, Pseudoknots: RNA structures with diverse functions, PLoS Biol. 3 (2005) e213.

[4] D. Konings, R. Gutell, A comparison of thermodynamic foldings with comparatively derived structures of 16s and 16s-like rRNAs, RNA 1 (1995) 559–574.

[5] C. Tuerk, S. MacDougal, L. Gold, RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase, Proc. Natl. Acad. Sci. USA 89(15) (1992) 6988–6992.

[6] M. Chamorro, N. Parkin, H. E. Varmus, An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA, Proc. Natl. Acad. Sci. USA 89(2) (1992) 6988–6992.

[7] R. B. Lyngsø, C. N. Pedersen, RNA pseudoknot prediction in energy-based models, J. Comp. Biol. 7 (2000) 409–427.

[8] E. Rivas, S. R. Eddy, A dynamic programming algorithm for RNA structure prediction including pseudoknots, J. Mol. Biol. 285 (1999) 2053–2068.

[9] M. E. Nebel, F. Weinberg, Algebraic and combinatorial properties of common RNA pseudoknot classes with applications, Journal of Computational Biology 19 (2012) 1134–1150.

[10] W. S. Massey, Algebraic Topology: An Introduction, Springer-Veriag, New York, 1967.

[11] M. S. Waterman, Combinatorics of RNA hairpins and cloverleaves, Stud. Appl. Math. 60 (1979) 91–96.

[12] R. Nussinov, G. Piecznik, J. R. Griggs, D. J. Kleitman, Algorithms for loop matching, SIAM J. Appl. Math. 35 (1) (1978) 68–82.

[13] D. Kleitman, Proportions of irreducible diagrams, Studies in Appl. Math. 49 (1970) 297–299.

[14] R. C. Penner, M. S. Waterman, Spaces of RNA secondary structures, Adv. Math. 101 (1993) 31–49.

[15] H. Orland, A. Zee, RNA folding and large n matrix theory, Nuclear Physics B 620 (2002) 456–476.

[16] M. Bon, G. Vernizzi, H. Orland, A. Zee, Topological classification of RNA structures, J. Mol. Biol. 379 (2008) 900–911.

[17] C. M. Reidys, F. Huang, J. E. Andersen, R. C. Penner, P. F. Stadler, M. E. Nebel, Topology and prediction of RNA pseudoknots, Bioinformatics 27 (2011) 1076–1085.

[18] J. E. Andersen, R. C. Penner, C. M. Reidys, M. S. Waterman, Topological classification and enumeration of RNA structrues by genus, J. Math. Biol.Prepreint.

[19] D. Zagier, On the distribution of the number of cycles of elements in symmetric groups, Nieuw Arch. Wisk. IV 13 (1995) 489–495.

[20] G. Chapuy, A new combinatorial identity for unicellular maps, via a direct bijective approach, Adv. Appl. Math. 47(4) (2011) 874–893.

[21] J. S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers 29 (1990) 1105–1119.

[22] Y. Ding, C. E. Lawrence, A statistical sampling algorithm for RNA secondary structure prediction, Nucleic Acids Res. 31 (2003) 7280–7301.

[23] M. Bon, C. Micheletti, H. Orland, Mcgenus: a monte carlo algorithm to predict RNA secondary structures with pseudoknots, Nucleic Acids Res. 41(3) (2013) 1895–900.

[24] R. M. Dirks, N. A. Pierce, A partition function algorithm for nucleic acid secondary structure including pseudoknots, J. Comput. Chem. 24 (2003) 1664–1677.

[25] W. Y. C. Chen, E. Y. P. Deng, R. R. X. Du, R. P. Stanley, C. H. Yan, Crossings and nestings of matchings and partitions, Trans. Amer. Math. Soc. 359 (2005) 1555–1575.

[26] W. Y. C. Chen, H. S. W. Han, C. M. Reidys, Random k-noncrossing RNA structures, Proc. Natl. Acad. Sci. USA 106(52) (2009) 22061–22066.

[27] M. Loebl, I. Moffatt, The chromatic polynomial of fatgraphs and its categorification, Adv. Math. 217 (2008) 1558–1587.

[28] R. C. Penner, M. Knudsen, C. Wiuf, J. E. Andersen, Fatgraph models of proteins, Comm. Pure Appl. Math. 63 (2010) 1249–1297.

[29] A. Hatcher, Algebraic Topology, Cambridge University Press, 2002.

[30] P. Flajolet, P. Zimmerman, B. V. Custem, A calculus for the random generation of labelled combinatorial structures, Theor. Comp. Sci. 132 (1994) 1–35.

[31] P. Duchon, P. Flajolet, G. Louchard, G. Schaeffer, Boltzmann samplers for the random generation of combinatorial structures, Combinatorics, Probability and Computing 13 (2004) 2004.

[32] D. Mathews, J. Sabina, M. Zuker, D. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, J. Mol. Biol. 288 (1999) 911–940.

[33] F. Huang, J. Qin, C. M. Reidys, P. F. Stadler, Target prediction and a statistical sampling algorithm for RNA-RNA interaction, Bioinformatics 26 (2010) 175–181.

[34] F. Huang, W. Peng, C. M. Reidys, Folding 3-noncrossing RNA pseudoknot structures, J. Comp. Biol. 16 (2009) 1549–1575.

[35] M. S. Waterman, Secondary structure of single-stranded nucleic acids, Adv. Math. (Suppl. Studies) 1 (1978) 167–212.

[36] M. E. Nebel, A. Scheid, Evaluation of a sophisticated scfg design for RNA secondary structure prediction., 2011, pp. 313–336.

[37] M. E. Nebel, A. Scheid, A n^2 RNA secondary structure prediction algorithm., in: Bioinformatics, 2012, pp. 66–75.

[38] A. Taneda, Multi-objective genetic algorithm for pseudoknotted RNA sequence design, Front Genet 3(36) (2012) PMCID: PMC3337422.

[39] J. Z. Gao, L. Y. Li, C. M. Reidys, Inverse folding of RNA pseudoknot structures, Algorithms Mol Biol. 5 (2010) R27.

Figures:

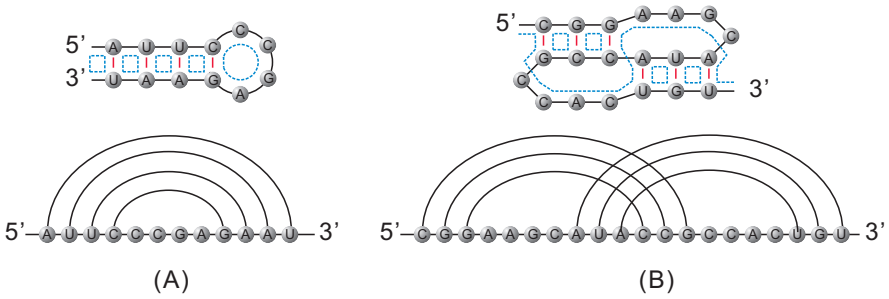

Figure 1: A pseudoknot-free secondary structure (A) and a pseudoknot structure (B) and their diagram representation.

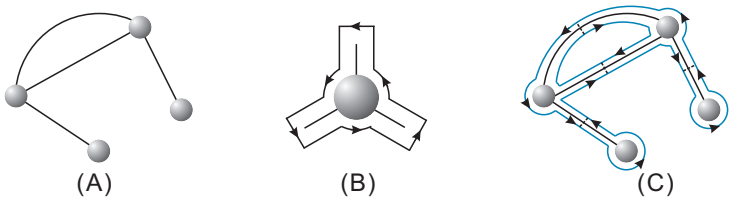

Figure 2: From graphs to fatgraphs: (A) A graph with 4 vertexes and 4 edges. (B) Inflation of a vertex. (C) A fatgraph derived from (A) induces a topological surface.

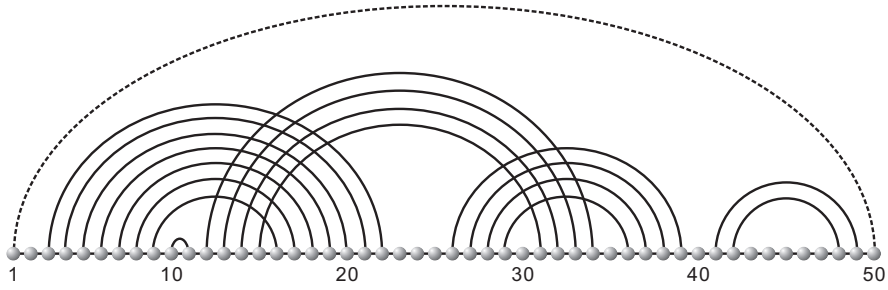

Figure 3: A diagram over 50 vertices. The arc $(10, 11)$ is a 1-arc. The arcs $(3, 22)$ and $(12, 34)$ are crossing. The dashed arc $(1, 50)$ is the rainbow.

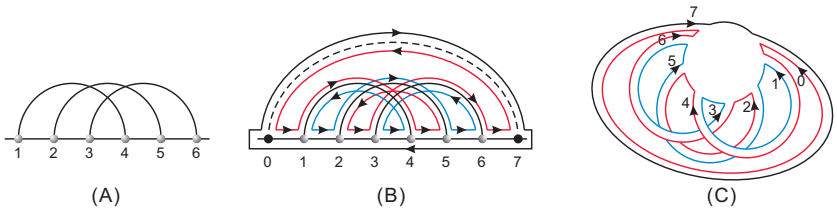

Figure 4: (A) A diagram. (B) the fattening of (A) augmented by the rainbow $(0, 7)$. Here $\sigma = (0, 1, 2, 3, 4, 5, 6, 7)$, $\alpha = (0, 7)(1, 4)(2, 5)(3, 6)$. Accordingly $\gamma = \alpha \circ \sigma = (0, 4, 2, 6)(1, 5, 3)(7)$ has two cycles. (C) Collapsing the backbone into a vertex.

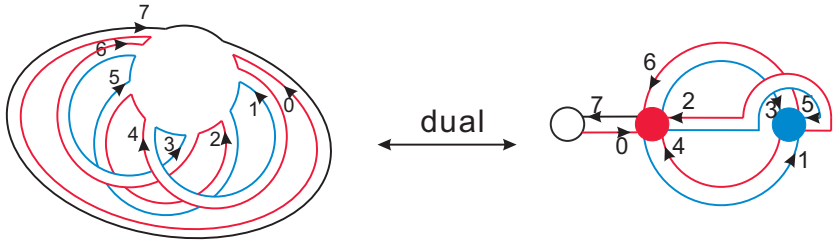

Figure 5: The Poincaré dual: we map a fatgraph with 1 vertex and 3 boundary components into a fatgraph with 3 vertexes and 1 boundary component.

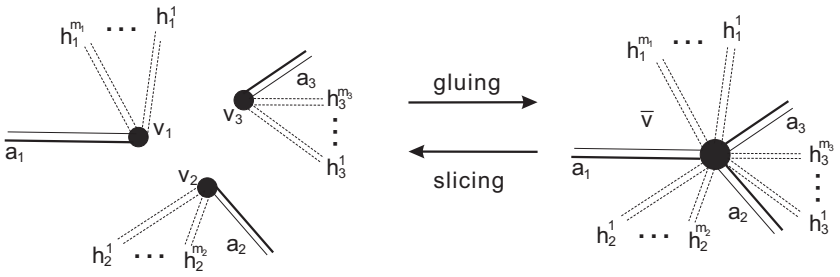

Figure 6: Illustration of gluing and slicing in a unicellular map.

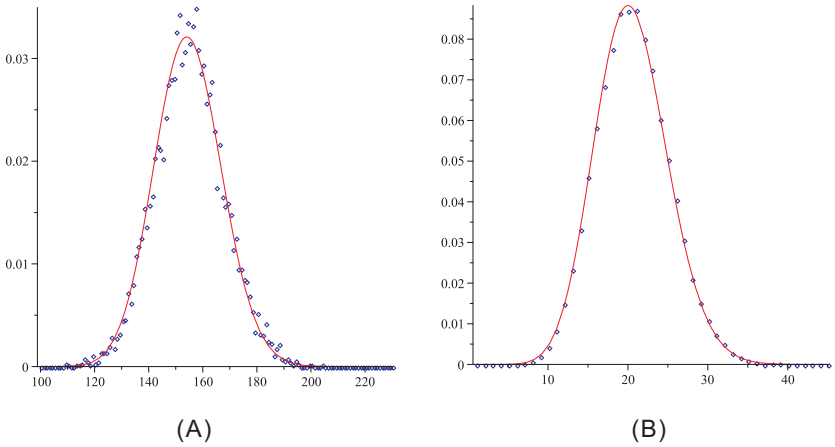

Figure 7: Uniform generation: (A) matchings, $n = 12$, $g = 2$ and $\epsilon_2(n/2) = 6468$. We generate $N = 10^6$ matchings and display the frequencies of their multiplicities (blue dots) together with the binomial coefficient of the uniform sampling $\binom{N}{\ell}(1/\epsilon_2(n/2))^\ell(1 - 1/\epsilon_2(n/2))^{N-\ell}$ (red). (B) The analog of (A) for diagrams. Here we have $n = 12$, $g = 2$ and $\delta_2(n) = 48741$. We generate $N = 10^6$ diagrams and display the frequencies of their multiplicities (blue dots) together with the binomial coefficients $\binom{N}{\ell}(1/\delta_2(n))^\ell(1 - 1/\delta_2(n))^{N-\ell}$ (red).

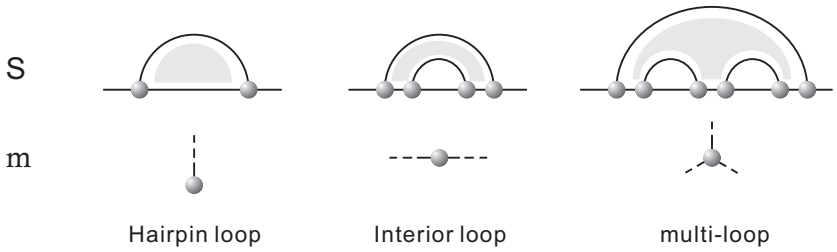

Figure 8: Hairpin loops, interior loops and multi-loops in secondary structures and their translation into unicellular maps.

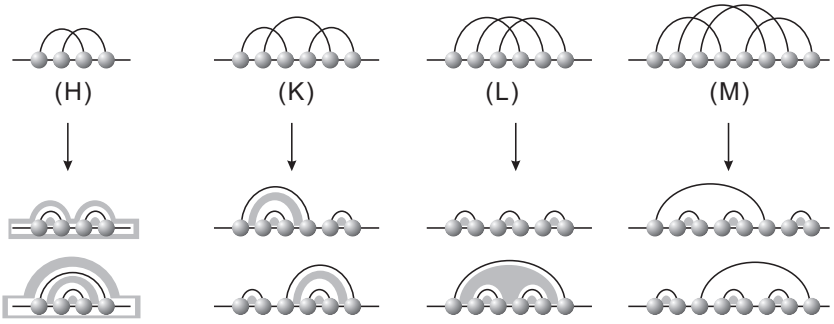

Figure 9: Correspondence between shadows of genus one and their labeled secondary structures.

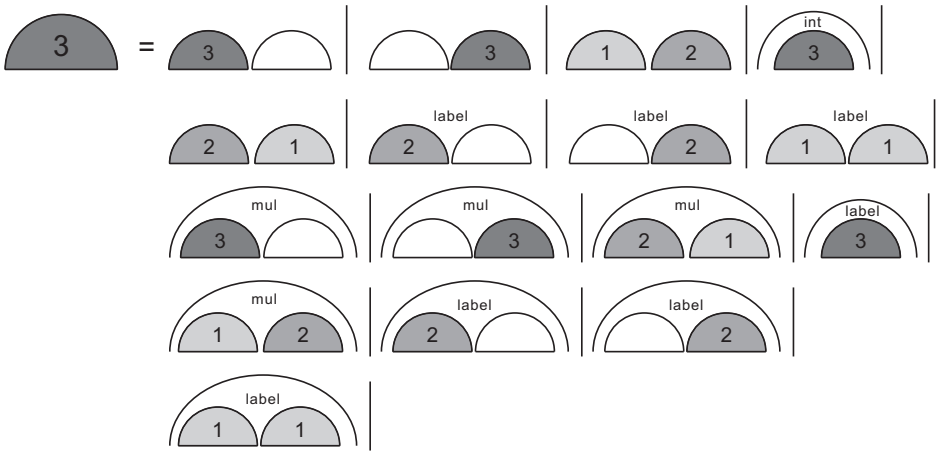

Figure 10: The recursion for $\theta_0^{(3)}$.

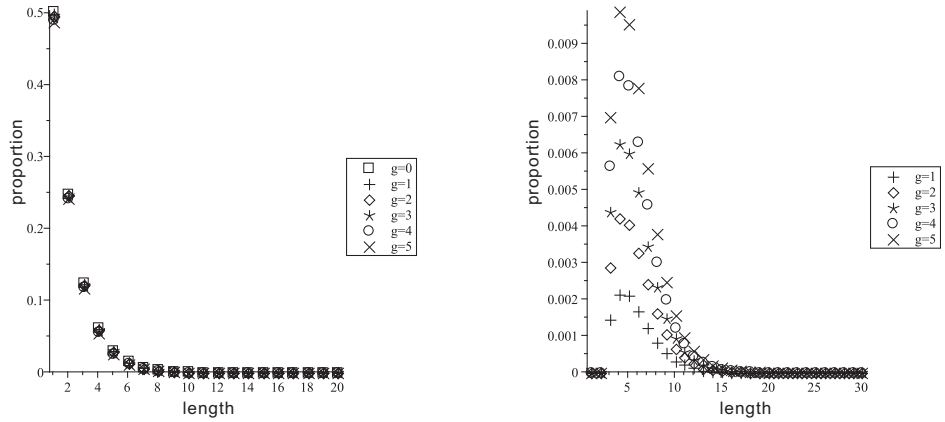

Figure 11: Loops in uniformly generated, genus filtered, RNA structures. We run the sampler 10000 times over a sequence of length 500 with genus ranging from 0 to 5. Left: distribution of standard loops, where x-axis is the length of boundary component and y-axis is frequency. Right: distribution of pseudoknot loops.