Bulletin of Mathematical Biology (2023) 85:21
https://doi.org/10.1007/s11538-023-01128-5

Society for Mathematical Biology

ORIGINAL ARTICLE

RNA Secondary Structures with Given Motif Specification: Combinatorics and Algorithms

Ricky X. F. Chen[1] · **Christian M. Reidys**[2,3] · **Michael S. Waterman**[2,4]

Received: 17 September 2021 / Accepted: 26 January 2023 / Published online: 13 February 2023

Abstract
The study of native motifs of RNA secondary structures helps us better understand the formation and eventually the functions of these molecules. Commonly known structural motifs include helices, hairpin loops, bulges, interior loops, exterior loops and multiloops. However, enumerative results and generating algorithms taking into account the joint distribution of these motifs are sparse. In this paper, we present progress on deriving such distributions employing a tree-bijection of RNA secondary structures obtained by Schmitt and Waterman and a novel rake decomposition of plane trees. The key feature of the latter is that the derived components encode motifs of the RNA secondary structures without pseudoknots associated with the plane trees very well. As an application, we present an algorithm (*RakeSamp*) generating uniformly random secondary structures without pseudoknots that satisfy fine motif specifications on the length and degree of various types of loops as well as helices.

Keywords RNA secondary structure · Plane tree · Rake decomposition · Helix · Loop · Uniform sampling

1 Introduction

Ribonucleic acid (RNA) plays an important role in various biological processes within cells, ranging from catalytic activity to gene expression. RNA can be described at different resolutions. The coarsest way of referring to an RNA molecule employs its

✉ Ricky X. F. Chen
chenshu731@sina.com

1 School of Mathematics, Hefei University of Technology, Hefei 230601, Anhui, People's Republic of China

2 Biocomplexity Institute and Initiative, Arlington, USA

3 Department of Mathematics, University of Virginia, Charlottesville, VA 22904, USA

4 Department Quantitative and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

primary structure that is its sequence of bases: A (adenine), U (uracil), G (guanine), and C (cytosine). A finer description specifies the chemical bonds between non-adjacent bases. In particular, the structure consisting of the bases and the bonds if consistent with a certain planar graph is known as a secondary structure. The finest 3D structure of an RNA molecule is called its tertiary structure. It is well-known that the functions of RNAs are deeply related to their structures rather than their primary sequences. However, there are major difficulties in determining RNA tertiary structures due to their high complexity and experimental costs. Instead, we frequently perform computational prediction of RNA secondary structures which requires a good understanding of the space of secondary structures.

In this work, we only focus on secondary structures without pseudoknots, although pseudoknots playing important roles have been found (e.g., Staple and Butcher 2005; Tuerk et al. 1992). More than four decades ago, Waterman and his coworkers pioneered the combinatorics of these RNA secondary structures (Smith and Waterman 1978; Stein and Waterman 1979; Waterman 1978, 1979). Since then, the combinatorics of RNA secondary structures taking into account various features has been one of the most important topics in computational biology, see Chen (2019), Clote (2006), Clote et al. (2012), Došlić et al. (2004), Han and Reidys (2012), Heitsch and Poznanović (2014), Fontana et al. (2004), Hofacker et al. (1998), Lorenz et al. (2008), Nebel (2003) and references therein. In particular, based on a bijection, Schmitt and Waterman (1994) obtained the exact formula for the number of RNA secondary structures with b base pairs and k isolated bases which is given by the Narayana number

$$\frac{1}{b+k}\binom{b+k}{k}\binom{b+k}{k-1}.$$

Among native motifs, loops and helices are the most studied, as they are closely related to prediction of RNA secondary structures based on thermodynamic models (Hofacker 2003; Mathews et al. 1999; Sloma and Mathews 2016; Zuker 1989; Zuker and Sankoff 1984; Zuker 2003). Loops are further distinguished into hairpin loops, bulges, interior loops, multiloops and exterior loops. Asymptotic enumerative results on a single type of motif such as helices and loops have been obtained (e.g., Fontana et al. 2004; Hofacker et al. 1998; Poznanović and Heitsch 2014) by exploring the recursions and analytical analyses on the associated generating functions. However, this approach cannot handle joint distributions of multiple motifs easily because simple recursion may not exist and analytical analyses of multi-variable functions are difficult.

In addition to enumerative studies, algorithms for generating or sampling random structures are also desired, as sometimes the obtained structures can be accepted as the predicted structures. See some samplers in Ding and Lawrence (2013), Nebel et al. (2011) and Ponty (2008) as well as a general framework on Boltzmann sampler in Duchon et al. (2004). To the best of our knowledge, however, there exists no algorithm generating a random RNA secondary structure with a given joint distribution of distinct loops and helices. The contribution that is most closely related to this paper is the work of Nebel et al. (2011), where a stochastic context-free grammar taking account of the probabilities of emergence of distinct loops was developed.

The paper is organized as follows. In Sect. 2, we propose a new decomposition algorithm for labelled plane trees that is called rake decomposition. We then review all native motifs interested in detail and apply the rake decomposition as well as the Schmitt–Waterman bijection in order for obtaining the most detailed and exact results in the literature on enumerating RNA secondary structures in Sect. 3. To be specific, we obtain formulas for the number of RNA secondary structures having prescribed numbers of helices, hairpin loops, bulges, interior loops, exterior loops and multiloops. In Sect. 4, we present an algorithm (RakeSamp) that produces an RNA secondary structure without pseudoknots satisfying a given joint distribution of all concerned motifs uniformly at random. The underlying idea and pseudocode of the algorithm are discussed in detail. We also illustrate RakeSamp using the motif parameters of the secondary structure of human mitochondrial tRNA-Ser1 (Jühling et al. 2009). Finally, we conclude the paper with some discussion and future directions in Sect. 5.

2 Rake Decompositions

A *plane tree* is an unlabelled tree with a distinguished vertex which is referred to as the *root*, and the subtrees incident to the root is also plane trees and linearly ordered. In a plane tree T, the number of edges in the unique path from a vertex v to the root of T is called the *level* of v. The vertices adjacent to v but further away from the root are called the *children* of v. A vertex other than the root having no children is called a *leaf*, and otherwise called an *internal vertex*. The root is always treated as internal. The *outdegree* of a vertex is the number of children of the vertex, i.e., one less than the degree of the vertex for any non-root vertex. We will draw a plane tree with its root on the top level, i.e., level 0, and with the children of a level i vertex arranged on level $i + 1$, left-to-right, following their linear order.

Definition 2.1 A *rake* is a plane tree where all leaves are children of the same vertex.

See Fig. 1 (left) for an example of a rake. A plane tree composed of a single vertex is also treated as a rake. In a rake T, the induced graph by the set of vertices that are not leaves is called the *stem* of T and the number of vertices contained in the stem is called the length of the stem.

Definition 2.2 An internal vertex of a plane tree that is not the only child of its parent is called a *companion*.

The root of a plane tree is always considered to be a companion. A companion in a plane tree is said to determine a rake if the vertex and its descendants form a rake. It is worth noting that a companion may not determine a rake and a non-companion vertex may determine a rake.

Lemma 2.1 *In any plane tree T, there exists at least one companion v which determines a rake. Suppose T has m companions. Then, deleting all descendants of v and their incident edges from T produces a plane tree T' having $m - 1$ companions.*

Lemma 2.1 is illustrated in Fig. 1 (right). It is beneficial to view deleting all descendants of v as removing the rake determined by v from T, where v splits into two vertices,

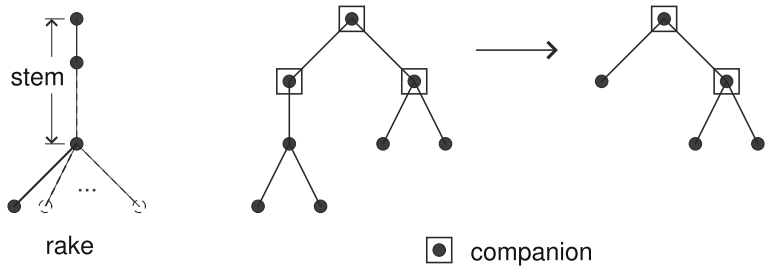

Fig. 1 A rake and an illustration of Lemma 2.1

one goes to be the root of the removed rake and the other stays in T. Lemma 2.1 suggests an iterative procedure to decompose a plane tree into rakes. However, in order to better control the process, we need to consider labelled plane trees. A *labelled plane tree* is a plane tree where each vertex has a unique label from some set of labels. In a labelled tree or forest, we may identify a vertex and its label. Now, we are in position to present the rake decompositions of labelled plane trees.

Rake decomposition procedure. Let T be a plane tree of $n+1$ vertices with the vertex set $[n+1] = \{1, 2, \ldots, n+1\}$ and $k > 0$ companion vertices. We decompose T into a set $\mathcal{F}(T)$ of k labelled rakes using the following procedure:

(i) Let l be an integer initializing $l = 1$.
(ii) Find the minimum companion vertex i in T that together with its descendants form a rake, and delete the rake rooted on i from T, and then put a vertex with label $(n+1+l)^*$ at the original position of i, and finally update T to the resulting plane tree and increase l by one.
(iii) Iterate (ii) until T is a labelled rake.

In view of Lemma 2.1, it is clear that the above procedure works. In fact, we have

Theorem 2.2 *The vertex labels of $\mathcal{F}(T)$ constitute the set $[n+1] \bigcup \{(n+2)^*, \ldots, (n+k)^*\}$. Furthermore, let v be the root of T. Then,*

- *the marked labels (with *) only appear as leaves in the rakes in $\mathcal{F}(T)$,*
- *the largest marked label, if any, is contained in the rake having v as its root.*

In addition, the sets of labelled rakes subject to these conditions are in one-to-one correspondence with the labelled plane trees.

Reverse of the rake decomposition. Let $\mathcal{F}$ be a set of labelled rakes with the vertex set $[n+1] \bigcup \{(n+2)^*, \ldots, (n+k)^*\}$, and with the marked labels (with *) only appear as leaves in the rakes. Then, a labelled plane tree on $[n+1]$ results from the procedure:

(a) In $\mathcal{F}$, among the trees having no marked labels, find the one with the minimum root.
(b) Replace the remaining minimum marked vertex with the found tree such that the root of the found tree is adjacent to the parent of the marked vertex, and update $\mathcal{F}$ as the resulting set of trees.
(c) Iterate (a)–(b) until $\mathcal{F}$ contains a single tree.

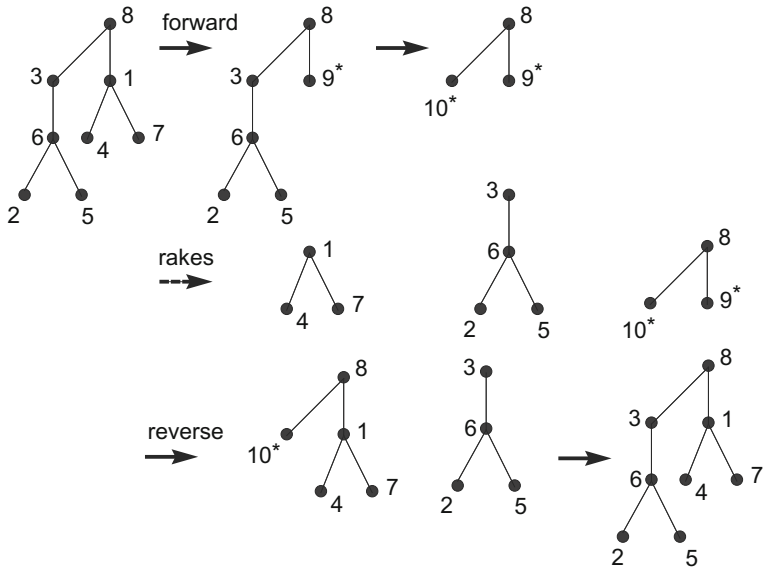

Fig. 2 A labelled plane tree and its rake decomposition

We visualize the rake decomposition and its reverse step by step in Fig. 2. We remark that the mark (*) is not essential and is used only for convenience and the bijection in Theorem 2.2 is an analog of the bijection presented in Chen (1990). But, the key here is to realize in the first place that rakes are the right pieces in order for studying RNA native motifs which will be explained later.

3 Enumeration of RNA Secondary Structures

3.1 Preliminaries

We first recall the definition of RNA secondary structures (without pseudoknots) following Waterman (1978).

Definition 3.1 (RNA secondary structure) An *RNA secondary structure* of length n is a simple graph with vertices in $[n]$ and edges in E that satisfies:

- if $(i, j) \in E$, then $|i - j| \geq 2$;
- if $(i, j) \in E$ and $(k, l) \in E$, where $i < j$ and $k < l$, and $[i, j] \bigcap [k, l] \neq \varnothing$, then either $[i, j] \subset [k, l]$ or $[k, l] \subset [i, j]$ (where $[i, j]$ denotes the interval $\{r : i \leq r \leq j\}$).

The vertices represent the bases, while the edges represent the *base pairs*. One way for representing an RNA secondary structure is to draw a diagram placing all vertices in a horizontal line and drawing its edges as arcs in the upper half-plane. See the upper picture in Fig. 3 as an example. By construction, any two arcs do not cross. (The

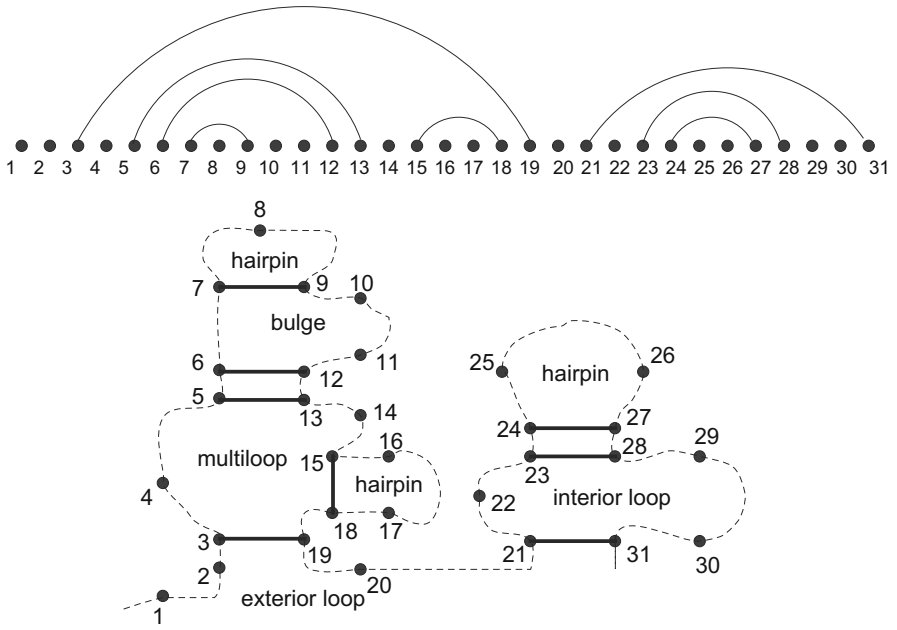

Fig. 3 An RNA secondary structure represented in two ways

structures with crossing arcs are those with pseudoknots.) A vertex not incident to any edge is called an *isolated base*. We say an arc (i_1, j_1) (resp. an isolated base k) is covered by an arc (i, j) if $[i_1, j_1] \subset [i, j]$ (resp. $k \in [i, j]$). In this case, we also say that the arcs (i, j) and (i_1, j_1) nest with each other. Moreover, once the diagram is ready, the vertex (base) labels in $[n]$ are implicitly determined and can sometimes be omitted.

RNA secondary structures of length n can be viewed as a special class of noncrossing partitions on $[n]$, for which we refer to Klazar (1998), Simion (2000) and references therein. Although some statistics, e.g., number of blocks, block sizes, are studied in the context of noncrossing partitions, we are interested in another family of motifs for RNA secondary structures.

Definition 3.2 (Helix) A *helix* in an RNA secondary structure is a maximal set of arcs that mutually nest with each other and whose left-ends and right-ends are, respectively, consecutive. The size of this arc set is called the length of the helix.

For the RNA in Fig. 3, the arcs (5, 13) and (6, 12) give a helix. A helix may consist of only one arc. For example, in Fig. 3, (3, 19) gives a helix. Loops in RNA secondary structures have been extensively studied, as it is important for certain energy models predicting the folded secondary structure of a given RNA (primary) base sequence. We follow the definition in Hofacker et al. (1998).

Definition 3.3 (Loop) A *loop* consists of a set of isolated bases that are either directly covered by the same arc (i, j) or not covered by any arc. The *length* of the loop is

the size of the set of isolated bases, and we refer to (i, j) as *the arc of the loop*. The *degree* of a loop is one larger than the number of arcs directly covered by the arc of the loop.

By that an isolated base or arc is directly covered by an arc (i, j), we mean no other arc that covers the isolated base or arc is covered by (i, j). For the RNA in Fig. 3, {4, 14} gives a loop since the isolated bases 4 and 14 are both directly covered by the arc (3, 19). It is a degree three loop, since the arc of the loop, (3, 19), directly covers two arcs (5, 13) and (15, 18). Loops are classified into different types: hairpin loops, interior loops, exterior loop, bulges, and multiloops.

Definition 3.4 (Hairpin loop) A *hairpin loop* (or simply hairpin) is a loop whose arc does not cover any arcs.

Definition 3.5 (Interior loop) An *interior loop* is a loop where there exists exactly one arc directly covered by the arc of the loop and separating at least two isolated bases (i.e., one is to the left and one is to the right).

An interior loop has a length at least two and degree exactly two. In Fig. 3, {25, 26} and {22, 29, 30}, respectively, provide a hairpin and an interior loop.

Definition 3.6 (Bulge) A *bulge* is a degree two loop such that either the left-ends or the right-ends (but not both) of the arc of the loop and the arc directly covered by it are consecutive.

Definition 3.7 (Multiloop) Loops of degree larger than two are called *multiloops*.

In Fig. 3, {10, 11} gives a bulge, while {4, 14} yields a multiloop. Generally, a loop cannot be empty. However, we allow for an empty multiloop that is determined by an arc directly covering more than two arcs but no isolated bases.

Definition 3.8 (Exterior loop) The isolated bases that are not covered by any arcs form the unique *exterior loop* of an RNA secondary structure.

The exterior loop need not have isolated bases, i.e., it could be empty. The degree of the exterior loop is one larger than the number of arcs that are not covered by any arcs, and this number is also referred to as the number of *components*. Sometimes an RNA secondary structure is depicted as the lower picture in Fig. 3 from which the terms like hairpin may be more intuitive.

Note that the arc of a loop other than the exterior loop is contained in a unique helix. Thus, each helix is associated with a unique non-exterior loop.

Enumeration of secondary structures is usually facilitated by translating them into plane trees via certain bijections. Such bijections can be found in Schmitt and Waterman (1994) as well as Chen (2019). Here, we make use of the former which works as follows. For a given RNA secondary structure R of length n represented as a diagram that has b base pairs and k isolated bases (so $n = 2b + k$), add an auxiliary arc $(0, n + 1)$ at first. Next, generate a vertex for each arc and each isolated base in the resulting diagram. Equivalently, this may be understood as replacing each arc and each

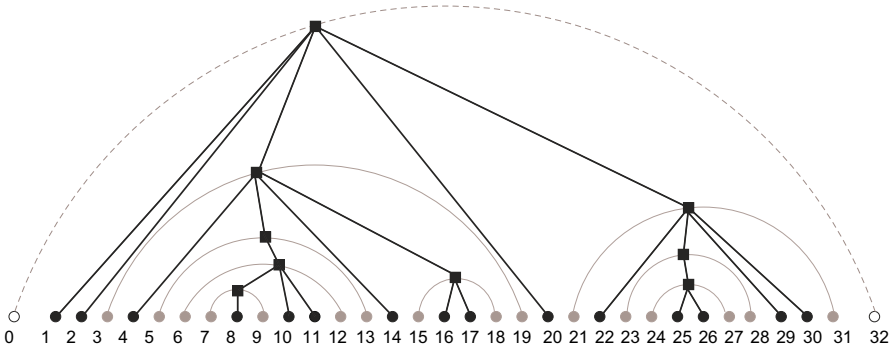

Fig. 4 The Schmitt–Waterman bijection: the RNA secondary structure (see Fig. 3) in gray and its corresponding plane tree in dark. Note that the plane tree has no vertex or edge labels (Color figure online)

isolated base, respectively, with a vertex. Thus, we have $n+1$ vertices now. Next, we connect these vertices such that the resulting structure is a plane tree with the vertex (corresponding to the arc) $(0, n+1)$ as its root, which can be done recursively: We start with the vertex $(0, n+1)$ and connect it by an edge with any vertex corresponding to either an arc or an isolated base that is directly covered by $(0, n+1)$ in R. In this manner, the children of the vertex $(0, n+1)$ are completely determined. Next, we deal with these children one by one. If the current child v corresponds to an isolated base in R, then as a vertex in the desired plane tree, v is a leaf having no children; If v corresponds to an arc, then its children will be found and connected to it in a similar way as we dealt with $(0, n+1)$. We iterate this procedure until the children of all $n+1$ vertices (including those with 0 children) have been found and connected, and we obtain a plane tree with $b+1$ internal vertices and k leaves.

Conversely, given a plane tree T with $b+1$ internal vertices and k leaves, we iteratively construct the diagram of an RNA secondary structure as follows. Draw an arc e and leave the two ends of the arc unlabelled for the moment. Inspect the children of the root of T from left to right: If an inspected child v is a leaf in T, then put a vertex under the arc e; If v is an internal vertex, then put an arc (with ends unlabelled) under e. These vertices and arcs will be the only elements that are directly covered by e in the desired diagram, and their left-to-right order under e follows the left-to-right order of the corresponding children. Next, for each newly generated arc, we check the children of its corresponding internal vertex in T and generate the elements directly covered by it analogously. Iterate the above procedure so that each internal vertex in T uniquely corresponds to an arc and each leaf uniquely corresponds to an isolated vertex in the constructed diagram. Finally, we delete the arc e and label the vertices in the remaining diagram from left to right with labels $1, 2, \ldots, 2b+k$. See Fig. 4 for an illustration.

3.2 Exact Formulas

In the following, we shall provide a novel enumeration of RNA secondary structures, employing the Schmitt–Waterman bijection and the rake decompositions of labelled

plane trees. Although in the Schmitt–Waterman bijection RNA secondary structures are in one-to-one correspondence with (unlabelled) plane trees instead of labelled plane trees, this can be easily "remedied" as follows. From each plane tree of $b+k$ edges, we can obtain $(b+k)!$ different labelled plane trees on $[b+k+1]$ that have $b+k+1$ as the root. Equivalently, an RNA secondary structure of b base pairs and k isolated bases is associated with $(b+k)!$ different labelled plane trees. Then, we can translate the enumeration problem of RNA secondary structures into enumerating labelled plane trees which is eventually tantamount to enumerating the corresponding sets of rakes. What matters is whether the rakes encode any meaningful structural information of the RNA secondary structures.

To that end, we first provide several lemmas by which we analyze the aforementioned two bijections. By abuse of notation, we shall use the same notation for a vertex in a plane tree and its corresponding arc or isolated base in the corresponding secondary structure.

Lemma 3.1 *Given an RNA secondary structure R with b base pairs and k isolated bases, suppose u is the outermost arc of a helix. Then, by the Schmitt–Waterman bijection,*

- *if u is not $(1, 2b+k)$, then u is mapped to a companion vertex that is not the root, and the helix as well as the elements directly covered by the innermost arc of the helix induces a rake (i.e., as the induced graph) rooted at u;*
- *if u is $(1, 2b+k)$, then the helix together with the auxiliary arc $(0, 2b+k+1)$ and the elements directly covered by the innermost arc of the helix induces a rake rooted at the root of the plane tree.*

Suppose R has h helices. Then, the corresponding plane tree has h companion vertices if $(1, 2b+k)$ is a base pair, and $h+1$ companion vertices otherwise.

Lemma 3.2 *Let T be a labelled plane tree with a distinguished companion v, and let u_v be the closest v-descendant having more than one child. Processing T via the rake-decomposition, v induces a labelled rake with root v, whose stem equals the induced subgraph of the vertices on the path from v to u_v. The leaves of this rake are obtained from the left-to-right children of u_v, replacing any companion vertices by certain marked vertices.*

Proof In the course of the rake decomposition, the children of u_v that are companions will determine some rakes that will be removed at some point. After their removal, v determines by construction the rake described in the lemma, completing the proof. □

Combining Lemma 3.1 and 3.2, it can be seen that a helix and the associated loop give rise to a unique labelled rake (through a labelled plane tree). Moreover, the attribute of the loop determined by a labelled rake depends on the number and distribution of unmarked and marked leaves in the rake as summarized in the following lemma.

Lemma 3.3 *An RNA secondary structure with b base pairs and k isolated bases is associated with $(b+k)!$ sets of labelled rakes with the vertex set $[b+k+1] \bigcup \{(b+k+2)^*, \ldots, (b+k+l)^*\}$ for some appropriate l where*

- *a helix of length q that does not contain the arc $(1, 2b+k)$ uniquely corresponds to the stem of length q of a rake, and a helix of length q that contains the arc $(1, 2b+k)$ uniquely corresponds to the stem of length $q+1$ of a rake,*
- *a hairpin loop of length q uniquely corresponds to a rake whose leaves consist of q unmarked labels (without *),*
- *a bulge loop of length q uniquely corresponds to a rake whose leaves consist of a marked label (with *) and q unmarked labels all either to the left or to the right of the marked label,*
- *an interior loop of length q uniquely corresponds to a rake whose leaves consist of a marked label and q unmarked labels surrounding the marked label,*
- *a multiloop of length q and degree d uniquely corresponds to a rake whose leaves consist of q unmarked labels and $d-1$ marked labels,*
- *if $(1, 2b+k)$ is not a pair, then the exterior loop of length e_l and degree e_d corresponds to a rake with root $b+k+1$ where the stem has length one and there are $e_d - 1$ marked including the maximum marked label and e_l unmarked labels; otherwise, no rake is associated with the exterior loop.*

The number of marked labels (i.e., l) is determined by the number of helices according to Lemma 3.1 and the rake decomposition procedure. Note that there may be no marked label at all. It is worth noting that the base labels of an RNA secondary structure are not related to the vertex labels of the associated labelled rakes. In Fig. 5, we depict the relation between helices as well as loops of secondary structures and rakes.

Theorem 3.4 *The number of RNA secondary structures with b base pairs, k isolated bases, an exterior loop of length e_l and degree e_d, and*

h_i	helices of length $i > 0$
p_i	hairpin loops of length $i > 0$
g_i	bulges of length $i > 0$
t_i	interior loops of length $i > 1$
m_{ij}	multiloops of length $i \geq 0$ and degree $j \geq 3$

is given by

$$\frac{(h-1)!h!(e_l+e_d-1)2^g\prod_{i>1}(i-1)^{t_i}}{\prod_{i>0}p_i!g_i!t_i!h_i!\prod_{i\geq 0,j\geq 3}m_{ij}!}\binom{e_l+e_d-2}{e_l}\prod_{i\geq 0,j\geq 3}\binom{i+j-1}{j-1}^{m_{ij}},$$

where $h=\sum_{i>0}h_i$ and $g=\sum_{i>0}g_i$.

We remark that the parameters in Theorem 3.4 are not independent and satisfy

$$\sum_i ih_i = b, \quad k = e_l + \sum_i i(p_i+g_i+t_i) + \sum_{i,j} im_{ij},$$

$$\sum_i h_i = \sum_i p_i + g_i + t_i + \sum_{i,j} m_{ij} = \sum_i g_i + t_i + \sum_{i,j}(j-1)m_{ij} + e_d - 1.$$

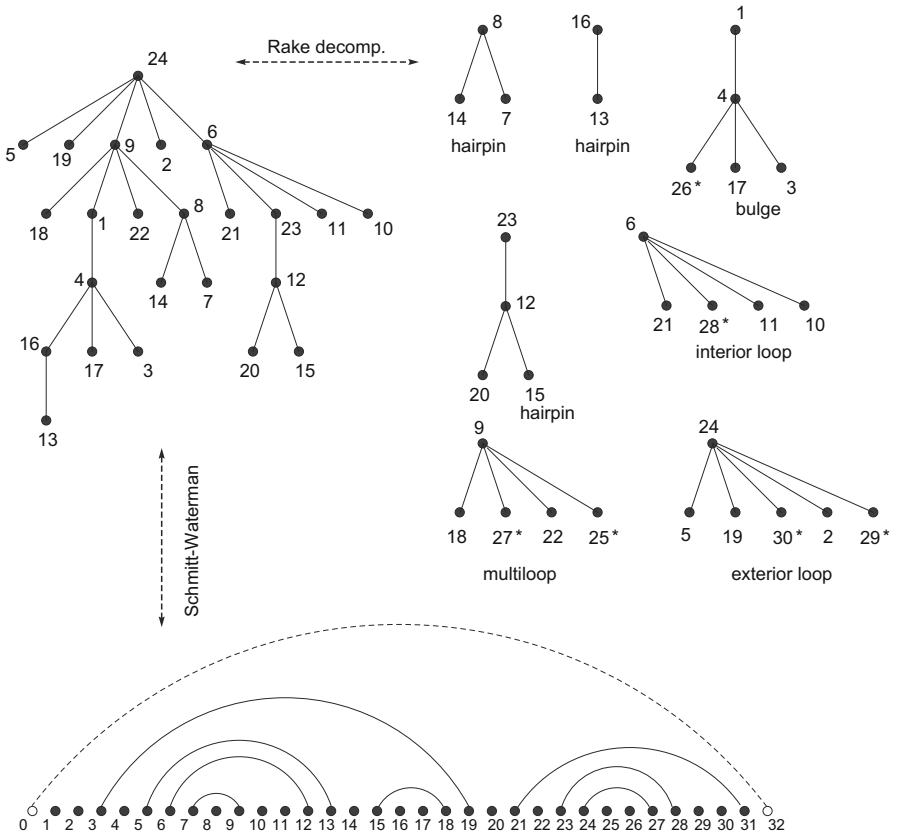

Fig. 5 From motifs of secondary structure to rakes. Note: only one labelling of the plane tree corresponding to the secondary structure under the Schmitt–Waterman bijection is illustrated and in particular the labelling has no relation to the base labels

Biological secondary structures exhibit specific constraints on the minimum length σ (typically $\sigma \geq 2$) of a helix and the minimum distance θ (typically $\theta \geq 3$) of the two bases of a base pair, i.e., $|j - i| > \theta$ if bases i and j form a base pair. Note that the latter is equivalent to requiring the minimum length of a hairpin loop to be θ as every base pair is either an arc of a hairpin loop or covering a hairpin loop. The following takes these constraints into account.

Theorem 3.5 *Let σ and θ be any positive integers. Then, the number of RNA secondary structures with b base pairs, k isolated bases, h helices with each having a length at least σ, p hairpin loops with each having a length at least θ, g bulges, t interior loops, m_j multiloops of degree $j \geq 3$, and an exterior loop of length e_l and degree e_d, is given by*

$$\frac{(e_l+e_d-1)(h-1)!2^g}{p!g!t!\prod_j m_j!}\binom{h+(b-h\sigma)-1}{h-1}\binom{e_l+e_d-2}{e_l}$$
$$\binom{\sum_j jm_j+(k-e_l)-1+p-p\theta}{k-e_l-p\theta-g-2t}.$$

The proofs of Theorem 3.4 and 3.5 are similar, and both can be found in Appendix.

4 Uniform Sampling

4.1 Strategy

We next derive an algorithm for uniformly sampling a subset of RNA secondary structures (without pseudoknots) satisfying given motif parameters. Our idea is to first uniformly generate a random set of labelled rakes, then transform them into a plane tree using our rake decomposition correspondence and finally, transform the plane tree into an RNA secondary structure via the Schmitt–Waterman bijection.

Let us consider the subset of secondary structures enumerated in Theorem 3.4 in the case of $e_d > 2$, or $e_l > 0$ and $e_d = 2$. All other scenarios can be treated analogously. Indeed, once the structural parameters are given, the sizes of the leaves and stems of the corresponding rakes are determined, respectively. It remains to assign labels to the vertices of the rakes employing Lemma 3.3. We note that for the case here there are in total $h+1$ rakes and the vertex labels for the rakes are from the set $X \bigcup\{(b+k+1+h)^*\}$ of marked labels (with *) as well as the set $Y \bigcup\{b+k+1\}$ of unmarked labels, where

$$X = \{(b+k+2)^*, \ldots, (b+k+h)^*\},$$
$$Y = \{1, \ldots, b+k\}.$$

As long as we assign the labels uniformly at random, we can produce a uniform random secondary structure exhibiting a specified set of features. In fact, this is easier than deriving the exact formula as in Theorem 3.4, as we can neglect certain multiplicity issues that we needed to address in our computations due to the following.

Lemma 4.1 *Suppose A is a set of objects and B (a multiset) is a disjoint set-union of several copies of A. Then, uniformly sampling the objects in A is equivalent to uniformly sampling the objects in B.*

There are two places where we introduce multiplicity: one is from unlabelled plane trees (which are in one-to-one correspondence with RNA secondary structures) to labelled plane trees, the other is from sets of rakes to sequences of rakes as we shall see shortly.

Note that each labelled rake can be represented by a pair of sequences (S, L) where S encodes the top-down vertices (labels) of its stem while L encodes its left-to-right leaves. Accordingly, our objective becomes uniformly sampling the space A of **sets** of $h+1$ pairs (S, L) with the unmarked and marked vertex labels from $Y \bigcup X \bigcup\{(b+k+1+h)^*, b+k+1\}$ subject to Lemma 3.3. Thanks to Lemma 4.1,

this is achieved in *RakeSamp* by uniformly sampling the space B of **sequences** (i.e., the second-place-multiplicity) of $h+1$ pairs $(S[1], L[1]), \ldots, (S[h+1], L[h+1])$ with the unmarked and marked vertex labels subject to Lemma 3.3, and additionally with

- the first p pairs corresponding to hairpins,
- the next g pairs corresponding to bulges,
- the next t pairs corresponding to interior loops,
- the following $\sum_{i,j} m_{ij}$ pairs corresponding to multiloops,
- the last pair corresponding to the exterior loop with $S[h+1] = b+k+1$ (i.e., length one) and $(b+k+1+h)^*$ being an entry in $L[h+1]$, and
- pairs corresponding to the same type of loops are arranged according to the number of unmarked labels (and degrees for multiloops) in L increasingly.

For such a random sequence of pairs (i.e., a random sample), in the concatenation $L[1]L[2]\cdots L[h+1]$, the unmarked labels form a random sequence of length k with the entries from the set $[b+k]$, and the marked labels form a random sequence of length h from the set $X \bigcup \{(b+k+1+h)^*\}$ with $(b+k+1+h)^*$ being an entry in $L[h+1]$. On the other hand, the concatenation $S[1]S[2]\cdots S[h]$ gives a random permutation of the remaining unmarked labels other than $b+k+1$ (which is contained in $S[h+1]$).

Once the motif parameters are given, the loop type and its length (the number of unmarked labels) as well as degree (one greater than the number of marked labels) associated with $L[i]$ are completely determined. The only thing that is free is the locational relation among unmarked and marked labels, and the only restriction in this regard is clearly specified in Lemma 3.3 (e.g., the unique marked label is either to the left or right of all unmarked labels for bulges.) As for the length distribution of $S[i]$ $(1 \le i \le h)$, the only requirement is that it agrees with the multiset $\{1^{h_1}, 2^{h_2}, \ldots\}$ (i.e., h_1 helices of length one, h_2 helices of length two, etc.) We denote the number of unmarked labels contained in $L[i]$ by $y(L[i])$ and the number of marked labels by $x(L[i])$. As such, for the implementation, we work backwards in two steps:

Step 1 Generate a random list of desired sequences $L[1], L[2], \ldots, L[h+1]$. This is done by first generating (via Rdm function) a random sequence Lev of length k using unmarked labels from Y and generating (via Rdm) a random sequence Mlv of the marked labels (where the largest one, $(b+k+1+h)^*$, will be separately handled as can be seen in the algorithm code later in detail). Then, for $1 \le i \le h+1$, get the first $y(L[i])$-element subsequence of the remaining sequence from Lev and get the first $x(L[i])$-element subsequence of the remaining sequence from Mlv. Finally, we interlace the two subsequences randomly respecting the loop type of $L[i]$ (that is, for instance, if it is an interior loop, the resulting sequence cannot have a marked label to be the leftmost or the rightmost entry) but keep the relative order of the labels from the same subsequence unchanged. The last task is done by various Mix functions. This produces $L[1], \ldots, L[h+1]$.

Step 2 Generate a random list of sequences $S[1], S[2], \ldots, S[h]$ using the unused unmarked labels other than $b+k+1$ from Step 1. This is accomplished by first generating a random permutation Z' of the unused unmarked labels other than $b+k+1$ from Step 1, and generating a random "permutation"

$Hsz = Hsz[1]Hsz[2]\cdots Hsz[h]$ of the multiset $\{1^{h_1}, 2^{h_2}, \ldots\}$. Finally, for $i = 1$ to h, assign the first $Hsz[i]$-element subsequence of the remaining sequence from Z' to $S[i]$. This gives us a uniform random sequence $S[1], S[2], \ldots, S[h]$.

4.2 Algorithm: RakeSamp

Here, we provide the pseudocode of *RakeSamp*. The functions are defined below:

- $Rdm(S, k)$: returns a uniform random permutation of a uniform random k-element subset of the set or multiset S. This can be done by generating a uniform random permutation of S and picking the subsequence consisting of the first k entries. The time complexity of doing this is at most $O(|S|)$, e.g., using the well-known Fisher–Yates shuffle algorithm (see Algorithm 3.4.2P in Knuth 1997). If $k = |S|$, we simply write $Rdm(S, \circ)$.
- $firstK(U, k)$: returns the subsequence consisting of the first k entries of U and updates U to the remaining sequence after deleting the first k entries of U.
- $Mix(U, V)$: returns the sequence resulting from interlacing the sequences U and V uniformly at random but keeping the relative order of the entries from U and, respectively, V unchanged. This is essentially generating a random $|U|$-element subset from the set $[\,|U| + |V|\,]$. For example, for $U = 142$ and $V = 536$, one possible outcome would be 154362. If V is empty, then the returned sequence is just U; The case U being empty is similar.
- $Mix_e(U, v)$: inserts the single element v before the first entry or after the last entry of U, each with probability $1/2$, and returns the resulting sequence.
- $Mix_m(U, v)$: inserts the single element v between two consecutive entries of U, each possible space with probability $\frac{1}{|U|-1}$, and outputs the outcome.

As for the complexity, all involved functions are elementary and have roughly linear-time implementation. Accordingly, *RakeSamp* has approximately $O(n)$ time complexity where n is the length of the RNA sequences in study. As for the space complexity, we merely need to store a finite number of arrays of labels, each of length at most $O(n)$.

Example 4.1 We evaluate *RakeSamp* using the following parameters:

exterior loops	degree $e_d = 2$, length $e_l = 4$
helices h_i	$h_2 = 2,\ h_5 = 1,\ h_7 = 1$
hairpins p_i	$p_8 = 1,\ p_9 = 1$
bulges g_i	no bulges, i.e., $g_i = 0$
interior loops t_i	$t_2 = 1$
multiloops m_{ij}	$m_{7,3} = 1$ length seven multiloop of degree three

These parameters are specified in the file "trnaser1.txt" on Github. Three secondary structures satisfying these motif parameters are shown in Fig. 6.

Algorithm 1 RakeSamp

Input: parameters specified in Theorem 3.4 where $e_d > 2$, or $e_l > 0$ and $e_d = 2$
// e_d *and* e_l *are resp. the degree and length of the exterior loop*
Output: a sequence of $h+1$ S-L pairs // h *is the number of helices*
$X \leftarrow \{(b+k+2)^*, (b+k+3)^*, \ldots, (b+k+h)^*\}$ // *marked labels other than*
// $(b+k+1+h)^*$, b *and* k *are resp. the number of base pairs and isolated bases*
$Y \leftarrow \{1, 2, \ldots, b+k\}$ // *unmarked labels excluding* $b+k+1$
$Lev \leftarrow Rdm(Y, k)$ // *the subsequence of the unmarked labels in* $L[1]\cdots L[h+1]$
$Mlv \leftarrow Rdm(X, \circ)$ // *the subsequence of the marked labels*
// *in* $L[1]\cdots L[h+1]$ *excluding* $(b+k+1+h)^*$
$Z \leftarrow$ labels in Y that not used in Lev
for $i = 1$ to h **do**
  $U \leftarrow firstK(Lev, y(L[i]))$ // $y(L[i])$ *is the number of unmarked labels in* $L[i]$
  $V \leftarrow firstK(Mlv, x(L[i]))$ // $x(L[i])$ *is the number of marked labels in* $L[i]$
  // *a preprocess of the input parameters can determine the loop type associated*
  // *to* $L[i]$ *as well as* $y(L[i])$ *and* $x(L[i])$
  if $L[i]$ corresponds to a bulge **then**
    $L[i] \leftarrow Mix_e(U, V)$ // V *is a sequence of length one in this case*
  else if $L[i]$ corresponds to an interior loop **then**
    $L[i] \leftarrow Mix_m(U, V)$ // V *is a sequence of length one in this case*
  else
    $L[i] \leftarrow Mix(U, V)$ // V *and* U *maybe empty for hairpins and multiloops resp.*
  end if
end for// *now the remaining labels in* Lev *and* Mlv *are all for* $L[h+1]$
$V \leftarrow Mix(Mlv, (b+k+h+1)^*)$ // *the largest marked label is always in* $L[h+1]$
$L[h+1] \leftarrow Mix(Lev, V)$ // *end of Step* 1
$Hsz \leftarrow Rdm(\{1^{h_1}, 2^{h_2}, \ldots\}, \circ)$ // *a random permutation of the lengths of helices*
$Z' \leftarrow Rdm(Z, \circ)$ // *a random permutation of the unused unmarked labels from* Y
for $i = 1$ to h **do**
  $S[i] \leftarrow firstK(Z'; Hsz[i])$
end for
$S[h+1] \leftarrow b+k+1$ // $S[h+1]$ *only contains the largest unmarked label*
return $(S[1], L[1]), \ldots, (S[h+1], L[h+1])$

In fact, the leftmost one is the secondary structure of *human mitochondrial tRNA-Ser1* (Jühling et al. 2009).

According to Theorem 3.4, there exist 12960 secondary structures with the parameters. We ran *RakeSamp* for 7980000 iterations which took 152672 seconds (with CPU: Intel Xeon Silver 4210R @2.4GHz; RAM: 64GB; Compiler: DEV C++ 6.3). The observed frequencies for all structures are depicted in Fig. 7 (a) and (b).

To provide further evidence for uniformity, we used $e_l = 2$ as well as $m_{2,3} = 1$ (and discarded $m_{7,3} = 1$) and kept others unchanged (just for reducing the size of the sample space to saving time) and ran 72 simulations where the expected occurrence of each structure was 100. The empirical distribution of the corresponding p-values for the chi-square goodness of fit statistic is plotted in Fig. 7 (c). Under uniform sampling, the p values will be uniformly distributed as the empirical distribution function in Fig. 7 (c) illustrates. All of these clearly suggest the uniformity of our algorithm. We refer to the instruction on Github for using *RakeSamp*.

Often when sampling for specific properties within a large space of objects, importance sampling such as Markov Chain Monte Carlo is employed. This is expensive computationally and convergence properties are seldom well understood. Our algo-

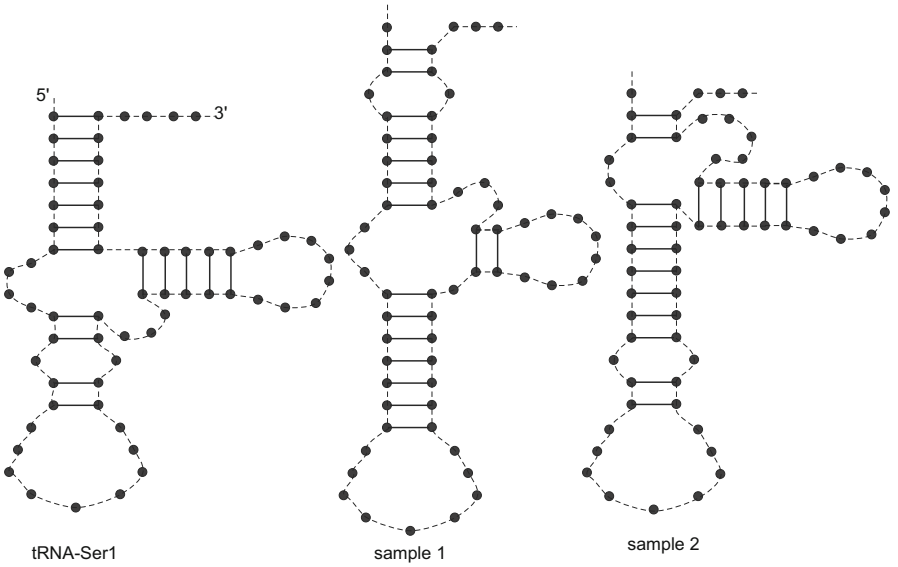

Fig. 6 Secondary structure for tRNA-Ser1 (left) and two structures from simulation

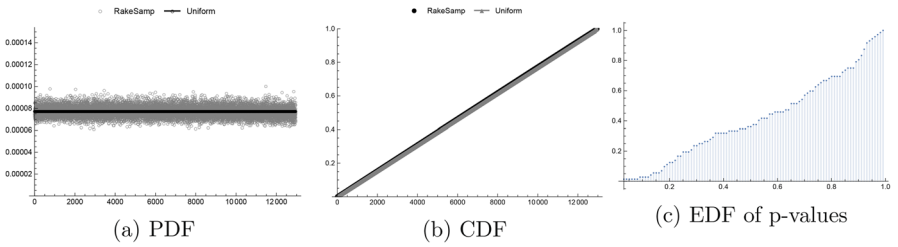

Fig. 7 Test RakeSamp: (a) plot of frequencies of all structures; (b) cumulated frequencies of all structures; (c) empirical distribution function of p-values

rithm which depends on a careful analysis of the combinatorial space is fast and exact and illustrates the value of such an understanding.

4.3 Software

The sampling algorithm for all cases (of e_d and e_l) is implemented in *RakeSamp*. In addition, the obtained sequence of S-L pairs has been transformed into RNA secondary structure in the form of bracket-dot representation as well. The program is written in C and available on Github https://github.com/RickyXFChen/RakeSamp.

5 Conclusion

In this paper, we presented a new bijection between plane trees and certain forests of rakes. The key was the notion of a companion, a vertex which is not the only child of its parent, which during the recursive rake decomposition was shown to become the root of the rakes.

We then employed the Schmitt–Waterman bijection between RNA secondary structures and plane trees in order to interpret the rakes which in turn produces exact enumerative results of RNA secondary structures satisfying specific, biologically relevant structural motifs. These results allow one to uniformly sample RNA secondary structures which have specific, structural motifs including helices, hairpins, bulges, interior loops, exterior loops and multiloops. We then implemented such a uniform sampling algorithm, via *RakeSamp*.

As a direction regarding application, we can first learn the distribution (or ratio) of structural features in known RNA secondary structures, e.g., through machine learning as in Sato et al. (2021) and references therein and then, generate a random structure respecting the distribution using our algorithms as a prediction. This could provide an alternative approach to energy-based prediction of secondary structures. This sort of thing has been done in bioinformatics, from generating synthetic RNA structures both using energy models and stochastic grammars, to reconstructing phylogenetic trees, to detecting CpG islands in genomes, only to name a few applications. Our work opens the door to doing this with RNA structures in a different manner. Moreover, structures with the same motif parameters have approximately the same free energy in the common energy models. Thus, we can carry out comparative study of a natural secondary structure and a random structure satisfying the same set of motif parameters as the natural one (via *RakeSamp*) in order for better understanding why nature selects one over others even though they all have similar free energy. As another potential application, *RakeSamp* may facilitate calibrating or extracting finer energy parameters for distinct motifs in the context of RNA folding as the current energy parameters do not distinguish multiloops of different lengths and degrees. All of these deserve further investigations and some of them are in progress.

Acknowledgements We are grateful for the valuable comments and suggestions of the anonymous referees which improved the presentation of the work. The first author also acknowledges the support by the Yellow Mountain Distinguished Scholar Program at the Hefei University of Technology.

Data Availability The software *RakeSamp* and its source code are freely available at Github: https://github.com/RickyXFChen/RakeSamp.

Declarations

Conflicts of interest: none.

Funding The first author was supported by the Anhui Provincial Natural Science Foundation of China (No. 2208085MA02) and Overseas Returnee Support Project on Innovation and Entrepreneurship of Anhui Province (No. 11190-46252022001).

Availability of data and material: not applicable.

Code availability: RakeSamp is available on Github https://github.com/RickyXFChen/RakeSamp.

Appendix Proofs

Proof of Lemma 2.1 Note that in a finite plane tree, there always exists an internal vertex u whose children are all leaves. Starting with u and traveling along the unique path from u to the root (which could be u itself), the first encountered vertex that has siblings, v, is a companion that determines a rake. In case no such v exists, the root represents the companion that determines a rake.

It is clear that deleting all descendants of v and their incident edges will not affect any other remaining vertex's attribute of being a companion or not. It may affect whether or not such a companion determines a rake after the deletion. Furthermore, v itself is a leaf in T', and by definition, none of the v-descendants are companions in T, whence we have $m - 1$ companions in T'.

Proof of Lemma 3.1 Taking the auxiliary arc $(0, 2b+k+1)$ into consideration, there is at least one arc covering u. Then, by construction of the Schmitt–Waterman bijection, u is a child of the innermost arc v covering u. Since u is the outermost arc of a helix and is not $(1, 2b+k)$, v must directly cover some other isolated bases or arcs whence u is not the only child of v. Moreover, only isolated bases are mapped to leaves, thus u is mapped to an internal vertex. Therefore, u corresponds to a companion. If $(1, 2b + k)$ is a base pair, it will be clearly mapped to the unique child of the vertex $(0, 2b+k+1)$. The remaining statements follow directly from the Schmitt–Waterman bijection.

If $(1, 2b+k)$ is a base pair, then it is the outermost arc of a helix. As a consequence of the above discussion, the other $h - 1$ helices determine $h - 1$ companions that are not the root. Taking account of the root, there are h companion in total. Analogously, if $(1, 2b + k)$ is not a pair, there are $h + 1$ companions, and the proof follows.

Proof of Lemma 3.3 For a bulge loop of length q, by definition, the arc v of the bulge directly covers q isolated bases and an arc which is either to the left or right of all q isolated bases. Note that the corresponding vertices covered by v are leaves in the corresponding rake and their left-to-right order stays the same as in the secondary structure.

We next consider the exterior loop. By definition, the number of arcs directly covered by $(0, 2b + k + 1)$ is $e_d - 1$ and the number of directly covered isolated bases is exactly e_l. If $(1, 2b + k)$ is not a pair, then, according to Lemma 3.1, the arcs directly covered by $(0, 2b + k + 1)$ are mapped to companions that will give rise to marked leaves in the corresponding rake rooted on $b + k + 1$ in view of Lemma 3.2. The isolated bases in the exterior loop are mapped to the unmarked leaves of the rake. If $(1, 2b + k)$ is a base pair, then $e_d = 2$ and $e_l = 0$. In this case, the helix having $(1, 2b+k)$ as the outermost arc and the associated loop together with $(0, 2b+k+1)$ give rise to the rake rooted on $b + k + 1$, and no rake is associated with the exterior loop. The remaining statements follow analogously.

Proof of Theorem 3.4 Note that the length of the RNA secondary structures under consideration is $2b + k$. Let

$$p = \sum_i p_i, \quad P = \sum_i i p_i,$$
$$g = \sum_i g_i, \quad G = \sum_i i g_i,$$
$$t = \sum_i t_i, \quad T = \sum_i i t_i,$$
$$m = \sum_{i,j} (j-1) m_{ij}, \quad M = \sum_{i,j} i m_{ij}.$$

We first consider the case that $(1, 2b+k)$ is not a base pair, i.e., $e_d > 2$, or, $e_d = 2$ and $e_l > 0$. As discussed above, we turn to enumerating the corresponding labelled plane trees on $[b+k+1]$ with $b+k+1$ as the root. In view of Theorem 2.2 and Lemma 3.1, the corresponding labelled plane trees for the RNA secondary structures in this case can be decomposed into $h+1$ rakes, where the rake rooted on $b+k+1$ corresponds to the exterior loop and the label $(b+k+1+h)^*$ is a leaf there. Based on Lemma 3.3, we enumerate the corresponding forests of rakes according to the following sequential construction.

(I) Determine the number of ways for constructing leaves of the rakes corresponding to hairpin loops. We assume those labelled rakes are arranged linearly according to the length of the corresponding hairpin loops increasingly, and those of the same length are arranged according to the minimum unmarked leaves increasingly. This is done as follows: choose P unmarked labels out of $[b+k]$ in $\binom{b+k}{P}$ different ways, and arrange them linearly in $P!$ ways, then cut the obtained sequence into segments such that the first p_1 segments have length one, the next p_2 segments have length two, etc. Here, a segment gives leaves of a hairpin-rake. However, the respective minimum unmarked leaves in the p_i segments of the same length i could yield any relative order. Thus, we next need to divide $\prod_{i>0} p_i!$ so that only the ones in increasing order are counted.

(II) Determine the number of ways for constructing leaves of the rakes corresponding to bulges. This can be first done in

$$\binom{b+k-P}{G} G! \frac{1}{\prod_{i>0} g_i!}$$

different ways to place the unmarked leaves. Next, we need to place a marked leaf other than $(b+k+1+h)^*$ either to the left or to the right of each segment of unmarked leaves. This can be achieved in $\binom{h-1}{g} 2^g g!$ different ways.

(III) Determine the leaves of the rakes corresponding to interior loops. This is obtained analogously to the case of bulges, with the exception that for an interior loop of length i, there are $i-1$ possible spaces between unmarked leaves to place a marked

leaf. Hence, we arrive at

$$\binom{b+k-P-G}{T} T! \frac{1}{\prod_{i>0} t_i!} \binom{h-1-g}{t} t! \prod_{i>1} (i-1)^{t_i}.$$

(IV) Determine the number of ways for constructing leaves of the rakes corresponding to multiloops. Suppose those rakes are arranged linearly according to the degree and then, the length of the corresponding multiloops increasingly, and those of the same length and degree are arranged according to the minimum marked label in increasing order. The enumeration is analogous, specific differences being the following: the length of a multiloop may be zero, and the marked leaves could be at any positions relative to the unmarked ones. Accordingly, we have

$$\binom{b+k-P-G-T}{M} \frac{M!}{\prod_{i\geq 0, j\geq 3} m_{ij}!} \prod_{i\geq 0, j\geq 3} \binom{i+j-1}{j-1}^{m_{ij}} \binom{h-1-g-t}{m} m!.$$

(V) Determine the number of ways for constructing leaves of the rake corresponding to the exterior loop. Recall that there is exactly one rake corresponding to the exterior loop. This can be done as follows: pick e_l unused unmarked labels from $[b+k]$ and arrange them together with the remaining unused marked labels linearly. This results in the multiplicity

$$\binom{b+k-P-G-T-M}{e_l} (e_l + e_d - 1)!.$$

(VI) Determine the number of ways for constructing stems for all rakes. Note that the stem of the rake corresponding to any loop except the exterior loop could have any size respecting the length distribution of the helices. This is equivalent to the number of ways of first arranging the remaining unused (unmarked) labels linearly and then dividing the resulting sequence into segments with the lengths respecting the length distribution of the helices, given by

$$b! \frac{h!}{\prod_{i>0} h_i!}.$$

Accordingly, the total number of distinct forests of rakes is given by

$$\begin{aligned}
&\binom{b+k}{P} \frac{P!}{\prod_{i>0} p_i!} \binom{b+k-P}{G} \frac{G!}{\prod_{i>0} g_i!} 2^g \binom{h-1}{g} g! \\
&\binom{b+k-P-G}{T} \frac{T!}{\prod_{i>0} t_i!} \prod_{i>1} (i-1)^{t_i} \binom{h-1-g}{t} t! \\
&\binom{b+k-P-G-T}{M} \frac{M!}{\prod_{i\geq 0, j\geq 3} m_{ij}!} \prod_{i\geq 0, j\geq 3} \binom{i+j-1}{j-1}^{m_{ij}} \binom{h-1-g-t}{m} m! \\
&\binom{b+k-P-G-T-M}{e_l} (e_l + e_d - 1)! b! \frac{h!}{\prod_{i>0} h_i!}.
\end{aligned}$$

Dividing the last number by $(b+k)!$ and expanding the involved binomial coefficients, i.e., using $\binom{m}{n} = \frac{m!}{n!(m-n)!}$, we observe lots of cancellation. For example, the first line of the above four-line expression (after dividing $(b+k)!$) becomes

$$\frac{1}{(b+k)!}\frac{(b+k)!}{P!(b+k-P)!}\frac{P!}{\prod_{i>0} p_i!}\frac{(b+k-P)!}{G!(b+k-P-G)!}\frac{G!}{\prod_{i>0} g_i!}2^g\frac{(h-1)!}{g!(h-1-g)!}g! = \frac{(h-1)!2^g}{(b+k-P-G)!(h-1-g)!\prod_{i>0} p_i!\prod_{i>0} g_i!}.$$

With this, we eventually obtain the desired formula.

It remains to consider the case of $(1, 2b+k)$ being a pair, i.e., $e_l = 0$ and $e_d = 2$. Then, the corresponding forests consist of h rakes where no rake is associated with the exterior loop. If there is only one helix (hence one loop being of a hairpin), i.e., $h = 1$ and $p = 1$, then there is only one rake in the corresponding forest and no marked label, and the desired number is one. Otherwise, we analogously enumerate the corresponding forests of rakes according to the following construction.

(a) Determine the number of ways for constructing leaves of the rakes corresponding to hairpin loops. This is analogous to the computation of hairpin loops in the previous case.

(b) Determine the number of ways for constructing leaves of the rakes corresponding to bulges. Note that we have only $h-1$ marked leaves in total. But in difference to the situation analyzed in the previous case, it is possible for the largest marked label $(b+k+1+h-1)^*$ to be contained in a rake that is associated with a bulge. Accordingly, the number of ways to construct bulge-rakes is given by

$$\binom{b+k-P}{G} G! \frac{1}{\prod_{i>0} g_i!} \binom{h-1}{g} 2^g g!.$$

(c) Determine the number of ways for constructing leaves of the rakes corresponding to interior loops and multi-loops. This is analogous to the previous case.

(d) Determine the number of ways for constructing stems for all rakes. The stem of the rake corresponding to any loop could have any length respecting the length distribution of the helices, with the exception that the length of one helix increases by one due to the auxiliary arc $(0, 2b+k+1)$. This is equivalent to the number of ways of first arranging the remaining unused unmarked labels (other than $b+k+1$) linearly and then cutting the resulting sequence into segments with the lengths respecting the original length distribution of the helices, and finally associating $b+k+1$ (as the root) to the rake that contains $(b+k+1+h-1)^*$, which is given by

$$b! \frac{h!}{\prod_{i>0} h_i!}.$$

As a result, the total number of distinct forests of rakes for this case is given by

$$\binom{b+k}{P}\frac{P!}{\prod_{i>0}p_i!}\binom{b+k-P}{G}\frac{G!}{\prod_{i>0}g_i!}2^g\binom{h-1}{g}g!b!\frac{h!}{\prod_{i>0}h_i!}$$
$$\binom{b+k-P-G}{T}\frac{T!}{\prod_{i>0}t_i!}\prod_{i>1}(i-1)^{t_i}\binom{h-1-g}{t}t!$$
$$\binom{b+k-P-G-T}{M}\frac{M!}{\prod_{i\geq 0,j\geq 3}m_{ij}!}\prod_{i\geq 0,j\geq 3}\binom{i+j-1}{j-1}^{m_{ij}}\binom{h-1-g-t}{m}m!.$$

Dividing the last number by $(b+k)!$ and subsequent simplification (involving lots of cancellation) yields the formula which agrees with the formula obtained for the previous case.

Proof of Theorem 3.5 The proof is similar to that of Theorem 3.4. We first consider the case $e_d > 2$, or, $e_l > 0$ and $e_d = 2$.

(I) Determine the number of ways for constructing stems of the rakes. There is one rake rooted in $b+k+1$ whose stem has length one. For the remaining h rakes, we arrange them according to the minimum elements contained in the stems in increasing order. This can be done by first picking b elements out of $[b+k]$ in $\binom{b+k}{b}$ possible ways and arranging them linearly in $b!$ ways. Next, we dissect the resulting sequence into h segments such that each segment has a length at least σ in $\binom{h+(b-h\sigma)-1}{b-h\sigma}$ different ways, and finally, we normalize by the factor $\frac{1}{h!}$.

(II) Determine the number of ways for constructing loop types associated with the rakes. Evidently, by construction the exterior loop is associated with the rake rooted on $b+k+1$. Among the remaining h rakes, we choose p of them for hairpin loops, g of them for bulges, t of them for interior loops, and m_j of them for multiloops of degree j. The number of ways of doing this is given by

$$\binom{h}{p}\binom{h-p}{g}\binom{h-p-g}{t}\frac{(\sum_j m_j)!}{\prod_j m_j!}.$$

(III) Place the marked leaves. The marked label $(b+k+1+h)^*$ is contained in the exterior loop. There are

$$\binom{h-1}{e_d-2}(h-e_d+1)!$$

ways to next place one marked leaf to each bulge and each interior loop, and place $j-1$ marked leaves to each multiloop of degree j.

(IV) Place the unmarked leaves. In each hairpin loop, there are at least θ unmarked leaves, and in each bulge, there is at least one unmarked leaf either to the left or to the right of the marked leaf. In each interior loop, there exists at least one unmarked leaf on both sides of the marked leaf. Subject to these constraints, the number of distinct

placements is given by

$$\binom{k}{e_l}\binom{p+g+2t+\sum_j jm_j+(k-e_l-p\theta-g-2t)-1}{k-e_l-p\theta-g-2t}(k-e_l)!2^g.$$

As for the exterior loop, the remaining labels are contained in the rake corresponding to the exterior loop, and there are $(e_l+e_d-1)!$ different ways to arrange them.

In conclusion, the number of ways for constructing such forests is given by

$$\begin{aligned}
&\binom{h+b-h\sigma-1}{b-h\sigma}\binom{b+k}{b}b!\frac{1}{h!}\\
&\binom{h}{p}\binom{h-p}{g}\binom{h-p-g}{t}\frac{(\sum_j m_j)!}{\prod_j m_j!}\\
&\binom{h-1}{e_d-2}(h-e_d+1)!(e_l+e_d-1)!\\
&\binom{k}{e_l}\binom{p+g+2t+\sum_j jm_j+(k-e_l-p\theta-g-2t)-1}{k-e_l-p\theta-g-2t}(k-e_l)!2^g
\end{aligned}$$

and dividing by $(b+k)!$ produces the formula.

Next, we consider the case $e_d=2$ and $e_l=0$.

If there is only one helix (hence one loop), then the desired number is obviously one. Otherwise, there are in total $h>1$ rakes, and $b+k+1$ and $(b+k+1+h-1)^*$ are contained in the same rake. Analogously, the desired number in this case reads

$$\begin{aligned}
&\frac{1}{(b+k)!}\binom{h+b-h\sigma-1}{b-h\sigma}\binom{b+k}{b}b!\frac{1}{h!}\\
&\binom{h}{p}\binom{h-p}{g}\binom{h-p-g}{t}\frac{(\sum_j m_j)!}{\prod_j m_j!}(h-1)!\\
&\binom{p+g+2t+\sum_j jm_j+(k-p\theta-g-2t)-1}{k-p\theta-g-2t}k!2^g.
\end{aligned}$$

Simplifying the last expression produces the formula in the theorem, completing the proof.

References

Chen RXF (2019) A new bijection between RNA secondary structures and plane trees and its consequences. Electron J Combin 26(4):4–48

Chen WYC (1990) A general bijective algorithm for trees. Proc Natl Acad Sci USA 87:9635–9639

Clote P (2006) Combinatorics of saturated secondary structures of RNA. J Comp Biol 13:1640–1657

Clote P, Ponty Y, Steyaert JM (2012) Expected distance between terminal nucleotides of RNA secondary structures. J Math Biol 65:581–599

Došlić T, Svrtan D, Veljan D (2004) Enumerative aspects of secondary structures. Discrete Math 285(2004):67–82

Ding Y, Lawrence CE (2003) A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res 31:7280–7301
Duchon P, Flajolet P, Louchard G, Schaeffer G (2004) Boltzmann samplers for the random generation of combinatorial structures. Combin Probab Comput 13:577–625
Fontana W, Konings D, Stadler PF, Schuster P (2004) Statistics of RNA secondary structures. Biopolymers 33:1389–1404
Hofacker IL, Schuster P, Stadler PF (1998) Combinatorics of RNA secondary structures. Discrete Appl Math 88:207–237
Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429–3431
Heitsch C, Poznanović S (2014) Combinatorial insights into RNA secondary structure. In: Jonoska N, Saito M (eds) Discrete and topological models in molecular biology. Springer, pp 145–166
Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res 37:D159–D162
Knuth D (1997) The Art of Computer Programming, Vol. 2 (3rd Ed.) Addison-Wesley Longman: Boston
Poznanović S, Heitsch C (2014) Asymptotic distribution of motifs in a stochastic context-free grammar model of RNA folding. J Math Biol 69:1743–1772
Han HSW, Reidys CM (2012) The $5'$-$3'$ distance of RNA secondary structures. J Comp Biol 19:867–878
Klazar M (1998) On trees and noncrossing partitions. Discrete Appl Math 82:263–269
Lorenz W, Ponty Y, Clote P (2008) Asymptotics of RNA shapes. J Comp Biol 15:31–63
Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911–940
Nebel ME (2003) Combinatorial properties of RNA secondary structures. J Comp Biol 9(3):541–574
Nebel ME, Scheid A, Weinberg F (2011) Random generation of RNA secondary structures according to native distributions. Algorithms Mol Biol 6:24
Ponty Y (2008) Efficient sampling of RNA secondary structures from the Boltzmann ensemble of low-energy: the boustrophedon method. J Math Biol 56:107–127
Schmitt WR, Waterman MS (1994) Linear trees and RNA secondary structure. Discrete Appl Math 51(3):317–323
Smith TF, Waterman MS (1978) RNA secondary structure. Math Biol 42:31–49
Stein PR, Waterman MS (1979) On some new sequences generalizing the Catalan and Motzkin numbers. Discrete Math 26:261–272
Simion R (2000) Noncrossing partitions. Discrete Math 217:367–409
Sloma MF, Mathews DH (2016) Exact calculation of loop formation probability identifies folding motifs in RNA secondary structures. RNA 22:1808–1818
Sato K, Akiyama M, Sakakibara Y (2021) RNA secondary structure prediction using deep learning with thermodynamic integration. Nat Comm 12:941
Staple DW, Butcher SE (2005) Pseudoknots: RNA Structures with Diverse Functions. PLOS Biol 3(6):e213
Tuerk C, MacDougal S, Gold L (1992) RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. Proc Natl Acad Sci USA 89(15):6988–6992
Waterman MS (1978) Secondary structure of single-stranded nucleic acids. In: Rota G-C (ed) Studies on foundations and combinatorics, Advances in mathematics supplementary studies. Academic Press, New York, pp 167–212
Waterman MS (1979) Combinatorics of RNA hairpins and cloverleaves. Stud Appl Math 60:91–98
Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244:48–52
Zuker M, Sankoff D (1984) RNA secondary structures and their prediction. Bull Math Bio 46:591–621
Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406–3415