A new bijection between RNA secondary structures and plane trees and its consequences

Ricky X. F. Chen

Biocomplexity Institute and Initiative
University of Virginia
Charlottesville, Virginia, U.S.A.

chen.ricky1982@gmail.com

Submitted: Feb 28, 2019; Accepted: Dec 5, 2019; Published: Dec 22, 2019

Abstract

In this paper, we first present a new bijection between RNA secondary structures and plane trees. Combined with the Schmitt-Waterman bijection between these objects, we then obtain a bijection on plane trees that relates the horizontal fiber decomposition associated to internal vertices to the degrees of odd-level vertices while the vertical path decomposition associated to leaves is related to the degrees of even-level vertices. To the best of our knowledge, only the former relation (i.e., horizontal vs odd-level) due to Deutsch is known. As a consequence, we obtain enumeration results for various classes of plane trees, e.g., refining the Narayana numbers and the enumeration involving young leaves due to Chen, Deutsch and Elizalde, and counting a newly introduced 'vertical' version of k-ary trees. The enumeration results can be also formulated in terms of RNA secondary structures with certain parameterized features, which might have some biological significance.

Mathematics Subject Classifications: 05C05, 05A19, 05A15

1 Introduction

Ribonucleic acid (RNA) plays an important role in various biological processes within cells, ranging from catalytic activity to gene expression. RNA is described by its sequence of bases: A (adenine), U (uracil), G (guanine), and C (cytosine). These single-stranded molecules fold onto themselves forming helical structures, by forming base pairs where A pairs with U while G pairs with C. The sequence of bases of the RNA molecule is known as primary structure, and it is determined experimentally. A subset of the helical structure consistent with a planar graph is known as a secondary structure.

More than three decades ago, Waterman and his coworkers pioneered the combinatorics and prediction of RNA secondary structures [6–10]. In particular, enumeration of

the number of secondary structures over a sequence of length n that have k base pairs has been done in Schmitt and Waterman [6] by establishing a bijection between secondary structures and plane trees.

In this paper, we present a new bijection between RNA secondary structures and plane trees. Combining our new bijection and the Schmitt-Waterman bijection [6] leads to a new bijection φ on plane trees which enables us to obtain many interesting results. The most relevant studies on plane trees in the literature are as follows. In his paper [4], Deutsch presented an implicit, iteratively constructed bijection on plane trees which allowed him to show that the number of vertices of degree $q > 0$ and the number of odd-level vertices of degree $q - 1$ are equidistributed on the set of all plane trees. In particular, the case $q = 1$ implies that the number of plane trees with k leaves is the same as the number of plane trees with k even-level vertices, where the former is well known to be the Narayana number. In Chen, Deutsch and Elizalde [2], the authors classified leaves of a plane tree into old and young leaves, where a leaf is called old if it is the leftmost child of its parent and young otherwise, and they obtained enumerative results with respect to these bijectively.

In a plane tree, the horizontal elementary substructures are fibers associated to its internal vertices, i.e., internal vertices and their respective children. This horizontal fiber decomposition has been well understood through extensive studies of plane trees according to the number of internal vertices and their degree distribution. A dual perspective which appears to be ignored (at least less-studied) is that, vertically, a plane tree can be decomposed into paths associated to its leaves. Through our new bijection φ on plane trees, not only can we show the correspondence between horizontal fibers and odd-level vertices established by Deutsch [4], but we can also show that the vertical paths associated to leaves actually correspond to even-level vertices at the same time. Namely, we discover that the joint distribution of horizontal fibers and vertical paths is the same as the joint distribution of odd- and even-level vertices.

As a consequence, based on these equidistribution results, we can compute the number of plane trees with certain restricted path lengths in the vertical path decomposition (and with restriction on the horizontal fiber decomposition) via the multivariate Lagrange inversion formula, which refines the Narayana numbers and gives rise to some new results. For example, k-ary trees are plane trees with every horizontal fiber having a size k, which is known to be counted by certain generalized Catalan numbers, see, e.g., Chen [3]. Here we are interested in their vertical duals, i.e., plane trees where any vertical path associated to a leaf somehow has a size k. We show that these 'vertical' k-ary trees are counted by numbers very similar to the generalized Catalan numbers. In addition, we observe that the lengths of the paths associated to leaves can enable us to distinguish between old and young leaves. Accordingly, we refine some results obtained in Chen, Deutsch and Elizalde [2].

2 A new bijection

We first recall the definition of RNA secondary structures. Let $[n] = \{1, 2, \ldots, n\}$. An RNA secondary structure of length n is a simple graph with vertices in $[n]$ and edges in

E satisfying

- if $(i,j) \in E$, then $|i-j| \geqslant 2$;
- 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 $\{r : i \leqslant r \leqslant j\}$).

We typically draw an RNA secondary structure in the following manner: we place all vertices in a horizontal line and we draw an edge as an arc in the upper half-plane. Then, the second condition in the above definition guarantees that any two arcs do not cross. The vertex of an arc with a smaller label is called the left-end of the arc, and a vertex not adjacent to any edge is called an isolated base. In addition, if (i,j) is an arc, we say that an arc (i_1,j_1) (resp. an isolated base k) is covered by (i,j) if $[i_1,j_1] \subset [i,j]$ (resp. $k \in [i,j]$), and we also say that the arcs (i,j) and (i_1,j_1) nest with each other.

A plane tree T can be recursively defined as an unlabeled tree with one distinguished vertex v called the root of T, where the unlabeled trees T' obtained by deleting v as well as its incident edges from T are linearly ordered, and T' is a plane tree with the vertex adjacent to v in T as its root. 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, and the vertices adjacent to v on a lower level are called the children of v. The vertices on level $2i$ for $i \geqslant 0$ are called even-level vertices and the rest are called odd-level vertices. A vertex is called a leaf if it has no children, and is called an internal vertex otherwise. We will draw plane trees with the 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.

Theorem 1. *There is a bijection ϕ between the set of RNA secondary structures of length $2a+k$ with k isolated bases and the set of plane trees with $a+k$ edges and k even-level vertices.*

Proof. Let R be an RNA secondary structure of length $2a+k$ with k isolated bases. We construct a plane tree $\phi(R)$ as follows:

S1: Put a big arc covering all existing arcs and isolated bases of R and still refer to the obtained structure as R in the following. Label the isolated bases in R with $b_1, b_2, \ldots, b_k$ left-to-right, and label the arcs with $e_0, e_1, \ldots, e_a$ based on the left-to-right order of their left-ends;

S2: Start with a vertex that will be the root of $\phi(R)$ and label the vertex with b_1, and generate k_1 children for b_1 if there are k_1 arcs covering the isolated base b_1 in R, where the children from left to right correspond to these k_1 arcs from the outermost to the innermost and are labeled correspondingly, respectively;

S3: Set $j = 2$. While $j \leqslant k$, put a new child to the left of all existing children of the vertex that corresponds to the innermost arc covering the isolated base b_j in the current partially constructed tree and label the newly generated child with b_j, and next generate k_j children for the vertex b_j if there are k_j unused arcs (i.e., those with

labels not appearing in the current partial tree) covering the isolated base b_j, where again the children from left to right correspond to these k_j arcs from the outermost to the innermost and are labeled correspondingly, respectively, and set $j = j + 1$.

The following properties are observed in the above construction: (i) the vertices b_i for all i are even-level vertices, and vice versa; (ii) the sequence $e_0e_1 \cdots e_a$ will be obtained if the children of the even-level vertices (in the order $b_1b_2 \cdots b_k$) are collected left-to-right sequentially; (iii) the sequence $b_1b_2 \cdots b_k$ will be obtained if the even-level vertices are searched by depth-first search from right to left. (i) and (ii) should be straightforward, and (iii) can be shown by induction. Hence, the labels of the vertices can be easily and uniquely recovered after being removed whence the obtained structure with labels removed is a plane tree with $a + k$ edges.

Before we specify the reverse algorithm, we mention two additional properties in the above forward algorithm which are important to better understand the reverse algorithm to come: (iv) the number of children of an isolated base (as a vertex in $\phi(R)$) is the number of left-ends of arcs between the present isolated base and the one immediately to the left of it if any; (v) the parent of an isolated base if any is the innermost arc, excluding those with the left-ends identified in (iv) if any, that covers the isolated base.

Let T be a plane tree with $a + k$ edges and k even-level vertices. It is not hard to verify that we can construct $\phi^{-1}(T)$ following the steps below:

SS1: Label the even-level vertices of T with $b_1, b_2, \ldots, b_k$ respectively in the depth-first searching manner from right to left, and label the left-to-right children of even-level vertices arranged in the order $b_1b_2 \cdots b_k$ sequentially with $e_0, e_1, \ldots, e_a$;

SS2: On a horizontal line, start with an isolated base labeled with b_1, cover b_1 with k_1 mutually nesting arcs if the vertex b_1 in T has k_1 children, and label these arcs based on the order of their left-ends left-to-right with $e_0, e_1, \ldots, e_{k_1-1}$;

SS3: Set $j = 2$. While $j \leqslant k$, place an isolated base with a label b_j to the right of all existing isolated bases such that, (I) the newly placed isolated base is covered by the arc e_t but not by e_s for any $s > t$ if the vertex b_j is a child of the vertex e_t in T, and (II) generate k_j mutually nesting arcs to cover the isolated base b_j if b_j has k_j children in T, (II1) without crossing with any existing arcs, as well as, (II2) without covering b_{j-1}, and label these k_j arcs left-to-right correspondingly, and set $j = j+1$. Suppose we have just completed all steps though $j - 1$. Then, it is clear that the left-ends of all already generated arcs are to the left of b_{j-1}. Next consider j. If the vertex b_j is a child of the vertex e_t in T, then the innermost arc, excluding those later added by (II), that covers b_j should be e_t. Thus the condition (I) is necessary. Next, if b_j has k_j children in T, then according to (iv), in $\phi^{-1}(T)$, there should be k_j arcs whose left-ends lie between b_j and b_{j-1}. In order to guarantee this, we need to generate k_j arcs to cover b_j. The condition (II1) is clearly required to not violate the definition of secondary structures, while (II2) is essentially the same as (iv).

Finally, removing the arc e_0 as well as all e-labels and b-labels will give us a secondary structure $\phi^{-1}(T)$. □

See Figure 1 for an illustration of the bijection ϕ.

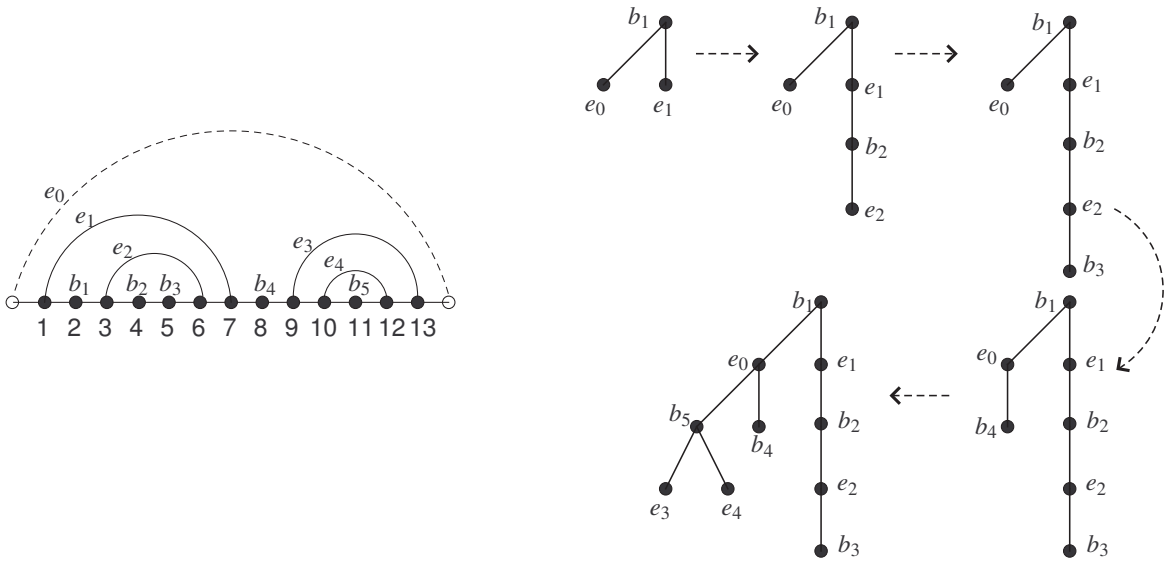

Figure 1: An RNA secondary structure (left) and the process of constructing its corresponding plane tree (right).

Remark 2. There are several variations of the bijection ϕ. For instance, we can put the covering arcs of an isolated base right-to-left as children, or we can put a newly generated child of an arc to the right of existing ones, or we can read the isolated bases from right to left, or different combinations of these.

3 Consequences

In this section, we present a number of applications by combining the bijection ϕ and the Schmitt-Waterman bijection [6].

The Schmitt-Waterman bijection from RNA secondary structures to plane trees can be briefly summarized as follows: for a given RNA secondary structure, put a big arc covering everything. Next, view each arc and isolated base as a vertex in a tree rooted at the vertex corresponding to the big arc, where the left-to-right children of a vertex v in the tree are the vertices corresponding to the left-to-right arcs and isolated bases directly covered by v (if v is an arc). Therefore, the Schmitt-Waterman bijection maps an RNA secondary structure with k isolated bases to a plane tree with k leaves. Thus, combined with our bijection ϕ, with RNA secondary structures serving as intermediate objects, we immediately obtain the following well-known result [4].

Corollary 3. *The number of plane trees with n edges and k leaves equals the number of plane trees with n edges and k even-level vertices.*

Proof. Let f be the Schmitt-Waterman bijection from RNA secondary structures to plane trees. Clearly, $\varphi = \phi \circ f^{-1}$ gives a bijection from the set of plane trees with n edges and k leaves to the set of plane trees with n edges and k even-level vertices. □

Figure 2 gives an example of the bijection φ.

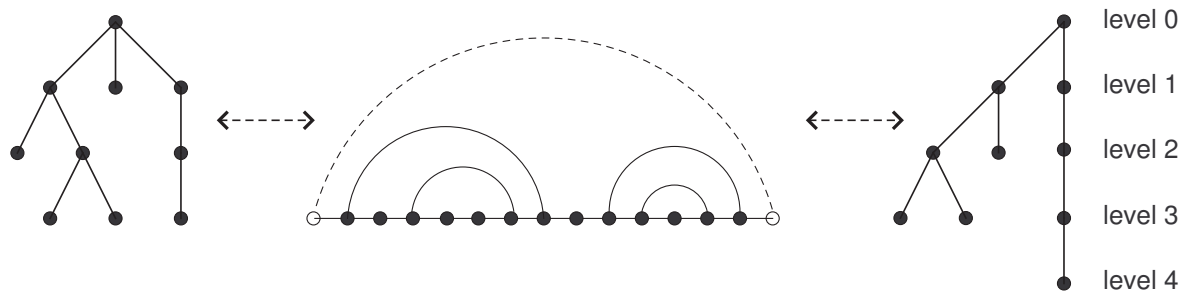

Figure 2: Correspondence between a plane tree with 5 leaves and a plane tree with 5 even-level vertices.

The following corollary which is implied in [4] can be obtained as well.

Corollary 4. *The bijection φ restricts to a bijection between the set of plane trees of n edges with a k-element multiset $\mathcal{M}$ as its outdegree distribution of the internal vertices and the set of plane trees of n edges having the multiset $\mathcal{M}' = \{z - 1 \mid z \in \mathcal{M}\}$ as the outdegree distribution of the odd-level vertices.*

Proof. Let R be a secondary structure (with the big arc added). An arc e corresponds to an internal vertex in the plane tree $f(R)$, and corresponds to an odd-level vertex in the plane tree $\phi(R)$. If e covers t disconnected components (here a component is either an isolated base or an arc and everything covered by the arc), then the outdegree of the corresponding vertex in $f(R)$ is t. Note that, in each component, by definition of RNA secondary structure, there is at least one isolated base. By construction of $\phi(R)$, the arc e is a child of the vertex corresponding to the first (left-to-right) isolated base in the first component covered by e, while the first isolated base in each other component covered by e must be a child of e as a vertex in $\phi(R)$. Thus, the outdegree of the odd-level vertex corresponding to e in $\phi(R)$ is $t - 1$. The converse can be argued analogously, completing the proof. □

Remark 5. The reader can check that the outcomes of the implicit, iterative bijection of Deutsch [4] are quite similar to those of our bijection φ. In fact, we believe that the former can be transformed into the latter by specifying further steps in the iterative construction there. However, our bijections in this paper are not motivated by revising the former bijection. Nevertheless, we believe that our bijection φ discovered in the study of RNA secondary structures is more explicit and more constructive. More importantly, the results in the rest of this paper are not discussed in [4].

By inspecting our bijections more carefully, we can obtain more properties on plane trees, which will be the main theme of the rest of the paper. Note that there is a unique path from a leaf to the root of a plane tree. Then, we can decompose a plane tree into a set of paths where each path has a leaf as a terminate vertex. The decomposition works as follows: suppose all leaves are ordered by their relative order in the depth-first search from left to right. The first path is the path from the first leaf to the root. For $t > 1$, the t-th path is the remaining part of the path from the t-th leaf to the root after the previously obtained paths are removed from the tree, or equivalently, the t-th path should go from the t-th leaf up to the first vertex that is already in a path that has been obtained. We refer to this decomposition as the vertical path decomposition associated to leaves. See Figure 3 (left) for an illustration. We will call the multiset consisting of the lengths of the obtained paths the path distribution of the given tree.

Theorem 6. *The bijection φ restricts to a bijection between the set of plane trees of n edges with a k-element multiset $\mathcal{M}$ as its path distribution and the set of plane trees of n edges having $\mathcal{M}$ as the degree distribution of the even-level vertices.*

Proof. Let R be a secondary structure with an added big arc. In the Schmitt-Waterman bijection f, the path from a leaf to the root in $f(R)$ consists of the leaf itself (an isolated base) and all arcs (including the added big arc) covering the isolated base. Thus, the first path is determined by the first isolated base b_1 and all arcs covering b_1. So the length of the first path is the number of these arcs which equals the number of children (hence the degree) of b_1 in $\phi(R)$. It is not hard to see that the length of the i-th ($i > 1$) path is the number one larger than the number of 'unused' arcs covering the i-th leaf after the first $i-1$ paths have been obtained in the decomposition process. Thus, the length of the i-th path is one larger than the number of children of b_i in $\phi(R)$ which is the degree of b_i in $\phi(R)$. The converse is also not hard to see whence the theorem. □

Based on Corollary 4 and Theorem 6, we can conclude that, in a sense, the vertical determines the even-levels while the horizontal determines the odd-levels. Although the horizontal-odd relation is known, to the best of our knowledge, the two relations as a whole have not been addressed. We also remark that it seems not easy to motivate the vertical-even relation from Deutsch's bijection [4] due to its implicit, iterative nature.

Let T be a plane tree with k leaves. Let $\mathfrak{l}_t$ be the number one less than the length of the t-th path in the path decomposition of T for $1 < t \leqslant k$, and let $\mathfrak{l}_1$ be the length of the first path in the path decomposition. We denote the multiset consisting of these numbers $\mathfrak{l}_t$ $(1 \leqslant t \leqslant k)$ as $\widetilde{\mathcal{M}}(T)$. With an application of the multivariate Lagrange inversion formula, we will obtain the forthcoming theorem. Let us first recall the following version of the multivariate (bivariate) Lagrange inversion formula [1, 5]:
Let $g(x_1, x_2),\ f_1(x_1, x_2),\ f_2(x_1, x_2)$ be formal power series in $x_1,\ x_2$ such that $f_i(0, 0) \neq 0$. Then, the set of equations $w_i = t_i f_i(w_1, w_2)$ for $1 \leqslant i \leqslant 2$ uniquely determine the w_i as formal power series in $t_1,\ t_2$, and

$$[t_1^p t_2^q] g(w_1, w_2) = [x_1^p x_2^q] g(x_1, x_2) f_1^p(x_1, x_2) f_2^q(x_1, x_2) \det \begin{Bmatrix} 1 - \frac{x_1}{f_1}\frac{\partial f_1}{\partial x_1} & -\frac{x_1}{f_2}\frac{\partial f_2}{\partial x_1} \\ -\frac{x_2}{f_1}\frac{\partial f_1}{\partial x_2} & 1 - \frac{x_2}{f_2}\frac{\partial f_2}{\partial x_2} \end{Bmatrix},$$

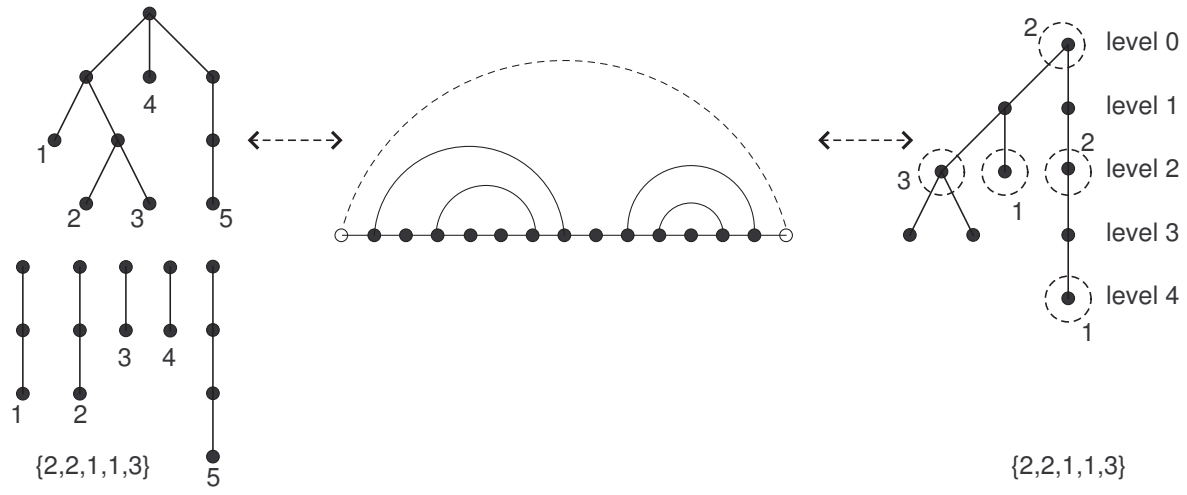

Figure 3: The bijection φ preserves the multiset $\{2, 2, 1, 1, 3\}$.

where $[t_1^p t_2^q]$ denotes the coefficient of $t_1^p t_2^q$.

Theorem 7. *The number $C_{k,h}(n)$ of plane trees T with $n > 0$ edges and k leaves such that $\max \widetilde{\mathcal{M}}(T) \leqslant h$ is given by*

$$C_{k,h}(n) = \frac{1}{n}\binom{n}{k} \sum_{i \geqslant 0}^{i \leqslant \frac{n+1-k}{h+1}} (-1)^i \binom{k}{i} \binom{n - i(h+1)}{k-1}. \tag{1}$$

Proof. Based on Theorem 6, the number of plane trees T with n edges and k leaves where $\max \widetilde{\mathcal{M}}(T) \leqslant h$ is equal to the number of plane trees of n edges with k even-level vertices such that every even-level vertex has at most h children. The latter can be computed as shown below:

Given two sets E and O of vertices, we call a plane tree T on $E \bigcup O$ a set-alternating tree if vertices on any path starting from the root of T alternate in the two sets. Let

$$w_1(t_1, t_2) = \sum_{T \in \mathscr{P}_E} t_1^{\#\text{vertices in } E \text{ in } T} t_2^{\#\text{vertices in } O \text{ in } T},$$
$$w_2(t_1, t_2) = \sum_{T \in \mathscr{P}_O} t_1^{\#\text{vertices in } E \text{ in } T} t_2^{\#\text{vertices in } O \text{ in } T},$$

where $\mathscr{P}_E$ denotes the set of set-alternating plane trees with root in E and every E-vertex having at most h children while $\mathscr{P}_O$ denotes the set of set-alternating plane trees with root in O and every E-vertex having at most h children. Then, it is obvious that

$$w_1 = t_1 \frac{1 - w_2^{h+1}}{1 - w_2}, \quad w_2 = t_2 \frac{1}{1 - w_1}.$$

Clearly, the number of plane trees with n edges and k even-level vertices such that every even-level vertex has at most h children is the same as the number of set-alternating trees of n edges with root in E and with every E-vertex having at most h children, which is obviously $[t_1^k t_2^{n+1-k}]w_1$.

In terms of the above bivariate Lagrange inversion formula, we have

$$g(x_1,x_2)=x_1,\quad f_1(x_1,x_2)=\frac{1-x_2^{h+1}}{1-x_2},\quad f_2(x_1,x_2)=\frac{1}{1-x_1}.$$

$$\begin{aligned}
[t_1^p t_2^q]w_1 &= [x_1^p x_2^q]g\cdot f_1^p\cdot f_2^q\cdot\det\begin{Bmatrix}1-\frac{x_1}{f_1}\frac{\partial f_1}{\partial x_1} & -\frac{x_1}{f_2}\frac{\partial f_2}{\partial x_1}\\ -\frac{x_2}{f_1}\frac{\partial f_1}{\partial x_2} & 1-\frac{x_2}{f_2}\frac{\partial f_2}{\partial x_2}\end{Bmatrix}\\
&= [x_1^p x_2^q]\frac{(1-x_2^{h+1})^p}{(1-x_2)^p}\frac{x_1}{(1-x_1)^q}\left(1-\frac{x_1x_2(1-x_2)}{(1-x_1)(1-x_2^{h+1})}\Big[\frac{1-x_2^h}{(1-x_2)^2}-\frac{hx_2^h}{1-x_2}\Big]\right)\\
&= [x_1^{p-1}x_2^q](1-x_1)^{-q-1}(1-x_2)^{-p-1}(1-x_2^{h+1})^{p-1}\\
&\qquad\times[(1-x_1)-x_2(1-x_2^{h+1})-x_2^{h+1}+(h+2)x_1x_2^{h+1}-(h+1)x_1x_2^{h+2}]\\
&= \binom{q+p-2}{p-1}\sum_{i\geqslant 0}^{i\leqslant\frac{q}{h+1}}(-1)^i\binom{p-1}{i}\binom{p+q-i(h+1)}{q-i(h+1)} &\text{(A1)}\\
&\quad-\binom{q+p-1}{p-1}\sum_{i\geqslant 0}^{i\leqslant\frac{q-1}{h+1}}(-1)^i\binom{p}{i}\binom{p+q-1-i(h+1)}{q-1-i(h+1)} &\text{(A2)}\\
&\quad-\binom{q+p-1}{p-1}\sum_{i\geqslant 0}^{i\leqslant\frac{q-1-h}{h+1}}(-1)^i\binom{p-1}{i}\binom{p+q-h-1-i(h+1)}{q-h-1-i(h+1)} &\text{(A3)}\\
&\quad+(h+2)\binom{q+p-2}{p-2}\sum_{i\geqslant 0}^{i\leqslant\frac{q-1-h}{h+1}}(-1)^i\binom{p-1}{i}\binom{p+q-h-1-i(h+1)}{q-h-1-i(h+1)} &\text{(A4)}\\
&\quad-(h+1)\binom{q+p-2}{p-2}\sum_{i\geqslant 0}^{i\leqslant\frac{q-2-h}{h+1}}(-1)^i\binom{p-1}{i}\binom{p+q-h-2-i(h+1)}{q-h-2-i(h+1)}. &\text{(A5)}
\end{aligned}$$

The last quantity can be simplified into

$$\begin{aligned}
&\binom{p+q-2}{p-1}\sum_{i\geqslant 0}^{i\leqslant\frac{q-1}{h+1}}(-1)^i\binom{p}{i}\binom{p+q-i(h+1)}{p}\frac{1}{p+q-i(h+1)} &\text{(B1)}\\
&\quad+(h+1)\binom{q+p-2}{p-2}\sum_{i>\frac{q-1}{h+1}}^{i\leqslant\frac{q}{h+1}}(-1)^{i-1}\binom{p-1}{i-1}\binom{p+q-i(h+1)}{p} &\text{(B2)}\\
&\quad+\binom{q+p-2}{p-1}\sum_{i>\frac{q-1}{h+1}}^{i\leqslant\frac{q}{h+1}}(-1)^i\binom{p}{i}\binom{p+q-i(h+1)}{p}, &\text{(B3)}
\end{aligned}$$

where more detailed manipulations can be found in Appendix A.

Note that there is at most one integer in the interval $(\frac{q-1}{h+1}, \frac{q}{h+1}]$ for any integer $h \geqslant 0$. Specifically, if $h+1 \mid q$, there exists exactly one integer m in the interval such that $m(h+1) = q$. In this case, the sum of B2 and B3 is

$$(-1)^m \left[\binom{q+p-2}{p-1} \binom{p}{m} - (h+1) \binom{q+p-2}{p-2} \binom{p-1}{m-1} \right] = \frac{(-1)^m}{p} \binom{p+q-2}{p-1} \binom{p}{m},$$

which can be merged into B1 by changing $i \leqslant \frac{q-1}{h+1}$ into $i \leqslant \frac{q}{h+1}$. If $h+1 \nmid q$, there is no integer in that interval, thus B2 and B3 are both zero. Furthermore, any integer $i \leqslant \frac{q}{h+1}$ must satisfy $i \leqslant \frac{q-1}{h+1}$. Hence, changing $i \leqslant \frac{q-1}{h+1}$ into $i \leqslant \frac{q}{h+1}$ in B1 makes no difference. Therefore, the sum of B1, B2 and B3 can be written in a unified form in all cases, which gives the quantity in the theorem after setting $p = k$, $q = n+1-k$. □

For example, there are 14 plane trees T with five edges and three leaves such that $\max \widetilde{\mathcal{M}}(T) \leqslant 2$, which are shown below:

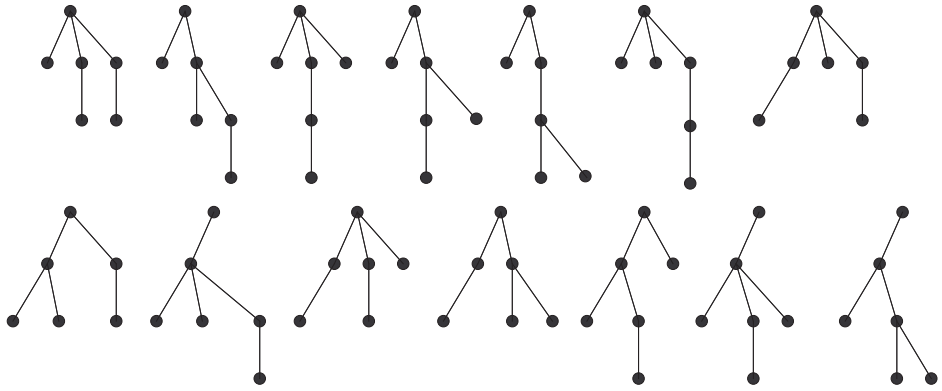

Figure 4: The 14 plane trees.

Note that for any plane tree T with n edges, we have $\max \widetilde{\mathcal{M}}(T) \leqslant n$. Then, we immediately have

Corollary 8. *The number of plane trees with n edges and k leaves is given by the Narayana number*

$$\frac{1}{n} \binom{n}{k} \binom{n}{k-1} = C_{n,k}(n).$$

Theorem 9. *The number of plane trees T with $n > 0$ edges and k leaves such that $\mathfrak{l}_t = h$ for $1 \leqslant t \leqslant k$ is given by $\frac{h}{n}\binom{n}{k-1}$ if $(h+1)k = n+1$, and 0 otherwise.*

Proof. It is easy to see that $n+1 = (h+1)k$ if $\mathfrak{l}_t = h$ for $1 \leqslant t \leqslant k$. The remaining part can be shown analogously as Theorem 7. □

An equivalent formulation of Theorem 9 is that the number of plane trees T with k leaves such that $\mathfrak{l}_t = h$ for $1 \leqslant t \leqslant k$ is given by $\frac{h}{(h+1)k-1}\binom{(h+1)k-1}{k-1}$, which is very similar to the number $\frac{1}{hk+1}\binom{hk+1}{k}$ of h-ary trees with k internal vertices (e.g., see [3]). So, to some extent, these trees can be viewed as 'h-ary' trees defined from another angle, i.e., vertically.

As a corollary of Theorem 7 and Theorem 9, we obtain a new curious identity below.

Corollary 10. *For $n \geqslant 0,\ k \geqslant 1$, we have*

$$\sum_{i\geqslant 0}^{i\leqslant \frac{n+1-k}{h+1}} (-1)^i \binom{k}{i}\binom{n-i(h+1)}{k-1} = \begin{cases} 0, & \text{for } n+1 > (h+1)k,\ h \geqslant 0; \\ 1, & \text{for } n+1 = (h+1)k,\ h \geqslant 0. \end{cases} \tag{2}$$

Proof. Note that a plane tree T with k leaves and $\max \widetilde{\mathcal{M}}(T) \leqslant h$ can have at most $(h+1)k$ vertices. Thus, $C_{k,h}(n) = 0$ for $n+1 > (h+1)k$, which gives the first case. If the number of vertices $n+1 = (h+1)k$, then $\mathfrak{l}_t = h$ for $1 \leqslant t \leqslant k$. Applying Theorem 7 and Theorem 9 gives the second case, completing the proof. □

Note that Eq. (2) can be rewritten as

$$\sum_{i\geqslant 0}^{i\leqslant \frac{hk+\delta}{h+1}} (-1)^i \binom{k}{i}\binom{(k-i)(h+1)+\delta-1}{k-1} = \begin{cases} 0, & \text{for } \delta > 0,\ h \geqslant 0; \\ 1, & \text{for } \delta = 0,\ h \geqslant 0. \end{cases} \tag{3}$$

It might be interesting to find a direct combinatorial proof for the identity.

The leaves of a plane tree are classified into old and young leaves in Chen, Deutsch and Elizalde [2]: a leaf is an old leaf if it is the leftmost child of its parent, and it is a young leaf otherwise. We can identify young and old leaves from the above path decomposition of a plane tree.

Lemma 11. *In the path decomposition of a plane tree, a leaf contained in a length one path other than the first path is a young leaf, and vice versa.*

Proof. Let T be a plane tree and v is an old leaf there. Suppose the parent of v is u_1. We have the following two cases: (i) If u_1 is the root of T, then the path containing v is the first path and has length 1 since v is the leftmost child of the root; (ii) Otherwise, u_1 has a parent u_2. By definition, v is the leftmost child of u_1. Then, the path from v to the root is the 'leftmost' path containing the edge (u_1, u_2). Thus, the edge (u_1, u_2) can not be contained in any previous path in the path decomposition. So, the path containing v has length at least 2. In summary, an old leaf either induces a path of length at least two or the first path with a length one. Conversely, the first leaf is clearly always an old leaf regardless of the length of the first path. For $t > 1$, if the t-th path has a length at least two, then we can conclude the t-th leaf to be old by arguing analogously as the case (ii), whence the lemma. □

Corollary 12. *The number of plane trees with n edges, k leaves and i young leaves is the same as the number of plane trees with n edges and k even-level vertices where i of them are leaves.*

Proof. Considering Theorem 6 and Lemma 11 together completes the proof. □

Based on Corollary 12, we can compute the number of plane trees with restrictions on path lengths and the number of young leaves.

Theorem 13. *The number $C_{h,y,k}(n)$ of plane trees T with $n > 0$ edges and k leaves such that $\max \widetilde{\mathcal{M}}(T) \leqslant h$ and y out of k leaves are young leaves is given by*

$$\frac{1}{k-y}\binom{n-y-1}{k-y-1}\binom{n-1}{y}\sum_{i\geqslant 0}^{i\leqslant \frac{n+1+y-2k}{h}}(-1)^i\binom{k-y}{i}\binom{n-k-ih}{n+1+y-2k-ih}. \tag{4}$$

Proof. Let $\bar{C}_{h,j}(m)$ be the number of plane trees with m edges and j leaves such that $1 \leqslant \mathfrak{l}_t \leqslant h$ for $1 \leqslant t \leqslant j$. We first show that

$$C_{h,y,k}(n) = \binom{n-1}{y}\bar{C}_{h,k-y}(n-y).$$

This can be seen as follows: On the one hand, for each plane tree T with n edges and k leaves such that $\max \widetilde{\mathcal{M}}(T) \leqslant h$ and y out of k leaves are young leaves, if we delete the young leaves, we will obtain a plane tree with $n-y$ edges and $k-y$ leaves such that $1 \leqslant \mathfrak{l}_t \leqslant h$ for $1 \leqslant t \leqslant k-y$ due to Lemma 11. On the other hand, for each plane tree of the latter case, inserting y leaves into the sectors other than the leftmost ones around these $n+1-k$ internal vertices will generate a plane tree of the former case. There are $2(n-y)-(k-y)-(n+1-k)+1 = n-y$ such sectors, which gives in total $\binom{n-1}{y}$ different ways of inserting y leaves, whence we have the desired relation.

Next, based on Theorem 6, the number $\bar{C}_{h,k-y}(n-y)$ also counts plane trees T of $n-y$ edges with $k-y$ even-level vertices such that every even-level vertex has at least one child and at most h children. Employing an analogous computation as in Theorem 7, we obtain

$$\bar{C}_{h,k-y}(n-y) = \frac{1}{k-y}\binom{n-1-y}{k-y-1}\sum_{i\geqslant 0}^{i\leqslant \frac{n+1+y-2k}{h}}(-1)^i\binom{k-y}{i}\binom{n-k-ih}{n+1+y-2k-ih},$$

and the proof follows. □

As an immediate consequence, we recover the following result obtained in [2].

Corollary 14. *The number of plane trees with $n > 0$ edges, i old leaves and j young leaves is $\frac{1}{n}\binom{n}{i}\binom{n-i}{j}\binom{n-i-j}{i-1}$.*

Proof. Obviously, the desired number is

$$C_{n,j,i+j}(n) = \frac{1}{i}\binom{n-1-j}{i-1}\binom{n-1}{j}\binom{n-i-j}{i-1} = \frac{1}{n}\binom{n}{i}\binom{n-i}{j}\binom{n-i-j}{i-1},$$

and the proof follows. □

The number of plane trees with n edges and i old (resp. young) leaves can be obtained by summing over all possible j's (resp. i's) in Corollary 14, which can be found in Chen, Deutsch and Elizalde [2] as well.

Based on Theorem 6 and Corollary 4, we can also count plane trees with both vertical and horizontal restrictions. It should be noted that it is generally not possible to have a plane tree with every vertex having k children while every path (from the path decomposition) has exactly length k, i.e., in a sense being regular both 'horizontally' and 'vertically'. However, we can have a weaker version of these regular trees, called strong k-ary trees. A k-ary tree T is called strong if $\max \widetilde{\mathcal{M}}(T) \leqslant k$. Applying an analogous computation as in Theorem 7, we obtain

Theorem 15. *The number of strong k-ary trees with n internal vertices is given by*

$$\sum_{i\geqslant 0}^{i\leqslant \frac{n}{k+1}} (-1)^i \frac{1}{(k-1)n+1}\binom{(k-1)n+1}{i}\binom{kn-i(k+1)}{n-i(k+1)}.$$

Finally, we remark that the computational results in this paper can be also formulated in terms of RNA secondary structures with certain parameterized features (similar to, e.g., hairpins and cloverleaves [10]), which might have some biological significance. For instance, the path distribution represents the distribution of the sizes of parallel base pairs (or arcs) 'induced' by isolated bases.

Acknowledgements

The author would like to thank the anonymous referees for valuable comments and suggestions which improved the presentation of the paper. The author would also like to thank Prof. Michael Waterman for providing comments and suggestions on the earlier manuscript.

A From A-terms to B-terms

Combining A1 and A2, we obtain

$$\frac{1}{p+q-1}\binom{p+q-1}{p}\binom{p+q-1}{p-1} \tag{C1}$$

$$+\binom{q+p-2}{p-1}\sum_{i\geqslant 1}^{i\leqslant \frac{q}{h+1}} (-1)^i\binom{p-1}{i}\binom{p+q-i(h+1)}{p} \tag{C2}$$

$$-\binom{q+p-1}{p-1}\sum_{i\geqslant 1}^{i\leqslant\frac{q-1}{h+1}}(-1)^i\frac{p}{p-i}\binom{p-1}{i}\binom{p+q-i(h+1)}{p}\frac{q-i(h+1)}{p+q-i(h+1)}, \quad \text{(C3)}$$

where C1 is the sum of the terms for $i=0$ in A1 and A2. Combining A3–A5 by writing $h+2=(h+1)+1$ and using $\binom{n}{k}=\binom{n-1}{k}+\binom{n-1}{k-1}$ twice, we obtain

$$\binom{q+p-2}{p-2}\sum_{i\geqslant 1}^{i\leqslant\frac{q-1}{h+1}}(-1)^{i-1}\frac{i(h+1)}{p-i}\binom{p-1}{i}\binom{p+q-i(h+1)}{p}\frac{p}{p+q-i(h+1)} \quad \text{(C4)}$$

$$-\binom{q+p-2}{p-1}\sum_{i\geqslant 1}^{i\leqslant\frac{q}{h+1}}(-1)^{i-1}\frac{i}{p-i}\binom{p-1}{i}\binom{p+q-i(h+1)}{p} \quad \text{(C5)}$$

$$+\text{B2}.$$

Combining C2 and C5, we have

$$\binom{q+p-2}{p-2}\sum_{i\geqslant 1}^{i\leqslant\frac{q}{h+1}}(-1)^{i}\frac{qp[p+q-i(h+1)]}{(p-1)(p-i)[p+q-i(h+1)]}\binom{p-1}{i}\binom{p+q-i(h+1)}{p}. \quad \text{(D1)}$$

Note that C3 can be rewritten as

$$-\binom{q+p-2}{p-2}\sum_{i\geqslant 1}^{i\leqslant\frac{q-1}{h+1}}(-1)^{i}\binom{p-1}{i}\binom{p+q-i(h+1)}{p}\frac{q-i(h+1)}{p+q-i(h+1)}\frac{p}{p-i}\frac{p+q-1}{p-1}. \quad \text{(C3)}$$

Combining C4, D1 (i.e., C2 and C5) and C3 together, C4 will be cancelled, D1 will be cancelled except for the part B3, and the remaining part of C3 is

$$\binom{p+q-2}{p-2}\sum_{i\geqslant 1}^{i\leqslant\frac{q-1}{h+1}}(-1)^{i}\binom{p}{i}\binom{p+q-i(h+1)}{p}\frac{q}{[p+q-i(h+1)](p-1)}. \quad \text{(D3)}$$

It is easy to see that C1 equals the term for $i=0$ of D3. Thus, C1 and D3 give B1. In conclusion, we have arrived at the B-terms from the A-terms.

References

[1] E. A. Bender and L. B. Richmond. A multivariate Lagrange inversion formula for asymptotic calculations. *Electron. J. Combin.*, 5(1):#R33, 1998.

[2] W. Y. C. Chen, E. Deutsch, and S. Elizalde. Old and young leaves on plane trees. *European J. Combin.*, 27:414–427, 2006.

[3] R. X. F. Chen. A refinement of the formula for k-ary trees and the Gould-Vandermonde's convolution. *Electron. J. Combin.*, 15(1):#R52, 2008.

[4] E. Deutsch. A bijection on ordered trees and its consequences. *J. Combin. Theory Ser. A*, 90:210–215, 2000.

[5] I. M. Gessel. A combinatorial proof of the multivariate Lagrange inversion formula. *J. Combin. Theory Ser. A*, 45:178–195, 1987.

[6] W. R. Schmitt and M. S. Waterman. Linear trees and RNA secondary structure. *Discrete Appl. Math.*, 51(3):317–323, 1994.

[7] T. F. Smith and M. S. Waterman. RNA secondary structure. *Math. Biol.*, 42:31–49, 1978.

[8] P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Math.*, 26:261–272, 1979.

[9] M. S. Waterman. Secondary structure of single-stranded nucleic acids. In *Rota G.-C. (ed) Studies on foundations and combinatorics*, vol 1 of *Advances in mathematics supplementary studies*, pages 167–212. Academic Press, N.Y., 1978.

[10] M. S. Waterman. Combinatorics of RNA Hairpins and Cloverleaves. *Stud. Appl. Math.*, 60(2):91–98, 1979.