# Shapes of RNA pseudoknot structures

Christian M. Reidys$^\star$ and Rita R. Wang

Center for Combinatorics, LPMC-TJKLC
Nankai University
Tianjin 300071
P.R. China
Phone: *86-22-2350-6800
Fax: *86-22-2350-9272
duck@santafe.edu

## Abstract

In this paper we study abstract shapes of $k$-noncrossing, $\sigma$-canonical RNA pseudoknot structures. We consider $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes, which represent a generalization of the abstract $\pi'$- and $\pi$-shapes of RNA secondary structures introduced by Giegerich et al. (2004). Using a novel approach we compute the generating functions of $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes as well as the generating functions of all $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes induced by all $k$-noncrossing, $\sigma$-canonical RNA structures for fixed length $n$. By means of singularity analysis of the generating functions, we derive explicit asymptotic expressions.

**Keywords**: $k$-noncrossing RNA structure, $\sigma$-canonical, shape, singularity analysis, generating function

## 1. INTRODUCTION

Pseudoknots have long been known as important structural elements [Westhof and Jaeger 1992], see Fig. 1. They represent cross-serial interactions between RNA nucleotides and are an important functionally in tRNAs, RNaseP [Loria and Pan 1996], telomerase RNA [Staple and Butcher 2005], and ribosomal RNAs [Konings and Gutell 1995]. Pseudoknots in plant virus RNAs mimic tRNA structures, and *in vitro* selection experiments have produced pseudoknotted RNA families that bind to the HIV-1 reverse transcriptase [Tuerk et al. 1992]. Import general mechanism, such as ribosomal frame shifting, are dependent upon pseudoknots [Chamorro et al. 1992].

Despite their biological importance, pseudoknots are typically excluded from large-scale computational studies. Although the problem has attracted considerable attention in the last decade, and several software tools [Huang et al. 2009, Rivas and Eddy 1999] have become available, the required resources have remained prohibitive for applications beyond individual molecules.

An RNA molecule is a sequence of the four nucleotides **A**, **G**, **U** and **C** together with the Watson-Crick (**A-U**, **G-C**) and **U-G** base pairing rules. The sequence of bases is called the primary structure of the RNA molecule. Two bases in the primary structure which are not adjacent may form hydrogen bonds following the Watson-Crick base pairing rules. Three decades ago Waterman *et al.* [Kleitman 1970, Nussinov et al. 1978, Waterman 1978b] analyzed RNA secondary structures. Secondary structures are coarse grained RNA contact structures. They can be represented as diagrams, planar graphs as well as Motzkin-paths, see Fig. 2. Diagrams are labeled graphs over the vertex set $[n] = \{1, \dots, n\}$ with vertex degrees $\leq 1$, represented by drawing its vertices on a horizontal line and its arcs $(i, j)$ $(i < j)$, in the upper half-plane, see Fig. 2 and Fig. 3. Here, vertices and arcs correspond to the nucleotides **A**, **G**, **U** and **C** and Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairs, respectively. In a diagram two arcs $(i_1, j_1)$ and $(i_2, j_2)$ are called crossing if $i_1 < i_2 < j_1 < j_2$ holds. Accordingly, a $k$-crossing is a sequence of arcs $(i_1, j_1), \dots, (i_k, j_k)$ such that $i_1 < i_2 < \cdots < i_k < j_1 < j_2 < \cdots < j_k$, see Fig. 3. We call diagrams containing at most $(k-1)$-crossings, $k$-noncrossing diagrams ($k$-noncrossing partial matchings).

An important observation in this context is that RNA secondary structures have no crossings in their diagram representation, see Fig. 3 (l.h.s.) and Fig. 2, and are therefore 2-noncrossing diagrams. The length of an arc $(i, j)$ is given by $j - i$, characterizing the minimal length of a hairpin loop. A stack of length $\sigma$ is a sequence of “parallel” arcs of the form

$$((i,j),(i+1,j-1),\dots,(i+(\sigma-1),j-(\sigma-1))). \tag{1.1}$$

In the context of minimum-free energy pseudoknot structures [Huang et al. 2009], a minimum stack length $\sigma$ or either two or three is stipulated. We remark that RNA secondary structures are 2-noncrossing, 2-canonical diagrams, whose numbers are asymptotically given by [Hofacker et al. 1998]

$$S_{2,2}(n) \sim c\, n^{-3/2}\, 1.96798^n, \quad c>0. \tag{1.2}$$

We call an arc of length one a 1-arc. A $k$-noncrossing, $\sigma$-canonical RNA structure is a $k$-noncrossing diagram without 1-arcs, having a minimum stack-size of $\sigma$.

The efficient minimum free energy (mfe) folding of secondary structures is a consequence of the following relation of the numbers of RNA secondary structures over $n$ nucleotides, $S_2(n)$, [Waterman 1978b]

$$S_2(n) = S_2(n-1) + \sum_{j=0}^{n-2} S_2(n-2-j)S_2(j), \tag{1.3}$$

where $S_2(n) = 1$ for $0 \le n \le 2$. Accordingly, RNA secondary structures satisfy a constructive recursion. As mentioned above, this relation is the key for deriving the fundamental DP-recursions used for the polynomial time folding of secondary structures [Hofacker 2003, Nussinov et al. 1978] and has therefore profound algorithmic implications. In addition, eq. (1.3) is of central importance for the analysis of abstract shapes [Nebel and Scheid 2009]. In addition, for a given RNA sequence, we have not only one but an ensemble of structures, quantified via the partition function generated by the (Boltzman weighted) probability space of all structures [McCaskill 1990]. In view of the fact that the number of the mfe and suboptimal foldings of an RNA sequence is large, Giegerich *et al.* [Giegerich et al. 2004] introduced the notion of abstract shapes of secondary structures. Two particularly important shape levels are the important level-1 ($\pi'$-) and level-5 ($\pi$-) shapes were studied in [Giegerich et al. 2004]. In [Voß et al. 2006], the authors compute the probability of a shape by means of the partition function, where the probability of a shape is the induced probability of all the structures inducing it.

The problem with pseudoknotted structures is, that they do not satisfy a recursion of the type of eq. (1.3), rendering the *ab initio* folding into mfe configurations [Huang et al. 2009, Lyngso and Pedersen 2000] as well as the derivation of any other properties a nontrivial task. Here, we generalize the $\pi'$- and $\pi$-shapes of [Giegerich et al. 2004], by introducing $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes, see Fig. 4. Our results are not new in case of $k = 2$, since we have $\mathsf{lv}_2^1 = \pi'$ and $\mathsf{lv}_2^5 = \pi$. In two beautiful papers [Lorenz et al. 2008, Nebel and Scheid 2009] $\pi'$- and $\pi$-shapes have been analyzed. The results of [Lorenz et al. 2008, Nebel and Scheid 2009] explicitly make use of the constructive recurrence relation given in eq. (1.3). Their approach can consequently not be generalized to RNA pseudoknot structures, as the latter are genuinely nonrecursive. Our framework therefore identifies the combinatorial “heart” of the results of [Lorenz et al. 2008, Nebel and Scheid 2009] and provides a new approach avoiding any notion of recursiveness. The key idea behind the construction of $\mathsf{lv}_k^5$-shapes is a projection onto so called $k$-noncrossing core-structures [Jin and Reidys 2009].

The paper is organized as follows: after introducing all necessary background we give a detailed computation of the generating functions and study their singularities. We derive simple asymptotic expressions for the numbers of $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes as well as the numbers of theses shapes, induced by $k$-noncrossing, $\sigma$-canonical RNA structures of fixed length $n$. Finally we put our results into context.

## 2. Some basic facts

Let $f_k(n,\ell)$ denote the number of $k$-noncrossing diagrams on $n$ vertices having exactly $\ell$ isolated vertices. A diagram without isolated points is called a matching. The exponential generating function of $k$-noncrossing matchings satisfies the following identity [Chen et al. 2007, Grabiner and Magyar 1993, Jin et al. 2008]

$$\sum_{n\geq 0} f_k(2n,0)\cdot\frac{z^{2n}}{(2n)!} \;=\; \det[I_{i-j}(2z)-I_{i+j}(2z)]|_{i,j=1}^{k-1} \tag{2.1}$$

where $I_r(2z)=\sum_{j\geq 0}\frac{z^{2j+r}}{j!(j+r)!}$ is the hyperbolic Bessel function of the first kind of order $r$. Eq. (2.1) allows to conclude that the ordinary generating function

$$\mathbf{F}_k(z)=\sum_{n\geq 0} f_k(2n,0)z^n$$

is $D$-finite [Stanley 1980], i.e. there exists some $e\in\mathbb{N}$ such that

$$q_{0,k}(z)\frac{d^e}{dz^e}\mathbf{F}_k(z)+q_{1,k}(z)\frac{d^{e-1}}{dz^{e-1}}\mathbf{F}_k(z)+\cdots+q_{e,k}(z)\mathbf{F}_k(z)=0, \tag{2.2}$$

where $q_{j,k}(z)$ are polynomials. Since $I_r(2z)$ is $D$-finite by its definition and $D$-finite power series are algebraic closed [Stanley 1980]. The key point is that any singularity of $\mathbf{F}_k(z)$ is contained in the set of roots of $q_{0,k}(z)$ [Stanley 1980], which we denote by $R_k$. For $2\leq k\leq 7$, we give the polynomials $q_{0,k}(z)$ and their roots in Table 1.

In [Jin et al. 2008] we showed that for arbitrary $k$

$$f_k(2n,0) \sim \widetilde{c}_k\, n^{-((k-1)^2+(k-1)/2)}\,(2(k-1))^{2n}, \qquad \widetilde{c}_k > 0\ . \tag{2.3}$$

in accordance with the fact that $\mathbf{F}_k(z)$ has the unique dominant singularity $\rho_k^2$, where $\rho_k = 1/(2k-2)$.

According to Pringsheim's Theorem [Flajolet and Sedgewick 2009, Titchmarsh 1939], each power series $f(z) = \sum_{n\geq 0} a_n\, z^n$ with nonnegative coefficients and a radius of convergence $R > 0$ has a positive real dominant singularity at $z = R$. This singularity plays a key role for the asymptotics of the coefficients. The class of theorems that deal with such deductions are called transfer-theorems [Flajolet and Sedgewick 2009]. One key ingredient in this framework is a specific domain in which the functions in question are analytic, which is "slightly" bigger than their respective radius of convergence. It is tailored for extracting the coefficients via Cauchy's integral formula. Details on the method can be found in [Flajolet and Sedgewick 2009, Stanley 1980]. In case of $D$-finite functions we have analytic continuation in any simply connected domain containing zero and avoiding the singularities [Wasow 1987] and all prerequisites of singularity analysis are met. We use the notation

$$(f(z) = \Theta\left(g(z)\right) \ \text{as } z \to \rho) \iff (f(z)/g(z) \to c \text{ as } z \to \rho)\,, \tag{2.4}$$

where $c$ is some constant. Let $[z^n]f(z)$ denote the $n$-th coefficient of the power series $f(z)$ at $z = 0$. Since the Taylor coefficients have the property

$$\forall \gamma \in \mathbb{C} \setminus 0; \quad [z^n]f(z) = \gamma^n [z^n] f(\frac{z}{\gamma}), \tag{2.5}$$

we can, without loss of generality, reduce our analysis to the case where $z = 1$ is the unique dominant singularity.

**Theorem 1.** [Flajolet and Sedgewick 2009] *(a) Suppose* $f(z) = (1-z)^{-\alpha}$, $\alpha \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0}$, *then*

$$\begin{aligned} [z^n]\, f(z) \ \sim\ & \frac{n^{\alpha-1}}{\Gamma(\alpha)} \left[ 1 + \frac{\alpha(\alpha-1)}{2n} + \frac{\alpha(\alpha-1)(\alpha-2)(3\alpha-1)}{24n^2} + \right. \\ & \left. \frac{\alpha^2(\alpha-1)^2(\alpha-2)(\alpha-3)}{48n^3} + O\left(\frac{1}{n^4}\right) \right]. \end{aligned} \tag{2.6}$$

*(b) Suppose* $f(z) = (1-z)^r \log(\frac{1}{1-z})$, $r \in \mathbb{Z}_{\geq 0}$, *then we have*

$$[z^n]f(z) \sim (-1)^r \frac{r!}{n(n-1)\dots(n-r)}. \tag{2.7}$$

**Theorem 2.** [Flajolet and Sedgewick 2009] *Let* $f(z)$ *be a D-finite function having a unique dominant singularity* $z = 1$. *Let*

$$g(z) = (1-z)^\alpha \log^\beta \left( \frac{1}{1-z} \right), \quad \alpha, \beta \in \mathbb{R}.$$

*That is we have in the intersection of a neighborhood of* 1

$$f(z) = \Theta(g(z)) \quad \textit{for } z \to 1. \tag{2.8}$$

*Then we have*

$$[z^n]f(z) = \Theta\left([z^n]g(z)\right). \tag{2.9}$$

Theorem 1 and Theorem 2 facilitate to derive the asymptotics of the coefficients of the $D$-finite function $\mathbf{F}_k(z)$ at its unique dominant singularity. More precisely we shall encounter a particular instance of the supercritical paradigm, where we have the following situation: we are given a $D$-finite function, $f(z)$ and an algebraic function $g(u)$ satisfying $g(0) = 0$. Furthermore we suppose that $f(g(u))$ has a unique real valued dominant singularity $\gamma$ and $g$ is regular in a disc with radius slightly larger than $\gamma$. Then the supercritical paradigm stipulates that the subexponential factors of $f(g(u))$ at $u = 0$, given that $g(u)$ satisfies certain conditions, coincide with those of $f(z)$.

**Proposition 1.** *Let $\psi(z)$ be an algebraic, analytic function in a domain $\mathcal{D} = \{z||z| \le r\}$ such that $\psi(0) = 0$. In addition suppose $\gamma$ is the unique dominant singularity of $\mathbf{F}_k(\psi(z))$ and minimum positive real solution of $\psi(\gamma) = \rho_k^2$, $|\gamma| < r$, $\psi'(\gamma) \neq 0$. Then $\mathbf{F}_k(\psi(z))$ has a singular expansion and*

$$[z^n]\mathbf{F}_k(\psi(z)) \sim A\, n^{-((k-1)^2+(k-1)/2)} \left(\frac{1}{\gamma}\right)^n, \tag{2.10}$$

*where $A$ is some constant.*

In the following we will compute the generating functions via the symbolic enumeration method [Flajolet and Sedgewick 2009]. For this purpose we need the notion of a combinatorial class. A combinatorial class $(\mathcal{C}, w_{\mathcal{C}})$ is a set together with a size-function, $w_{\mathcal{C}} \colon \mathcal{C} \longrightarrow \mathbb{Z}^{+}$ such that $\mathcal{C}_n = w_{\mathcal{C}}^{-1}(n)$ is finite for any $n \in \mathbb{Z}^{+}$. We write $w$ instead of $w_{\mathcal{C}}$ and set $C_n = |\mathcal{C}_n|$. Two special combinatorial classes are $\mathcal{E}$ and $\mathcal{Z}$ which contain only one element of size 0 and 1, respectively. The generating function of a combinatorial class $\mathcal{C}$ is given by

$$\mathbf{C}(z) = \sum_{c \in \mathcal{C}} z^{w(c)} = \sum_{n \geq 0} C_n \, z^n, \tag{2.11}$$

where $\mathcal{C}_n \subset \mathcal{C}$. In particular, the generating functions of the classes $\mathcal{E}$ and $\mathcal{Z}$ are $\mathbf{E}(z) = 1$ and $\mathbf{Z}(z) = z$. Suppose $\mathcal{C}, \mathcal{D}$ are combinatorial classes. Then $\mathcal{C}$ is isomorphic to $\mathcal{D}$, $\mathcal{C} \cong \mathcal{D}$, if and only if $\forall\, n \geq 0, |\mathcal{C}_n| = |\mathcal{D}_n|$. In the following we shall identify isomorphic combinatorial classes and write $\mathcal{C} = \mathcal{D}$ if $\mathcal{C} \cong \mathcal{D}$. We set

- $\mathcal{C} + \mathcal{D} := \mathcal{C} \cup \mathcal{D}$, if $\mathcal{C} \cap \mathcal{D} = \varnothing$ and for $\alpha \in \mathcal{C} + \mathcal{D}$,

$$w_{\mathcal{C}+\mathcal{D}}(\alpha) = \begin{cases} w_{\mathcal{C}}(\alpha) & \text{if } \alpha \in \mathcal{C} \\ w_{\mathcal{D}}(\alpha) & \text{if } \alpha \in \mathcal{D}. \end{cases} \tag{2.12}$$

- $\mathcal{C} \times \mathcal{D} := \{\alpha = (c, d) \mid c \in \mathcal{C}, d \in \mathcal{D}\}$ and for $\alpha \in \mathcal{C} \times \mathcal{D}$,

$$w_{\mathcal{C} \times \mathcal{D}}((c, d)) = w_{\mathcal{C}}(c) + w_{\mathcal{D}}(d). \tag{2.13}$$

and furthermore $\mathcal{C}^m := \prod_{h=1}^{m} \mathcal{C}$ and $\textsc{Seq}(\mathcal{C}) := \mathcal{E} + \mathcal{C} + \mathcal{C}^2 + \cdots$. Plainly, $\textsc{Seq}(\mathcal{C})$ is a combinatorial class if and only if there is no element in $\mathcal{C}$ of size 0. We immediately observe

**Proposition 2.** *Suppose $\mathcal{A}$, $\mathcal{C}$ and $\mathcal{D}$ are combinatorial classes with generating functions $\mathbf{A}(\mathbf{z})$, $\mathbf{C}(z)$ and $\mathbf{D}(z)$. Then*

(a) $\mathcal{A} = \mathcal{C} + \mathcal{D} \Longrightarrow \mathbf{A}(z) = \mathbf{C}(z) + \mathbf{D}(z)$

(b) $\mathcal{A} = \mathcal{C} \times \mathcal{D} \Longrightarrow \mathbf{A}(z) = \mathbf{C}(z) \cdot \mathbf{D}(z)$

(c) $\mathcal{A} = \textsc{Seq}(\mathcal{C}) \Longrightarrow \mathbf{A}(z) = \frac{1}{1-\mathbf{C}(z)}$.

## 3. COMBINATORICS OF $\mathsf{lv}_k^5$-SHAPES

Let $\mathcal{G}_k(s,m)$ denote the set of the $k$-noncrossing matchings of length $2s$ with $m$ 1-arcs. In our first lemma, we will compute the bivariate generating function of $g_k(s,m)$, i.e. the number of $k$-noncrossing matchings of length $2s$ with exactly $m$ 1-arcs.

**Lemma 1.** *Suppose $k,s,m \in \mathbb{N}$, $k \geq 2$. Then $g_k(s,m)$ has the following properties*

$$g_k(s,m) \quad = \quad 0 \quad \text{for } m > s, \tag{3.1}$$

$$\sum_{m=0}^{s} g_k(s,m) \quad = \quad f_k(2s,0) \tag{3.2}$$

*and we have the recursion*

$$(m+1)g_k(s+1,m+1) = (m+1)g_k(s,m+1) + (2s+1-m)g_k(s,m). \tag{3.3}$$

*Furthermore, the generating function $\mathbf{G}_k(x,y) = \sum_{s\geq 0}\sum_{m=0}^{s} g_k(s,m)x^s y^m$ is given by*

$$\mathbf{G}_k(x,y) = \frac{1}{x+1-yx}\mathbf{F}_k\left(\frac{x}{(x+1-yx)^2}\right). \tag{3.4}$$

*Proof.* By construction eq. (3.1) and $\sum_{m=0}^{s} g_k(s,m) = f_k(2s,0)$ hold, the latter being equivalent to

$$\mathbf{G}_k(x,1) = \mathbf{F}_k(x). \tag{3.5}$$

Choose a $k$-noncrossing matching $\delta \in \mathcal{G}_k(s+1, m+1)$ and label one 1-arc. We have $(m+1)g_k(s+1, m+1)$ different such labeled $k$-noncrossing matchings. On the other hand, in order to obtain such a labeled matching, we can also insert one labeled 1-arc in a $k$-noncrossing matching $\delta' \in \mathcal{G}_k(s, m+1)$. In this case, we can only put it inside one original 1-arc in $\delta'$ in order to preserve the number of the 1-arcs. We may also insert a labeled 1-arc in a $k$-noncrossing matching $\delta'' \in \mathcal{G}_k(s, m)$. In this case, we can only insert the 1-arc between two vertices not forming a 1-arc, see Fig. 5. Therefore, we arrive at $(m+1)g_k(s, m+1) + (2s+1-m)g_k(s, m)$ different such labeled matchings and

$$(m+1)g_k(s+1, m+1) = (m+1)g_k(s, m+1) + (2s+1-m)g_k(s, m).$$

The above recursion is equivalent to the partial differential equation

$$\frac{\partial \mathbf{G}_k(x,y)}{\partial y} = x\frac{\partial \mathbf{G}_k(x,y)}{\partial y} + 2x^2\frac{\partial \mathbf{G}_k(x,y)}{\partial x} + x\mathbf{G}_k(x,y) - xy\frac{\partial \mathbf{G}_k(x,y)}{\partial y}. \tag{3.6}$$

We next claim

- The function

$$\mathbf{G}_k^*(x,y) = \frac{1}{x+1-yx}\mathbf{F}_k\left(\frac{x}{(x+1-yx)^2}\right)$$

  is a solution of eq. (3.6),
- its coefficients, $g_k^*(s,m) = [x^s y^m]\mathbf{G}_k^*(x,y)$, satisfy $g_k^*(s,m) = 0$ for $m > s$,
- $\mathbf{G}_k^*(x,1) = \mathbf{F}_k(x)$.

Indeed,

$$\frac{\partial \mathbf{G}^*_k(x,y)}{\partial y} = xu\,\mathbf{F}_k\,(xu) + 2xu\,\mathbf{F}'_k\,(xu) \tag{3.7}$$

$$\frac{\partial \mathbf{G}^*_k(x,y)}{\partial x} = (y-1)u\,\mathbf{F}_k\,(xu) + \frac{(1+yx-x)u}{x}\mathbf{F}'_k\,(xu)\,, \tag{3.8}$$

where $u = (x+1-yx)^{-2}$ and $\mathbf{F}'_k\,(xu) = \sum_{n\geq 0} n f_k(2n,0)(xu)^n$. Consequently,

$$(1+xy-x)\frac{\partial \mathbf{G}^*_k(x,y)}{\partial y} = 2x^2\frac{\partial \mathbf{G}^*_k(x,y)}{\partial x} + x\mathbf{G}^*_k(x,y) \tag{3.9}$$

which coincides with eq. (3.6). In order to prove $g^*_k(s,m) = 0$ for $m > s$, we first observe that $\mathbf{G}^*_k(x,y)$ is a power series, since it is analytic in $(0,0)$. Note that the indeterminant $y$ only appears in form of products $xy$, from which the assertion follows. The equality $\mathbf{G}^*_k(x,1) = \mathbf{F}_k(x)$ is obvious. We next claim

$$\mathbf{G}^*_k(x,y) = \mathbf{G}_k(x,y). \tag{3.10}$$

By construction $g^*_k(s,m)$ satisfies

$$\begin{aligned} g^*_k(s,m) &= 0 \quad \text{for } m > s \\ \sum_{m=0}^{s} g^*_k(s,m) &= f_k(2s,0) \\ (m+1)g^*_k(s+1,m+1) &= (m+1)g^*_k(s,m+1) + (2s+1-m)g^*_k(s,m), \end{aligned}$$

Using these properties we can prove by induction over $s$

$$\forall\, s,m \geq 0; \qquad g^*_k(s,m) = g_k(s,m),$$

whence eq. (3.4) and the lemma is proved. □

We now show how to derive the $\mathsf{lv}_k^{\mathsf{5}}$-shape of a given $k$-noncrossing, $\sigma$-canonical RNA structures. This construction is based on the notion of $k$-noncrossing cores [Jin and Reidys 2009]. A $k$-noncrossing core is a $k$-noncrossing RNA structure in which each stack has size exactly one. The cores of a $k$-noncrossing, $\sigma$-canonical RNA structure, $\delta$, denoted by $c(\delta)$ is obtained in two steps: first we map arcs and isolated vertices as follows:

$$(3.11)\qquad \forall \ell \geq \sigma - 1; \quad ((i-\ell, j+\ell), \ldots, (i,j)) \mapsto (i,j) \text{ and } j \mapsto j \text{ if } j \text{ is isolated}$$

and second we relabel the vertices of the resulting diagram from left to right in increasing order, see Fig.6. We are now in position to define $\mathsf{lv}_k^{\mathsf{5}}$-shapes.

**Definition 1.** **($\mathsf{lv}_k^{\mathsf{5}}$-shape)** A $\mathsf{lv}_k^{\mathsf{5}}$-shape is a $k$-noncrossing matching with stacks of size exactly one.

That is, given a $k$-noncrossing, $\sigma$-canonical RNA structure $\delta$, its $\mathsf{lv}_k^{\mathsf{5}}$-shape, $\mathsf{lv}_k^{\mathsf{5}}(\delta)$, is obtained by first removing all isolated vertices and second collapsing all stacks into a single arc, see Fig.7.

By construction, $\mathsf{lv}_k^5$ shapes do not preserve stack-lengths, unpaired regions, and interior loops, i.e. sequences of parallel arcs of the form

$$((i_1, j_1), [i_1 + 1, i_2 - 1], (i_2, j_2), [j_2 + 1, j_1 - 1]),$$

where $(i_2, j_2)$ is an arc nested in $(i_1, j_1)$ and $[i, j]$ is an interval of unpair region.

Let $\mathcal{I}_k(s, m)$ ($i_k(s, m)$) denote the set (number) of the $\mathsf{lv}_k^5$-shapes of length $2s$ with $m$ 1-arcs and

$$\mathbf{I}_k(z, u) = \sum_{s \geq 0} \sum_{m=0}^{s} i_k(s, m) z^s u^m \tag{3.12}$$

be the bivariate generating function. Furthermore, let $i_k(s)$ denote the number of the $\mathsf{lv}_k^5$-shapes of length $2s$ with generating function

$$\mathbf{I}_k(z) = \sum_{s \geq 0} i_k(s) z^s. \tag{3.13}$$

Since any $\mathsf{lv}_k^5$-shape is in particular the core of some $k$-noncrossing matching, Lemma 1 allows us to establish a relation between the bivariate generating function of $i_k(s, m)$ and the generating function of $\mathbf{F}_k(z)$.

**Theorem 3.** *Let $k, s, m$ be natural numbers where $k \geq 2$, then the following assertions hold*

(a) *the generating functions* $\mathbf{I}_k(z,u)$ *and* $\mathbf{I}_k(z)$ *satisfy*

$$\mathbf{I}_k(z,u) = \frac{1+z}{1+2z-zu}\mathbf{F}_k\left(\frac{z(1+z)}{(1+2z-zu)^2}\right) \tag{3.14}$$

$$\mathbf{I}_k(z) = \mathbf{F}_k\left(\frac{z}{1+z}\right). \tag{3.15}$$

(b) *for* $2 \le k \le 7$*, the number of* $\mathsf{lv}_k^5$*-shapes of length* $2s$ *is asymptotically given by*

$$i_k(s) \sim c_k s^{-((k-1)^2+(k-1)/2)} \left(\mu_k^{-1}\right)^s, \tag{3.16}$$

*where* $\mu_k$ *is the unique minimum positive real solution of* $\frac{z}{1+z} = \rho_k^2$ *and* $c_k$ *is some positive constant.*

*Proof.* We first prove (a). For this purpose we define a map between $k$-noncrossing matchings with $m$ 1-arcs and $\mathsf{lv}_k^5$-shapes

$$g\colon\ \mathcal{G}_k(s,m) \to \dot{\bigcup_{0\le b\le s-m}} \left[\mathcal{I}_k(s-b,m) \times \left\{(a_j)_{1\le j\le s-b} \mid \sum_{j=1}^{s-b} a_j = b,\ a_j \ge 0\right\}\right],$$

where $s \ge 1$. Here, for every $\delta \in \mathcal{G}_k(s,m)$, we have $g(\delta) = (c(\delta),(a_j)_{1\le j\le s-b})$, where $c(\delta)$ is the core structure of $\delta$ obtained according to eq. (3.11) and where $(a_j)_{1\le j\le s-b}$ keeps track of the deleted arcs. It is straightforward to check that the map $g$ is well defined, since all the 1-arcs of $c(\delta)$ are just the 1-arcs of $\delta$. By construction, $g$ is a bijection and then we have that

$$\mathbf{G}_k(x,y) = \sum_{s\ge 0}\sum_{m=0}^{s} g_k(s,m)x^s y^m = \sum_{m\ge 0}\sum_{\gamma\in\mathcal{I}_k(m)} \mathbf{G}_\gamma(x,y), \tag{3.17}$$

where $\mathcal{I}_k(m)$ is the set of $\mathsf{lv}_k^{\mathsf{5}}$-shapes having $m$ 1-arcs and $\mathbf{G}_\gamma(x,y)$ is the generating function of all $k$-noncrossing matchings having $m$ 1-arcs that are mapped into the shape $\gamma$. Suppose $\gamma$ has $s$ arcs. We consider the combinatorial classes of arcs $\mathcal{R}$ and 1-arcs $\mathcal{R}^*$ with generating functions $\mathbf{R}(x) = x$ and $\mathbf{R}^*(x,y) = yx$. Then

- each $k$-noncrossing matching having shape $\gamma$ is obtained by inflating $\gamma$-arcs to stacks and the combinatorial class of stacks is given by

  $$\mathcal{R} \times \textsc{Seq}(\mathcal{R})$$

- the inflation of arcs does not affect the number of 1-arcs.

Therefore we derive

$$\mathbf{G}_\gamma(x,y) \quad = \quad \left(\frac{x}{1-x}\right)^s y^m. \tag{3.18}$$

For any $\gamma, \gamma_1 \in \mathcal{I}_k(m)$, having $s$ arcs we have $\mathbf{G}_\gamma(x,y) = \mathbf{G}_{\gamma_1}(x,y)$, whence

$$\mathbf{G}_k(x,y) = \sum_{m\geq 0}\sum_{\gamma\in\mathcal{I}_k(m)} \mathbf{G}_\gamma(x,y) = \sum_{s\geq 0}\sum_{m=0}^{s} i_k(s,m)\left(\frac{x}{1-x}\right)^s y^m. \tag{3.19}$$

According to Lemma 1, we have

$$\mathbf{G}_k(x,y) = \frac{1}{x+1-yx}\mathbf{F}_k\left(\frac{x}{(x+1-yx)^2}\right)$$

and setting $z = \frac{x}{1-x}$ and $u = y$, we arrive substituting for $\mathbf{G}_k(x,y)$ in eq. (3.19) at

$$\mathbf{I}_k(z,u) = \frac{1+z}{1+2z-zu}\mathbf{F}_k\left(\frac{z(1+z)}{(1+2z-zu)^2}\right).$$

In particular, setting $u = 1$, we derive

$$\mathbf{I}_k(z) = \mathbf{F}_k\left(\frac{z}{1+z}\right),$$

whence (a) follows.

Assertion (b) is a direct consequence of the supercritical paradigm, see Proposition 1. As mentioned before, the ordinary generating function $\mathbf{F}_k(z) = \sum_{n\geq 0} f_k(2n,0)z^n$ is $D$-finite [Stanley 1980] and the inner function $\vartheta(z) = \frac{z}{1+z}$ is algebraic, satisfies $\vartheta(0) = 0$ and is analytic for $|z| < 1$. By direct calculation, using the fact that all singularities of $\mathbf{F}_k(z)$ are contained within the set of zeros of $q_{0,k}(z)$, see Tab. 1, we can then verify that $\mathbf{F}_k(\vartheta(z))$ has the unique dominant real singularity $\mu_k < 1$ satisfying $\vartheta(\mu_k) = \rho_k^2$ for $2 \leq k \leq 7$ by Maple, see the Supplemental Materials (SM). In view of $\vartheta'(\mu_k) \neq 0$, Proposition 1 guarantees eq. (3.16)

$$i_k(s) \sim c_k s^{-((k-1)^2+(k-1)/2)}\left(\mu_k^{-1}\right)^s.$$

This proves (b) completing the proof of the theorem. $\square$

We next studying the number of $\mathsf{lv}_k^{\mathsf{5}}$-shapes induced by $k$-noncrossing, $\sigma$-canonical RNA structures of fixed length $n$, $\mathsf{lv}_{k,\sigma}^{\mathsf{5}}(n)$, setting

$$\mathbf{Lv}_{k,\sigma}^{\mathsf{5}}(x) = \sum_{n\geq 0} \mathsf{lv}_{k,\sigma}^{\mathsf{5}}(n)x^n. \tag{3.20}$$

**Theorem 4.** *Let $k, \sigma \in \mathbb{N}$, where $k \geq 2$. Then the following assertions hold*

(a) *the generating function $\mathbf{Lv}_{k,\sigma}^{\mathsf{5}}(x)$ is given by*

$$\mathbf{Lv}_{k,\sigma}^{\mathsf{5}}(x) = \frac{(1+x^{2\sigma})}{(1-x)(1+2x^{2\sigma}-x^{2\sigma+1})}\mathbf{F}_k\left(\frac{x^{2\sigma}(1+x^{2\sigma})}{(1+2x^{2\sigma}-x^{2\sigma+1})^2}\right). \tag{3.21}$$

(b) *for $2 \leq k \leq 7$ and $1 \leq \sigma \leq 10$*

$$\mathsf{lv}_{k,\sigma}^{\mathsf{5}}(n) \sim c_{k,\sigma} n^{-((k-1)^2+(k-1)/2)} \left(\zeta_{k,\sigma}^{-1}\right)^n, \tag{3.22}$$

*where $c_{k,\sigma} > 0$ and $\zeta_{k,\sigma}$ is the unique minimum positive real solution of*

$$\frac{x^{2\sigma}(1+x^{2\sigma})}{(1+2x^{2\sigma}-x^{2\sigma+1})^2} = \rho_k^2. \tag{3.23}$$

*Proof.* In order to proof of (a) we observe that we can always inflate a structure by adding arcs to stacks or isolated vertices without changing its $\mathsf{lv}_k^{\mathsf{5}}$-shape. In fact, for any given $\mathsf{lv}_k^{\mathsf{5}}$-shape, $\beta$, adding the minimal number of arcs to each stack such that every stack has $\sigma$ arcs, and inserting one isolated vertex in any 1-arc, we derive a $k$-noncrossing, $\sigma$-canonical RNA structure having arc-length$\geq 2$, of minimal length. If a $k$-noncrosing, $\sigma$-canonical RNA structure of length $n$ with a shape in $\mathcal{I}_k(s,m)$, then $n \geq 2\sigma s + m$. We can therefore derive $\mathbf{Lv}_{k,\sigma}^{\mathsf{5}}(x)$, see eq.(3.20), from the bivariate generating function $\mathbf{I}_k(z,u)$ as follows

$$\mathbf{Lv}^{\mathsf{5}}_{k,\sigma}(x) = \sum_{n\geq 0}\sum_{s=0}^{\lfloor \frac{n}{2\sigma}\rfloor}\sum_{m=0}^{\min\{s,n-2\sigma s\}} i_k(s,m)x^n = \sum_{s\geq 0}\sum_{m=0}^{s}\sum_{n\geq 2\sigma s+m} i_k(s,m)x^n,$$

whence

$$\mathbf{Lv}^{\mathsf{5}}_{k,\sigma}(x) = \frac{1}{1-x}\sum_{s\geq 0}\sum_{m=0}^{s} i_k(s,m)x^{2\sigma s+m}$$

and in view of eq. (3.14), $\mathbf{I}_k(z,u) = \frac{1+z}{1+2z-zu}\mathbf{F}_k\left(\frac{z(1+z)}{(1+2z-zu)^2}\right)$, we derive

$$\mathbf{Lv}^{\mathsf{5}}_{k,\sigma}(x) = \frac{(1+x^{2\sigma})}{(1-x)(1+2x^{2\sigma}-x^{2\sigma+1})}\mathbf{F}_k\left(\frac{x^{2\sigma}(1+x^{2\sigma})}{(1+2x^{2\sigma}-x^{2\sigma+1})^2}\right),$$

and (a) follows. As for assertion (b), we observe that

$$\varphi_\sigma(x) = \frac{x^{2\sigma}(1+x^{2\sigma})}{(1+2x^{2\sigma}-x^{2\sigma+1})^2}$$

is algebraic and $\varphi_\sigma(0) = 0$. We verify that $\varphi_\sigma(x)$ is for $1 \leq \sigma \leq 10$ analytic when $|x| < r_\sigma$, where $r_\sigma < 1$ and furthermore

$$\phi_\sigma(x) = \frac{(1+x^{2\sigma})}{(1-x)(1+2x^{2\sigma}-x^{2\sigma+1})}$$

is analytic for $|x| < r_\sigma$. We distinguish the cases $k > 2$ and $k = 2$.

For $2 < k \leq 7$ and $1 \leq \sigma \leq 10$, the minimum positive real solution of eq. (3.23), $\zeta_{k,\sigma}$, is the unique dominant singularity of $\mathbf{Lv}^{\mathsf{5}}_{k,\sigma}(x)$, $|\zeta_{k,\sigma}| < r_\sigma$ and $\varphi'_\sigma(\zeta_{k,\sigma}) \neq 0$. Therefore, Proposition 1 implies

$$\mathsf{lv}^{\mathsf{5}}_{k,\sigma}(n) \sim c_{k,\sigma} n^{-((k-1)^2+(k-1)/2)} \left(\zeta_{k,\sigma}^{-1}\right)^n,$$

where $c_{k,\sigma}$ is some positive constant. In case of $k=2$, we have

$$\mathbf{F}_2(z) = \sum_{n\geq 0} f_2(2n,0)z^n = \frac{2}{1+\sqrt{1-4\,z}}. \tag{3.24}$$

Substituting $\varphi_\sigma(x)$ into the eq. (3.24), we observe that the poles of $\varphi_\sigma(x)$ are not singularities of $\mathbf{Lv}^{\mathsf{5}}_{2,\sigma}(x)$ and the dominant singularity of $\mathbf{Lv}^{\mathsf{5}}_{2,\sigma}(x)$ is the minimum positive solution of $\varphi_\sigma(x) = \rho_2^2$. Employing Theorem 1 and Theorem 2, we derive eq. (3.22) and the proof of the theorem is complete. $\square$

## 4. COMBINATORICS OF $\mathsf{lv}_k^1$-SHAPES

**Definition 2.** **($\mathsf{lv}_k^1$-shape)** A $\mathsf{lv}_k^1$-shape is a $k$-noncrossing structure in which each stack and each segment of isolated vertices have length exactly one.

That is, given a $k$-noncrossing, $\sigma$-canonical RNA structure its $\mathsf{lv}_k^1$-shape is derived as follows: first we apply the core map, second we replace a segment of isolated vertices by a single isolated vertex and third relabel the vertices of the resulting diagram, see Fig.8. $\mathsf{lv}_k^1$-shapes do not preserve stack-lengths and project intervals of isolated vertices into singletons. Let $\mathfrak{J}_k$ and $\mathfrak{I}_k$ denote the set of $\mathsf{lv}_k^1$-shapes and $\mathsf{lv}_k^5$-shapes, respectively. There is a map between $\mathsf{lv}_k^1$-shapes and $\mathsf{lv}_k^5$-shapes

$$\phi: \quad \mathfrak{J}_k \to \mathfrak{I}_k, \tag{4.1}$$

obtained by removing all isolated vertices from $\mathsf{lv}_k^1$-shapes and then collapsing each stack into a single arc, see Fig. 8. By construction, $\phi$ is surjective (for any $\mathsf{lv}_k^5$-shape, we can, inserting one isolated vertex in any 1-arc, obtain a $\mathsf{lv}_k^1$-shape).

Let $\mathfrak{J}_k(n,h)$ ($j_k(n,h)$) denote the set (number) of $\mathsf{lv}_k^1$-shapes of length $n$ having $h$-arcs, and let $j_k(n)$ be the number of all $\mathsf{lv}_k^1$-shapes of length $n$ and

$$\mathbf{J}_k(z,u) = \sum_{h\geq 0}\sum_{n=2h}^{4h+1} j_k(n,h) z^n u^h \quad \text{and} \quad \mathbf{J}_k(z) = \sum_{n\geq 0} j_k(n) z^n. \tag{4.2}$$

**Theorem 5.** *For $k, n, h \in \mathbb{N}$, $k \geq 2$, the following assertions hold*

(a) *the generating functions $\mathbf{J}_k(z,u)$ and $\mathbf{J}_k(z)$ are given by*

$$\mathbf{J}_k(z,u) = \frac{(1+z)(1+uz^2)}{uz^3+2uz^2+1}\mathbf{F}_k\left(\frac{(1+z)^2(1+uz^2)uz^2}{(uz^3+2uz^2+1)^2}\right) \tag{4.3}$$

$$\mathbf{J}_k(z) = \frac{(1+z)(1+z^2)}{z^3+2z^2+1}\mathbf{F}_k\left(\frac{(1+z)^2(1+z^2)z^2}{(z^3+2z^2+1)^2}\right). \tag{4.4}$$

(b) *for $2 \leq k \leq 7$, the number of $\mathsf{lv}_k^1$-shapes of length $n$ satisfies*

$$j_k(n) \sim c_k' n^{-((k-1)^2+(k-1)/2)} \left(\mu_k'^{-1}\right)^n, \tag{4.5}$$

*where $c_k' > 0$ and $\mu_k'$ is the unique minimum positive real solution of*

$$\frac{(1+z)^2(1+z^2)z^2}{(z^3+2z^2+1)^2} = \rho_k^2. \tag{4.6}$$

*Proof.* We prove (a) via symbolic enumeration noticing that we can represent an $\mathsf{lv}_k^1$-shape as an inflation of an $\mathsf{lv}_k^5$-shape. By definition of the map $\phi$, the generating function $\mathbf{J}_k(z,u)$ can be rewritten as

$$\mathbf{J}_k(z,u) = \sum_{m\geq 0}\sum_{\zeta\in\mathcal{I}_k(m)} \mathbf{S}_\zeta(z,u), \tag{4.7}$$

where $\mathbf{S}_\zeta(z,u)$ is the generating function of all $\mathsf{lv}_k^1$-shapes whose images of $\phi$ are $\zeta$ and $\mathcal{I}_k(m)$ denotes the set of $\mathsf{lv}_k^5$-shapes with $m$ 1-arcs. For any $\zeta \in \mathcal{I}_k(s,m)$, we can inflate it to a $\mathsf{lv}_k^1$-shapes by the following steps, see Fig. 9

- we inflate each arc of the shape to a stem of stacks of size 1. Between stacks we can insert isolated vertices to the left or the right, or on both sides in order to separate the stacks and for each such insertion exactly one isolated vertex is used.
- we insert at most one isolated vertex in any of the remaining $(2s+1)$ positions. Furthermore one isolated vertex is inserted into each 1-arc.

We call the newly added stacks of size 1 induced. We introduce the combinatorial classes $\mathcal{E}$ (neutral class which include only one element of size 0), $\mathcal{Z}$ (vertex), $\mathcal{R}$ (arc), $\mathcal{N}$ (induced stacks), $\mathcal{M}$ (stems), and $\mathcal{S}_\zeta$ ($\mathsf{lv}_k^1$-shapes whose images are $\zeta$ under $\phi$), where $\mathbf{E}(z)=1$, $\mathbf{Z}(z)=z$, $\mathbf{R}(z,u)=uz^2$.

$$\mathcal{S}_\zeta = (\mathcal{M})^s \times (\mathcal{E}+\mathcal{Z})^{2s+1-m} \times (\mathcal{Z})^m\,, \tag{4.8}$$

$$\mathcal{M} = \mathcal{R} \times \textsc{Seq}\,(\mathcal{N})\,, \tag{4.9}$$

$$\mathcal{N} = \mathcal{R} \times \left(\mathcal{Z}+\mathcal{Z}+\mathcal{Z}^2\right)\,, \tag{4.10}$$

Thus the generating function $\mathbf{S}_\zeta(z,u)$ is given by

$$\mathbf{S}_\zeta(z,u) = \left(\frac{uz^2}{1-uz^2(2z+z^2)}\right)^s (1+z)^{2s+1-m} z^m, \tag{4.11}$$

By construction, for any two shapes $\zeta_1, \zeta_2 \in \mathcal{I}(s,m)$ we have $\mathbf{S}_{\zeta_1}(z,u) = \mathbf{S}_{\zeta_2}(z,u)$, whence

$$\mathbf{J}_k(z,u) = \sum_{s\geq 0} \sum_{m=0}^{s} i_k(s,m)\mathbf{S}_{\zeta}(z,u). \tag{4.12}$$

According to Theorem 3 we have

$$\sum_{s\geq 0} \sum_{m=0}^{s} i_k(s,m)\, x^s\, y^m = \frac{1+x}{1+2x-xy}\mathbf{F}_k\left(\frac{x(1+x)}{(1+2x-xy)^2}\right). \tag{4.13}$$

Therefore substituting

$$x = \frac{uz^2(1+z)^2}{1-uz^2(2z+z^2)} \quad \text{and} \quad y = \frac{z}{1+z}$$

into eq. (4.13) we derive eq. (4.3). In particular, setting $u = 1$, we derive $\mathbf{J}_k(z)$, whence assertion (a).

Assertion (b) follows in complete analogy to the proof of Theorem 4. First we note that the factor

$$\tau(z) = \frac{(1+z)^2(1+z^2)z^2}{(z^3+2z^2+1)^2} \tag{4.14}$$

is algebraic and $\tau(0) = 0$. We next verify that $\tau(z)$ is analytic for $|z| < r'$, where $r' < 1$ and the unique dominant singularity of $\mathbf{J}_k(z)$ is the minimum positive real solution $\mu'_k$ of

$$\frac{(1+z)^2(1+z^2)z^2}{(z^3+2z^2+1)^2} = \rho_k^2$$

for $2 \leq k \leq 7$, $|\mu'_k| < r'$ and $\tau'(\mu'_k) \neq 0$, see the Supplemental Material. Now (b) follows from Proposition 1. $\square$

We finally compute the number of $\mathsf{lv}_k^1$-shapes induced by $k$-noncrossing, $\sigma$-canonical RNA structures of fixed length $n$, $\mathsf{lv}_{k,\sigma}^1(n)$. Let

$$\mathbf{Lv}_{k,\sigma}^1(x) = \sum_{n\geq 0} \mathsf{lv}_{k,\sigma}^1(n)x^n. \tag{4.15}$$

**Theorem 6.** *Let $k, \sigma \in \mathbb{N}$, where $k \geq 2$. Then the following assertions hold*
(a) *the generating function $\mathbf{Lv}_{k,\sigma}^1(x)$ is given by*

$$\mathbf{Lv}_{k,\sigma}^1(x) = \frac{(1+x)(1+x^{2\sigma})}{(1-x)(x^{2\sigma+1}+2x^{2\sigma}+1)}\mathbf{F}_k\left(\frac{(1+x)^2x^{2\sigma}(1+x^{2\sigma})}{(x^{2\sigma+1}+2x^{2\sigma}+1)^2}\right). \tag{4.16}$$

(b) *for $2 \leq k \leq 7$ and $1 \leq \sigma \leq 10$, we have*

$$\mathsf{lv}_{k,\sigma}^1(n) \sim c'_{k,\sigma} n^{-((k-1)^2+(k-1)/2)} \left(\chi_{k,\sigma}^{-1}\right)^n, \tag{4.17}$$

*where $c'_{k,\sigma} > 0$ and $\chi_{k,\sigma}$ is the unique minimum positive real solution of*

$$\frac{(1+x)^2x^{2\sigma}(1+x^{2\sigma})}{(x^{2\sigma+1}+2x^{2\sigma}+1)^2} = \rho_k^2. \tag{4.18}$$

*Proof.* Obviously, we can inflate any structure by adding arcs into its stacks or duplicating isolated vertices without changing its $\mathsf{lv}_k^1$-shape. As a result, we can derive from any $\mathsf{lv}_k^1$-shape by inflating its stacks to $\sigma$ arcs, a unique, minimal, $k$-noncrossing, $\sigma$-canonical RNA structure inducing it. If a $k$-noncrossing, $\sigma$-canonical

RNA structure of length $n$ has a shape in $\mathfrak{J}_k(s,h)$, then we have $n \geq 2h(\sigma-1)+s$. We can use this observation to conclude

$$\mathsf{lv}^{1}_{k,\sigma}(n) = \sum_{h=0}^{\lfloor \frac{n}{2\sigma} \rfloor} \sum_{s=2h}^{\min\{4h+1,n-2(\sigma-1)h\}} j_k(s,h).$$

Accordingly, we can rewrite the generating function

$$\mathbf{Lv}^{1}_{k,\sigma}(x) = \sum_{h\geq 0} \sum_{s=2h}^{4h+1} \sum_{n\geq 2h(\sigma-1)+s} j_k(s,h)x^n = \frac{1}{1-x}\sum_{h\geq 0}\sum_{s=2h}^{4h+1} j_k(s,h)x^{2h(\sigma-1)+s}.$$

Employing eq. (4.3), we derive

$$\mathbf{Lv}^{1}_{k,\sigma}(x) = \frac{(1+x)(1+x^{2\sigma})}{(1-x)(x^{2\sigma+1}+2x^{2\sigma}+1)}\mathbf{F}_k\left(\frac{(1+x)^2x^{2\sigma}(1+x^{2\sigma})}{(x^{2\sigma+1}+2x^{2\sigma}+1)^2}\right)$$

and assertion (a) follows. As for assertion (b), we proceed in analogy to the proof of Theorem 4. We verify that for $2 \leq k \leq 7$ and $1 \leq \sigma \leq 10$, the unique minimum positive real solution, $\chi_{k,\sigma}$, of eq. (4.18) is the unique dominant singularity of generating function $\mathbf{Lv}^{1}_{k,\sigma}(x)$ and that the derivative of

$$\frac{(1+x)^2x^{2\sigma}(1+x^{2\sigma})}{(x^{2\sigma+1}+2x^{2\sigma}+1)^2}$$

is nonzero at $z = \chi_{k,\sigma}$. Consequently, Proposition 1 implies that

$$\mathsf{lv}^{1}_{k,\sigma}(n) \sim c'_{k,\sigma} n^{-((k-1)^2+(k-1)/2)} \left(\chi_{k,\sigma}^{-1}\right)^n,$$

where $c'_{k,\sigma}$ is some positive constant, whence (b) and the theorem is proved. □

## 5. Conclusion

$\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes of $k$-noncrossing, $\sigma$-canonical RNA pseudoknot structures provide a significant simplification of complicated molecular configurations with cross-serial interactions. The asymptotic formulas presented in Theorem 4 and Theorem 6

$$\begin{aligned} \mathsf{lv}_{k,\sigma}^5(n) &\sim c_{k,\sigma} n^{-((k-1)^2+(k-1)/2)} \left(\zeta_{k,\sigma}^{-1}\right)^n \\ \mathsf{lv}_{k,\sigma}^1(n) &\sim c'_{k,\sigma} n^{-((k-1)^2+(k-1)/2)} \left(\chi_{k,\sigma}^{-1}\right)^n, \end{aligned}$$

imply all asymptotic results on abstract shapes of secondary structures in the literature (note that for $k = 2$, we have $n^{-((k-1)^2+(k-1)/2)} = n^{-3/2}$).

The growth rates of $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes of $k$-noncrossing, $\sigma$-canonical RNA structures, are displayed in Tab. 4 and Tab. 5, where they are contrasted with the exponential growth rates of $k$-noncrossing, $\sigma$-canonical RNA structures, $\gamma_{k,\sigma}$.

Table 5 shows that the exponential growth rate of $\mathsf{lv}_k^5$-shapes of $k$-noncrossing 3-canonical structures are significantly smaller than that of all $k$-noncrossing 3-canonical structures. Therefore, the abstract $\mathsf{lv}_k^5$-shapes represent a meaningful reduction. At `http://www.combinatorics.cn/cbpc/paper.html`, we provide supplemental material for our results.

**Acknowledgments.** This work was supported by the 973 Project, the PCSIRT Project of the Ministry of Education, the Ministry of Science and Technology, and the National Science Foundation of China.

## REFERENCES

[Chamorro et al. 1992] Chamorro, M., Parkin, N., and Varmus, H.E. 1992. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc. Natl. Acad. Sci. USA* 89, 713–717.

[Chen et al. 2007] Chen, W.Y.C., Deng, E.Y.P., Du, R.R.X., Stanley, R.P., Yan, C.H. 2007. Crossings and nestings of matchings and partitions. *Trans. Amer. Math. Soc.* 359, 1555–1575.

[Flajolet and Sedgewick 2009] Flajolet, P. and Sedgewick, R. 2009. *Analytic combinatorics.* Cambridge University Press, New York.

[Giegerich et al. 2004] Giegerich, R., Voß, B. and Rehmsmeier, M. 2004. Abstract shapes of RNA. *Nucleic Acids Res.* 32, 4843–4851.

[Grabiner and Magyar 1993] Grabiner, D.J. and Magyar, P. 1993. Random walks in Weyl chambers and the decomposition of tensor powers. *J. Algebr. Comb.* 2, 239–260.

[Hofacker et al. 1998] Hofacker, I.L., Schuster, P. and Stadler, P.F. 1998. Combinatorics of RNA secondary structures. *Discr. Appl. Math.* 88, 207–237.

[Hofacker 2003] Hofacker, I.L. 2003. Vienna RNA secondary structure server. *Nucl. Acids. Res.* 31, 3429–3431.

[Huang et al. 2009] Huang, F.W.D., Peng, W.W.J. and Reidys, C.M. 2009. Folding 3-noncrossing RNA pseudoknot structures. *J. Comput. Biol.* 16, 1549-1575.

[Jin et al. 2008] Jin, E.Y., Qin, J. and Reidys, C.M. 2008. Combinatorics of RNA structures with pseudoknots. *Bull. Math. Biol.* 70, 45–67.

[Jin and Reidys 2008] Jin, E.Y. and Reidys, C.M. 2008. Asymptotic enumeration of RNA structures with pseudoknots. *Bull. Math. Biol.* 70, 951–970.

[Jin and Reidys 2009] Jin, E.Y. and Reidys, C.M. 2009. Combinatorial design of pseudoknot RNA. *Adv. Appl. Math.* 42, 135–151.

[Jin et al. 2008] Jin, E.Y., Reidys, C.M. and Wang, R.R. 2008. Asympotic analysis of $k$-noncrossing matchings. *arXiv:0803.0848*.

[Kleitman 1970] Kleitman, D.J. 1970. Proportions of irreducible diagrams. *Stud. Appl. Math.* 49, 297–299.

[Konings and Gutell 1995] Konings, D.A.M. and Gutell, R.R. 1995. A compariosn of thermodynamic folidngs with comparatively derived structures of 16S and 16S-like rRNAs. *RNA* 1, 559–574.

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

[Lorenz et al. 2008] Lorenz, W.A., Ponty, Y. and Clote, P. 2008. Asymptotics of RNA shapes. *J. Comput. Biol.* 15, 31–63.

[Lyngso and Pedersen 2000] Lyngso, R.B. and Pedersen, C.N.S. 2000. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* 7, 409–427.

[Ma and Reidys 2008] Ma, G. and Reidys, C.M. 2008. Canonical RNA pseudoknot structures. *J. Comput. Biol.* 15, 1257–1273.

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

[Mapping 2005] Mapping RNA form and function, 2005, *Sicence* 309, 1441–1632.

[Nebel and Scheid 2009] Nebel, M.E. and Scheid, A. 2009. On quantitative effects of RNA shape abstraction. *Theory in Biosciences* 128, 211–225.

[Nussinov et al. 1978] Nussinov, R., Pieczenik, G., Griggs, J.R., and Kleitman, D.J. 1978. Algorithms for loop matchings. *SIAM J. Appl. Math.* 35, 68–82.

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

[Stanley 1980] Stanley, R. 1980. Differentiably finite power series. *Europ. J. Combinatorics* 1, 175–188.

[Staple and Butcher 2005] Staple, D.W. and Butcher, S.E. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* 3, 956–959.

[Titchmarsh 1939] Titchmarsh, E.C. 1939. *The theory of functions*, Oxford Uninversity Press, Oxford, UK.

[Tuerk et al. 1992] Tuerk, C., MacDougal, S. and Gold, L. 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl. Acad. Sci. USA* 89, 6988–6992.

[Voß et al. 2006] Voß, B., Giegerich, R. and Rehmsmeier, M. 2006. Complete probabilistic analysis of RNA shapes. *BMC Biology* 4, 1–23.

[Wasow 1987] Wasow, W. 1987. *Asymptotic expansions for ordinary differential equations.* Dover, New York.

[Waterman 1978a] Waterman, M.S. 1978. Secondary structure of single-stranded nucleic acids. *Adv. Math.I (suppl.)* 1, 167–212.

[Waterman 1978b] Waterman, M.S. 1978. Combinatorics of RNA hairpins and cloverleafs. *Stud. Appl. Math.* 60, 91–96.

[Schmitt and Waterman 1994] Schmitt, W.R. and Waterman, M.S. 1994. Linear trees and RNA secondary structure. *Discr. Appl. Math.* 51, 317–323.

[Westhof and Jaeger 1992] Westhof, E. and Jaeger, L. 1992. RNA pseudoknots. *Curr. Opin. Struct. Biol.* 2, 327–333.

| $k$ | $q_{0,k}(z)$ | $R_k$ |
|---|---|---|
| 2 | $(4z-1)z$ | $\{\frac{1}{4}\}$ |
| 3 | $(16z-1)z^2$ | $\{\frac{1}{16}\}$ |
| 4 | $(144z^2-40z+1)z^3$ | $\{\frac{1}{4},\frac{1}{36}\}$ |
| 5 | $(1024z^2-80z+1)z^4$ | $\{\frac{1}{16},\frac{1}{64}\}$ |
| 6 | $(14400z^3-4144z^2+140z-1)z^5$ | $\{\frac{1}{4},\frac{1}{36},\frac{1}{100}\}$ |
| 7 | $(147456z^3-12544z^2+224z-1)z^6$ | $\{\frac{1}{16},\frac{1}{64},\frac{1}{144}\}$ |

Table 1. We present the polynomials $q_{0,k}(z)$ and their nonzero roots obtained by the MAPLE package `GFUN`.

| $\sigma/k$ | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 1 | 1.51243 | 3.67528 | 5.77291 | 7.82581 | 9.85873 | 11.88118 |
| 2 | 1.26585 | 1.93496 | 2.41152 | 2.80275 | 3.14338 | 3.44943 |
| 3 | 1.17928 | 1.55752 | 1.80082 | 1.98945 | 2.14693 | 2.28376 |

Table 2. The exponential growth rates $\zeta_{k,\sigma}^{-1}$ of $\mathsf{lv}_k^{\mathsf{5}}$-shapes induced by $k$-noncrossing, $\sigma$-canonical RNA structures of length $n$.

| $\sigma/k$ | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| 1 | 2.09188 | 4.51263 | 6.65586 | 8.73227 | 10.7804 | 12.8137 |
| 2 | 1.56947 | 2.31767 | 2.81092 | 3.21184 | 3.55939 | 3.87079 |
| 3 | 1.38475 | 1.80408 | 2.05600 | 2.24968 | 2.41081 | 2.55050 |

TABLE 3. The exponential growth rates $\chi_{k,\sigma}^{-1}$ of $\mathsf{lv}_k^1$-shapes induced by $k$-noncrossing, $\sigma$-canonical RNA structures of length $n$.

| $k$ | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| $\gamma_{k,2}^{-1}$ | 1.96798 | 2.58808 | 3.03825 | 3.41383 | 3.74381 | 4.04195 |
| $\chi_{k,2}^{-1}$ | 1.56947 | 2.31767 | 2.81092 | 3.21184 | 3.55939 | 3.87079 |
| $\zeta_{k,2}^{-1}$ | 1.26585 | 1.93496 | 2.41152 | 2.80275 | 3.14338 | 3.44943 |

TABLE 4. The exponential growth rates of arbitrary $k$-noncrossing, 2-canonical RNA structures of length $n$ and the numbers of their induced $\mathsf{lv}_k^1$ and $\mathsf{lv}_k^5$ shapes.

| $k$ | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|
| $\gamma_{k,3}^{-1}$ | 1.71599 | 2.04771 | 2.27036 | 2.44664 | 2.59554 | 2.72590 |
| $\chi_{k,3}^{-1}$ | 1.38475 | 1.80408 | 2.05600 | 2.24968 | 2.41081 | 2.55050 |
| $\zeta_{k,3}^{-1}$ | 1.17928 | 1.55752 | 1.80082 | 1.98945 | 2.14693 | 2.28376 |

TABLE 5. The exponential growth rates of arbitrary $k$-noncrossing, 3-canonical RNA structures of length $n$ and the numbers of their induced $\mathsf{lv}_k^1$ and $\mathsf{lv}_k^5$ shapes.

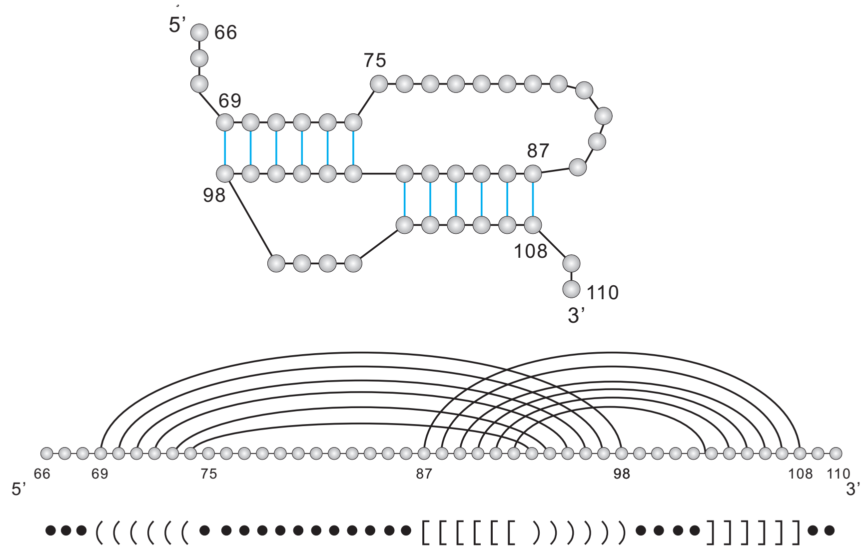

Figure 1. The pseudoknot structure of the PrP-encoding mRNA.

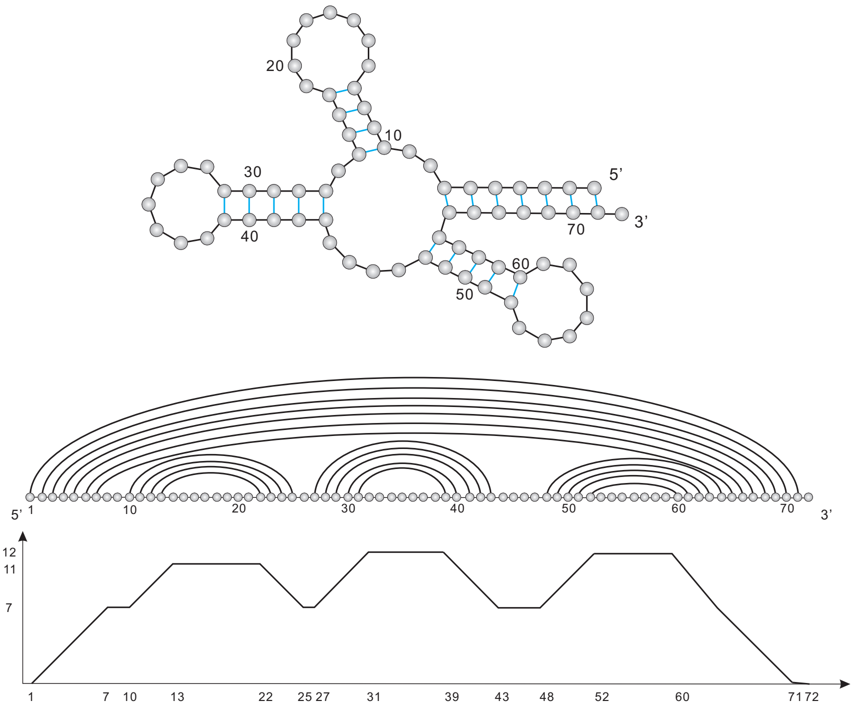

Figure 2. The Sprinzl tRNA RD7550 secondary structure represented as a planar graph (top), 2-noncrossing diagram (middle) and Motzkin-path (bottom), where up/down/horizontal-steps correspond to start/end/unpaired vertices, respectively.

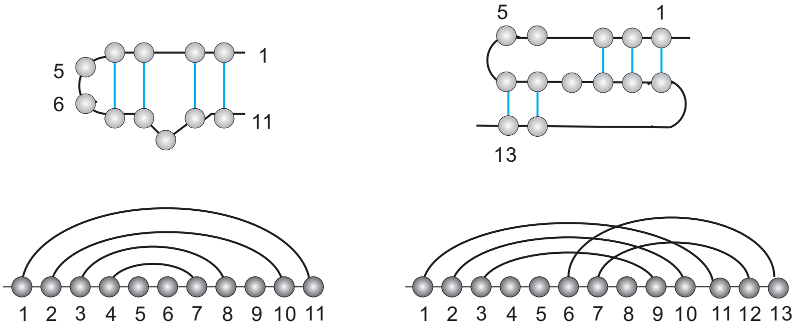

Figure 3. A 2-noncrossing, 2-canonical RNA structure (left) and a 3-noncrossing, 2-canonical RNA structure (right).

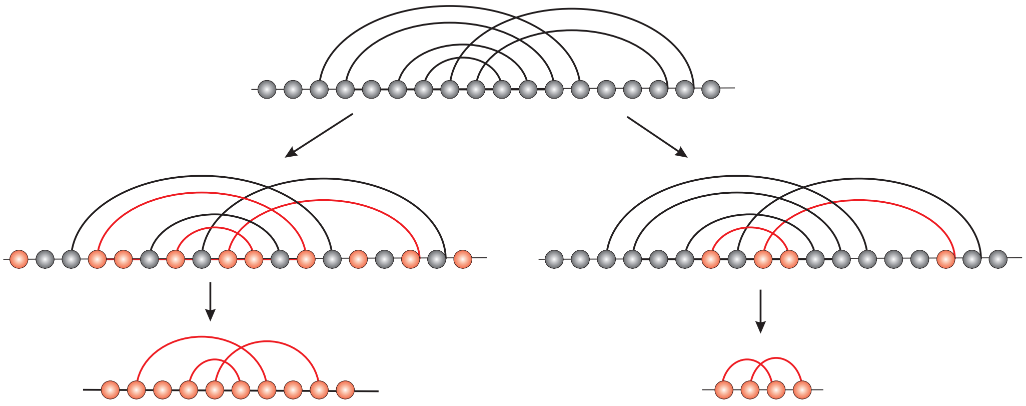

Figure 4. $\mathsf{lv}_k^1$- and $\mathsf{lv}_k^5$-shapes: a 3-noncrossing, 2-canonical RNA structure (top), its $\mathsf{lv}_3^1$-shape (bottom left) and its $\mathsf{lv}_3^5$-shape (bottom right).

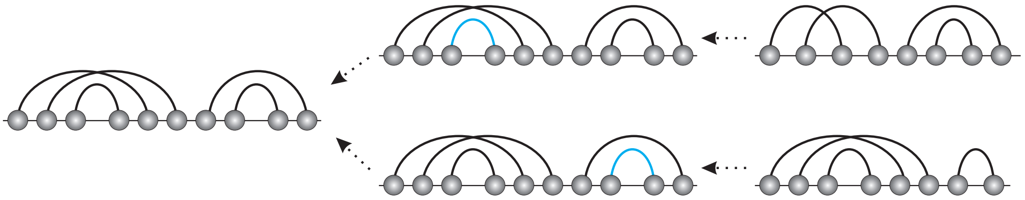

Figure 5. Labeling the 1-arcs allows to trace how certain arc configurations arise.

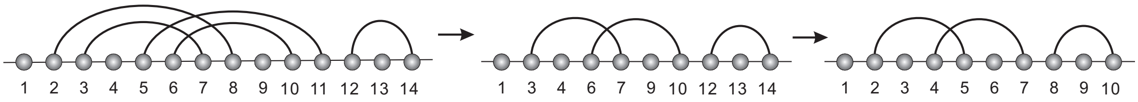

Figure 6. A 3-noncrossing core structure is obtained from a 3-noncrossing, 1-canonical RNA structure in two steps.

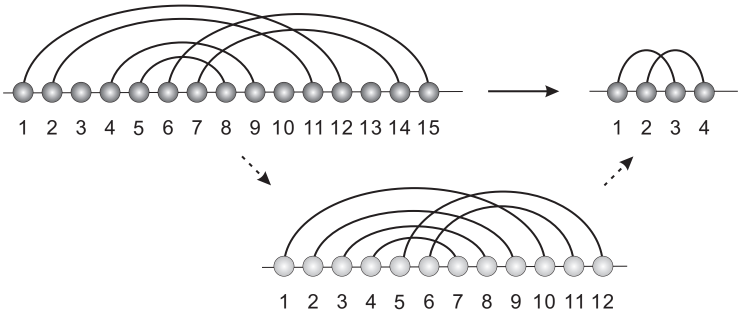

Figure 7. Generation of the $\mathsf{lv}_3^5$-shape. A 3-noncrossing, 2-canonical RNA structure (top-left) is mapped in two steps into its $\mathsf{lv}_3^5$-shape (top-right).

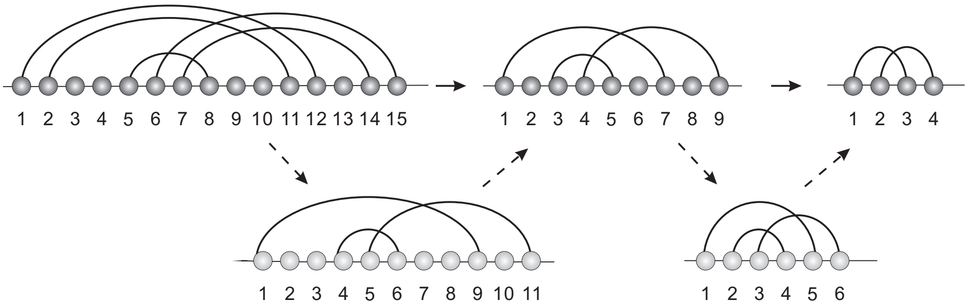

Figure 8. $\mathsf{lv}_k^1$-shapes via the core map and subsequent identification of unpaired nucleotides: A 3-noncrossing, 1-canonical RNA structure (top-left) is mapped into its $\mathsf{lv}_3^1$-shape (top-middle). The $\mathsf{lv}_3^1$-shape is projected into its $\mathsf{lv}_3^5$-shape (top-right) via $\phi$ in two steps.

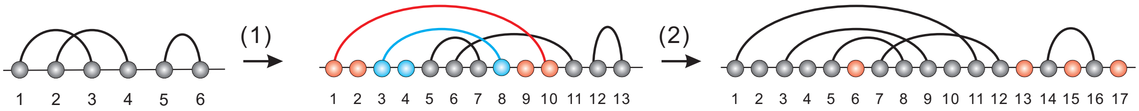

Figure 9. An $\mathsf{lv}_3^5$-shape (left) is inflated to an $\mathsf{lv}_3^1$-shape (right) in two steps. In (1) we add stacks of size one. Stacks are separated by isolated vertices to the left (blue) or both sides (red). In (2) we insert at most one isolated vertex (red) at the remaining positions.