STATISTICS OF CANONICAL RNA PSEUDOKNOT STRUCTURES

FENIX W.D. HUANG AND CHRISTIAN M. REIDYS*

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

Abstract. In this paper we study canonical RNA pseudoknot structures. We prove central limit theorems for the distributions of the arc-numbers of k-noncrossing RNA structures with given minimum stack-size τ over n nucleotides. Furthermore we compare the space of all canonical structures with canonical minimum free energy pseudoknot structures. Our results generalize the analysis of Schuster *et.al.* obtained for RNA secondary structures [11] and [15, 14] to k-noncrossing RNA structures. Here $k \geq 2$ and τ are arbitrary natural numbers. We compare canonical pseudoknot structures to arbitrary structures and show that canonical pseudoknot structures exhibit significantly smaller exponential growth rates. We then compute the asymptotic distribution of their arc-numbers. Finally we analyze how the minimum stack-size and crossing number factor into the distributions.

Date: April 2008.
Key words and phrases. k-noncrossing RNA structure, singularity analysis, central limit theorem, D-finite.

1. Introduction

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. RNA molecules form "helical" structures by pairing their nucleotides and thereby lowering their minimum free energy. The biochemistry of these nucleotide-pairings favors parallel stacking of bonds due to entropy. The resulting 3-dimensional configuration of the nucleotides is the RNA tertiary structure which, in many cases, determines the functionality of the molecule. The prediction of RNA structures is a central question and of crucial importance for finding and designing new RNA functionalities.

In this paper we study canonical RNA structures. Let us first explain what notion of coarse grained structure we employ. We represent RNA structures as diagrams and focus on the intramolecular nucleotide pair-interactions. Accordingly, we do not consider the particular embedding of the nucleotides in 3-dimensional space. In the diagram representation we identify a Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairing with an arc drawn in the upper halfplane and ignore the bonds of the primary sequence. A diagram is a labeled graph over the vertex set $[n] = \{1, \dots, n\}$ in which each vertex has degree ≤ 1, represented by drawing its in a horizontal line and its arcs (i, j), where $i < j$, in the upper half-plane. The vertices and arcs correspond to nucleotides and Watson-Crick (**A-U**, **G-C**) and (**U-G**) base pairs, respectively. Diagrams have the two key parameters k and τ. Here $k - 1$ is the maximum number of mutually crossing arcs, and τ the minimum length of a stack. By a stack of length τ we mean a sequence of "parallel" arcs of the form $((i, j), (i + 1, j - 1), \dots, (i + (\tau - 1), j - (\tau - 1)))$, see Figure 1. The length of an arc (i, j) is given by $j - i$. We call a k-noncrossing diagram with arc-length ≥ 2 and stack-length $\geq \tau$ a k-noncrossing τ-canonical RNA structure. A 2-canonical structure is called a *canonical* structure. In other words, canonical structures are those without isolated arcs. In Figure 1 we illustrate the properties k-noncrossing and canonicity. We denote the set (number) of k-noncrossing τ-canonical RNA structures by $T_{k,\tau}(n)$ ($\mathsf{T}_{k,\tau}(n)$). k-noncrossing RNA structures for $k \geq 3$ are called pseudoknot RNA structures. In Figure 2 we give the diagram representation of the hammerhead ribozyme [2].

Three decades ago Waterman *et.al.* pioneered the concept of RNA secondary structures [27, 29]. The latter are subject to the most strict combinatorial constraints: there exist no two arcs that cross in the diagram representation of the structure. It is well-known, however, that there exist crossing base pairs [20]. These configurations are called pseudoknots [30] and occur in functional RNA (RNAseP [19]), ribosomal RNA [18] and are conserved in the catalytic core of group I introns. Pseudoknots appear in plant viral RNAs pseudo-knots and in *in vitro* RNA evolution [25] experiments have produced families of RNA structures with pseudoknot motifs, when binding HIV-1 reverse transcriptase. Important mechanisms like ribosomal frame shifting [3] also involve pseudoknot interactions. k-noncrossing RNA structures [16] allow to express pseudoknots and generalize the concept of the RNA secondary structures in a natural way.

We are ultimately interested in designing new computer algorithms for the prediction of pseudoknot RNA [21]. The prediction of RNA pseudoknots is at present time difficult. We are (a) in lack of energy-parameters for specific pseudoknot motifs and (b) the combinatorial search-problem of structural configurations is nontrivial. In contrast, for RNA secondary structures there exists a dynamic programming routine which computes the minimum free energy structure for sequences of length $n = 100$ within seconds. The complexity of this algorithm is $O(n^3)$ in time and $O(n^2)$ in space [27] and in addition detailed energy parameters are available. Unfortunately, the dynamic programming routine developed by Waterman [27, 28] does not work for pseudoknot RNA. Due to the crossing of arcs no obvious recursive routine exists. Consequently, even if perfect data on pseudoknot RNA energy parameters are available their prediction would still be hard–as a combinatorial problem in its own right. Let us next explain why k-noncrossing RNA structures are conceptionally so different from RNA secondary structures. In Figure 3 we illustrate main steps in the enumeration of k-noncrossing RNA structures [13]. A structure is translated into a sequence of tableaux [4, 5] which corresponds to a $\mathbb{Z}^{k-1}$-walk that remains in the region $\{(x_1, \dots, x_{k-1}) \in \mathbb{Z}^{k-1} \mid x_1 \geq x_2 \geq \dots x_{k-1} \geq 0\}$ which starts and ends at 0. The boundaries of the above region are called walls. The enumeration is obtained employing the reflection principle. This method is due to D. André in 1887 [1] and has subsequently been generalized by Gessel and Zeilberger [8]. In the reflection principle "bad" walks cancel themselves. I.e. one enumerates all walks and due to cancellation only the ones survive that never touch the walls. This method does not trigger any algorithmic intuition and is nonconstructive. The asymptotic analysis [14] showed in particular that 3-noncrossing RNA structures grow at a rate of $(5+\sqrt{21})/2$. Accordingly, there exist more 3-noncrossing RNA structures than sequences over the natural alphabet! Of course this means that not all 3-noncrossing RNA structures can possibly be realized as minimum free-energy structures of natural sequences. This observation gives rise to the following question: can we identify the structures which occur as minimum free energy structures? To answer this question we can get some intuition from biophysics. Due to minimum free energy considerations isolated bonds do (practically) not occur in pseudoknot RNA, leading to the notion of canonical structures. It is therefore easy to identify a subset of pseudoknot structures containing minimum free energy structures. The difficult part is their analysis which required the concept of core-structures [16]. The key observation about canonical k-noncrossing RNA structures is their small exponential growth rates, see Table 2. In particular, there are significantly less k-noncrossing canonical RNA pseudoknot structures than sequences over the natural alphabet. This gives some first clue on the existence of neutral networks of RNA pseudoknot structures.

Let us put our results into context: Schuster *et.al.* [11] derived several asymptotic formulas for the numbers of RNA secondary structures with given minimum stack- and arc-length. Furthermore it has been shown that the arc-numbers of all 3-noncrossing RNA structures satisfy a central limit theorem [15]. In this paper we present a (k, τ)-matrix of central limit theorems, $k - 1$ being the number of mutually crossing arcs and τ the minimum stack-length, see Figure 4 and Table 1. Our results generalize those in [11] to arbitrary crossing number k. In comparison to [15] we present a new sequence of arguments based on the D-finiteness of the generating functions. Our arguments are much more "structural" and work for arbitrary crossing number k. For instance, we use general structure theorems about asymptotic solutions of ODE in order to show the uniformity of the error bounds in s and z. This uniformity is critical for the existence of central limit theorems. Our results are derived from a new functional equation, proved in Lemma 3 which relates k-noncrossing structures for arbitrary k and τ and a certain bivariate generating function of k-noncrossing matchings. We give in Table 1 key information about our central limit theorems listing means and variances for various k and τ. Table 1 shows that for secondary structures the mean simply increases with increasing τ. This observation is intuitively clear: higher minimum stack-size requirements result in more and more densely packed arc-configurations. In contrast, pseudoknot structures for $k > 2$ exhibit a *drop* when passing from arbitrary to canonical structures. Here the large stacks typical for canonical structures are antagonistic to complex crossing motifs. We can paraphrase the situation by saying that canonical structures are more "ordered" and that this order limits their numbers, see Figure 2 and Section 5

Let us finally discuss the relation between the statistics of all canonical pseudoknot structures and those realized by some minimum free energy pseudoknot folding algorithm. Of course, in lack of detailed energy parameters such a discussion is to some extend speculative. However, we believe that statistically meaningful results can be produced. For a given sequence we search exhaustively all canonical structures. This search is not trivial and utilizes the interpretation of RNA pseudoknot structures as tableaux-sequences [13], see Figure 3. In addition we are guaranteed to search *all* possible canonical structures. As for energy parameters we use the standard energy assignments from Vienna RNAfold [10]. We then compute the arc-distributions induced by the top 5% pseudoknot structures. Here a "top"-structure is one that is optimal w.r.t. the energy parameters of Vienna RNAfold [10]. Since the energy parameters for pseudoknot structures are not sufficiently explored this averaging procedure is necessary in order to guarantee generic results. In Figure 5 we present average arc-frequency distributions of canonical minimum free energy pseudoknot structures for 3000 uniform (uni) and random (rand) sequences, respectively. Here uniform means that the ratios of the nucleotides within the sequence are equal. The data show that for $k = 2$ unimodal curves with a mean of $\mu_{\text{uni}} = 11.8380$ and $\mu_{\text{rand}} = 11.4125$, respectively. For $k = 3$ we observe $\mu_{\text{uni}} = 13.7807$ and $\mu_{\text{rand}} = 13.2456$. Theory predicts for $k = 2$, $\tau = 2$, $\mu = 12.688$ and for $k = 3$, $\tau = 2$, $\mu = 15.268$, respectively.

2. Background

In this Section we discuss several basic facts instrumental for our arguments. For particular background on crossings and nestings in diagrams and partitions we recommend the paper of Chen *et.al.* [4, 5]. Analytic combinatorics and singularity analysis can be found in the book of Flajolet [7]. In the following we will discuss the generating function of k-noncrossing RNA structures [13], analytic continuation [14] and asymptotic analysis of k-noncrossing RNA structures [14, 15]. We recall that $T_{k,\tau}(n)$ ($\mathsf{T}_{k,\tau}(n)$) denotes the set (number) of k-noncrossing RNA structures with minimum stack length τ. $T_{k,\tau}(n)$ is identified with a set of diagrams. Such a diagram is obtained by drawing its vertice $1, \dots, n$ in a horizontal line and its arcs (i, j), where $i < j$, in the upper half plane. All arcs have a minimum length ≥ 2 and stack-length $\geq \tau$ and $k - 1$ is the maximum number of mutually crossing arcs. Furthermore let $T_{k,\tau}(n, h)$ denote the set of k-noncrossing RNA structures stack-length $\geq \tau$ having exactly h arcs and let $\mathsf{T}_{k,\tau}(n, h)$ denote their number. A k-noncrossing core-structure is a k-noncrossing RNA structures in which there exists *no* two arcs of the form $(i,j),(i+1,j-1)$. We denote the set (number) of core-structures having h arcs by $C_k(n,h)$ ($\mathsf{C}_k(n,h)$) and $C_k(n)$ ($\mathsf{C}_k(n)$) denotes the set (number) of core-structures. We will use core-structures in order to prove the central functional equation in Lemma 3. Let finally $f_k(n,\ell)$ be the number of k-noncrossing diagrams with arbitrary arc-length and ℓ isolated points. These diagrams are also called partial matchings. For $\ell=0$ we refer to them as matchings since they have no isolated vertices. To provide some intuition we present in Figure 6 the various types of diagrams involved. Let us begin by stating two identities for the generating function of k-noncrossing diagrams without (matchings) and with isolated points (partial matchings), due to Grabiner *et.al.* [9]

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

$$\sum_{n\geq 0}\left\{\sum_{\ell=0}^{n} f_k(n,\ell)\right\}\cdot\frac{x^n}{n!} \quad = \quad e^x\det[I_{i-j}(2x)-I_{i+j}(2x)]|_{i,j=1}^{k-1} \tag{2.2}$$

where $I_r(2x)=\sum_{j\geq 0}(x^{2j+r})/(j!(r+j)!)$ denotes the hyperbolic Bessel function of the first kind of order r. Eq. (2.1) and (2.2) allow to prove that $\sum_{n\geq 0} f_k(n,0)x^n$ is D-finite since the hyperbolic Bessel function is D-finite and D-finite functions form an algebra closed under taking Hadamard products. A power series $u(x)$ is D-finite if $\dim_{K(x)}\{u,u',\dots\}<\infty$ [22]. In addition, eq. (2.1) and (2.2) allow "in principle" for explicit computation of the numbers $f_k(n,\ell)$. In particular for $k=2$ and $k=3$ we have the formulas

$$f_2(n,\ell)=\binom{n}{\ell}C_{(n-\ell)/2} \quad \text{and} \quad f_3(n,\ell)=\binom{n}{\ell}\left[C_{(n-\ell)/2+2}C_{(n-\ell)/2}-C^2_{(n-\ell)/2+1}\right] \ , \tag{2.3}$$

where C_m denotes the m-th Catalan number. The second formula results from a determinant formula enumerating pairs of nonintersecting Dyck-paths. In view of

$$f_k(n,\ell)=\binom{n}{\ell}f_k(n-\ell,0)$$

everything can be reduced to matchings, where we have the asymptotic formula [17]

$$f_k(n) \sim c_k\, n^{-((k-1)^2+(k-1)/2)}\, (2(k-1))^{2n}, \qquad \text{for some } c_k > 0\ . \tag{2.4}$$

The number of k-noncrossing RNA structures with $((n-\ell)/2)$ arcs, $\mathsf{T}_{k,1}(n,(n-\ell)/2)$, and the number of k-noncrossing RNA structures, $\mathsf{T}_{k,1}(n)$, are given by [13]

$$\mathsf{T}_{k,1}(n,(n-\ell)/2) \;=\; \sum_{b=0}^{\lfloor n/2\rfloor} (-1)^b \binom{n-b}{b} f_k(n-2b,\ell) \tag{2.5}$$

$$\mathsf{T}_{k,1}(n) \;=\; \sum_{b=0}^{\lfloor n/2\rfloor} (-1)^b \binom{n-b}{b} \left\{ \sum_{\ell=0}^{n-2b} f_k(n-2b,\ell) \right\}\ , \tag{2.6}$$

where $\{\sum_{\ell=0}^{n-2b} f_k(n-2b,\ell)\}$ is given via eq. (2.2). The following functional identity is due to [14] and relates the bivariate generating function for $\mathsf{T}_{k,1}(n,h)$, the number of RNA pseudoknot structures with h arcs to the generating function of k-noncrossing matchings.

Lemma 1. *Let $k \in \mathbb{N}$, $k \geq 2$ and z, u be indeterminants over $\mathbb{C}$. Then we have the following identity of analytic functions*

$$\sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{T}_{k,1}(n,h)\; u^{2h} z^n = \frac{1}{u^2z^2-z+1} \sum_{n\geq 0} f_k(2n,0) \left(\frac{uz}{u^2z^2-z+1}\right)^{2n}\ . \tag{2.7}$$

It will be important to deduce relations between the coefficients from the equality of generating functions. The class of theorems that deal with this deduction are called transfer-theorems [7]. 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 [7] and its application to 3-noncrossing RNA in [14]. To be precise, given two numbers ϕ, R, where

$R > 1$ and $0 < \phi < \pi/2$ and $\rho \in \mathbb{R}$ the open domain $\Delta_\rho(\phi, R)$ is defined as

$$\Delta_\rho(\phi, R) = \{z \mid |z| < R, z \neq \rho, \, |\mathrm{Arg}(z - \rho)| > \phi\} \tag{2.8}$$

A domain is a Δ_ρ-domain if it is of the form $\Delta_\rho(\phi, R)$ for some R and ϕ. A function is Δ_ρ-analytic if it is analytic in some Δ_ρ-domain. We use the notation

$$(f(z) = O\left(g(z)\right) \text{ as } z \to \rho) \quad \Longleftrightarrow \quad (f(z)/g(z) \text{ is bounded as } z \to \rho) \tag{2.9}$$

and if we write $f(z) = O(g(z))$ it is implicitly assumed that z tends to a (unique) singularity. $[z^n]\, f(z)$ denotes the coefficient of z^n in the power series expansion of $f(z)$ around 0. Theorem 1 allows us to obtain key information about the coefficients of a power series based on its behavior locally at its dominant singularities.

Theorem 1. [7] *Let $f(z)$ be a Δ-analytic function and $g(z)$ its singular expansion at a singularity ρ. That is we have in the intersection of a neighborhood of ρ with the Δ-domain*

$$f(z) = O(g(z)) \quad \textit{for } z \to \rho \ . \tag{2.10}$$

Then we have

$$[z^n]f(z) = A\ (1 - O(1/n))\ [z^n]g(z) \quad \textit{for some } A \in \mathbb{C} \ . \tag{2.11}$$

Let $S(\rho, n)$ denote the subexponential factor of $[z^n]\, g(z)$ of the singularity ρ. Note that in general $[z^n]\, g(z)$ is a sum over all dominant singularities of the form $[z^n]\, g(z) \sim \sum_i S(\rho_i, n)\rho_i^n$. The second result is a consequence of Theorem 1 and the uniformity lemma of singularity analysis see [7], Lemma XI.2, p. 635.

Theorem 2. *Using the notation of Theorem 1, let $\psi(z,s)$ be an algebraic, analytic function in z and s such that $\psi(0,s)=0$ and for $|s|<\epsilon$ all $\psi(z,s)$-singularities have modulus strictly greater than ρ. In addition suppose $\gamma(s)$ is the unique dominant singularity of $f(\psi(z,s))$ and unique analytic solution of $\psi(\gamma(s),s)=\rho$ for $|s|<\epsilon$. Then $f(\psi(z,s))$ has a singular expansion and*

$$[z^n]f(\psi(z,s)) = A(s)\ (1-O(1/n))\ S(\rho,n)\left(\frac{1}{\gamma(s)}\right)^n \quad \textit{for some } A(s)\in\mathbb{C}\ , \tag{2.12}$$

uniformly for s contained in a neighborhood of 0.

The key property of the singular expansion of Theorem 2 is the uniformity of eq. (2.12) in the parameter s. The observation that $f(\psi(z,s))$ has a singular expansion is due to Stanley, [22] proved in the context of closure properties of D-finite functions. Transfer theorems are accordingly a translation of error terms from functions to coefficients and guaranteed when the functions in question are analytic in some Δ_ρ-domain. For our purposes the D-finiteness of the ordinary generating function $\sum_{n\geq 0} f_k(2n,0)x^{2n}$ implies its Δ_{ρ_k}-analyticity and the existence of a singular expansion, see Lemma 5.

3. Functional equations

The first Lemma relates the number of canonical structures to core-structures [16]. Core-structures here serve as an intermediate step via which we can relate the numbers $\mathsf{T}_{k,\tau}(n)$ and $\mathsf{T}_{k,1}(n)$. Lemma 3 rewrites this bivariate generating function as a composition of two “simple” functions. This is crucial for the subsequent singularity analysis insofar as we encounter a phenomenon known as persistence of the singularity of the “outer” function, i.e. we have the *supercritical case* [6]. The type of singularity coincides with that of the generating function of k-noncrossing matchings. In Lemma 4 we use Lemma 3 in order to draw first conclusions about the singularities. Finally

Lemma 5 asserts that we have an unique dominant singularity and that the subexponential factors coincide with those from $f_k(2n, 0)$ and are independent of s.

Lemma 2. [16] *Let $k, \tau \in \mathbb{N}$, $k \geq 2$ and let u, x be indeterminants. Then we have the functional relation*

$$\sum_{n\geq 0}\sum_{h\leq \frac{n}{2}} \mathsf{T}_{k,\tau}(n,h)u^h x^n = \sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{C}_k(n,h)\left(\frac{u\cdot(ux^2)^{\tau-1}}{1-ux^2}\right)^h x^n + \frac{x}{1-x} \tag{3.1}$$

and in particular, for $u = 1$

$$\sum_{n\geq 0} \mathsf{T}_{k,\tau}(n)x^n = \sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{C}_k(n,h)\left(\frac{(x^2)^{\tau-1}}{1-x^2}\right)^h x^n + \frac{x}{1-x}\ . \tag{3.2}$$

The key idea is now to combine Lemma 2 and Lemma 1 as follows:

Lemma 3. *Let $k, \tau \in \mathbb{N}$ $k \geq 2$ and suppose u, x are indeterminants. Then we have the functional relation of formal power series*

$$\sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{T}_{k,\tau}(n,h)u^h x^n = \frac{1}{u_0x^2-x+1}\sum_{n\geq 0} f_k(2n,0)\left(\frac{\sqrt{u_0}x}{u_ox^2-x+1}\right)^{2n} \tag{3.3}$$

where $u_0 = u_0(x, u)$ is given by

$$u_0 = \frac{u\,(ux^2)^{\tau-1}}{(ux^2)^{\tau}-ux^2+1}\ . \tag{3.4}$$

Considered as a relation between analytic functions, eq. (3.3) holds for $u = e^s$ and $|s| \leq \epsilon$ for ϵ sufficiently small and $|x| \leq 1/2$.

Proof. According to Lemma 2

$$\sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{T}_{k,\tau}(n,h)u^h x^n = \sum_{n\geq 0}\sum_{h\leq n/2} \mathsf{C}_k(n,h)\left(\frac{u(ux^2)^{\tau-1}}{1-ux^2}\right)^h x^n + \frac{x}{1-x} \tag{3.5}$$

holds and in particular, for $\tau = 1$

$$\sum_{n\geq 0}\sum_{h\leq n/2}\mathsf{T}_{k,1}(n,h)w^h x^n = \sum_{n\geq 0}\sum_{h\leq n/2}\mathsf{C}_k(n,h)\left(\frac{w}{1-wx^2}\right)^h x^n + \frac{x}{1-x}\ . \tag{3.6}$$

According to Lemma 1 we have

$$\sum_{n\geq 0}\sum_{h\leq n}\mathsf{T}_{k,1}(n)w^{2h}x^n = \frac{1}{w^2x^2-x+1}\sum_{n\geq 0} f_k(2n,0)\left(\frac{wx}{w^2x^2-x+1}\right)^{2n} \tag{3.7}$$

Setting $w^2 = (u(ux^2)^{\tau-1})/(1-ux^2+(ux^2)^\tau)$ we obtain

$$\frac{u\,(ux^2)^{\tau-1}}{1-ux^2} = \frac{w^2}{1-w^2x^2}\ .$$

We proceed by using the expression for w^2 in order to relate eq. (3.5) and eq. (3.7). This "connection" is facilitated via eq. (3.6)

$$\begin{aligned}\sum_{n\geq 0}\sum_{n\leq n/2} T_{k,\tau}(n,h)u^h x^n &= \sum_{n\geq 0}\sum_{h\leq n/2} C_k(n,h)\left(\frac{u(ux^2)^{\tau-1}}{1-ux^2}\right)^h x^n + \frac{x}{1-x}\\ &= \sum_{n\geq 0}\sum_{h\leq n/2} T_{k,1}(n,h)u_0^h x^n\\ &= \frac{1}{u_0x^2-x+1}\sum_{n\geq 0} f_k(2n,0)\left(\frac{\sqrt{u_0}x}{u_0x^2-x+1}\right)^{2n}\ ,\end{aligned}$$

where $u_0 = (u(ux^2)^{\tau-1})/(1-ux^2+(ux^2)^\tau)$. Eq. (3.3) can be considered as a relation between analytic functions for x, u with the property $1-ux^2+(ux^2)^\tau \neq 0$. The conditions $u = e^s$, $|s| \leq \epsilon$ for ϵ sufficiently small, $|x| \leq 1/2$ and the continuity (in s) of the roots of the s parametrized family of polynomials

$$p_s(X) = (e^s\,X^2)^\tau - e^s\,X^2 + 1$$

guarantee that $p_s(x) \neq 0$ for $|x| \leq 1/2$, whence the Lemma. □

Suppose $\epsilon > 0$, $k \in \mathbb{N}$, $k \geq 2$ and $u = e^s$, where $|s| < \epsilon$. We set

$$\varphi_{n,k,\tau}(s) = \sum_{h \leq n/2} T_{k,\tau}(n,h) e^{hs} \tag{3.8}$$

$$U_k(z,s) = \sum_{n \geq 0} \varphi_{n,k,\tau}(s) z^n \tag{3.9}$$

Via Lemma 3 the generating function $\sum_n f(2n,0)z^{2n}$ becomes of interest. According to the theorem of Pfringsheim [24] it has a dominant positive real singularity, which we denote by ρ_k.

Lemma 4. *Suppose $\epsilon > 0$, $k \in \mathbb{N}$, $k \geq 2$ and $u = e^s$, where $|s| < \epsilon$. Then we have for $|s| < \epsilon$, $z \in \mathbb{C}$ the identity of formal power series:*

$$U_k(z,s) = \frac{1}{u_0 z^2 - z + 1} \sum_{n \geq 0} f_k(2n,0) \left(\frac{\sqrt{u_0} z}{u_0 z^2 - z + 1} \right)^{2n} \tag{3.10}$$

where $u_0 = (e^s(e^s z^2)^{\tau-1})/(1 - e^s z^2 + (e^s z^2)^{\tau})$. Furthermore, any dominant singularity of $U_k(z,s)$ is a singularity of $\sum_{n \geq 0} f_k(2n,0) \left((\sqrt{u_0} z)/(u_0 z^2 - z + 1)\right)^{2n}$. Let $\gamma_{k,\tau}(s)$ be the solution of the equation

$$\frac{\sqrt{u_0} z}{u_0 z^2 - z + 1} - \rho_k = 0 \ , \tag{3.11}$$

such that $\gamma_{k,\tau}(0)$ is the minimal real positive solution of eq. (3.11). Then there exists an analytic function $\gamma_{k,\tau}(s)$ such that $\gamma_{k,\tau}(s)$, is a dominant singularity of $U_k(z,s)$.

Proof. The formal identity of eq. (3.10) follows from Lemma 3 setting $u = e^s$. We next prove the existence of $\gamma_{k,\tau}(s)$. For this purpose we consider the equation

$$F(z,s) = \left((\sqrt{u_0} z)/(u_0 z^2 - z + 1)\right) - \rho_k \ . \tag{3.12}$$

For $s = 0$ we easily derive that there exists a unique minimal real solution ω. For $|s| < \epsilon$, we observe $F(\omega, 0) = 0$, $F_z(\omega, 0) \neq 0$ and the partial derivatives $F_z(z,s)$ and $F_s(z,s)$ are continuous.

According to the analytic implicit function theorem [7], there exists an unique analytic function $\gamma_{k,\tau}(s)$ that satisfies

$$F(\gamma_{k,\tau}(s), s) = 0 \ , \quad \text{and} \quad \gamma_{k,\tau}(0) = \omega$$

which proves that $\gamma_{k,\tau}(s)$ exists. Let us denote $W_k(z,s) = \sum_{n\geq 0} f_k(2n,0)\left((\sqrt{u_0}z)/(u_0z^2 - z + 1)\right)^{2n}$.

Claim. For $|s| < \epsilon$, all dominant singularities of $U_k(z,s)$ are singularities of $W_k(z,s)$ and $\gamma_{k,\tau}(s)$ is a dominant singularity.

Let $\zeta(s)$ be a dominant singularity of $U_k(z,s)$. Eq. (3.10) shows that $\zeta(s)$ is either a dominant singularity of

$$W_k(z,s) \quad \text{or} \quad 1/(u_0z^2 - z + 1) \ .$$

If $\zeta(s)$ is a singularity of $1/(u_0z^2 - z + 1)$, then $\zeta(s)$ is also a singularity of

$$\psi_\tau(z,s) = (\sqrt{u_0}z)/(u_0z^2 - z + 1) \tag{3.13}$$

and $W_k(z,s)$ is non-finite at $\zeta(s)$. We set now $s = 0$. Since $W_k(z,0)$ has positive coefficients and $\gamma_{k,\tau}(0)$ is real and positive, $|\zeta(0)| \leq \gamma_{k,\tau}(0)$ implies

$$|W_k(\zeta(0),0)| \leq W(\gamma_{k,\tau}(0),0) \ ,$$

which is impossible since

$$W(\gamma_{k,\tau}(0),0) = \sum_n f_k(2n,0)\rho_k^{2n} \tag{3.14}$$

and ρ_k is an algebraic singularity of $\sum_n f_k(2n,0)z^{2n}$. We can conclude from this that the singularity $\zeta(0)$ has modulus strictly larger than $\gamma_{k,\tau}(0)$, i.e. $|\zeta(0)| > \gamma_{k,\tau}(0)$. We proceed by applying an continuity argument. For ϵ sufficiently small and $|s| < \epsilon$ the singularities of $1/(u_0z^2 - z + 1)$ and $\gamma_{k,\tau}(s)$ are continuous in s. Therefore we can conclude that for sufficiently small ϵ

$$|\zeta(s)| > |\gamma_{k,\tau}(s)| \tag{3.15}$$

holds and we have proved that for $|s| < \epsilon$ and ϵ sufficiently small all dominant singularities of $U_k(z,s)$ are singularities of $W(z,s)$. By construction $\gamma_{k,\tau}(s)$ is a singularity of $U(z,s)$ and $\gamma_{k,\tau}(0)$ is a dominant singularity of $U(z,0)$. Since $\gamma_{k,\tau}(s)$ is continuous we can conclude from this that for ϵ sufficiently small $\gamma_{k,\tau}(s)$ is dominant, whence the Claim and the Lemma follows. □

Lemma 5. *Suppose $\epsilon > 0$, $k \in \mathbb{N}$, $k \geq 2$ and $u = e^s$, where $|s| < \epsilon$. Then $\gamma_{k,\tau}(s)$ is the unique dominant singularity of $U_k(z,s)$ and*

$$[z^n]\,U_k(z,s) = A(s)\;(1 - O(1/n))\;n^{-((k-1)^2+(k-1)/2)}\left(\frac{1}{\gamma_{k,\tau}(s)}\right)^n \quad \textit{for some } A(s) \in \mathbb{C}\ , \tag{3.16}$$

uniformly in s in a neighborhood of 0. In particular, the subexponential factors of the coefficients of $U_k(z,s)$ coincide with those of $F_k(z) = \sum_n f_k(2n,0)z^{2n}$ and are independent of s.

Proof. Claim. $\gamma_{k,\tau}$ is the unique.
In view of Lemma 4 we analyze the dominant singularities of $F_k(z) = \sum_n f_k(2n,0)z^{2n}$. For this purpose we observe that $F_k(z) = \sum_n f_k(2n,0)z^{2n}$ is D-finite. Accordingly there exists some $e \in \mathbb{N}$ for which $F_k(z)$ satisfies an ODE of the form

$$q_{0,k}(z)\frac{d^e}{dz^e}F_k(z) + q_{1,k}(z)\frac{d^{e-1}}{dz^{e-1}}F_k(z) + q_{e,k}(z)F_k(z) = 0\ , \tag{3.17}$$

where $q_{j,k}(z)$ are polynomials. The key point is now that any dominant singularities of $F_k(z)$ is contained in the set of roots of $q_{0,k}(z)$, which we denote by M_k [22]. We then compute, see eq. (3.22)-(3.26), that for $k = 3,\dots,7$, $\gamma_{k,\tau}$ is the *unique* solution with minimal modulus of

$$\psi_\tau(z,s) = |\rho_k| \tag{3.18}$$

(see eq. (3.13)) and $\gamma_{k,\tau}$ is in fact a solution of $\psi_\tau(z,s) = \rho_k$. This proves the Claim. Let $Q_{\gamma_{k,\tau}(s)}(z,s)$ denote the singular expansion of $U_k(z,s)$ at $\gamma_{k,\tau}(s)$. According to eq. (2.4) we have $f_k(n) \sim c_k\, n^{-((k-1)^2+(k-1)/2)}\,(2(k-1))^{2n}$ for some $c_k > 0$. In combination with Theorem 1 we can conclude

$$F_k(z) = \begin{cases} O((z-\rho_k)^{(k-1)^2+(k-1)/2)-1}\ln(z-\rho_k)) & \text{for } k \text{ odd, } z \longrightarrow \rho_k \\ O((z-\rho_k)^{(k-1)^2+(k-1)/2)-1}) & \text{for } k \text{ even, } z \longrightarrow \rho_k, \end{cases}$$

in accordance with basic structure theorems for singularities of solutions of eq. (3.17) [7], p. 499. According to Lemma 4 we have

$$U_k(z,s) = \frac{1}{(u_0 z^2 - z + 1)}\, F_k(\psi_\tau(z,s)) \tag{3.19}$$

where $\psi_\tau(z,s)$ is given by eq. (3.13). We showed in Lemma 4 that $\psi_\tau(z,s)$ does not induce any dominant singularities and is regular at ρ_k. Let $Q_{\rho_k}(z)$ denote the singular expansion of $F_k(z)$ at the dominant singularity ρ_k, i.e.

$$F_k(z) = O(Q_{\rho_k}(z)) \quad \text{for } z \longrightarrow \rho_k \ .$$

The singular expansion of $F_k(\psi_\tau(z,s))$, $Q_{\gamma_{k,\tau}(s)}(z,s)$, is derived by substituting the Taylor-expansion of $\psi_\tau(z,s)$ into $Q_{\rho_k}(z)$ and we observe

$$Q_{\gamma_{k,\tau}(s)}(z,s) = Q_{\rho_k}(\psi_\tau(\zeta_k(s),s)) = O(Q_{\gamma_{k,\tau}(s)}(z)) \ . \tag{3.20}$$

Indeed, eq. (3.20) follows immediately substituting $\psi_\tau(z,s) - \psi_\tau(\gamma_{k,\tau}(s),s)$ for $z - \rho_k$ which does not change the singular expansion. According to Theorem 2 we can conclude

$$[z^n]\, U_k(z,s) = A(s)\ (1 - O(1/n))\ n^{-((k-1)^2+(k-1)/2)} \left(\frac{1}{\gamma_{k,\tau}(s)}\right)^n \quad \text{for some } A(s) \in \mathbb{C} \ , \tag{3.21}$$

uniformly in s in a neighborhood of 0. Therefore the asymptotic expansion is uniform in s and eq. (3.16) follows. The proof shows in addition that the subexponential factors of the coefficients of $U_k(z,s)$ coincide with those of $F_k(z)$ and are independent of s. □

In the following we give the polynomials $q_{0,k}(z)$ and their sets of roots for $k = 3, \ldots, 7$. Note that the following data confirm $\rho_k = (2(k-1))^{-1}$ as given in eq. (2.4)

(3.22)
$$q_{0,3}(z) = (1/4 - 4z^2)\, z^2 \qquad M_3 = \{1/4, -1/4\}$$

(3.23)
$$q_{0,4}(z) = (144\, z^4 - 40\, z^2 + 1)\, z^6 \qquad M_4 = \{1/2, -1/2, 1/6, -1/6\}$$

(3.24)
$$q_{0,5}(z) = (-80\, z^2 + 1024\, z^4 + 1)\, z^8 \qquad M_5 = \{1/4, -1/4, 1/8, -1/8\}$$

(3.25)
$$q_{0,6}(z) = (-4144\, z^4 + 140\, z^2 + 14400\, z^6 + 1)\, z^{10} \qquad M_6 = \{1/2, -1/2, 1/6, -1/6, 1/10, -1/10\}$$

(3.26)
$$q_{0,7}(z) = (-1 - 12544\, z^4 + 224\, z^2 + 147456\, z^6)\, z^{12} \qquad M_7 = \{1/4, -1/4, 1/8, -1/8, 1/12, -1/12\}$$

Analysis of $\psi_\tau(z, s) = |\rho_k|$ allows us to conclude that $\rho_3 = 1/4$, $\rho_4 = 1/6$, $\rho_5 = 1/8$, $\rho_6 = 1/10$ and finally $\rho_7 = 1/12$.

4. Central limit theorems

Before we state our main result we give a classic result on limit distributions which is instrumental for its proof.

Theorem 3. (Lévy-Cramér) *Let $\{\xi_n\}$ be a sequence of random variables and let $\{\varphi_n(x)\}$ and $\{F_n(x)\}$ be the corresponding sequences of characteristic and distribution functions. If there exists a function $\varphi(t)$, such that $\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$ uniformly over an arbitrary finite interval enclosing* *the origin, then there exists a random variable ξ with distribution function $F(x)$ such that*

$$F_n(x) \Longrightarrow F(x)$$

uniformly over any finite or infinite interval of continuity of $F(x)$.

We now consider the random variable $X_{n,k,\tau}$ having the distribution

$$\mathbb{P}(X_{n,k,\tau} = h) = \mathsf{T}_{k,\tau}(n,h)/\mathsf{T}_{k,\tau}(n)$$

where $h = 0, 1, \dots \lfloor n/2 \rfloor$. Remarkably, the particular distribution is determined by the shift of the singularity parametrized by s. Lemma 4 and Lemma 5 provide the essential information about the bivariate generating function $U_k(z, s)$. The key point in Theorem 4 below consists in analyzing the characteristic function and then to apply the Lévy-Cramér Theorem.

Theorem 4. *Let $k, \tau \in \mathbb{N}$, $k \geq 2$. Then for given k and τ there exist a pair $(\mu_{k,\tau}, \sigma_{k,\tau})$ such that the normalized random variable*

$$Y_{n,k,\tau} = \frac{X_{n,k,\tau} - \mu_{k,\tau}\, n}{\sqrt{n\, \sigma_{k,\tau}{}^2}} \tag{4.1}$$

has asymptotically normal distribution with parameter $(0,1)$, i.e. we have

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_{n,k,\tau} - \mu_{k,\tau} n}{\sqrt{n\, \sigma_{k,\tau}^2}} < x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-1/2t^2} dt \ , \tag{4.2}$$

where $\mu_{k,\tau}$ and $\sigma_{k,\tau}^2$ are given by

$$\mu_{k,\tau} = -\frac{\gamma_{k,\tau}'(0)}{\gamma_{k,\tau}(0)}, \qquad \sigma_{k,\tau}^2 = \left(\frac{\gamma_{k,\tau}'(0)}{\gamma_{k,\tau}(0)}\right)^2 - \frac{\gamma_{k,\tau}''(0)}{\gamma_{k,\tau}(0)}. \tag{4.3}$$

Proof. Let us recall

$$\varphi_{n,k,\tau}(s)=\sum_{h\le n/2} T_{k,\tau}(n,h)e^{hs} \qquad U_k(z,s)=\sum_{n\ge 0}\varphi_{n,k,\tau}(s)z^n \ .$$

Suppose we are given the random variable (r.v.) ξ_n with mean μ_n and variance σ_n^2. We consider the rescaled r.v. $\eta_n=(\xi_n-\mu_n)\sigma_n^{-1}$ and the characteristic function of η_n:

$$f_{\eta_n}(t)=\mathbb{E}[e^{it\eta_n}]=\mathbb{E}[e^{it\frac{\xi_n}{\sigma_n}}]e^{-i\frac{\mu_n}{\sigma_n}t} \ . \tag{4.4}$$

Writing X_n instead of $X_{n,k,\tau}$ we derive for $\xi_n=X_n$, substituting the term $\mathbb{E}[e^{it\eta_n}]$

$$f_{X_n}(t)=\left(\sum_{h=0}^{n}\frac{\mathsf{T}_{k,\tau}(n,h)}{\mathsf{T}_{k,\tau}(n)}e^{it\frac{h}{\sigma_n}}\right)\,e^{-i\frac{\mu_n}{\sigma_n}t} \ . \tag{4.5}$$

In view of

$$\varphi_{n,k,\tau}(s)=\sum_{h\le n/2}\mathsf{T}_{k,\tau}(n,h)e^{hs} \ ,$$

we interpret $\sum_{h\le n/2}\mathsf{T}_{k,\tau}(n,h)$ and $\sum_{h\le n/2}\mathsf{T}_{k,\tau}(n,h)e^{h(it)/(\sigma_n)}$, as $\varphi_{n,k,\tau}(0)$ and $\varphi_{n,k,\tau}((it)/(\sigma_n))$, respectively. Writing φ_n instead of $\varphi_{n,k,\tau}$, we accordingly obtain

$$f_{X_n}(t)=\frac{1}{\varphi_n(0)}\,\varphi_n(\frac{it}{\sigma_n})\,e^{-i\frac{\mu_n}{\sigma_n}t} \ . \tag{4.6}$$

Now we have arrived at the crucial point of the proof: we have to provide the interpretation of $\varphi_n(0)$ and $\varphi_n((it)/(\sigma_n))$. This is obtained via Lemma 5:

$$[z^n]\,U_k(z,s)=K(s)\,\theta_k(n)\,\left(\gamma_{k,\tau}(s)^{-1}\right)^n(1-O(1/n)) \quad \text{for some } K(s)\in\mathbb{C} \ , \tag{4.7}$$

uniformly in s and where $\theta_k(n)$ is some subexponential factor, independent of s (we showed that the singular expansion remains invariant when substituting $\psi_\tau(z,s)$ for z). Therefore

$$f_{X_n}(t)\sim\frac{K(\frac{it}{\sigma_n})}{K(0)}\left[\frac{\gamma_{k,\tau}(\frac{it}{\sigma_n})}{\gamma_{k,\tau}(0)}\right]^{-n}e^{-i\frac{\mu_n}{\sigma_n}t} \ , \tag{4.8}$$

uniformly in t, where t is contained in an arbitrary, bounded interval. The rest of the proof is analogous to [15]: taking the logarithm we obtain

$$\ln f_{X_n}(t) \sim \ln \frac{K(\frac{it}{\sigma_n})}{K(0)} - n \ln \frac{\gamma_{k,\tau}(\frac{it}{\sigma_n})}{\gamma_{k,\tau}(0)} - i\frac{\mu_n}{\sigma_n}t \ . \tag{4.9}$$

Expanding $g(s) = \ln(\gamma_{k,\tau}(s))/(\gamma_{k,\tau}(0))$ in its Taylor series at $s = 0$, (note that $g(0) = 0$ holds) yields

$$\ln \frac{\gamma_{k,\tau}(\frac{it}{\sigma_n})}{\gamma_{k,\tau}(0)} = \frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)} \frac{it}{\sigma_n} - \left[\frac{\gamma''_{k,\tau}(0)}{\gamma_{k,\tau}(0)} - \left(\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)}\right)^2\right] \frac{t^2}{2\sigma_n^2} + O(\left(\frac{it}{\sigma_n}\right)^3) \tag{4.10}$$

and therefore $\ln f_{X_n}(t)$ becomes asymptotically

$$\ln \frac{K(\frac{it}{\sigma_n})}{K(0)} - n \left\{\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)} \frac{it}{\sigma_n} - \frac{1}{2}\left[\frac{\gamma''_{k,\tau}(0)}{\gamma_{k,\tau}(0)} - \left(\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)}\right)^2\right] \frac{t^2}{\sigma_n^2} + O(\left(\frac{it}{\sigma_n}\right)^3)\right\} - \frac{i\mu_n t}{\sigma_n} \ . \tag{4.11}$$

$U_k(z, s)$ is analytic in s where s is contained in a disc of radius ϵ around 0 and therefore in particular continuous in s for $|s| < \epsilon$. In view of eq. (4.11) we introduce

$$\mu = -\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)}, \qquad \sigma^2 = \left\{\left(\frac{\gamma'_{k,\tau}(0)}{\gamma_{k,\tau}(0)}\right)^2 - \frac{\gamma''_{k,\tau}(0)}{\gamma_{k,\tau}(0)}\right\}$$

Setting $\mu_n = n\mu$ and $\sigma_n^2 = n\sigma^2$ we can conclude from eq. (4.7) for fixed $t \in]-\infty, \infty[$

$$\lim_{n\to\infty} (\ln K((it)/(\sigma_n)) - \ln K(0)) = 0 \tag{4.12}$$

and eq. (4.11) becomes

$$\ln f_{X_n}(t) \sim -\, t^2/2 + O(((it)/\sigma_n)^3) \tag{4.13}$$

uniform for t from any bounded interval. This is equivalent to $\lim_{n\to\infty} f_{X_n}(t) = \exp(-t^2/2)$, uniformly in t. The Lévy-Cramér Theorem (Theorem 3) implies now eq. (4.2) and the proof of Theorem 4 is complete. □

5. Discussion

Let us begin with the case $k = 2$, i.e. RNA secondary structures. Sequence to structure mappings into secondary structures have been analyzed by Schuster *et.al.* [11] where detailed asymptotics has been derived. To our knowledge our central limit theorems are new results even for canonical secondary structures. They have been observed [23] in data produced by the folding algorithm Vienna RNAfold [10].

Canonical RNA pseudoknot structures exhibit significantly smaller growth rates than pseudoknot structures with isolated bonds as shown in Table 2 and discussed in the Introduction. They have been studied in [16] where their asymptotic numbers are derived. In Table 2 we provide a complete overview of all relevant growth rates indexed by k and τ, i.e. the maximal number of crossing arcs, $k-1$ and the minimum stack-size, τ. For instance, for $\mathsf{T}_{3,2}(n)$ and $\mathsf{T}_{4,2}(n)$, the numbers of canonical 3- and 4-noncrossing pseudoknot structures we have

$$\mathsf{T}_{3,2}(n) \sim \frac{311.2470 \cdot 4!}{n(n-1)\cdots(n-4)}\, 2.5881^n \quad \text{and} \quad \mathsf{T}_{4,2}(n) \sim 1.217 \cdot 10^7 n^{-\frac{21}{2}}\, 3.0382^n \ . \tag{5.1}$$

In other words: there are less canonical 3-noncrossing structures (2.5881) than arbitrary secondary structures (2.6180). Furthermore it is remarkable that 6-noncrossing canonical pseudoknot structures still grow at a growth rate of less than 4: this structure class allows for very complex pseudoknot configurations.

One important implication of the central limit theorems is that the numbers of arcs of canonical secondary structures are concentrated at $0.3172\, n$ and for 3-(4-)noncrossing, canonical pseudoknot structures at $0.3817\, n (0.4035\, n)$, respectively. This concentration result proves two nontrivial points: (a) the existence of neutral networks of any folding map into the latter structure classes and (b) nontrivial sequence to structure maps. Indeed, first they immediately imply that neutral networks are exponentially small compared to sequence space over the natural alphabet: there are only 6 choices for base pairs in Watson-Crick and **G-U** pairs as opposed to $4^2 = 16$ choices for two unpaired positions. Our results show that a certain fraction of positions in a pseudoknot structure is paired and hence restrict the size of the set of sequences which fold into it. Secondly, since only exponentially small subsets of sequence space can fold into a particular canonical structure, the number of canonical structures of a typical sequence to structure map grows exponentially.

We shall proceed by discussing a structural difference between the parameters τ and k w.r.t. the central limit theorems. Lemma 4 and Lemma 5 show that the minimum stack-size τ only appears as a parameter of the inner function $\psi_\tau(z,s)$:

$$\psi_\tau(z,s) = \frac{\sqrt{\frac{e^s(e^sz^2)^{\tau-1}}{1-e^sz^2+(e^sz^2)^\tau}z}}{\frac{e^s(e^sz^2)^{\tau-1}}{1-e^sz^2+(e^sz^2)^\tau}z^2 - z + 1}\ .$$

Therefore, varying the minimum stack-size, τ, does only result in a shift of the singularity but does not change its type. In contrast the crossing number k, does affect *both*: type and location of the singularity, since for different k we have different singular expansions and singularities ρ_k. Most remarkably, for odd k there exists a logarithmic term in the singular expansion of the generating function which controls the distribution (Lemma 5):

$$F_k(z) = \begin{cases} O((z-\rho_k)^{(k-1)^2+(k-1)/2)-1}\ln(z-\rho_k)) & \text{for } k \text{ odd, } z \longrightarrow \rho_k \\ O((z-\rho_k)^{(k-1)^2+(k-1)/2)-1}) & \text{for } k \text{ even, } z \longrightarrow \rho_k. \end{cases}$$

This shows how the concept of k-noncrossing, τ-stable RNA structures generalizes that of secondary structures. The crossing number k alone controls the class of structures. A minimum stack-size larger than 2 then leads to structure classes with moderate growth rates. Our results imply furthermore explicit formulas for all singularities for arbitrary k and τ.

Table 1 shows another intriguing feature: for $k > 2$ there is a unique minimum for the mean number of arcs for $\tau = 2$. Only for $k = 2$ we observe a monotone increase of μ as a function of τ. This results from the fact that canonicity enforces stacking which is "by nature" antagonistic to the complex crossing motifs, responsible for the high number of unrestricted pseudoknot structures. This is intuitively clear: unrestricted structures can "pack" arcs more densely because for stacks restrictive sequence symmetries are needed. We remark that canonical structures represent an exponentially small subset of all unrestricted structures. I.e. we sample a set of practically "zero" measure and increasing τ the sets becomes smaller and smaller. μ then increases with τ since one considers smaller and smaller subsets that are by definition more and more densely packed. $\mu = 0.5$ would correspond to a perfect matching, i.e. a structure in which all bases are paired. In view of this we observe a key difference between pseudoknot and secondary structures: the former have significantly less unpaired bases. Of course this has to be considered in context with energy parameters, see Figure 5, where we show that this finding remains valid for minimum free energy pseudoknot structures.

A generic feature of folding algorithms into pseudoknot RNA structures [21] is that they cannot control the number of mutually crossing arcs: a consequence of employing dynamic programming paradigms. One look at Table 2 shows that these algorithms necessarily generate structure classes which grow at rates much larger than 4. As a result only small subsets of structures can be realized by folding maps as there are simply not enough sequences. Our analysis suggests to consider designing folding maps into k-noncrossing, canonical structures. It offers the prospect of deriving such algorithms in the near future as the absolute growth rates are small. In this context the number of bonds in canonical pseudoknot RNA is of interest since the number of bonds entails central information about the minimum free energy. Indeed, for canonical structures bonds can only occur in stacks where main energy contributions originate. Therefore our findings quantify how much lower free energies canonical pseudoknot RNA achieve. In addition our analysis shows that increasing the crossing number, k, does not linearly increase the mean number of bonds, see Table 1. This suggests that 3-noncrossing canonical structures (exhibiting an exponential growth rate of 2.5881) are a class of structures which could serve as a paradigm for RNA pseudoknot structures. We remark that already 3-noncrossing canonical structures contain "motifs" that cannot be assigned an energy-value.

Acknowledgments. We are grateful to Emma Y. Jin, Jing Qin and Rita Wang for their help. 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

[1] André. D., 1887. Solution directed du probleème, résolu par M. Bertrand. C. R. Acad. Sci. Paris, 105, 436–437.

[2] Batey. R. T., Rambo. R. P., Doudna. J. A., 1999. Tertiary Motifs in RNA Structure and Folding Angew. Chem. Int. Ed., 38, 2326-2343.

[3] Chamorro. M., Parkin. N., 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. Jan 15;89 (2):713-7 1309954

[4] 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, No. 4, 1555–1575.

[5] Chen, W.Y.C., Qin. J., Reidys, C.M., 2007d. Crossings and Nestings of tangled-diagrams. arXiv:0710.4053v2

[6] Flajolet. P., Fill. J. A., Kapur. N. 2005. Singularity analysis, Hadamard products, and tree recurrences J. Comp. Appl. Math. 174, 271-313.

[7] Flajolet. P., Sedgewick. R., 2007. Analytic combinatorics,

[8] Gessel. I. M., Zeilberger. D., 1992. Random walk in a Weyl chamber, Proc. Amer. Math. Soc. 115 27–31.

[9] Grabiner. D.J., Magyar. P., 1993. Random walks in Weyl chambers and the decomposition of tensor powers Discr. Appl. Math. 2, 239-260.

[10] Hofacker. I.L., Fontana. W., Stadler. P.F., Bonhoeffer. L.S., Tacker. M., Schuster. P., 1994. Fast Folding and Comparison of RNA Secondary Structures. Monatsh. Chem. 125: 167-188.

[11] Hofacker. I.L., Schuster. P., Stadler. P.F., 1998. Combinatorics of RNA Secondary Structures, Discr. Appl. Math., 88, 207-237.

[12] Howell. J.A., Smith. T.F., Waterman. M.S., 1980. Computation of generating functions for biological molecules. SIAM J. Appl. Math. 39, 119-133.

[13] Jin. E.Y., Qin. J., Reidys. C.M., 2007a. Combinatorics of RNA structures with pseudoknots. Bull. Math. Biol. **70**(2008), 45-67, PMID: 17896159.

[14] Jin. E.Y., Reidys. C.M., 2007b. Asymptotic enumberation of RNA structures with pseudoknots. Bull. Math. Biol. (2007), DOI 10.1007/s11538-007-9265-2.

[15] Jin. E.Y. Reidys. C.M., 2007c. Central and Local Limit Theorems for RNA Structures J. Theoret. Biol. **250**(2008), 547-559.

[16] Jin. E.Y., Reidys, C.M., 2007e RNA-Lego: Combinatorial Design Of Pseudoknot RNA, arXiv:0711.1405v2

[17] Jin, E.Y., Reidys. C.M., Wang. R., 2008. Asymptotic analysis of *k*-noncrossing matchings, arXiv:0803.0848

[18] Konings. D.A.M., Gutell. R.R., 1995. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. RNA 1, 559-574.

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

[20] 2005. Mapping RNA Form and Function. Science, 2.

[21] Rivas. E., Eddy. S., 1999. A Dynamic Programming Algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 285, 2053-2068.

[22] Stanley. R.P., 1980. Differentiably finite power series. Europ. J. Combinatorics 1, 175-188.

[23] Stadler. P.F., private communication 2007.

[24] Titchmarsh. E.C., 1939. The theory of functions. Oxford University Press, London.

[25] Tuerk. C., MacDougal. S., Gold. L., 1992. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. Proc. Natl. Acad. Sci. USA, 89:6988-6992.

[26] Wasow. W., 1987. Asymptotic expansions for ordinary differential equations, Dover, A reprint of the John Wiley edition, 1965

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

[28] Waterman. M.S., 1979. Combinatorics of RNA hairpins and cloverleafs. Stud. Appl. Math. 60, 91-96.

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

[30] Westhof. E., Jaeger. L., 1992. RNA pseudoknots. Current Opinion Struct. Biol. 2, 327-333.

Figure Captions

Figure 1

k-noncrossing and canonical: top diagram: the red/blue/green arcs mutually cross and the arcs $(1,5)$ and $(2,6)$ are isolated. Accordingly, this is a 4-noncrossing, $\tau = 1$ diagram without isolated vertex. Bottom diagram: 3-noncrossing (no red/green cross), $\tau = 2$ (canonical) diagram with isolated vertex 6.

Figure 2

Diagram representation of the hammerhead ribozyme [2]. Two tertiary interactions are shown in green arcs. The gap after **C**25 indicates that some nucleotides are omitted, which are involved in an unrelated structural motif.

Figure 3

Counting RNA pseudoknot structures: from diagrams to tableaux-sequences and then to walks. The enumeration is non-constructive and based on the reflection-principle. Here we choose $(1,0)$ as start and endpoint of the walk. The resulting walk does not touch the walls $x = y$ and $x = -1$.

Figure 4

Central limit theorems versus exact enumeration data for sequences of length $n = 200$ for canonical 2-, 3-, and 4-noncrossing RNA pseudoknot structures. We display the asymptotic arc-length distributions (solid curves: red/blue/green) and actual frequencies (dots) computed for $n = 200$.

Table 1

Central limit theorems for $\mathsf{T}_{k,\tau}(n,h)$ any k and τ. We list mean (μ) and variance (σ^2). The mean drops for pseudoknot RNA from $\tau = 1$ to $\tau = 2$ for $k > 2$. This indicates that canonical pseudoknot structures have less arcs.

Figure 5

Distribution of arc-numbers of canonical minimum free energy pseudoknot structures. We display the arc-frequency distributions for $k = 2, 3$ and $n = 40$, for uniform sequences (lhs) i.e. sequences in which the ratios of all nucleotides are equal and random sequences (rhs). The shift in distributions indicates that pseudoknot structures achieve lower minimum free energies than secondary structures.

Figure 6

Basic diagram types: (a) a matching ($f_3(8, 0)$), (b) partial matching with 1-arc $(5, 6)$ and isolated points $2, 7$ ($f_3(8, 2)$), (c) structure (i.e. minimum arc-length ≥ 2) with minimum stack-length 2 and no isolated point ($\mathsf{T}_{3,2}(8)$) and (d) structure with minimum stack-length 3 and isolated points $1, 5$ ($\mathsf{T}_{2,3}(8)$).

Table 2

The exponential growth rates for some important classes of pseudoknot RNA [16]: $\tau = 1$ corresponds to structures with isolated arcs, $\tau = 2$ are canonical structures. Increasing τ means to have larger and larger minimum stack-sizes.

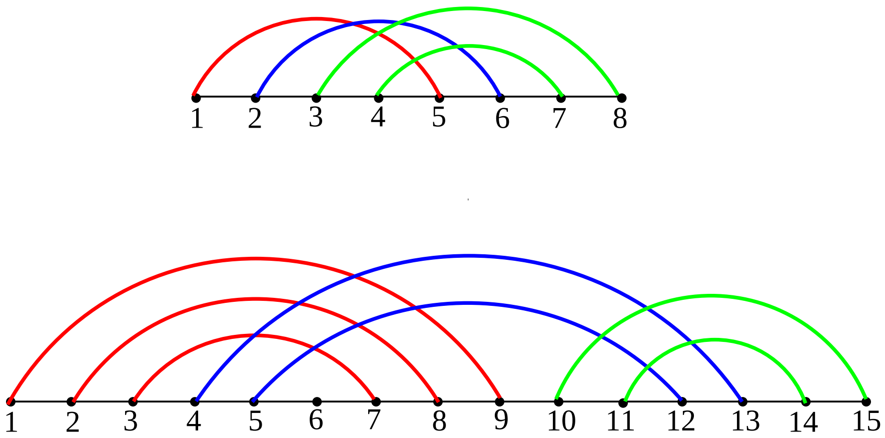

Figure 1.

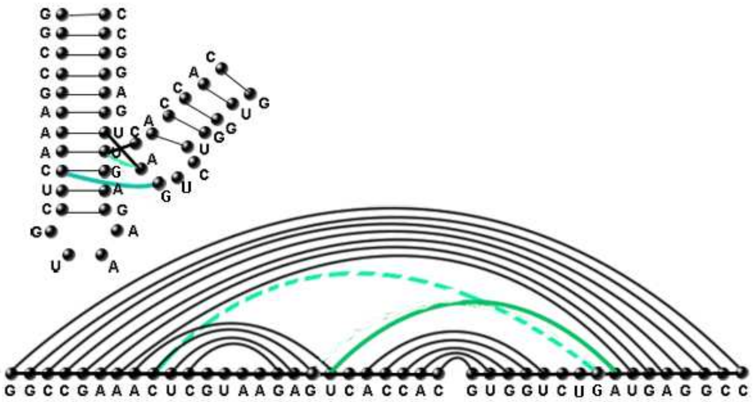
G G C C G A A A C U C G U A A G A G U C A C C A C G U G G U C U G A U G A G G C C

Figure 2.

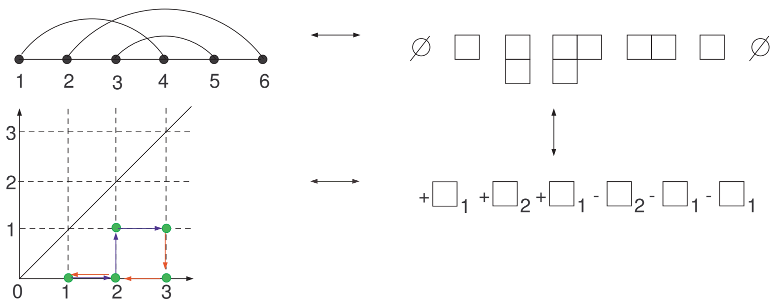

Figure 3.

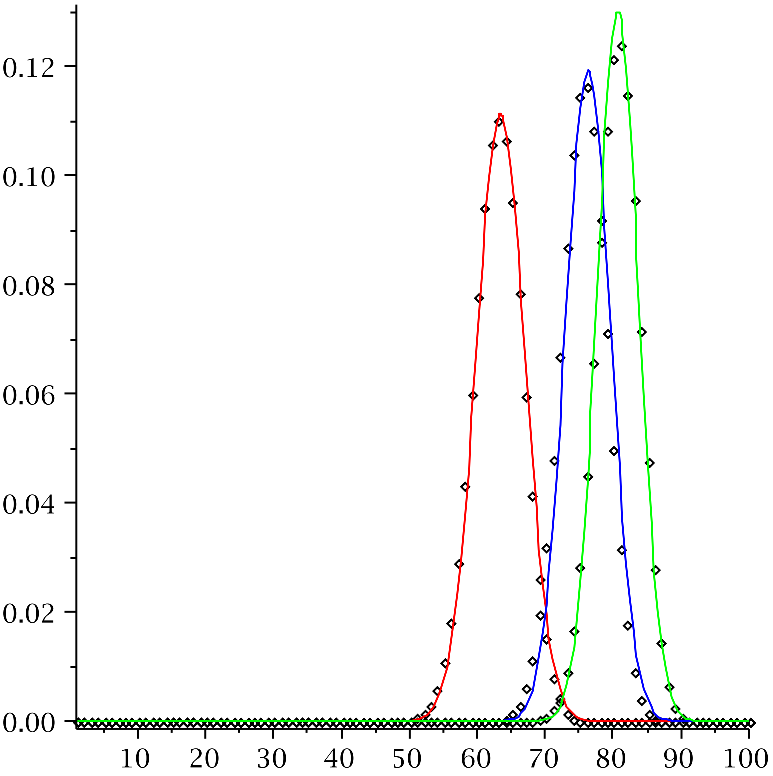

Figure 4.

	$k = 2$		$k = 3$		$k = 4$	
	μ	σ^2	μ	σ^2	μ	σ^2
$\tau = 1$	0.276393	0.0447214	0.390891	0.0415653	0.425464	0.0314706
$\tau = 2$	0.317240	0.0643144	0.381701	0.0559928	0.403574	0.0470546
$\tau = 3$	0.336417	0.0791378	0.383555	0.0670987	0.400288	0.0559818
$\tau = 4$	0.348222	0.0916871	0.386408	0.0767872	0.400412	0.0667094
$\tau = 5$	0.356484	0.1028563	0.389134	0.0855937	0.401402	0.0748305
$\tau = 6$	0.362717	0.1130777	0.391573	0.0937749	0.402640	0.0823440
$\tau = 7$	0.367658	0.1225974	0.393733	0.1014803	0.403908	0.0894075
	$k = 5$		$k = 6$		$k = 7$	
	μ	σ^2	μ	σ^2	μ	σ^2
$\tau = 1$	0.443020	0.0251601	0.453775	0.0209395	0.461750	0.0179291
$\tau = 2$	0.416068	0.0413361	0.424531	0.0373179	0.430788	0.0342976
$\tau = 3$	0.410087	0.0517052	0.416860	0.0474929	0.421957	0.0443150
$\tau = 4$	0.408701	0.0603242	0.414487	0.0558238	0.418872	0.0524231
$\tau = 5$	0.408741	0.0680229	0.413886	0.0632201	0.417800	0.0595864
$\tau = 6$	0.409306	0.0751211	0.413996	0.0700206	0.417575	0.0661575
$\tau = 7$	0.410071	0.0817830	0.414421	0.0763943	0.417747	0.0723092

Table 1.

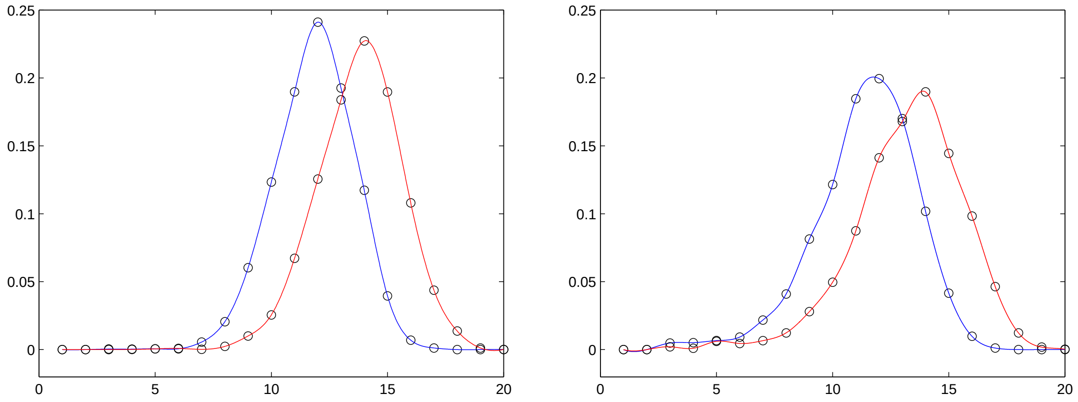

Figure 5.

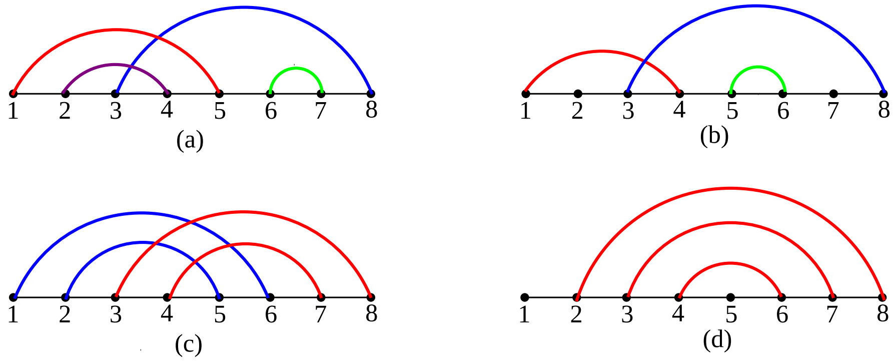

Figure 6.

k	2	3	4	5	6	7	8	9	10
$\tau = 1$	2.6180	4.7913	6.8541	8.8875	10.9083	12.9226	14.9330	16.9410	18.9472
$\tau = 2$	1.9680	2.5881	3.0382	3.4138	3.7438	4.0420	4.3162	4.5715	4.8115
$\tau = 3$	1.7160	2.0477	2.2704	2.4466	2.5955	2.7259	2.8427	2.9490	3.0469
$\tau = 4$	1.5782	1.7984	1.9410	2.0511	2.1423	2.2209	2.2904	2.3529	2.4100
$\tau = 5$	1.4899	1.6528	1.7561	1.8347	1.8991	1.9540	2.0022	2.0454	2.0845
$\tau = 6$	1.4278	1.5563	1.6368	1.6973	1.7466	1.7883	1.8248	1.8573	1.8866
$\tau = 7$	1.3815	1.4872	1.5528	1.6019	1.6415	1.6750	1.7041	1.7300	1.7533
$\tau = 8$	1.3454	1.4351	1.4903	1.5314	1.5645	1.5923	1.6165	1.6378	1.6571
$\tau = 9$	1.3164	1.3941	1.4417	1.4770	1.5054	1.5291	1.5497	1.5679	1.5842
$\tau = 10$	1.2925	1.3610	1.4028	1.4337	1.4585	1.4792	1.4971	1.5129	1.5270

Table 2.