# The Expected Order of Saturated RNA Secondary Structures

Emma Yu Jin* Markus E. Nebel†

Department of Computer Science
University of Kaiserslautern
67663 Kaiserslautern, Germany
Email: {jin,nebel}@cs.uni-kl.de

**Abstract**

Regarding so-called *hairpin-loops* as the building blocks of a RNA secondary structure, the *order* (as introduced by Waterman as a parameter on graphs in 1978) provides information on the (balanced) nesting-depth of hairpin-loops and thus on the overall complexity of the structure.

Subsequent to Waterman's seminal work, Zucker *et al.* and Clote introduced a more realistic combinatorial model for RNA secondary structures, the so-called *saturated secondary structures*. Compared to the traditional model of Waterman, unpaired nucleotides (vertices) which are in favorable position for a pairing do not exist, i.e. no base pair (edge) can be added without violating at least one restriction for the graphs. That way, one major shortcoming of the traditional model has been cleared. However, the resulting model gets much more challenging from a mathematical point of view. As a consequence, the current state of knowledge concerning saturated structures is limited to (1) their asymptotic number, (2) the expected number of base pairs, (3) the asymptotic normal density of states [4]. Nothing is known about the nature of the branching topology of saturated structures – a topic that the current paper completely solves.

In this paper we show how it is possible to attack saturated structures and especially how to analyze their order. This is of special interest since in the past it has been proven to be one of the most demanding parameters to address (for the traditional model it has been an open problem for more than 20 years to find asymptotic results for the number of structures of given order and similar). We show the expected order of RNA saturated secondary structures of size $n$ is $\log_4 n \left(1 + \mathcal{O}\left(\frac{1}{\log_2 n}\right)\right)$, if we select the saturated secondary structure uniformly at random. Furthermore, the order of saturated secondary structures is sharply concentrated around its mean. As a consequence saturated structures and structures in the traditional model behave the same with respect to the expected order. Thus we may conclude that the traditional model has already drawn the right picture and conclusions inferred from it with respect to the order (the overall shape) of a structure remain valid even if enforcing saturation (at least in expectation).

## 1 Introduction

Over the last 30 years the development of RNA secondary structure prediction algorithms have been guided and inspired by corresponding combinatorial studies where the RNA molecules are modeled as certain kind of planar graphs. The other way round, new algorithmic ideas gave rise to interesting combinatorial problems asking for a deeper understanding of the structures processed. The building blocks of RNA are four different nucleotides $\{a, c, g, u\} = \Sigma$ which are linked to each other in a linear fashion. Accordingly, the so-called primary structure of RNA, i.e. the linear sequence of building blocks, is modeled as string over $\Sigma$. In addition, non-neighboring nucleotides have a second means of binding by which certain combinations of nucleotides ($a - u$, $c - g$ and $g - u$) may form pairs, i.e. stick to each other. This gives rise to a 3D folding of the molecule which in many cases determines its biological function. Each such pair reduces the so-called free energy of the molecule and the conformation of minimal free energy is adopted in nature. Today, lab techniques to determine the primary structure of RNA are cheap and efficient while determining the 3D structure still is a time-consuming and expensive task. Accordingly one aims for algorithms to predict the structure from the sequence. However, Lyngso and Pedersen [12] have shown that determining the minimum energy pseudoknotted structure is NP-complete, when the energy model is a form of nearest neighbor model. As a consequence, the set of considered structures is constrained and so-called secondary structures are considered as the first step towards understanding RNA biological function. There, only non-crossing pairings of nucleotides are allowed such that – ignoring types of nucleotides – the

*The work of this author has been supported by the Alexander von Humboldt Foundation by a postdoctoral research fellowship.

†Author to whom correspondence should be addressed.

molecule can be represented as a planar graph [18] (see Figure 1) or alternatively by strings over $\{.,(,)\}$ where a . represents an unpaired nucleotide and a pair of corresponding brackets represents two paired nucleotides (the left structure of Figure 1 is in correspondence with ((..(((.....))).)) ). Even if computing a structure of minimum free energy (mfe) becomes efficient for secondary structures (algorithms with cubic time-bounds are well-known), the empiric thermodynamic data used are incomplete and erroneous such that suboptimal solutions need to be taken into consideration [13]. Computing the suboptimal structures is not difficult, however, the number of potentially interesting suboptimal conformation grows exponentially with the length of the nucleotide sequence. As one possible solution, Zucker suggested to restrict secondary structure folding to structures whose stacking regions (runs of consecutive brackets) extend maximally in both directions. This led to the definition of saturated structures for which no base pair can be added without violating the restrictions for secondary structures, see Figure 1. Please note that these two notions are not equivalent since there are structures for which one can add a base pair, but for which one cannot extend a stacking region. Extending the runs

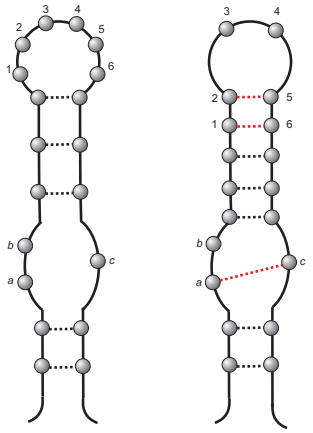


Figure 1: Secondary structure (left) and its saturated counterpart (right) where three additional links have been added (highlighted in red). The primary structure is given by the chain of vertices along the solid, pairs of nucleotides are represented by dotted edges. Note that 3 and 4 cannot be paired since both are neighbored with respect to primary structure.

of consecutive brackets clears one major shortcoming of the traditional model, i.e. of secondary structures, which compared to native molecules tends to have way too short stacking regions. Furthermore, in light of the asymptotic number of saturated structures determined by Clote *et al.* [4], the run time of RNA prediction algorithm should be substantially reduced if the search for suboptimal foldings is limited to saturated structures only, as observed by Bompfünewerer *et al.* for so-called canonical structures [1].

Clote initiated the combinatorial study of saturated structures [3] which gets much more challenging than that for secondary structures from a mathematical point of view. He estimated the number of saturated structures by applying implicit function theory to the functional equations of its generating function $S(z)$ [4], i.e.,

$$-S(z)^3z^4-S(z)^2z^2(-2+z^2)+S(z)(-1+z^2)+z(1+z)=0,$$

whereas the functional equation for secondary structures is relatively simple and given by

$$T(z)=z+zT(z)+z^2T(z)+z^2T(z)^2,$$

for $T(z)$ the generating function of secondary structures. Of course we observe variations of local parameters of the structures like the length and number of stacking regions or the length and number of loops (runs of symbol '.'). However, it is not at all obvious whether saturation has an effect on the overall shape of the structures. One parameter which allows to measure their overall shape is the so-called *order*, originally introduced by Waterman in 1978 for algorithmic purposes. A secondary structure $s$ (saturated or not, represented in dot-bracket form) has order $p$ if we need exactly $p$ iterations of deleting all maximal substrings $(^k)^k$ within $\phi(s)$ in parallel to find the empty string $\varepsilon$. Here $\phi$ is the homomorphism implied by $\phi('(')='(', \phi(')')=')'$ and $\phi('.')=\varepsilon$. Accordingly, the order provides information on the (balanced) nesting-depth of so-called hairpin-loops (substring with $\phi$-image $(^n)^n$ which e.g. holds for the structures depicted in Figure 1) and thus on the overall complexity of the structure (it was used by algorithms to increasingly consider more and more complex foldings starting with a search space restricted to structures of order 1).

In this paper we show one way to approach the combinatorics of saturated structures and especially how to analyze their order. This – besides the motivating remarks from above – is of special interest since in the past it has been proven to be one of the most demanding parameters to address (for secondary structures it has been an open problem for more than 20 years to find asymptotic results for the number of structures of given order and similar). For that purpose we discuss the generating function of saturated structures having order $\geq p$, denoted by $S_p(z)$, from which we extract the information of the expected order of a saturated structure of given size. We find that in expectation the order behaves the same for secondary and saturated structures such that we may conclude that the traditional model

(secondary structures) has already drawn the right picture and conclusions inferred from it with respect to the order (the overall shape) of a structure remain valid even if enforcing saturation (at least in expectation).

The paper is organized as follows. We first present our main results. Afterwards we describe a streamlined analysis with details delayed till the last sections (or the appendix available at the corresponding author's homepage).

## 2 Main Results

Let $S(n)$ be the number of saturated RNA secondary structures of size $n$ and $S_p(n)$ (resp. $S_{=p}(n)$) be the number of saturated RNA secondary structures of size $n$ and having order $\geq p$ (resp. exactly order $p$), then we set $\xi_n$ to be the random variable having probability distribution

$$\mathbb{P}(\xi_n = p) = \frac{S_p(n) - S_{p+1}(n)}{S(n)},$$

namely we select each saturated structure uniformly at random among the family of saturated RNA secondary structures of size $n$. Our main results are summarized as

THEOREM 2.1. (MAIN THEOREM) *The expected order of a saturated secondary structures of size $n$ is*

$$\mathbb{E}\xi_n = \log_4 n \cdot \left(1 + \mathcal{O}\left(\frac{1}{\log_2 n}\right)\right).$$

Our Main Theorem indicates that although the saturation of secondary structures increases the expected number of paired bases (and therefore increases the number of hairpin-loops possible) and scales down the search space, the complexity of the folding algorithm for saturated structures as given by the order stays almost the same. We refer the reader to [14] where it has been shown that the expected order of ordinary secondary structures is asymptotically given by $\frac{1}{2}\log_2\left(\frac{2\pi^2}{\rho^2}n\right) - \frac{\gamma+2}{2\ln(2)} + \Delta\left(\log_2\left(\frac{n}{\rho^2}\right)\right) + \mathcal{O}(n^{-1})$, $\Delta(x)$ a periodic function of very small modulus ($|\Delta(x)| \leq 0.040597\ldots$) and known expansion as a Fourier series. We may conclude that the traditional secondary structure model has already drawn the right picture and conclusions inferred from it with respect to the order of a structure (its overall shape) remain valid even if enforcing saturation (at least in expectation).

Theorem 3.3 below proves $\xi_n$ to be highly concentrated around the expected order $\mathbb{E}\xi_n$.

THEOREM 2.2. *Assume we choose $0 \leq x \leq (\frac{1}{2} - \beta)\log_4 n$ for arbitrary $\beta > 0$, then we have*

$$\mathbb{P}(|\xi_n - \mathbb{E}\xi_n| \geq x) = \mathcal{O}(2^{-x}).$$

It is worthwhile to mention Drmota's theorem [5] which makes it possible to prove that the number of base pairs of saturated secondary structures with fixed size $n$ is Gaussian distributed with mean approximately $0.337361n$ and variance $0.017636n^2$. The proof is based on a system of functional equations for saturated structures $S(z,u)$ having each base labeled by $z$ and each base pair labeled by $u$ [4]. However, even if this suggests it should be possible to apply Drmota's theorem to show the distribution of base pairs of saturated structures having fixed order to be asymptotically normal, we so far do not have an appropriate functional identity of generating function $S_{=p}(z) = \sum_{n\geq 0} S_{=p}(n)z^n$ since the family of saturated structures having fixed order $p$ is not closed under decomposition.

## 3 Road Map of the Proof

In this section, we shall address the major steps and difficulties of analyzing the expected order of saturated structures by tools from analytic combinatorics [6, 8]. We start by deriving the key recursions for saturated structures of order $p$.

Let $S(z)$ (resp. $\mathcal{S}$) be the generating function (resp. the family) of saturated RNA structures and $R(z)$ (resp. $\mathcal{R}$) be the generating function (resp. the family) of saturated structures having the first and the last position paired, i.e., $\mathcal{R} = (\mathcal{S})$ where the parenthesis represents the paired bases and $R(z) = z^2S(z)$. Furthermore, let $S_p(z)$ and $R_p(z)$ represent the corresponding generating function assuming order $\geq p$, $p \geq 1$. By decomposing the saturated structure into independent $\mathcal{R}$-type structures, we obtain the functional equation for $S(z)$, see Figure 2 for details.

$$\begin{aligned} S(z) &= \sum_{i=0}^{\infty}\left(1 + (i+1)(z+z^2)\right)R(z)^i - 1 \\ &= \frac{z^2S(z)}{1-z^2S(z)} + \frac{z+z^2}{(1-z^2S(z))^2}. \end{aligned} \tag{3.1}$$

Now, taking the order into account (omitting variable

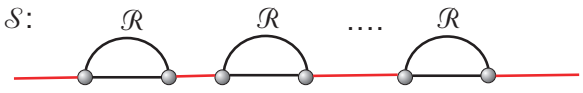


Figure 2: The decomposition of saturated structure $\mathcal{S}$ into independent $\mathcal{R}$-type structures: For any saturated structure $\mathcal{S}$ having $i$ ($i \geq 1$) $\mathcal{R}$-type structures, at most one of the $i+1$ intervals (colored in red) are allowed to add unpaired bases of at most 2, which results the term $\left(1 + (i+1)(z+z^2)\right)$ in eq. (3.1).

$z$ for the ease of notation), see Figure 3 for details, we

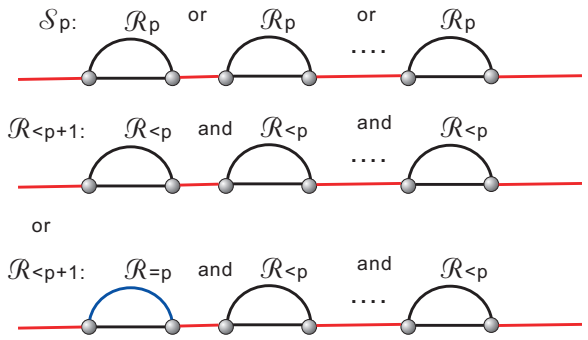


Figure 3: Top: The decomposition of saturated structure $\mathcal{S}_p$ into independent $\mathcal{R}_p$-type structures: For any saturated structure $\mathcal{S}$ having order $\geq p$, at least one of its $\mathcal{R}$-type structures must have order $\geq p$, which results the term $(1+(i+1)(z+z^2))(R^i-(R-R_p)^i)$ in eq. (3.2). Middle and bottom: The decomposition of $\mathcal{R}_{<p+1}$-type structures: For any $\mathcal{R}$-structures having order $< p+1$, at most one of its sub $\mathcal{R}$-structures having exactly order $p$, which contributes to the term $(R-R_p)^i+(R_p-R_{p+1})i\times(R-R_p)^{i-1}$.

find the following recurrences for $S_p$ and $R_{p+1}$, $p\geq 1$,

$$\begin{aligned} S_p &= \sum_{i\geq 1}(1+(i+1)(z+z^2))(R^i-(R-R_p)^i) \\ &= \frac{R_p}{(R-1)^2(R-R_p-1)^2} \\ &\quad\times[1+2z^2+2z-2R-2Rz-2Rz^2+R^2 \\ &\quad+(1+z+z^2-R)R_p]. \end{aligned} \tag{3.2}$$

Let $\mathcal{R}_{<p}$ be the $\mathcal{R}$-type structures having order $< p$, and see Figure 3 for the decomposition of $\mathcal{R}_p$-type structures.

$$\begin{aligned} R_{p+1} &= R-z^2\left[\sum_{i=1}^{\infty}(1+(i+1)(z+z^2))\left((R-R_p)^i\right.\right. \\ &\quad\left.\left.+(R_p-R_{p+1})i\times(R-R_p)^{i-1}\right)+z+z^2\right] \\ &= \frac{(-R-z^2)R_p^3}{\eta} \\ &\quad+\frac{(-3R+3R^2+3Rz^2-z^2)R_p^2}{\eta}, \end{aligned} \tag{3.3}$$

where $\eta=-R_p^3+(3R-3)R_p^2+(6R-3-3R^2+z^2)R_p+(R-1)P_R$, $P=R^3+(z^2-2)R^2+(1-z^2)R-z^3-z^4$ and $P_R=\partial P/\partial R=3R^2+2(z^2-2)R+(1-z^2)$ and the initial conditions are $R_1=R$ and $S_0=S$.

Unlike for secondary structures[1], due to the non-local dependencies imposed for saturation neither the appropriate symbolic substitution nor the closed form solution of recurrence (3.2) could possibly exist, for which we have to decode the information of expected order from the recurrence itself other than attempting to solve it. Therefore, the proof for the expected order of saturated structures consists of locating the dominant singularities of $S_p(z)$ for $p\geq 0$, verifying the analytic continuation of $S_p(z)$ for some $\Delta$-domain, which guarantees the validness of integration along Hankel contour, see Figure 4, and finding the singular expansion of $S_p(z)$ within the intersection of $\Delta$-domain and a small neighborhood of the dominant singularity. Finally we apply a transfer theorem on the singular expansions of $S_p(z)$ and $S(z)$ to extract the $n$-th coefficient of $\sum_{p\geq 1}S_p(z)$ and $S(z)$, and conclude the expected order $\mathbb{E}\xi_n$ via

$$\mathbb{E}\xi_n=\frac{[z^n]\sum_{p\geq 1}S_p(z)}{[z^n]S(z)}.$$

The results on the deviation to the expected order follows similarly.

Before we proceed, we present the Transfer Theorem by Flajolet and Odlyzko [8]. The central point of this theorem is to use Cauchy's formula by integrating along the Hankel contour depicted in Figure 4, which is guaranteed by the analytic continuation within a $\Delta$-domain. We set

$$\Delta_{z_0}(M,\phi)=\{z\mid |z|<M, z\neq z_0, |\arg(z-z_0)|>\phi\}$$

where $M>z_0$ and $0<\phi<\frac{\pi}{2}$. Let $U_{z_0}(r,\phi)$ be the intersection of $\Delta_{z_0}(M,\phi)$ and the neighborhood of $z_0$, i.e.,

$$U_{z_0}(r,\phi)=\{z\mid 0<|z-z_0|<r, |\arg(z-z_0)|>\phi\}.$$

Let $\mathcal{U}=\{(1-z)^{-\alpha}\lambda(z)^{\beta}:\alpha,\beta\in\mathbb{C}\}$ and $\lambda(z)=\frac{1}{z}\log_2\frac{1}{1-z}$, then we have:

THEOREM 3.1. **(Transfer Theorem)[8]** *Assume that $f(z)$ is analytic at 0 with a singularity at $z_0$, such that $f(z)$ can be continued to some $\Delta_{z_0}(M,\phi)$ domain, and there exists two functions $\sigma$, $\tau$ where $\sigma$ is a finite linear combination of functions in $\mathcal{U}$ and $\tau\in\mathcal{U}$, such that*

$$f(z)=\sigma\left(\frac{z}{z_0}\right)+\mathcal{O}\left(\tau\left(\frac{z}{z_0}\right)\right).$$

*Then we have $[z^n]f(z)=z_0^{-n}\sigma_n+\mathcal{O}(z_0^{-n}\tau_n)$ where $\sigma_n=[z^n]\sigma(z)$ and $\gamma_n=n^{a-1}(\log n)^b$ if $\tau(z)=(1-z)^{-a}\lambda(z)^b$ for $b\in\mathbb{Z}_{\geq 0}$ and $a\notin\mathbb{Z}_{\leq 0}$.*

[1]For secondary structures the expected order has been analyzed by making use of well-known closed form representations of multivariate generating function for binary trees having Horton-Strahler number $p$. By the use of appropriate symbolic substitutions for the different variables the binary trees with Horton-Strahler number $p$ were expanded into the secondary structures of order $p$ and a closed form for the corresponding generating function followed [14].

For the full proof of Theorem 3.1, we refer to Theorem VI.4 at page 393 of the book "Analytic Combinatorics" by Flajolet and Sedgewick [8].

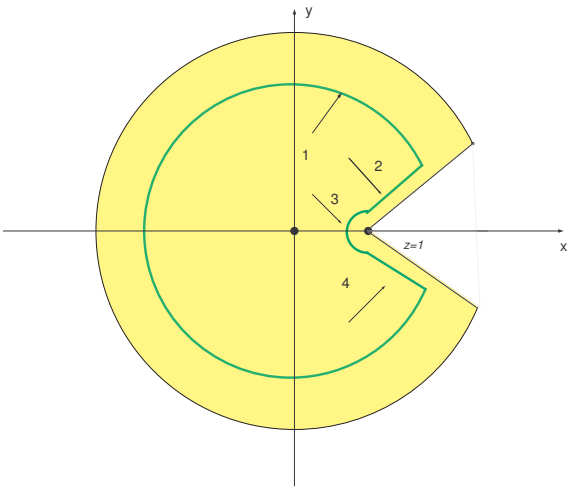


Figure 4: $\Delta_1$-domain (yellow) and Hankel contour (green): Transfer theorem applies Cauchy's formula by integrating along the Hankel contour, colored in green. The inner incomplete circle 3, together with two rectilinear lines 2 and 4 mainly contribute to the integral. Here we assume the dominant singularity is at $z = 1$.

In what follows we detail the steps that are needed for the singularity analysis of $\sum_{p\geq 1} S_p(z)$.

**Step 1: Locate dominant singularities:** We first observe that the dominant singularity of $S(z)$ is unique. Let $z_0$ be the unique dominant singularity of $S(z)$, then we shall show that $z_0$ is also the unique dominant singularity of $S_p(z)$ for $p \geq 0$. Indeed, we can inductively prove that the $R_{<p}(z)$'s are rational functions via the recurrence

$$R_{<p} = \frac{-z^2(R^3_{<p-1} - R^2_{<p-1} - 3aR_{<p-1} + a)}{R^3_{<p-1} - 3R^2_{<p-1} + (3-z^2)R_{<p-1} + b}$$

where $a = z + z^2$ and $b = -1 + z^2 + 2z^3 + 2z^4$, with the initial condition $R_{<1}(z) = -\frac{z^3}{2z^3+z-1}$. Accordingly, the $S_{<p}(z)$'s are rational functions via the recurrence

$$S_{<p} = -\frac{R_{<p} - 1 - z - z^2}{(R_{<p} - 1)^2}.$$

In view of $S_p(z) = S(z) - S_{<p+1}(z)$, we can claim that $S_p(z)$ $(p \geq 0)$ have the same unique dominant singularity as $S(z)$, otherwise, suppose $z = \gamma < z_0$ is the dominant singularity of $S_p(z)$ and therefore $S_p(\gamma) < S_p(z_0) < S(z_0) < \infty$, which contradicts to the fact that $S_p(\gamma) = S(\gamma) - S_{<p+1}(\gamma) = \infty$ since $S_{<p+1}(z)$ is a rational function and $z = \gamma$ must be one of the poles of $S_{<p+1}(z)$. We can conclude that $z_0$ is also the unique dominant singularity of $S_p(z)$.

LEMMA 3.1. *Let $z_0$ be the unique dominant singularity of $S_p(z)$ $(p \geq 0)$, then $z_0 \approx 0.424687$.*

We apply the implicit function theorem on eq. (3.1) to extract the unique dominant singularity of $S(z)$, which is also the unique dominant singularity of $S_p(z)$.

**Step 2: Establish the analytic continuation in some $\Delta_{z_0}$-domain:** Since $S_p(z)$ is an algebraic function of degree 3 over the rational function field $Q(z)$, $S_p(z)$ must be D-finite, which allows for analytic continuation in any $\Delta_{z_0}$-domain containing zero [16].

**Step 3: Singular expansion:** We shall show the singular expansion of $S_p(z)$ within $U_{z_0}(\epsilon, \phi)$ for sufficiently small $\epsilon > 0$ and $0 < \phi < \frac{\pi}{2}$. Our strategy is to transform the fractional form of the recursion for $R_p(z)$ (eq. (3.3)) into "linear" form, based on the contributions of individual terms to the behavior of $R_p(z)$ for different $p$.

**Case 1:** $p \leq p_M = \max\left\{p : |P_R(R-1)| \leq \left|\frac{a_2}{4} \cdot R_p\right|\right\}$ for $a_2 = -3R + 3R^2 + 3Rz^2 - z^2$.

LEMMA 3.2. *The sequence $\{S_p(z_0)\}$ satisfies $S_p(z_0) \to 0$ and $\frac{S_{p+1}(z_0)}{S_p(z_0)} \to \frac{1}{2}$ as $p \to \infty$. Furthermore,*

$$S_{p+1}(z_0) = \left(s + \mathcal{O}(2^{-p})\right) 2^{-p}$$

*where $s > 0$ is a constant.*

LEMMA 3.3. *Suppose $0 < \phi < \frac{\pi}{2}$ and $\epsilon > 0$ is sufficiently small constant, then there exist constants $d_1, d_2$ such that for $z \in U_{z_0}(\epsilon, \phi)$ and $p \leq p_M$,*

$$d_1 \cdot 2^{-p} \leq |S_{p+1}(z)| \leq d_2 \cdot 2^{-p}.$$

LEMMA 3.4. *Setting $s_p(z_0) = 2^p \cdot S_{p+1}(z_0)$ and assuming that $z \in U_{z_0}(\epsilon, \phi)$ and $a_2 = -3R + 3R^2 + 3Rz^2 - z^2$, then*

$$S_{p+1}(z) = s_p(z_0)\, 2^{-p} - \frac{1}{z_0^2} \cdot \frac{P_R(R-1)}{2a_2}$$
$$+\mathcal{O}\left(\frac{p}{2^p}\left|\frac{P_R(R-1)}{a_2}\right|\right) + \mathcal{O}\left(2^p\left|\frac{P_R(R-1)}{a_2}\right|^2\right)$$

*holds for $p \leq p_M$ and $s_p(z_0) = s + \mathcal{O}(2^{-p})$ where $s > 0$ is constant.*

**Case 2:** $p > p_M$. We continue analyzing the recurrence relations for $S_p(z)$ and $R_p(z)$ for $p > p_M$. Let $A'_p = -\frac{R+z^2}{a_2} \cdot R_p^3$ and $B'_p = -\frac{1}{a_2}R_p^3 + \frac{3R-3}{a_2}R_p^2 - \frac{3P_R R_p}{a_2}$. Note

that $A'_p \to 0$ and $B'_p \to 0$ as $p \to \infty$ and $z \in U_{z_0}(\epsilon, \phi)$. Then we simply have

$$R_{p+1}(z) = \frac{R_p^2 + A'_p}{\frac{P_R(R-1)}{a_2} + 2R_p + B'_p}. \tag{3.4}$$

We observe that $B'_p$ and $A'_p$ converge to 0 faster than $R_p$ as $p \to \infty$, and it only remains to determine the major contribution between $R_p$ and $\frac{P_R(R-1)}{a_2}$ from the denominator to the behavior of $R_p$ for different $p$. Here we all reduce the recursions to the function $h(x, \mu, \nu) = \frac{x^2+\mu}{1+2x+\nu}$, from which we can prove $h(x, \mu, \nu) = h(x, 0, 0)+\mathcal{O}(\max\{|\mu|, |\nu|\})$ holds uniformly for $x \neq -\frac{1}{2}$ as $\max\{|\mu|, |\nu|\} \to 0$. In order to asymptotically solve eq. (3.4), we need to avoid $R_p = -\frac{1}{2}\frac{P_R(R-1)}{a_2}$, which may occur when $p$ is sufficiently large. To this aim, we select $\lambda_1 > 0$ and $\lambda_2 > 0$ such that for $p \leq p_M + \lambda_2$, $\left|\frac{P_R(R-1)}{a_2}\right| \leq |R_p|$ and for $p \geq p_M - \lambda_1$, $|R_p| \geq 8\left|\frac{P_R(R-1)}{a_2}\right|$. Lemma 3.5 below shows the "continuity" of the phase transition around $p = p_M$.

LEMMA 3.5. *Assume $z \in U_{z_0}(\epsilon, \phi)$ and $p_0 = p_M - \lambda_1$, then for arbitrary but fixed $\delta \leq \lambda_2$, we have uniformly for $z$ and for $0 \leq k \leq \lambda_1 + \delta$ that,*

$$S_{p_0+k} = \frac{1}{z_0^2} \frac{\frac{P_R(R-1)}{a_2}}{\left(\frac{\frac{P_R(R-1)}{a_2}}{R_{p_0}} + 1\right)^{2^k} - 1} + \mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right|^2\right),$$

*where $a_2 = -3R + 3R^2 + 3Rz^2 - z^2$.*

LEMMA 3.6. *Assume that $z \in U_{z_0}(\epsilon, \phi)$, there exists $\kappa_0 \geq \lambda_2$ such that for $p > p_M + \kappa_0$,*

$$S_{p+1}(z) = \mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right| \exp(-\ln 2 \cdot 2^p)\right).$$

**Step 4: Transfer to coefficients:** It only remains to translate the singular expansion of the function into an asymptotic estimate of its coefficients.

THEOREM 3.2. *The expected order of a saturated secondary structures of size $n$ is*

$$\mathbb{E}\xi_n = \log_4 n \cdot \left(1 + \mathcal{O}\left(\frac{1}{\log_2 n}\right)\right).$$

*Proof.* We first analyze the expectation function $F(z) = \sum_{p\geq 1} S_p(z)$ for $z \in U_{z_0}(\epsilon, \phi)$. According to Lemma 3.4, Lemma 3.5 and Lemma 3.6, we have for $p \geq 1$,

$$\begin{aligned}
&\sum_{p\leq p_M} S_{p+1}(z) \\
=& \sum_{p\leq p_M} s_p(z_0)\, 2^{-p} - \frac{p_M}{z_0^2} \cdot \frac{P_R(R-1)}{2a_2} \\
&+\mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right|\right) \\
=& \sum_{p\geq 1} s_p(z_0)\, 2^{-p} - \frac{p_M}{z_0^2} \cdot \frac{P_R(R-1)}{2a_2} \\
&+\mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right|\right). \\
&\sum_{p>p_M} S_{p+1}(z) \\
=& \sum_{p_M<p\leq p_M+\kappa_0} S_{p+1}(z) + \sum_{p>p_M+\kappa_0} S_{p+1}(z) \\
=& \mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right|\right) \\
&+ \sum_{p>p_M+\kappa_0} \mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right| \exp(-\ln 2 \cdot 2^p)\right) \\
=& \mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right|\right).
\end{aligned}$$

In combination of the cases $p \leq p_M$ and $p > p_M$, we obtain

$$\begin{aligned}
F(z) &= \sum_{p\leq p_M} S_{p+1}(z) + \sum_{p>p_M} S_{p+1}(z) + S_1(z) \\
&= \sum_{p\geq 0} S_{p+1}(z_0) + \left(S(z) - \frac{1}{1-z} - S_1(z_0)\right) \\
&\quad -\frac{p_M}{z_0^2} \cdot \frac{P_R(R-1)}{2a_2} + \mathcal{O}\left(\left|\frac{P_R(R-1)}{a_2}\right|\right).
\end{aligned}$$

Recall that $p_M$ is given by

$$p_M = \max\left\{p : |P_R(R-1)| \leq \left|\frac{a_2}{4} \cdot R_p\right|\right\}$$

and we need to find an appropriate representation for it. For $z \in U_{z_0}(\epsilon, \phi)$, $p_M \approx -\log_2\left|\frac{P_R(R-1)}{a_2}\right|$. By setting $F_0 = F(z_0)$ and recalling $S(z) = S(z_0) - \frac{1}{z_0^2}\frac{P_R(R-1)}{2a_2} +$

$\mathcal{O}\left(1-\frac{z}{z_0}\right)^{\frac{3}{2}}$, we simplify $F(z)$ into

$$\begin{aligned}
F(z) &= F_0 - \frac{1}{z_0^2}\frac{P_R(R-1)}{2a_2} + \mathcal{O}(z-z_0) \\
&\quad -\frac{p_M}{z_0^2}\frac{P_R(R-1)}{2a_2} + \mathcal{O}\left(1-\frac{z}{z_0}\right)^{\frac{3}{2}} \\
&= F_0 + \frac{\log_2\left|\frac{P_R(R-1)}{a_2}\right|}{z_0^2}\frac{P_R(R-1)}{2a_2} \\
&\quad +\mathcal{O}\left(1-\frac{z}{z_0}\right)^{\frac{3}{2}} + \mathcal{O}(z-z_0) \\
&= F_0 + \mathcal{O}\left(1-\frac{z}{z_0}\right)^{\frac{3}{2}} + \mathcal{O}(z-z_0) \\
&\quad +\frac{1}{z_0^2}\frac{P_R(R-1)}{2a_2}\log_2\left(\frac{P_R(R-1)}{a_2}\right) \\
&= F_0 + \mathcal{O}\left(1-\frac{z}{z_0}\right)^{\frac{1}{2}} + \mathcal{O}(z-z_0) \\
&\quad -\sqrt{\frac{P_z(z_0)}{2P_{RR}(z_0)z_0^3}}\sqrt{1-\frac{z}{z_0}}\log_2\left(\frac{1}{1-\frac{z}{z_0}}\right) \\
&\quad +\mathcal{O}\left(1-\frac{z}{z_0}\right)^{\frac{3}{2}}\log_2\left(\frac{1}{1-\frac{z}{z_0}}\right).
\end{aligned}$$

According to Theorem 3.1 (Transfer Theorem), for $n \geq 1$, the expected order of a saturated secondary structure is thus given by

$$\begin{aligned}
\mathbb{E}\xi_n &= \frac{[z^n]F(z)}{[z^n]S(z)} \\
&= \frac{-n^{-\frac{3}{2}}}{\Gamma(-\frac{1}{2})}\cdot\log_2 n\cdot z_0^{-n}\cdot z_0^2\sqrt{\frac{P_z(z_0)}{2P_{RR}(z_0)z_0^3}}n^{\frac{3}{2}}z_0^n \\
&\quad \times\sqrt{\frac{2\pi\,P_{RR}(z_0)}{z_0\,P_z(z_0)}}\left(1+\mathcal{O}\left(\frac{1}{\log_2 n}\right)\right)+\mathcal{O}(n^{-\frac{3}{2}}) \\
&= \log_4 n\cdot\left(1+\mathcal{O}\left(\frac{1}{\log_2 n}\right)\right),
\end{aligned}$$

whence the proof is complete.

Finally we discuss the large deviation of the random variable $\xi_n$.

THEOREM 3.3. *Assume we choose* $0 \leq x \leq (\frac{1}{2}-\beta)\log_4 n$ *for arbitrary* $\beta > 0$*, then we have*

$$\mathbb{P}(|\xi_n - \mathbb{E}(\xi_n)| \geq x) = \mathcal{O}(2^{-x}).$$

*Proof.* For $p \leq \log_4 n$, Lemma 3.4 in combination with the Transfer Theorem implies

$$\mathbb{P}(\xi_n \geq p) = 1 + \mathcal{O}\left(\frac{2^p}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{p}{2^p}\right).$$

For $p > \log_4 n$, Lemma 3.5 indicates that

$$\mathbb{P}(\xi_n \geq p) = \mathcal{O}\left(\exp\left(-\frac{\beta' 2^p}{\sqrt{n}}\right)\right) \quad \text{for } \beta' > 0.$$

Consequently the theorem follows.

## References


[1] A.F. Bompfünewerer, R. Backofen, S.H. Bernhart, J. Hertel, I.L.,Hofacker, P.F. Stadler and S. Will, Variation on RNA folding and alignment: lessons from Banasque. *J. Math. Biol.* **56**(1-2) (2008), 129-144.

[2] T.R. Cech, RNA as an enzyme, *Sci.Am.* **255**(5) (1986), 64-75.

[3] P. Clote, Combinatorics of Saturated Secondary Structures of RNA, *J. Comp. Biol.*, **13**(9) (2006), 1640-1657.

[4] P. Clote, E. Kranakis, D. Krizanc and B. Salvy, Asymptotics of Canonical and Saturated RNA secondary structures, *J. Bioinformatics and Comp. Biol.*, **5** (2009), 869-893.

[5] M. Drmota, Systems of funtional equations, *Random structures and algorithms*, **10** (1999), 103-124.

[6] M. Drmota and H. Prodinger, The register function for $t$-ary trees, *ACM Transactions on Algorithms*, **2** (2006), 318-334.

[7] D.J. Evers and R. Giegerich, Reducing the conformation space in RNA structure prediction, *German Conf. Bioinfom.*, (2001), 118-124.

[8] P. Flajolet and R. Sedgewick, Analytic combinatorics, ISBN-13:9780521898065 Cambridge University Press, 2009.

[9] De Gennes in C. Domb and M.S. Green eds., Phase Transition and Critical Phenomena, **3**, Academic Press, London, 1976.

[10] M. Gô, Statistical Mechanics of Biopolymers and its application to the melting transition of polynucleotides, *J. Phys. Soc. Jpn*, **23** (1967), 597-608.

[11] A.M. Lesk, A combinatorial study of the effects of admitting non-Watson-Crick base pairing and of base composition on the helix-forming potential of polynucleotides of random sequences, *J. Theor. Biol.* **44** (1974), 7-17.

[12] R.B. Lyngsø and C.N.S. Pedersen, RNA Pseudoknot Prediction in Energy Based Models, *Journal of Computational Biology* **7** (2000), 409-427.

[13] D.H. Mathews, J. Sabina, M. Zucker and D.H. Turner, Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure, *J. Mol. Biol.* **288**(5) (1999), 911-940.

[14] Markus E. Nebel, Combinatorial Properties of RNA secondary structures, *J. Comp. Biol.*, **9** (2001), 541-573.

[15] J.M. Pipas and J.E. McMahon, Method for predicting RNA secondary structure, *Proc. Nat. Acad. Sci. U.S.A.* **72** (1975), 2017-2021.

[16] R.P. Stanley, Differentiably finite power series, *Eur. J. Combinator.* **1** (1980), 175-188.

[17] X.G. Viennot, A Strahler bijection between Dyck paths and planar trees, *Discr. Math* **246** (2002), 317-329.

[18] M. S. Waterman, Secondary Structure of Single-Stranded Nucleic Acids, *Adv. in Math. Suppl. Stud.* **1** (1978), 167-212.

[19] M. Zucker, RNA folding prediction: The continued need for interaction between biologists and mathematicians, *Lectures on Mathematics in the Life Sciences* **17**, 87-124. Springer-Verlag, 1986.