Asymptotics of Canonical RNA Secondary Structures

Peter Clote
Biology Department
Boston College
Chestnut Hill, MA 02467 USA
clote@bc.edu

Evangelos Kranakis
School of Computer Science
Carleton University
K1S 5B6, Ottawa, Ontario, Canada
kranakis@scs.carleton.ca

Danny Krizanc
Department of Mathematics
Wesleyan University
Middletown CT 06459, USA
dkrizanc@wesleyan.edu

Abstract

It is a classical result of Stein and Waterman [16] that the asymptotic number $S(n)$ of RNA secondary structures is $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$, where the combinatorial model of RNA concerns a length n homopolymer, such that any base can pair with any other base, subject to the usual convention that hairpin loops must contain at least $\theta = 1$ unpaired bases. The result of Stein and Waterman is proved by developing recursions, using generating functions and applying Bender's theorem [1]. These recursions form the basis to compute the minimum free energy secondary structure for a given RNA sequence, with respect to the Nussinov energy model [15], later extended by Zuker [20] to substantially more complicated resursions for the Turner nearest neighbor energy model [14].

In this paper, we study combinatorial asymptotics for two special subclasses of RNA secondary structures – canonical *and* saturated *structures. Canonical secondary structures are defined to have no lonely (isolated) base pairs. This class of secondary structures was introduced by Bompfünewerer et al. [2], who noted that the run time of Vienna RNA Package is substantially decreased when restricting computations to canonical structures. Here we provide an explanation for the speed-up, by proving that the asymptotic number of canonical RNA secondary structures is $2.1614 \cdot n^{-3/2} \cdot 1.96798^n$. Saturated secondary structures have the property that no base pairs can be added without violating the definition of secondary structure (i.e. introducing a pseudoknot or base triple). In the Nussinov energy model, where the energy for a base pair is -1, saturated structures [3] correspond to kinetic traps. In [3], we showed that the asymptotic number of saturated structures of a length n homopolymer is $1.07427 \cdot n^{-3/2} \cdot 2.35467^n$. In this paper, we show that the expected number of base pairs of* random *saturated structures, generated by a natural stochastic procedure, is*

$$\frac{z^{\theta+1}}{(1-z)^2} e^{\left(-z-\sum_{i=0}^{\theta} \frac{z^i}{i+1}\right)} \left(\int e^{\left(z+\sum_{i=0}^{\theta} \frac{z^i}{i+1}\right)} dz \right).$$

1. Introduction

An RNA secondary structure, formally defined later, is an outerplanar graph (no pseudoknots) with the property that no vertex is incident to more than one edge (no base triples) and that for every chord between vertices i, j, there exist at least $\theta = 1$ many vertices that are not incident to any edge (hairpin requirement). RNA secondary structure is equivalently defined to be a well-balanced parenthesis expression $s_1, \ldots, s_n$, containing left parenthesis (, right parenthesis) , and dot $\bullet$, where if nucleotide i is unpaired then $s_i = \bullet$, while if there is a base pair between nucleotides $i < j$ then $s_i =$ (and $s_j =$) . This representation is known as the *Vienna representation* or *dot bracket notation* (dbn).

In this paper, we are interested in specific classes of secondary structure: *canonical* and *saturated* structures. A secondary structure is canonical [2] if it has no lonely (isolated) base pairs. A secondary structure is saturated [18] if no base pairs can be added without violating the notion of secondary structure. In order to compute parameters like asymptotic value for number of structures, expected number of base pairs, etc. we adopt the *homopolymer* model of Stein and Waterman [16]. By homopolymer, we mean that any position (nucleotide, also known as base) can base-pair with any other position. Since there are steric constraints for RNA to fold back on itself within a hairpin region, following Stein and Waterman we additionally require that every hairpin loop contain at least $\theta = 1$ unpaired bases; i.e. if i, j are base-paired, then $j - i > \theta$.

1.1. Examples of representations of secondary structures

Below, we display the sequence and consensus secondary structure of a selenocysteine insertion (SECIS) sequence taken from the Rfam database [8]. SECIS sequences are responsible for a *re-translation* event, whereby the UGA stop codon does not cause a termination of protein translation, but rather causes the incorporation of selenocysteine (cysteine whose side chain contains selenium instead of sulfur) into the growing polypeptide. This sequence, whose GenBank accession number S79854.1/1605-1666, and its consensus secondary structure in (Vienna) dot bracket notation are given as follows:

```
CACUGCUGAUGACGAACUAUCUCUAACUGGUCUUGACCACGAGCUAGUUCUGAAUUGCAGGG
(((((....(((((((((((((((.....(((....)))..)))))))))))))))...)))))
```

Figure 1 depicts an equivalent representation of the RNA secondary structure, in Feynman linear form.

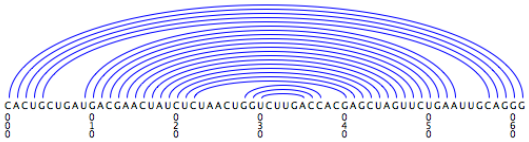

Figure 1. Depiction of SECIS sequence with GenBank accession number S79854.1/1605-1666, represented as a Feynman linear diagram. The sequence and secondary structure were taken from the Rfam Database [8], and the graph was created using `jViz` [17].

1.2. Outline and results of the paper

In Section 2, we review a combinatorial method, known as DSV methodology, which we apply in Section 2.1 to compute the asymptotic number of canonical secondary structures for a homopolymer of length n with $\theta = 1$. Our description of the DSV methodology with its application of the main theorem of Flajolet and Odlyzko [6] is not meant to be self-contained, and we refer the reader to [12] for a detailed overview of this method, along with a number of example applications. Subsequently, we determine the expected number of base pairs for canonical structures; due to lack of space, we can only state the result, proved by similar techniques.

Applying DSV methodology for the expected number of base pairs for such saturated structures leads to complex expressions, which may be intractable. For that reason, we consider a natural stochastic process to generate random saturated structures. Our stochastic process adds base pairs, one at a time, according to the uniform distribution, without violating any of the constraints of a structure. See Section 3 for more details. At the web site http://bioinformatics.bc.edu/clotelab/SUPPLEMENTS/RNAasymptoticsCanonicalStr/, we have placed Python programs and Mathematica code used in computing and checking the asymptotic number of canonical secondary structures.

2. DSV methodology

In this section, we describe a combinatorial method usually called DSV methodology, apparently due to Delest, Schützenberger and Viennot. See especially the appendix of [12] for a detailed presentation of this method.

If A is a finite alphabet, then A^* denotes the set of all finites sequences of characters drawn from A. Let Σ be the set consisting of the symbols for left parenthesis `(`, right parenthesis `)`, and dot $\bullet$, used to represent a secondary structure in Vienna notation.

A context-free grammar [11] for RNA secondary structures is given by $G = (V, \Sigma, \mathcal{R}, S_0)$, where V is a finite set of nonterminal symbols (also called variables), $\Sigma = \{\bullet, \texttt{(}, \texttt{)}\}$, $S_0 \in V$ is the *start* nonterminal, and

$$\mathcal{R} \subseteq V \times (V \cup \Sigma)^*$$

is a finite set of production rules. Elements of $\mathcal{R}$ are usually denoted by $A \to w$, rather than (A, w). If rules $A \to \alpha_1, \ldots, A \to \alpha_m$ all have the same left hand side, then this is usually abbreviated by $A \to \alpha_1 \| \cdots \| \alpha_m$.

If $x, y \in (V \cup \Sigma)^*$ and $A \to w$ is a rule, then by replacing the occurrence of A in xAy we obtain xwy. Such a derivation in one step is denoted by $xAy \Rightarrow_G xwy$, while the reflexive, transitive closure of $\Rightarrow_G$ is denoted $\Rightarrow_G^*$. The language generated by context-free grammar G is denoted by $L(G)$, and defined by

$$L(G) = \{w \in \Sigma^* : S_0 \Rightarrow_G^* w\}.$$

For any nonterminal $S \in V$, we also write $L(S)$ to denote the language generated by rules from G when using start symbol S. A derivation of word w from start symbol S_0 using grammar G

Type of nonterminal	Equation for the l.g.f.
$S \to T \mid U$	$S(z) = T(z) + U(z)$
$S \to T\,U$	$S(z) = T(z)U(z)$
$S \to t$	$S(z) = z$
$S \to \varepsilon$	$S(z) = 1$

Table 1. Translation between context-free grammars and generating functions. Here, $G = (V, \Sigma, \mathcal{R}, S_0)$ is a given context-free grammar, S, T and U are any nonterminal symbols in V, and t is a terminal symbol in Σ. The generating functions for the languages $L(S)$, $L(T)$, $L(U)$ are respectively denoted by $S(z)$, $T(z)$, $U(z)$. Table is reproduced with permission from [12].

is a *leftmost* derivation, if each successive rule application is applied to replace the leftmost nonterminal occurring in the intermediate expression. A context-free grammar G is *non-ambiguous*, if there is no word $w \in L(G)$ which admits two distinct leftmost derivations.

Suppose that $G = (V, \Sigma, \mathcal{R}, S)$ is a non-ambiguous context-free grammar which generates a collection $L(S)$ of objects (e.g. canonical secondary structures) and that $S(z) = \sum_{n=0}^{\infty} s_n z^n$ is a (complex) generating function, such that the nth Taylor coefficient $[z^n]S(z) = s_n$ represents the number of objects we wish to count. In the sequel, s_n will represent the number of canonical secondary structures for a homopolymer of length n. The DSV method uses Table 1, taken from [12], in order to translate the grammar rules of $\mathcal{R}$ into a system of equations, which can be solved to find an expression for $S(z)$. The appendix of [12] explains at length how to perform *dominant singularity analysis* in order to determine the asymptotic value of s_n.

2.1. Asymptotic number of canonical secondary structures

In [2] the notion of *canonical secondary structure* S is defined as a secondary structure having no *lonely* (isolated) base pairs; i.e. formally, there are no base pairs $(i, j) \in S$ for which both $(i-1, j+1) \notin S$ and $(i+1, j-1) \notin S$. In this section, we compute the asymptotic number of canonical secondary structures for a homopolymer of length n with $\theta = 1$; i.e. for which there exists at least one unpaired base in every hairpin loop. Throughout this section, secondary structure is interpreted to mean a secondary structure on a homopolymer with minimum number θ of unpaired bases in every hairpin loop set to be 1. At the cost of working with more complex expressions, by the same method, one could analyze the case when $\theta = 3$, which is assumed for the software `mfold` [19] and `RNAfold` [9].

Consider the context-free grammar $G = (V, \Sigma, \mathcal{R}, S)$, where V consists of nonterminals S, R, Σ consists of the terminals $\bullet$, $($, $)$, S is the start symbol and $\mathcal{R}$ consists of the following rules:

$$\begin{aligned} S &\to \bullet \mid S\bullet \mid (R) \mid S(R) \\ R &\to (\bullet) \mid (R) \mid (S(R)) \mid (S\bullet) \end{aligned} \tag{1}$$

The nonterminal S is intended to generate all *nonempty canonical* secondary structures. In contrast, the nonterminal R is intended to generate all secondary structures which become canonical when surrounded by a closing set of parentheses. We claim that the grammar G is non-ambiguous and generates all nonempty canonical secondary structures (inductive proof not given due to space constraints).

By DSV methodology, the non-ambiguous grammar (1) gives the following equations

$$S(z) = z + S(z)z + R(z)z^2 + S(z)R(z)z^2 \tag{2}$$
$$R(z) = z^3 + R(z)z^2 + S(z)R(z)z^4 + S(z)z^3 \tag{3}$$

which can be solved using Mathematica™ software to give the solutions

$$S(z) = \frac{1 - z - z^2 + z^3 - z^5 - \sqrt{F(z)}}{2z^4} \tag{4}$$

and

$$S(z) = \frac{1 - z - z^2 + z^3 - z^5 + \sqrt{F(z)}}{2z^4} \tag{5}$$

where $F(z) = 4z^5\left(-1 + z^2 - z^4\right) + \left(-1 + z + z^2 - z^3 + z^5\right)^2$. Note that when evaluated at $z = 0$, the equation (4) yields $S = 0/0$, while equation (5) yields $2/0$. Since $S(z) = \sum_{n=0}^{\infty} s_n z^n$ is assumed to be analytic at 0, we retain only the first solution, for which it is easily verified that $\lim_{z \to 0} S(z) = 0$.

The square root function $\sqrt{z}$ has a singularity at $z = 0$, so we are led to investigate the roots of $F(z)$. Mathematica™ computes the 10 roots 0.508136, 4.11674, $-0.868214 - 0.619448i$, $-0.868214 + 0.619448i$, $-0.799805 - 0.367046i$, $-0.799805 + 0.367046i$, $0.410134 - 0.564104i$, $0.410134 + 0.564104i$, $0.945448 - 0.470929i$, $0.945448 + 0.470929i$. It follows that $\rho = 0.508136$ is the root of $F(z)$ having smallest (complex) modulus; such a singularity is known as the *dominant singularity*.

Let $T(z) = \frac{1-z-z^2+z^3-z^5}{2z^4}$ and factor $1 - z/\rho$ out of $F(z)$ to obtain $Q(z)(1 - z/\rho) = F(z)$. It follows that

$$S(z) = T(z) + \frac{\sqrt{Q(z)}}{2z^4} \cdot (1 - z/\rho)^{\alpha}$$

where $\alpha = 1/2$. As explained in [12], for asymptotics we can ignore the term $T(z)$. By the main theorem of Flajolet and Odlyzko [6], discussed at length in the appendix of [12],

$$s_n \sim \frac{K(\rho)}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot (1/\rho)^n \tag{6}$$

where $\alpha = 1/2$ and $K(z) = \frac{\sqrt{Q(z)}}{2z^4}$. Plugging $\rho = 0.508136$ into equation (6), we derive the following theorem.

Theorem 1: The asymptotic number of canonical secondary structures for a homopolymer of length n is

$$2.1614 \cdot n^{-3/2} \cdot 1.96798^n. \tag{7}$$

We now derive the expected number of base pairs in canonical secondary structures of a homopolymer of length n with $\theta = 1$. Modify the equations (2,3) by adding a new variable u, intended to count the number of base pairs, thus yielding $S(z, u) =$

$$z + S(z, u)z + R(z, u)uz^2 + S(z, u)R(z, u)uz^2$$

and $R(z, u) =$

$$uz^3+R(z, u)uz^2+S(z, u)R(z, u)u^2z^4+S(z, u)uz^3.$$

which can be solved by the software Mathematica™to yield the solution[1]

$S(z, u) = \sum_{n\geq 0} \sum_{k\geq 0} s_{n,k} z^n u^k$
$= 2u^2z^4 \left(1 - z - uz^2 + uz^3 - u^2z^5 - \sqrt{4u^2z^5 \left(uz^2 - u^2z^4 - 1\right) + \left(z + uz^2 - uz^3 + u^2z^5 - 1\right)^2}\right).$

Using a classical observation [5], note that the expected number $s_{n,k}$ of base pairs in a canonical secondary structure on a homopolymer of length n is related to the partial derivative of $S(z, u)$; indeed,

$$\begin{aligned}
& \frac{[z^n]\frac{\partial S(z,u)}{\partial u}(z, 1)}{[z^n]S(z, 1)} \\
= & \frac{[z^n]\left(\sum_{i\geq 0}\sum_{k\geq 0} s_{i,k} z^i k u^{k-1}\right)(z, 1)}{s_n} \\
= & \frac{\sum_{k\geq 0} s_{n,k} k}{s_n} \\
= & \sum_{k\geq 0} k\mathbb{P}(X_n = k) \\
= & \mathbb{E}(X_n) \\
= & s_{n,k}.
\end{aligned}$$

Here, $\mathbb{P}(X_n = k) = \frac{s_{n,k}}{s_n}$ is the (uniform) probability that a canonical secondary structure of a homopolymer of length n has exactly k base pairs.

We compute that $G(z) = \frac{\partial S(z,u)}{\partial u}(z, 1)$ satisfies

$$G(z) = \frac{-(z^2 - 2)(T(z) - \sqrt{F(z)} + z\sqrt{F(z)})}{2z^4\sqrt{F(z)}}$$

where $T(z) = (1-2z+2z^3-z^4-3z^5+z^6)$ and $F(z) = -4z^5(1-z^2+z^4)+(-1+z+z^2-z^3+z^5)^2$. Simplification yields $G(z) =$

$$\frac{-(z^2 - 2)(z - 1)}{2z^4} - \frac{T(z)(z^2 - 2)}{2z^4} \cdot \left(\frac{1}{\sqrt{F(z)}}\right).$$

We are thus led to investigate the dominant singularity ρ, i.e., the root of $F(z)$ having smallest (complex) modulus. As before, we obtain $\rho = 0.508136$. Factor $(1 - z/\rho)$ out of $F(z)$ so that $F(z) = Q(z)(1 - z/\rho)$. It follows that $G(z) =$

$$\frac{-(z^2 - 2)(z - 1)}{2z^4} - \frac{T(z)(z^2 - 2)}{2z^4} \cdot Q(z)^{\alpha} \cdot (1-z/\rho)^{\alpha}$$

where $\alpha = -1/2$. By the theorem of Flajolet and Odlyzko [6], we obtain the asymptotic value

$$\frac{K(\rho)}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot (1/\rho)^n \tag{8}$$

where $\alpha = -1/2$ and $K(z) = -\frac{Q(z)^{-1/2}T(z)(z^2-2)}{2z^4}$. Plugging $\rho = 0.508136$ into equation (8), we find the asymptotic value of $[z^n]\frac{\partial S(z,u)}{\partial u}(z, 1)$ is

$$0.68568 \cdot n^{-1/2} \cdot 1.96798^n. \tag{9}$$

Dividing (9) by the asymptotic number $[z^n]S(z)$ of canonical secondary structures, given in (7), we have the following theorem.

Theorem 2: The asymptotic expected number of base pairs in canonical secondary structures on a homopolymer of length n with $\theta = 1$ is $0.31724 \cdot n$.

1. Since $S(z, u)$ is assumed to be analytic at 0, we have discarded one of the two solutions as before.

3. Random saturated structures

An RNA secondary structure is *saturated* if $\theta+1$ is the maximum size of a contiguous sequence of unpaired nucleotides; i.e. it is not possible to add any base pairs without violating the definition of secondary structures. If one models the folding of an RNA secondary structure as a random walk on a Markov chain (i.e. by the Metropolis-Hastings algorithm), then saturated structures correspond to *kinetic traps* with respect to the Nussinov energy model [15].

In [3], we computed the asymptotic number $N(n) = 1.07427 \cdot n^{-3/2} \cdot 2.35467^n$ of saturated secondary structures on $[1, n]$ for the homopolymer with $\theta = 1$. An important question is related to whether or not the distribution of the number of base pairs of saturated structures is Gaussian.[2] It is currently an open question to determine the asymptotic number $N(n, k)$ of saturated secondary structures having k base pairs, although this was (essentially) solved in [3] for k equal to the largest and next to largest possible value. If values of $N(n, k)$ were available, then the expected number of base pairs in a random saturated secondary structure would be $\sum_k kN(n, k)/N(n)$. However, since we are currently unable to determine the asymptotic value of $N(n, k)$ or to otherwise determine the asymptotic number of base pairs in random saturated secondary structures, in this section and the next, we consider natural algorithmic approaches of generating (a subclass of) random saturated secondary structures. In particular, in this section we are interested in the asymptotic expected number of base pairs in saturated structures. We define a stochastic greedy process to generate random saturated structures.

We look at the expected number of base pairs of random saturated structures, generated by the following stochastic process. We begin with n bases in sequential order arranged on a line. Select the base pair $(1, u)$ by choosing u, where $\theta + 2 \leq u \leq n$, at random with probability $1/(n - \theta - 1)$. The base pair joining 1 and u partitions the line into two parts. The left region has k bases strictly between 1 and u, where $k \geq \theta$, and the right region contains the remaining $n - k - 2$ bases

2. Since the energy of a secondary structure S under the Nussinov energy model [15] is equal to -1 times the number of base pairs, this question is equivalent to whether the energy distribution for saturated secondary structures is Gaussian. Note that it remains an open question whether the energy distribution of (not necessarily saturated) secondary structures is Gaussian under the Turner energy model [14].

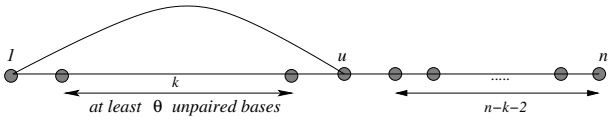

Figure 2. Base 1 is base-paired by selecting a random base u such there are at least θ unpaired bases enclosed between 1 and u.

properly contained within endpoints $k + 2$ and n (see Figure 2). Proceed recursively on each of the two parts. Observe that the secondary structures produced by our stochastic process will always base pair with the leftmost available base.

Let U_n^θ be the expected number of basespairs of the saturated secondary structure generated by this recursive procedure. In general, we have the following recursive equation

$$U_n^\theta = 1 + \frac{1}{n-\theta-1} \sum_{k=\theta}^{n-2} (U_k^\theta + U_{n-k-2}^\theta) \quad (10)$$

for all $n \geq \theta + 2$. Observe that we have the initial conditions $U_0^\theta = U_1^\theta = \cdots = U_{\theta+1}^\theta = 0, U_{\theta+2}^\theta = U_{\theta+3}^\theta = 1$. If we write equation (10) for U_{n+1}^θ and substitute in it the value for U_n^θ we derive

$$\begin{aligned} &U_{n+1}^\theta = \\ &1 + \frac{1}{n-\theta} \sum_{k=\theta}^{n-1} (U_k^\theta + U_{n-k-1}^\theta) = \\ &1 + \frac{1}{n-\theta} \left(U_{n-1}^\theta + U_{n-\theta-1}^\theta + \sum_{k=\theta}^{n-2} (U_k^\theta + U_{n-k-2}^\theta) \right) \\ &1 + \frac{1}{n-\theta} \left(U_{n-1}^\theta + U_{n-\theta-1}^\theta \right) + \frac{n-\theta-1}{n-\theta} (U_n^\theta - 1). \end{aligned}$$

If we multiply out by $n-\theta$ and simplify we obtain

$$(n-\theta)U_{n+1}^\theta = 1 + (n-\theta-1)U_n^\theta + U_{n-1}^\theta + U_{n-\theta-1}^\theta, \quad (11)$$

which is valid for $n \geq \theta + 1$.

3.1. Asymptotic behavior

In the sequel we look at asymptotics. In particular we prove the following result.

Theorem 3: If the threshold satisfies $\theta = o(n)$ then the limit $\lim_{n\to\infty} \frac{U_n^\theta}{n}$ exists.

Before continuing with the proof of Theorem 3 we mention an alternative approach in trying to establish the convergence of U_n^θ/n as n approaches infinity. As explained in [4], where we established the existence of an asymptotic limit for the expected minimum free energy As explained in [4],

where we established the existence of an asymptotic limit for the expected minimum free energy for randomly generated RNA *sequences* (not random structures!), Kingman's ergodicity theorem [10] requires (here) a superadditive sequence of doubly-indexed random variables $X_{i,j}$. Indeed, if one defines $X^\theta_{i,j}$ to be the expected number of base pairs among random saturated random secondary structures on the nucleotide sequence $[i,j]$, then Kingman's theorem would require superadditivity, i.e. $X^\theta_{i,k} \geq X^\theta_{i,j} + X^\theta_{j,k}$. If one could establish superadditivity, then since $X^\theta_{1,n}/n$ is bounded above by $1/2$, it would follow from Kingman's theorem that $\lim_{n\to\infty} X_{1,n}/n$ exists. Since it is not obvious that one has superadditivity in the case at hand, we have proceeded otherwise.

Proof: **(of Theorem 3)** Consider the two sequences $(\frac{U^\theta_n}{n} : n \geq 1)$ and $(U^\theta_{n+1} - U^\theta_n : n \geq 1)$. We prove that if one of these sequences has a limit, as n approaches infinity, then so does the other; moreover, we claim that $\lim_{n\to\infty} \frac{U^\theta_n}{n} = \lim_{n\to\infty} \left(U^\theta_{n+1} - U^\theta_n\right)$. The idea is to use the following equation which is derived from the basic identity (11): $U^\theta_{n+1} =$

$$\frac{1}{n-\theta} + \frac{n-\theta-1}{n-\theta} U^\theta_n + \frac{1}{n-\theta} U^\theta_{n-1} + \frac{1}{n-\theta} U^\theta_{n-\theta-1}.$$

Collecting terms we have $U^\theta_{n+1} - U^\theta_n =$

$$\frac{1}{n-\theta} - \frac{U^\theta_n}{n-\theta} + \frac{U^\theta_{n-1}}{n-\theta} + \frac{U^\theta_{n-\theta-1}}{n-\theta} \quad (12)$$

$$= \frac{1+U^\theta_{n-\theta-1}}{n-\theta} - \frac{U^\theta_n - U^\theta_{n-1}}{n-\theta} \quad (13)$$

$$= \frac{1+U^\theta_{n-\theta-1}}{n-\theta} - \frac{1}{n-\theta}(U^\theta_n - U^\theta_{n-1}) \quad (14)$$

Repeating equation (14) to the term $U^\theta_n - U^\theta_{n-1}$ we derive that $U^\theta_{n+1} - U^\theta_n =$

$$\frac{1+U^\theta_{n-\theta-1}}{n-\theta} - \frac{1}{n-\theta}\left(\frac{1+U^\theta_{n-\theta-2}}{n-1-\theta} - \frac{U^\theta_{n-1} - U^\theta_{n-2}}{n-1-\theta}\right) \quad (15)$$

Since there are at most $n/2$ base pairs in any secondary structure on a sequence of length n, we conclude from equation (15) that the limit of $U^\theta_{n+1} - U^\theta_n$ exists if and only if the limit of U^θ_n/n exists.

Next we show that the sequence $\frac{U^\theta_n}{n}$ has a limit as $n \to \infty$. First of all we prove that

$$\left|\frac{U^\theta_n}{n} - \frac{U^\theta_{n+1}}{n+1}\right| \in \frac{O(\theta)}{n^2}. \quad (16)$$

To prove this, observe that by equation (11),

$$nU^\theta_{n+1} = \theta U^\theta_{n+1} + 1 + (n-\theta-1)U^\theta_n + U^\theta_{n-1} + U^\theta_{n-\theta-1} \quad (17)$$

and hence $(n+1)U^\theta_n nU^\theta_{n+1} =$

$$nU^\theta_n + U^\theta_n - \theta U^\theta_{n+1} - 1 - (n-\theta-1)U^\theta_n - U^\theta_{n-1} - U^\theta_{n-\theta-1}$$

$$= 2U^\theta_n + \theta U^\theta_n - \theta U^\theta_{n+1} - U^\theta_{n-1} - U^\theta_{n-\theta-1}$$

$$= \theta(U^\theta_n - U^\theta_{n+1}) + (U^\theta_n - U^\theta_{n-1}) + (U^\theta_n - U^\theta_{n-\theta-1})$$

Dividing both sides by $n(n+1)$, we have $\frac{U^\theta_n}{n} - \frac{U^\theta_{n+1}}{n+1} =$

$$\frac{\theta(U^\theta_n - U^\theta_{n+1})}{n(n+1)} + \frac{(U^\theta_n - U^\theta_{n-1})}{n(n+1)} + \frac{(U^\theta_n - U^\theta_{n-\theta-1})}{n(n+1)}$$

Equation (15) clearly implies that $|U^\theta_{n+1} - U^\theta_n| \in O(1)$ and $|U^\theta_n - U^\theta_{n-\theta-1}| \in O(1)$, and so we have established Equation (16), that

$$\left|\frac{U^\theta_n}{n} - \frac{U^\theta_{n+1}}{n+1}\right| \in \frac{O(\theta)}{n^2}.$$

Now we can complete the proof of Theorem 3. Indeed, equality (16) implies that the sequence $\frac{U^\theta_n}{n}$ is uniformly convergent, in the sense that for any $\epsilon > 0$ there is an integer n_0 such that form $n, m \geq n_0$ we have that $\left|\frac{U^\theta_n}{n} - \frac{U^\theta_m}{m}\right| < \epsilon$. Since every uniformly convergent sequence has a limit the proof of the theorem is complete. □

3.2. Expected number of base pairs for arbitrary threshold $\theta \geq 0$

In this section we determine the generating function

$$y = \sum_{n=0}^{\infty} U^\theta_n z^n \quad (18)$$

of the sequence U^θ_n of the expected number of basespairs of random saturated secondary structures, generated by the stochastic process described earlier. The main result of this section is summarized in the following theorem.

Theorem 4: The generating function y for U^θ_n, i.e., the expected number of basespairs of the random saturated secondary structure with threshold $\theta \geq 0$, generated by the recursive procedure described in equation (10), is

$$\frac{z^{\theta+1}}{(1-z)^2} e^{\left(-z - \sum_{i=0}^{\theta} \frac{z^i}{i+1}\right)} \int e^{\left(z + \sum_{i=0}^{\theta} \frac{z^i}{i+1}\right)} dz. \quad (19)$$

Proof: In order to use the method of generating functions we multiply equation (11) by z^{n+1} to obtain that $(n-\theta)U^\theta_{n+1} z^{n+1} =$

$$z^{n+1} + (n-\theta-1)U^\theta_n z^{n+1} + U^\theta_{n-1} z^{n+1} + U^\theta_{n-\theta-1} z^{n+1},$$

which is valid for $n \geq \theta+1$, and then sum the resulting equation for $n \geq \theta+1$ to derive a functional equation

$$\begin{aligned}&\sum_{n=\theta+1}^{\infty}(n-\theta)U_{n+1}^{\theta}z^{n+1}\\ =\;&\sum_{n=\theta+1}^{\infty}z^{n+1}+\sum_{n=\theta+1}^{\infty}(n-\theta-1)U_{n}^{\theta}z^{n+1}+\\ &\sum_{n=\theta+1}^{\infty}U_{n-1}^{\theta}z^{n+1}+\sum_{n=\theta+1}^{\infty}U_{n-\theta-1}^{\theta}z^{n+1}.\end{aligned}$$

Next we express the above equation using the generating function y given in equation (18) as a Taylor series expansion around the origin 0. Note that the first derivative of y is

$$y'=\sum_{n=1}^{\infty}nU_{n}^{\theta}z^{n-1}.$$

Observe that using the initial conditions $U_0^\theta = U_1^\theta = \cdots = U_{\theta+1}^\theta = 0$ and $U_{\theta+2}^\theta = U_{\theta+3}^\theta = 1$ we can make the following substitutions

$$\begin{aligned}\sum_{n=\theta+1}^{\infty}(n-\theta)U_{n+1}^{\theta}z^{n+1} &= zy'-(\theta+1)y,\\ \sum_{n=\theta+1}^{\infty}z^{n+1} &= \frac{z^{\theta+2}}{1-z},\\ \sum_{n=\theta+1}^{\infty}(n-\theta-1)U_{n}^{\theta}z^{n+1} &= z^2y'-(\theta+1)zy,\\ \sum_{n=\theta+1}^{\infty}U_{n-1}^{\theta}z^{n+1} &= z^2y\\ \sum_{n=\theta+1}^{\infty}U_{n-\theta-1}^{\theta}z^{n+1} &= z^{\theta+2}y.\end{aligned}$$

If we substitute these values into the above equation we derive the differential equation

$$zy'-(\theta+1)y=\frac{z^{\theta+2}}{1-z}+z^2y'-(\theta+1)zy+z^2y+z^{\theta+2}y.$$

Finally, if we collect terms and simplify we obtain $y' =$

$$\frac{(\theta+1)-(\theta+1)z+z^2+z^{\theta+2}}{z-z^2}y+\frac{z^{\theta+1}}{(1-z)^2}. \tag{20}$$

It remains to solve the resulting differential equation (20). To this effect, multiply both sides of equation (20) by an (as yet unknown) function $\phi(z)$ in order to obtain $\phi(z)y' =$

$$\phi(z)\frac{(\theta+1)-(\theta+1)z+z^2+z^{\theta+2}}{z-z^2}y+\phi(z)\frac{z^{\theta+1}}{(1-z)^2}.$$

If we could find a function $\phi(z)$ such that

$$\frac{d\phi(z)}{dz}=-\phi(z)\frac{(\theta+1)-(\theta+1)z+z^2+z^{\theta+2}}{z-z^2} \tag{21}$$

then using the product rule for derivatives, namely $\phi(z)y'+\frac{d\phi(z)}{dz}y=\frac{d}{dz}(\phi(z)y)$, we conclude that

$$\frac{d}{dz}(\phi(z)y)=\phi(z)\frac{z^{\theta+1}}{(1-z)^2}. \tag{22}$$

A solution of the homogeneous differential equation (21) is

$$\phi(z)=e^{-\int\frac{(\theta+1)-(\theta+1)z+z^2+z^{\theta+2}}{z-z^2}dz}. \tag{23}$$

After doing some elementary calculations we can calculate the integral

$$\int\frac{(\theta+1)-(\theta+1)z+z^2+z^{\theta+2}}{z-z^2}dz$$

and derive up to a constant the following formula

$$(\theta+1)\ln z-z-2\ln(1-z)-\sum_{i=0}^{\theta}\frac{z^i}{i+1}. \tag{24}$$

Using equation (23), this implies that

$$\phi(z)=\frac{(1-z)^2\exp\left(z+\sum_{i=0}^{\theta}\frac{z^i}{i+1}\right)}{z^{\theta+1}} \tag{25}$$

up to a multiplicative constant >0. Finally, equation (22) yields the following closed form solution for the desired function

$$y=\frac{1}{\phi(z)}\left(\int\frac{\phi(z)z^{\theta+1}}{(1-z)^2}dz+c\right), \tag{26}$$

where c is a constant. Substituting $\phi(z)$ from equation (25) into equation (26) we derive

$$y=\frac{z^{\theta+1}}{(1-z)^2}\cdot e^{-z-\sum_{i=0}^{\theta}\frac{z^i}{i+1}}\cdot\int e^{z+\sum_{i=0}^{\theta}\frac{z^i}{i+1}}dz+c, \tag{27}$$

for some constant c. The constant c is easily determined to be 0 in view of the initial condition $U_{\theta+2}^\theta = 1$. This completes the proof of Theorem 4. □

Elementary calculations show that our recurrence is identical to a formula derived for a seating arrangement problem for which Rothman in [7] (see also [13]) gives the following asymptotic formula

$$U_{n+2\theta+1}^{\theta}\sim(n+2(\theta+1)+1)\ell(\theta)-1, \tag{28}$$

which is valid for $\theta \geq 0$, where

$$\ell(\theta)=\int_0^1\exp\left\{2\left[\sum_{i=1}^{\theta+1}\frac{t^i-1}{i}\right]\right\}dt. \tag{29}$$

The previous discussion implies that the asymptotic limit $\lim_{n\to\infty} \frac{U_n^\theta}{n}$ is equal to $\ell(\theta)$ and therefore we have the following theorem.

Theorem 5: The expected number U_n^θ of base-spairs of the saturated secondary structures with threshold $\theta \geq 0$, generated by the recursive procedure above, satisfies

$$\lim_{n\to\infty} \frac{U_n^\theta}{n} = \int_0^1 \exp\left\{2\left[\sum_{i=1}^{\theta+1} \frac{t^i - 1}{i}\right]\right\} dt. \quad (30)$$

4. Conclusion

In this paper we applied the DSV methodology to enumeration problems concerning canonical and saturated secondary structures. For instance, we showed that the asymptotic number of canonical RNA secondary structures for the homopolymer model with $\theta = 1$ is equal to $2.1614 \cdot n^{-3/2} \cdot 1.96798^n$, which provides a theoretical explanation for the speed-up observed for Vienna RNA Package when restricted to canonical structures [2]. We then developed a new method to determine certain structural properties for random saturated RNA secondary structures; in particular, we constructed generating functions for the expected number of base pairs and hairpins.

References

[1] E. A. Bender. Asymptotic methods in enumeration. *SIAM Rev.*, 16(4):485–515, 1974.

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

[3] P. Clote. Combinatorics of saturated secondary structures of RNA. *J. Comput. Biol.*, 13(9):1640–1657, November 2006.

[4] P. Clote, F. Ferré, E. Kranakis, and D. Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, 2005.

[5] A. Denise, O. Roques, and M. Termier. Random generation of words of context-free languages according to the frequencies of letters. In D. Gardy and A. Mokkadem, editors, *Mathematics and Computer Science: Algorithms, Trees, Combinatorics and probabilities*, Trends in Mathematics, pages 113–125. Birkhaüser, 2000.

[6] P. Flajolet and A. M. Odlyzko. Singularity analysis of generating functions. *SIAM Journal of Discrete Mathematics*, 3:216–240, 1990.

[7] H. D. Friedman and D. Rothman. Solution to an Unfriendly Seating Arrangement (Problem 62-3), *SIAM Review*, 1964, 2, Vol. 6, pp. 180–182.

[8] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S.R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441, 2003.

[9] I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31(13):3429–3431, 2003.

[10] J.F.C. Kingman. Subadditive ergodic theory. *Annals of Probability Theory*, 1(6):893–909, 1973.

[11] H.R. Lewis and C.H. Papadimitriou. *Elements of the Theory of Computation*. Prentice-Hall, 1997. Second edition.

[12] W.A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of rna shapes. *J Compu Biol.*, 2007. in press.

[13] J. K. MacKenzie. Sequential filling of a line by intervals placed at random and its application to linear adsorption. *The Journal of Chemical Physics*, 37(4):723–728, August 1962.

[14] D.H. Matthews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[15] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.

[16] P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics*, 26:261–272, 1978.

[17] K. C. Wiese, E. Glen, and A. Vasudevan. JViz.Rna–a Java tool for RNA secondary structure visualization. *IEEE. Trans. Nanobioscience.*, 4(3):212–218, September 2005.

[18] M. Zuker. RNA folding prediction: The continued need for interaction between biologists and mathematicians. In *Lectures on Mathematics in the Life Sciences*, volume 17, pages 87–124. Springer-Verlage, 1986.

[19] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.

[20] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.