MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **66** (2011) 399-414

ISSN 0340 - 6253

Enumerating RNA Structures, Including Pseudoknots of any Topology

Mohammad GANJTABESH[*,a,b] and **Jean-Marc STEYAERT**[b]

[a] *Center of Excellence in Biomathematics, School of Mathematics, Statistics and Computer Science, University of Tehran, Tehran 14155-6455, Iran.*
e-mail: *mgtabesh@ut.ac.ir*

[b] *Laboratoire d'Informatique, Ecole Polytechnique, 91128 Palaiseau Cedex, France.*
e-mail: mohammad.ganjtabesh@polytechnique.edu; steyaert@lix.polytechnique.fr

(Received October 19, 2010)

Abstract

Counting the number of RNA structures is an important combinatorial problem in computational biology. In this paper, we enumerate the number of RNA structures for a special case, in which any arbitrary number of base pairs are allowed in the given RNA sequence. The only criteria considered in our model, is the minimum length condition for hairpin loops, assumed to be 1. The asymptotic behavior and its relation with the number of involutions are presented from the analytical and combinatorial points of view.

1. Introduction

The most important problem and the greatest challenge in bioinformatics deals with deciphering the code transforming sequences of biopolymers (such as RNAs, Proteins, etc.) into special molecular structures. A sequence can be visualized as a string of symbols,

together with the environment, encodes the molecular structure of the biopolymer. Several methods are available to predict the RNA secondary structure without pseudoknots [11, 14, 16, 21] or with some types of pseudoknots [1, 9, 13, 17]. Some of these models are statistical models with roots in combinatorial problems. Although these models are much simpler than the energy based models [5, 16] (dealing with the thermodynamical parameters), they often provide exact analytical solutions about the structure and entropy. For these reasons, the combinatorics of the biopolymers' structures has been considered extensively during the past 30 years [4].

Extensive theoretical analysis for the number of RNA secondary structures, their complexity, and their composition, has been studied previously [4, 10, 15, 19, 20]. Based on graph-theoritical properties of RNA structures, some enumeration and classification methods have been presented [2, 3].

Also, the combinatorial aspects of RNA secondary structures have been studied in detail by Waterman [20]. A recursive formula for the number of distinct RNA secondary structures is obtained and the analytical results, as well as the asymptotic behavior about this recursive formula are presented. For pseudoknotted RNA structure, the combinatorial properties have been studied by Hofacker et al. [4]. Also a recursive formula for the number of pseudoknotted RNA structures is presented.

Jin et al. developed a general framework based on generating functions for the asymptotic expansion of the number of k-noncrossing RNA structures [6, 7]. They have proved that for an arbitrary k the expansions exist, and via transfer theorem of analytic combinatorics, it is possible to obtain the asymptotic expression for coefficients of the generating function. Asymptotic expansions for $k = 2$ and $k = 3$ were also presented.

In [12], the relation between the RNA secondary structures and the Feynman diagrams is made more explicit by formulating a matrix field theory model, whose Feynman diagrams give exactly all the pseudoknotted RNA structures. By using this matrix model formulation, Vernizzi et al. [18] enumerated the number of RNA contact structures for the simple case of an RNA molecule with a flexible backbone, in which any arbitrary base pairs are allowed.

In this paper, we present some analytical results about the number of RNA structures for a sequence of length n. Then by using the combinatorial methods, we introduce a new

way to enumerate the number of RNA structures and provide its asymptotic behavior.

The rest of this paper is organized as follows: In Section 2, we review the basic definitions of RNA structures. Section 3 presents the methodology of counting the number of different RNA structures and its recursive formula. The analytical results of the recursive formula are presented in Section 4. In Section 5, the problem of counting the number of RNA structures is studied from the combinatorial point of view and a new asymptotic behavior is presented. Finally, the conclusions are presented in Section 6.

2. RNA Structure: Basic Definitions

An RNA molecule is a sequence of nucleotides of four possible types, denoted by the letters A, C, G, and U (stand for Adenine, Cytosine, Guanine, and Uracil, respectively), connected by a backbone which is called RNA Primary Structure. For an RNA molecule of length n, we index the nucleotides from 1 to n, starting from left ($5'-$ end). Two nucleotides that are connected via hydrogen bonds are called a base pair. There are two kinds of base pairing, Watson-Crick and Wobble. In the Watson-Crick base pairing, A always forms a base pair with U, as does G with C, and vice versa. In the Wobble base pairing, G can form a base pair with U as well as U with G. We write $i.j$ if the nucleotide with index i is paired with the nucleotide with index j ($i < j$). For an RNA sequence of length n, its structure is a set S of base pairs $i.j$ with $1 \leq i < j \leq n$, such that for all $i_1.j_1, i_2.j_2 \in S$ we have $i_1 = i_2$ if and only if $j_1 = j_2$ (each base can take part in at most one base pairing). The set S is called pseudoknot-free structure if for all $i_1.j_1, i_2.j_2 \in S$ they are either nested ($i_1 < i_2 < j_2 < j_1$) or disjoint ($i_1 < j_1 < i_2 < j_2$). In many situations these conditions allow us to first handle one base pair and then the other one (if they are nested) or handle them independently (if they are disjoint). Two base pairs $i_1.j_1, i_2.j_2 \in S$ form a pseudoknot if $i_1 < i_2 < j_1 < j_2$ and S is called pseudoknotted structure if it contains at least two base pairs which form pseudoknot. The difficulty behind the pseudoknotted structures is that in many situations we can not handle its base pairs separately and we should consider them all together. Therefore, dealing with pseudoknotted structures is much more difficult than the pseudoknot-free structures.

In the rest of this paper, an RNA sequence of length n is assumed to be a sequence of n points ($1 - 2 - \cdots - n$), where each point i is connected to the points $i-1$ and $i+1$

$(1 < i < n)$. A point i is called unpaired if it is not connected to any points other than $i-1$ and $i+1$. A point i is paired if there exists just one point j, other than $i-1$ and $i+1$, in which i and j make a base pair.

3. A Recursion Formula for the Number of RNA Structures

Suppose that $P(n)$ denotes the number of RNA structures for a given RNA sequence of length n. In order to formulate the $P(n)$, two different situations should be considered. First, suppose that the last point (n^{th} point) does not make any base pair. In this situation, there are exactly $P(n-1)$ different structures. In second situation, suppose that the last point makes a base pair with another point k, where $1 \leq k \leq n-2$. By removing the points labeled n and k from the sequence, there are $P(n-2)$ different structures for the remaining $n-2$ points. In this situation the points $k-1$ and $k+1$ become neighbors and in our formalism, they can not make a base pair. Therefore, some extra structures corresponding to the case where the point $k-1$ make a base pair with the point $k+1$ should be considered separately. By extending this formalism, again for the situation when the point $k-2$ and $k+2$ make a base pair should be considered separately, and so on. The schematic representation of our formalism is shown in figure 1. The solid arcs in the figure show the base pairings between two end points and the dashed arcs show the situation in which we do not know whether or not there is any base paring.

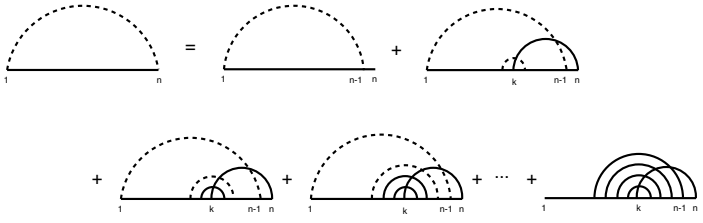

Figure 1: Schematic computation of $P(n)$.

By summarizing the above discussion, the following formula is obtained:

$$P(n) = \begin{cases} 0 & \text{if } n < 0, \\ 1 & \text{if } 0 \le n \le 2, \\ 2 & \text{if } n = 3, \\ P(n-1) + \sum_{k=1}^{n-2} \left(\sum_{t=0}^{min(k-1,n-k-1)} P(n-2t-2) \right) & \text{otherwise.} \end{cases} \quad (1)$$

By using the elementary calculus, this formula is simplified and for $n \ge 4$ a nice recursive formula is obtained as follows:

$$\begin{aligned} P(n) &= P(n-1) + (n-2)P(n-2) + \sum_{k=2}^{n-2} \left(\sum_{t=1}^{min(k-1,n-k-1)} P(n-2t-2) \right) \\ &= P(n-1) + (n-2)P(n-2) + \sum_{k=1}^{n-3} \left(\sum_{t=1}^{min(k,n-k-2)} P(n-2t-2) \right) \\ &= P(n-1) + (n-2)P(n-2) + \sum_{k=1}^{n-3} \left(\sum_{t=0}^{min(k-1,n-k-3)} P(n-2t-4) \right) \\ &= P(n-1) + (n-2)P(n-2) + P(n-4) + \sum_{k=1}^{n-4} \left(\sum_{t=0}^{min(k-1,n-k-3)} P(n-2t-4) \right) \\ &= P(n-1) + (n-2)P(n-2) + P(n-4) + P(n-2) - P(n-3) \\ &= P(n-1) + (n-1)P(n-2) - P(n-3) + P(n-4). \end{aligned} \quad (2)$$

In the next section, the analytical results of formula (2) are discussed. In table 1, the number of different RNA structures for sequences of length n ($1 \le n \le 20$) are presented.

Table 1: The number of different RNA structures.

n	$P(n)$	n	$P(n)$	n	$P(n)$	n	$P(n)$
1	1	6	37	11	16526	16	20732609
2	1	7	112	12	64351	17	94607409
3	2	8	363	13	259471	18	443476993
4	5	9	1235	14	1083935	19	2130346450
5	13	10	4427	15	4668704	20	10482534517

4. Analytical Results

In this section, we first present some properties of the recursive formula (2) and then we show that $I(n)/P(n)$ remains bounded for sufficiently large n, where $I(n)$ is the number

of involutions on a set of size n. An involution on a set S is a permutation $\pi : S \mapsto S$ such that for each $s \in S$, $\pi^2(s) = s$. From [8], it is well known that the asymptotic behavior of $I(n)$ is approximated by the following formula:

$$\frac{1}{\sqrt{2}} n^{\frac{n}{2}} \mathrm{e}^{(\frac{-n}{2}+\sqrt{n}-\frac{1}{4})}.$$

First of all, we present two lemmas without proof to indicate that $P(n)$ is a positive and strictly increasing function.

Lemma 1. *$P(n) > 0$, for each $n \geq 0$.*

Lemma 2. *$P(n) > P(n-1)$, for each $n \geq 3$.*

Moreover, the following stronger statement indicates that the increasing rate of $P(n)$ is at least exponential.

Lemma 3. *$P(n) \geq \sum_{i=1}^{n-1} P(i)$, for each $n \geq 4$.*

Proof. (The proof is based on induction) It is easy to verify that $P(4) \geq P(1) + P(2) + P(3)$, so the statement is true for $n = 4$. Now, suppose that the statement is true for each number k, where $4 \leq k < n$. We show that the statement is also true for n. From the formula (2) and previous lemmas, the term $P(n-2) - P(n-3) + P(n-4)$ is always positive and we have the following inequality:

$$\begin{aligned} P(n) &= P(n-1) + (n-1)P(n-2) - P(n-3) + P(n-4) \\ &\geq P(n-1) + (n-2)P(n-2). \end{aligned} \tag{3}$$

Since $(n-2)P(n-2) \geq 2P(n-2)$ (for each $n \geq 4$), we can write (3) as follows:

$$P(n) \geq P(n-1) + 2P(n-2) \tag{4}$$

Now, from the induction hypothesis, the inequality (4) becomes as follows:

$$\begin{aligned} P(n) &\geq P(n-1) + P(n-2) + \sum_{i=1}^{n-3} P(i) \\ &\geq \sum_{i=1}^{n-1} P(i). \end{aligned}$$

□

The number of involutions on a set S of size n is given by the recursive formula $I(n) = I(n-1) + (n-1)I(n-2)$ [8]. The term $-P(n-3) + P(n-4)$ is always negative (see Lemma 2), which implies that $I(n) > P(n)$ and so $I(n)/P(n) > 1$. Now, we show that $I(n)/P(n)$ has an upper bound. To do this, we need the following lemmas.

Lemma 4. *For $n \geq 1$ we have*

$$4\left(n^{1/4} - (n-1)^{1/4}\right) \leq \frac{1}{(n-1)^{3/4}}.$$

Proof. From the elementary calculus, we can write the following equivalencies:

$$\begin{aligned}
4\left(n^{1/4} - (n-1)^{1/4}\right) &\leq \frac{1}{(n-1)^{3/4}} \\
&\Longleftrightarrow \\
4n^{1/4}(n-1)^{1/4} &\leq 4n - 3 \\
&\Longleftrightarrow \\
4^4 3n^2 - 4^4 n &\leq 64^2 3^2 n^2 - 43^4 n + 3^4 \\
&\Longleftrightarrow \\
13056n^2 + 148n + 81 &\geq 0
\end{aligned}$$

, where the last inequality is always correct for $n \geq 1$. □

Lemma 5. *For $n \geq 1$ we have*

$$2\left(n^{1/4} - (n-2)^{1/4}\right) \leq \frac{1}{(n-2)^{3/4}}.$$

Proof. Using the same calculation performed in the proof of Lemma 4, we can do as follows:

$$\begin{aligned}
2\left(n^{1/4} - (n-2)^{1/4}\right) &\leq \frac{1}{(n-2)^{3/4}} \\
&\Longleftrightarrow \\
2n^{1/4}(n-2)^{1/4} &\leq 2n - 3 \\
&\Longleftrightarrow \\
2^6 3n^2 - 2^7 n &\leq 2^3 3^3 n^2 - 23^4 n + 3^4 \\
&\Longleftrightarrow \\
24n^2 - 34n + 81 &\geq 0.
\end{aligned}$$

Again, the last inequality is always correct for $n \geq 1$. □

Lemma 6. *$P(n) \geq 2n^{1/4}P(n-1)$, for each $n \geq 10$.*

Proof. (The proof is based on induction) For $n = 10$ the statement is true (it is easy to check). Now, suppose that the statement is true for $n-1$ and $n-2$. We show that the statement is true for n. By removing the positive term $P(n-4)$ from the recursive formula (2), the following inequality is obtained:

$$P(n) \geq P(n-1) + (n-1)P(n-2) - P(n-3) \tag{5}$$

By replacing n by $n-1$ in the recursive formula of $P(n)$, the following recursive formula is obtained for $P(n-1)$:

$$P(n-1) = P(n-2) + (n-2)P(n-3) - P(n-4) + P(n-5)$$

Now by removing the negative term $-P(n-4) + P(n-5)$ from the recursive formula of $P(n-1)$ and multiplying the obtained inequality by $2n^{1/4}$, the following inequality is produced:

$$2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \geq 2n^{1/4}P(n-1) \tag{6}$$

By combining (5) and (6), it is sufficient to prove:

$$P(n-1) + (n-1)P(n-2) - P(n-3) \geq 2n^{1/4}P(n-2) + 2n^{1/4}(n-2)P(n-3) \tag{7}$$

Now, from the induction hypothesis we have:

$$P(n-1) \geq 2(n-1)^{1/4}P(n-2) \tag{8}$$

and

$$(n-2)P(n-2) \geq 2(n-2)(n-2)^{1/4}P(n-3) \tag{9}$$

(where the term $n-2$ is multiplied in both sides of the last inequality). By adding inequalities (8), (9) and the following inequality

$$P(n-2) - P(n-3) \geq P(n-2) - P(n-3)$$

we obtain the following inequality:

$$\begin{aligned} P(n-1) + (n-1)P(n-2) - P(n-3) \;\geq\; & 2(n-1)^{1/4}P(n-2) + P(n-2) \\ & +2(n-2)(n-2)^{1/4}P(n-3) \\ & -P(n-3). \end{aligned} \tag{10}$$

Now, in order to prove (7), it is sufficient to prove that the right side of (10) is greater than or equal to the right side of (7), where the simplified inequality can be written as follows:

$$\begin{aligned} P(n-2) \;\geq\; & 2\left[n^{1/4}-(n-1)^{1/4}\right]P(n-2) \\ & +2\left[(n-2)n^{1/4}-(n-2)(n-2)^{1/4}\right]P(n-3)+P(n-3). \end{aligned} \tag{11}$$

Instead, by using the Lemmas 4 and 5, the inequality (11) becomes correct if the following inequality holds:

$$P(n-2) \geq \frac{P(n-2)}{2(n-1)^{3/4}} + (n-2)\frac{P(n-3)}{(n-2)^{3/4}} + P(n-3). \tag{12}$$

The inequality (12) is equivalence with the following series of inequalities:

$$\begin{aligned} \left[1-\frac{1}{2(n-1)^{3/4}}\right]P(n-2) \quad &\geq \quad (n-2)^{1/4}P(n-3)+P(n-3) \\ &\Longleftrightarrow \\ 2\left[1-\frac{1}{2(n-1)^{3/4}}\right](n-2)^{1/4}P(n-3) \quad &\geq \quad (n-2)^{1/4}P(n-3)+P(n-3) \\ &\Longleftrightarrow \\ 2\left[1-\frac{1}{2(n-1)^{3/4}}\right](n-2)^{1/4} \quad &\geq \quad (n-2)^{1/4}+1 \\ &\Longleftrightarrow \\ \left[1-\frac{1}{(n-1)^{3/4}}\right](n-2)^{1/4} \quad &\geq \quad 1, \end{aligned}$$

where the last one is satisfied for any $n \geq 7$ and this completes the proof. □

Previously, we proved that $I(n)/P(n) > 1$, for any $n \geq 3$. Now, we present the main theorem for the upper bound of $I(n)/P(n)$.

Theorem 1. *There exists a constant number $M > 1$ and a positive integer N, such that $I(n)/P(n) \leq M$, for each $n \geq N$.*

Proof. Recall from [8] that the number of involutions on a set of size n is expressed by $I(n) = I(n-1) + (n-1)I(n-2)$. Suppose that $R(n)$ denotes the fraction $I(n)/P(n)$. We can write $R(n)$ as follows:

$$\begin{aligned} R(n) &= \frac{I(n)}{P(n)} + \frac{I(n)}{P(n-1)+(n-1)P(n-2)} - \frac{I(n)}{P(n-1)+(n-1)P(n-2)} \\ &= \frac{I(n-1)+(n-1)I(n-2)}{P(n-1)+(n-1)P(n-2)} + \frac{I(n)}{P(n)}\left(\frac{P(n-3)-P(n-4)}{P(n-1)+(n-1)P(n-2)}\right). \end{aligned}$$

Now, suppose that $S(n)$ denotes the following fraction:

$$\frac{I(n-1)+(n-1)I(n-2)}{P(n-1)+(n-1)P(n-2)}.$$

Since $R(n-1)=I(n-1)/P(n-1)$ and $R(n-2)=I(n-2)/P(n-2)$, therefore from the elementary calculus we can drive that $S(n)$ is placed between $R(n-1)$ and $R(n-2)$, i.e. $S(n)\leq max\{R(n-1),R(n-2)\}$. So we have:

$$\begin{aligned}
R(n) &= S(n)+\frac{I(n)}{P(n)}\left(\frac{P(n-3)-P(n-4)}{P(n-1)+(n-1)P(n-2)}\right)\\
&= S(n)+R(n)\left(\frac{P(n-3)-P(n-4)}{P(n-1)+(n-1)P(n-2)}\right)\\
&\leq S(n)+R(n)\left(\frac{P(n-3)}{(n-1)P(n-2)}\right)\\
&\leq S(n)+R(n)\left(\frac{1}{(n-1)(n-2)^{1/4}}\right)\\
&\leq S(n)\left(1-\frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}\\
&\leq max\{R(n-1),R(n-2)\}\left(1-\frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}. \qquad (13)
\end{aligned}$$

From the elementary calculus, we know that there exists a positive integer N such that for each $i\geq N$ the following inequality is correct:

$$\left(1-\frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1}\leq\left(1+\frac{2}{(i-1)(i-2)^{1/4}}\right). \qquad (14)$$

From this inequality we have:

$$\prod_{i=N}^{\infty}\left(1-\frac{1}{(i-1)(i-2)^{1/4}}\right)^{-1}\leq\prod_{i=N}^{\infty}\left(1+\frac{2}{(i-1)(i-2)^{1/4}}\right). \qquad (15)$$

The right side of (15) converges to a number, say M_0. So we can write:

$$\begin{aligned}
R(n) &\leq max\{R(N-1),R(N-2)\}\prod_{i=N}^{\infty}\left(1-\frac{1}{(n-1)(n-2)^{1/4}}\right)^{-1}\\
&\leq M_0\times max\{R(N-1),R(N-2)\}. \qquad (16)
\end{aligned}$$

Therefore the theorem is correct if we let $M=M_0\times max\{R(N-1),R(N-2)\}$, where M is a constant number. □

5. Combinatorics

Another way for counting the number of RNA structures is by using the combinatorial properties of this problem and the principle if inclusion-exclusion. To do this, let $I_r(n)$ denotes the number of involutions on a set of n points in which at least r points among n are mapped to their immediate next point (see Figure 5.). To compute $I_r(n)$, we can choose r points among $n-r$ and then insert r extra points right after the selected points without taking care about the other ones. By this formalism, we have the following formula for $I_r(n)$:

$$I_r(n) = \binom{n-r}{r} I(n-2r), \tag{17}$$

where $I(n)$ denotes the number of involutions. In the RNA structure, no point is allowed to make a base pair with its immediate next point, i. e. the number of RNA structures is exactly the number of involutions in which no point is mapped to its immediate next point. By using the inclusion-exclusion principle, the following formula is obtained for the number of RNA structures:

$$\begin{aligned} P(n) &= I_0(n) - I_1(n) + I_2(n) \cdots + (-1)^{\lfloor \frac{n}{2} \rfloor} I_{\lfloor \frac{n}{2} \rfloor}(n) \\ &= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i I_i(n) \\ &= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i \binom{n-i}{i} I(n-2i). \end{aligned} \tag{18}$$

The following theorem indicates that the leading term of the serie (18), i.e. $I_0(n)$, can be considered as an asymptotic behavior for $P(n)$ (It should be noted that $I_0(n) = I(n)$).

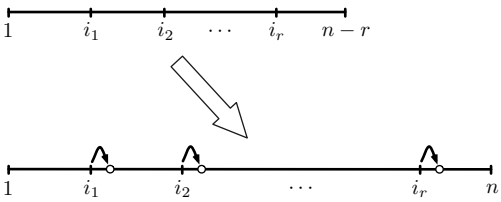

Figure 2: Schematic representation of $I_r(n)$, where the circular points are inserted right after the selected points.

Theorem 2. *For any large value of n,* $I_0(n)/3 \leq P(n) \leq I_0(n)/2$.

Proof. By using the induction, it is easy to show that $I_{i+1}(n) \leq I_i(n)$ $(0 \leq i \leq \lfloor n/2 \rfloor)$. To show that $I_0(n)/3 \leq P(n)$ we do as follows:

$$\begin{aligned}\frac{I_0(n)}{I_2(n)} &= \frac{\binom{n}{0} I(n)}{\binom{n-2}{2} I(n-4)} = \frac{2n^{\frac{n}{2}} \mathrm{e}^{(\frac{-n}{2}+\sqrt{n}-\frac{1}{4})}}{(n-2)(n-3)(n-4)^{\frac{n-4}{2}} \mathrm{e}^{(\frac{-n-4}{2}+\sqrt{n-4}-\frac{1}{4})}} \\ &= \frac{2(\frac{n}{n-4})^{(\frac{n-4}{2})} n^2 \mathrm{e}^{(\sqrt{n}-\sqrt{n-4}-2)}}{(n-2)(n-3)} = \frac{2\mathrm{e}^2 n^2 \mathrm{e}^{-2}}{(n-2)(n-3)} \simeq 2. \end{aligned} \tag{19}$$

By performing the same computation, it is easy to show that $I_0(n)/I_1(n) \simeq 1$ and $I_2(n)/I_3(n) \simeq 3$. Therefore we have:

$$\begin{aligned} P(n) &= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i I_i(n) \\ &\geq I_0(n) - I_1(n) + I_2(n) - I_3(n) \\ &\geq \frac{2}{3} I_2(n) = \frac{1}{3} I_0(n). \end{aligned} \tag{20}$$

On the other hand, we have:

$$\begin{aligned} P(n) &= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i I_i(n) \\ &\leq I_0(n) - I_1(n) + I_2(n) \simeq \frac{1}{2} I_0(n). \end{aligned} \tag{21}$$

The proof is completed by combining the inequalities (20) and (21). □

Since the number of RNA structures grows exponentially, the logarithmic behavior of its growing rate with respect to the boundaries given in the theorem 2 is represented in Figure 3. By looking to this figure, we then conjectured that $P(n) = I(n)/e$ as it is proved in the following theorem.

Theorem 3. *For any large value of n,* $P(n) \simeq I(n)/e$.

Proof. By using the induction, it is easy to show that $I_{i+1}(n) \leq I_i(n)$ $(0 \leq i < \lfloor n/2 \rfloor)$. To prove the theorem, we have:

$$\begin{aligned} P(n) &= \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i I_i(n) \\ &= I_0(n) \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (-1)^i \frac{I_i(n)}{I_0(n)}. \end{aligned} \tag{22}$$

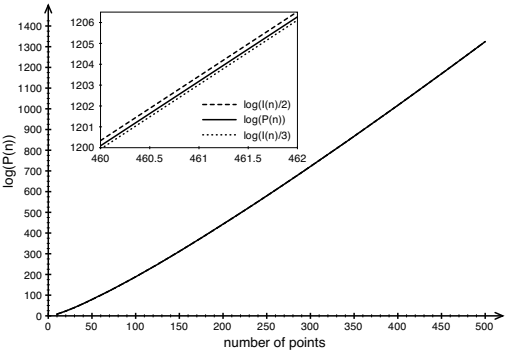

Figure 3: Logarithmic behavior of $P(n)$ with respect to its boundary functions.

Suppose that $S_n = \sum_{i=0}^{\lfloor n/2 \rfloor} (-1)^i (I_i(n)/I_0(n)$. Since $I_{i+1}(n)/I_0(n) < I_i(n)/I_0(n)$ and $\lim_{n\to\infty} I_i(n)/I_0(n) = 0$, so S_n is a convergence serie. Now it is sufficient to show that $\lim_{n\to\infty} S_n = 1/e$. By using the asymptotic formula of $I(n)$ we have:

$$\begin{aligned}
\lim_{n\to\infty} \frac{I_1(n)}{I_0(n)} &= \lim_{n\to\infty} \frac{\binom{n-1}{1} I(n-2)}{I_0(n)} \\
&\simeq \lim_{n\to\infty} \frac{\frac{1}{\sqrt{2}}(n-1)(n-2)^{\frac{n-2}{2}} e^{(-\frac{n-2}{2}+\sqrt{n-2}-\frac{1}{4})}}{\frac{1}{\sqrt{2}} n^{\frac{n}{2}} e^{(-\frac{n}{2}+\sqrt{n}-\frac{1}{4})}} \\
&\simeq \lim_{n\to\infty} \frac{(n-1)(n-2)^{\frac{n}{2}-1} e^{(1+\sqrt{n-2})}}{n^{\frac{n}{2}} e^{\sqrt{n}}} \\
&\simeq \lim_{n\to\infty} \frac{(n-2)^{\frac{n}{2}} e^1}{n^{\frac{n}{2}}} \\
&\simeq 1. \qquad (23)
\end{aligned}$$

By performing the similar calculations, we can show that:

$$\begin{aligned}
\lim_{n\to\infty} \frac{I_2(n)}{I_0(n)} &= \lim_{n\to\infty} \frac{\binom{n-2}{2} I(n-4)}{I_0(n)} \\
&\simeq \lim_{n\to\infty} \frac{1}{2!} \frac{(n-2)(n-3)(n-4)^{\frac{n}{2}-2} e^{(2+\sqrt{n-4})}}{n^{\frac{n}{2}} e^{\sqrt{n}}} \\
&\simeq \lim_{n\to\infty} \frac{1}{2!} \frac{(n-4)^{\frac{n}{2}} e^2}{n^{\frac{n}{2}}} \\
&\simeq \frac{1}{2!}. \qquad (24)
\end{aligned}$$

And in general, we have $\lim_{n\to\infty} I_i(n)/I_0(n) = 1/i!$. By using these results, the proof is

completed as follows:

$$\begin{aligned}\lim_{n\to\infty} S_n &= \sum_{i=0}^{\infty} (-1)^i \frac{I_i(n)}{I_0(n)} \\ &\simeq \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} \\ &\simeq \frac{1}{e}. \end{aligned} \tag{25}$$

□

6. Conclusions

In this paper we have discussed about enumerating the number of RNA structures for a sequence of length n. Some properties of the presented recursive formula (2) are introduced and proved. Also the relation between the number of RNA structures for a sequence of length n and the number of involutions for a set of size n is discussed. The only criteria considered for the RNA structures was that the loops should contain at least one base. Perhaps one of the interesting problems in this area is how to determine the number of RNA structures for a sequence of length n such that each loop has at least l bases and each stem has at least h base pairs.

References

[1] T. Akutsu, Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, *Discr. Appl. Math.* **104** (2000) 45–62.

[2] H. Gan, S. Pasquali, T. Schlick, Exploring the repertoire of RNA secondary motifs using graph theory; implications for rna design, *Nucleic Acids Res.* **31** (2003) 2926–2943.

[3] C. Haslinger, P. Stadler, RNA structures with pseudo-knots: Graph-theoretical, combinatorial and statistical properties, *Bull. Math. Biol.* **61** (1999) 437–467.

[4] I. Hofacker, P. Schuster, P. Stadler, Combinatorics of RNA secondary structures, *Discr. Appl. Math.* **88** (1998) 207–237.

[5] H. Isambert, E. Siggia, Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme, *Proc. Natl. Acad. Sci. USA* **97** (2000) 6515–6520.

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

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

[8] D. E. Knuth, *The Art of Computer Programming: Sorting and Searching*, Vol. 3, Addison–Wesley Longman, Amsterdam, 1998.

[9] R. B. Lyngso, C. N. Pedersen, RNA pseudoknot prediction in energy based models, *J. Comput. Biol.* **7** (2000) 409–428.

[10] M. Nebel, Combinatorial properties of RNA secondary structures, *J. Comput. Biol.* **9** (2002) 541–573.

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

[12] H. Orland, A. Zee, RNA folding and large n matrix theory, *Nucl. Phys. B.* **620** (2002) 456–476.

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

[14] G. Studnicka, G. Rahn, I. Cummings, W. Salser, Computer method for predicting the secondary structure of single-stranded RNA, *Nucleic Acids Res.* **5** (1978) 3365–3387.

[15] M. Tacker, P. Stadler, E. Bornberg-Bauer, I. Hofacker, P. Schuster, Algorithm independent properties of RNA secondary structure prediction, *Eur. Biophys. J.* **25** (1996) 115–130.

[16] I. J. Tinoco, O. Uhlenbeck, M. Levine, Estimation of secondary structure in ribonucleic acids, *Nature* **230** (1971) 362–367.

[17] Y. Uemura, A. Hasegawa, S. Kobayashi, T. Yokomori, Tree adjoining grammars for RNA structure prediction, *Theor. Comput. Sci.* **210** (1999) 277–303.

[18] G. Vernizzi, H. Orland, A. Zee, Enumeration of RNA structures by matrix models, *Phys. Rev. Lett.* **94** (2005) 168103.

[19] G. Viennot, M. Vauchaussade de Chaumont, Enumeration of RNA secondary structures by complexity, *Lect. Notes Biomath.* **57** (1985) 360–365.

[20] M. S. Waterman, Secondary structure of single-stranded nucleic acids, in: *Studies on Foundations and Combinatorics*, Vol. 1, Academic Press, New York, 1978, pp. 167–212.

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