International Journal of Combinatorial GraphTheory and Applications
Vol. 6, No. 1, (January-June, 2021)

Reviewed by: Martin Trinks

Received: 15th March 2021 *Revised: 20th April 2021* *Accepted: 10th May 2021*

ENUMERATION PROBLEM OF RNA SECONDARY STRUCTURES UNDER NEW REPRESENTATION

Wenwen Wang, Tianming Wang

Department of Applied Mathematics, Dalian University of Technology
Dalian 116024, P.R.China
e-mail: wenandwxj@163.com, wangtm@dlut.edu.cn

Abstract. Under new representation, the number of RNA secondary structures has been further studied with the technique of generating function. And according to [7], we get the combinatorial expression of $S^*(n,k)$, then $S^*(n)$ is computed. Finally, we give another simple explicit formulas on $S^*(n)$ and $S^*(n,k)$.

1. Introduction

RNA, DNA and protein are the basic composition composed of lives in the earth [5]. RNA is an important molecule that performs a wide range of functions in biological systems. RNA has recently become the center of much attention because of its catalytic properties [1], leading to an increased interest in obtaining structural information.

RNA molecules are typically described at three different levels: first, the *primary structure* of RNA is a single strand made of the ribonucleotides adenine, cytosine, guanine and uracil. The secondary level by indicating the bonds between pairs of nucleotides. These three levels give the topology of the molecule, and its geometric shape [3].

RNA can fold back on itself. The pairing rules for its sequences in RNA alphabet is A pairs with U and G pairs with C. In addition, frequently G is thought to pair with U and G-U base pairs called wobble base pairs. In this paper, we ignore G-U pairs. The two-dimensional self-pairing is called *secondary structure*. Here, we are only concerned with the enumeration problem of RNA

[0]2000 Mathematics Subject Classification: 39B82, 39B52, 39B55, 46C05, 81Q05.

[0]Keywords: RNA secondary structures, Combinatorial enumeration, Generating function, linear trees.

secondary structures. Previous results on the number of different secondary structures of RNA molecules are due to M. S. Waterman with their coworkers [6, 7, 8, 11, 12] and C. D. Svrtan, et [2]. Particularly important work reported by I.L. Hofacker, P. Schuster, P.F. Stadler is the recursion for the number of secondary structures with limited length loop [4]. E. A. RØLand discussed the enumeration about the secondary structures with Pseudoknots [3].

In this paper, we consider the secondary structures without Pseudoknots. According to a new representation, we compute the total number of RNA secondary structure of a given length $S^*(n)$, and the explicit expression of $S^*(n)$ from the generating function is obtained. Then we give an asymptotic analysis about $S^*(n)$. Finally, And according to [7], we get the combinatorial expression of $S^*(n,k)$, then $S^*(n)$ is computed.

2. The basic Definition

Definition 2.1 ([9] Definition 3.1). *Let $R = r_1r_2\cdots r_n, r_i \in \{A,U\} or \{G,C\}, i = 1,2,\cdots,n$ be the RNA sequence. The* secondary structure *is a vertex-labelled graph on n vertices with an adjacency matrix $A = (r_{ij})$ fulfilling : (1) $r_{i,i+1} = 1, 1 \le i \le n-1$; (2) If $r_{i,k} = 1, k \ne i-1, i+1$, r_i pairs with r_k; (3) For each i there is at most a single $k \ne i-1, i+1$ such that $r_{i,k} = 1$; (4) If $r_{i,j} = r_{k,l} = 1$ and $i < k < j$, then $i < l < j$.*

We will call an edge (i,j), $|i-j| \ne 1$ a bond *(or a base pair). A vertex i connected only to $i-1$ and $i+1$ will be called* unpaired*. A vertex i is said to be* interior *to the base pair (k,l) if $k < i < l$. If, in addition, there is no base pair (p,q) such that $k < p < i < q < l$, we will say that i is* immediately interior *to the base pair (k,l).*

Definition 2.2 ([4] Definition 2.2). *A* stack *consists of subsequent base pairs $(p-k,q+k),(p-k+1,q+k-1),\cdots,(p,q)$ such that neither $(p-k-1,q+k+1)$ nor $(p+1,q-1)$ is a base pair. $k+1$ is the length of the stack. $(p-k,q+k)$ is the terminal base pair of the stack.*

Definition 2.3 ([4] Definition 2.2). *A* bonding loop *consists of a terminal base pair and unpaired vertices. The number of unpaired vertices is the length of the bonding loop.*

Definition 2.4 ([4] Definition 2.4). *A stack $[(p,q),\cdots(p+k,q-k)]$ is called* terminal *if $p-1=0$ or $q+1=n+1$ or if the two vertices $p-1$ and $q+1$ are not interior to any base pair. The sub-structure enclosed by the terminal base pair (p,q) of a terminal stack will be called a* component *of secondary structure. We will say that a structure on n vertices has a* terminal base pair *if $(1,n)$ is a base pair.*

Definition 2.5 ([4]Definition 2.2)**.** *A* external vertex *is an unpaired vertex which dose not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex* 1 *or* n *it is a free end, otherwise it is called joint.*

Definition 2.6 ([4])**.** *A* internal vertex *is an unpaired vertex which is interior to a base pair.*

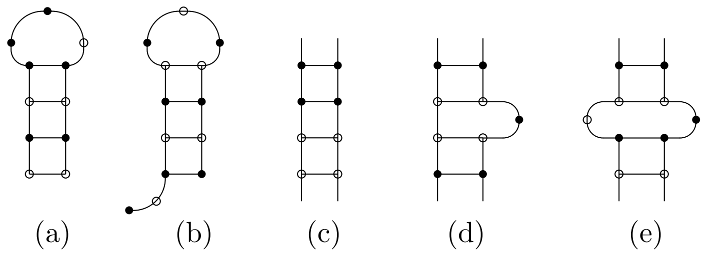

Fig.1 the elements of RNA Secondary Structures.

We consider the difference of the paired bases, $A-U$ and $G-C$. Generally, we ignore the paired bases $G-U$. The notation (a) is a hairpin; the end of (b) is called tail; (c) is a stack; (d) is a convex loop; (e) is called a interior loop.

Now, considered the sequence $r = r_1r_2r_3$, the total possible secondary structures is 12 kinds. It is shown in Fig.2.

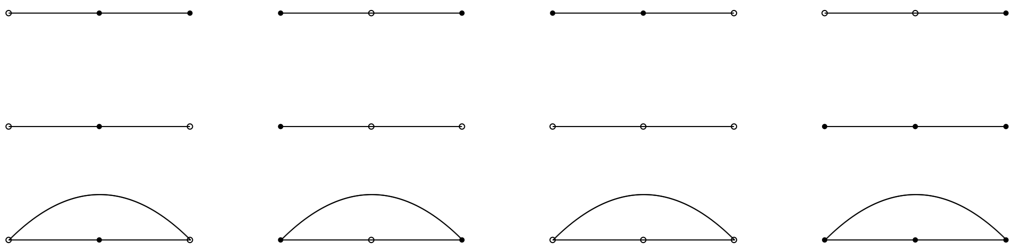

Fig.2 Twelve kinds secondary structures on [3]

3. The recursion formulas

Lemma 3.1 ([9])**.** *Let ordered set* $[n] = \{1, 2, \cdots, n\}$ *and* $S^*(n)$ *be the total number of RNA secondary structures on* $[n]$*, then* $S^*(n)$ *satisfies the recurrence relation:*

$$S^*(n+1) = 2[S^*(n) + \sum_{k=1}^{n-1} S^*(k-1)S^*(n-k)],\ n \geq 1$$

with the boundary values $S^*(0) = 1, S^*(1) = 2$.

Theorem 3.2. *Let* $\varphi(x) = \sum\limits_{n\geq 0} S^*(n)x^n$*, then it fulfills the functional equation*

$$2x^2\varphi(x)^2 - [1 - 2x + 2x^2]\varphi(x) + 1 = 0.$$

In [9], let $m = 1$ in recurrence (2), we can get the desired result.

Theorem 3.3. *The explicit expression of* $S^*(n)$ *is*

$$S^*(n) = \sum_{\substack{4v_1+3v_2\leq n+2\\ v_1,v_2\geq 0}} \frac{(\frac{1}{2})_{n+2-3v_1-2v_2}(-1)^{n+1-4v_1-2v_2}2^{2n+2-6v_1-3v_2}}{(v_1)!(v_2)!(n+2-4v_1-3v_2)!},$$

where $(x)_k = x(x-1)\cdots(x-k+1)$ *for any* $x \in \mathbb{C}$ *and* $k \in \mathbb{N}$, $(x)_0 = 1$.

To prove this theorem let the generating function of $S^*(n)$ be $y(x) = \sum_{n\geq 0} S^*(n)x^n$. According to Theorem 2, we get the algebraic function satisfied by $y(x)$, that is, $2x^2y^2 - (1-2x+2x^2)y + 1 = 0$. By the initial condition, we can deduce that

$$y(x) = \frac{1-2x+2x^2}{4x^2} - \frac{1}{4x^2}(1+4x^4-8x^3-4x)^{\frac{1}{2}}$$

Computing the above function equation, we get multinomial identity

$$y(x) = \frac{1-2x+2x^2}{4x^2} - \sum_{\substack{4v_1+3v_2\leq n+2\\ v_1,v_2\geq 0}} \binom{\frac{1}{2}}{v_1\ v_2\ n+2-4v_1-3v_2} (-1)^{n+1-4v_1-2v_2}2^{2n+2-6v_1-3v_2}\cdot x^n$$

Hence, by virtue of the above identity , we obtain the coefficient of x^n, which is just the explicit value of $S^*(n)$. That is,

$$S^*(n) = \sum_{\substack{4v_1+3v_2\leq n+2\\ v_1,v_2\geq 0}} \frac{(\frac{1}{2})_{n+2-3v_1-2v_2}\cdot(-1)^{n+1-4v_1-2v_2}\cdot 2^{2n+2-6v_1-3v_2}}{(v_1)!(v_2)!(n+2-4v_1-3v_2)!}.$$

Lemma 3.4 ([8])**.** *Let* $y(x) = \sum_{n=0} a_nx^n$ *be the ordinary generating function of the sequence* a_n *which is known to have the property* $a_n \geq 0$*. let* y *satisfy the functional equation* $F(x,y) = 0$*, and let* $r > 0$, $s > a_0$ *be the unique real solutions of the system* $F(r,s) = 0$, $F_y(r,s) = 0$*. Then*

$$a_n \sim \sqrt{\frac{rF_x(r,s)}{2\pi F_{yy}(r,s)}}\, n^{-\frac{3}{2}}r^{-n}.$$

Theorem 3.5. $S^*(n) \sim \sqrt{\dfrac{1+(1-\sqrt{2})r}{4\pi r^2}}\, n^{-\frac{3}{2}}r^{-n}$.

In the present case, by Theorem 3.2, we can let

$$F(x,y) = 2x^2y^2(x) - (1-2x+2x^2)y(x) + 1,$$

and we have the following relations

$$2r^2s^2 - (1-2r+2r^2)s + 1 = 0\,,$$
$$4r^2s - (1-2r+2r^2) = 0\,, \quad r \neq 0\,,$$

then $s = \frac{1}{\sqrt{2}r}$. Hence, according to Lemma 4, we get

$$F_x(r,s) = 4rs^2 - 4rs + 2,\; F_{yy}(r,s) = 4r^2.$$

Then the desired result is obtained.

$S^*_{n,k}$ is the set of secondary structures on n vertices that have exactly k base pairs and the bonding loop with limited length 1. In [9], let $m = 1$ in Theorem 5.1.1, we can get the following result.

Theorem 3.6. *Let $S^*(n,k) = |S_{n,k}|$, then*

$$\begin{aligned} S^*(n,k) =& 2[S^*(n-1,k) + S^*(n-2,k-1) + \sum_{j=2}^{n-2}\sum_{i=0}^{k-1} S^*(j-1,i) \\ & S^*(n-j-1,k-1-i)],\; n \geq 4\,, \quad 0 \leq k \leq \lfloor \frac{n-1}{2} \rfloor. \end{aligned}$$

with the boundary values $S^*(n,0) = 2^n,\; n \geq 1,\; S^*(0,k) = 0,\; k \geq 0.$

Theorem 3.7. *The generating function of $S^*(n,k)$ is denoted by $\phi(x,y)$, then it satisfied the following recurrence:*

$$\phi(x,y) = \frac{1}{4x^2y}[1-2x-2x^2y - \sqrt{(1-2x-2x^2y)^2 - 16x^3y}].$$

From Theorem 3.6, we get

$$\begin{aligned} \phi(x,y) &= \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} S^*(n,k)x^ny^k \\ &= \sum_{n=0}^{3}\sum_{k=0}^{\infty} S^*(n,k)x^ny^k + \sum_{n=4}^{\infty}\sum_{k=0}^{\infty} S^*(n,k)x^ny^k \\ &= 2x + (2x)^2 + (2x)^3 + 2x\phi + 2x^2y\phi - 2x(2x + (2x)^2) + 2x^2y\phi^2 \\ &= 2x + 2x\phi + 2x^2y\phi + 2x^2y\phi^2 \end{aligned}$$

From which we get the following identity,

$$2x^2y\phi^2 - [1-2x-2x^2y]\phi + 2x = 0.$$

By the initial condition, we can get the desired result.

Theorem 3.8.

$$S^*(n,k) = \sum_{\substack{2v_1+v_2\le n-2k\\ v_2+2v_3\le k\\ v_1,v_2,v_3\ge 0}} (2)^{n-k}(-2)^{n-k-2v_1-v_2-2v_3} \binom{\frac{1}{2}}{v_1\ n-2k-2v_1-v_2\ k-v_2-2v_3\ v_2\ v_3}$$

According to Theorem 3.7, $\phi(x,y)$ can be written in the form

$$\phi(x,y) = \frac{1-2x-2x^2y}{4x^2y} - \frac{1}{4x^2y}[(1-2x-2x^2y)^2-16x^3y]^{\frac{1}{2}},$$

We only need decompose the part

$$g(x,y) = -\frac{1}{4x^2y}[(1-2x-2x^2y)^2-16x^3y]^{\frac{1}{2}}.$$

By virtue of multinomial identity, we have

$$\begin{aligned}
&g(x,y)\\
&= -\frac{1}{4x^2y}\sum_{v_1v_2v_3v_4v_5}\binom{\frac{1}{2}}{v_1\ v_2\ v_3\ v_4\ v_5}(4x^2)^{v_1}(-4x)^{v_2}(-4x^2y)^{v_3}(-8x^3y)^{v_4}(4x^4y^2)^{v_5}\\
&= \sum_{v_1v_4v_5}\binom{\frac{1}{2}}{v_1\ n-2v_1-v_4-2k\ k-v_4-2v_5\ v_4\ v_5}2^{n-k-1}(-2)^{n-k-2v_1-v_4-2v_5-1}x^{n-2}y^{k-1}\\
&= \sum_{\substack{2v_1+v_2\le n-2k\\ v_2+2v_3\le k\\ v_1,v_2,v_3\ge 0}} 2^{n-k}(-2)^{n-k-2v_1-v_2-2v_3}\binom{\frac{1}{2}}{v_1\ n-2k-2v_1-v_2\ k-v_2-2v_3\ v_2\ v_3}x^ny^k.
\end{aligned}$$

The coefficient of x^ny^k in the above recurrence is the expected result.

Theorem 3.9. $S^*(n) = \sum\limits_{k=0}^{\lfloor\frac{n-1}{2}\rfloor} S^*(n,k)$.

It is much easier to compute $S^*(n,k)$ than $S^*(n)$. In [7], Waterman had been constructed a bijection between $\Phi_{n,k}$ to one set of tree denoted by $\Gamma_{n,k}$. Let $s^*(n,k) = |\Phi_{n,k}|$ and $t(n,k) = |\Gamma_{n,k}|$. Now we give another simple explicit formula on $S^*(n)$ and $S^*(n,k)$.

Lemma 3.10 ([7] Proposition 2.1)**.** *For all $n,k\ge 1$, there exists a bijection*

$$\varphi : \Phi_{n+k-2,\ k-1} \to \Gamma_{n,k}.$$

In [7], the explicit expression about $s^*(n,k)$ is given according to Lemma 3.10. It is shown as the following Lemma.

Lemma 3.11 ([7] Theorem 2.2.)**.** *The number of secondary structures over a sequence of length n having exactly k pairs is given by*

$$s^*(n,k) = \frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1}$$

for $n, k \geq 0$.

In this paper, we consider the difference between $A-G$ base pairs and $G-U$ base pairs. By the construction of the $\Gamma_{n,k}$, we can dye two colors in each branch of the tree. Then we get the combinatoric expression of $S^*(n,k)$.

Theorem 3.12.

$$S^*(n,k) = 2^{n-k}\frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1}.$$

The proof is omitted.

Corollary 3.13. $S^*(n) = \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} 2^{n-k}\frac{1}{k}\binom{n-k}{k+1}\binom{n-k-1}{k-1}$.

By virtue of Theorem 3.9. and Theorem 3.12., we can easily get the result.

References

[1] T. Cech, B. Bass, *Biological catalysis by RNA*, Annual Review of Biochemistry **55** (1988) 599–629.
[2] C. D. Svrtan, D. Veljan, *Enumerative aspects of secondary structures*, Discr. Math. **285** (2004) 67–82.
[3] E. A. RØLand, *Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence*, J. Comp. Biol. **13** (2006) 1197–1213.
[4] I. L. Hofacker, P. Schuster, P. F. Standler,*Combinatorics of RNA secondary structures*, Disc. Appl. Math. **88** (1998) 207–237.
[5] J. Setubal, J. Meidanis, *Introduction to computational molecular biology*, 1997.
[6] J. A. Howell, T. F. Smith, M. S. Waterman, *Computation of generating function for biological molecules*, SIAM J. Appl. Math. **39** (1980) 119–133.
[7] W. R. Schmitt, M. S. Waterman, *Liner trees and RNA secondary structure*, Discr. Appl. Math. **12** (1994) 317–323.
[8] P. R. Stein, M. S. Waterman, *On some new sequences generalizing the Catalan and Motzkin numbers*, Disc. Math. **26** (1978) 261–272.
[9] W. W. Wang, M. Zhang, T. M. Wang, *A new enumeration problem of RNA secondary structure*, Utilitas Mathematics accepted.
[10] W. W. Wang, T. M. Wang, *The enumeration of various types of constrained secondary structure*, WSEAS Transactions on Biology and Biomedicine, **3** (2006) 77–80.
[11] M. S. Waterman, *Secondary structure of single-stranded nucleic acids*, Adv. Math. Suppl. Stud. **1** (1978) 167–212.
[12] M. S. Waterman, T. F. Smith, *RNA secondary structure: A complete mathematical analysis*, Math. Biosic. **42** (1978) 257–266.