Counting RNA loop interaction networks of homology group rank zero

Zi-Wei Bai, Ricky X. F. Chen *

School of Mathematics, Hefei University of Technology

Hefei, Anhui 230601, P. R. China

xiaofengchen@hfut.edu.cn

Michael Fuchs †

Department of Mathematical Sciences, National Chengchi University

Taipei, 116, Taiwan

mfuchs@nccu.edu.tw

Abstract. Enumerative study of RNA secondary structures according to various characteristics is a topic of key importance in computational biology. RNA secondary structure pairs have been also studied in various contexts. Recently, the homology groups of the simplicial complices induced by pairs of secondary structures have been studied by Bura, He and Reidys, providing a new way for characterizing these structure pairs. In particular, the homology group H_2 corresponding to any pair has been shown to be a free group. In this paper, we provide enumerative results, both exactly and asymptotically, for those pairs giving a free group of rank zero. The asymptotic number of these structure pairs of length n is shown to be $cn^{-3/2}4.8105752536^n$. We also prove that the distribution of the

*ORCID: 0000-0003-1061-3049

†ORCID: 0000-0001-8891-6897

number of base pairs in those pairs of secondary structures is asymptotically normal.

Keywords: RNA secondary structure, homology group, loop, central limit theorem
MSC 2010: 05C05, 05A19, 05A15

1 Introduction

Ribonucleic acid (RNA) plays an important role in various biological processes within cells, ranging from catalytic activity to gene expression. An RNA molecule may consist of four types of bases: A (adenine), U (uracil), G (guanine), and C (cytosine). These bases form base pairs where A pairs with U while G pairs with C (and sometimes the non-Watson-Crick base pair G with U). The configuration representing the pairing relation of the bases in an RNA is referred to as the secondary structure of the RNA.

More than four decades ago, Waterman and his coworkers pioneered the combinatorics of RNA secondary structures [21–23]. Since then, the combinatorics of RNA secondary structures has been one of the most important topics in computational biology, see [3–6, 10–12,15–20] and references therein. The notion of bi-secondary structures was introduced by Haslinger and Stadler [13] in order to study pseudoknotted structures. Informally, a bi-secondary structure is a secondary structure with possibly crossing base pairs such that there exists a way of partitioning the base pairs into two families and each family does not have crossing base pairs. As such, bi-secondary structures can be viewed as pairs of secondary structures of the same length without pseudoknots (i.e., without crossing pairs), and they are also studied in other contexts, for example, RNA riboswitches (i.e., base sequences that exhibit two stable configurations [9]). Recently, the homology of simplicial complices built from pairs of RNA secondary structures has been studied by Bura, He and Reidys [1, 2]. As for the associated homology groups, H_2 is particularly interesting. It is proved that for any pair of secondary structures, H_2 is a free group of

rank k for some $k \geq 0$. In [2], a combinatorial characterization for the pairs of secondary structures which give rise to a rank k group H_2 is provided. As such, the problem of enumerating pairs of secondary structures according to the ranks of their corresponding group H_2 is naturally motivated. In this paper, we only enumerate these pairs with H_2 being of rank zero. It is worth noting that a bi-secondary structure may admit multiple different partitions into a pair of secondary structures without pseudoknots, but all such pairs give the same H_2. See [2]. Our computation here distinguishes two pairs of secondary structures representing the same bi-secondary structure, and explicitly enumerating bi-secondary structures according to their H_2 ranks is open.

The paper is organized as follows. In Section 2, we give a brief review of RNA secondary structures and their loop homology. We obtain some functional equations satisfied by the generating functions counting pairs of secondary structures giving a rank zero H_2, with or without tracking the number of base pairs, in Section 3. In Section 4, based on the functional equations obtained in the previous section, we derive some asymptotics and a central limit theorem regarding the distribution of the number of base pairs.

2 RNA loop homology

Recall RNA secondary structures (without pseudoknots) defined in Waterman [23]. An *RNA secondary structure* of length n is a simple labeled graph with vertices in $[n] = \{1, 2, \ldots, n\}$ and edges in E that satisfies:

- if $(i, j) \in E$, then $|i - j| \geq 2$;
- if $(i, j) \in E$ and $(k, l) \in E$, where $i < j$ and $k < l$, and $[i, j] \bigcap [k, l] \neq \varnothing$, then either $[i, j] \subset [k, l]$ or $[k, l] \subset [i, j]$ (where $[i, j]$ denotes the interval $\{r : i \leq r \leq j\}$).

The vertices represent the bases while the edges represent the *base pairs* of an RNA. We represent an RNA secondary structure as a diagram with all vertices arranged in a

horizontal line and its edges as arcs in the upper (or lower) half-plane. According to the above definition, any two arcs do not cross. A vertex not incident to any arc is called an *isolated base*. We say an arc (i_1, j_1) (resp. an isolated base k) is covered by an arc (i, j) if $[i_1, j_1] \subset [i, j]$ (resp. $k \in [i, j]$).

Loops in RNA secondary structures have been extensively studied due to their importance for certain energy models predicting the folded secondary structure of a given RNA base sequence. Following [1], a *loop* s is a subset of vertices, represented as a union of intervals, $s = \bigcup_{i=1}^{k}[a_i, b_i]$, such that (a_1, b_k) and (b_i, a_{i+1}), for $1 \leq i < k$, are arcs and such that any other interval-vertices are isolated bases. If $k + 1$ loops (as subsets) have a non-empty intersection, then they induce a k-simplex. As such, a simplicial complex can be constructed from the disjoint union of loops of a pair of secondary structures of the same length by including all possible induced simplices. In [1], the homology group of the constructed simplicial complex of a pair of RNA secondary structures were investigated. It was proved H_2 is a free group of rank k for some $k \geq 0$. In a later paper [2], a combinatorial characterization for the pairs of secondary structures which give rise to a rank k group H_2 is provided. In summary, the rank k is determined by the number of crossing components in the pair of structures under consideration, and we will not go to details and refer to [1, 2].

Here we merely deal with the secondary structure pairs whose H_2 group is rank 0, i.e., trivial. According to [2], these pairs are those with no crossing arcs where one arc is from the first secondary structure and the other belongs to the secondary structure. That is, for a pair (S, T) of RNA secondary structures of length n, there do not exist a base pair (i_1, i_2) from S and a base pair (j_1, j_2) from T such that $i_1 < j_1 < i_2 < j_2$. See an example in Figure 1, where the arcs of S are placed in the upper half-plane while the arcs of T are placed in the lower half-plane. The main task of this paper is enumerating these pairs of RNA secondary structures.

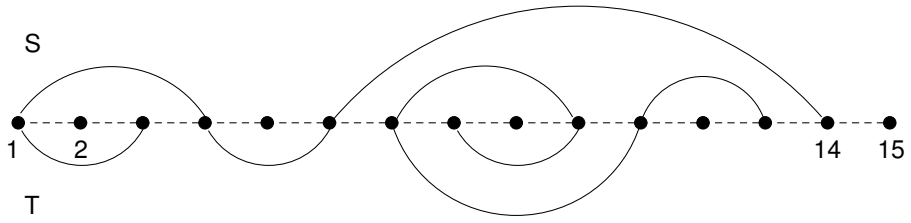


Figure 1: An example of a pair of RNA secondary structures of length 15 with its homology group H_2 being of rank zero.

3 Exact enumeration

Let

$$F(x,y) = \sum_{n\geq 0, k\geq 0} f_{n,k} x^n y^k,$$

where $f_{n,k}$ is the number of pairs of secondary structures of length n and having in total k base pairs which give a rank zero H_2. We make the convention $f_{0,0} = 1$. Thus, $F(x,y)-1$ is the generating function for the corresponding pairs of non-empty (i.e., with at least one base) secondary structures.

Now we are ready to present our first main result.

Theorem 3.1. *The function $F(x,y)$ satisfies the following relation:*

$$\begin{aligned} F(x,y) &= 1 + xF(x,y) + x^2y(y+2)F(x,y)[F(x,y)-1] \\ &\quad + 2xyG(x,y)[3F(x,y)-1] + 2yG(x,y)^2 + \frac{2x^4y^4[F(x,y)-1]^3F(x,y)}{1-x^2y^2[F(x,y)-1]^2} \\ &\quad + \frac{2yG(x,y)F(x,y)(x^2y[F(x,y)-1] + xyG(x,y))}{1-xy[F(x,y)-1]}, \end{aligned} \tag{1}$$

where $G(x,y) = \frac{x^2yF(x,y)[F(x,y)-1]}{1+xy-2xyF(x,y)}$.

Proof. It is crucial to define an auxiliary function

$$G(x,y) = \sum_{n\geq 3, k\geq 1} g_{n,k} x^n y^k,$$

where $g_{n,k}$ is the number of pairs of secondary structures (S,T) of length n and k arcs in total such that base 1 (i.e., the leftmost base) is paired and only paired in S. Obviously, $G(x,y)$ also counts the number of pairs (S,T) where the rightmost base is paired and only paired in S.

For a pair of non-empty secondary structures (S,T), base 1 (the leftmost base) is clearly either not paired in both S and T, or paired in at least one structure. We next consider their respective contributions to $F(x,y)$. These pairs of the former obviously contribute $xF(x,y)$. For the latter, we distinguish the cases as follows:

- Case 1: If both S and T have $(1,k)$ for some $k>1$ as a base pair, then these structure pairs contribute $x^2y^2F(x,y)[F(x,y)-1]$. Namely, the pair $(1,k)$ contributes x^2y^2, the structure covered by the arc $(1,k)$ which cannot be empty by definition (i.e., two adjacent bases cannot form an arc) contributes $F(x,y)-1$, and the structure to the right of the arc $(1,k)$ (which may be empty) contributes $F(x,y)$.

- Case 2: $(1,k)$ is a base pair in S and $j<k$ if $(1,j)$ is a base pair in T.

- Case 3: $(1,k)$ is a base pair in T and $j<k$ if $(1,j)$ is a base pair in S.

Case 2 and Case 3 are essentially the same since turning a pair in Case 2 upside down gives rise to a pair in Case 3. So we shall only discuss Case 2 here. We classify the structure pairs of Case 2 into categories below:

2(i) The category that both 1 and k are not paired in T gives $x^2yF(x,y)[F(x,y)-1]$. See Figure 2 for an illustration. In Figure 2 and those coming later, F stands for $F(x,y)$ and G stands for $G(x,y)$.

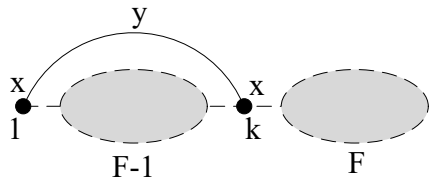


Figure 2: An illustration for 2(i) structure pairs.

2(ii) Note that in Case 2, if 1 is paired in T, it can be only paired with j for $j < k$ by assumption. Suppose $(1, j)$ for $j < k$ is a pair in T and k is not paired in T. Then, the contribution of these structures is $xyG(x, y)F(x, y)$. Due to the noncrossing property, these structure pairs can be first regarded as two independent parts: the structure on $[k]$ and the remaining part which is counted by $F(x, y)$. For the part on $[k]$, if the arc $(1, k)$ and base k are removed, the remaining is easily seen to be counted by $G(x, y)$. Since there is only one way to put base k and the arc $(1, k)$ back, we just need to add a factor xy (for base k and the arc).

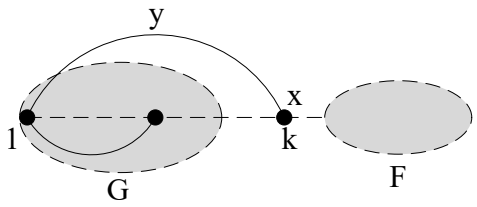
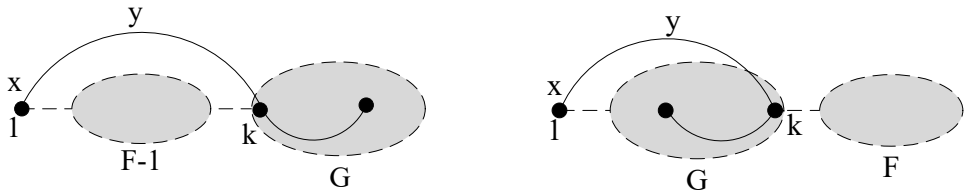


Figure 3: An illustration for 2(ii) structure pairs.

2(iii) Suppose 1 is not paired in T and k is paired in T. There are two cases: (a) k is paired with a base on its righthand side, and (b) k is paired with a base on its lefthand side. For a pair in case (a), base 1 and the arc $(1, k)$ contributes xy, the necessarily non-empty structure covered by $(1, k)$ contributes $F(x, y) - 1$, and the remaining part may be viewed as a pair with k as the leftmost base and paired and only paired in T whence contributing $G(x, y)$. Thus, the pairs in case (a) are counted by $xyG(x, y)[F(x, y) - 1]$. For a pair in case (b), base 1 and the arc $(1, k)$ contributes xy, the structure on $[k]$ with base 1 and the arc $(1, k)$ removed contributes $G(x, y)$,

and the structure (possibly empty) to the right of k is clearly counted by $F(x, y)$. Thus, the pairs in case (b) are counted by $xyG(x, y)F(x, y)$. In summary, the two cases (a) and (b) together contribute $xyG(x, y)[2F(x, y) - 1]$.



(a) 2(iii)(a) (b) 2(iii)(b)

Figure 4: Illustrations for 2(iii) structure pairs.

2(iv) Consider the pairs where both 1 and k are paired in T but $(1, k)$ is not a base pair in T (i.e., excluding those of Case 1). We distinguish three subcases:

(a) k is paired with a base on its righthand side. This subcase is similar to the case 2(iii)(a) and contributes $yG(x, y)^2$.

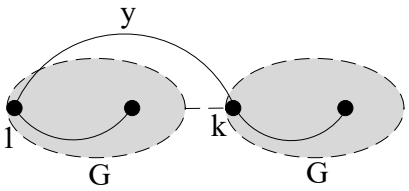


Figure 5: An illustration for 2(iv)(a) structure pairs.

(b) k is paired with a base on its lefthand side, and 1 and k are connected by a sequence of alternating arcs, i.e., for some $r \geq 1$ and for $1 < t_1 < t_2 < \cdots < t_{2r+1} = k$, $(1, t_1)$ and (t_i, t_{i+1}) (for $1 \leq i \leq 2r$) are base pairs. Note that by construction, $(1, t_1)$ is a base pair in T, (t_1, t_2) is a base pair in S, (t_2, t_3) is a base pair in T, and so on. Moreover, inside each such an arc there is a non-empty structure. Thus, the structure on $[k]$ contributes

$$\sum_{r \geq 1} x^{2r+2} y^{2r+2} [F(x, y) - 1]^{2r+1} = \frac{x^4 y^4 [F(x, y) - 1]^3}{1 - x^2 y^2 [F(x, y) - 1]^2}.$$

Taking into account the possible empty structure to the right of k, this subcase contributes $\frac{x^4y^4[F(x,y)-1]^3F(x,y)}{1-x^2y^2[F(x,y)-1]^2}$.

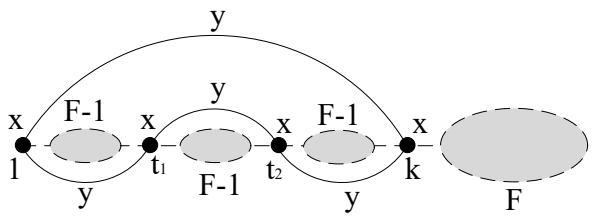


Figure 6: An illustration for 2(iv)(b) structure pairs.

(c) k is paired with a base on its lefthand side, and 1 and k are not connected by a sequence of alternating arcs. Suppose for some $r \geq 0$ and for $t < t_1 < t_2 < \cdots < t_r < k$, $(1,t)$, (t,t_1) and (t_i,t_{i+1}) (for $1 \leq i < r$) are base pairs, and (t_r,z) is not a base pair for any $z > t_r$. First, the part to the right of base k is obviously $F(x,y)$. As for the part on $[k]$, it can be viewed as three independent components: the arc $(1,k)$, the one on $[t_r]$ (t_0 is treated as t) and the rest. For the one on $[t_r]$, depending on whether t_r is paired or not, the contributions are $G(x,y)x^{r+1}y^{r+1}[F(x,y)-1]^r$ and $x^{r+2}y^{r+1}[F(x,y)-1]^{r+1}$, respectively. The arc $(1,k)$ and the rest component together contribute $yG(x,y)$. Hence, for a fixed $r \geq 0$, the contribution of those structure pairs is given by

$$yF(x,y)G(x,y)\Big(x^{r+2}y^{r+1}[F(x,y)-1]^{r+1} + G(x,y)x^{r+1}y^{r+1}[F(x,y)-1]^r\Big).$$

Thus, the overall contribution of this subcase is

$$\begin{aligned} &yG(x,y)F(x,y)(x^2y[F(x,y)-1] + xyG(x,y))\sum_{r\geq 0} x^r y^r [F(x,y)-1]^r \\ =&\frac{yG(x,y)F(x,y)(x^2y[F(x,y)-1] + xyG(x,y))}{1-xy[F(x,y)-1]}. \end{aligned}$$

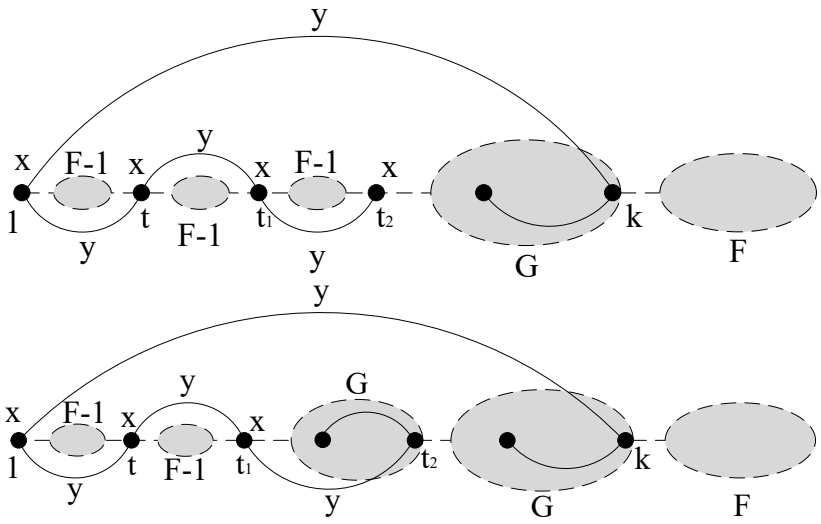


Figure 7: Illustrations for 2(iv)(c) structure pairs.

In summary, we have

$$
\begin{aligned}
F(x,y) &= 1 + xF(x,y) + x^2y^2F(x,y)[F(x,y)-1] \\
&\quad + 2 \times \Bigg\{ x^2yF(x,y)[F(x,y)-1] + xyG(x,y)[3F(x,y)-1] + yG(x,y)^2 \\
&\quad + \frac{x^4y^4[F(x,y)-1]^3F(x,y)}{1-x^2y^2[F(x,y)-1]^2} + \frac{yG(x,y)F(x,y)[x^2y(F(x,y)-1)+xyG(x,y)]}{1-xy[F(x,y)-1]} \Bigg\}.
\end{aligned}
$$

Analogously, for $G(x,y)$, we have the following three cases.

- Case $1'$: If $(1,k)$ for some $k>1$ is a pair in S and k is not paired in T, then the contribution is $x^2y[F(x,y)-1]F(x,y)$.

- Case $2'$: If $(1,k)$ is a base pair in S and (j,k) is a base pair in T for some $1<j<k$, then this case contributes $xyG(x,y)F(x,y)$.

- Case $3'$: If $(1,k)$ for some $k>1$ is a pair in S and k is paired to a base on its righthand side in T, the contribution is $xy[F(x,y)-1]G(x,y)$.

Together we obtain

$$G(x,y) = \frac{x^2[F(x,y)-1]F(x,y)}{1-x[2F(x,y)-1]},$$

completing the proof of the theorem. □

We can of course derive from eq. (1) a recurrence for $f_{n,k}$ which is not elegant though. It seems not easy to derive an explicit formula for $F(x,y)$ and thus for $f_{n,k}$. But, we will still be able to obtain some information for $f_{n,k}$ in the next section.

By setting $y = 1$ in Theorem 3.1, we immediately obtain the following corollary.

Corollary 3.2. *Let $F(x) = \sum_{n\geq 0} f_n x^n$, where f_n is the number of pairs of secondary structures of length n which give a rank zero H_2 with the convention $f_0 = 1$. Then, we have*

$$\begin{aligned} F(x) = 1 + xF(x) + 3x^2F(x)[F(x)-1] + 2xG(x)[3F(x)-1] + 2G(x)^2 \\ + \frac{2x^4F(x)[F(x)-1]^3}{1-x^2[F(x)-1]^2} + \frac{2G(x)F(x)(x^2[F(x)-1]+xG(x))}{1-x[F(x)-1]} \end{aligned} \tag{2}$$

where $G(x) = \frac{x^2F(x)[F(x)-1]}{1+x-2xF(x)}$.

The first few entries of f_n's are given by

$$F(x) = 1 + x + x^2 + 4x^3 + 14x^4 + 50x^5 + 191x^6 + 751x^7 + 3018x^8 + 12371x^9 + \cdots .$$

Again, although we are not able to obtain an explicit exact formula for f_n, we will present its asymptotical behavior shortly.

4 Asymptotics

In this section, we derive the first-order asymptotics of f_n from Corollary 3.2 and show that $f_{n,k}$ from Theorem 3.1 satisfies a central limit theorem. For this, we need the following

result from [7]; see also Section 2.2.3 in [8].

Theorem 4.1 (Drmota)**.** *Let $H(z,w,u)$ be a function which is analytic at $z=w=0$ and $u=1$ and satisfies that $H(0,w,u)\equiv 0$ and $H(z,0,u)\not\equiv 0$. Assume that the Taylor coefficients of H at $z=w=0$ and $u=1$ are non-negative. Moreover, assume that there exist positive z_0, w_0 inside the domain of analyticity of $H(z,w,u)$ such that*

$$w_0 = H(z_0,w_0,1), \quad 1 = H_w(z_0,w_0,1), \quad H_z(z_0,w_0,1)\neq 0, \quad H_{ww}(z_0,w_0,1)\neq 0. \tag{3}$$

Then, there exits a unique solution of

$$w(z,u) = H(z,w(z,u),u)$$

which satisfies $w(0,u)=0$ and is analytic at $z=0$ and $u=1$. Moreover, if the Taylor coefficients of $w(z,1)$ at $z=0$ are eventually positive, then

$$[z^n]w(z,u) = \sqrt{\frac{f(u)H_z(f(u),w(f(u),u),u)}{2\pi H_{ww}(f(u),w(f(u),u),u)}}\, f(u)^{-n} n^{-3/2}\left(1+\mathcal{O}(n^{-1})\right) \tag{4}$$

uniformly for u sufficiently close to 1, where $f(u)$ is an analytic function at $u=1$ with $f(1)=z_0$ and $w(z_0,1)=w_0$.

Applying this result to $F(x)$ from Corollary 3.2, we obtain the following theorem.

Theorem 4.2. *For the number f_n of pairs of secondary structures of length n which give rank zero H_2, we have, as $n\to\infty$,*

$$f_n = (0.2774624151\cdots)(4.8105752536\cdots)^n n^{-3/2}\left(1+\mathcal{O}(n^{-1})\right).$$

Proof. Set $\tilde{F}(x) = F(x) - 1$. Then, the result in Corollary 3.2 becomes

$$\begin{aligned}\tilde{F}(x) = x[\tilde{F}(x)+1] + 3x^2[\tilde{F}(x)+1]\tilde{F}(x) + 2x\tilde{G}(x)[3\tilde{F}(x)+2] + 2\tilde{G}(x)^2 \\ + \frac{2x^4[\tilde{F}(x)+1]\tilde{F}(x)^3}{1-x^2\tilde{F}(x)^2} + \frac{2\tilde{G}(x)[\tilde{F}(x)+1][x^2\tilde{F}(x)+x\tilde{G}(x)]}{1-x\tilde{F}(x)},\end{aligned}$$

where $\tilde{G}(x) = \frac{x^2[\tilde{F}(x)+1]\tilde{F}(x)}{1-x-2x\tilde{F}(x)}$. Thus, if we set

$$\begin{aligned}H(z,w) = z(w+1) + 3z^2(w+1)w + \frac{2z^3(3w+2)(w+1)w}{1-z-2zw} + \frac{2z^4(w+1)^2w^2}{(1-z-2zw)^2} \\ + \frac{2z^4(w+1)w^3}{1-z^2w^2} + \frac{2z^4(w+1)^2w^2}{(1-zw)(1-z-2zw)} + \frac{2z^5(w+1)^3w^2}{(1-zw)(1-z-2zw)^2},\end{aligned}$$

then

$$\tilde{F}(x) = H(x, \tilde{F}(x)).$$

Consequently, we are in the setting of Theorem 4.1 (without the dependence on u). Note also that f_n is eventually positive. Thus, we only have to check that the function $H(z,w)$ above satisfies all the required assumptions from Theorem 4.1.

First, $H(z,w)$ clearly satisfies $H(0,w) \equiv 0$ and $H(z,0) \not\equiv 0$. Moreover, $H(z,w)$ is analytic, e.g., in the region

$$\mathcal{D} = \{(z,w) \ : \ |z| < 1/3, |w| < 3/5\}$$

as on $\mathcal{D}$, we have that

$$|zw| < 1/5 < 1, \quad |z^2w^2| < 1/25 < 1, \quad \text{and} \quad |z+2zw| = |z| \cdot |1+2\omega| < 11/15 < 1.$$

In addition, the Taylor coefficients of $H(z,w)$ are all positive as

$$\frac{1}{1-zw}=\sum_{n\geq 0} z^n w^n, \qquad \frac{1}{1-z^2w^2}=\sum_{n\geq 0} z^{2n} w^{2n},$$

and

$$\frac{1}{1-z-2zw}=\sum_{n\geq 0}\frac{2^n z^n}{(1-z)^{n+1}}w^n=\sum_{n,m\geq 0} 2^n\binom{n+m}{m} z^{n+m}w^n.$$

Thus, what remains is to check (3). First, by solving the first two equations of (3) with mathematical software, we find that

$$z_0=0.2078753469\cdots \qquad \text{and} \qquad w_0=0.5525053505\cdots$$

which lie in $\mathcal{D}$. Next, again with mathematical software,

$$H_z(z_0,w_0)=4.9272739169\cdots\neq 0 \qquad \text{and} \qquad H_{ww}(z_0,w_0)=2.1174906457\cdots\neq 0.$$

Finally, by plugging everything into (4), we obtain that (where we set $u=1$):

$$f_n=(0.2774624151\cdots)(4.8105752536\cdots)^n n^{-3/2}(1+\mathcal{O}(n^{-1}))$$

which is the claimed result. □

In Table 1, we present the exact numbers and the asymptotics of f_n for some n's.

Table 1: The exact and asymptotic values of f_n for some n.

n	exact	asymptotic
10	51495	58232.5
20	128387692013	1.36643×10^{11}
30	473467997674019937	4.93645×10^{17}
40	2062303553810701768953425	2.128×10^{24}
50	985527416952109411645329409722 1	1.01058×10^{31}
60	499665117387106225401946051045444795 49	5.1023×10^{37}

Similarly, we can apply Theorem 4.1 to $F(x, y)$ from Theorem 3.1, where we now have to consider the dependence in u (with y playing the role of u).

Corollary 4.3. *As $n \to \infty$,*

$$\sum_{k\geq 0} f_{n,k} u^k = a(u)b(u)^{-n} n^{-3/2} \left(1 + \mathcal{O}(n^{-1})\right)$$

uniformly for u sufficiently close to 1, where $a(u)$ and $b(u)$ are analytic functions at $u = 1$.

Proof. The proof is as above with the function $H(z, w)$ replaced by

$$\begin{aligned} H(z, w, u) = z(w+1) + z^2u(u+2)(w+1)w + \frac{2z^3u^2(3w+2)(w+1)w}{1-zu-2zuw} \\ + \frac{2z^4u^3(w+1)^2w^2}{(1-zu-2zuw)^2} + \frac{2z^4u^4(w+1)w^3}{1-z^2u^2w^2} + \frac{2z^4u^3(w+1)^2w^2}{(1-zuw)(1-zu-2zuw)} \\ + \frac{2z^5u^4(w+1)^3w^2}{(1-zuw)(1-zu-2zuw)^2} \end{aligned}$$

which gives the claimed result with $b(u) = f(u)$ and $a(u)$ as in (4). □

From this, we can deduce a central limit theorem with Hwang's quasi-power theorem; see [14].

Theorem 4.4 (Hwang)**.** *Let X_n be a non-negative integer-valued sequence of random variables with probability generating function $p_n(u)$. Assume that*

$$p_n(u) = A(u)B(u)^n \left(1 + \mathcal{O}(n^{-1})\right), \tag{5}$$

where $A(u)$ and $B(u)$ are analytic functions at $u = 1$. Define

$$\mu := B'(1) \quad \text{and} \quad \sigma^2 := B''(1) + B'(1) - B'(1)^2.$$

Then, if $\sigma^2 > 0$*, we have*

$$\sup_x \left| \mathbb{P}\left(\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)}} \le x \right) - \Phi(x) \right| = \mathcal{O}\left(n^{-1/2}\right),$$

where $\Phi(x)$ *denotes the distribution function of the standard normal distribution. Moreover,*

$$\mathbb{E}(X_n) \sim \mu n \qquad \textit{and} \qquad \mathrm{Var}(X_n) \sim \sigma^2 n.$$

Since we are interested in the number of base pairs of a randomly chosen pair of secondary structures of length n which give a rank zero H_2, we define a sequence of random variables as

$$\mathbb{P}(X_n = k) = \frac{f_{n,k}}{f_n}.$$

Thus, the probability generating function of X_n is given as

$$p_n(u) := \mathbb{E}(u^{X_n}) = \frac{\sum_{k\ge 0} f_{n,k} u^k}{f_n}.$$

By Theorem 4.2 and Corollary 4.3, we see that $p_n(u)$ satisfies the expansion required in the quasi-power theorem. Applying it gives the following central limit theorem with rate. See Figure 8 for an illustration.

Theorem 4.5. *For the number* X_n *of base pairs of a randomly chosen pair of secondary structures of length* n *which give a rank zero* H_2*, as* $n \to \infty$*,*

$$\sup_x \left| \mathbb{P}\left(\frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathrm{Var}(X_n)}} \le x \right) - \Phi(x) \right| = \mathcal{O}\left(n^{-1/2}\right),$$

where

$$\mathbb{E}(X_n) \sim (0.4977358975\cdots)n \qquad \textit{and} \qquad \mathrm{Var}(X_n) \sim (0.1030331255\cdots)n.$$

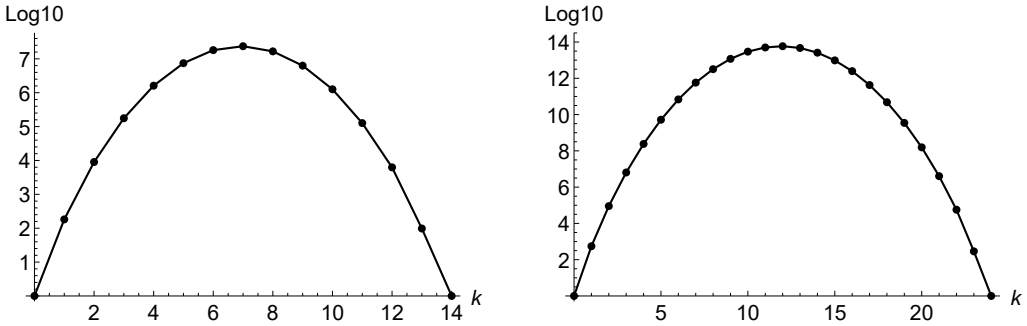
Proof. By Theorem 4.2 and Corollary 4.3, we see that (5) holds with $B(u) = z_0/f(u)$ where z_0 and $f(u)$ can be found in the proof of Theorem 4.2 and Corollary 4.3, respectively. In order to find μ and σ^2, note that that $f(1) = z_0$ and derivatives of $f(u)$ at $u = 1$ can be computed from the relations

$$w(f(u), u) = H(f(u), w(f(u), u), u) \quad \text{and} \quad 1 = H_w(f(u), w(f(u), u), u),$$

which have been established in the proof of Theorem 4.1, and implicit differentiation; see Theorem 2.23 in [8]. This gives,

$$\mu = 0.4977358975\cdots \qquad \text{and} \qquad \sigma^2 = 0.1030331255\cdots .$$

In particular, $\sigma^2 > 0$ as required and thus, the claimed limit law follows from the quasi-power theorem. □



(a) $n = 15$ (b) $n = 25$

Figure 8: The distribution of the number of base pairs for $n = 15$ and 25, where the points $(k, \log_{10} f_{n,k})$ are plotted.

Author Disclosure Statement

The authors report no conflict of interest.

Acknowledgments

The authors are grateful to the anonymous referees for valuable comments and suggestions, particularly for pointing out a missing case in our computation in Section 3. We thank Emma Yu Jin for inviting RXFC and MF to a workshop on mathematical biology at Xiamen University in September 2024. RXFC was supported by the Anhui Provincial Natural Science Foundation of China (No. 2208085MA02) and Overseas Returnee Support Project on Innovation and Entrepreneurship of Anhui Province (No. 11190-46252022001). MF acknowledges partial support by the grant NSTC-113-2115-M-004-004-MY3 of the National Science and Technology Council, Taiwan.

Appendix

For the interested reader, here we briefly present an alternative way for computing $F(x)$ without introducing any auxiliary functions according to an idea suggested by one of the anonymous referees. According to the "position" of base 1, we can classify the studied pairs (S, T) of non-empty secondary structures into three cases:

- Case 1″: base 1 is neither paired in S nor in T. The pairs in this case are clearly counted by $xF(x, y)$.

- Case 2″: base 1 is contained in a cycle when viewing the pair (S, T) as a conventional graph (with multiple edges). This case consists of pairs in Case 1, Case 2(iv)(b) and their upside-down version in Section 3 of the paper. The contribution of this case is then given by

$$x^2y^2F(x, y)[F(x, y) - 1] + \frac{2x^4y^4[F(x, y) - 1]^3F(x, y)}{1 - x^2y^2[F(x, y) - 1]^2}.$$

- Case 3″: base 1 is contained in a path. This is the most subtle case to handle

without the aid of introducing auxiliary functions. We distinguish two situations below.

(a) Suppose base 1 is the leftmost terminal point of a path of length $i \geq 1$. Note that this path contains $i+1$ bases and i edges (arcs), contributing a factor $x^{i+1}y^i$ at first. Next, in each of these i intervals separated by the $i+1$ bases, there may be a possibly empty substructure. Identifying intervals which are allowed to contain an empty substructure is necessary. Similar analysis was carried out in studying γ structures [10, 17]. It is the key to realize that every time we switch the proceeding direction when traveling along the path starting with base 1, we "create" an interval which is allowed to contain an empty substructure. See Figure 9 for an illustration. Suppose along the path, there

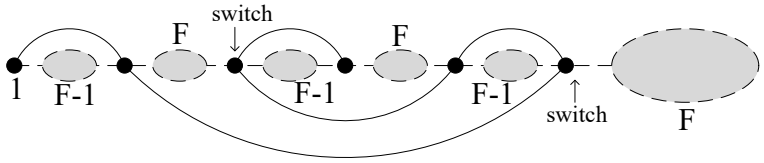


Figure 9: Base 1 is the leftmost terminal point of a path of length 5 with 2 direction switches resulting in two possibly empty substructures F.

are j times of switching direction. Since at the end point (with respect to the traveling direction) of each arc except the last arc there is a possibility of switching direction, the contribution of these i intervals is

$$\binom{i-1}{j} F(x,y)^j [F(x,y)-1]^{i-j}.$$

Note that to the right of the rightmost base of the path containing 1 there may be a possibly empty substructure which contributes an additional factor $F(x,y)$, and that base 1 may be paired either in S or T, yielding a factor 2.

Hence, the pairs in this situation are counted by

$$\begin{aligned}&2F(x,y)\sum_{i\geq 1}\sum_{j=0}^{i-1}x^{i+1}y^i\binom{i-1}{j}F(x,y)^j[F(x,y)-1]^{i-j}\\ =&\frac{2x^2yF(x,y)[F(x,y)-1]}{1-xy[2F(x,y)-1]}.\end{aligned}$$

(b) Suppose base 1 is a middle point of a path, i.e., 1 is paired in both S and T. This path could be viewed as two independent paths, one starting from the arc containing 1 in S and the other starting from the arc containing 1 in T. These two paths may be analogously processed as in situation (a), and the detail is left to the interested reader. Finally, the pairs in this situation are enumerated by

$$\frac{2x^2yF(x,y)[F(x,y)-1]}{1-xy[2F(x,y)-1]}xyF(x,y)+\frac{4y(x^2yF(x,y)[F(x,y)-1])^2}{(1-xy[2F(x,y)-1])^2}.$$

In summary, we have

$$\begin{aligned}F(x,y)=&1+xF(x,y)+x^2y^2F(x,y)[F(x,y)-1]\\ &+\frac{2x^4y^4[F(x,y)-1]^3F(x,y)}{1-x^2y^2[F(x,y)-1]^2}+\frac{2x^2yF(x,y)[F(x,y)-1]}{1-xy[2F(x,y)-1]}\\ &+\frac{2x^3y^2F(x,y)^2[F(x,y)-1]}{1-xy[2F(x,y)-1]}\left(1+\frac{2xy[F(x,y)-1]}{1-xy[2F(x,y)-1]}\right),\end{aligned}\tag{6}$$

which is equivalent to Theorem 3.1.

Setting $y=1$ in the last equation, we also obtain

$$\begin{aligned}F(x)=&1+xF(x)+x^2F(x)[F(x)-1]+\frac{2x^4[F(x)-1]^3F(x)}{1-x^2[F(x)-1]^2}\\ &+\frac{2x^2F(x)[F(x)-1]}{1-x[2F(x)-1]}+\frac{2x^3F(x)^2[F(x)-1]}{1-x[2F(x)-1]}\left(1+\frac{2x[F(x)-1]}{1-x[2F(x)-1]}\right).\end{aligned}\tag{7}$$

References

[1] A. C. Bura, Q. He, C. M. Reidys, Loop homology of bi-secondary structures, Discrete Math. 344(6) (2021), 112371.

[2] A. C. Bura, Q. He, C. M. Reidys, Loop homology of bi-secondary structures II, J. Algebr. Comb. 56 (2022), 785–798.

[3] R. X. F. Chen, A new bijection between RNA secondary structures and plane trees and its consequences, Electron. J. Combin. 26(4) (2019), P4.48.

[4] R. X. F. Chen, C. M. Reidys, M. S. Waterman, RNA secondary structures with given motif specification: combinatorics and algorithms, Bull. Math. Biol. 85 (2023), #21.

[5] P. Clote, Combinatorics of saturated secondary structures of RNA, J. Comp. Biol. 13 (2006), 1640–1657.

[6] T. Došlić, D. Svrtan, D. Veljan, Enumerative aspects of secondary structures, Discrete Math. 285 (2004), 67–82.

[7] M. Drmota, Systems of functional equations, Random Struc. Algor. 10(1-2) (1997), 103–124.

[8] M. Drmota, Random Trees: An Interplay between Combinatoris and Probability, Springer, 2009.

[9] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, M. Zehl, Design of multistable RNA molecules, RNA 7(2) (2001), 254–265.

[10] H. S. W. Han, T. J. X. Li, C. M. Reidys, Combinatorics of γ-structures, J. Comp. Biol. 21 (2014), 591–608.

[11] I. L. Hofacker, P. Schuster, P. F. Stadler, Combinatorics of RNA secondary structures, Discrete Appl. Math. 88 (1998), 207–237.

[12] C. Heitsch, S. Poznanović, Combinatorial insights into RNA secondary structure, Discrete and Topological Models in Molecular Biology (2013), 145–166.

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

[14] H. K. Hwang, On convergence rates in the central limit theorems for combinatorial structures, European J. Combin. 19(3) (1998), 329–343.

[15] W. Lorenz, Y. Ponty, P. Clote, Asymptotics of RNA shapes, J. Comp. Biol. 15 (2008), 31–63.

[16] T. J. X. Li, C. M. Reidys, The rainbow spectrum of RNA secondary structures, Bull. Math. Biol. 80 (2018), 1514–1538.

[17] T. J. X. Li, C. M. Reidys, The topological filtration of γ-structures, Math. Biosci. 241(1) (2013), 24–33.

[18] B. Liao, T. Wang, General combinatorics of RNA secondary structure, Math. Biosci. 191 (2004), 69–81.

[19] M. E. Nebel, Combinatorial properties of RNA secondary structures, J. Comp. Biol. 9(3) (2003), 541–574.

[20] W. R. Schmitt, M. S. Waterman, Linear trees and RNA secondary structure, Discrete Appl. Math. 51(3) (1994), 317–323.

[21] T. F. Smith, M. S. Waterman, RNA secondary structure, Math. Biol. 42 (1978), 31–49.

[22] P. R. Stein, M. S. Waterman, On some new sequences generalizing the Catalan and Motzkin numbers, Discrete Math. 26 (1979), 261–272.

[23] M. S. Waterman, Secondary structure of single-stranded nucleic acids, in Rota G.-C. (ed) Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y., vol 1, pp. 167–212, 1978.