Endhered patterns in matchings and RNA

Célia Biane[1], Greg Hampikian[2], Sergey Kirgizov[1], Khaydar Nurligareev[1]

[1]LIB, Université de Bourgogne, Dijon, France
[2]Department of Biological Sciences, Boise State University, Boise, ID, USA
Corresponding author: sergey.kirgizov@u-bourgogne.fr

May 2, 2024

Abstract

An *endhered (end-adhered) pattern* is a subset of arcs in matchings, such that the corresponding starting points are consecutive and the same holds for the ending points. Such patterns are in one-to-one correspondence with the permutations. We focus on the occurrence frequency of such patterns in matchings and real-world RNA structures with pseudoknots. We present combinatorial results related to the distribution and asymptotic behavior of the pattern 21, which corresponds to two consecutive stacked bonds frequently encountered in RNA, and the pattern 12, representing the archetypal minimal pseudoknot. We show that in matchings these two patterns are equidistributed, which is quite different from what we can find in real-world RNAs. We also examine the distribution of endhered patterns of size 3, showing how the patterns change under the transformation called *endhered twist*. Finally, we compute the distributions of endhered patterns of size 2 and 3 in real-world secondary RNA structures with pseudoknots and discuss possible outcomes of our study.

1 Introduction

Ribonucleic acids (RNAs) are macromolecules fulfilling many biological functions: they code for protein, are involved in the regulation of gene expression, can have catalytic activities and store the genetic information of certain viruses. The structure of RNAs is defined at the primary level as sequences of four nucleotides: Adenine (A), Uracil (U), Guanine (G) and Cytosine (C). The secondary structure abstracts from the nature of the nucleotide and considers only the bonds forming between nucleotides during the synthesis of RNAs and shaping how the molecules folds in space. Two types of bonds are formed during the RNA folding process: phospho-diester bonds (known as strong bonds) are formed between pairs of consecutive nucleotides in the sequence forming the RNA chain, and hydrogen bonds (also known as weak bonds) are formed between pairs of nucleotides distant in the sequence. The secondary structure represents an intermediate level between the primary sequence and the shape, and has the advantage of being both relevant from a biological perspective and tractable from a computational point of view.

1.1 Models of RNA secondary structures

RNA secondary structures have been formalized as graphs primarily by Waterman [39] to tackle the problem of prediction of secondary structures from the primary structure, which is more easily measurable. Ponty [34] gives a variant of Waterman definition of RNA secondary structure without pseudoknots and having a minimal number θ of unpaired positions between pairs positions. Formally, *Waterman-Ponty RNA secondary structure S* of size n is defined as a set of base-pairs $(i, j), 1 \leqslant i < j \leqslant n$, such that:

1. Each position is monogamous, $\forall (i, j) \neq (i', j') \in S : \{i, j\} \cap \{i', j'\} = \varnothing$.

2. Minimal distance θ between paired nucleotides, $\forall (i, j) \in S : j - i \geqslant \theta$.

3. No pseudoknot allowed, $\forall (i, j), (i', j') \in S, i < i' : (j', j)$ or $(j < i')$.

These structures can be represented using the *dot-bracket notation*. A secondary structure of an n-nucleotide RNA is encoded as an n-length sequence of parentheses "()" and dots ".", where an open parenthesis represents a nucleotide bonded to another nucleotide represented by a closed parenthesis, and dots correspond to unpaired nucleotides. *The extended dot-bracket notation* includes also other types of parentheses: "[]", "{}", "<>", "aA", etc. The extended dot-bracket notation enables the representation of pairings in pseudoknots. Figure 1 shows 4 different representations of an example of RNA secondary structure.

...(((....[[[.)))...((((.]]]....))))...

(4,17),(5,16),(6,15),(11,28),(12,27)(13,26),(21,36),(22,35),(23,34),(24,33)
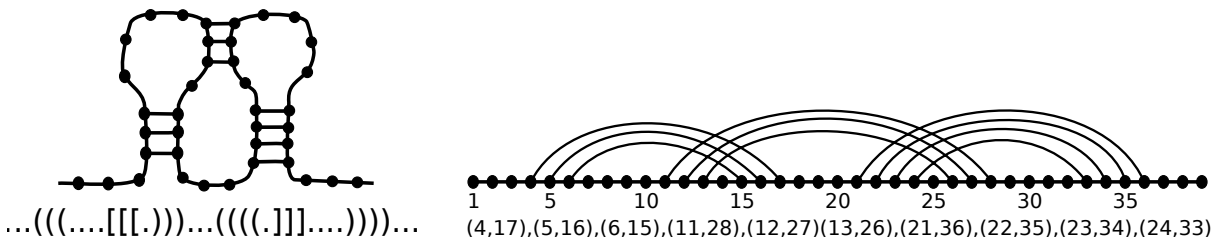
Figure 1: A drawing, an extended dot bracket notation, an arc diagram and a set of pairs representing an example of an RNA secondary structure.

Other models of RNA secondary structures were studied from combinatorial point of view. Haslinger and Standler [22] examined enumerative and asymptotic properties of *bi-secondary structures,* i.e. arc diagrams with arcs both in upper and lower half of the plane but without arc intersections. The concept of RNA shape was introduced by Giegerich, Voss and Rehmsmeier [17] who developed an algorithm for the computation of minimum free energy RNA shapes. The number and asymptotics of RNA shapes was studied by Lorenz, Ponty and Clote [27]. Reidys and Wang [36] considered a generalization of RNA shapes defined on k mutually-crossing arcs.

In this paper, we look at RNA structures with no restrictions on the number of arc crossings.

1.2 Notion of pattern in RNA secondary structures

At the biological level, common patterns have been observed in orthodox structures: single strand regions, hairpin and internal loops, buldges and various computational tools exist

for detecting them from primary sequences, including the work of Macke *et al.* [30]. More and more information is being gathered from three dimensional reconstructions of RNA molecules paving the way to a better comprehension of the laws governing the RNA folding process and the formation of RNA patterns.

Rødland [37] proposed a classification of RNA secondary structures in four level of abstractions: nucleotide, ladder, stem and collapsed level, based on the considered internal patterns. In his work, the nucleotide level corresponded to structures with arc diagrams containing unpaired nucleotides, the ladder level corresponded to structures abstracted from unpaired nucleotides, the stem level was abstracted from bulges and internal loops, and the collapsed level was abstracted from nested loops. Rødland studied different kind of pseudoknot patterns of increasing complexity: H-pseudoknot, double hairpin pseudoknot and pseudotrefoil. He counted these patterns in RNA secondary structures of increasing complexity and studied their asymptotics. He also showed that the theoretical number of pseudoknots in secondary structures is higher than in real secondary structure of the Rfam database. Note that Rødland's collapsed structures correspond to RNA shapes studied by Giegerich, Voss and Rehmsmeier [17].

Quadrini [35] addressed the problem of searching a given structural pattern, defined as a sequence of crossing loops in a RNA secondary structure or shape and characterized by arbitrary number of pseudoknots. She proposed polynomial time algorithms for their identification. A beautiful paper by Gan, Pasquali and Schlick [16] studied RNA structures and their patterns using graph representations. In the future, it would be interesting to compare the relationships between different kinds of patterns.

In this paper, we study the formation of endhered patterns (formally defined below) in matchings. In Section 2, we do it from a theoretical perspective. In Section 3, we observe the number of occurrences of such patterns in RNA secondary structures data derived from 3D experiments. We conclude by discussing possible outcomes of our study in Section 4.

2 Endhered patterns in matchings

2.1 Basic definitions

By (perfect) *matching* of size n we mean the sequence of $2n$ points $(1, 2, \ldots, 2n-1, 2n)$ endowed with a set of n arcs, such that every point is linked to one and only one another point. The left side of Figure 2 shows an example of a matching of size 4. Matchings of size n can be considered as fixed-point-free involutions in the symmetric group S_{2n}. Thus, the matching of Figure 2 can be represented by the permutation $3\,6\,1\,5\,4\,2\,8\,7$, which is the product of non-intersecting transpositions $(13)(26)(45)(78)$.

For a positive integer n, any matching of size n can be uniquely constructed from some matching of size $n-1$ by the following procedure. We add a new arc starting at the left of the already existing $2(n-1)$ points and ending at some of $2(n-1)+1$ possible new positions (see the right side of Figure 2). This observation leads us to the following recurrence relation for the number of matchings of size n: $a_n = (2n-1)a_{n-1}$ with $a_0 = 1$. As a consequence, $a_n = (2n-1)!! = (2n-1)\cdot(2n-3)\cdot\ldots\cdot3\cdot1$. The corresponding sequence starting with $1, 1, 3, 15, 105, 945, 10395$ is known as A1147 in Sloane's Encyclopedia [32].
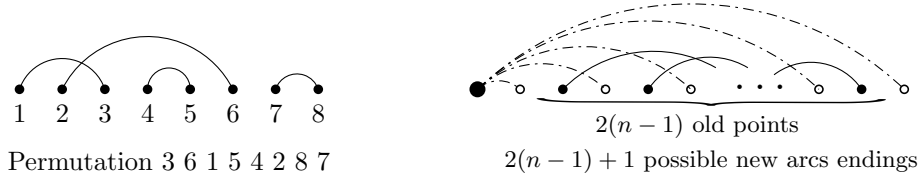
3

Figure 2: An example of matching of size 4, the corresponding permutation, and recursive construction of matchings.

In mathematical literature, matchings often appear in different contexts, from the representation theory of Lie algebras [8] to the geometry of moduli spaces of flat connections on surfaces [2]. Efficient generating algorithms for involutions with (without) fixed points were established by Vajnovszki [38]. Ardnt's book [3] also presents interesting generating algorithms for involutions and other combinatorial structures.

Different kinds of patterns in matchings are being actively studied in combinatorics for the past twenty years. Initially, the interest to this topic came from the rapidly developing study of permutation patterns, since a matching can be thought as a permutation of a specific form. From this perspective, we say that a matching σ is a *pattern* in a matching μ if σ can be obtained from μ by deleting some of its arcs (and consistently relabelling the remaining vertices). For instance, Chen, Deng, Du, Stanley and Yan [10] studied distributions of crossings and nestings. Jelínek [24], as well as Bloom and Elizalde [7], considered pattern avoiding matchings in the case when σ is a permutational matching of size 3. An extension for more general patterns was elaborated by Cervetti and Ferrari [9], while other authors, such as Chen, Mansour and Yan [11], Jelínek and Mansour [25], considered partial patterns.

As we see, what matters in the above investigations is the relative positions of arcs that form a pattern. At the same time, the distances between starting and ending points of these arcs are not fixed. The main object of our study concerns specific restrictions imposed on the arcs. Namely, the starting points of a pattern, as well as its ending points, form an interval, while the distance between these two intervals may vary. We call such patterns *endhered* (end-adhered) to emphasize the nature of these restrictions.

Definition 2.1. An *endhered pattern* is a matching, such that the starting point of any of its arcs precedes the ending point of any other arc. In other words, a matching σ of size p written as a permutation $\sigma = \sigma_1 \ldots \sigma_{2p}$ is an endhered pattern if $\sigma_{p+1} \ldots \sigma_{2p}$ is a permutation of size p (such matchings are also called *permutational*). The left side of Figure 3 presents an example of an endhered pattern of size 3. In the following, we identify endhered patterns with the corresponding permutations.

We say that a matching $\mu = \mu_1 \ldots \mu_{2n}$ *contains* an endhered pattern $\pi = \pi_1 \ldots \pi_p$ at position $(i + 1, j + 1)$, where $i \geqslant 0$ and $i + p \leqslant j \leqslant n - p$, if

$$\mu_{s+i} = \pi_s^{-1} + j, \qquad s = 1, \ldots, p$$

(here, $\pi^{-1} = \pi_1^{-1} \ldots \pi_p^{-1}$ is the inverse to the permutation π). In other words, μ contains p arcs whose starting points are $i + 1, \ldots, i + p$, whose ending points are $j + 1, \ldots, j + p$, and that form the endhered pattern π. The right side of Figure 3 shows an example of a matching containing pattern shown on the left side.

4

$$\sigma = 6\,4\,5\,\underline{2}\,\underline{3}\,\underline{1} \qquad \pi = 2\,3\,1 \qquad \mu = 5\,\underline{10}\,\underline{8}\,9\,1\,7\,6\,\underline{3}\,\underline{4}\,2\,14\,13\,12\,11 \qquad \pi^{-1} = 3\,1\,2$$
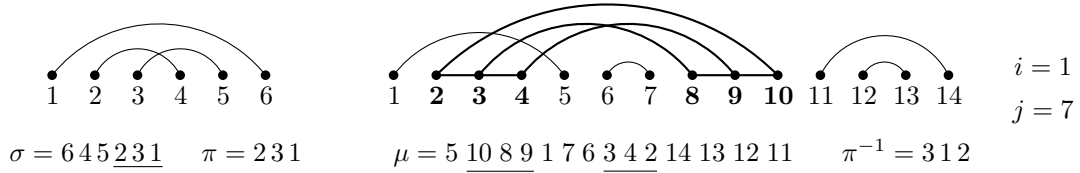
$$i = 1$$
$$j = 7$$

Figure 3: Endhered pattern 231 and an example of its occurrence.

The endhered patterns have been rarely studied in the literature, although they may shed light on the formation of collapsed RNA structures from Rødland's paper about pseudoknots [37]. The only work in this direction that we are aware of is the paper of Baril [4] who examined one of two endhered patterns of size 2 in his study on irreducible involutions and permutations.

2.2 Endhered twists and engendered symmetries

Given an endhered pattern π, let us denote by $a_{n,k}(\pi)$ the number of matchings of size n with k occurrences of π. If, additionally, σ is another endhered pattern, then we designate by $a_{n,k,m}(\pi, \sigma)$ the number of matchings of size n with k and m occurrences of patterns π and σ, respectively. Certain patterns, for instance $\pi = $ ⌢ and $\sigma = $ ⌢, have the same distribution, meaning that $a_{n,k}(\pi) = a_{n,k}(\sigma)$. The goal of this Subsection is to establish such equidistributed classes of endhered patterns with the help of bijections, i.e. without direct enumeration. To this end, we apply matching transformations that we call *endhered twists*.

Definition 2.2. The *left endhered twist* (resp. *right endhered twist*) is a transformation that takes a matching μ and produces a matching $\texttt{letw}(\mu)$ (resp. $\texttt{retw}(\mu)$) such that all runs of consecutive starting (resp. ending) points are reversed. Figure 4 shows an example of the right twist.
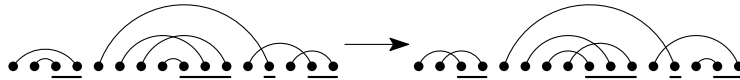
Figure 4: An example of the right endhered twist, runs of right points are underlined.

Observation 2.3. *Endhered twists of endhered patterns correspond to classical symmetries on permutations. Thus, the right endhered twist applied to an endhered pattern $\pi = \pi_1 \ldots \pi_p$ is its reverse: $\texttt{retw}(\pi) = \pi_p \ldots \pi_1$. At the same time, the left endhered twist is the complement: $\texttt{letw}(\pi) = (p + 1 - \pi_1) \ldots (p + 1 - \pi_p)$. For example,*

$$\texttt{retw}(321) = 123, \qquad \texttt{retw}(231) = 132, \qquad \texttt{retw}(312) = 213,$$
$$\texttt{letw}(321) = 123, \qquad \texttt{letw}(132) = 312, \qquad \texttt{letw}(213) = 231.$$

Geometrically, if we represent permutations as square tables, the endhered twists are axial symmetries (see Figure 5).
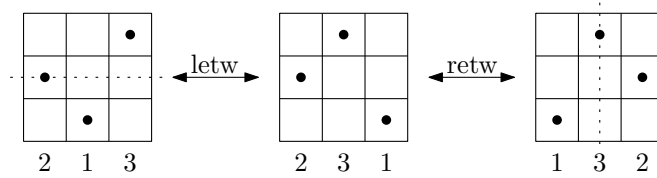
5

Figure 5: Geometrical meaning of endhered twists.

Lemma 2.4. *Two endhered patterns π and σ of the same size have the same joint distribution if they are identical under the left or right endhered twists. In other words, if $\pi = \mathtt{letw}(\sigma)$ or $\pi = \mathtt{retw}(\sigma)$, then*

$$a_{n,k,m}(\pi,\sigma) = a_{n,k,m}(\sigma,\pi) \tag{1}$$

for any integers n, k and m. In particular, $a_{n,k}(\pi) = a_{n,k}(\sigma)$.

Proof. Suppose that π is obtained from σ by applying the left endhered twist. Let μ be a matching containing k occurrences of π and m occurrences of σ. Applying the left twist to the whole μ, we transform every occurrence of π to σ and vice versa. No other occurrence of π or σ are created. This implies relation (1). The cases of the right right twist is obtained *mutatis mutandis*. $\square$

Remark 2.5. Given a positive integer n and an endhered pattern π, the value $a_{n,0}(\pi)$ is the number of matchings of size n avoiding π. We say that two endhered patterns π and σ are *Wilf equivalent* if $a_{n,0}(\pi) = a_{n,0}(\sigma)$ for every n. For example, Wilf equivalent patterns of size 2 and 3 are shown in Figure 6. Matchings avoiding the pattern ⌢⌢ correspond to involutions with no fixed points and no successions considered by Baril [4].

For the endhered patterns of size 2 and 3, as we will see in this paper, their Wilf equivalence implies that the corresponding patterns are equidistributed. In other words, if $a_{n,0}(\pi) = a_{n,0}(\sigma)$ for every n, then $a_{n,k}(\pi) = a_{n,k}(\sigma)$ for all n and k. In the general case, this question is not trivial. We will discuss this in another paper [6].
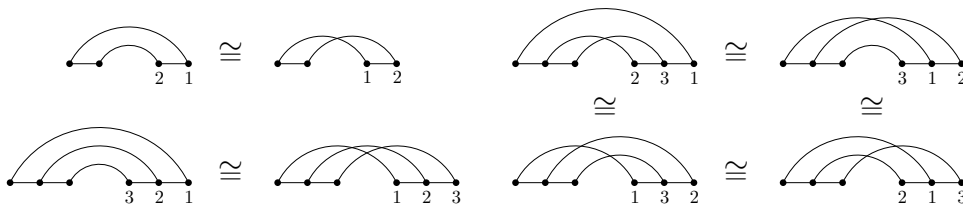
Figure 6: Wilf equivalence classes of endhered patterns of size 2 and 3.

Corollary 2.6. *The joint distribution of the endhered patterns $\pi = 1\ldots p$ and $\sigma = p\ldots 1$ is symmetric. In particular, the number of size-n matchings containing k ⌢⌢ and m ⌢⌢ equals the number of size-n matchings containing m ⌢⌢ and k ⌢⌢.*

Remark 2.7. Corollary 2.6 is consistent with the result of Chen, Deng, Du, Stanley and Yan [10] who showed that, in the case of classical patterns, the joint distribution of crossings and nestings is symmetric.

6

2.3 Enumeration of endhered patterns of size 2

The goal of this Subsection is to study the distributions of endhered patterns of size 2. It follows from Lemma 2.4 that patterns ⌢⌢ and ⌢⌢ have the same distribution, so we need only to enumerate one of them. For simplicity, we will write $a_{n,k}$ instead of $a_{n,k}(21)$ throughout the rest of the paper. Note that some of results presented here (Lemma 2.9, Corollary 2.10) were proved by Baril [4]. For the sake of completeness, we provide their proofs as well.

Theorem 2.8. *For $n \geqslant 1$, the number of size-n matchings containing exactly k occurrences of pattern ⌢⌢ (resp. ⌢⌢) satisfies*

$$
\begin{cases}
a_{n+1,k} &= a_{n,k-1} + 2(n-k) \cdot a_{n,k} + 2(k+1) \cdot a_{n,k+1} \\
a_{1,0} &= 1 \\
a_{1,k} &= 0, \quad k \neq 0.
\end{cases}
\tag{2}
$$

Formally, we allow negative values of k. It follows from (2) that $a_{n,k} = 0$ whenever $k < 0$.

Proof. The boundary values corresponding to $n = 1$ are easily obtained. To construct a matching of size $n + 1$ having k occurrence of ⌢⌢, we take a non-empty matching μ of size n and add a new arc E together with its two ending points (left and right). We always put the left point of E before the first point of μ. For the right point of E, there are three different cases, since the new arc either creates a new pattern, destroys an existing pattern, or leaves the number of occurrences of the pattern unchanged.

1. The right point of E is placed just after the right point of the arc $(1, i)$ of μ (see Figure 7). In this case, the number of occurrences increases by one. Hence, the matching μ must possess $k - 1$ patterns ⌢⌢, so that the resulting matching has k occurrences of this pattern. This gives us the first summand in relation (2).
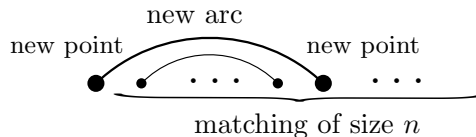
Figure 7: Adding a new arc creates a new pattern.

2. The right end of E is put between the left (resp. right) ends of any two arcs that make up the pattern ⌢⌢. In this case, one of the existing patterns is destroyed, and the number of occurrences decreases by one (see Figure 8). Therefore, to have k occurrences in the resulting matching, the initial matching μ must contain $k + 1$ occurrences of ⌢⌢. Since there are $2(k + 1)$ destructive positions, this gives us the last summand in relation (2).

3. The right point of E does not create, nor remove any pattern ⌢⌢. In this case, the matching μ must possess k occurrences of this pattern. There are $2n + 1$ possible positions for the right end of E, but $2k + 1$ of them are forbidden. Indeed, there are
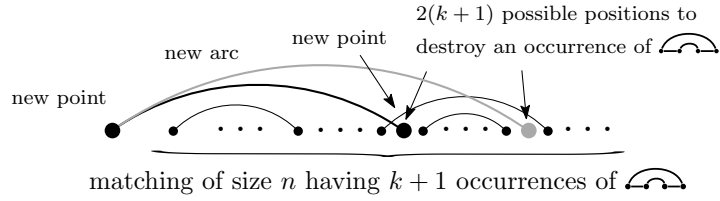
Figure 8: Adding a new arc destroys an existing pattern.

k already existing occurrence of ⌢⌣ that we do not want to destroy, and one more position is banned because we are not allowed to create any new ⌢⌣ in this case (see Figure 9).
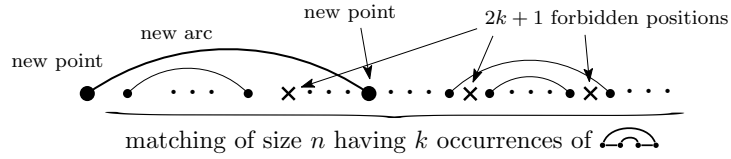

Figure 9: Adding a new arc does not change the number of pattern occurrences.

The distribution of the pattern ⌢⌣ respects the same recurrence. To establish the latter fact, it is sufficient to twist the corresponding ends using Lemma 2.4. □

Table 1 shows first values of $a_{n,k}$.

k \ n	1	2	3	4	5	6	7	8	9	OEIS
0	1	2	10	68	604	6584	85048	1269680	21505552	A165968
1	0	1	4	30	272	3020	39504	595336	10157440	A179540
2	0	0	1	6	60	680	9060	138264	2381344	
3	0	0	0	1	8	100	1360	21140	368704	
4	0	0	0	0	1	10	150	2380	42280	
5	0	0	0	0	0	1	12	210	3808	
6	0	0	0	0	0	0	1	14	280	
7	0	0	0	0	0	0	0	1	16	
8	0	0	0	0	0	0	0	0	1	

Table 1: Distribution of pattern ⌢⌣ (21).

Lemma 2.9. *For any $n > k \geqslant 0$,*

$$a_{n,k} = \binom{n-1}{k} a_{n-k,0}. \tag{3}$$

8

Proof. Each matching of size n with k occurrences of ⌢⌢ can be uniquely obtained by the following procedure. Take a matching of size $n - k$ avoiding ⌢⌢ pattern. Add k new arcs expanding some of the original arcs into double, triple, ..., and $(k+1)$-uple arcs. Since each t-uple arc contains $t - 1$ occurrences of ⌢⌢, we have k occurrences of ⌢⌢ in total (see Figure 10). Therefore, relation (3) comes from the classical stars-and-bars argument [14]: the binomial coefficient $\binom{n-1}{k}$ is the number of ways to add k new arcs by expanding some of the $n - k$ original arcs.
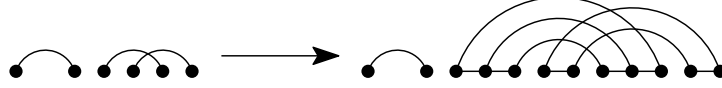
Figure 10: Arc expansion. Adding 3 new arcs, we create 3 new occurrences of ⌢⌢.

□

Corollary 2.10. *The sequence $(a_{n,0})$ satisfies the recurrence relation*

$$a_{n+1,0} = 2n a_{n,0} + 2(n-1) a_{n-1,0} \tag{4}$$

with the initial conditions

$$a_{1,0} = 1 \qquad and \qquad a_{2,0} = 2.$$

In particular, this sequence corresponds to a shift of A165968 in OEIS [32].

Proof. Recurrence (4) follows immediately from relations (2) and (3). □

Remark 2.11. There is a constructive combinatorial explanation of relation (4) (see also the work of Baril [4]). Indeed, if we remove the first arc in a mathing of size $n + 1$ avoiding ⌢⌢, then we obtain a matching of size n that either avoids ⌢⌢ or contains exactly one pattern ⌢⌢. This observation shows that any mathing of size $n + 1$ avoiding ⌢⌢ can be uniquely constructed in the following way. Either we take a matching μ of size n avoiding ⌢⌢ and add a new arc starting at the left of the already existing $2n$ points of μ and ending at some of $2n$ possible new positions (it cannot arrive just after the right point of the first arc of μ). Or we take a matching ν of size $n - 1$ avoiding ⌢⌢, double one of its $n - 1$ arcs, and then destroy the just created pattern ⌢⌢ by a new arc drawn in one of two possible ways (see Figure 11).
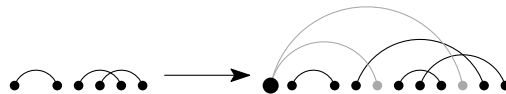
Figure 11: Two possible positions for destroying the newly created pattern ⌢⌢.

Corollary 2.12. *The sequence $(a_{n,1})$ satisfies the recurrence relation*

$$a_{n+1,1} = 2n(a_{n,1} + a_{n-1,1}) \tag{5}$$

with the initial conditions

$$a_{1,1} = 0 \qquad and \qquad a_{2,1} = 1.$$

In particular, this sequence corresponds to a shift of A179540 in OEIS [32].

9

Proof. Apply relations (2) and (3):

$$a_{n+1,1} = na_{n,0} = n\Big(2(n-1)a_{n-1,0} + 2a_{n-1,1}\Big) = 2n(a_{n,1} + a_{n-1,1}).$$

$\square$

Lemma 2.13. *For any $n \geqslant 0$,*

$$a_{n+1,0} = \sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}(2k+1)!!. \tag{6}$$

Proof. Being rewritten as

$$a_{n+1,0} = \sum_{k=0}^{n}(-1)^{k}\binom{n}{k}\big(2(n-k)+1\big)!!,$$

relation (6) can be proved with the help of the inclusion-exclusion principle. Indeed, there are $(2n+1)!!$ matchings of size $n+1$. In $(2n-1)!!$ among them, i-th arc belongs to the pattern ⌢⌢ as an outer arc, $i = 1, \ldots, n$. Hence, to get the number of pattern avoiding matchings, we need to deduce $\binom{n}{1}(2n-1)!!$ from $(2n+1)!!$. However, matchings with two occurrences of ⌢⌢ are deduced twice. There are $(2n-3)!!$ matchings with i-th and j-th arcs belonging to the pattern ⌢⌢ as outer arcs, and $\binom{n}{2}$ possible pairs (i,j), so we add $\binom{n}{2}(2n-3)!!$. The same way, we treat matchings with three occurrences of ⌢⌢, with four occurrences and so on.

$\square$

For any $n, k \geqslant 0$, let us denote

$$b_{n,k} := a_{n+1,k}.$$

Define also

$$B(z, u) := \sum_{n=0}^{\infty}\sum_{k=0}^{n} b_{n,k}\frac{z^n}{n!}u^k.$$

Lemma 2.14. *The exponential generating function $B(z, u)$ satisfies*

$$B(z, u) = \frac{e^{z(u-1)}}{\sqrt{(1-2z)^3}}. \tag{7}$$

In particular, the exponential generating function of the shifted k-th row of Table 1 is

$$[u^k]B(z, u) = \frac{z^k}{k!} \cdot \frac{e^{-z}}{\sqrt{(1-2z)^3}}.$$

Proof. Recall that

$$\sum_{n=0}^{\infty}(2n+1)!!\frac{z^n}{n!} = \frac{1}{\sqrt{(1-2z)^3}}.$$

Therefore, according to Lemma 2.13, we have

$$B(z,0) = \sum_{n=0}^{\infty} b_{n,0}\frac{z^n}{n!} = \sum_{n=0}^{\infty}\frac{z^n}{n!}\sum_{k=0}^{n}(-1)^{n-k}\binom{n}{k}(2k+1)!!$$

$$= \left(\sum_{n=0}^{\infty}\frac{(-z)^n}{n!}\right)\left(\sum_{n=0}^{\infty}(2n+1)!!\frac{z^n}{n!}\right) = \frac{e^{-z}}{\sqrt{(1-2z)^3}}.$$

Hence, due to Lemma 2.9,

$$B(z,u) = \sum_{n=0}^{\infty}\frac{z^n}{n!}\sum_{k=0}^{n}\binom{n}{k}b_{n-k,0}u^k$$

$$= \left(\sum_{n=0}^{\infty}b_{n,0}\frac{z^n}{n!}\right)\left(\sum_{n=0}^{\infty}\frac{(zu)^n}{n!}\right) = \frac{e^{-z}}{\sqrt{(1-2z)^3}}\cdot e^{zu}.$$

$\square$

Remark 2.15. Combinatorially, relation (7) represents another facet of the inclusion-exclusion principle. This can be proved using the symbolic method, see [15, Section III.7.4]. Indeed, the exponential generating function $e^{zv}/\sqrt{(1-2z)^3}$ corresponds to the labeled product of the combinatorial class of urns $\mathcal{U}$ and the derivative of the combinatorial class of matchings $\mathcal{M}$. Urns can be interpreted as distinguished arcs (marked by the variable v) that are inserted into a matching to form double arcs, that is, patterns ⌢⌢. Passing to the derivative $\mathcal{M}'$ corresponds to adding a supplementary arc to all matchings. This operation guarantees that every matching possesses at least one arc (and hence, there is something to double). Finally, the variable substitution $v = u - 1$ has the meaning of the inclusion-exclusion itself.

2.4 Asymptotics for patterns of size 2

Here, we provide the asymptotics of the distribution of size-2 patterns.

Theorem 2.16. *The asymptotic behavior of the numbers of matchings with k occurrences of pattern ⌢⌢ (resp. ⌢⌢), as $n \to \infty$, is*

$$a_{n,k} \sim \frac{1}{2^k k!}\left(\frac{2}{e}\right)^{n+1/2}n^n.$$

Proof. Let us start with establishing the asymptotics of $b_{n,0}$. Due to Lemma 2.13, we have

$$b_{n,0} = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (2k+1)!!$$

$$= (2n+1)!! \sum_{k=0}^{n} \frac{(-1)^k}{k!} \cdot \frac{n(n-1)\dots(n-k+1)}{(2n+1)(2n-1)\dots(2n-2k+3)}$$

$$\sim (2n+1)!! \sum_{k=0}^{n} \frac{(-1)^k}{2^k k!} \sim \frac{(2n)(2n)!}{2^n n! e^{1/2}} \sim \frac{2^{n+3/2}}{e^{n+1/2}} \cdot n^{n+1},$$

where the last passage is done with the help of Stirling's formula. Now, the general case can be done by Lemma 2.9:

$$b_{n,k} = \binom{n}{k} b_{n-k,0} \sim \frac{n^k}{k!} \cdot \frac{2^{n-k+3/2}}{e^{n-k+1/2}} \cdot (n-k)^{n-k+1} \sim \frac{2^{n-k+3/2}}{k!} \cdot \frac{n^{n+1}}{e^{n+1/2}}.$$

Passing to $a_{n,k} = b_{n-1,k}$, we get the final result. $\square$

Remark 2.17. There is an alternative proof that relies on the singularity analysis of the exact form of the series $B(z,u)$ given by Lemma 2.14. For details of this technique, see [15, Theorem VI.4].

Corollary 2.18. *The limit distribution of pattern* $\overset{\frown}{}$ *(resp.* $\overset{\frown}{}$*) in a uniform random matching of size n follows asymptotically a Poisson law with parameter* $1/2$.

Proof. This follows from Theorem 2.16 and Stirling's formula:

$$\frac{a_{n,k}}{(2n-1)!!} \sim \frac{e^{-1/2}}{2^k k!}.$$

$\square$

Corollary 2.19. *The ratio of numbers from the k-th row to the numbers of the $(k+1)$-th row of Table 1 tends to $2(k+1)$. In other words, for any $k \in \mathbb{N}$,*

$$\frac{a_{n,k}}{a_{n,k+1}} \sim 2(k+1).$$

Proof. This is an immediate consequence of Corollary 2.18. $\square$

2.5 Patterns of size 3

In this Subsection, we examine endhered patterns of size 3. There are six of them and, as it follows from Lemma 2.4, they are divided into two equivalence classes with respect to distributions of these patterns in matchings (see Fig. 6). Thus, our goal is to study these two different distributions represented, for instance, by patterns 321 and 132. For simplicity, we will use the following notations:

$$c_{n,k} := a_{n,k}(321) = a_{n,k}(123)$$

to denote the number of matchings of size n containing k patterns 321 ($\overset{\frown}{}$), and

$$d_{n,k} := a_{n,k}(132) = a_{n,k}(213) = a_{n,k}(231) = a_{n,k}(312)$$

to designate the number of matchings of size n containing k patterns 132 ($\overset{\frown}{}$).

Theorem 2.20. *For any $n > 0$ and $k \geqslant 0$,*

$$c_{n,0} = \sum_{s=0}^{\lfloor n/2 \rfloor} \binom{n-s}{s} a_{n-s,0},$$

$$c_{n,k} = \sum_{s=1}^{\lfloor (n-k)/2 \rfloor} \binom{k+s-1}{k}\binom{n-k-s}{s} a_{n-k-s,0} \quad \text{if } k > 0, \tag{8}$$

where $a_{n,0}$ is the number of matchings of size n avoiding pattern 21 (⌒⌒).

Proof. Each matchings avoiding pattern ⌒⌒⌒ can be uniquely obtained from a matching avoiding pattern ⌒⌒ by doubling some of its arcs (possibly, none). In particular, to get a matching of size n by doubling exacty s arcs, where $0 \leqslant s \leqslant \lfloor n/2 \rfloor$, we take a matching of size $n-s$ avoiding ⌒⌒, and then choose s of its arcs to transform them into s occurrences of ⌒⌒. Since this can be done in $\binom{n-s}{s}$ ways, we obtain relation (8) for $k = 0$.

Similarly, any matching of size n with $k > 0$ occurrences of pattern ⌒⌒⌒ can be uniquely obtained by the following procedure. First, we fix $0 < s \leqslant \lfloor (n-k)/2 \rfloor$ and take a matching of size $n - k - s$ avoiding ⌒⌒. Second, we choose s arcs in this matching and double them, there are $\binom{n-k-s}{s}$ ways to do it. Finally, we stack additional k arcs onto just created s occurrences of ⌒⌒, which can be done in $\binom{k+s-1}{k}$ ways (apply the stars-and-bars argument [14]). $\qquad\square$

Table 2 shows the first values of $c_{n,k}$.

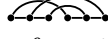| n \ k	1	2	3	4	5	6	7	8	9
0	1	3	14	100	906	10022	130864	1969884	33583700
1	0	0	1	4	34	332	3866	52400	811248
2	0	0	0	1	4	36	362	4304	59256
3	0	0	0	0	1	4	38	392	4752
4	0	0	0	0	0	1	4	40	422
5	0	0	0	0	0	0	1	4	42
6	0	0	0	0	0	0	0	1	4
7	0	0	0	0	0	0	0	0	1

Table 2: Distribution of pattern ⌒⌒⌒ (321).

In order to study the distribution of the occurrences of pattern ⌒⌒, we apply Goulden-Jackson *cluster method* [19, 20] and the inclusion-exclusion principle in the framework of the symbolic method [15, Section III.7.4]. Let us trace the main steps.

We start by considering all possible matchings; the corresponding generating function is given by

$$F(z) = \sum_{n=0}^{\infty} (2n-1)!! \, z^n = 1 + z + 3z^2 + 15z^3 + 105z^4 + \dots$$

13

In every matching, we distinguish certain arcs, say, by coloring them violet. Algebraically, this is done by passing to the generating function

$$G(z, v) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} g_{n,k} z^n v^k = F(z + zv),$$

where $g_{n,k}$ is the number of matchings of size n with k violet arcs (marked by the variable v). Next, we replace violet arcs by "thick arcs" constructed from 3 arcs forming pattern (in general case, we should replace them by *clusters* consisting of intersected copies of a given pattern; pattern , however, admits no possible self-intersection). This corresponds to the generating function

$$H(z, v) = G(z, z^2 v).$$

And finally, according to the inclusion-exclusion principle, we obtain the generating function

$$D(z, u) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} d_{n,k} z^n u^k = H(z, u - 1)$$

whose coefficients $d_{n,k}$ enumerate the matchings of size n having k occurrences of the pattern (see Figure 12 illustrating the process).
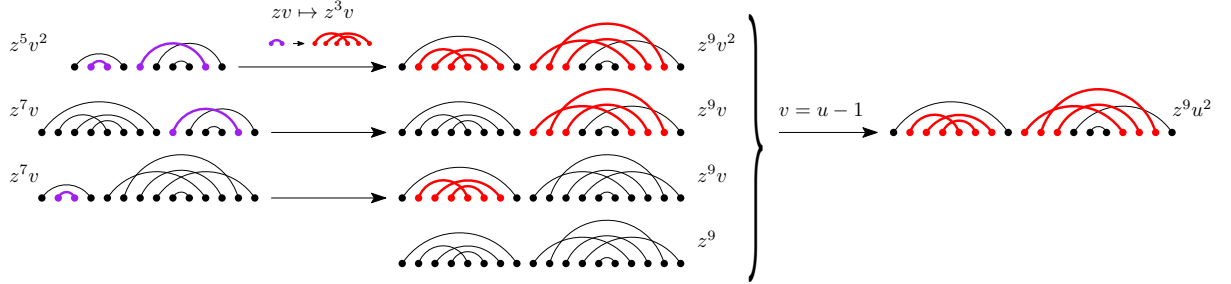
Figure 12: Cluster method and the inclusion-exclusion principle.

Thus, the following result takes place.

Theorem 2.21. *The bivariate generating function*

$$D(z, u) = \sum_{n=0}^{\infty} \sum_{k=0}^{n} d_{n,k} z^n u^k,$$

where $d_{n,k}$ is the number of matchings of size n containing k patterns can be expressed as

$$D(z, u) = F\big(z + (u - 1)z^3\big) = \sum_{n=0}^{\infty} (2n - 1)!! \big(z + (u - 1)z^3\big)^n.$$

k \ n	1	2	3	4	5	6	7	8	9
0	1	3	14	99	900	9978	130455	1965285	33522915
1	0	0	1	6	45	414	4635	61110	927090
2	0	0	0	0	0	3	45	630	9405
3	0	0	0	0	0	0	0	0	15

Table 3: Distribution of pattern ⌢⌢⌢ (132).

The limit distributions of patterns ⌢⌢⌢ and ⌢⌢⌢ are different. In the case of ⌢⌢⌢, for a fixed $k \geqslant 0$, as $n \to \infty$, we have

Table 3 shows the first values of $d_{n,k}$.

$$c_{n,k} \sim \sqrt{2} \cdot C_k \cdot \frac{(2n)^{n-k}}{e^n},$$

where

$$C_0 = 1 \qquad \text{and} \qquad C_k = \sum_{s=1}^{k} \binom{k-1}{s-1} \frac{1}{2^s s!} \quad \text{if } k > 0.$$

On the other hand, for ⌢⌢⌢ takes place

$$d_{n,k} \sim \frac{2^{n-2k+1/2}}{k! e^n} \cdot n^{n-k},$$

as $n \to \infty$. In other words, for large values of n, we have

$$\frac{c_{n,k}}{(2n-1)!!} \sim \frac{C_k}{2^k} \cdot \frac{1}{n^k} \qquad \text{and} \qquad \frac{d_{n,k}}{(2n-1)!!} \sim \frac{1}{2^{2k} k!} \cdot \frac{1}{n^k},$$

meaning that a large uniform random matching avoids both patterns with high probability. The proof of this asymptotic behavior requires other technical means and is beyond the scope of this paper. We will present this proof as part of the analysis of the asymptotic behavior of any endhered pattern in our next work [6].

3 Occurrences of endhered patterns in RNA secondary structures and shapes

In this section, we shift our attention to the real-world data and discuss how endhered patterns are distributed in various representations of the secondary RNA structures obtained from Protein Data Bank [5] using our Python script (https://gitlab.com/celiabiane/endhered_pattern) and x3DNA-DSSR computer program [28, 29].

From the RNA secondary structures obtained with x3NA-DSSR, we collapsed unpaired nucleotides in order to keep only paired ones. When keeping only RNA structures composed of one chain, this leads to 1408 RNA structures, 881 of which are non empty. These structures in the extended dot-bracket notation are converted in matchings using
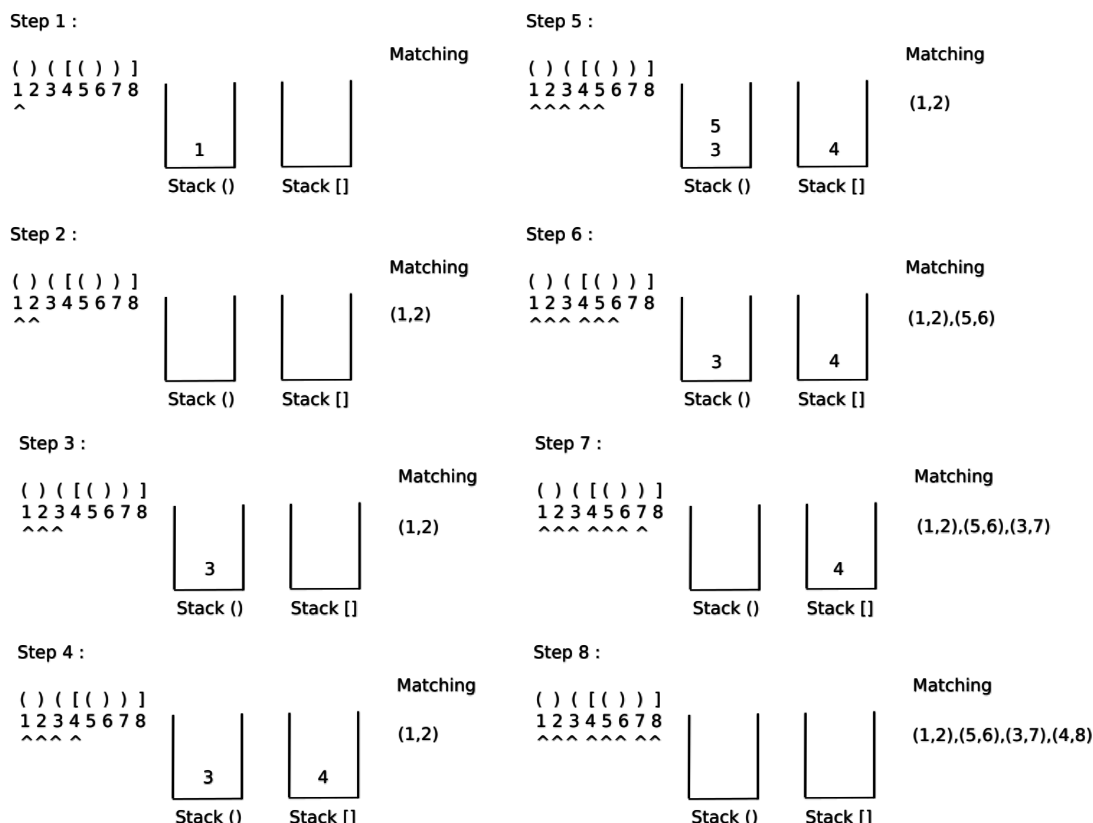
Figure 13: Example of conversion from dot-bracket notation to a matching

an algorithm based on stacks. The algorithm works as follows: the word composed of parentheses is read from left to right, when an opening character is met its index is stacked in a stack corresponding to the nature of the character. When a closing character is met, a pair corresponding to the last index in the corresponding stack and the current index is added to the matching, and the index of the opening parenthesis is unstacked. Figure 13 shows an example of this conversion.

Among the 881 paired and non empty secondary structures, 874 contains the pattern ⌢. The seven remaining structures, with PDB identifiers 3IZY, 5UZ9, 6B45, 6B46, 6B47, 6B48, 6VQX, contain a single base pairing.

None of the considered 881 structures are encoded with more than 4 types of parentheses. Only 2 structures (4DS6 and 4E8V) are written using 4 types of parentheses, 25 employ 3 types of parentheses (4DS6, 4E8V, 5KMZ, 5TPY, 6AGB, 6AHR, 6UES, 6WLQ, 6WLR, 6WLS, 7EZ0, 7K16, 7U4A, 7UO5, 7UVT, 7XD3, 7XD4, 7XD7, 7XSK, 7XSL, 7XSM, 7XSN, 8HD6, 8HD7, 8I7N), and 212 structures are encoded with 2 types of parentheses.

The distribution of patterns ⌢ and ⌢ in RNA secondary structure is shown in Figure 14. We can observe that the majority of RNA secondary structures are of small size (between 0 and 100 nucleotides), and there seems to be a linear relation between the length of the matching and the number of occurrence of the pattern. This can be explained by the large number of hairpin patterns in RNA secondary structures. Only 2 RNA structures contain the ⌢ pattern: 4M4O, 5U3G in PDB. One of them, 4M4O, corresponds to the structure of the minE aptamer involved in a complex with a lysozyme.

It has the following secondary structure:

$$(((((((.(((((.((.......((((.(.(.[..)]..).)))))).)))))).))))))$$

Another one, 5U3G, is the Dickeya dadantii ykkC riboswitch, which has the following secondary structure:

$$((((((..((...(((((....))))).)).....))))))(((((((....))))))))·(.[...(((.......))).).·].·$$

Pattern 231 is contained only in 4M4O, while pattern 312 belongs only to 5U3G. Pattern 321 is contained in 866 structures. It is surprising that pattern ⌒⌒ appears so rarely in RNA structures while pseudoknots are thought to have important biological functions. We observe that occurrence of this patter are "hidden" by the high frequency of nested bonds. To neutralise this effect, we pass to RNA shapes.

The concept of RNA shapes have been introduced by Giegerich, Voss and Rehmsmeier in 2004 [17]. In these shapes, no unpaired regions are included and nested bonds are combined. For instance, the secondary structure
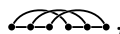
$$..(((.((..(((....)))·(((.....)))))))))..$$

has the following RNA shape:

$$(()())$$

Originally, the RNA shapes have been defined in words with a single type of parenthesis. They are counted by Motzkin numbers [12, 13, 17] and correspond exactly to non-crossing matchings avoiding the endhered pattern ⌒⌒.

We adapt the Giegerich-Voss-Rehmsmeier reduction to matchings with crossings represented by words with different types of parentheses. The result of this adaptation is what Rødland called *collapsed structures* [37]. This is done by keeping only (i, j) pairs in matching such that $(i + 1, j - 1)$ does not belong to the matching and then reindexing the pairs. The number of ⌒⌒ pattern in resulting reduced matchings of RNA shapes is obviously 0. Interestingly, the number of RNA shapes with at least one occurrence of pattern ⌒⌒ increases up to 55. Among those, 8 RNA structures have pattern ⌒⌒⌒ and 13 have pattern ⌒⌒⌒. Other size 3 patterns (123, 213, 132) are not detected in RNA shapes. It is expected for patterns 213 and 132, as they contain the pattern ⌒⌒ forbidden in RNA shapes.

4 Conclusion and discussions

We have examined distributions of endhered patterns of sizes 2 and 3 in matchings from theoretico-combinatorial and data-driven points of view. In matchings, patterns ⌒⌒ and ⌒⌒ have the same distribution. Six endhered patterns of size 3 are divided into 2 equidistributed groups: ⌒⌒⌒, ⌒⌒⌒ and ⌒⌒⌒, ⌒⌒⌒, ⌒⌒⌒, ⌒⌒⌒. Moreover, the joint distribution of patterns of the same size is symmetric if the patterns are equivalent under the twist operation. We have also provided corresponding asymptotic behavior of these distributions.

In this work, we deliberately abstract from the nucleic acid sequences, and model the secondary structures directly by matchings. Our results show that there is a big

Pattern	RNA secondary structure	Extended RNA shape
	874	0
	2 (4M4O, 5U3G)	55 (1A60, 1E95, 1KAJ, 1KPD, 1KPY, 1KPZ, 1L2X, 1L3D, 1RNK, 1YG3, 1YG4, 1YMO, 2A43, 2AP0, 2AP5, 2G1W, 2K95, 2K96, 2LC8, 2M58, 2M8K, 2RP0, 2RP1, 2TPK, 2XDB, 3IYQ, 3IYR, 3IZ4, 3U4M, 3U56, 3UMY, 437D, 4M4O, 4PQV, 4QG3, 4QVI, 4R8I, 5KMZ, 5NPM, 5TPY, 5U3G, 6AGB, 6AHR, 6DLQ, 6DLR, 6DLS, 6DLT, 6DNR, 6E1T, 6E1V, 6VUH, 7K16, 7U4A, 7UO5, 8HIO)
	1 (4M4O)	8 (3U4M, 3U56, 3UMY, 4M4O 4QG3, 4QVI, 4R8I, 5NPM)
	1 (5U3G)	13 (3U4M, 3U56, 3UMY, 4QG3, 4QVI, 5KMZ, 5NPM, 5U3G, 6DLQ, 6DLR, 6DLS, 6DLT, 6DNR)
	866	0
	0	0

Table 4: Number of pattern occurrences in RNA secondary structures and RNA shapes.

difference between observed and modelled pattern distributions. This means that non-restricted matchings are too permissive and new models should be developed to get closer to the observed pattern distributions. We wonder if it is possible to describe essential features of RNA secondary structures with pseudoknots, using pattern-based restrictions. The classical Waterman's definition of RNA structures and its generalizations (see Subsection 1.1) can be regarded as an example of pattern-based restriction used, among other things, to control the prediction algorithms for secondary structures from nucleic acid sequences. More insights into patterns frequencies in the real-world secondary RNA structures may guide us towards mixture of two definitons, complex enough to cover different pseudoknot-like structures presented in real data, but at the same time quite simple and neat to prevent the uncontrolled combinatorial explosion.

The non-existence of certain short sequences in genomic and protein sequences is a well-known fact [21, 23, 26, 33]. Applications include cancer research [1, 31] and forensic science [18]. Less is known about the forbidden secondary structures, although some interesting works about theoretical (im)possibility of inverse RNA folding have been published recently [40, 41]. One of the following research directions would be to determine what influence the distribution of endhered patterns in the real-world RNA. Some configurations are probably forbidden due to physicochemical constraints on the bending of RNA (something like Waterman-Ponty restrictions, but for the case that include pseudoknots), others may not be present because of biological reasons. Are there some evolutionary mechanisms that divert the distribution of patterns from the theoretically observed in the equi-probabilistic model of matchings? How pattern distributions in secondary structures are related to RNA dynamics and function?

For any new combinatorial characterization of RNA structures, we need to develop a method for estimating their affinity with structures observed in real-world molecules. One such method could also be pattern-based: compare the distribution of patterns in real-world RNA with theoretically calculated distributions over matchings avoiding certain patterns. Our exploratory study suggest, for instance, that the patterns , and never appear in RNA. Moreover, PDB references presented in Table 4 look very interesting, especially 4M4O, 5U3G.
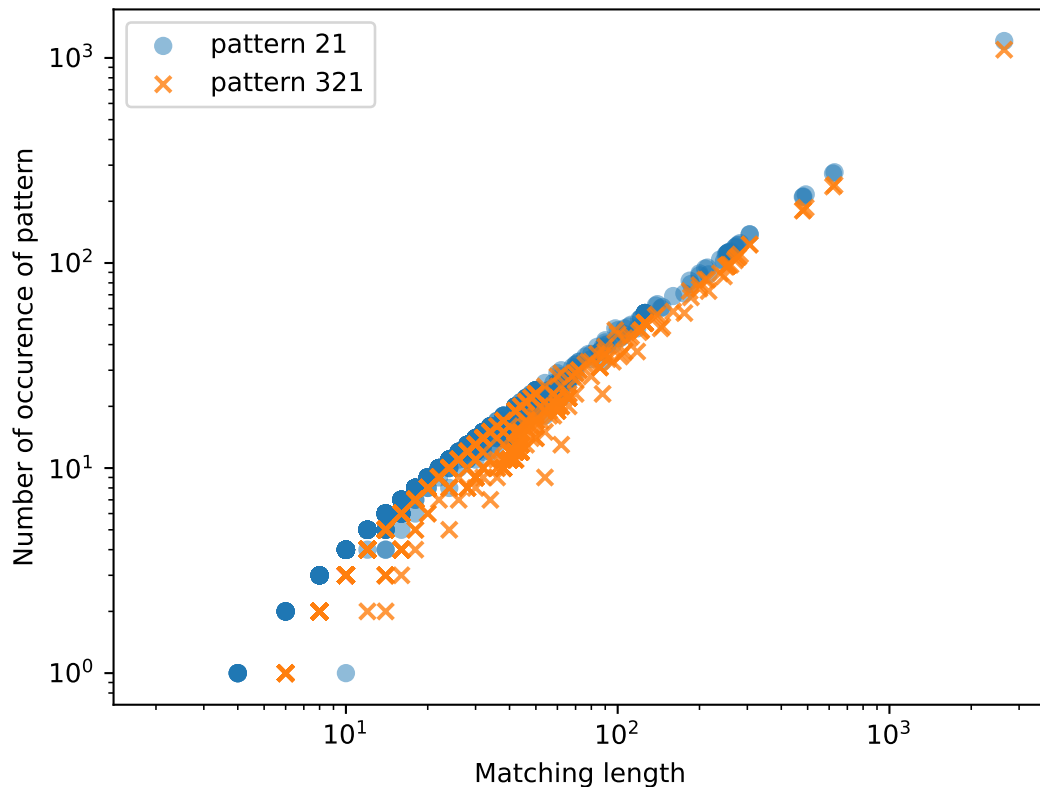
Figure 14: Scatter plot of the number of occurrences of patterns 21 and 321 as a function of the length of the matching. Axes are in logarithmic scale. RNA with no occurrences of patterns 21 and 321 are not displayed.

Acknowledgments

References

[1] Nilufar Ali, Cody Wolf, Swarna Kanchan, Shivakumar R. Veerabhadraiah, Laura Bond, Matthew W. Turner, Cheryl L. Jorcyk, and Greg Hampikian. 9S1R nullomer peptide induces mitochondrial pathology, metabolic suppression, and enhanced immune cell infiltration, in triple-negative breast cancer mouse model. *Biomedicine & Pharmacotherapy*, 170:115997, 2024.

[2] Jørgen Ellegaard Andersen, Josef Mattes, and Nicolai Reshetikhin. The poisson structure on the moduli space of flat connections and chord diagrams. *Topology*, 35(4):1069–1083, 1996.

[3] Jörg Arndt. *Matters Computational: Ideas, Algorithms, Source Code*. Springer, Berlin, Heidelberg, 2011.

[4] Jean-Luc Baril. Avoiding patterns in irreducible permutations. *Discrete Mathematics & Theoretical Computer Science*, 17(3):2158, 2016.

[5] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[6] C. Biane, S. Kirgizov, and K. Nurligareev. Clusters of endhered patterns in matchings. In preparation.

[7] Jonathan Bloom and Sergi Elizalde. Pattern avoidance in matchings and partitions. *The electronic journal of combinatorics*, 20(2), 2013.

[8] R Campoamor-Stursberg and VO Manturov. Invariant tensor formulas via chord diagrams. *Journal of Mathematical Sciences*, 128(4):3018–3029, 2005.

[9] Matteo Cervetti and Luca Ferrari. Enumeration of some classes of pattern avoiding matchings, with a glimpse into the matching pattern poset. *Annals of Combinatorics*, 26(4):971–995, 2022.

[10] William Y. C. Chen, Eva Y. P. Deng, Rosena R. X. Du, Richard P. Stanley, and Catherine H. Yan. Crossings and Nestings of Matchings and Partitions. *Transactions of the American Mathematical Society*, 359(4):1555–1575, 2007.

[11] William YC Chen, Toufik Mansour, and Sherry HF Yan. Matchings avoiding partial patterns. *The Electronic Journal of Combinatorics*, 13(1):R112, 2006.

[12] Sang Kwan Choi, Chaiho Rim, and Hwajin Um. Narayana Number, Chebyshev Polynomial and Motzkin Path on RNA Abstract Shapes. In Jan de Gier, Cheryl E. Praeger, and Terence Tao, editors, *2017 MATRIX Annals*, MATRIX Book Series, pages 153–166. Springer International Publishing, Cham, 2019.

[13] Marie-Pierre Delest and Gérard Viennot. Algebraic languages and polyominoes enumeration. *Theoretical Computer Science*, 34(1):169–206, 1984.

[14] William Feller. *An introduction to probability theory and its applications, volume 2*. John Wiley & Sons, 1991.

[15] Philippe Flajolet and Robert Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge ; New York, 2009. OCLC: ocn244767782.

[16] Hin Hark Gan, Samuela Pasquali, and Tamar Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. *Nucleic acids research*, 31(11):2926–2943, 2003.

[17] Robert Giegerich, Björn Voss, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Research*, 32(16):4843–4851, 2004.

[18] Jayita Goswami, Michael C. Davis, Tim Andersen, Abdelkrim Alileche, and Greg Hampikian. Safeguarding forensic DNA reference samples with nullomer barcodes. *Journal of Forensic and Legal Medicine*, 20(5):513–519, 2013.

[19] Ian P. Goulden and David M. Jackson. An inversion theorem for cluster decompositions of sequences with distinguished subsequences. *Journal of the London Mathematical Society*, s2-20(3):567–576, 1979.

[20] Ian P. Goulden and David M. Jackson. *Combinatorial enumeration*. Wiley, 1983.

[21] Greg Hampikian and Tim Andersen. Absent sequences: nullomers and primes. In *Biocomputing*. World Scientific, 2007.

[22] Christian Haslinger and Peter F. Stadler. RNA structures with pseudo-knots: graph-theoretical, combinatorial, and statistical properties. *Bulletin of mathematical biology*, 61:437–467, 1999.

[23] Julia Herold, Stefan Kurtz, and Robert Giegerich. Efficient computation of absent words in genomic sequences. *BMC Bioinformatics*, 9(1), 2008.

[24] Vít Jelínek. Dyck paths and pattern-avoiding matchings. *European Journal of Combinatorics*, 28(1):202–213, 2007.

[25] Vít Jelínek and Toufik Mansour. Matchings and Partial Patterns. *The Electronic Journal of Combinatorics*, pages R158–R158, 2010.

[26] Grigorios Koulouras and Martin C Frith. Significant non-existence of sequences in genomes and proteomes. *Nucleic Acids Research*, 49(6):3139–3155, 2021.

[27] W. A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 15(1):31–63, 2008.

[28] Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, 43(21):e142, 2015.

[29] Xiang-Jun Lu and Wilma K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, 2003.

[30] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research*, 29(22):4724–4735, 2001.

[31] Austin Montgomery, Georgios Christos Tsiatsianis, Ioannis Mouratidis, Candace S. Y. Chan, Maria Athanasiou, Anastasios D. Papanastasiou, Verena Kantere, Nikos Syrigos, Ioannis Vathiotis, Konstantinos Syrigos, Nelson S. Yee, and Ilias Georgakopoulos-Soares. Utilizing nullomers in cell-free RNA for early cancer detection. *Cancer Gene Therapy*, 2024.

[32] OEIS Foundation Inc. The On-Line Encyclopedia of Integer Sequences, 2024. Published electronically at https://oeis.org.

[33] Armando J Pinho, Paulo JSG Ferreira, Sara P Garcia, and João MOS Rodrigues. On finding minimal absent words. *BMC Bioinformatics*, 10(1), May 2009.

[34] Yann Ponty. Ensemble algorithms and analytic combinatorics in RNA bioinformatics and beyond, 2020.

[35] Michela Quadrini. Structural relation matching: an algorithm to identify structural patterns into RNAs and their interactions. *Journal of Integrative Bioinformatics*, 18(2):111–126, 2021.

[36] Christian M. Reidys and Rita R. Wang. Shapes of RNA pseudoknot structures. *Journal of Computational Biology*, 17(11):1575–1590, 2010.

[37] Einar Andreas Rødland. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *Journal of Computational Biology*, 13(6):1197–1213, 2006.

[38] Vincent Vajnovszki. Generating involutions, derangements, and relatives by ECO. *Discrete Mathematics & Theoretical Computer Science*, 12(1):479, 2010.

[39] Michael S. Waterman. Secondary structure of single-stranded nucleic acids. *Adv. math. suppl. studies*, 1:167–212, 1978.

[40] Hua-Ting Yao, Cédric Chauve, Mireille Regnier, and Yann Ponty. Exponentially few RNA structures are designable. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 289–298. Association for Computing Machinery, 2019.

[41] Tianshuo Zhou, Wei Yu Tang, David H. Mathews, and Liang Huang. Undesignable RNA structure identification via rival structure generation and structure decomposition. 2024. To appear in proceedings of RECOMB 2024, preprint https://arxiv.org/abs/2311.08339v3.