ASYMPTOTICS OF RNA SHAPES

W.A. LORENZ, Y. PONTY, AND P. CLOTE

ABSTRACT. RNA shapes, introduced by Giegerich et al. (17), provide a useful classification of the branching complexity for RNA secondary structures. In this paper, we derive an exact value for the asymptotic number of RNA shapes, by relying on an elegant relation between non-ambiguous, context-free grammars and generating functions. Our results provide a theoretical upper bound on the length of RNA sequences amenable to probabilistic shape analysis (37; 41), under the assumption that any base can basepair with any other base. Since the relation between context-free grammars and asymptotic enumeration is simple yet not well-known in bioinformatics, we give a self-contained presentation with illustrative examples. Additionally, we prove a surprising 1-to-1 correspondence between π-shapes and Motzkin numbers.

1. Introduction

Recently, there has been an intense interest in RNA due to the surprising, previously unsuspected regulatory and catalytic roles played by ribonucleic acid in what until now has been primarily a predominantly protein-centric view of molecular biology. Apart from its long-understood roles as mRNA and tRNA, ribonucleic acid molecules play a catalytic role in peptide bond formation (45; 45) and in intron splicing (40), both examples of enzymatic RNAs called *ribozymes* (14). RNA also plays a role in post-transcriptional gene regulation by RNA interference (RNAi), for which discovery, A.Z. Fire and C.C. Mello were awarded the 2006 Nobel Prize in Physiology or Medicine. By quite different means, RNA performs transcriptional and translational gene regulation by allostery, where a portion of the $5'$ untranslated region ($5'$ UTR) of mRNA known as a *riboswitch* (33; 46) can undergo a conformational change upon binding a specific ligand such as adenine, guanine, lysine, etc. RNA is known as well to play critical roles in various other cellular mechanisms including dosage compensation (7), protein shuttling (42), retranslation events such as selenocysteine insertion (12) and ribosomal frameshift (4; 29), etc.

As in the case of protein, the function of RNA often depends on its tertiary structure.[1] Since such tertiary contacts disappear much earlier than stacked base pairs when temperature is raised (3), it is commonly believed that RNA secondary structure serves as a scaffold for tertiary structure formation. For this reason, accurate prediction of RNA secondary structure is an important problem of computational biology.

Ab initio RNA secondary structure prediction by free energy minimization (Zuker (49)) is one of the real successes of bioinformatics, along with sequence alignment (Smith-Waterman (36), BLAST (1), PSI-BLAST (2)). Indeed, minimum free energy (MFE) secondary structure prediction algorithms currently average 73% accuracy for sequences of length bounded by 700 (24). Reasons for this success depend on a combination of techniques deriving from physical chemistry, mathematics and computer science: *(i)* a realistic *nearest neighbor* energy model pioneered by Tinoco (19; 20), *(ii)* improved, experimentally determined free energies for stacked base pairs and loops (25; 47), *(iii)* a simple mathematical representation of secondary structures as generalized *balanced parenthesis* expressions, which are generated by

---

[1]An exception to this statement is afforded by mRNA and small RNAs, such as the approximately 21 nt. microRNAs (22), which effect post-transcriptional gene regulation by hybridizing to mRNA.

a *context-free grammar* (21), *(iv)* an efficient dynamic programming algorithm which runs in time $O(n^3)$ and space $O(n^2)$, where n is the length of the input RNA sequence (26; 49).

Due to the simple combinatorial representation of secondary structures, it is possible to apply methods of enumerative combinatorics to determine the asymptotic number of RNA secondary structures, a result first obtained by Stein and Waterman (38; 44) by using a result known as Bender's theorem (5). Although the general idea is sound, the hypotheses given in (5) are not sufficient for the conclusion of the theorem to hold; indeed Canfield (8) gave a counterexample to the statement of Bender's theorem, and Meir and Moon (28) provided a somewhat less general result, which nevertheless covers many enumeration problems.

The following version of Bender-Meir-Moon is stated as Theorem 10.13 on page 1162 of (32).

Theorem 1.1 (Bender, Meir and Moon, Odlyzko)**.** *Suppose that $f(z) = \sum_{n=1}^{\infty} f_n z^n$ is analytic at $z = 0$, that $f_n \geq 0$ for all n, and that $f(z) = G(z, f(z))$, where $G(z, w) = \sum_{m,n \geq 0} g_{m,n} z^m w^n$. Suppose that there exist real numbers $\delta, r, s > 0$ such that*

- $G(z, w)$ *is analytic in* $|z| < r + \delta$ *and* $|w| < s + \delta$.
- $G(r, s) = s$, $G_w(r, s) = 1$,
- $G_z(r, s) \neq 0$ *and* $G_{w,w}(r, s) \neq 0$.

Suppose that $g_{m,n}$ is real and non-negative for all m, n, that $g_{0,0} = 0$, $g_{0,1} \neq 1$,[2] and $g_{m,n} > 0$ for some m and some $n \geq 2$. Assume further that there exist $h > j > i \geq 1$ such that $f_h f_i f_j \neq 0$ while the greatest common divisor of $j - i$ and $h - i$ is 1. Then $f(z)$ converges at $z = r$, $f(r) = s$, and

$$f_n = [z^n] f(z) \sim \sqrt{\frac{r G_z(r, s)}{2\pi G_{w,w}(r, s)}} r^{-n} n^{-3/2}.$$

In (18), Hofacker et al. extended results of Stein and Waterman to determine the asymptotic number of various parameters related to RNA secondary structure – parameters such as the expected number of base pairs, average number of hairpin loops, expected size of bulges, etc. In (34), Rodland applied the Bender-Meir-Moon Theorem to compute the asymptotic number of RNA secondary structures including certain

---

[2]In Theorem 10.13 on page 1162 of (32), this condition is (incorrectly) stated as $g_{0,1} = 1$, a typographic error, as evidenced by the example 10.14 on pages 1162-1163, for which $g_{0,1} \neq 1$. Odlyzko mentions that his statement of the theorem of Meir and Moon includes some of his own corrections to (28).

types of pseudoknots. Finally, in (10), we applied the theorem of Meir
and Moon (28) to determine the asymptotic number of *saturated* RNA
secondary structures; here, a structure S is saturated (48) if no base
pairs can be added to S without violating the definition of secondary
structure; equivalently S is saturated if it is locally optimal with respect
to the Nussinov-Jacobson energy model (31).

All of the previous asymptotic results were obtained by the following
approach.

Method 1.2.

(a) *Inductively define the number a_n of objects of interest for length
   n RNA by a recurrence relation[3] usually involving a convolution
   – i.e. a sum of the general form $\sum_{1 \leq k < n} S_k \cdot S_{n-k}$.*
(b) *For the generating function $w = \sum_{n=0}^{\infty} a_n z^n$, determine a si-
   multaneous solution $z = r$, $w = s$ for the (in general nonlinear)
   functional equations $G(z, w) = w$ and $G_w(z, w) = 1$, where G_w
   denotes the partial derivative of G with respect to w.*
(c) *If G and a solution $x = r$, $y = s$ satisfy the hypotheses of the
   Bender-Meir-Moon Theorem 1.1, then*

$$a_n \sim \sqrt{\frac{r G_z(r, s)}{2\pi G_{ww}(r, s)}} \; n^{-3/2} \; r^{-n}$$

In place of the Bender-Meir-Moon Theorem 1.1, we make use of
Corollary 2 of Flajolet and Odlyzko (part (i) of (16) on page 224),
restated here as the following theorem. (Undefined concepts will be
explained later.)

Theorem 1.3 (Flajolet and Odlyzko)*. Assume that $f(z)$ has a singu-
larity at $z = 1$ and is analytic in the region $\triangle \backslash 1$, depicted in Figure
A1 in the Appendix, and that as $z \to 1$ in $\triangle$,*

$$f(z) \sim K(1 - z)^\alpha$$

Then, as $n \to \infty$, if $\alpha \notin 0, 1, 2, ...,$

$$f_n \sim \frac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1}.$$

---

[3]Such recurrence relations form the basis for dynamic programming algorithms
to count the number of structures, to determine the minimum free energy (MFE)
structure (31; 49) and to compute the Boltzmann partition function (27), which
latter yields thermodynamic parameters such as free energy, heat capacity, expected
internal energy, etc.

In contrast with Method 1.2, the approach taken in this article is as follows.

Method 1.4.

(a) *Define a non-ambiguous context-free grammar G which generates the set of all combinatorial objects, regardless of length.*

(b) *Use the DSV methodology[4] to immediately write down an explicit function for the generating function $w = f(z) = \sum_{n=0}^{\infty} a_n z^n$. In applications, it often happens that $f(z)$ is a quotient of functions involving fractional powers of polynomials.*

(c) *Determine the dominant singularity ρ of $f(z)$. Rescale so that ρ may be assumed to equal 1, and apply the Flajolet-Odlyzko Theorem 1.3 to obtain an explicit formula for the asymptotic value of a_n.*

See Vauchaussade (39) for additional explanation of the DSV method, and see Nebel (30) for an application of the Flajolet O-transfer method and singularity analysis.

Advantages of the latter method are twofold. First, derivation of the non-ambiguous context-free grammar and application of DSV methodology as summarized in (a),(b) of Method 1.4 is much easier than complicated error-prone algebraic manipulations required to obtain (a),(b) of Method 1.2. Second, it is often difficult or impossible to explicitly verify the hypotheses of the Bender-Meir-Moon Theorem 1.1. In contrast, it is more straightforward to verify the hypotheses of the Flajolet-Odlyzko Theorem 1.3.

The plan of this paper is as follows. In Section 2 we explain the relation between context-free grammars and generating functions, known as the DSV method, and we show how to rescale the dominant singularity ρ to 1 in order to apply the Flajolet-Odlyzko Theorem. (See the Appendix for a clear explanation of any skipped details.) We illustrate Method 1.4 by providing a simpler derivation for the well-known asymptotic number of secondary structures (38). Our goal in reviewing this material is to provide a broad understanding to the bioinformatics community of the power and simplicity of the DSV method in finding generating functions of combinatorial objects, and of the singularity analysis of Flajolet and Odlyzko (16) to determine the asymptotic number of combinatorial objects described by these generating functions. In Section 3, we present our new results concerning the asymptotic

---

[4]Presumably named after Dyck, Schützenberger and Viennot.

number of RNA shapes. First, Section 3.1 presents background material on RNA shapes (17; 37; 41). Second, in Section 3.2, we derive the asymptotic number of π-shapes and of π-shapes compatible with some secondary structure for a length n RNA sequence, under the assumption that any base can basepair with any other base, and that there is a minimum of one unpaired base in every hairpin loop. Third, In Section 3.3, we derive the asymptotic number of π'-shapes and of π'-shapes compatible with some secondary structure for a length n RNA sequence. Section 3.4 presents a surprising one-to-one correspondence between π-shapes having size $2n + 2$ and and Motzkin words having size n. In Section 4, we present a sharper asymptotic count on the number of π-shapes having k stems or pairs of brackets. Taken together, our results provide evidence for the exponential time required by the program `RNAshapes` of Giegerich and co-workers, which latter computes the Boltzmann probability for occurrences of various RNA shapes for a given sequence. Finally, in the Appendix, we present a detailed, self-contained proof from basic principles of how to apply the method of Flajolet and Odlyzko (16).

Source code for programs developed in this paper is available at the web supplement `bioinformatics.bc.edu/clotelab/RNAshapes/webSupplement`.

2. Method and Materials

In this section, we define non-ambiguous context-free grammars and describe the DSV methodology. Since the asymptotic number of RNA secondary structures on n is both well-known (38) and not difficult to obtain, we illustrate the classic approach (recurrence relations and Bender's Theorem) with our current approach (DSV methodology and Theorem of Flajolet and Odlyzko). We begin by recalling the definition of RNA secondary structure.

Definition 2.1. *A secondary structure S on RNA sequence $s_1, \ldots, s_n$ is defined to be a set of ordered pairs (i, j), such that $1 \leq i < j \leq n$ and the following are satisfied.*

(1) Watson-Crick or GU wobble pairs: *If (i, j) belongs to S, then pair (a_i, a_j) must be one of the following canonical basepairs: (A, U), (U, A), (G, C), (C, G), (G, U), (U, G).*
(2) Threshold requirement: *If (i, j) belongs to S, then $j - i > \theta$, where θ, generally taken to be equal to 3, is the minimum number of unpaired bases in a hairpin loop; i.e. there must be at least θ unpaired bases in a hairpin loop.*
(3) Nonexistence of pseudoknots: *If (i, j) and (k, ℓ) belong to S, then it is not the case that $i < k < j < \ell$.*
(4) No base triples: *If (i, j) and (i, k) belong to S, then $j = k$; if (i, j) and (k, j) belong to S, then $i = k$.*

In this paper, we are interested in the asymptotic number of structures and of shapes of an RNA sequence of length n, so we follow the convention of Stein and Waterman (38; 44) by assuming that any position i can base-pair with any any position j, provide only that $|j - i| > \theta$; i.e. condition (1) of Definition 2.1 is dropped. From this point on, we will speak of a secondary structure S on the sequence $1, \ldots, n$, rather than on the nucleotide sequence $s_1, \ldots, s_n$. For brevity, we may say that S is a secondary structure on n. The size of secondary structure S is the number of base pairs belonging to S, whereas the length of S is the length of the Vienna dot bracket notation equivalent to S. Thus S is a secondary structure on n exactly when S has length n. Since the nature of the nucleotide or base a_i located at position i is not pertinent to the combinatorial study in this paper, by abuse of notation, we may say that i is a *base*.

Following (38; 44), we illustrate Method 1.2 consisting of recurrence relations and Bender-Meir-Moon by outlining the derivation of the asymptotic number

$$S(n) \sim \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3 + \sqrt{5}}{2} \right)^n \sim 1.104366 \cdot 2.618034^n / n^{3/2}.$$

of secondary structures on n. As explained above, this assumes that the minimum number θ of unpaired bases in a hairpin loop is taken to be 1 and that each bases can basepair with any other base.

2.1. Context-free grammars and DSV method.
In this section, we illustrate Method 1.4 consisting of the DSV method and Flajolet-Odlyzko. We recall the definition of non-ambiguous context-free grammars and explain the DSV method which relates such grammars with generating functions.

2.1.1. Some context on context-free grammars.
Let Σ be a finite set of symbols. A language is a subset of Σ^*, the set of all words $a_1, \ldots, a_n$, where $a_i \in \Sigma$ for all $0 \le i \le n$ and n is an arbitrary integer. In this paper, Σ will consist of left parenthesis (, right parenthesis), and dot • when discussing secondary structures and of left bracket [, right bracket], and dot • when discussing shapes. (Giegerich et al. (17) use an underscore _ to denote an unpaired shape region, while we use dot • to denote this.)

A context-free grammar is given by $G = (V, \Sigma, R, S_0)$, where V is a finite set of nonterminal symbols (also called variables), Σ is a disjoint finite set of terminal symbols, $S_0 \in V$ is the *start* nonterminal, and

$$R \subset V \times (V \cup \Sigma)^*$$

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

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

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

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

A context-free grammar $G = (V, \Sigma, R, S_0)$ is in Chomsky normal form when all rules in R are of the form $A \to BC$, or $A \to a$, where

$A, B, C \in V$ and $a \in \Sigma$. Grammar G is said to be ε-free if either *(i)* $L(G)$ does not contain the empty word, ε, and G contains no rule of the form $A \to \varepsilon$, or *(ii)* $L(G)$ contains the empty word ε, and the only rule occurrence of ε is $S_0 \to \varepsilon$. It is a classical result that every context-free language is generated by a context-free ε-free grammar in Chomsky normal form (21). Note that there do exist context-free languages L which are inherently ambiguous, in the sense that no non-ambiguous context-free grammar generates L.

If $w = w_1 \cdots w_n$ is a word of length n in $L(G)$, where G is a context-free grammar, then a *parse tree* for w is a multifurcating tree T, such that:

(1) w is the word formed by reading from left to right the leaves of T.

(2) The root of T is labeled by S, the initial, "start" variable for the grammar G.

(3) If a node of T is labeled by A, then
   (a) either that node has only one child, which is labeled a and $A \to a$ is a rule of G,
   (b) or that node has k children, labeled by $B_1 | \ldots | B_k$, and $A \to B_1 \cdots B_k$ is a rule of G.

A context-free grammar G is called *non-ambiguous*, if there is no word $w \in L(G)$ which admits two distinct parse trees.

2.1.2. *From grammars to generating functions.* A general approach to the enumeration of combinatorial objects relies on generating functions. The so-called *length generating function* for an object class $\mathcal{C}$ is defined by $C(z) := \sum_{i \geq 0} C_n z^n$, where C_n is the finite number of objects having size n in the class $\mathcal{C}$. From such a function, it is sometimes possible to derive a closed-form formula for the coefficient of order n, denoted by $[z^n]C(z)$, which is also the number C_n of objects of size n. Furthermore, it is almost always possible to efficiently derive the behavior of C_n when n approaches infinity (16), as is described later in the paper. Sometimes, an explicit expression for $C(z)$ is unnecessary, and the asymptotic value of C_n can be derived, for instance by means of Lagrange inversion, from a functional equation involving $C(z)$.

A generating function can be obtained through recurrence relations, which may involve long and arduous calculations; for instance, see (38; 43) for the enumeration of RNA secondary structures, as summarized in (a),(b) of Method 1.2. However, an alternative technique, due to M. Schützenberger and summarized in (a),(b) of Method 1.4, can be used to derive the generating function of $\mathcal{C}$. This technique is known as the DSV method; see (6) for more details. The key idea is

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

TABLE 1. Translation between context-free grammars and generating functions. Here, $G = (V, \Sigma, S_0, R)$ is a given context-free grammar, S, T and U are any nonterminal symbols in V, and t is a terminal symbol in Σ. The length generating functions (l.g.f) for the languages $L(S)$, $L(T)$, $L(U)$ are respectively denoted by $S(z)$, $T(z)$, $U(z)$.

as follows. Instead of counting the objects of $\mathcal{C}$, one may instead count the number of words of a language $\mathcal{L}$ that encodes the objects of $\mathcal{C}$. An ambiguous generative process for the language can then be directly transposed into a set of equations involving $L(z)$, where $L(z)$ is the generating function of $\mathcal{L}$ and $L(z) = C(z)$. See (15) for a survey of actual admissible constructions. When the language $\mathcal{L}$ is context-free, generated by a non-ambiguous context-free grammar $\mathcal{G}$, such equations can be deduced directly from the rules of $\mathcal{G}$, using the scheme in Table 1. The correctness of this translation scheme is given in the following theorem.

Theorem 2.2. *Let $G = (V, \Sigma, R, S_0)$ be a non-ambiguous, ε-free, context-free grammar in Chomsky normal form. For each nonterminal symbol S, let $S(z)$ be the corresponding generating function, defined by applying the translation scheme from Table 1. If $L(z)$ denotes the length generating function for the language $L(G)$, then $S_0(z) = L(z)$.*

PROOF. In order to prove the validity of the previous equations, we introduce the notation $\mathcal{S}$ for the language generated from a given nonterminal S and $\mathcal{S}_n$ for its restriction to words having size n. From the definition of a grammar, we directly get:

$$
\begin{aligned}
S \to T \mid U &\Rightarrow \mathcal{S} = \mathcal{T} \cup \mathcal{U} \\
S \to T\,U &\Rightarrow \mathcal{S} = \mathcal{T} \cdot \mathcal{U} \\
S \to t &\Rightarrow \mathcal{S} = \{t\} &\Rightarrow S(z) = z \\
S \to \varepsilon &\Rightarrow \mathcal{S} = \emptyset &\Rightarrow S(z) = z^0 = 1
\end{aligned}
$$

where the operator *dot* $\cdot$ denotes language concatenation; i.e. the extension to sets of the concatenation operation.

Since the grammar is *non-ambiguous*, the *union* involved in $\mathcal{S} = \mathcal{T} \cup \mathcal{U}$ is disjoint, and the equation can be transposed to the cardinalities:

$$\mathcal{S}_n = \mathcal{T}_n + \mathcal{U}_n$$

After recalling that $T(z) = \sum_{n \geq 0} \mathcal{T}_n z^n$ and $U(z) = \sum_{n \geq 0} \mathcal{U}_n z^n$, we get:

$$T(z) + U(z) = \sum_{n \geq 0} \mathcal{T}_n z^n + \sum_{n \geq 0} \mathcal{U}_n z^n = \sum_{n \geq 0} (\mathcal{T}_n + \mathcal{U}_n) z^n = \sum_{n \geq 0} \mathcal{S}_n z^n = S(z)$$

Moreover, in the case of language concatenation, $\mathcal{S} = \mathcal{T} \cdot \mathcal{U}$, the non-ambiguity of the grammar ensures that each word ω in $\mathcal{S}$ admits a unique decomposition $\omega = \omega_p \omega_s$ such that prefix $\omega_p \in \mathcal{T}$ and suffix $\omega_s \in \mathcal{U}$. Thus, we have

$$\mathcal{S}_n = \sum_{i=0}^{n} \mathcal{T}_i \cdot \mathcal{U}_{n-i}$$

and

$$T(z)U(z) = \sum_{n \geq 0} \mathcal{T}_n z^n \sum_{n \geq 0} \mathcal{U}_n z^n = \sum_{n \geq 0} \sum_{i=0}^{n} \mathcal{T}_i \cdot \mathcal{U}_{n-i} z^n = \sum_{n \geq 0} \mathcal{S}_n z^n = S(z)$$

$\square$

Theorem 2.2 assumes that G is an ε-free grammar for simplicity of notation. There is no loss of generality, since it is well known that such an equivalent form exists for any given non-ambiguous grammar. However, it is unnecessary to put the grammar into ε-free form before applying the translation rules from Table 1, since the proof above can easily be extended to *general rules* of the form $S \rightarrow \alpha_1 \mid \ldots \mid \alpha_k$, where $\alpha_i \in (V \cup \Sigma)^*$ are words over the alphabet of both terminal and nonterminal symbols. Such an extension would involve the introduction of new *dummy* nonterminal characters, each of which appears on the left side in Chomsky-style rules. In fact, this is the basic principle of the Chomsky normal form construction.

Recurrence relations and Bender-Meir-Moon. An alternative to the method of Bender, Meir and Moon is that developed in the paper by Flajolet and Odlyzko (16). We outline the technique here; details and necessary background are given in the Appendix. This alternative, as mentioned in the introduction, is very general, and does not require all of the technical conditions of theorems based on Bender's (5) theorem. This alternative is well suited to a wide class of problems, including problems described in this paper.

In this section, we illustrate the application of Method 1.2 in order to establish a classic result of Stein and Waterman (38) concerning the

asymptotic number

$$s_n \sim \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3 + \sqrt{5}}{2}\right)^n \sim 1.104366 \cdot n^{-3/2} \cdot 2.618034^n$$

of secondary structures on n. As earlier mentioned, it is here assumed that the minimum number θ of unpaired bases in a hairpin loop is 1; i.e. s_n is the number of balanced-parenthesis expressions with dot, such that if i, j form a base pair, then $|j - i| > 1$.

Proposition 2.3 (Stein and Waterman (38)). *We have $s_0 = s_1 = s_2 = 1$ and for all $n > 2$,*

$$s_n = s_{n-1} + \sum_{k=1}^{n-2} s_{k-1} \cdot s_{n-k-1}.$$

PROOF. By induction on n. There is only one empty word, so $s_0 = 1$, and clearly $s_1 = 1 = s_2$. For the inductive case, there are two subcases: either n is not basepaired, or n basepairs with some $k \in \{1, \ldots, n-2\}$. In the former case, the contribution is s_{n-1}. Suppose that n basepairs with some $k \in \{1, \ldots, n-2\}$. Since there are no pseudoknots, if (x, y) is a base pair different than (k, n), then either $1 \le x < y < k$ or $k + 1 \le x < y < n$, hence the contribution is $s_{k-1} \cdot s_{n-k-1}$.   □

Lemma 2.4 (Stein and Waterman (38)). *Letting $w = f(z) = \sum_{n=1}^{\infty} s_n z^n$, we have $w^2 z^2 - w(1 - z - z^2) + z = 0$.*

PROOF.

$$(2) \qquad w^2 = \left(\sum_{n=1}^{\infty} s_n z^n\right)^2 = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{n-1} s_k s_{n-k}\right) z^n.$$

By Proposition 2.3, $s_n = s_{n-1} + \sum_{k=1}^{n-2} s_{k-1} \cdot s_{n-k-1}$. Replacing n by $n + 2$, we have $s_{n+2} = s_{n+1} + \sum_{k=1}^{n} s_{k-1} \cdot s_{n-(k-1)}$. Substituting r for $k - 1$, we have $s_{n+2} = s_{n+1} + \sum_{r=0}^{n} s_r \cdot s_{n-r}$. Since $s_0 = 1$, we have $\sum_{r=0}^{n} s_r \cdot s_{n-r} = s_n + \sum_{r=0}^{n-1} s_r \cdot s_{n-r}$, so

$$s_{n+2} - s_{n+1} - s_n = \sum_{r=0}^{n-1} s_r \cdot s_{n-r}.$$

Now

$$w^2 = \left(\sum_{n=1}^{\infty} s_n z^n\right)^2 = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{n-1} s_k s_{n-k}\right) z^n$$

so

$$w^2 = \sum_{n=1}^{\infty}(s_{n+2} - s_{n+1} - s_n)z^n = \sum_{n=1}^{\infty}s_{n+2}z^n - \sum_{n=1}^{\infty}s_{n+1}z^n - \sum_{n=1}^{\infty}s_n z^n.$$

Note that

$$\frac{w - s_1 z - s_2 z^2}{z^2} = \sum_{n=1}^{\infty}s_{n+2}z^n$$

and

$$\frac{w - s_1 z}{z} = \sum_{n=1}^{\infty}s_{n+1}z^n.$$

Thus

$$w^2 = \frac{w - z - z^2}{z^2} - \frac{w - z}{z} - w$$

Multiply by z^2 to get

$$(3) \qquad z^2 w^2 = w - z - z^2 - zw + z^2 - wz^2$$

so

$$z^2 w^2 - w(1 - z - z^2) + z = 0$$

$$\square$$

Theorem 2.5 (Stein and Waterman (38))**.**

$$s_n \sim \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3 + \sqrt{5}}{2}\right)^n$$

PROOF. Noting that the golden ratio $\alpha = \frac{1+\sqrt{5}}{2}$, the theorem states that s_n has growth rate $\Theta\left(\frac{(1+\alpha)^n}{n^{3/2}}\right)$. From equation (3), we have that the generating function $w = \sum_{n=1}^{\infty}s_n z^n$ satisfies $G(z, w) = w$ where G is defined by

$$\begin{aligned} G(z, w) &= w^2 z^2 - w(1 - z - z^2) + z + w \\ &= w^2 z^2 + wz + wz^2 + z. \end{aligned}$$

Solve the system $G(z, w) = w$, $G_w(z, w) = 1$, i.e.

$$(4) \qquad w^2 z^2 + wz + wz^2 + z = w$$
$$(5) \qquad 2wz^2 + z + z^2 = 1.$$

A solution of equations (4,5) is given by $z = r$, $w = s$, where $r = \frac{2}{3+\sqrt{5}}$ and $s = \frac{1+\sqrt{5}}{2}$. If we can apply Theorem 1.1, then we obtain the desired

$$s_n \sim \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3 + \sqrt{5}}{2}\right)^n \sim 1.104366 \cdot n^{-3/2} \cdot 2.618034^n.$$

Let's verify the hypotheses of Bender-Meir-Moon Theorem 1.1. Clearly $S(z) = \sum_{n=1}^{\infty} s_n z^n$ is analytic at 0, with $s_n \geq 0$ for all n. Since

(6) $$G(z, w) = w^2 z^2 + w z^2 + w z + z$$

we've seen that $G(z, w) = w$. As a polynomial in variables z, w, G is clearly analytic in $|z| < r + \delta$ and $|w| < s + \delta$, and since r, s is a solution of equations (4) and (5), we have $G(r, s) = s$, $G_w(r, s) = 1$. From equation (6), $g_{0,0} = 0$ and $g_{m,n} \geq 0$ for all m, n. The Taylor coefficient $g_{0,1}$ of $z^0 w^1$ is 0, hence $g_{0,1} \neq 1$, and the Taylor coefficient $g_{1,2}$ of $z w^2$ is 1, hence $g_{m,n} > 0$ for some m and some $n \geq 2$. Taking $i = 1$, $j = 2$, $h = 3$, the greatest common divisor of $j - i$ and $h - i$ is 1 and we have $s_i s_j s_h \neq 0$. We have verified all the conditions of Theorem 1.1, and so conclude the proof of equation (1).

DVS and Flajolet-Odlyzko. In this section, we illustrate the application of Method 1.4 and give an alternate proof for the classic result of Stein and Waterman (38) concerning the asymptotic number

(7) $$s_n \sim \sqrt{\frac{15 + 7\sqrt{5}}{8\pi}} n^{-3/2} \left(\frac{3 + \sqrt{5}}{2}\right)^n \sim 1.104366 \cdot n^{-3/2} \cdot 2.618034^n$$

of secondary structures on n.

Consider the context-free grammar G with the following rules:

$$S \to \bullet \mid S \bullet \mid (S) \mid S (S)$$

Motivated by the Nussinov-Jacobson algorithm (31), it is easy to establish by induction on word length that G is a non-ambiguous grammar which generates all non-empty secondary structures. (A minor modification of the grammar generates all secondary structures where $\theta = 3$.) By DVS methodology, the generating function for non-empty Vienna notation expressions for RNA secondary structures is a solution of the following equation:

$$S = z + Sz + Sz^2 + S^2 z^2$$

Notice that this equation is identical to equation (6), and that its derivation took two lines, in contrast with the rather lengthy algebra involving convolutions. By the quadratic formula, the roots of this

equation are S_+ and S_- where

$$S_+ = \frac{1 - z - z^2 + \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}$$

$$S_- = \frac{1 - z - z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}.$$

Since $S(z)$ is analytic at $z = 0$ and S_+ blows up at the origin, we must choose S_-.[5]

The dominant singularity $z = r$ will be that root of $P(z)$ having least modulus, for the polynomial $P(z) = 1 - 2z - z^2 - 2z^3 + z^4$ occurring within the radical. Mathematica computes that the roots are two imaginary roots with modulus 1 and the real roots $\frac{3-\sqrt{5}}{2}, \frac{3+\sqrt{5}}{2}$. It follows that the dominant singularity[6] is $\rho := \frac{3-\sqrt{5}}{2}$. The asymptotic value of the coefficients s_n of the generating series $S(z) = \sum_n s_n z^n$ is determined by the comportment of the function $f(z) := S_-(z) = \frac{1-z-z^2-\sqrt{1-2z-z^2-2z^3+z^4}}{2z^2}$ about the dominant singularity ρ. (See the Appendix for detailed justification of this and other points.) Define $G(z) = \frac{1-z-z^2}{2z^2}$ and $H(z) = -\frac{\sqrt{1-2z-z^2-2z^3+z^4}}{2z^2}$, so that $f(z) = G(z) + H(z)$. Since $G(z)$ is of slow growth, the asymptotic value of s_n is in fact determined by the comportment of $H(z)$ about ρ (see the Appendix for justification). In order to apply Theorem 1.3, we rescale the dominant singularity from ρ to 1 by making the change of variable $x = z/\rho$. This ensures that x approaches 1 exactly when z approaches ρ. Since ρ is a root of $P(z) = 1 - 2z - z^2 - 2z^3 + z^4$, and we are working over the complex numbers, we can factor $(1 - z/\rho)$ out of $P(z)$ to obtain $Q(z) = 1 + 0.618z + 0.618z^2 - 0.382z^3$. Thus

$$\begin{aligned} P(z) &= 1 - 2z - z^2 - 2z^3 + z^4 \\ &= Q(z) \cdot (1 - z/\rho) \\ &= (1 + 0.618z + 0.618z^2 - 0.382z^3) \cdot (1 - z/\rho). \end{aligned}$$

---

[5]Note that the Taylor expansion of $\sqrt{1 - 2z - z^2 - 2z^3 + z^4}$ about $z = 0$ is $1 - z - z^2 - 2z^3 - 2z^4 - 4z^5 - 8z^6 - 16z^7 + \cdots$, where all coefficients of z are negative. Since S_- has a minus sign before the term $\sqrt{1 - 2z - z^2 - 2z^3 + z^4}$, its Taylor expansion at 0 has non-negative coefficients for each term z^n, as required for the generating function $\sum_n s_n z^n$. This is the case for all applications of DSV methodology in this paper.

[6]The dominant singularity is that singularity ρ, which is the only singularity on the circle of convergence $z = |\rho|$; i.e. ρ is the isolated singularity having least modulus $|\rho|$. Later, an example will be given where singularities of a different function include both r and $-r$ of smallest modulus. In such cases, r is not isolated and Theorem 1.3 cannot be directly applied.

It follows that

$$(8) \qquad H(z) \;=\; -\frac{\sqrt{P(z)}}{2z^2} = -\frac{\sqrt{Q(z)}}{2z^2} \cdot (1 - z/\rho)^{-1/2}$$

$$(9) \qquad\qquad = \; -\frac{\sqrt{Q(\rho x)}}{2(\rho x)^2} \cdot (1 - x)^{-1/2}.$$

hence

$$(10) \qquad\qquad H(z) \sim -\frac{\sqrt{Q(\rho x)}}{2(\rho x)^2} \cdot (1 - x)^{-1/2}$$

as x approaches 1, or equivalently z approaches ρ. Let $K = -\frac{\sqrt{Q(\rho)}}{2\rho^2} = -3.91487$ and let $\alpha = -1/2$, and compute that $\Gamma(-1/2) = -2\pi$. The hypotheses of Theorem 1.3 hold, so we conclude that

$$s_n \sim \frac{K}{\Gamma(-1/2)} \cdot n^{-3/2} \cdot (1/\rho)^n = 1.104366 \cdot n^{-3/2} \cdot 2.61803^n$$

which agrees with equation (1). $\square$

   We now use DSV plus Flajolet-Odlyzko to obtain asymptotics for RNA shapes.

3. Asymptotic number of RNA shapes

In this section, we begin by presenting some background material on RNA shapes (17; 37; 41). In Section 3.2, we derive the asymptotic number of π-shapes and of π-shapes compatible with a length n sequence, and in Section 3.3, we derive corresponding values for π'-shapes. Section 3.4 presents a surprising one-to-one correspondence between π-shapes having size $2n + 2$ and and Motzkin words having size n.

3.1. **Computing the shape of a secondary structure.** In (17), Giegerich and co-workers defined an RNA shape to be a particular compact representation of the branching structure of a given RNA secondary structure. From (17), a *shape abstraction* is defined to be a homomorphic mapping from the set of all secondary structures (considered as parse trees with respect to a given context-free grammar over the terminal symbols $\bullet$, **(**, **)**) into the set of well-balanced dot-bracket expressions (considered as parse trees with respect to another given context-free grammar over the terminal symbols $\bullet$, **[**, **]**).[7] Although (17) considered five different shape abstractions, details were given only for the two shape abstractions π and π'; see (17) for the formal definition using tree homomorphisms. For example, the π-shape of the usual cloverleaf secondary structure of tRNA is **[[ ] [ ] [ ]]**, while the less succinct π'-shape is **[$\bullet$ [$\bullet$] $\bullet$ [$\bullet$] $\bullet$ [$\bullet$]] $\bullet$**, since a typical tRNA structure has no unpaired bases on the $5'$ end or between the T-stem and acceptor stem. Another example is given in Figure 1, which depicts two different secondary structures, both having the same π-shape.

If s is a given secondary structure, then to compute the corresponding π-shape, one first removes all dots $\bullet$ and then replaces all stems (base-paired regions possibly interrupted by bulges and internal loops) by a single base pair **[$\cdots$]**. To obtain the corresponding π'-shape, contract all maximal consecutive dots $\bullet^k$ by a single dot $\bullet$, and replace all maximal nested, uninterrupted stacks of base pairs **($^k \cdots$) ^{k}** by a single base pair **[$\cdots$]**. Formally, we have the following linear time algorithm to compute the π- and π'-shape of a secondary structure s, where s is given in Vienna notation.

Algorithm 3.1. function secStr2shape(s,shapeType)
//Input: sec str s in Vienna notation and shape type π, π'

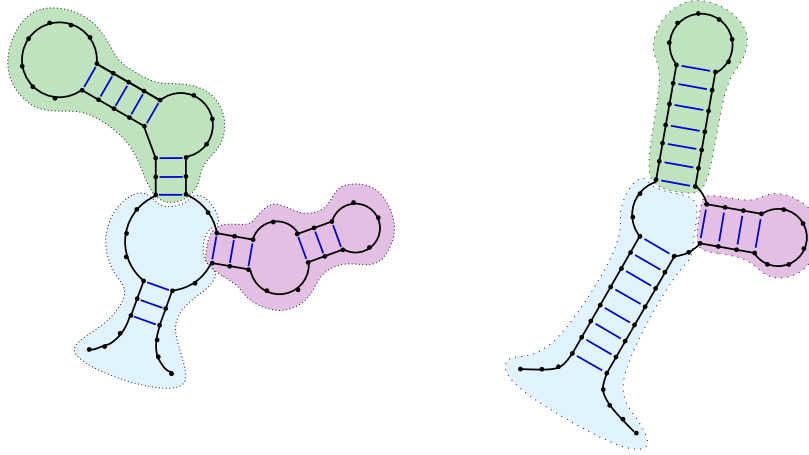[7] In this paper, we use the dot $\bullet$ in place of the underscore symbol $\lrcorner$, which latter is used in (17).

FIGURE 1. Two RNA secondary structures described by
the π-shape `[ [ ] [ ] ]`.

```
//Output:   π-shape or π′-shape of s, depending on shape type
1  if shapeType = π
2    remove all dots from s
3  else if shapeType = π′
4    replace each group of consecutive dots in s by single dot
5  n = |s| //  s = s₁,...,sₙ where dots have been contracted
6  if n ≤ 2 return ●
7  use stack to convert Vienna notation s into list S of base pairs
8  for i = 1 to n,  A[i] = 0
9  for (i,j) in S,  A[i] = j
10 // Array A satisfies A[i] = j if (i,j) ∈ S, else A[i] = 0
11 x = y = 0 // x (resp.  y) denotes lastLeftPos (resp.  lastRightPos)
12 //last pos of left, right paren used to contract adjacent parentheses
13 for i = 1 to n
14   if A[i] > 0 //i is base paired to j = A[i]
15     if i = x + 1 and A[i] = y − 1
16       j = A[i];  x = i;  y = j
17       //update last viewed base paired positions
18       sᵢ = sⱼ = 0 //mark positions i,j by 0 for subsequent deletion
19     else
20       x = i;  y = j //update last viewed base paired positions
21   else //i is not base-paired
22       x = y = 0 //reset positions
23 strip all occurrences of 0 from s
24 return s
```

3.2. **Combinatorics for π-shapes.** In this section we derive the number of π-shapes by first using the Bender-Meir-Moon Theorem 1.1 and then using the Flajolet-Odlyzko Theorem 1.3.

Let a_n [resp. b_n] denote the number of π-shapes [resp. number of π-shapes which can be placed within an external bracket pair $\mathtt{[\ \cdot\ ]}$ and which have n pairs of brackets. It is not difficult to prove that $a_0 = 1 = a_1$, $b_0 = 1$, $b_1 = 0$ and for $n \geq 2$,

$$(11) \qquad a_n = \sum_{i=0}^{n-1-i} a_i \cdot b_{n-1-i}$$

$$(12) \qquad b_n = \sum_{i=1}^{n-1-i} a_i \cdot b_{n-1-i}.$$

By a lot of algebra, we could derive a functional relation of the form $G(x, y) = y$, where $y = \sum_{n=0}^{\infty} a_n x^n$. Since this is tedious and error-prone, we instead use the DSV methodology.

Let $G = (V, \Sigma, R, S)$ be the context-free grammar, where V is the set consisting of S, T, Σ is the set consisting of $\mathtt{[}, \mathtt{]}$, and the rules in R are given as follows

$$(13) \qquad \begin{aligned} S &\rightarrow \ \mathtt{[}T\mathtt{]}S \mid \mathtt{[}T\mathtt{]} \\ T &\rightarrow \ \mathtt{[}T\mathtt{]}S \mid \epsilon \end{aligned}$$

By induction on length, it follows that G is a non-ambiguous grammar for the collection of all *nonempty* π-shapes, and after some algebra, the DSV method yields the equation

$$(14) \qquad S(z) = z^2 S(z)^2 + z^2 S(z) + z^2$$

With the aim of applying the Bender-Meir-Moon Theorem 1.1, we define the function $G(z, w) = z^2 w^2 + z^2 w + z^2$ and would like to obtain that the asymptotic number s_n of π-shapes of length n is

$$(15) \qquad s_n = [z^n]S(z) \sim \sqrt{\frac{rG_z(r, s)}{2\pi G_{w,w}(r, s)}} n^{-3/2} r^{-n}$$

$$(16) \qquad = \sqrt{\frac{3}{2\pi}} \cdot n^{-3/2} \cdot \sqrt{3}^{\,n}$$

However, the hypotheses of the Bender-Meir-Moon Theorem 1.1 are not satisfied, since there are no values of $1 \leq i < j < h$ for which the greatest common divisor of $j - i$ and $h - i$ is 1 and $f_h f_i f_j \neq 0$; indeed, since every π-shape has even length, $s_n = 0$ for n odd. Moreover, as we'll soon see, the value $\sqrt{\frac{3}{2\pi}} \cdot n^{-3/2} \cdot \sqrt{3}^{\,n}$ is off by a factor of 2.

We proceed as follows. Make the variable change $x = z^2$, and define $R(x) = \sum_n r_n x^n$. Since $s_n = 0$ for odd n, we have

$$
\begin{aligned}
R(x) &= \sum_{n=0}^{\infty} r_n x^n = \sum_{n=0}^{\infty} s_{2n} z^{2n} \\
&= \left(\sum_{n=0}^{\infty} s_{2n} z^{2n} \right) + \left(\sum_{n=0}^{\infty} s_{2n+1} z^{2n+1} \right) \\
&= \sum_{n=0}^{\infty} s_n z^n = S(z).
\end{aligned}
$$

Next, it follows from (14) that

(17) $$R(x) = xR(x)^2 + xR(x) + x.$$

Letting w denote $R(x)$, if we define $G(x, w) = xw^2 + xw + x$, then it is straightforward to verify the hypotheses of the Bender-Meir-Moon Theorem 1.1 for the values $r = 1/3$, $s = 1$, which satisfy

$$
\begin{aligned}
G(r, s) &= s \\
G_w(r, s) &= 1.
\end{aligned}
$$

Hence it follows that

$$
\begin{aligned}
[z^{2n}]S(z) &= [x^n]R(x) \sim \sqrt{\frac{rG_z(r, s)}{2\pi G_{w,w}(r, s)}} \cdot n^{-3/2} \cdot r^{-n} \\
&= \sqrt{\frac{3}{4\pi}} \cdot n^{-3/2} \cdot 3^n \\
&= \sqrt{\frac{3}{4\pi}} \cdot \left(\frac{2n}{2} \right)^{-3/2} \cdot \sqrt{3}^{2n} \\
&= \sqrt{\frac{6}{\pi}} \cdot (2n)^{-3/2} \cdot \sqrt{3}^{2n}.
\end{aligned}
$$

Thus $[z^{2n}]S(z) = \sqrt{6/\pi} \cdot (2n)^{-3/2} \cdot \sqrt{3}^{2n}$. Since there are no π-shapes of odd length, $[z^{2n+1}]S(z) = 0$ and it follows that the number of π-shapes is $\sqrt{6/\pi} \cdot n^{-3/2} \cdot \sqrt{3}^n$, provided n is even. This value has been verified by simulation of equations (11) and (12).

Now we derive the same result using DSV methodology and the Flajolet-Odlyzko Theorem 1.3. From (14), we use the quadratic formula to solve for S in $S(z) = z^2 S(z)^2 + z^2 S(z) + z^2$ and obtain

(18) $$S(z) = \frac{1 - z^2 \pm \sqrt{1 - 2z^2 - 3z^4}}{2z^2}.$$

Since $S(x) = \sum_{n=0}^{\infty} s_n z^n$ is the length generating function for π-shapes, obtained by a Taylor expansion of $S(x)$ at $z = 0$, we clearly must choose the solution with a minus sign before the radical, i.e.

$$(19) \qquad S(z) = \frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2}.$$

The dominant singularity will occur where the square root evaluates to 0, or where the denominator is 0. However since a generating function is always analytic at $z = 0$, the dominant singularity must be that root of the polynomial $1 - 2z^2 - 3z^4$ having least modulus. The roots are 0.57735, -0.57735, $\pm i$; however, since $|0.57735| = |-0.57735|$, there does not exist a unique singularity isolated within a circle of convergence about the origin, hence Theorem 1.3 cannot be applied. As before, we make the variable change $x = z^2$, and define $R(x) = \sum_n r_n x^n$. Since $s_n = 0$ for odd n, as before we have

$$R(x) \;\; = \;\; \sum_{n=0}^{\infty} r_n x^n = \sum_{n=0}^{\infty} s_{2n} z^{2n}$$

$$= \;\; \sum_{n=0}^{\infty} s_n z^n = S(z)$$

and

$$(20) \qquad R(x) = \frac{1 - x - \sqrt{1 - 2x - 3x^2}}{2x}.$$

The roots of $P(x) = 1 - 2x - 3x^2$ are $-1, 1/3$, hence the dominant singularity of $R(x)$ is $x = \rho = 1/3$. Factor $(1 - x/\rho)$ out of $P(x)$ to obtain $P(x) = Q(x) \cdot (1 - 3x)$, where $Q(x) = 1 + x$. Define $H(x) = -\frac{\sqrt{1+x}}{2x}$, and let $K = \frac{H(\rho)}{\Gamma(-1/2)} = \frac{\sqrt{3}}{4\pi}$. The hypotheses of Theorem 1.3 are satisfied so we deduce that

$$[x^n]R(x) \sim \frac{K}{\Gamma(-1/2)} \cdot n^{-3/2} \cdot (1/\rho)^n \approx \frac{\sqrt{3}}{4\pi} \cdot n^{-3/2} \cdot 3^n.$$

As before,

$$[z^{2n}]S(z) = [x^n]R(x) = \sqrt{\frac{6}{\pi}} \cdot (2n)^{-3/2} \cdot \sqrt{3}^{2n}$$

and we conclude that the number of π-shapes is $\sqrt{6/\pi} \cdot n^{-3/2} \cdot \sqrt{3}^n$, provided n is even.

It is often non-trivial to verify that the hypotheses necessary for application of the theorem of Meir and Moon (28), as well as of Odlyzko's correction of Meir-Moon given in Theorem 10.13 on page 1162 of (32).

In some cases, like the example in the next subsection, they are not satisfied.

We now compute the number of π-shapes compatible with RNA secondary structures of length n. (Recall that the length of a secondary structure is the number of symbols in its dot-parenthesis Vienna notation.)

π-shapes compatible with RNA sequences of length n. Our main interest is to compute the asymptotic number of π-shapes compatible with secondary structures of length n.

The following grammar non-ambiguously generates all nonempty expressions that begin with an arbitrary number of occurrences of the dummy character $\Box$, followed (essentially) by a nonempty π-shape.[8]

Since the software `RNAshapes` (17; 37; 41), assumes a minimum of $\theta = 3$ unpaired bases in a hairpin loop, we consider the non-ambiguous context-free grammar G_π with the following rules:

$$
\begin{aligned}
S &\rightarrow \Box S \mid A \\
A &\rightarrow A\,\texttt{[}\,B\,\texttt{]} \mid \texttt{[}\,B\,\texttt{]} \\
B &\rightarrow A\,\texttt{[}\,B\,\texttt{]} \mid \bullet^3
\end{aligned}
$$

where $\bullet^3$ abbreviates $\bullet\,\bullet\,\bullet$. Although there are no dots $\bullet$ occurring in π-shapes, our grammar requires $\bullet^3$ to properly count the number of π-shapes compatible with length n secondary structures. Note that the grammar rules correspond to the various cases in the Nussinov-Jacobson algorithm (11; 31).

We claim that the collection $\mathcal{P}$ of π-shapes of nonempty secondary structures of length n is in one-to-one correspondence with the set $\mathcal{R}$ of words of length n generated by the grammar G_π.

To see that $|\mathcal{P}| \leq |\mathcal{R}|$, let $\phi_0 \in \mathcal{P}$ be a π-shape of a nonempty secondary structure of length n. Let ϕ_1 be obtained from ϕ_0 by replacing opposing symbols $\texttt{[}\,\texttt{]}$ (with no intervening symbols between $\texttt{[}$ and $\texttt{]}$) by $\texttt{[}\,\bullet^3\,\texttt{]}$. Let $\phi_2 = \Box^k \phi_1$ be obtained by prefixing $k = n - |\phi_1|$ many occurrences of the symbol $\Box$ to ϕ_1. Clearly ϕ_2 is a length n expression which is generated by the grammar G_π. This correspondence is one-to-one.

To see that $|\mathcal{R}| \leq |\mathcal{P}|$, let $\phi_0 \in \mathcal{R}$. Replacing all occurrences of the symbol $\texttt{[}$ resp. $\texttt{]}$ by $\texttt{(}$ resp. $\texttt{)}$, and replacing occurrences of the symbol $\Box$ by $\bullet$, we obtain a secondary structure S_0 of length n having π-shape ϕ_0. This correspondence is one-to-one. It follows that $|\mathcal{P}| = |\mathcal{R}|$,

---

[8]Essentially, in the sense that opposing symbols $\texttt{[}\,\texttt{]}$ (with no intervening symbols between $\texttt{[}$ and $\texttt{]}$) are replaced by $\texttt{[}\,\bullet^3\,\texttt{]}$, as required by the grammar G_π about to be defined.

hence by using DSV methodology and the Odlyzko-Flajolet theorem to count the number of length n expressions generated by the grammar G_π, we obtain the asymptotic number of π-shapes corresponding to secondary structures of length n.

By DSV, we have the equations

$$
\begin{aligned}
S &= zS + A \\
A &= z^2 AB + z^2 B \\
B &= z^2 AB + z^3
\end{aligned}
$$

For notational simplification in the previous equations, we write S in place of $S(z)$, and similarly for A, B. Such notational simplifications will be tacitly applied without mention. Solving for S using substitution we find that

$$(21) \qquad S(z) = z^2(1-z)^2 S(z)^2 + (z + z^5 - z^6)S(z) + z^5$$

Define the function $G(z, w) = z^2(1-z)^2 w^2 + (z + z^5 - z^6)w + z^5$. The hypotheses of the Bender-Meir-Moon Theorem 1.1 are not satisfied. In particular, for the power series expansion of $G(z, w) = \sum_{m,n \geq 0} g_{m,n} z^m w^n$, it is required that $g_{m,n} \geq 0$, but by taking partial derivatives, we can calculate that $g_{6,1}$ is negative.

Until now we have seen the superior simplicity of the DSV method over algebraic manipulations, in order to obtain a functional relation of the form (21). Now we will see the the usefulness of the Flajolet-Odlyzko Theorem 1.3.

We solve equation (21) using Mathematica to obtain two solutions for S, given by

$$(22) \qquad S_+(z) = \frac{-1 + z^5 + \sqrt{1 - 2z^5 - 4z^7 + z^{10}}}{2(-1+z)z^2}$$

$$(23) \qquad S_-(z) = \frac{-1 + z^5 - \sqrt{1 - 2z^5 - 4z^7 + z^{10}}}{2(-1+z)z^2}.$$

Since $S(z) = \sum_{n=0}^{\infty} s_n z^n$ is a generating function obtained by a Taylor expansion about $z = 0$, as before we must choose the solution $S(z) = S_-(z)$. The function $S(z)$ will be analytic except possibly where the denominator is zero, or where the value inside the square root is zero. The dominant singularity, which determines the exponential growth, is the singularity closest to 0 in the complex plane, and is almost always a real number. In the present case, the dominant singularity, ρ is a solution to the equation $1 - 2z^5 - 4z^7 + z^{10} = 0$, and using Mathematica, we find $\rho \approx 0.756$ from which we immediately

deduce that the exponential growth is $(1/\rho)^n \approx (1.322)^n$. For many applications this is enough, and no deeper analysis is needed.

To obtain more precise asymptotics, we will first ignore the part of the equation without the dominant singularity, since this part grows more slowly as n approaches infinity (see the Appendix for justification of this point). Thus $S(z) = G(z) + H(z)$, where $G(z) = \frac{-1+z^5}{2(-1+z)z^2}$ and

$$(24) \qquad H(z) = -\frac{\sqrt{1 - 2z^5 - 4z^7 + z^{10}}}{2(-1 + z)z^2}.$$

Factor the singularity $\sqrt{1 - z/\rho}$ out of $H(z)$ so that

$$(25) \qquad H(z) = \frac{\sqrt{1 - z/\rho}\sqrt{Q(z)}}{2(-1 + z)z^2}$$

where $Q(z)$ can be gotten by simply dividing polynomials. Since singularity ρ is isolated, we can apply the Flajolet-Odlyzko Theorem 1.3. Make the variable change $x = z/\rho$ and define $J(x) = H(z)$, so that

$$(26) \qquad J(x) = -\frac{\sqrt{Q(\rho x)}}{2(\rho x - 1)\rho^2 x^2}(1 - x)^{1/2}.$$

We now have $J(x)$ in the required form to apply the Flajolet-Odlyzko Theorem 1.3, where the (rescaled) singularity is $x = 1$, and the power of $(1-x)$ is $\alpha = 1/2$. The location of the singularity gives the exponential growth, as mentioned. We pull out the singularity from H and evaluate the rest at ρ, dividing by $\Gamma(-\alpha) = \Gamma(-1/2)$ to get the constant for the asymptotics, given by the following calculations.

$$(27) \qquad K = \frac{\sqrt{Q(\rho)}}{2(\rho - 1)\rho^2} \approx -8.65846$$

$$(28) \qquad s_n \sim \frac{K}{\Gamma(-1/2)} \cdot n^{-3/2} \cdot \left(\frac{1}{\rho}\right)^n$$

$$(29) \qquad s_n \sim 2.44251 \cdot n^{-3/2} \cdot 1.32218^n.$$

This last equation gives the asymptotic number s_n of π-shapes compatible with secondary structures of length n; i.e. π-shapes of secondary structures for an RNA sequence of length n, assuming that every base can basepair with every other base and that there is a minimum of $\theta = 3$ unpaired bases in every hairpin loop.

See the web supplement for full justification of all details concerning application of the Flajolet-Odlyzko Theorem 1.3 to compute the number s_n given in (29) of π-shapes compatible with secondary structures of length n.

3.3. **Combinatorics for π'-shapes.** Let $G = (V, \Sigma, R, S)$ be the context-free grammar, where the set V of nonterminalsconsists of S, T, U, the set Σ of terminals consists of $\textbf{[}, \textbf{]}, \bullet$, and the rules in R are given by the following.

$$
\begin{aligned}
S &\rightarrow U \textbf{[} T \textbf{]} S \mid U \\
T &\rightarrow U \textbf{[} T \textbf{]} U \textbf{[} T \textbf{]} S \mid \bullet \textbf{[} T \textbf{]} \mid \textbf{[} T \textbf{]} \bullet \mid \bullet \textbf{[} T \textbf{]} \bullet \mid \varepsilon \\
U &\rightarrow \bullet \mid \varepsilon
\end{aligned}
$$

By induction on length it can be shown that G is a non-ambiguous grammar for the collection of all π'-shapes, including the empty word ε. In particular T generates all π'-shapes for secondary structures which can appear within an external base pair – i.e. either a hairpin loop, left or right bulge, internal loop or multi-loop. Note the close similarity of the grammar rules with the treatment of various cases in McCaskill's algorithm (27) for the partition function over all secondary structures.

By DSV we obtain the corresponding equations (see web supplement) and solve for them with Mathematica to get the generating series function. The asymptotics are then obtained by again using the same method as in the last section to obtain

(30) $$ S_n \sim 0.985542 \cdot n^{-3/2} \cdot 2.40591^n. $$

Let $G = (V, \Sigma, R, S_0)$ be the context-free grammar, where $V = \{S_0, S, T, U\}$, $\Sigma = \{ \textbf{[}, \textbf{]}, \bullet \}$, and the rules in R are given by the following.

$$
\begin{aligned}
S_0 &\rightarrow \square S_0 \mid S \\
S &\rightarrow U \textbf{[} T \textbf{]} S \mid U \\
T &\rightarrow U \textbf{[} T \textbf{]} U \textbf{[} T \textbf{]} S \mid \bullet \textbf{[} T \textbf{]} \mid \textbf{[} T \textbf{]} \bullet \mid \bullet \textbf{[} T \textbf{]} \bullet \mid \bullet^3 \\
U &\rightarrow \bullet \mid \varepsilon
\end{aligned}
$$

By induction on length it can be shown that G is a non-ambiguous grammar which generates all π'-shapes possibly prefixed by a finite number of occurrences of the dummy variable $\square$, where $\bullet^3$ appears in each hairpin loop. It follows that the number of π'-shapes corresponding to secondary structures of length n is equal to the number of words of length n generated by G. By misuse of terminology, we may at times say that G is a grammar which generates the collection of π'-shapes compatible with secondary structures on n. As before, note that T generates all π'-shapes for secondary structures which can appear within

an external base pair – i.e. either a hairpin loop, left or right bulge, internal loop or multi-loop. By using DSV and Mathematica, we obtain

$$(31) \qquad S_n \sim 1.27613 \cdot 1.80776^n \cdot n^{-3/2}.$$

3.4. Correspondence between π-shapes and Motzkin numbers.

Motzkin words are well-balanced words in the alphabet $\textbf{(}, \textbf{)}, \bullet$, i.e. those for which θ in Definition 2.1 is 0. We denote the set of all Motzkin words by $\mathcal{M}$. Motzkin words are generated by the non-ambiguous context-free grammar having the rules:

$$(32) \qquad M \quad \rightarrow \quad \textbf{(}\, M \,\textbf{)}\, M \mid \bullet M \mid \varepsilon$$

The following theorem establishes a surprising correspondence between Motzkin numbers and π-shapes.

Theorem 3.1. *Let s_n be the number of π-shapes of size n and m_n the number of Motzkin words of size n. Then*

$$(33) \qquad s_{2n+2} = m_n$$

PROOF. A Dyck word is a well-balanced parenthesis expression, with no occurrences of dot $\bullet$. Clearly, π-shapes are exactly those Dyck words not containing *doubly nested* $\textbf{[}\,\textbf{[}\,\mathcal{D}\,\textbf{]}\,\textbf{]}$ patterns, where $\mathcal{D}$ is a Dyck word. The grammar given in (13) at the beginning of Section 3.2 generates the collection of non-empty π-shapes. By a small modification, we obtain the following non-ambiguous grammar $G = (V, \Sigma, R, S_0)$, which generates π-shapes, including the empty shape ε.

$$
\begin{aligned}
S \quad &\rightarrow \quad R \mid \varepsilon \\
R \quad &\rightarrow \quad \textbf{[}\,T\,\textbf{]}\,R \mid \textbf{[}\,T\,\textbf{]} \\
T \quad &\rightarrow \quad \textbf{[}\,T\,\textbf{]}\,R \mid \epsilon
\end{aligned}
$$

This grammar is equivalent to the grammar

$$
\begin{aligned}
S \quad &\rightarrow \quad \textbf{[}\,T\,\textbf{]}\,S \mid \varepsilon \\
T \quad &\rightarrow \quad \textbf{[}\,T\,\textbf{]}\,\textbf{[}\,T\,\textbf{]}\,S \mid \varepsilon
\end{aligned}
$$

where T generates π-shapes which can be placed within an exterior bracket $\textbf{[}\,\cdots\,\textbf{]}$. By DSV methodology, the length-generating function $S(z)$ for π-shapes is the solution of

$$
\begin{aligned}
S(z) \quad &= \quad R(z) + 1 \\
R(z) \quad &= \quad R(z)T(z)z^2 + T(z)z^2 \\
T(z) \quad &= \quad R(z)T(z)z^2 + 1
\end{aligned}
$$

Using Mathematica, we eliminate $R(z), T(z)$ to obtain

$$S_+(z) = \frac{1 - z^2 + \sqrt{1 - 2z^2 - 3z^4}}{2z^2}$$

$$S_-(z) = \frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2}$$

Since $S(z) = \sum_{n=0}^{\infty} s_n z^n$ is a generating function, we have $S(z) = S_-(z).$[9]

$$(34) \qquad S(z) = \frac{1 + z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2}$$

On the other hand the grammar in (32) for the Motzkin words, including the empty word, yields the following equation for the length generating function $M(z)$ for Motzkin words

$$M(z) = z^2 M(z)^2 + zM(z) + 1$$

The solution for this equation is

$$(35) \qquad M(z) = \frac{1 - z - \sqrt{1 - 2z - 3z^2}}{2z^2}$$

The generating functions $S(z)$ for π-shapes and $M(z)$ for Motzkin numbers turn out to be surprisingly similar. More precisely, we have

$$(36) \qquad S(z) = 1 + z^2 M(z^2)$$

After recalling that $S(z) = \sum_{n \geq 0} s_n z^n$ and $M(z) = \sum_{n \geq 0} m_n z^n$, where s_n is the number of π-shapes of size n and m_n the number of Motzkin words of size n, we get

$$\sum_{n \geq 0} s_n z^n = 1 + \sum_{n \geq 0} m_n z^{2n+2}$$

$$\sum_{\substack{n \geq 0 \\ n \text{ even}}} s_n z^n + \sum_{\substack{n \geq 1 \\ n \text{ odd}}} s_n z^n = 1 + \sum_{n \geq 0} m_n z^{2n+2}$$

$$\sum_{n \geq 0} s_{2n} z^{2n} + \sum_{\substack{n \geq 1 \\ n \text{ odd}}} s_n z^n = 1 + \sum_{n \geq 0} m_n z^{2n+2}$$

$$s_0 + \sum_{n \geq 0} s_{2n+2} z^{2n+2} + \sum_{\substack{n \geq 1 \\ n \text{ odd}}} s_n z^n = 1 + \sum_{n \geq 0} m_n z^{2n+2}$$

---

[9]The solution of equation (14) in Section 3.2 is $\frac{1 - z^2 - \sqrt{1 - 2z^2 - 3z^4}}{2z^2}$, which is the generating function of π-shapes without the empty word. Since the current grammar generates the empty word, the right side of equation (34) differs by 1.

Notice that Dyck words (well-balanced parenthesis words) are of even length, so that

$$\sum_{\substack{n \geq 1 \\ n \text{ odd}}} s_n z^n = 0.$$

Thus, for even $n \geq 0$

$$s_{2n+2} = m_n.$$

$\square$

3.5. **Hairpin loops where $\theta > 0$.** From Theorem 3.1 in the previous subsection, it is tempting to conjecture the existence of a similar one-to-one correspondence between secondary structures of length n, assuming that hairpin loops contain at least $\theta > 0$ unpaired bases, and π-shapes, assuming that a minimum number θ of dots $\bullet$ appear in closing brackets $[\ ]$. However, as shown in Figure 2, no such correspondence exists. In Figure 2, we see that the number of π-shapes of
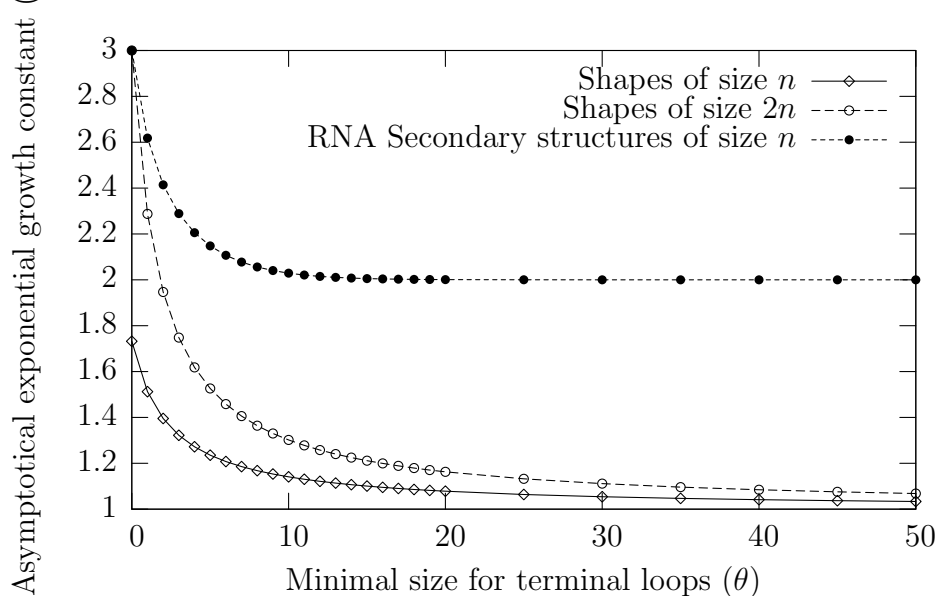
FIGURE 2. Asymptotic exponential growth factors for both π-shapes and Motzkin words/RNA secondary structures for increasing values of θ. These numbers are computed from the generating series for each θ using the function `infsing` of the Maple package `GFun`(35).

size $2n$ with $\theta = 3$ grows more slowly than the number of RNA secondary structures having a minimum of θ' unpaired base inside each hairpin loops, for all values of θ'.

This phenomenon could be explained if RNA secondary structures of length n have significantly fewer hairpin loops than do π-shapes of length $2n$. In such a case, a parameter (in this case θ) that impacts the number of hairpin loops would naturally have a radically different effect on the two combinatorial classes. However, the expected number of hairpin loops turns out to surprisingly similar.

Theorem 3.2 (Expected number of hairpin loops inside π-Shapes and Motzkin words). *Let X_n (resp. Y_{2n+2}) be the random variable which counts the number of hairpin loops in a random, uniformly drawn Motzkin word (resp. π-shape) of length n (resp. $2n+2$). Then the expected number $m_n^t = \mathbb{E}(X_n)$ of hairpin loops (resp. of terminal brackets* [ ] $s_{2n}^t = \mathbb{E}(Y_{2n+2})$*] satisfies*

$$m_n^t \sim \frac{n}{6} + \mathcal{O}(1) \qquad s_{2n+2}^t \sim \frac{2n}{3} + \mathcal{O}(1)$$

*Thus, there are **4 times more** terminal loops inside π-shapes than inside Motzkin words.*

PROOF. Consider the following grammar for the Motzkin words, adapted in order to *mark* each occurrence of a hairpin or terminal loop with a special dummy terminal symbol H, having size 0:

$$
\begin{aligned}
M &\rightarrow \ (\,N\,)\,M \mid \bullet M \mid \varepsilon \\
N &\rightarrow \ (\,N\,)\,M \mid \bullet N \mid H
\end{aligned}
$$

Following the DSV methodology introduced earlier and replacing each occurrence of H by a new variable u, we obtain the equations

$$
\begin{aligned}
M(z,u) &= M(z,u)N(z,u)z^2 + M(u,z)z + 1 \\
N(z,u) &= M(z,u)N(z,u)z^2 + N(u,z)z + u
\end{aligned}
$$

from which we obtain the solution

$$
\begin{aligned}
M(z,u) &= \sum_{\omega \in \mathcal{M}} z^{|\omega|} u^{\tau(\omega)} = \sum_{n \geq 0} \sum_{k \geq 0} m_{n,k} z^n u^k \\
&= \frac{1 - 2z + (2-u)z^2 - \sqrt{1 - 4z + (4-2u)z^2 + 4uz^3 + (u^2 - 4u)z^4}}{2z^2(1-z)}.
\end{aligned}
$$

Here $\tau : \mathcal{M} \to \mathbb{N}$ is the function which counts the number of occurrences of hairpin loops inside a Motzkin word, and $m_{n,k}$ is the number of Motzkin words having size n and k hairpin loops.

We now use the classical observation, found for instance in (13), that the expected number m_n^h of hairpin loops in Motzkin words of length n is closely related to the partial derivative of the multivariate length

generating function. More precisely,

$$
\frac{[z^n]\frac{\partial M(z,u)}{\partial u}(z,1)}{[z^n]M(z,1)} = \frac{[z^n]\left(\sum_{i\geq 0}\sum_{k\geq 0}m_{i,k}z^i k u^{k-1}\right)(z,1)}{m_n}
$$

$$
= \frac{\sum_{k\geq 0}m_{n,k}k}{m_n} = \sum_{k\geq 0}k\mathbb{P}(X_n = k)
$$

$$
= \mathbb{E}(X_n) = m_n^h
$$

Here, $\mathbb{P}(X_n = k) = \frac{m_{n,k}}{m_n}$ is the (uniform) probability that a Motzkin word of length has exactly k hairpin loops. Then we apply the asymptotic techniques extensively described throughout this article to $\frac{\partial M(z,u)}{\partial u}(z,1)$ and $M(z,1)$, and obtain

$$
[z^n]\frac{\partial M(z,u)}{\partial u}(z,1) \sim \frac{\sqrt{3}}{4\sqrt{\pi}}\frac{3^n}{\sqrt{n}} + \mathcal{O}(\frac{1}{n\sqrt{n}})
$$

$$
[z^n]M(z,1) \sim \frac{3\sqrt{3}}{2\sqrt{\pi}}\frac{3^n}{n\sqrt{n}} + \mathcal{O}(\frac{1}{n^2\sqrt{n}})
$$

from which the ratio $\frac{[z^n]\frac{\partial M(z,u)}{\partial u}(z,1)}{[z^n]M(z,1)}$ yields the claimed result.

This proof also holds for the π-shapes, using the grammar

$$
\begin{aligned}
S &\rightarrow \ [\,T\,]\,S \mid \varepsilon \\
T &\rightarrow \ [\,T\,]\,[\,T\,]\,S \mid H
\end{aligned}
$$

where H is a length 0 dummy symbol to mark hairpin loops. This yields the generating function

$$
S(z,u) = \frac{1 + (2-u)z^2 - \sqrt{1 - 2uz^2 - (4u - u^2)z^4}}{2z^2}
$$

and, using the DSV technique coupled with singularity analysis (15),

$$
[z^{2n+2}]\frac{\partial S(z,u)}{\partial u}(z,1) \sim \frac{\sqrt{3}}{\sqrt{\pi}}\frac{3^n}{\sqrt{n}} + \mathcal{O}(\frac{1}{n\sqrt{n}})
$$

$\square$

4. π-SHAPES WITH k STEMS

In this section, we apply the DSV method and the Flajolet-Odlyzko Theorem 1.3 in order to compute the number of π-shapes having k-stems, i.e. k pairs of brackets. Unlike other sections, the material makes use of more advanced singularity analysis techniques from (15).

π-**expansion.** Let $\mathbb{S}$ denote the set of secondary structures in Vienna notation, and let $\mathbb{P}$ denote the set of all π-shapes. First we consider the total number of secondary structures of size n compatible with a given shape π. We claim that the set of RNA structures compatible with a given shape π can be exhaustively built from π by the means of an operation called π-expansion, consisting in two consecutive transformations:

(1) **Helix expansion:** Replace each opening left bracket **[** resp. its corresponding right closing bracket **]**, by k open parentheses **(**^{k}, resp. right parentheses **)**^{k}, for $k \geq 1$.

(2) **Unpaired base insertion:** Insert any number of unpaired bases (symbol $\bullet$) at any position in the structure resulting from the previous operation, except among occurrences of the motif **( )** where at least θ must be added.

We claim that this transformation is non-ambiguous, meaning that there is at most one way to obtain a given structure r from a given shape π by applying the above two transformations.

Let us properly define these concepts and notions, starting with a factorization of RNA secondary structures into shapes, introduced already in Algorithm 3.1.

Definition 4.1 (π-factorization). *Define the factorization function* $\phi :$ $\mathbb{S} \to \mathbb{P}$, *mapping RNA secondary structures into π-shapes, given by* $\phi = \phi_2 \circ \phi_1$, *where*

$$\phi_1(\alpha\omega) \;=\; \begin{cases} \phi_1(\omega) & \text{if } \alpha = \bullet \\ \alpha\phi_1(\omega) & \text{if } \alpha \in \{\, \textbf{(} \,,\, \textbf{)} \,\} \end{cases}$$

$$\phi_1(\varepsilon) \;=\; \varepsilon$$

where $\omega \in \{\, \bullet \,,\, \textbf{(} \,,\, \textbf{)} \,\}^*$ *is a suffix of an RNA secondary structure,* $\alpha \in \{\, \bullet \,,\, \textbf{(} \,,\, \textbf{)} \,\}$ *and*

$$\phi_2(\, \textbf{(}^k\omega\,\textbf{)}^k\omega') \;=\; \textbf{[}\,\phi_2(\omega)\,\textbf{]}\,\phi_2(\omega')$$

$$\phi_2(\varepsilon) \;=\; \varepsilon$$

for $k \geq 0$, $\omega, \omega' \in \mathbb{S}$ *and* ω *not of the form* $\textbf{(}\omega''\textbf{)}$ *with* $\omega'' \in \mathbb{S}$. *Alternatively, this means that k is maximal such that* $\omega \in \mathbb{S}$.

Definition 4.2 (π-expansion). *Let $\mathcal{P}(\mathbb{S})$ be the set of all subsets of $\mathbb{S}$, also called the power set of $\mathbb{S}$. Define π-expansion to be the function $\psi : \mathbb{P} \to \mathcal{P}(\mathbb{S})$, given by $\psi = \psi_2 \circ \psi_1$, where*

$$\psi_1(\, [\, \omega\,]\, \omega') \;=\; \bigcup_{k \geq 1} \{\, (^k\} . \psi_1(\omega) . \{\,)^k\} . \psi_1(\omega')$$

$$\psi_1(\varepsilon) \;=\; \varepsilon$$

with $\omega, \omega' \in \mathbb{P}$ and

$$\psi_2(\alpha\alpha'\omega) \;=\; \begin{cases} \{\bullet^* (\, \bullet^\theta\} . \psi_2(\alpha'\omega) & \text{if } \alpha\alpha' = (\,) \\ \{\bullet^*\alpha\} . \psi_2(\alpha'\omega) & \text{otherwise} \end{cases}$$

$$\psi_2(\alpha) \;=\; \{\bullet^*\alpha\bullet^*\}$$

$$\psi_2(\varepsilon) \;=\; \{\bullet^*\}$$

where $\alpha \in \{\, (\, ,\,)\, \}$, $\alpha \in \alpha \in \{\, (\, ,\,)\, \}$ and $\bullet^*$ is a macro for the union language of any number of dots $\bullet$ corresponding to unpaired bases.

Note that the functions ψ_1 and ψ_2 correspond to the transformations (1) and (2) introduced above.

Proposition 4.3. *For all $\pi \in \mathbb{P}$, the π-expansion of π is exactly the set of all secondary structures of RNA that factor into π, i.e. all secondary structures having shape π, or more formally*

$$\psi(\pi) = \{r \in \mathbb{S} \mid \phi(r) = \pi\}, \quad \text{for all } \pi \in \mathbb{P}$$

Moreover, the construction ψ is non-ambiguous.

PROOF. For any $\pi \in \mathbb{P}$, let:
- $A_\pi \subset \mathbb{S}$ be the set of RNA structures ω such that $\phi(\omega) = \pi$
- $B_\pi = \phi_1(A_\pi)$ be the set of Dyck words ω such that $\phi_2(\omega) = \pi$
- $C_\pi = \psi_1(\pi)$
- $D_\pi = \psi_2(C_\pi) = \psi(\pi)$.

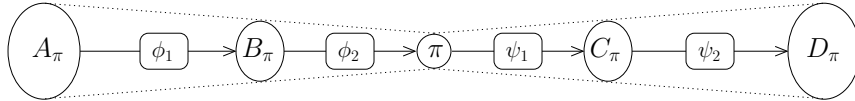Then proving the Proposition 4.3 is equivalent to proving that $A_\pi = D_\pi$.

FIGURE 3. Sketch of the proof goes as follows. We first identify B_π with C_π and then prove that the inverse image of any $\omega \in B_\pi = C_\pi$ under ϕ_1 is the same as the image of ω under ψ_2. Then it follows directly that $A_\pi = D_\pi$.

- $B_\pi = C_\pi$:

  This equality can be proved by induction on the size $|\pi|$ of the shape π. For $|\pi| = 0$, the only candidate for π is ε, therefore $B_\pi = C_\pi = \{\varepsilon\}$. Assume now that, $\forall \pi \in \mathbb{P}$ such that $|\pi| \le n$, $B_\pi = C_\pi$ holds. Then for π such that $|\pi| = k+1$, we know that $\pi = \mathbf{[}\,\pi'\,\mathbf{]}\,\pi''$ ($|\pi| > 0$). Recall that

  $$B_\pi = \{w \in \mathcal{D} \mid \phi_2(w) = \pi\}.$$

  Since $|w| > 0$ ($|w| \ge |\pi|$), w can be decomposed in $w = \mathbf{(}^k w'\mathbf{)}^k w''$, for given $w', w'' \in \mathcal{D}$ such that $w' \ne \mathbf{(}\,v\,\mathbf{)}$, $v \in \mathcal{D}$ (k is *maximal*). Thus we get the following equivalent definition for B_π:

  $$
  \begin{aligned}
  B_\pi &= \{\mathbf{(}^k w'\mathbf{)}^k w'' \in \mathcal{D} \mid w' \ne \mathbf{(}\,v\,\mathbf{)} \text{ and } \phi_2(\mathbf{(}^k w'\mathbf{)}^k w'') = \mathbf{[}\,\pi'\,\mathbf{]}\,\pi''\} \\
  &= \{\mathbf{(}^k w'\mathbf{)}^k w'' \in \mathcal{D} \mid w' \ne \mathbf{(}\,v\,\mathbf{)}, w' \in B_{\pi'} \text{ and } w'' \in B_{\pi''}\} \\
  &= \{\mathbf{(}^k w'\mathbf{)}^k w'' \in \mathcal{D} \mid w' \in B_{\pi'} \text{ and } w'' \in B_{\pi''}\}.
  \end{aligned}
  $$

  Let us now focus on B_π, which can be defined as:

  $$
  \begin{aligned}
  C_\pi &= \{v \in \mathcal{D} \mid v \in \psi_1(\pi)\} \\
  &= \{\mathbf{(}^{k'} v'\mathbf{)}^{k'} v'' \in \mathcal{D} \mid v' \in C_{\pi'} \text{ and } v'' \in C_{\pi''}\}.
  \end{aligned}
  $$

  After noting that $|\pi'| < |\pi|$ (resp. $|\pi''| < |\pi|$), we apply the induction hypothesis yields $(w' \in B_{\pi'}) \Leftrightarrow (w' \in C_{\pi'})$ and $(w'' \in B_{\pi''}) \Leftrightarrow (w'' \in C_{\pi''})$. This establishes the equality $B_\pi = C_\pi$.

- $A_\pi = D_\pi$:

  We will focus first on A'_ω, the inverse image of $\omega \in \mathcal{D}$ under ϕ_1, and on D'_ω, the image of $\omega \in \mathcal{D}$ under ψ_2. As ϕ_1 simply deletes each occurrence of an unpaired base $\bullet$, its inverse should consist of inserting any number of dot symbols $\bullet$ before or after any symbol in the shape expression. However, such a construction would also yield words over $\{\,\mathbf{(},\mathbf{)},\bullet\,\}^*$ that are not secondary structures, due to the constraint that there are at least θ unpaired bases symbolized by dots $\bullet$ in each hairpin (or terminal) loop. Therefore, a minimum number of at least θ dots $\bullet$ must occur within the $\mathbf{(}\,\mathbf{)}$ motif. The resulting construction is then exactly that of ψ_2, thus $A'_\omega = D'_\omega$. As

  $$A_\pi = \bigcup_{\omega \in B_\pi} A'_\omega \ , \quad D_\pi = \bigcup_{\omega \in C_\pi} D'_\omega \text{ and } B_\pi = C_\pi,$$

  then $A_\pi = D_\pi$.

Concerning the non-ambiguity of the construction, we first point out that in the definitions of ψ_1 and ψ_2, at most one rule can be applied at any time, and the unions involved in the definitions of the right-hand sides are obviously disjoint. The only potentially pathological case

would then consist of two shapes π and π', mapping to two distinct sets S and S' under ψ_1, and then mapping to a unique set T under ψ_2. Since $D_\pi = A_\pi$, the image of T under ϕ is a singleton, which makes such a case impossible to arise. $\qquad\square$

Theorem 4.4. *Let $\pi \in \mathbb{P}$ be a shape, having m base pairs and h occurrences of the motif* **[ ]**. *Let $\mathcal{L}(\pi) := \psi(\pi)$ be the language associated with π through the π-expansion. Then the length generating function $L_\pi(z) := \sum_{\omega \in \mathcal{L}(\pi)} z^{|\omega|}$ of $\mathcal{L}(\pi)$ is given by*

$$(37) \qquad L_\pi(z) = \frac{z^{\theta h}}{1 - z} \frac{z^{2m}}{(1 - 2z)^m}$$

Furthermore, the number s_n^π of RNA secondary structures that map under ϕ to a given π-shape π is asymptotically given by

$$(38) \qquad s_n^\pi = [z^n]L_\pi(z) \sim \frac{1}{2^{\theta h + 2m - 1}} \frac{2^n n^{m-1}}{(m-1)!}(1 + \mathcal{O}(1/n))$$

PROOF. Let $\mathbf{k} = \{k_1, \ldots, k_m\}$, $k_i \geq 1$ be the indices assigned by ψ_1 to the parentheses in a *left-to-right* fashion, and let $\psi_1^{\mathbf{k}}(\pi) \in \mathcal{D}$ be the Dyck word obtained from π under ψ_1 using values from $\mathbf{k}$ during the expansion of helices. The length generating function for the language $\psi_1(\pi)$ is then $\psi_{1\pi}(z)$ such that

$$\psi_{1\pi}(z) = \sum_{\omega \in \psi_1(\pi)} z^{|\omega|} = \sum_{\substack{\mathbf{k} \\ k_i \geq 1}} z^{|\psi_1^{\mathbf{k}}(\pi)|} = \sum_{n \geq 0} \sum_{\substack{\mathbf{k} \\ k_i \geq 1 \\ |\mathbf{k}| = n}} z^n = \frac{z^{2m}}{(1 - z^2)^m}$$

for $|\mathbf{k}| = \sum_{i=1}^m k_i$. The last part of the previous equation arises from the enumeration of the partitions of n into m non-empty parts. It can also be derived directly from the fact that generating functions are commutative images of languages, which means that it is possible to *remove the order* in a sequence. Let E_π be the *reconciliation* language built by reordering the words of $\psi_1(\pi)$ such that each opening parenthesis is immediately followed by its corresponding closing one. Thus, the languages $\psi_1(\pi)$ and E_π share the same generating function. Namely:

$$E_\pi = \bigcup_{\substack{\mathbf{k} \\ k_i \geq 1}} \{(\,\textbf{( )}\,)^{k_1} \ldots (\,\textbf{( )}\,)^{k_m}\} = \mathcal{L}(\underbrace{(\,\textbf{( )}\,)^+ \ldots (\,\textbf{( )}\,)^+}_{m \text{ times}})$$

It is well known that the generating function of the language having regular expression $(\,\textbf{( )}\,)^+$ is $\frac{z^2}{1-z^2}$, so we get the result[10].

---

[10]The language denoted by the previous regular expression is ambiguous. However, the multiplicity of a word generated from it exactly equals the number of words from $\psi_1(\pi)$, so that the generating functions are the same for $\psi_{1\pi}$ and E_π.

The transform ψ_2 applied to a Dyck word ω appends any number of dots $\bullet$ occurring at the end of ω as well as before each symbol, while ensuring a minimal number θ of dots $\bullet$ in each hairpin loop. Since the variable z in $\psi_{1\pi}(z)$ is the image of a dot $\bullet$ in $\psi_1(\pi)$, this substitution resp. concatenation transformation on the language amounts to a composition resp. product of the generating functions, according to DSV methodology. Recall that the language $\{\bullet\}^*$ of *any number of unpaired bases* has generating function $\frac{1}{1-z}$ and that the θ unpaired bases in each of the h hairpin loops can be gathered (commutativity). Thus obtaining a factor $z^{\theta h}$ in the generating function, we get:

$$
\begin{aligned}
\psi_\pi(z) &= \frac{z^{\theta h}}{1-z}\psi_{1\pi}\left(\frac{z}{1-z}\right) \\
&= \frac{z^{\theta h}}{1-z}\frac{z^{2m}}{(1-z)^{2m}}\frac{1}{\left(1-\left(\frac{z}{1-z}\right)^2\right)^m} \\
&= \frac{z^{\theta h}}{1-z}\frac{z^{2m}}{(1-z)^{2m}}\frac{(1-z)^{2m}}{((1-z)^2-z^2)^m} \\
&= \frac{z^{\theta h}}{1-z}\frac{z^{2m}}{(1-2z)^m} = L_\pi(z)
\end{aligned}
$$

Using singularity analysis techniques extensively described in (15), it is then possible to extract the asymptotic behavior of $s_n^\pi = [z^n]L_\pi(z)$, the number of secondary structures of size n associated with a given shape π.

The dominant singularity ρ is the pole of $\frac{1}{(1-2z)^m}$, thus $\rho = 1/2$. Observing that $[z^n]L_\pi(z) = \rho^{-n}[z^n]L_\pi(z\rho)$, we focus on the function $f(z) = L_\pi(z/2)$, namely

$$
f(z) = \frac{z^{2m}z^{\theta h}}{2^{\theta h+2m}(1-z/2)}\frac{1}{(1-z)^m}
$$

whose dominant singularity is now at $z = 1$. By defining

$$
g(z) = \frac{1}{2^{\theta h+2m-1}}\frac{1}{(1-z)^m}
$$

it follows that $f(z) \sim g(z)$. The function $g(z)$ is of the *basic-scale* type defined in (15), and thus admits an asymptotical expansion of the form

$$
[z^n]g(z) \sim \frac{1}{2^{\theta h+2m-1}}\frac{n^{m-1}}{\Gamma(m)} = \frac{1}{2^{\theta h+2m-1}}\frac{n^{m-1}}{(m-1)!}(1+\mathcal{O}(1/n))
$$

Since the generating function is of rational type, it meets the analyticity condition of (15), so that we can *transfer* the asymptotic behavior of

the coefficient of $g(z)$ into the behavior of $[z^n]f(z)$. The results follows, after recalling that $[z^n]L_\pi(z) = 2^n[z^n]f(z)$. $\qquad\square$

5. Discussion

In this paper, we determine the asymptotic number of π- and π'-shapes, as well as the number of shapes compatible with an RNA secondary structure of length n. We describe the DSV method which allows very simple determination of the function $S(z)$ whose power series $\sum_{n \geq 0} s_n z^n$ has the property that s_n is the number of combinatorial objects (secondary structures, π-shapes, π'-shapes, etc.) of length n. The DSV method begins with a non-ambiguous context-free grammar that generates all combinatorial objects, regardless of length, and applies a simple transform to obtain an implicit equation for $S(z)$, where $S(z) = \sum_{n \geq 0} s_n z^n$ is the length generating function for the combinatorial objects being counted. This implicit equation immediately gives rise to the functional equation $G(z, w) = w$, used in the Bender-Meir-Moon Theorem 1.1. Alternatively, this implicit equation can be solved to give an equation $S(z) = f(z)/g(z)$, and dominant singularity analysis can be carried out using the Flajolet-Odlyzko Theorem 1.3. Since the hypotheses for the Bender-Meir-Moon Theorem 1.1 do not hold in certain cases, and may be very difficult to verify in other cases, the approach using DSV and Flajolet-Odlyzko can be quite useful. Basically, one first determines the dominant singularity $z = \rho$, then performs a change of variable $x = z/\rho$, in order to rescale the dominant singularity to $x = 1$. In this form, the Flajolet-Odlyzko Theorem 1.3 can be applied to deduce the asymptotic value $s_n \sim K/\Gamma(-\alpha)\rho^{-n}n^{-3/2}$. The combination of DSV and Flajolet-Odlyzko is not well-known in the bioinformatics community, although there are some notable exceptions such as Nebel (30).

Table 5 of (41) presents heuristic approximations on the number of shapes for secondary structures of a given RNA sequence of length n. For π-shapes, the number obtained by repeated simulations as stated in (41) is 1.1^n, while for π'-shapes, the number is 1.16^n. Originally, our motivation in this paper was to give a rigorous asymptotic limit for the expected number of π- and π'-shapes compatible with secondary structures for random RNA sequences of length n, where the sequences are generated by a zero-order Markov process assuming a given composition frequency for each nucleotide. Such a value could then be compared directly with the experimentally obtained values of 1.1^n and 1.16^n. Unfortunately, we are not currently able to compute this expected value; however, in Sections 3.2 and 3.3, we compute the asymptotic number of π- and π'-shapes compatible with secondary structures for an RNA sequence of length n, under the assumption that any base can basepair with any other base. Those results are summarized in

Object counted	Asymptotic number a_n
num of sec str on n	$1.104366 \cdot 2.618034^n / n^{3/2}$
num of π-shapes of size n	$1.38198 \cdot 1.732051^n / n^{3/2}$
num of π-shapes compatible with sec str on n	$2.44251 \cdot 1.32218^n / n^{3/2}$
num of π'-shapes of size n	$0.985542 \cdot 2.40591^n / n^{3/2}$
num of π'-shapes compatible with sec str on n	$1.27613 \cdot 1.80776^n / n^{3/2}$

TABLE 2. Summary of asymptotic results concerning π- and π'-shapes. Asymptotic number of secondary structures is given in the first line for purpose of comparison. Asymptotic value s_n in the second line is for n even, since there are no π-shapes when n is odd. Asymptotic values in the third and fifth line assume a minimum of $\theta = 3$ unpaired bases in hairpin loops.

Table 2. Additionally, in Theorem 3.1 we establish an interesting one-one correspondence between π-shapes and Motzking numbers. Finally, performing a finer analysis, in Theorem 4.4 of Section 4, we give the asymptotic number of RNA secondary structures having any fixed, given π-shape π. This result may lead to a rigorous asymptotic limit for the expected number of π-shapes compatible with secondary structures for random RNA sequences of length n, where the sequences are generated by a zero-order Markov process assuming a given composition frequency for each nucleotide.

6. ACKNOWLEDGEMENTS

Appendix

In this section, we give a self-contained justification of the application of the theorem of Flajolet and Odlyzko (16) to obtain the asymptotic number of π-shapes compatible with secondary structures on n. Recall from Section 3.2 that if $S(z) = \sum_{n=0}^{\infty} s_n z^n$ is the generating function for the number of secondary structures on a sequence of length n, with minimum hairpin length 1, it is given by

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

We begin by discussing why the exponential growth rate of s_n is determined by the dominant singularity.

A1. **Determining the exponential growth factor.** The definition of a function f being analytic at a point z_0 is that the complex derivative of f is defined at z_0. Note that while the function $\sqrt{z}$ is defined at z=0, it is not analytic at z=0. The derivative of $\sqrt{z} = z^{1/2}$ is $\frac{1}{2} z^{-1/2}$. As is suggested by this, the derivative does not exist at zero. Thus, the function $\sqrt{z}$ is analytic everywhere except 0.

Similarly the function $\sqrt{1 - 2z - z^2 - 2z^3 + z^4}$ is not analytic exactly at the zeros of the polynomial $1 - 2z - z^2 - 2z^3 + z^4$. And the function, $S(z)$ is analytic everywhere except the zeros of the polynomial inside the square root, and possibly where the denominator equals 0. (In this case it is actually analytic at $z = 0$.)

It is known from introductory complex analysis that a power series converges in a circular region about the point of expansion out to the nearest non-analytic point, or singularity. In addition, if the singularity is not trivial[11] the power series always diverges outside of this circle. (See the chapter on power series in Churchill's Complex Variables and Applications (9) for a good and quick introduction.)

This fact gives an immediate answer for the exponential growth of the power series terms of a given function. In the case of generating series, we are expanding about the point $z = 0$. For a generating series with positive coefficients, it can be shown, using Pringsheim's theorem (23), that the singularity closest to the origin always occurs on the positive real axis at some value ρ. Then, we know that the power series converges for the circular region $|z| < \rho$, and so the exponential

---

[11] All singularities we deal with will be what we call non-trivial. A function f analytic inside a circle C has a non-trivial singularity at z_0 on C if either f or its derivative of some order has no limit as z tends to z_0 in C. An example of a trivial singularity is the singularity of the function $f(z) = e^z(z-1)/(z-1)$ at $z = 1$.

growth of the terms f_n cannot be greater than $(1/\rho)^n$. Otherwise, if the terms grow faster than this, it is clear that the series

$$f(z) \;=\; \sum_{n=0}^{\infty} f_n z^n$$

cannot converge near $z = \rho$ as the terms aren't going to zero. Similarly, since the power series diverges for any z such that $|z| > \rho$, the exponential growth rate of the terms cannot be less than $(1/\rho)^n$. Otherwise it is straightforward to show the series will converge for real $z > \rho$. Thus we immediately get that for generating functions the exponential rate of growth of terms is exactly $(1/\rho)$.

The singularity closest to the origin is called the dominant singularity. For our function S, the dominant singularity is at $\rho = \frac{1}{2}(3 - \sqrt{5}) \approx 0.381966$, one of the roots of the polynomial $1 - 2z - z^2 - 2z^3 + z^4$, which is inside the square root in $S(z)$. We get immediately that for large n, S scales as

$$S_n \approx (1/\rho)^n \approx (1/0.381966)^n \approx (2.61803)^n.$$

So, the above gives the exponential growth. In many cases, this is all that is desired. However, we still could be off by non-exponential growth factors. Thus, for example, if $\rho = 1$, all we know is that there is no exponential growth or decay. Within these bounds, anything, for example polynomial growth, is possible.

A2. **Finer asymptotics.** To get the asymptotics more exactly is not hard either, that is, using the results from the paper by Flajolet and Odlyzko (16).

To use these results, we have to verify that the generating series is analytic in the region $\triangle$ shown in Figure A1, except at the point ρ, thus analytic in $\triangle \backslash \rho$, where for the shape $\triangle$ we can choose any ε and $0 < \phi < \pi/2$. The region $\triangle$ is the solid circle about the origin with radius $\rho + \varepsilon$, with a symmetric wedge cut out of it, centered about the real axis, to the point ρ.

Since our singularities are isolated (this will always be true if you have only finitely many singularities), and our dominant singularity is unique, (that is, we do not have more than one singularity the same minimal distance from the origin) we can choose ε to make our function analytic in $\triangle \backslash \rho$. Simply note that $\triangle$ is a subset of the solid circle of radius $\rho + \varepsilon$ about the origin. Thus, if all of our singularities have larger magnitude than ρ, they will have larger magnitude than $\rho + \varepsilon$ for some ε, and will not be in $\triangle$.

Note that this method can be applied in any case in which the singularities are isolated and the dominant singularity is unique. There are usually ways to work around cases where the dominant singularity is not unique. (We saw an example in Section 3.2.)
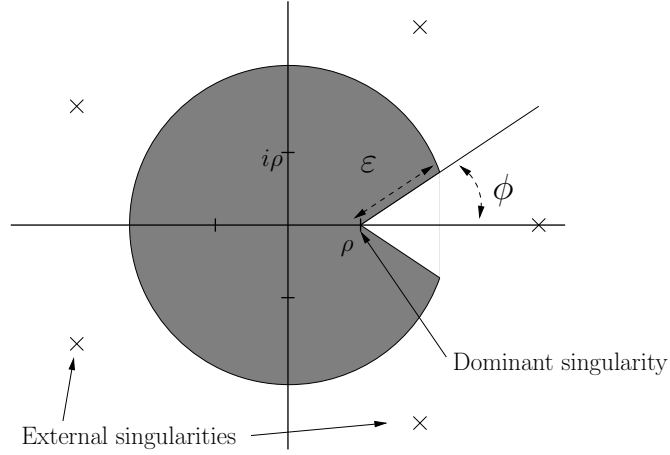

FIGURE A1. The shaded region $\triangle$ where, except at $z = \rho$, the generating function $S(z)$ must be analytic

First some setup. We have our function

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

Call the polynomial under the square root $P(z)$. Since $z = \rho$ is a root of $P(z)$ we can pull out the factor $\sqrt{1 - z/\rho}$ (using Mathematica or Maple) to get

$$\sqrt{P(z)} = \sqrt{1 - z/\rho}\sqrt{Q(z)}$$

where now $\sqrt{Q(z)}$ will be analytic for all z such that $|z| < \rho + \varepsilon$ for some ε, so that for where we're interested in, $\sqrt{Q(z)}$ is always analytic. Split S into 2 parts.

$$
\begin{aligned}
S(z) &= \frac{1 - z + z^2}{2z^2} - \frac{\sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2} \\
g(z) &= \frac{1 - z + z^2}{2z^2} \\
h(z) &= -\frac{\sqrt{1 - z/\rho}\sqrt{Q(z)}}{2z^2} \\
S(z) &= g(z) + h(z)
\end{aligned}
$$

If we don't worry about being rigorous, we can do some quick calculations to pull out the asymptotics. To go straight to these calculations, skip the next section.

A3. **A detailed analysis.** To apply the results of the paper by Flajolet and Odlyzko(16), we will need to rescale the relevant part of the function so that the dominant singularity is at 1 instead of at ρ.

Let

$$
\begin{aligned}
G(z) = z^2 g(z) &= \frac{1}{2}(1 - z + z^2) \\
H(z) = z^2 h(z) &= -\frac{1}{2}\sqrt{1 - z/\rho}\sqrt{Q(z)}
\end{aligned}
$$

That way, G(z) and H(z) are both defined, and analytic, at 0 and we can talk about their power series expansion about 0. Recall that Cauchy's formula is

$$
f_n = [z^n]f(z) = \frac{1}{2\pi i} \oint_{O^+} \frac{f(z)}{z^{n+1}} dz,
$$

where O^+ is any positively oriented contour in $\triangle$ (in an analytic region) that encloses the origin. In their proof, Flajolet and Odlyzko use a special contour, but we don't have to worry about that.

Then,

$$
\begin{aligned}
s_n &= \frac{1}{2\pi i} \oint \frac{S(z)}{z^{n+1}} dz \\
&= \frac{1}{2\pi i} \oint \frac{g(z)}{z^{n+1}} dz + \oint \frac{h(z)}{z^{n+1}} dz \\
&= \frac{1}{2\pi i} \oint \frac{G(z)}{z^{n+3}} dz + \oint \frac{H(z)}{z^{n+3}} dz \\
s_n &= G_{n+2} + H_{n+2}
\end{aligned}
$$

We figure out the asymptotics of G and H.

It is clear in this example that G_n is 0 for any large n (for any n larger than 2). But note that even if this were not the case, more generally we know that $G(z)$ will grow exponentially like $1/|\rho'|$, where ρ' is the location closest to the origin that the function $G(z)$ is not analytic (may be complex). Since $|\rho'| > \rho$, as ρ is our dominant singularity, this exponential growth rate will be slower than the growth rate of $H(z)$, so we can ignore it.

For $H(z)$, rescale so that the singularity occurs at $w = 1$ instead of $z = \rho$. To do this, simply substitute $z = \rho w$. We get

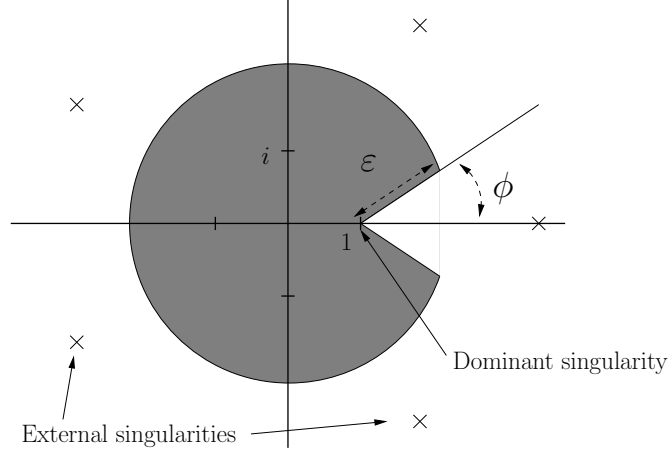$$H(w) \quad = - \quad \frac{1}{2}\sqrt{1-w}\sqrt{Q(\rho w)}$$

FIGURE A2. The rescaled region $\triangle$, where the dominant singularity of $H(z)$ has been moved out from ρ to 1

The function $H(w)$ has a singularity at $w = 1$, and is analytic in the required region, $\triangle\backslash 1$, where the rescaled region $\triangle$ is shown in figure A2. Note that external singularities that remain will scale to still be outside of the region $\triangle$. We now apply the following theorem (stated as Corollary 2, part (i) of (16) on page 224) which states

Theorem. *Assume that f(z) is analytic in $\triangle\backslash 1$, and that as $z \to 1$ in $\triangle$,*

$$f(z) \sim K(1-z)^\alpha$$

Then, as $n \to \infty$, if $\alpha \notin 0, 1, 2, ...$,

$$f_n \sim \frac{K}{\Gamma(-\alpha)}n^{-\alpha-1}.$$

We take $\alpha = +1/2$. Note that

$$f(z) \sim g(z)$$

as $z \to z_0$ means

$$\lim_{z \to z_0} \frac{f(z)}{g(z)} = 1$$

For our $H(w)$, we find

$$\frac{H(w)}{(1-w)^{1/2}} = -\frac{1}{2}\sqrt{Q(\rho w)}$$

so that

$$\lim_{w \to 1} \frac{H(w)}{(1-w)^{1/2}} = -\frac{1}{2}\sqrt{Q(\rho)} = K'$$

$$\lim_{w \to 1} \frac{H(w)}{K'(1-w)^{1/2}} = 1$$

This can be rewritten

$$H(w) \sim K'(1-w)^{1/2}$$

By the above theorem, we get

$$[w^n]H(w) \sim \frac{K'}{\Gamma(-1/2)} n^{-3/2}$$

Now we scale back. Note that

$$H(w) = \sum H_n^w w^n$$

where in the term $H_n^w = [w^n]H(w)$, the superscript w reminds us that these are the coefficients *when* we expand the function in terms of the variable w.

$$
\begin{aligned}
H(w) &= \sum H_n^w w^n \\
&= \sum H_n^w \frac{z^n}{\rho^n} \\
&= \sum \frac{H_n^w}{\rho^n} z^n
\end{aligned}
$$

Therefore, the $H_n = [z^n]H(z)$, the power series coefficients of H in terms of z, are given by,

$$H_n = \frac{H_n^w}{\rho^n}$$

so that

$$
\begin{aligned}
H_n &\sim \frac{H_n^w}{\rho^n} \\
H_n &\sim \frac{K'}{\Gamma(-1/2)}(\frac{1}{\rho})^n n^{-3/2}
\end{aligned}
$$

Remember that for large n, the G_n goes away so that

$$s_n = H_{n+2} \sim \frac{K'}{\Gamma(-1/2)}(\frac{1}{\rho})^{n+2}(n+2)^{-3/2}$$

And then note that

$$\lim_{n\to\infty} \frac{(n+2)^{-3/2}}{n^{3/2}} = \lim_{n\to\infty} (\frac{n+2}{n})^{3/2} = 1$$

so that

$$(n+2)^{-3/2} \sim n^{-3/2}$$

which means we can simplify to

$$s_n \sim \frac{K'}{\rho^2 \Gamma(-1/2)} (\frac{1}{\rho})^n n^{-3/2}$$

or letting $K = K'/\rho^2$,

$$s_n \sim \frac{K}{\Gamma(-1/2)} (\frac{1}{\rho})^n n^{-3/2}$$

Plugging in values ($\rho \approx 0.381966$) gives

$$s_n \sim 1.10437(2.61803)^n n^{-3/2}$$

A4. **The short way.** Now that we can see how the theorem applies, how rescaling works, and that splitting the generating function into parts that are not analytic at 0 does not cause problems, we can see that if we start with

$$\begin{aligned}
g(z) &= \frac{1 - z + z^2}{2z^2} \\
h(z) &= -\frac{\sqrt{1 - z/\rho}\sqrt{Q(z)}}{2z^2} \\
S(z) &= g(z) + h(z)
\end{aligned}$$

we can ignore $g(z)$ as it doesn't have the dominant singularity. Then we simply get K by taking out the $\sqrt{1 - z/\rho}$ term and evaluating the rest of $h(z)$ at the dominant singularity ρ to get

$$K = -\frac{\sqrt{Q(\rho)}}{2\rho^2} \approx -3.91487$$

Since the singularity is of the form $(1 - z/\rho)^{1/2}$, we read off $\alpha = 1/2$. We then take the general equation

$$s_n \sim \frac{K}{\Gamma(-1/2)} (\frac{1}{\rho})^n n^{-1-\alpha}$$

and plug in our values to obtain our final answer.

$$s_n \sim 1.10437(2.61803)^n n^{-3/2}.$$

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman DJ. Basic local alignment search tool. *J Mol Biol.*, 215(3):403–410, 1990.

[2] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17):3389–3402, 1997.

[3] A.R. Banerjee, J.A. Jaeger, and D.H. Turner. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32:153–163, 1993.

[4] M. Bekaert, L. Bidou, A. Denise, G. Duchateau-Nguyen, J. Forest, C. Froidevaux, I. Hatin, J. Rousset, and M. Termier. Towards a computational model for -1 eukaryotic frameshifting sites. *Bioinformatics*, 19:327–335, 2003.

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

[6] M. Bousquet-Melou. Convex polyominoes and algebraic languages. *Journal of Physics A: Mathematical and General*, 25:1935–1944, 1992.

[7] C. Brown, B. Hendrich, J. Rupert, R. Lafreniere, Y. Xing, J. Lawrence, and H. Willard. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71:527–542, 1992.

[8] E.R. Canfield. Remarks on an asymptotic method in combinatorics. *Journal of Combinatorial Theory*, 37:348–352, 1984. Series A.

[9] R. V. Churchill. *Complex Variables and Applications*. McGraw-Hill, 1960.

[10] P. Clote. Combinatorics of saturated secondary structures of RNA. *J. Comp Biol.*, 13:1640–1657, 2006. 9.

[11] P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, 2000. 279 pages.

[12] S. Commans and A. Böck. Selenocysteine inserting tRNAs: an overview. *FEMS Microbiology Reviews*, 23:333–351, 1999.

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

[14] J.A. Doudna and T.R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418(6894):222–228, 2002.

[15] P. Flajolet. Singular combinatorics. In *Proceedings of the International Congress of Mathematicians*, volume 3, pages 561–571, 2002.

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

[17] R. Giegerich, B. Voss, and M. Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Res.*, 32(16):4843–4851, 2004.

[18] I.L. Hofacker, P. Schuster, and P.F. Stadler. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88:207–237, 1998.

[19] Jr. I. Tinoco, P.N. Borer, B. Dengler, M.D. Levin, O.C. Uhlenbeck, D.M. Crothers, and J.Bralla. Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol.*, 246(150):40–41, 1973.

[20] Jr. I. Tinoco and M. Schmitz. Thermodynamics of formation of secondary structure in nucleic acids. In E.D. Cera, editor, *Thermodynamics in Biology*, pages 131–176. Oxford University Press, 2000.

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

[22] L.P. Lim, M.E. Glasner, S. Yekta, C.B. Burge, and D.P. Bartel. Vertebrate microRNA genes. *Science*, 299(5612):1540, 2003.

[23] A. I. Markushevich. *Theory of Functions of a Complex Variable*. Chelsea Publishing Company, 1977.

[24] D.H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10:1178–1190, 2004.

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

[26] D.H. Mathews, D.H. Turner, and M. Zuker. Secondary structure prediction. In S. Beaucage, D.E. Bergstrom, G.D. Glick, and R.A. Jones, editors, *Current Protocols in Nucleic Acid Chemistry*, pages 11.2.1–11.2.10. John Wiley & Sons, New York, 2000.

[27] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[28] A. Meir and J.W. Moon. On an asymptotic method in enumeration. *Journal of Combinatorial Theory*, 51:77–89, 1989. Series A.

[29] S. Moon, Y. Byun, H.-J. Kim, S. Jeong, and K. Han. Predicting genes expressed via −1 and +1 frameshifts. *Nucleic Acids Res.*, 32(16):4884–4892, 2004.

[30] M. Nebel. Combinatorial properties of rna secondary structures. *Journal of Computational Biology*, 3(9):541–574, 2003.

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

[32] A.M. Odlyzko. Asymptotic enumeration methods. In R.L. Graham, M. Grötschel, and L. Lovász, editors, *Handbook of Combinatorics*, pages 1063–1230. Elsevier Science B.V. and MIT Press, Amsterdam and Cambridge, 1995. Volume II.

[33] R. Penchovsky and R.R. Breaker. Computational design and experimental validation of oligonucleotide-sensing allosteric ribozymes. *Nature Biotechnology*, 23(11):1424–1431, 2005.

[34] E.A. Rodland. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *J Comput Biol*, 13(6):1197–1213, 2006.

[35] B. Salvy and P. Zimmerman. Gfun: a maple package for the manipulation of generating and holonomic functions in one variable. *ACM Transactions on Mathematical Softwares*, 20(2):163–177, 1994.

[36] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *J Mol Biol.*, 147(1):195–197, 1981.

[37] P. Steffen, B.Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

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

[39] M. Vauchaussade de Chaumont and X.G. Viennot. Enumeration of RNA's secondary structures by complexity. In V. Capasso, E. Grosso, and S.L. Paven-Fontana, editors, *Mathematics in Medecine and Biology*, volume 57 of *Lecture Notes in Biomathematics*, pages 360–365, 1985.

[40] Q. Vicens and T.R. Cech. Atomic level architecture of group I introns revealed. *Trends Biochem Sci.*, 31(1):41–51, 2006.

[41] B. Voss, R. Giegerich, and M. Rehmsmeier. Complete probabilistic analysis of RNA shapes. *BMC Biol.*, 4(5), 2006.

[42] P. Walter and G. Blobel. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, 299(5885):691–698, 1982.

[43] M. S. Waterman. Secondary structure of single stranded nucleic acids. *Advances in Mathematics Supplementary Studies*, 1(1):167–212, 1978.

[44] M. S. Waterman. *Introduction to Computational Biology - Maps, Sequences and Genomes*. Chapman & Hall, 1995.

[45] J.S. Weinger, K.M. Parnell, S. Dorner, R. Green, and S.A. Strobel. Substrate-assisted catalysis of peptide bond formation by the ribosome. *Nature Structural & Molecular Biology*, 11:1101–1106, 2004.

[46] W. C. Winkler, S. Cohen-Chalamish, and R. R. Breaker. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. U.S.A.*, 99:15908–15913, 2002.

[47] T. Xia, Jr. J. SantaLucia, M.E. Burkard, R. Kierzek, S.J. Schroeder, X. Jiao, C. Cox, and D.H. Turner. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37:14719–35, 1999.

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

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

BIOLOGY DEPARTMENT, BOSTON COLLEGE, HIGGINS 577, CHESTNUT HILL, MA 02467

E-mail address: {lorenzwi,ponty,clote}@bc.edu