Asymptotic number of hairpins of saturated RNA secondary structures

Peter Clote[*] Evangelos Kranakis[†] Danny Krizanc[‡]

Abstract

In the absence of chaperone molecules, RNA folding is believed to depend on the distribution of kinetic traps in the energy landscape of all secondary structures. Kinetic traps in the Nussinov energy model are precisely those secondary structures that are *saturated*, meaning that no base pair can be added without introducing either a pseudoknot or base triple. In this paper, we compute the asymptotic expected number of hairpins in saturated structures. For instance, if every hairpin is required to contain at least $\theta = 3$ unpaired bases and the probability that any two positions can base-pair is $p = 3/8$, then the asymptotic number of saturated structures is $1.34685 \cdot n^{-3/2} \cdot 1.62178^n$, and the asymptotic expected number of hairpins follows a normal distribution with mean $0.06695640 \cdot n + 0.01909350 \cdot \sqrt{n} \cdot \mathcal{N}$. Similar results are given for values $\theta = 1, 3$ and $p = 1, 1/2, 3/8$; for instance, when $\theta = 1$ and $p = 1$, the asymptotic expected number of hairpins in saturated secondary structures is $0.123194 \cdot n$, a value greater than the asymptotic expected number $0.105573 \cdot n$ of hairpins over all secondary structures. Since RNA binding targets are often found in hairpin regions, it follows that saturated structures present potentially more binding targets than non-saturated structures, on average. Next, we describe a novel algorithm to compute the *hairpin profile* of a given RNA sequence: given RNA sequence $a_1, \ldots, a_n$, for each integer k, we compute that secondary structure S_k having minimum energy in the Nussinov energy model, taken over all secondary structures having k hairpins. We expect that an extension of our algorithm to the Turner energy model may provide more accurate structure prediction for particular RNAs, such as tRNAs and purine riboswitches, known to have a particular number of hairpins. Mathematica™computations, C and Python source code, and additional supplementary information are available at the web site `http://bioinformatics.bc.edu/clotelab/RNAhairpinProfile/`.

1 Introduction

Since the function of RNA often depends on its structure, much work has been done on secondary structure prediction, using stochastic context free grammars [23, 18], thermodynamic algorithms [44, 15, 24], and kinetic folding algorithms [10, 42, 6].

Formally, a secondary structure for a given RNA nucleotide sequence $a_1, \ldots, a_n$ is a set S of base pairs (i, j), such that *(i)* if $(i, j) \in S$ then a_i, a_j form either a Watson-Crick (AU,UA,CG,GC) or wobble (GU,UG) base pair, *(ii)* if $(i, j) \in S$ then $j - i > \theta = 3$ (a steric constraint requiring that there be at least $\theta = 3$ unpaired bases between any two paired

[*] Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.

[†] School of Computer Science, Carleton University, K1S 5B6, Ottawa, Ontario, Canada.

[‡] Department of Mathematics and Computer Science, Wesleyan University, Middletown CT 06459, USA.

bases), *(iii)* if $(i, j) \in S$ then for all $j' \neq j$ and $i' \neq i$, $(i', j) \notin S$ and $(i, j') \notin S$ (nonexistence of base triples), *(iv)* if $(i, j) \in S$ and $(k, \ell) \in S$, then it is not the case that $i < k < j < \ell$ (nonexistence of pseudoknots). For the purposes of this paper, following Stein and Waterman [37], we consider the *homopolymer* model of RNA, in which condition *(i)* is dropped, thus entailing that any base can pair with any other base, and we modify condition *(ii)* so that $\theta = 1$. With inessential additional complications in the combinatorics, we can handle the situation where θ is any fixed positive constant, and where there is a fixed probability p, called *stickiness* [40, 16], that any two positions can pair. For simplicity of argument, in the homopolymer model, we take $\theta = 1$ and $p = 1$. See Table 2 for asymptotic values computed for $\theta = 1, 3$ and $p = 1, 1/2, 3/8$.

An RNA secondary structure is *saturated* [3, 39, 5] if no base pair can be added without violating the definition of secondary structure (i.e. without introducing either a pseudoknot or base triple). Recalling that in the Nussinov energy model [31], the energy of a secondary structure is -1 times the number of base pairs, it follows that saturated structures have a *maximal* number of base pairs, though not necessarily a *maximum* number of base pairs. If a given saturated structure S is not a minimum energy structure S_0,[*] then any folding pathway from S to S_0 must proceed by *removing* a base pair from S – an energetically unfavorable move with respect to the Nussinov energy model. It follows that saturated structures form kinetic traps in the Nussinov energy model. Since the kinetics of RNA structure formation is thought to depend on the distribution of kinetic traps (i.e. saturated structures), it is of theoretical interest to compute the number of saturated structures as well as structural features such as the expected number of base pairs and expected number of hairpins. In previous work, we determined the asymptotic number $1.07427 \cdot n^{-3/2} \cdot 2.35467^n$ of saturated secondary structures [3] and the expected number $0.337361 \cdot n$ of base pairs in saturated secondary structures [5]. These values should be compared with the asymptotic number $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$ of all secondary structures, as computed by Stein and Waterman [37], and the expected number $0.276393 \cdot n$ of base pairs over all secondary structures, as computed by Nebel [27].[†] In this paper, we show that the expected number of hairpins in saturated secondary structures is asymptotically equivalent to $0.123194 \cdot n$, which is greater than the asymptotic expected number $0.105573 \cdot n$ of hairpins over all secondary structures.[‡]

Secondary structures are conveniently displayed in Vienna *dot bracket notation*, consisting of a balanced parenthesis expression with dots, where an unpaired nucleotide at position i is depicted by a dot at that position, while a base pair (i, j) is depicted by the presence of matching left and right parentheses located respectively at positions i and j. The precursor microRNA with miRBase [13] accession code hvt-mir-H14 and ID MI0012627 has minimum free energy structure, as computed by RNAfold from Vienna RNA Package [15], is given by

```
CGGACUCAUUCAGCGGGCAAUGUAGACUGUGUACCAAGUGACAGCUACAUUGCCCGCUGGGUUUCUG
((((...((((((((((((((((((.((((.(((...))))))))))))))))))))))))).))))
```

[*] In the Nussinov energy landscape, due to *degeneracy* of the model, the minimum energy structure may not be unique. Indeed, in [3], we show that even RNA homopolymers have quadratically many minimum energy structures.

[†] In Theorem 10 of Nebel [27], it is shown that the number of *unpaired* nucleotides is asymptotically equal to $\frac{n}{\sqrt{5}}$, whence the stated result follows. One can compare as well with the asymptotic number of hairpins in k-noncrossing structures, given in Table 2 of Nebel, Reidys and Wang [30].

[‡] In Theorem 16 of Nebel [27], it is shown that the expected number of hairpins over all secondary structures is asymptotically equivalent to $(1 - \frac{2\sqrt{5}}{5}) \cdot n \sim 0.105573 \cdot n$.

Note that this structure is saturated. In the homopolymer model considered in this paper, there are precisely five saturated structures for a homopolymer sequence of length 5, separated by semi-colons:

((•));•(••);(••)•;(•)••;••(•)

while there are eight homopolymer structures for RNA of length 5, separated by semi-colons:

•••••; (•)••; (••)•; (•••); •(•)•; •(••); ••(•); ((•));

A *hairpin* is defined to be a base pair (i, j), such that all positions from $i+1, \dots, j-1$ are unpaired. It follows that the expected number of hairpins, over all saturated (homopolymer) structures of length 5, is $\frac{5}{5} = 1$, while the expected number of hairpins, over all structures of length 5, is $\frac{0+7 \cdot 1}{8} = \frac{7}{8}$. Since the exhaustive list of all eight saturated structures of length 6 is given by

((•))•;•((•));((•)•);(•(•));((••));(•)(•);(••)••;••(••)

it follows that the expected number of hairpins, over all saturated structures of length 6, is $\frac{5 \cdot 1 + 1 \cdot 2}{6} = \frac{7}{6}$.

2 Context-free grammars and DSV method

In this section, we define non-ambiguous context-free grammars, describe the DSV methodology, and state the Flajolet-Odlyzko theorem, from which we derive asymptotic results. Since we have previously provided a detailed description of this method in Lorenz et al. [22], we only sketch a brief overview, referring the reader to [22] for details.

Context-free grammars

Let Σ be a finite set of symbols. A language is a subset of Σ^*, the set of all words $a_1, \dots, a_n$, where $a_i \in \Sigma$ for all $0 \leq i \leq n$ and n is an arbitrary integer. In our application, Σ will consist of left parenthesis (, right parenthesis), and dot • when discussing secondary structures. 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 $A \to \alpha_1, \dots, A \to \alpha_m$ 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 derivation is said to be a *leftmost* derivation, provided that each application of a rule is applied to the leftmost variable in the expression. A grammar is *non-ambiguous* provided that no word $w \in L(G)$ has two distinct leftmost derivations (this condition is equivalent to requiring that no $w \in L(G)$ have two distinct parse trees).

Type of nonterminal	Equation for generating function
$S \to T \mid U$	$S(z) = T(z) + U(z)$
$S \to T\,U$	$S(z) = T(z)U(z)$
$S \to t$	$S(z) = z$
$S \to \varepsilon$	$S(z) = 1$

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 generating functions for the languages $L(S)$, $L(T)$, $L(U)$ are respectively denoted by $S(z)$, $T(z)$, $U(z)$. Table taken from Lorenz et al. [22].

From grammars to generating functions

A general approach in enumerating combinatorial objects is to introduce generating functions. The generating function for class $\mathcal{C}$ of objects is a complex function defined by $C(z) = \sum_{i\geq 0} c_n z^n$, where c_n is the number of objects length n that belong to $\mathcal{C}$. For certain generating functions $C(z)$, it may be possible to derive a closed-form formula for the Taylor coefficient c_n of order n, denoted by $[z^n]C(z)$. Often, expecially when the collection of all combinatorial objects is generated by a context free grammar, it is possible to efficiently derive the behavior of c_n when n approaches infinity, i.e. to derive a function $g(n)$, such that $\lim_{n\to\infty} \frac{c_n}{g(n)} = 1$, denoted by asymptotic equality $c_n \sim g(n)$.

Theorem 2.1 *Let $G = (V, \Sigma, R, S_0)$ be a non-ambiguous context-free grammar. For each nonterminal symbol S, let $S(z)$ be the corresponding generating function, defined by applying the translation scheme from Table 1. If $C(z) = \sum_{i\geq 0} c_n z^n$ is the length generating function for $L(G)$, where c_n is the number of length n words in $L(G)$, then $S_0(z) = C(z)$.*

3 Expected number of hairpins in saturated structures

A *hairpin* is a base pair (i, j) such that no position strictly between i and j is paired. In the homopolymer model with $\theta = 1$ and $p = 1$, a hairpin occurring within a saturated structure must have exactly one or two unpaired bases between the closing base pair.

Define the context free grammar G_1 with nonterminal symbols S, R, start symbol S, and production rules

$$\begin{aligned} S &\to \bullet \mid \bullet\bullet \mid R\bullet \mid R\bullet\bullet \mid \boldsymbol{(}\,S\,\boldsymbol{)} \mid S\,\boldsymbol{(}\,S\,\boldsymbol{)} \\ R &\to \boldsymbol{(}\,S\,\boldsymbol{)} \mid R\,\boldsymbol{(}\,S\,\boldsymbol{)} \end{aligned}$$

A nucleotide position i in $\{1, \ldots, n\}$ is said to be *visible* in a given structure S if there is no base pair $(x, y) \in S$ for which $x \leq i \leq y$. In other words, visible positions are external to every base pair of S. A straightforward proof by induction on word length establishes that G_1 is a non-ambiguous grammar for the collection of all saturated secondary structures [5], and that the nonterminal S generates all saturated structures, while the nonterminal R generates all saturated structures which have no *visible* positions.

In order to count the expected number of hairpins in saturated structures, we need to *mark* occurrences of hairpins by a finer grammar. To that end, we define the alternate non-ambiguous grammar G_2 with production rules

$$\begin{array}{rcl} S & \rightarrow & D|N \\ D & \rightarrow & \bullet|\bullet\bullet \\ N & \rightarrow & RD|\,\mathbf{(}\,D\,\mathbf{)}\,|\,\mathbf{(}\,N\,\mathbf{)}\,|S\,\mathbf{(}\,D\,\mathbf{)}\,|S\,\mathbf{(}\,N\,\mathbf{)} \\ R & \rightarrow & \mathbf{(}\,D\,\mathbf{)}\,|\,\mathbf{(}\,N\,\mathbf{)}\,|R\,\mathbf{(}\,D\,\mathbf{)}\,|R\,\mathbf{(}\,N\,\mathbf{)} \end{array}$$

where S generates all saturated structures, R generates all saturated structures that have no visible positions, D generates a saturated empty structure (*dots*, i.e. only $\bullet$ or $\bullet\bullet$), and N generates saturated structures that contain at least one base pair (*not dots*). It easily follows by DSV methodology [5] and a Mathematica™computation that G_1 and G_2 are equivalent grammars, both non-ambiguously generating exactly the saturated secondary structures (see web supplement for computation). Define

$$S(z,u) = \sum_{n=0}^{\infty}\sum_{k=0}^{\infty} s_{n,k} z^n u^k \tag{1}$$

where the coefficient $s_{n,k}$ in the series expansion of S represents the number of secondary structures on $[1,n]$ having k hairpins. Thus $\sum_k k \cdot \frac{s_{n,k}}{s_n}$ is the expected number of hairpins in saturated secondary structures on $[1,n]$. By using the methods of [5], we can compute the expected number of hairpins in saturated structures by

$$\mathbb{E}(X_n) = \frac{[z^n]\frac{\partial S(z,u)}{\partial u}(z,1)}{[z^n]S(z,1)} \tag{2}$$

where X_n is the random variable for the number of hairpins in a saturated secondary structure (see web supplement for details of this complicated computation). However, a substantially simpler and more complete result can be obtained by application of Drmota's Theorem 2, described below.

Now, by applying the DSV methodology from Table table:DSV to grammar G_2, we have the system of equations

$$\begin{array}{rcl} S & = & D+N \\ D & = & z+z^2 \\ N & = & R\cdot D + Dz^2 + Nz^2 + S\cdot Dz^2 + S\cdot Nz^2 \\ R & = & Dz^2 + Nz^2 + R\cdot Dz^2 + R\cdot Nz^2 \end{array}$$

whence we can mark the introduction of a hairpin by the auxilliary variable u, thus yielding

$$\begin{array}{rcl} S & = & D+N \\ D & = & z+z^2 \\ N & = & R\cdot D + D\cdot uz^2 + N\cdot z^2 + S\cdot D\cdot uz^2 + S\cdot N\cdot z^2 \\ R & = & D\cdot uz^2 + N\cdot z^2 + R\cdot D\cdot uz^2 + R\cdot N\cdot z^2 \end{array}$$

Using three distinct approaches, in [3, 5, 11], we computed the asymptotic number of saturated secondary structures, for $\theta = 1$, $p = 1$. In our opinion, the simplest approach consists of: *(i)* giving a non-ambiguous context free grammar to generate all saturated structures, *(ii)* using the DSV methodology described below to obtain a functional relation of the form $\Phi(z, S(z)) = S(z)$, *(iii)* applying the Drmota-Lalley-Woods Theorem [9, VII.6], stated as Theorem 1 below. The following description of the theorems of Drmota-Lalley-Woods (Theorem 1) and of Drmota (Theorem 2), up to the end of Remark 2, is adapted from our paper [11].

For $r \geq 1$, a *weighted combinatorial class indexed by r parameters* is a set $\mathcal{A}$ together with a *weight function* W from $\mathcal{A}$ to $\mathbb{R}$ and r *parameter-functions* $P_1, \ldots, P_r$ from $\mathcal{A}$ to $\mathbb{N}$ such that for any fixed integers $n_1, \ldots, n_r$, the set of structures $\gamma \in \mathcal{A}$ such that $P_1(\gamma) = n_1, \ldots, P_r(\gamma) = n_r$ is finite. This set is denoted $\mathcal{A}[n_1, \ldots, n_r]$. For example, when $r = 1$, we could take $\mathcal{A}[n]$ to be the set of saturated secondary structures for a homopolymer of length n; when $r = 2$, we could take $\mathcal{A}[n_1, n_2]$ to be the set of saturated secondary structures for a homopolymer of length n_1 having n_2 hairpins.

For a weighted combinatorial class indexed by r parameters, the corresponding multivariate generating function is

$$A(z_1, \ldots, z_r) := \sum_{\gamma \in \mathcal{A}} z_1^{P_1(\gamma)} \cdots z_r^{P_r(\gamma)} W(\gamma). \tag{3}$$

Here, we say that variable z_i *marks* the parameter P_i, for $1 \leq i \leq r$. We also use the notation

$$[z_1^{n_1} \ldots z_r^{n_r}] A(z_1, \ldots, z_r) := \sum_{\gamma \in \mathcal{A}[n_1, \ldots, n_r]} W(\gamma).$$

In combinatorial analysis, one often considers a weight function, $W(S) = 1$, that assigns weight 1 to each structure S; however, structures S can be weighted; for instance, when considering *stickiness* p, we could assign a weight $W(S) = p^m$, where structure S has exactly m hairpins. The variables z_i are a priori considered as formal, but one can also evaluate a generating function at given values, provided the sum converges. The *convergence domain* of $A(z_1, \ldots, z_r)$ is the set of r-tuples $(z_1, \ldots, z_r)$ of nonnegative real values such that $A(z_1, \ldots, z_r)$ converges. In our applications, we consider only values 1, 2 for r, where $A(z_1) = S(z) = \sum_{n=0}^{\infty} s_n z^n$ is the generating function for the set of saturated structures, and $A(z_1, z_2) = S(z, u) = \sum_{n=0}^{\infty} \sum_{m=0}^{n} s_{n,m} z^n u^m$ is the bivariate generating function for saturated structures having m hairpins.

Consider a functional equation of the form

$$y = S(z) = \Phi(z, a(z)) = \Phi(z, y), \tag{4}$$

where $\Phi(z, y)$ is a rational expression in z, y. Such an equation is called *admissible* if the following conditions are satisfied:

- The rational expression $\Phi(z, y)$ has a series expansion in z and y with non-negative coefficients, is nonaffine in y, and satisfies[§] $\Phi(0, 0) = 0$ and $\Phi_y(0, 0) = 0$.
- The unique generating function $y = S(z)$ solution of (4) is aperiodic, i.e., can not be written as $S(z) = z^q \tilde{S}(z^p)$ for some integers p, q with $p \geq 2$.

[§] Subscript notation is used for partial derivatives.

There is an easy criterion to check the aperiodicity condition: it suffices to prove that there is some n_0 such that $[z^n]S(z) > 0$ for $n \geq n_0$.

Theorem 1 (Drmota-Lalley-Wood) *Let $y = S(z)$ be the generating function that is the unique solution of an admissible equation $y = \Phi(z, y)$. Then*

$$[z^n]S(z) \sim c\,\gamma^n n^{-3/2},$$

where $\gamma = 1/z_0$, with (z_0, y_0) the unique pair in the convergence domain of $\Phi(t, y)$ that is solution of the singularity system*:*

$$y = \Phi(z, y), \quad \Phi_y(z, y) = 1;$$

and where

$$c = \sqrt{z_0\Phi_z(z_0, y_0)/(2\pi\Phi_{y,y}(z_0, y_0))}.$$

Remark 1 *The Drmota-Lalley-Wood theorem is proved in [9, VII.6] where $\Phi(z, y)$ a polynomial; however, it can be checked that the same conclusions hold if $\Phi(z, y)$ is a bivariate series that diverges at all its singularities.*

An equation of the form

$$S(z, u) = \Phi(z, u, S(z, u)), \tag{5}$$

where $\Phi(z, u, y)$ is a rational expression in z, u and y, is called *simple*[¶] if $\Phi(z, u, y)$ is nonconstant in u, has a series expansion (in z, u, y) with non-negative coefficients, the equation $y = \Phi(z, 1, y)$ is admissible (as previously defined), and there is a 3×3-matrix $m[i, j]$ with integer coefficients and nonzero determinant such that $[z^{m[i,1]}u^{m[i,2]}y^{m[i,3]}]\Phi(z, u, y) > 0$ for all $i \in \{1, 2, 3\}$.

Theorem 2 (Drmota [7]) *Let $y = S(z, u)$ be a generating function that is the unique solution of a simple equation $y = \Phi(z, u, y)$. Assume that the generating function $b(z, u) = \sum_{\gamma\in\mathcal{G}} z^{|\gamma|}u^{\chi(\gamma)}W(\gamma)$ of a weighted combinatorial class $\mathcal{G}$ is given by $b(z, u) = \Psi(z, u, S(z, u))$, with $\Psi(z, u, y)$ a rational expression with non-negative coefficients (in the series expansion), nonconstant in y, and such that the convergence domain of $\Psi(z, 1, y)$ is included in the one of $\Phi(z, 1, y)$. For $n \geq 0$ let $\mathcal{G}_n := \{\gamma \in \mathcal{G},\ |\gamma| = n\}$, and define the random variable X_n as $\chi(\gamma)$, with γ a random structure in $\mathcal{G}_n$ under the distribution*

$$P(\gamma) = \frac{W(\gamma)}{\sum_{\gamma\in\mathcal{G}_n} W(\gamma)}.$$

For $u > 0$ in a neighborhood of 1, denote by $\rho(u)$ the radius of convergence of $y : z \to S(z, u)$, and let

$$\mu = -\frac{\rho'(1)}{\rho(1)}, \quad \sigma^2 = -\frac{\rho''(1)}{\rho(1)} + \mu^2 + \mu.$$

Then μ and σ are strictly positive and $\dfrac{X_n - \mu \cdot n}{\sigma\sqrt{n}}$ converges as a random variable to a normal distribution.

[¶] We follow Drmota [7], in using the term *simple*, whereas the term *admissible* was used in [11].

Remark 2 *Again the theorem was originally proved for polynomial systems, but the arguments of the proof hold more generally when* Φ *is rational. The role of the condition involving the existence of a nonsingular* 3×3 *matrix is to grant the strict positivity of* σ*, as proved in [8].*

We now describe how to compute the expected number of hairpins in saturated secondary structures, where $\theta = 1$ and $p = 1$. The computations for other values of θ, p in Table 2 proceed similarly. In particular, the Mathematica™computations and auxilliary data can be downloaded at the web supplement site

`http://bioinformatics.bc.edu/clotelab/RNAhairpinProfile/`.

Let the constant $p = 1$ denote stickiness. By DSV methodology, we obtain the equations

$$\begin{aligned} S &= D + N \\ D &= z + z^2 \\ N &= RD + pDz^2 + pNz^2 + pSDz^2 + pSNz^2 \\ R &= pDz^2 + pNz^2 + pRDz^2 + pRNz^2 \end{aligned}$$

corresponding to the context free grammar that generates the collection of saturated structures. This system contains 4 equations in 5 variables, hence we can eliminate variables N, D, R to obtain the equation

$$S^3 z^4 + S(1 - z^2) + S^2 z^2(-2 + z^2) = z(1 + z)$$

from which we obtain the functional relation

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

which satisfies $S(z) = \Phi(z, S(z))$. By numerical solution of the system of equations

$$\begin{aligned} \Phi(z, S) &= S \\ \Phi_S(z, S) &= \frac{\partial \Phi(z, S)}{\partial S} = 1 \end{aligned}$$

we obtain the solutions

$$\begin{aligned} z &= 3.2141, S = -0.587227 \\ z &= -0.854537, S = 0.988667 \\ z &= 0.424687, S = 1.6569 \\ z &= -2.29493, S = -0.513379 \\ z &= -0.244657 + 0.5601i, S = -0.741229 + 0.680476i \\ z &= -0.244657 - 0.5601i, S = -0.741229 - 0.680476i \end{aligned}$$

from which it follows that $z0 = 0.42468731042025953$ is the dominant singularity; i.e. having least modulus $|z|$. The corresponding value of S is $S0 = 1.6568963458689725$. We now apply the Drmota-Lalley-Woods Theorem, where $y = S(z)$. It follows that the asymptotic number of saturated secondary structures, for $\theta = 1$ and $p = 1$, is

$$1.07427 \cdot n^{-3/2} \cdot 2.35467^n.$$

This value agrees with that obtained in [3, 5, 11].

We now turn to the computation of the mean and standard deviation of the expected number of hairpins, using Drmota's Theorem. By weighting the previously given equations with an auxilliary variable u, used each introduction of a hairpin, we obtain the system of equations

$$\begin{aligned}
S &= D + N \\
D &= z + z^2 \\
N &= RD + puDz^2 + pNz^2 + puSDz^2 + pSNz^2 \\
R &= puDz^2 + pNz^2 + puRDz^2 + pRNz^2
\end{aligned}$$

This system contains 4 equations and 6 variables, thus we eliminate all variables except for z, u, S to obtain the functional $\Phi(z, u, S)$, defined to be equal to the following:

$$\frac{S^3z^4 + S^2z^2(-2 + z^2 + 2(-1+u)z^3 + 2(-1+u)z^4) - (-z(1+z)(-1 - (-1+u)z^2 + (-1+u)^2z^5 + (-1+u)^2z^6))}{-(1+z)(1 - z - 2(-1+u)z^3 + 2(-1+u)z^5 + (-1+u)^2z^6 + (-1+u)^2z^7)}$$

Express each of $S - \Phi(z, u, S)$ and $1 - \frac{\partial\Phi(z,u,S)}{\partial S}$ as rational expressions having the same common denomiator c; i.e. $\frac{a}{c} = S - \Phi(z, u, S)$ and $\frac{b}{c} = 1 - \frac{\partial\Phi(z,u,S)}{\partial S}$. Compute the resultant $Res(a, b)$ of a, b with respect to variable S, to obtain

$$\begin{aligned}
-4z^{11} - 5z^{12} + 6z^{13} + 23z^{14} + 12uz^{14} + 34z^{15} \\
+26uz^{15} + 12z^{16} + 20uz^{16} - 30z^{17} + 38uz^{17} - 12u^2z^{17} - \\
61z^{18} + 94uz^{18} - 37u^2z^{18} - 74z^{19} + 124uz^{19} - 50u^2z^{19} \\
-65z^{20} + 122uz^{20} - 61u^2z^{20} + 4u^3z^{20} - 52z^{21} + 120uz^{21} \\
-84u^2z^{21} + 16u^3z^{21} - 36z^{22} + 96uz^{22} - 84u^2z^{22} + 24u^3z^{22} \\
-16z^{23} + 48uz^{23} - 48u^2z^{23} + 16u^3z^{23} - 4z^{24} \\
+12uz^{24} - 12u^2z^{24} + 4u^3z^{24}
\end{aligned}$$

Let RES denote the expression obtained by replacing variable z in the previous expression by the function $z(u)$. From Drmota's Theorem we have that

$$\mu = -\frac{z'(1)}{z(1)}, \quad \sigma^2 = -\frac{z''(1)}{z(1)} + \mu^2 + \mu$$

which by computation (see web supplement) yields $\mu = 0.123194$ and $\sigma^2 = 0.0341867$. By Drmota's Theorem it follows that the number of hairpins in saturated structures with $\theta = 1$ and $p = 1$ is normally distributed with mean $0.12319400 \cdot n + 0.03418670 \cdot \sqrt{n} \cdot \mathcal{N}$.

We have just proved the following.

Theorem 3 *The asymptotic expected number of hairpins for saturated structures, where $\theta = 1$ and $p = 1$, is*

$$0.123194 \cdot n.$$

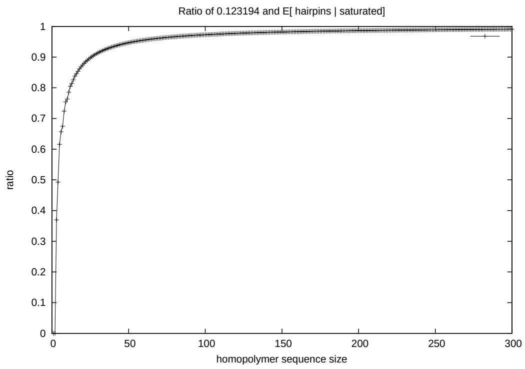

Figure 1: Ratio of the asymptotic expected number of hairpins of saturated secondary structures and the actual value, computed by dynamic programming.

This result should be compared with Theorem 16 of [27], where Nebel proves that the asymptotic expected number of hairpins in a secondary structure of a homopolymer of size n is

$$\left(1 - \frac{2 \cdot \sqrt{5}}{5}\right) n \approx 0.105573 \cdot n. \tag{6}$$

In other words, the expected number of hairpins in *saturated* structures is asymptotically approximately $16 - 17\%$ larger than the expected number of hairpins, taken over all structures.

Figure 1 shows the ratio of the asymptotic value $0.123194 \cdot n$ and the value, computed by dynamic programming, as explained in the Appendix (Python source code available on web supplement). Convergence is rather slow, compared to the rapid convergence of asymptotic results, such as those of Nebel [27], which concern expected values, when taken over all secondary structures.

Computations for different values of θ, p

By alternatively taking stickiness p to be 1/2 and 3/8, and by slightly modifying the context free grammar (for the case of $\theta = 3$), we obtain the number of saturated secondary structures given in Table 2. For example, when $\theta = 3$, the slightly modified non-ambiguous grammar G_3 has production rules

$$\begin{array}{rcl}
S & \rightarrow & D|N \\
D & \rightarrow & \bullet|\bullet\bullet|\bullet\bullet\bullet|\bullet\bullet\bullet\bullet \\
N & \rightarrow & RD|(\bullet\bullet\bullet)|(\bullet\bullet\bullet\bullet)|(N)|S(\bullet\bullet\bullet)|S(\bullet\bullet\bullet\bullet)|S(N) \\
R & \rightarrow & (\bullet\bullet\bullet)|(\bullet\bullet\bullet\bullet)|(N)|R(\bullet\bullet\bullet)|R(\bullet\bullet\bullet\bullet)|R(N)
\end{array}$$

where S generates all saturated structures, R generates all saturated structures that have no visible positions, D generates a structure of size 1,2,3 or 4 having no base pairs (hence saturated), and N generates saturated structures that contain at least one base pair. Using

θ	p	number of saturated structures	expected number of hairpins
1	1	$1.07427 \cdot n^{-3/2} \cdot 2.35467^n$	$0.12319400 \cdot n + 0.03418670 \cdot \sqrt{n} \cdot \mathcal{N}$
1	1/2	$1.37347 \cdot n^{-3/2} \cdot 1.87138^n$	$0.12426000 \cdot n + 0.03415170 \cdot \sqrt{n} \cdot \mathcal{N}$
1	3/8	$1.52744 \cdot n^{-3/2} \cdot 1.70513^n$	$0.12447200 \cdot n + 0.03410150 \cdot \sqrt{n} \cdot \mathcal{N}$
3	1	$0.76229 \cdot n^{-3/2} \cdot 2.10305^n$	$0.05983930 \cdot n + 0.01801440 \cdot \sqrt{n} \cdot \mathcal{N}$
3	1/2	$1.13709 \cdot n^{-3/2} \cdot 1.74543^n$	$0.06514370 \cdot n + 0.01882460 \cdot \sqrt{n} \cdot \mathcal{N}$
3	3/8	$1.34685 \cdot n^{-3/2} \cdot 1.62178^n$	$0.06695640 \cdot n + 0.01909350 \cdot \sqrt{n} \cdot \mathcal{N}$

Table 2: Asymptotic number of saturated secondary structures and asymptotic expected number of hairpins in saturated structures, for sample values of θ, p. Here, θ denotes the minimum required number of unpaired bases in every hairpin loop. The expected number of hairpins follows a normal distribution, as indicated, where $\mathcal{N}$ denotes the standard normal distribution with mean 0 and standard deviation of 1. For asymptotic analysis, following Stein and Waterman [37], θ is often taken to be 1, while in RNA structure prediction software UNAFOLD [24] and RNAfold [15], θ is taken to be 3. The parameter p, often called *stickiness*, denotes the probability that any two positions can base-pair. In asymptotic analysis, often p is taken to be 1; if RNA sequences are randomly generated with a probability of 50% for C and G, then $p = 1/2$, while if RNA sequences are randomly generated with probability 1/4 for each nucleotide A,C,G,U, then $p = 3/8$. The asymptotic number of saturated structures was previously computed in [3, 5, 11] for $\theta = 1$ and $p = 1$ and in [11] for $\theta = 1$ and $p = 3/8$.

θ	p	number of *all* structures	expected number of hairpins
1	1	$1.10437 \cdot n^{-3/2} \cdot 2.61803^n$	$0.1055730 \cdot n + 0.179611 \cdot \sqrt{n} \cdot \mathcal{N}$
1	1/2	$1.45030 \cdot n^{-3/2} \cdot 2.18543^n$	$0.0986392 \cdot n + 0.176918 \cdot \sqrt{n} \cdot \mathcal{N}$
1	3/8	$1.63740 \cdot n^{-3/2} \cdot 2.04101^n$	$0.0950281 \cdot n + 0.175330 \cdot \sqrt{n} \cdot \mathcal{N}$
3	1	$0.71312 \cdot n^{-3/2} \cdot 2.28879^n$	$0.0530486 \cdot n + 0.128013 \cdot \sqrt{n} \cdot \mathcal{N}$
3	1/2	$1.04267 \cdot n^{-3/2} \cdot 1.96401^n$	$0.0546750 \cdot n + 0.128864 \cdot \sqrt{n} \cdot \mathcal{N}$
3	3/8	$1.22479 \cdot n^{-3/2} \cdot 1.85479^n$	$0.0546382 \cdot n + 0.128845 \cdot \sqrt{n} \cdot \mathcal{N}$

Table 3: Asymptotic number of *all* secondary structures and asymptotic expected number of hairpins in *all* structures, for sample values of θ, p. Here, θ, p are as in Table 2. Values are presented here for comparison with those from Table 2.

DSV methodology and accounting for stickiness p, we have the corresponding equations

$$\begin{aligned} S &= D + N \\ D &= z + z^2 + z^3 + z^4 \\ N &= RD + p(z^3 + z^4)z^2 + pNz^2 + pS(z^3 + z^4)z^2 + pSNz^2 \\ R &= p(z^3 + z^4)z^2 + pNz^2 + pR(z^3 + z^4)z^2 + pRNz^2 \end{aligned}$$

The computation then follows as explained above. Full details of all computations from Table 2 are given on the web supplement.

Since all of the values in Table 2 were previously computed by other authors, we created this table as a cross check of our method, and as well to make available our substantially simpler computations available in the web supplement. The asymptotic number $1.10437 \cdot n^{-3/2} \cdot 2.61803^n$ of all structures was computed for $\theta = 1, p = 1$ in [37], while the values for $\theta = 1, 2, 3, 5$ and $p = 1$ can be found in Table 1 of [16]. The expected number of hairpins

Num hairpins	Energy	Structure
1	-5	(((((..)))))
2	-4	(((..)(..)))
3	-3	(..)(..)(..)

Table 4: Table of the energy E_k and the minimum energy structure S_k, taken over all structures having k hairpins, obtained by running our program from Algorithm 3.1 on the input RNA sequence `ACGUACGUACGU`.

was computed for $\theta = 1$ and *arbitrary* $0 \leq p \leq 1$ in Theorem 3 of [28], although the term $\sigma\sqrt{n}\mathcal{N}$ is not given in that paper, since a different method was used. The expected number of hairpins was computed for $\theta = 3$ and $p = 1, 1/2, 3/8, 1/4$ in Table 3 of [16]. In particular, for the expected number of hairpins can be computed as $\frac{N_n}{S_n} \cdot \frac{L_n(1)}{N_n} \cdot n$, where values $\frac{N_n}{S_n}$ and $\frac{L_n(1)}{N_n} \cdot n$ are found in Table 3 of [16]. Note that due to roundoff error in Table 3 of [16], there are slight discrepancies between values in our Table 3 and those from [16]. In particular, for $\theta = 3$ and $p = 1, 1/2, 3/8$ respectively we obtain expected number of hairpins $0.0530486 \cdot n$, $0.0546750 \cdot n$, and $0.0546382 \cdot n$ our Table 3, while corresponding results from [16] are $0.05302635 \cdot n$, $0.05468732 \cdot n$, and $0.05465211 \cdot n$.

Minimum energy structure having exactly k hairpins

In this section, we describe a novel $O(n^5)$ time and $O(n^3)$ space algorithm, to compute, given an RNA sequence, simultaneously for each value of k, the minimum energy E_k over all secondary structures having exactly k hairpins (recall that energy is with respect to the Nussinov energy model). Moreover, the algorithm computes for each k, the secondary structure S_k having k hairpins and energy E_k.

To fix ideas, we describe the output for a toy example, where for simplicity we define the minimum number θ of unpaired bases in a hairpin loop to be 1. Table 4 describes the energy E_k and the minimum energy structure S_k, taken over all structures having k hairpins, obtained by running our program from Algorithm 3.1 on the input RNA sequence `ACGUACGUACGU`. For this example, there are no structures having four or more hairpins. See Table 5 for the output of our C implementation of Algorithm 3.1 for a transfer RNA from Rfam family RF00005 [12], where $\theta = 3$.

For instance, among structures having $k = 1$ hairpin loop, the structure

```
ACGUACGUACGU
(((((..)))))
```

has the largest number (5) of base pairs; i.e. the Nussinov energy is -5.

The algorithm is described as follows. Let $a_1, \ldots, a_n$ be a given RNA sequence, and let $BP(i, j, k)$ denote the maximum number of base pairs for a k-hairpin structure in the region $a_i, \ldots, a_j$. In a manner reminiscent of our work in [2], we use dynamic programming to compute, for all intervals $[i, j]$, and all values of k, the maximum number of base pairs in a structure having k hairpins on subsequence $a_i, \ldots, a_j$. We treat three cases. Case 1 considers structures on $[i, j]$ in which j is unpaired in $[i, j]$. Case 2 considers structures on $[i, j]$ in which i, j form a base pair. Case 3 considers structures on $[i, j]$ in which r, j form a base pair, for some intermediate $i < r < j$.

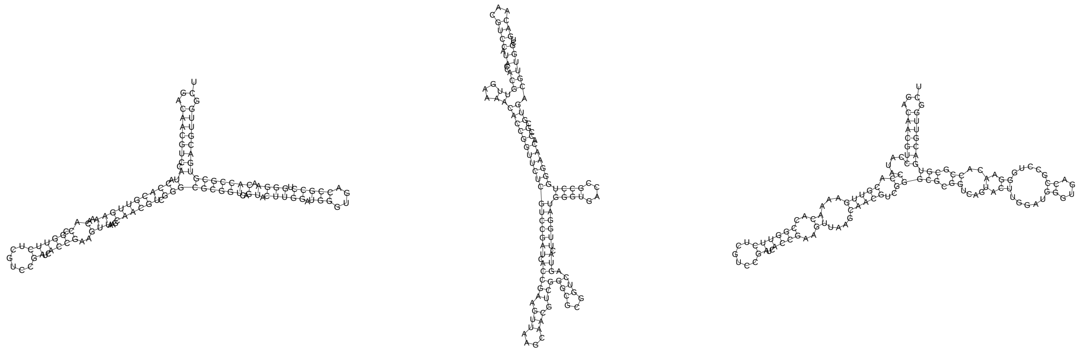

Figure 2: Structure prediction of the 119 nt 5S rRNA with EMBL accession code X13035.1/1-119, having sequence GACAACGUCC AUACCACGUU GAAAACACCG GUUCUCGUCC GAUCACCGAA GUUAAGCAAC GUCGGGCGCG GUCAGUACUU GGAUGGGUGA CCGCCUGGGA ACACCGCGUG ACGUUGGCU. *(A)* Minimum energy structure S_2 over all structures having exactly 2 hairpins, produced by our program. *(B)* Output of our implementation of the Nussinov-Jacobson algorithm [31], which computes the minimum energy structure. *(C)* Consensus structure from the Rfam database [12]. The base pair distance between structure A and C is 18, while the base pair distance between B and C is 77; moreover, visual inspection indicates that the output of our program (A) indeed closely resembles the Rfam consensus structure (C). This suggests that an implementation of Algorithm 3.1 using the Turner energy model [43] could improve structure prediction for RNA molecules, that are known to generally fold into structures with a given number of hairpins.

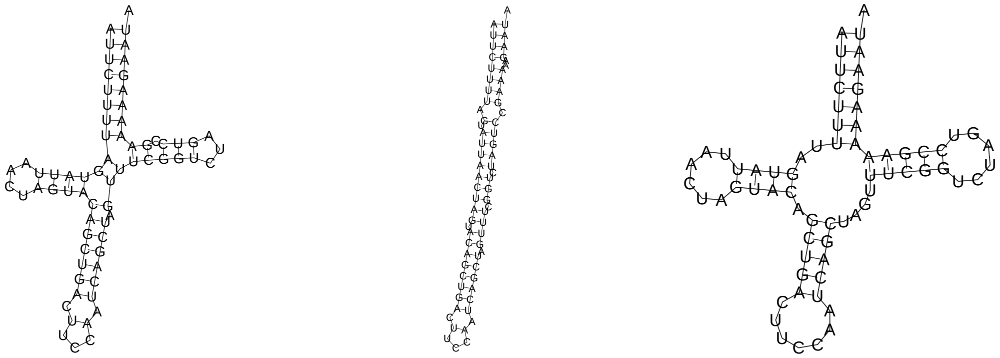

Figure 3: Structure prediction of the 79 nt tRNA with EMBL accession code AF493542.1/6654-6722, having sequence AUUCUUUUAG UAUUAACUAG UACAGCUGAC UUCCAAUCAG CUAGUUUCGG UCUAGUCCGA AAAAGAAUA. *(A)* Minimum energy structure S_3 over all structures having exactly 3 hairpins, produced by our program. *(B)* Output of our implementation of the Nussinov-Jacobson algorithm [31], which computes the minimum energy structure. *(C)* Consensus structure from the Rfam database [12]. The base pair distance between structure A and C is 14, while the base pair distance between B and C is 30; moreover, visual inspection indicates that the output of our program (A) indeed closely resembles the Rfam consensus structure (C).

Num hairpins	Energy	Structure
1	-27	((((((((.(.(((((((((.((((((((.(...).)))))).)))).))).))))).))))..)))).
2	-27	(((((((((((((((((.(((...))))..))..))).).))))(((((((...))).)))))))))).
3	-27	((((((((((((((....))))(((((((.(...).)))))).)))(((((...))).)))))))))).
4	-26	(((((((((((...)((((((...)((((.(...).)))))))))))((((...))).))))).)))).
5	-25	(((((((((...)((((((((...)((((.(...).))))))))))(.(...)))(...)))))))).
6	-24	(...)((((...)((((((((...)((((.(...).))))))))))(((((...))).)))))(...))
7	-22	(...)((((...)((((((((...)((((.(...).))))))))))(.(...)))(...))))(...).
8	-21	(...)((((...)((.(((((...)((((.(...).)))))))))(...)(...)(...))))(...))
9	-19	(...)((((...)(....)((...)((((.(...).)))))((...)((...).)(...))))(...))
10	-17	(...)((((((...).)(...)(((....))(...)(.....)))(...)(...)(...))))(...).
11	-14	(...)(...)(...).((...)(((....))(...)(.....)(...)))(...)(...)(......).

Table 5: Output of our C implementation of Algorithm 3.1 on the transfer RNA described in Figure 3.

Algorithm 3.1

```
for d=THETA+1 to n-1
    for i=1 to n-THETA
        j = i+d;
        if (j>n) break;
        for k=1 to n/2 //note n/2 max number of base pairs
            //CASE 1:  j does not pair in [i,j]
            max = BP(i,j-1,k)
            //CASE 2:  i and j pair together
            if a[i] can pair with a[j]
                num = BP(i+1,j-1,k)
                if (k=1 and num=0) or num>0
                    num += 1 //add 1 due to the basepair (i,j)
                    if (max<=num)
                        max = num
            //CASE 3:  r and j pair, for some i<r<j
            for r=i+1 to j-THETA-1
                if a[r] can pair with a[j]
                    //Case  3:  k = k0+k1 and r,j basepair
                    //Case 3a:  k0=0, k1=k and no bp in region i..r-1
                    if no positions can pair in region [i,r-1]
                        num = BP(r+1,j-1,k)
                        if (k=1 and num=0) or num>0
                            num += 1 //add 1 due to base pair (r,j)
                            if (max<num)
                                max = num
                        for k0=1 to kk0
```

```
                              k1=k-k0
                              x = BP(i,r-1,k0)
                              y = BP(r+1,j-1,k1)
                              if (k1=1 and x>0 and y=0) or (x>0 and y>0)
                                  num = x+y+1 //add 1 for basepair (r,j)
                                  if (max<num)
                                      max=num
       BP(i,j,k) = max
    return BP
```

Having computed the maximum number of base pairs $BP(1, n, k)$ (minimum energy $-BP(1, n, k)$) over all secondary structures having exactly k hairpins, a standard backtracking method produces the structures S_k having minimum energy over all structures having k hairpins.

4 Discussion

Formation of hairpins appears to be an important aspect in RNA structure evolution [26], a mechanism that could suggest an explanation for the *pervasively transcribed* RNA of unknown function (i.e. RNA *dark matter*) discovered by the Encode Consortium [38]. In this paper, we have used algebraic combinatorics to compute the asymptotic expected number of hairpins in *saturated* secondary structures of RNA. Though theoretical, our work adds to the growing literature of asymptotic results concerning RNA structure – see for instance [37, 27, 28, 34, 3, 5, 17, 32, 29, 20, 4, 21].

How does the asymptotic espected number of haipins (taken either over saturated structures as in Theorem 3, or taken over all structures as in Nebel [27]) compare with the number of hairpins in available RNA structure databases? We analyzed the average number of hairpins as a function of sequence length for RNA secondary structures in both the STRAND database [1] as well as a collection of D.H. Mathews [25]. The STRAND database consists of 4,666 RNA structures, deriving from the Protein Data Bank [35], Sprinzl's tRNA database [36], Gutell's database [14], etc. Mathews data collection consists of 1,192 RNA structures, deriving from most of the same databases. See [1] resp. [25] for references to those databases that contributed to each collection. Table 6 summarizes the equations of least-squares fit of the STRAND and Mathews' RNA databases, while Figure 4 presents a scatter plot for the STRAND database. Broadly speaking, there is a rough agreement between the asymptotic expected number of hairpins $0.0669564 \cdot n$ in saturated structures with $\theta = 3$ and $p = 3/8$ from Table 2 and the average number $0.0236 \cdot n$ of hairpins from the STRAND database, as given in Table 6. In the web supplement, we present a computation of the asymptotic number of *all* secondary structures and the asymptotic expected number of hairpins, taken over *all* secondary structures, where $\theta = 3$ and $p = 3/8$. Asymptotically, there are $1.22479 \cdot n^{-3/2} \cdot 1.85479^n$ many *arbitrary* secondary structures and $0.0546382 \cdot n + 0.0166011 \cdot \sqrt{n} \cdot \mathcal{N}$ expected hairpins, which is in somewhat better agreement with the value of $0.0236 \cdot n$ from the STRAND database – see Table 3. As previously mentioned in [16] the "distribution of loop sizes and loop degrees seems to be dominated by the combinatorics".

In [41], Weinberg and Nebel apply length-dependent context free grammars in RNA secondary structure prediction, while in [33], Rivas et al. describe stochastic context free grammars that incorporate frequencies of various motifs, such as hairpins, in a length-dependent

Database	Least-squares fit	R^2
STRAND (all RNAs)	$0.0236 \cdot n - 0.5312$	0.8153
STRAND (remove outliers)	$0.0258 \cdot n + 0.2531$	0.9535
Mathews (all RNAs)	$0.0215 \cdot n - 0.1642$	0.9616
Mathews (remove outliers)	$0.0134 \cdot n + 1.0691$	0.6549

Table 6: The average number of hairpins, as a function of sequence length n, computed for the STRAND [1] and Mathews [25] RNA structure collections. Values reported after removal of outliers represent averages taken over at least three distinct RNAs of the same length n.

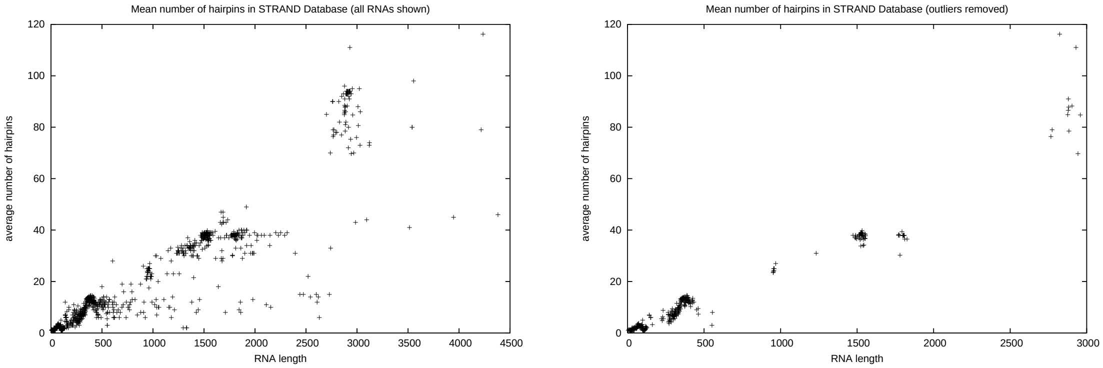

Figure 4: Scatter plot of average number of hairpins in the STRAND database [1], consisting of 4,666 RNA structures. Left panel taken over all data; right panel after removal of outliers, where outliers are defined to be averages taken over less than three distinct RNA sequences for a fixed length n.

manner. With such extensions, Rivas et al. have shown that stochastic context free grammar methods perform comparable to thermodynamics-based methods, such as UNAFOLD [24] and RNAfold [15]. Our new program, described in Algorithm 3.1 can be viewed as a computation of the minimum energy structure, with respect to the Nussinov model, dependent on the assumption that the structure has exactly k hairpins. Seen in this light, there is a loose connection with the work [41, 33]. We expect an even better prediction by extending Algorithm 3.1 to the Turner energy model.

Acknowledgements

We would like to thank D.H. Mathews for generously sharing his RNA data collection, and B. Salvy and E. Fusy for discussions of Drmota's Theorem. Partial support for the research of P. Clote is from NSF grants DMS-0817971 and DBI-1262439. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Funding for the research of E. Kranakis was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Mathematics of Information Technology and Complex Systems (MITACS).

Appendix

Computing the number of hairpins in saturated structures

To produce Figure 1, we computed by dynamic programming the expected number of hairpins in saturated structures for a homopolymer of size n. In the interests of brevity, we must refer the interested reader to [3] for background material on recurrence relations for the number of saturated structures. The recurrence relations require the auxiliary notion of saturated structure *with no visible positions*, defined as follows. A secondary structure S on sequence $a_1, \ldots, a_n$ has no visible positions, if for all $1 \le i \le n$ in which a_i is unpaired, there is no base pair (x, y) for which $x < i < y$.

Let $D(n, k)$ denote the number of saturated secondary structures having exactly k hairpins. Let $E(n, k)$ denote the number of saturated secondary structures having exactly k hairpins, which have no visible positions. Define $D(0, 0) = D(1, 0) = D(2, 0) = D(3, 0) = 1$ and $E(0, 0) = E(3, 1) = 1$; for all other values of $0 \le n \le 3$ and $0 \le k \le 3$, let $D(n, k) = E(n, k) = 0$.

The inductive case is given by:

$$\begin{aligned}
D(n,k) &= E(n-1,k) + E(n-2,k) + \sum_{r=1}^{n-2} D(r-1,k-1)D(n-r-1,0) + \\
&\quad \sum_{r=1}^{n-2}\sum_{s=0}^{k-1} D(r-1,s)D(n-r-1,k-s) \\
E(n,k) &= \sum_{r=1}^{n-2} E(r-1,k-1)D(n-r-1,0) + \\
&\quad \sum_{r=1}^{n-2}\sum_{s=0}^{k-1} E(r-1,s)D(n-r-1,k-s).
\end{aligned}$$

Since the justification for these recursion is similar to that of [3], we do not provide further details. These recursions are implemented using dynamic programming to compute the number of saturated structures on a homopolymer of size n having exactly k hairpins. It follows that the expected number of hairpins for a homopolymer of size n is

$$\sum_{k=0}^{n} k \cdot \frac{D(n,k)}{S(n)}$$

where $S(n) = \sum_{k=0}^{n} D(n, k)$ is the total number of saturated structures for a homopolymer of size n. The Python code is available on the web supplement.

Definition of resultant

In the proof of Theorem 3, we compute the *resultant* of two multivariable polynomials. For the benefit of the reader, we define this concept here. For any commutative ring A, indeterminate X and two multivariate polynomials

$$\begin{aligned}
p_1 &= v_n X^n + \cdots + v_1 X + v_0 \\
p_2 &= u_m X^m + \cdots + u_1 X + u_0
\end{aligned}$$

respectively having roots $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_m$ in the algebraic closure of A, the *resultant* of p_1, p_2 with respect to X is defined to be

$$v_n^n u_m^m \prod_{i=1}^{n} \prod_{j=1}^{m} (\alpha_i - \beta_j).$$

In applications, for instance g_1, g_2 could be functions in variables S, R, u, z, but construed to be polynomials over indeterminate R with coefficients from the ring $\mathbb{Z}(z, u, S)$. In such a case, the resultant $Res(g_1, g_2)$ of g_1, g_2 is a polynomial in $\mathbb{Z}[z, u, S]$, whose roots are the z-, u- and S-coordinates of the intersection of curves corresponding to g_1, g_2. Moreover, it is known that there exist polynomials $q_1, q_2 \in \mathbb{Z}[z, u, S][R]$ such that

$$g_1 \cdot q_1 + g_2 \cdot q_2 = Res(g_1, g_2). \tag{7}$$

For more background on resultants, see [19].

References

[1] M. Andronescu, V. Bereg, H. H. Hoos, and A. Condon. RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC. Bioinformatics*, 9:340, 2008.

[2] P. Clote. An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to the Nussinov-Jacobson energy model. *J. Comput. Biol.*, 12(1):83–101, 2005.

[3] P. Clote. Combinatorics of saturated secondary structures of RNA. *J. Comput. Biol.*, 13(9):1640–1657, November 2006.

[4] P. Clote, S. Dobrev, I. Dotu, E. Kranakis, D. Krizanc, and J. Urrutia. On the page number of RNA secondary structures with pseudoknots. *J Math Biol.*, 0(O):O, December 2011.

[5] P. Clote, E. Kranakis, D. Krizanc, and B. Salvy. Asymptotics of canonical and saturated RNA secondary structures. *J. Bioinform. Comput. Biol.*, 7(5):869–893, October 2009.

[6] L. V. Danilova, D. D. Pervouchine, A. V. Favorov, and A. A. Mironov. RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *J. Bioinform. Comput. Biol.*, 4(2):589–596, April 2006.

[7] M. Drmota. Systems of functional equations. *Random Structures Algorithms*, 10(1-2):103–124, 1997.

[8] M. Drmota, É. Fusy, J. Jué, M. Kang, and V. Kraus. Asymptotic study of subcritical graph classes. *SIAM J. Discrete Math.*, 25(4):1615–1651, 2011.

[9] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[10] C. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[11] E. Fusy and P. Clote. Combinatorics of locally optimal RNA secondary structures. *J Math Biol.*, 0(O):O, December 2012.

[12] P. P. Gardner, J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, R. D. Finn, E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, and A. Bateman. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic. Acids. Res.*, 39(Database):D141–D145, January 2011.

[13] S. Griffiths-Jones. mirbase: the microRNA sequence database. *Methods Mol. Biol.*, 342:129–138, 2006.

[14] R.R. Gutell. Collection of small subunit (16 S- and 16 S-like) ribosomal RNA structures. *Nucl. Acids Res.*, 22:3502–3507, 1994.

[15] I.L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31:3429–3431, 2003.

[16] I.L. Hofacker, P. Schuster, and P.F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998.

[17] E. Y. Jin and C. M. Reidys. Asymptotic enumeration of RNA structures with pseudoknots. *Bull. Math. Biol.*, 70(4):951–970, May 2008.

[18] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–3428, 2003.

[19] S. Lang. *Algebra*. Springer Verlage, 2002. Revised 3rd edition.

[20] T. J. Li and C. M. Reidys. Combinatorial analysis of interacting RNA molecules. *Math. Biosci.*, 233(1):47–58, September 2011.

[21] T. J. Li and C. M. Reidys. Combinatorics of RNA-RNA interaction. *J. Math. Biol.*, 64(3):529–556, February 2012.

[22] W. A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *J. Comput. Biol.*, 15(1):31–63, 2008.

[23] T. Lowe and S. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleid Acids Research*, 25(5):955–964, 1997.

[24] N. R. Markham and M. Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.

[25] D. H. Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA.*, 10(8):1178–1190, August 2004.

[26] U. R. Muller and W. M. Fitch. Evolutionary selection for perfect hairpin structures in viral DNAs. *Nature*, 298(5874):582–585, August 1982.

[27] M. E. Nebel. Combinatorial properties of RNA secondary structure. *Journal of Computational Biology*, 9(3):541–573, 2002.

[28] M. E. Nebel. Investigation of the Bernoulli model for RNA secondary structures. *Bull. Math. Biol.*, 66(5):925–964, September 2004.

[29] M. E. Nebel, C. M. Reidys, and R. R. Wang. Loops in canonical RNA pseudoknot structures. *J. Comput. Biol.*, 18(12):1793–1806, December 2011.

[30] N.E. Nebel, C.M. Reidys, and R.R. Wang. Loops in canonical RNA pseudoknot structures. *J. Comput. Biol.*, 18(12):1793–1806, 2011.

[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] C. M. Reidys and R. R. Wang. Shapes of RNA pseudoknot structures. *J. Comput. Biol.*, 17(11):1575–1590, November 2010.

[33] E. Rivas, R. Lang, and S. R. Eddy. A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more. *RNA.*, 18(2):193–212, February 2012.

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

[35] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, and P. E. Bourne. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic. Acids. Res.*, 39(Database):D392–D401, January 2011.

[36] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.*, 26:148–153, 1998.

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

[38] E. Torarinsson, Z. Yao, E. D. Wiklund, J. B. Bramsen, C. Hansen, J. Kjems, N. Tommerup, W. L. Ruzzo, and J. Gorodkin. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res.*, 0(O):O, December 2007.

[39] J. Waldispuhl and P. Clote. Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model. *J. Comput. Biol.*, 14(2):190–215, March 2007.

[40] M. S. Waterman. *Introduction to Computational Biology.* Chapman and Hall/CRC, 1995.

[41] F. Weinberg and N.E. Nebel. Applying length-dependent stochastic context-free grammars to RNA secondary structure prediction. *Algorithms*, 4(4):223–238, 2011.

[42] A. Xayaphoummine, T. Bucher, and H. Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic. Acids. Res.*, 33(Web):W605–W610, July 2005.

[43] 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.

[44] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.