Expected distance between terminal nucleotides of RNA secondary structures.

Peter Clote, Yann Ponty, Jean-Marc Steyaert

Journal of Mathematical Biology manuscript No.
(will be inserted by the editor)

Expected distance between terminal nucleotides of RNA secondary structures

Peter Clote · Yann Ponty · Jean-Marc Steyaert

Received: date / Accepted: date

Abstract In "The ends of a large RNA molecule are necessarily close", *Nucleic Acids Res.*, September 1, 2010, Yoffe et al. used the programs `RNAfold` [resp. `RNAsubopt`] from Vienna RNA Package to calculate the distance between $5'$ and $3'$ ends of the minimum free energy secondary structure [resp. thermal equilibrium structures] of viral and random RNA sequences. Here, the $5'-3'$ distance is defined to be the length of the shortest path from $5'$ node to $3'$ node in the undirected graph, whose edge set consists of edges $\{i, i+1\}$ corresponding to covalent backbone bonds and of edges $\{i, j\}$ corresponding to canonical base pairs. From repeated simulations and using a heuristic theoretical argument, Yoffe et al. conclude that the $5'-3'$ distance is less than a fixed constant, independent of RNA sequence length.

In this paper, we provide a rigorous, mathematical framework to study the expected distance from $5'$ to $3'$ ends of an RNA sequence. We present recurrence relations that precisely define the expected distance from $5'$ to $3'$ ends of an RNA sequence, both for the Turner nearest neighbor energy model, as well as for a simple homopolymer model first defined by Stein and Waterman. We implement dynamic programming algorithms to *compute* (rather than *approximate* by repeated application of Vienna RNA Package) the *expected distance* between $5'$ and $3'$ ends of a given RNA sequence, with respect to the Turner energy model. Using methods of analytical combinatorics, that depend on complex analysis, we prove that the asymptotic expected $5'-3'$ distance $\langle d_n \rangle$ of length n homopolymers is approximately equal to the constant 5.47211, while the asymptotic distance is 6.771096 if hairpins have a minimum of 3 unpaired bases and the probability that any two positions can form a base pair is 1/4. Finally, we analyze the $5'-3'$ distance

Research supported by the Digiteo Foundation, and National Science Foundation grants DMS-0817971, DBI-0543506 and DMS-1016618.

P. Clote
Department of Biology, Boston College, Chestnut Hill, MA 02467, USA. Tel.: +1-617-552-1332. Fax: +1-617-552-2011. E-mail: clote@bc.edu.

Y. Ponty, J.-M. Steyaert
Laboratoire d'Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France. E-mail: {ponty,steyaert}@lix.polytechnique.fr.

for secondary structures from the STRAND database, and conclude that the $5' - 3$ distance is correlated with RNA sequence length.

Source code (python, Maple, Mathematica and C programs) are available at `http://bioinformatics.bc.edu/clotelab/Expected5to3distance/`.

Keywords RNA · Boltzmann partition function · asymptotic combinatorics · dynamic programming

PACS 82.39.Pj · 87.14.gn · 02.10.Ox

Mathematics Subject Classification (2000) 05C30 · 49L20

1 Introduction

Yoffe et al. [24] point out that many biological processes require the effective circularization of linear RNA. In eukaryotes, circularization of messenger RNA is effected by the binding of eukaryotic initiation factor (bound to $5'$-cap of mRNA) with poly-A tail binding protein (bound to $3'$-polyadenylated tail of mRNA) [7]. In contrast, in plant viral mRNAs that lack both a $5'$-cap and $3'$-polyadenylated tail, there are reverse complementary bases in the $5'$ and $3'$ untranslated regions, which permit effective circularization [16,19]. Efficient replication of certain animal viral genomes depends on circularization, effected by the formation of short "panhandles" (15 nt in influenza A virus [15] and 21 nt in yellow fever virus [4]) that are responsible for keeping $5'$ and $3'$ ends close together.

Motivated by the biological importance of circularizing linear RNA, Yoffe et al. [24] provide a heuristic argument that the distance is small between the two ends of single-stranded RNA molecules, assuming there are approximately equal proportions of A, C, G and U. Using Zuker's `mfold` program [25] and the `RNAfold` program [12] from Vienna RNA package, the authors additionally conclude that the $5' - 3'$ distance of both viral RNA and random RNA sequences consisting of $1,000 - 10,000$ nt, is small (15-20 for viral sequences, 12 for random sequences) and appears to be bounded by a constant independent of sequence length.

In this paper, we provide a rigorous mathematical argument for the observations of Yoffe et al. We develop recurrence relations that precisely define the expected distance from $5'$ to $3'$ ends of an RNA sequence, both for the Turner nearest neighbor energy model, as well as for a simple homopolymer model first defined by Stein and Waterman.[1] We implement dynamic programming algorithms to *compute* (rather than *approximate* by repeated minimum free energy computations) the *expected distance* between $5'$ and $3'$ ends of a given RNA sequence, with respect to the Turner energy model. Using methods of analytical combinatorics, that depend on complex analysis, we prove that the asymptotic expected $5' - 3'$ distance $\langle d_n \rangle$ of length n homopolymers is approximately equal to the constant 5.47211. Our proof allows one to compute the asymptotic $5' - 3'$ distance for arbitrary values of θ, which represents the minimum number of unpaired bases in

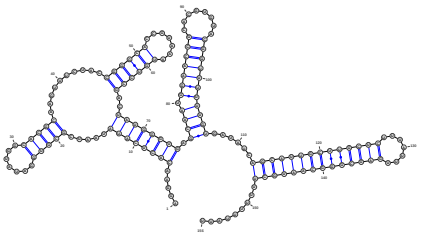
[1] The partition function and recurrence relations we give turn out to be essentially the same as those on page 1326 of Gerland et al. [9]. In that paper, kindly pointed out to us by an anonymous referee, the authors compute the expected number of *external* positions; in contrast, we compute the expected $5' - 3'$ distance, given by expected number of external positions plus twice the number of *components* minus 1. Technically, these values are different, but both methods are almost identical, though independently discovered.

hairpins, and p, a *stickiness* parameter that represents the probability that two positions can form a base pair. For instance, the asymptotic $5' - 3'$ distance is 6.771096 if hairpins have a minimum of 3 unpaired bases and the probability that any two positions can form a base pair is 1/4. Finally, we analyze the $5' - 3'$ distance for secondary structures from the STRAND database, and conclude that the $5' - 3$ distance is correlated with RNA sequence length.

2 Approach

An RNA secondary structure for a given RNA sequence $a_1, \ldots, a_n$ of length n is defined to be a set of unordered pairs consisting of *backbone covalent bonds* $\{i, i+1\}$ for $1 \leq i < n$, and hydrogen-bonded *canonical* base pairs $\{i, j\}$, for $1 \leq i, j \leq n$, which latter satisfy the following four conditions.

1. *Watson-Crick and wobble pairs:* If $\{i, j\} \in S$, then $\{a_i, a_j\} \in \{\{A, U\}, \{G, C\}, \{G, U\}\}$.
2. *Threshold requirement for hairpins:* If $\{i, j\}$ belongs to S, then $j - i > \theta$, for a fixed value of θ; i.e. there must be at least θ unpaired bases in a hairpin loop.
3. *No base triples:* If $\{i, j\}$ and $\{i, k\}$ belong to S, then $j = k$.
4. *Nonexistence of pseudoknots:* If $\{i, j\}$ and $\{k, \ell\}$ belong to S, where $i < j$ and $k < \ell$, then it is not the case that $i < k < j < \ell$.



AGGAACACUCAUAUAAUCGCGUGGAUAUGGCACGCAAGUUUCUACCGGGCACCGUAAAUGUCCGACUAUGGGUGAGCAAUGGAACCGCACGUGUACGGUUUUUUGUGAUAUCAGCAUUGCUUGCUCUUUAUUUGAGCGGGCAAUGCUUUUUUUAUU
.....(((((((((.....(((((.......)))))).........(((((.......)))))).)))))))).((((.(((((((.......)))))))))))......((((((((((((((.......))))))))))))))).....

Fig. 1 *(Top)* Low energy secondary structure of the purine riboswitch *xpt* from *B. subtilis*. Structure produced by applying `RNAfold` constrained by the aptamer from Rfam [8]; graphics produced using VARNA [5]. The shortest path from $5'$ to $3'$ nucleotide comprises 23 covalent bonds and 3 base pairs, hence is 26. *(Bottom)* Same secondary structure presented in Vienna dot bracket notation.

For software, such as `mfold` and `RNAfold`, that computes the minimum free energy (MFE) RNA secondary structure, θ is taken to be 3; i.e. due to steric constraints, every hairpin is required to contain at least three unpaired bases. For asymptotic analysis presented in Section 5 (and *only* in that section), we will not require *(1)* and the value θ in *(2)* can be fixed to any given value, such as 1. We will refer to this simple (theoretical) model as a *homopolymer*, since effectively there is no distinction between different nucleotides. The homopolymer model was first proposed by Stein and Waterman [22], who proved that the asymptotic number of secondary structures of a homopolymer of length n is $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$. See the appendix of Lorenz et al. [17] for an alternative proof of the Stein-Waterman result.

We define the $5' - 3'$ distance in a secondary structure S on $a_1, \ldots, a_n$ to be the *minimum path length* from a_1 to a_n, over all paths consisting of $x_1, \ldots, x_m$, where $x_1 = a_1$, $x_m = a_n$, and $\{x_i, x_{i+1}\} \in S$ for each $1 \leq i < m$. Minimum path length can be efficiently computed Dijkstra's single source minimum path length algorithm [3]; however, there is a much simpler method that is linear in the length of the Vienna dot bracket notation for the secondary structure. (Compute the number of external unpaired positions plus the number of components minus 1, where a component is a substructure designated by a closing, external base pair.)

3 Methods

Let $\mathbf{a} = a_1, \ldots, a_n$ be a given RNA sequence. Given any secondary structure $\mathcal{S}$ of a given RNA sequence $a_1, \ldots, a_n$, we let $\mathcal{S}[i,j]$ denote the *restriction* of $\mathcal{S}$ to interval $[i,j]$, defined by $\mathcal{S}[i,j] = \{\{x,y\} : i \leq x < y \leq j\}$. For $1 \leq i \leq j \leq n$, we define $d_{\mathcal{S}}(i,j)$ to be the minimum path length from i to j in $\mathcal{S}$. Define $D_{i,j}$ by

$$D_{i,j} = \sum_{\mathcal{S} \text{ on } [i,j]} \exp(-E(\mathcal{S})/RT) \cdot d_{\mathcal{S}}(i,j)$$

where the sum is taken over all secondary structures $\mathcal{S}$ of $a_i, \ldots, a_j$, where $E(\mathcal{S})$ is the free energy of secondary structure $\mathcal{S}$, as defined within the Turner nearest neighbor energy model [23], $R \approx 0.00198717$ kcal/mol is the universal gas constant, and T is absolute temperature. Finally, the *expected* distance $\langle d_{1,n} \rangle$ between $5'$ and $3'$ ends of RNA sequence $a_1, \ldots, a_n$ is defined by

$$\langle d_{1,n} \rangle = \sum_{\mathcal{S}} P(\mathcal{S}) \cdot d_S(1,n) = \sum_{\mathcal{S}} \frac{\exp(-E(S)/RT)}{Z_{1,n}} \cdot d_S(1,n) = \frac{D_{1,n}}{Z_{1,n}}$$

where $Z_{1,n}$ is the partition function

$$Z_{1,n} = \sum_{\mathcal{S}} \exp(-E(\mathcal{S})/RT)$$

where the sum is taken over all secondary structures $\mathcal{S}$.

To compute $\langle d_{1,n} \rangle$, we need to compute $D_{1,n}$ and $Z_{1,n}$. A dynamic programming $O(n^3)$ time algorithm to compute the partition function was first described by McCaskill [18] and forms part of the Vienna RNA Package. In the next section, we describe recurrence relations that lead to a dynamic programming algorithm for $D_{1,n}$.

4 Recurrence relations

Recall the definition of $D_{i,j}$

$$D_{i,j} = \sum_{\mathcal{S} \text{ on } [i,j]} \exp(-E(\mathcal{S})/RT) \cdot d_{\mathcal{S}}(i,j) \tag{1}$$

In the case that $j - i \leq \theta$, the only secondary structure on $[i, j]$ is the empty structure having zero free energy, hence it follows by definition that if $j - i \leq \theta$, then $D_{i,j} = j - i$.

We now suppose that $j - i > \theta$. Clearly we can compute $D_{i,j}$ as the sum

$$D_{i,j} = D1_{i,j} + D2_{i,j} + D3_{i,j}$$

where

1. $D1_{i,j}$ is the contribution arising from secondary structures $\mathcal{S}$ on $[i, j]$, in which j is unpaired in $[i, j]$.
2. $D2_{i,j}$ is the contribution arising from secondary structures $\mathcal{S}$ on $[i, j]$, in which i is base-paired with j.
3. $D3_{i,j}$ is the contribution arising from secondary structures $\mathcal{S}$ on $[i, j]$, in which r is base-paired with j, for some intermediate $r \in [i + 1, j]$.

More formally, we have

$$D1_{i,j} = \sum_{\mathcal{S} \text{ on } [i,j],\ j \text{ unpaired in } [i,j]} \exp(-E(\mathcal{S})/RT) \cdot [d_{\mathcal{S}}(i, j-1) + 1]$$

$$D2_{i,j} = \sum_{\mathcal{S} \text{ on } [i,j],\ (i,j) \in \mathcal{S}} \exp(-E(\mathcal{S})/RT) \cdot 1$$

$$D3_{i,j} = \sum_{r=i+1}^{j-\theta-1} \sum_{\mathcal{S} \text{ on } [i,r-1]} \sum_{\mathrm{T} \text{ on } [r,j],\ (r,j) \in T} \exp(-E(\mathcal{S})/RT) \cdot \exp(-E(T)/RT) \cdot [d_{\mathcal{S}}(i, r-1) + 2]$$

Clearly, the first sum can be written as

$$D1_{i,j} = D_{i,j-1} + Z_{i,j-1}.$$

The second sum satisfies

$$D2_{i,j} = ZB_{i,j}$$

where $ZB_{i,j}$ is defined as the partition function over all structures $\mathcal{S}$ on $[i, j]$ in which $(i, j) \in \mathcal{S}$; i.e.

$$ZB_{i,j} = \sum_{\mathcal{S} \text{ on } [i,j],\ (i,j) \in \mathcal{S}} \exp(-E(\mathcal{S})/RT).$$

The third sum satisfies

$$D3_{i,j} = \sum_{r=i+1}^{j-\theta-1} D_{i,r-1} \cdot ZB_{r,j} + \sum_{r=i+1}^{j-\theta-1} 2Z_{i,r-1} \cdot ZB_{r,j}$$

as justified in Figure 2.

Noting that

$$Z_{1,n} = Z_{1,n-1} + ZB_{1,n} + \sum_{r=2}^{n-\theta-1} Z_{1,r-1} \cdot ZB_{r,n}$$

it follows that

$$\begin{aligned} D_{1,n} &= (D_{1,n-1} + Z_{1,n-1}) + ZB_{1,n} + \\ &\quad \sum_{r=2}^{n-\theta-1} (D_{i,r-1} + 2Z_{1,r-1})\, ZB_{r,n} \\ &= D_{1,n-1} + \left(Z_{1,n-1} + ZB_{1,n} + \sum_{r=2}^{n-\theta-1} Z_{1,r-1} \cdot ZB_{r,n} \right) \\ &\quad + \sum_{r=2}^{n-\theta-1} (D_{i,r-1} + Z_{1,r-1})\, ZB_{r,n} \\ &= D_{1,n-1} + Z_{1,n} + \sum_{r=2}^{n-\theta-1} (D_{i,r-1} + Z_{1,r-1})\, ZB_{r,n} \end{aligned}$$

hence

$$\begin{aligned} \langle d_{1,n} \rangle &= \frac{D_{1,n}}{Z_{1,n}} \\ &= \frac{D_{1,n-1} + Z_{1,n} + \sum_{r=2}^{n-\theta-1} (D_{1,r-1} + Z_{1,r-1}) \cdot ZB_{r,n}}{Z_{1,n}} \\ &= \frac{D_{1,n-1} + (2Z_{1,n} - Z_{1,n-1} - ZB_{1,n})}{Z_{1,n}} + \\ &\quad \frac{\sum_{r=2}^{n-\theta-1} D_{1,r-1} \cdot ZB_{r,n}}{Z_{1,n}}. \end{aligned} \tag{2}$$

This latter recurrence admits a dynamic programming solution, for which we can compute in time $O(n^3)$ and space $O(n^2)$ the expected distance between $5'$ and $3'$ ends of an RNA secondary structure with respect to the Turner energy model. Notice that $d_{\mathcal{S}}(i,i) = 0$, $D_{i,i} = 0$, $D_{i,i+1} = 1$.

Turning to the homopolymer case, where we can simplify the energy model to consist of 0 for each base pair (equivalently, using the energy model of [20] with infinite temperature T). In this case, $Z_{1,n}$ equals simply the number N_n of structures on a homopolymer of length n, and $ZB_{1,n}$ equals the number NB_n of structures on a homopolymer, in which the first and last position are paired together. Moreover, for $n \leq \theta+1$, $NB_n = 0$, while for $n \geq \theta+2$, $NB_n = N_{n-2}$, since each such structure for a homopolymer of length $n-2$, when furnished with an additional enclosing base pair, constitutes a structure containing base pair $(1,n)$.

Using these observations, it follows that for $n \geq \theta + 2$,

$$\begin{aligned} D_n &= (D_{n-1} + 2N_n - N_{n-1} - NB_n) + \sum_{k=1}^{n-\theta-2} D_k \cdot NB_{n-k} \\ &= D_{n-1} + (2N_n - N_{n-1} - N_{n-2}) + \sum_{k=1}^{n-\theta-2} D_k \cdot N_{n-k-2} \end{aligned}$$

$$
\begin{aligned}
D1_{i,j} &= D_{i,j-1} + Z_{i,j-1} \\
D2_{i,j} &= ZB_{i,j} \\
D3_{i,j} &= \sum_{r=i+1}^{j-\theta-1} \sum_{\mathcal{S} \text{ on } [i,r-1]} (\exp(-E(\mathcal{S})/RT) \cdot d_{\mathcal{S}}(i,r-1)) \\
&\quad \cdot \sum_{T \text{ on } [r,j],\, (r,j) \in T} \exp(-E(T)/RT) \\
&\quad + \sum_{r=i+1}^{j-\theta-1} \sum_{\mathcal{S} \text{ on } [i,r-1]} \exp(-E(\mathcal{S})/RT) \cdot \sum_{\mathcal{T} \text{ on } [r,j],\, (r,j) \in \mathcal{T}} \exp(-E(\mathcal{T})/RT) \cdot 2 \\
&= \sum_{r=i+1}^{j-\theta-1} D_{i,r-1} \cdot ZB_{r,j} + \sum_{r=i+1}^{j-\theta-1} 2Z_{i,r-1} \cdot ZB_{r,j} \\
D_{i,j} &= (D_{i,j-1} + Z_{i,j-1}) + ZB_{i,j} + \sum_{r=i+1}^{j-\theta-1} (D_{i,r-1} \cdot ZB_{r,j}) + 2 \sum_{r=i+1}^{j-\theta-1} (Z_{i,r-1} \cdot ZB_{r,j}) \\
&= Z_{i,j} + D_{i,j-1} + \sum_{r=i+1}^{j-\theta-1} (D_{i,r-1} + Z_{i,r-1}) \cdot ZB_{r,j}
\end{aligned}
$$

Fig. 2 Recurrence relations for $D1_{i,j}$, $D2_{i,j}$, $D3_{i,j}$ and for $D_{i,j} = D1_{i,j} + D2_{i,j} + D3_{i,j}$, with respect to the Turner nearest neighbor energy model [23].

For the base case, $D_0 = 0$, $D_1 = 0$, $D_2 = 1$. Recalling that $NB_n = 0$ for $0 \leq n \leq \theta + 1$, it follows that for $2 \leq n \leq \theta + 1$,

$$D_n = D_{n-1} + 2N_n - N_{n-1}.$$

Putting everything together, we have the recursions given in Figure 3. Finally, if we redefine $D_0 = -1$, then the last equation of Figure 3 yields the simple recurrence relation

$$D_n = D_{n-1} + 2N_n - N_{n-1} + \sum_{k=0}^{n-\theta-2} D_k \cdot N_{n-k-2}. \tag{3}$$

valid for all $n \geq 1$.

5 Asymptotic analysis

In this section, we derive the asymptotic expected $5' - 3'$ distance in the RNA homopolymer model, where the minimum number θ of unpaired bases in a hairpin loop is fixed, but arbitrary. Our proof is self-contained, modulo knowledge of DSV methodology, explained in detail in Lorenz et al. [17], and modulo the major work of Flajolet and Sedgewick [6] on analytical combinatorics. The proof, which uses the notion of *marked, weighted grammar*, is simple, elegant, and provides notational

$$D_n = \begin{cases} 0 & n = 1 \\ 1 & n = 2 \\ D_{n-1} + 2N_n - N_{n-1} & n = 2, \ldots, \theta + 1 \\ D_{n-1} + (2N_n - N_{n-1} - N_{n-2}) + \sum_{k=1}^{n-\theta-2} D_k \cdot N_{n-k-2} & n \geq \theta + 2 \end{cases} \quad (4)$$

$$N_n = \begin{cases} 1 & n = 0, 1, \ldots, \theta + 1 \\ N_{n-1} + N_{n-2} + \sum_{k=1}^{n-\theta-2} N_k \cdot N_{n-k-2} & n \geq \theta + 2 \end{cases} \quad (5)$$

Fig. 3 Recurrence relation for to compute D_n, the homopolymer version of the relation defined in Equation (1) and in Figure 2. Formally, $D_n = \sum_{\mathcal{S} \text{ on } [1,n]} \exp(-E(\mathcal{S})/RT) \cdot d_{\mathcal{S}}(i,j)$, where in the case of the homopolymer, $E(\mathcal{S})$ equals -1 times the number of base pairs in $\mathcal{S}$. Additional recurrence for N_n, due to Stein and Waterman [22], is given here for convenience of the reader.

flexibility which permits us to compute the asymptotic expected $5' - 3'$ distance in the somewhat more realistic RNA model, where the probability that any two positions can form a base pair is given by a fixed probability p. (The value p, explained below, is called a *stickiness parameter*.)

First, we need some definitions. If $\mathcal{S}$ is a secondary structure on the homopolymer $[1, n]$, then a position $1 \leq x \leq n$ is *external in* $\mathcal{S}$ if there is no base pair $(i, j) \in \mathcal{S}$ such that $i \leq x \leq j$. In contrast, a position x is *exterior* if there is no base pair $(i, j) \in \mathcal{S}$ such that $i < x < j$. Exterior positions (or bases) are either external or are part of a closing base pair that defines a *component* of $\mathcal{S}$. For instance, in the structure $\bullet\,(\,\bullet\,)\,\bullet\,(\,\bullet\,)\,\bullet$, there are three external positions (bases 1,5,9), two components defined by base pairs $(2, 4)$ and $(6, 8)$, and 7 exterior bases (1,2,4,5,6,8,9). Letting $X(\mathcal{S})$ denote the number of *external* bases, $E(\mathcal{S})$ the number of *exterior* bases, and $J(\mathcal{S})$ the number of *components*, it is clear that for the $5' - 3'$ distance $D(\mathcal{S})$, we have $D(\mathcal{S}) = X(\mathcal{S}) + 2J(\mathcal{S}) - 1 = E(\mathcal{S}) - 1$.

A context-free grammar that marks exterior bases

We define the following context-free grammar G, which generates the collection of all secondary structures, in which hairpins have a minimum θ of unpaired bases, and in which exterior bases are *marked*. In the RNA structures generated by G, external unpaired bases are depicted by open circles, exterior base pairs are depicted by square brackets, while non-external, unpaired bases are depicted by filled circles and non-exterior base pairs by parentheses. Formally, we define the context-free grammar G with non-terminal symbols S_θ, R_θ, terminal symbols $\bullet, (,), \circ, [,]$, where S_θ is the start symbol, the symbols $\circ, [,]$ denote exterior bases and symbols $\bullet, (,)$ denote non-exterior bases. The rules, or productions, of G are as follows:

$$\begin{aligned} S_\theta &\to [R_\theta] S_\theta \mid \circ\, S_\theta \mid \varepsilon \\ R_\theta &\to \bullet^\theta \mid \bullet\, R_\theta \mid \{(R_\theta)\bullet^*\}^+ \end{aligned} \quad (6)$$

Here, $\{, \}$ are syntactic brackets, $*$ is Kleene's star (iteration) operator, and $+$ is the non-empty iteration operator. By using these well-known regular operators,

$*, +$, we have given a clear and succinct formulation for a (meta-) grammar that generates all RNA secondary structures. See the text of Hopcroft and Ullman [14] for more on regular operators and context-free grammars.

In the rules above, the empty sequence is denoted by ε, the symbol $\bullet^\theta = \bullet \cdots \bullet$, where there are θ occurrences of $\bullet$, and the expression $\bullet^*$ denotes zero or more occurrences of $\bullet$. The *initial* (or *start*) non-terminal, S_θ, generates all secondary structures in which any hairpin loop must have at least θ un unpaired bases, where exterior bases are marked with dedicated letters $\circ, [,]$. The symbol $\circ$ designates an unpaired, *external* base, while [,] designate an *exterior* base pair, i.e. the left and right base-paired positions that define a component. Finally, non-terminal R_θ generates all secondary structures of length at least θ, in which any hairpin loop must have at least θ unpaired bases.

Before continuing, we mention that an equivalent, but less concise grammar G' that does not use the (meta-) iterators $*, +$ is given by the following rules:

$$\begin{aligned} S_\theta &\to [R_\theta]S_\theta \mid \circ S_\theta \mid \varepsilon \\ T_\theta &\to (R_\theta)T_\theta \mid \bullet T_\theta \mid \varepsilon \\ R_\theta &\to (R_\theta)T_\theta \mid \bullet R_\theta \mid \bullet^\theta . \end{aligned} \tag{7}$$

In grammar G', the start symbol is S_θ, which generates all (marked) secondary structures, in which any hairpin must have at least θ unpaired bases. The symbol R_θ generates all secondary structures, of length at least θ, while the symbol T_θ is an auxilliary symbol for the grammar.

A straightforward proof by induction on lengths of expressions (left to the reader) establishes that the context free grammar G defined in (6) is unambiguous and satisfies the syntactical conditions of Theorem VII.5 on page 483 of [6].

Using DSV methodology on the grammar (6), we have that $T_\theta(z) = \frac{1}{1-z-z^2R_\theta(z)}$, hence

$$\begin{aligned} S_\theta(z,u) &= z^2u^2R_\theta(z)S_\theta(z,u) + zuS_\theta(z,u) + 1 \\ R_\theta(z) &= \frac{z^2R_\theta(z)}{1-z-z^2R_\theta(z)} + zR_\theta(z) + z^\theta \end{aligned}$$

Thus $R_\theta(z)$ is the solution of the equation

$$z^2(1-z)R_\theta^2(z) - \left(1 - 2z + z^{\theta+2}\right) R_\theta(z) + z^\theta(1-z) = 0 \tag{8}$$

whose MacLaurin series (i.e. Taylor expansion at $z = 0$) has positive coefficients. Solving the binomial equation, we have

$$R_\theta(z) = \frac{1 - 2z + z^{\theta+2} - \sqrt{\Delta_\theta}}{2z^2(1-z)} \tag{9}$$

with

$$\Delta_\theta = 1 - 4z + 4z^2 - 2z^{\theta+2} + 4z^{\theta+3} - 4z^{\theta+4} + z^{2\theta+4} \tag{10}$$

and hence

$$S_\theta(z,u) = \frac{1}{1 - zu - \frac{u^2\left(1-2z+z^{\theta+2}-\sqrt{\Delta_\theta}\right)}{2(1-z)}}. \tag{11}$$

Without risk of confusion, we write $S_\theta(z) = S_\theta(z,1)$ for the generating function for the (unmarked) collection of all secondary structures, in which all hairpins have at least θ unpaired bases. It follows that we obtain the following expression for the case $\theta = 1$:

$$S_1(z) = \frac{1 - z + z^2 - \sqrt{1 - 2z - z^2 - 2z^3 + z^4}}{2z^2}. \tag{12}$$

By different methods, Stein and Waterman [22] reported equation (12) as the generating function for the collection of secondary structures, in which all hairpins have at least one unpaired base.

The generating function for the expected number $E_{\theta,n}$ of exterior bases in secondary structures of length n can be expressed as a partial derivative of $S_\theta(z,u)$. Before giving this derivation, we need some notation. Let $[z^n u^k]S_\theta(z,u)$ denote the Taylor coefficient $s_{n,k}$ in the MacLaurin series $S_\theta(z,u) = \sum_n \sum_k s_{n,k} u^k z^n$, where $s_{n,k}$ denotes the number of secondary structures of length n with k exterior (marked) bases. Without risk of confusion, we let $[z^n]S_\theta(z)$ denote the Taylor coefficient s_n of z^n in the MacLaurin series $S_\theta(z) = \sum_n s_n z^n$, where s_n denotes the number of secondary structures of length n. Then

$$E_{\theta,n} = \frac{\sum_{k \geq 0} k \cdot s_{n,k}}{[z^n]S_\theta(z)} = \frac{[z^n]\frac{\partial S_\theta(z,u)}{\partial u}|_{u=1}}{[z^n]S_\theta(z)}.$$

Consequently, letting $E_\theta(z) = \frac{\partial S_\theta(z,u)}{\partial u}|_{u=1}$, we have

$$E_\theta(z) = \frac{P(z) - (2 - 5z + 4z^2 - 2z^{\theta+2} + z^{\theta+3})\sqrt{\Delta_\theta}}{2(1-z)^2 z^4} \tag{13}$$

where $P(z)$ is the polynomial

$$2 - 9z + 14z^2 - 8z^3 + 2z^5 + z^{\theta+2}(-4 + 10z - 10z^2 + 2z^3) + z^{2\theta+4}(2 - z).$$

Concerning the asymptotic limit of the Taylor coefficients of $E_\theta(z)$ in the expansion about $z = 0$, we apply Theorem VII.5 from page 483 of [6]. This theorem states that the solutions of an irreducible (positive) CF (context free) schema have a square-root singularity at the corresponding radius of convergence

$$C(z) = \tau - \gamma\sqrt{1 - \frac{z}{\rho}} + O\left(\sqrt{1 - \frac{z}{\rho}}\right) \tag{14}$$

so that

$$C_n \sim \frac{\gamma}{2\sqrt{\pi n^3}} \cdot \rho^{-n}. \tag{15}$$

Let $w = E_\theta(z)$ and let

$$F(z,w) = w \cdot (2(1-z)^2 z^4) - P(z) + (2 - 5z + 4z^2 - 2z^{\theta+2} + z^{\theta+3})\sqrt{\Delta_\theta}$$

Then $F(z,w)$ is equal to the expression obtained by multiplying equation (13) by $2(1-z)^2 z^4$, and rearranging all terms to one side of the equality. It follows that $F(z,w) = 0$. Then Lemma VII.3 on page 469 of [6] determines that the value γ,

referred to in equations (14) and (15), must satisfy $\gamma = \sqrt{\frac{2z_0 F_z(z_0,w_0)}{F_{ww}(z_0,w_0)}}$, where F_z is the partial derivative of F with respect to z, and F_{ww} is the second partial derivative of F with respect to w. In fact, Theorem VII.6 on page 489 of [6] (i.e. the theorem of Drmota, Lalley, and Woods) is even more precise and states that for irreducible positive CF systems, the solutions can be developed as Taylor series in $\sqrt{1-\frac{z}{\rho}}$. Hence there exists a full asymptotic expansion of the coefficients (see [6], p. 489–490). We have thus proved the following.

Theorem 1 *The asymptotic expected $5' - 3'$ distance D_n over all RNA secondary structures of length n, in which hairpins have a minimum of θ unpaired bases, is equal to*

$$D_n = E_{\theta,n} - 1 \sim \frac{2 - 5\rho + 4\rho^2 - 2\rho^{\theta+2} + \rho^{\theta+3}}{(1-\rho)\rho^2} - 1 \tag{16}$$

where $z = \rho$ is the smallest modulus real solution of

$$\Delta_\theta \equiv 1 - 4z + 4z^2 - 2z^{\theta+2} + 4z^{\theta+3} - 4z^{\theta+4} + z^{2\theta+4} = 0. \tag{17}$$

Proof. We obtain

$$[z^n]S_\theta(z) \sim -\frac{1}{(1-\rho)2\rho^2}[z^n]\sqrt{\Delta_\theta} \tag{18}$$

$$[z^n]E_\theta(z) \sim -\frac{\left(2 - 5\rho + 4\rho^2 - 2\rho^{\theta+2} + \rho^{\theta+3}\right)}{2(1-\rho)^2\rho^4}[z^n]\sqrt{\Delta_\theta}. \tag{19}$$

Taking the ratio of the two above expressions gives the expected number of exterior bases, from which we subtract 1 to obtain D_n as stated in Equation 16. □ Sample values of ρ and $\langle d_n \rangle$ are given for $\theta = 1, 2, 3$:

$$\begin{aligned} \theta = 1, \rho \approx 0.38196601, D_n \approx 5.47213595 \\ \theta = 2, \rho \approx 0.41421356, D_n \approx 4.65685424 \\ \theta = 3, \rho \approx 0.43691112, D_n \approx 4.15517996 \end{aligned} \tag{20}$$

See Table 1 for additional values.

Boltzmann-weighted ensemble and stickiness

A strength of the current grammar-based method is its robustness to the adjunction of *Boltzmann-like* probability distributions. Following Hofacker et al. [13], we define the *stickiness* $\mathbf{w} = 2\,(p_A \cdot p_U + p_G \cdot p_U + p_G \cdot p_C)$, where p_A (resp. p_C, p_G, p_U) is the mononucleotide frequency of A (resp. C, G, U). In random RNA with compositional frequency given by p_A, p_C, p_G, p_U, the probability that any two positions i, j can base-pair is equal to the stickiness parameter p. We model stickiness by associating weight $\mathbf{w} > 0$ with each base pair, and by defining the weight $w(\mathcal{S})$ of a secondary structure $\mathcal{S}$ to be the *product* of the weights of its base pairs; i.e. $w(\mathcal{S}) := \mathbf{w}^{bp(\mathcal{S})}$, where $bp(\mathcal{S})$ is the number of base pairs in $\mathcal{S}$. We define a probability distribution where each secondary structure $\mathcal{S}$ has probability $P(\mathcal{S}) = w(\mathcal{S})/Z_{|\mathcal{S}|}$, where Z_n is the cumulative weight over all structures of length n.

While the grammar (6) remains the same, its translation into a system of equations now includes a weight increment $\mathbf{w}$ for each base pair:

$$\begin{aligned} S_{\theta,\mathbf{w}}(z,u) &= \mathbf{w}z^2u^2R_\theta(z)S_\theta(z) + zuS_\theta(z,u) + 1 \\ R_\theta(z) &= z^\theta + zR_\theta(z) + \frac{\mathbf{w}z^2R_\theta(z)}{1-z} + \frac{\mathbf{w}z^2R_\theta(z)}{1-z} \cdot \frac{1}{1-\frac{\mathbf{w}z^2R_\theta(z)}{1-z}}. \end{aligned}$$

Note that the coefficient $[z^n]S_{\theta,\mathbf{w}}(z)$ is no longer the number of secondary structures of length n, but rather their cumulative weight.

Solving the system gives the following positive solutions for $S_{\theta,\mathbf{w}}(z)$ and $E_{\theta,\mathbf{w}}(z) = \frac{\partial S_{\theta,\mathbf{w}}(z,u)}{\partial u}|_{u=1}$:

$$\begin{aligned} E_{\theta,\mathbf{w}}(z) &= \frac{\Omega_{\theta,\mathbf{w}}(z) - \Phi_{\theta,\mathbf{w}}(z) \cdot \sqrt{\Delta_{\theta,\mathbf{w}}}}{4(z-1)^3\mathbf{w}^2z^4} \\ S_{\theta,\mathbf{w}}(z) &= \frac{1-2z+(\mathbf{w}+1)z^2 - \mathbf{w}z^{\theta+2} - \sqrt{\Delta_{\theta,\mathbf{w}}}}{(1-z)\mathbf{w}2z^2} \\ \Delta_{\theta,\mathbf{w}} &:= 1-4z+(6-2\mathbf{w})z^2+4(\mathbf{w}-1)z^3+(\mathbf{w}-1)^2z^4 \\ &\quad -2\mathbf{w}z^{\theta+2}+4\mathbf{w}z^{\theta+3}-2\mathbf{w}(1+\mathbf{w})z^{\theta+4}+\mathbf{w}^2z^{2\theta+4}, \\ \Phi_{\theta,\mathbf{w}}(z) &:= 4-14z-18z^2+(2\mathbf{w}-10)z^3+(2-2\mathbf{w})z^4-4\mathbf{w}z^{\theta+2} \\ &\quad +6\mathbf{w}z^{\theta+3}-2\mathbf{w}z^{\theta+4}, \\ \Omega_{\theta,\mathbf{w}}(z) &:= 4-22z-(4\mathbf{w}-50)z^2-(60-16\mathbf{w})z^3-(24\mathbf{w}-40)z^4 \\ &\quad -(14-16\mathbf{w}-2\mathbf{w}^2)z^5+2(1-2\mathbf{w}-\mathbf{w}^2)z^6-8\mathbf{w}z^{\theta+2}+28\mathbf{w}z^{\theta+3} \\ &\quad -4\mathbf{w}(9+\mathbf{w})z^{\theta+4}+4\mathbf{w}(5+\mathbf{w})z^{\theta+5}-4\mathbf{w}z^{\theta+6}+4\mathbf{w}^2z^{2\theta+4} \\ &\quad -6\mathbf{w}^2z^{2\theta+5}+2\mathbf{w}^2z^{2\theta+6}. \end{aligned}$$

Once again, the expected number of exterior bases is given by

$$\frac{[z^n]\frac{\partial S_{\theta,\mathbf{w}}(z,u)}{\partial u}|_{u=1}}{[z^n]S_{\theta,\mathbf{w}}(z)} = \sum_{s\in S_{\theta,n}} \frac{E(s)}{Z_n} = \frac{\sum_{k\geq 0} k \cdot s_{n,k}}{Z_n} = E_{\theta,n}.$$

Now, let D_n denote the expected $5'-3'$ distance over all secondary structures of a length n sequence, where hairpins are required to contain at least θ unpaired bases, and for which the *stickiness* $\mathbf{w} = 2\,(p_A \cdot p_U + p_G \cdot p_U + p_G \cdot p_C)$. Using the same argument used in the proof of Theorem 1, we have the following.

Theorem 2 *Let ρ be the root of $\Delta_{\theta,\mathbf{w}}$ having smallest modulus, where $\Delta_{\theta,\mathbf{w}}$ is a polynomial in variable z, defined by*

$$\begin{aligned} &1-4z+(6-2\mathbf{w})z^2+4(\mathbf{w}-1)z^3+(\mathbf{w}-1)^2z^4 \\ &-2\mathbf{w}z^{\theta+2}+4\mathbf{w}z^{\theta+3}-2\mathbf{w}(1+\mathbf{w})z^{\theta+4}+\mathbf{w}^2z^{2\theta+4}. \end{aligned}$$

Then the asymptotic expected $5'-3'$ distance D_n over all RNA secondary structures of length n with stickiness parameter $\mathbf{w}$, in which hairpins have a minimum of θ unpaired bases, is given by

$$D_n \sim \frac{\Phi_{\theta,\mathbf{w}}(\rho)}{4(\rho-1)^3\mathbf{w}^2\rho^4} - 1. \tag{21}$$

Table 1 presents sample values for different parameters of θ and of *stickiness parameter* p.

θ	p	$\langle d_n \rangle$
1	1/1	5.472136
1	3/8	7.065852
1	1/4	7.928203
2	1/1	4.656854
2	3/8	6.338835
2	1/4	7.246211
3	1/1	4.155180
3	3/8	5.849480
3	1/4	6.771096

Table 1 Values for the asymptotic expected $5'-3'$ distance for various values of $\theta \in [1,3]$ and stickiness p. Expected distance is computed over all secondary structures, in which hairpins have at least θ unpaired bases, and for which any two positions can form a base pair with stickiness probability $p = 2(p_C \cdot p_G) + 2(p_A \cdot p_U) + 2(p_G \cdot p_U)$, where p_A (resp. p_C, etc.) is the mononucleotide frequency.

6 Discussion

Yoffe et al. [24] observe that it appears that certain RNA molecules (viral genomes, certain messenger RNAs) have been under selective pressure to maintain a small distance between the $5'$ and $3'$ ends of the molecule [16,19]. The authors showed this result by repeatedly executing Vienna RNA Package `RNAfold` and `RNAsubopt` on very long RNA sequences (both random and viral RNA). In addition, Yoffe et al. provided a heuristic argument that the $5' - 3'$ distance is small, bounded by a constant independent of sequence length.

Hofacker et al. [13] used a different approach than that described in this paper in order to compute the asymptotic expected number of *external* bases E_n/S_n and the expected number I_n/S_n of components, both for length n secondary structures for the homopolymer model. From Table 3 on page 24 of [13], we have in the case of $\theta = 1$ that $E_n/S_n \sim 2$. From Theorem 14 on page 15 of [13], we have that $I_n/S_n \sim \frac{2}{\alpha} - 3$, where α is the dominant singularity (denoted by ρ in our paper). For $\theta = 1$, $\rho \sim 0.38196601$, so $I_n/S_n \sim 2.23606798$. It follows that

$$E_n/S_n + 2 \cdot I_n/S_n - 1 \sim 2 + 2 \cdot 2.23606798 - 1 = 5.472135$$

which is what we derived in equation (20). Since there seems to be a mistake in Table 3 of [13] for the value of I_n/S_n, as well as typos in Theorem 4.23 and equation (84) on page 19 of [13],[2] use of results from [13] may not be unproblematic in deriving results of our paper.

In this paper, we have described recurrence relations, displayed in Figure 2, for the exact computation of the expected $5'-3'$ distance over all secondary structures of a given RNA sequence. Similar methods could have been used to compute the expectation of the second moment, hence variance, as well as higher moments for the distribution of $5'-3'$ distances. The recurrence relations in Figure 2 are essentially the same as those previously developed by Gerland et al. [9], who computed

[2] The proportion $J_n(b)/S_n$ of structures of length n having exactly b components is computed in Theorem 4.23 of [13]. The sum over all values of b, $0 \leq b \leq n/2$ should equal 1.0; however, the sum is 1.618033. Presumably the expression $\frac{x^2}{1-x}$ in equation (84) should be raised to the power b, while $J_n(b)/S_n$ in the statement of Theorem 4.23 should equal $(\alpha^2 * b/(1-\alpha)^2) * ((1-2*\alpha)/(1-\alpha))^(b-1)$; i.e. there is a missing $1/(1-\alpha)$ factor.

the expected expected number of *external* positions in the secondary structures of an RNA sequence. The work of Gerland et al. was done to theoretically explain experimental force-extension curves obtained by single-molecule experiments, where the ends of an RNA molecule are pulled apart by optical tweezers. The recurrence relations are almost the same, though technically different, since Gerland et al. compute the expected number of *external* positions, while we compute the expected number of *external* positions plus twice the number of *components* minus 1, which latter gives the expected $5' - 3'$ distance.

In contrast to the repeated simulations of [24], we have implemented the recurrence relations of Figure 2 in a dynamic programming algorithm in C. Considering the analogous recurrence relations for the (theoretical) RNA homopolymer model of [22], we have used methods from complex analysis to determine the asymptotic expected $5' - 3'$ distance. In particular, Table 1 presents some numerical values for the asymptotic $5' - 3'$ distance for different choices of θ, the minimum number of unpaired bases in a hairpin loop, and p, the probability that any two positions can form a base pair.

The work of Yoffe et al. [24] and the exact results obtained by our dynamic programming algorithm and theoretical analysis show only that the *average* $5' - 3'$ distance is small and independent of sequence length, where the average is taken over all secondary structures in the Boltzmann ensemble (resp. uniform ensemble). Nevertheless, this does not prove that biologically functional RNA molecules (viral genomes, certain messenger RNAs) have small $5' - 3'$ distance, independendent of sequence length. Indeed, one cannot identify *native secondary structure* with the *minimum energy structure*, since benchmarking studies have shown that for RNA sequences of 700 nt or less, *at most* 73% of base pairs are correctly predicted by free energy minimization methods (`mfold`, `RNAfold`). Moreover, it is known that accuracy is increasingly lost as sequence length increases. It follows that $5' - 3'$ distance, computed by applying `RNAfold` or `RNAsubopt` to long RNA sequences (viral genomes and random RNA) [24], may be quite different that by analyzing the *real* RNA secondary structures, derived from X-ray diffraction (gold standard) or by *comparative sequence analysis* (e.g. [10], *silver standard*). STRAND [1] is a database of 4666 RNA secondary structures, obtained from X-ray structures taken from the Nucleic Acid Database [11], the Protein Data Bank [2] and from secondary structures derived from comparative sequence analysis [21, 10], etc. – see [1] for citation of original data sources.

We determined the $5' - 3'$ distance of all *secondary structures* from STRAND, obtaining a mean of 27.56, standard deviation of 77.29, maximum of 2086, minimum of 1. The Spearman rank correlation between RNA sequence length and $5' - 3'$ distance for the STRAND database is 0.644061.[3] In contrast, the average value for expected $5' - 3'$ distance, obtained by running our dynamic programming program implemented in C, with respect to the Turner energy model (see equation 2), yields a value of 4.82201 with standard deviation of 2.76071.

While the STRAND database depicts secondary structure base pairs by parentheses (Vienna dot bracket notation), the base pairs from pseudoknots are represented by angle brackets, curly brackets, etc. If we recompute the Spearman rank

[3] Pearson correlation coefficient for the data was 0.2612; however, the distribution of $5' - 3'$ distances is *not* normal – see Figure 5. For this reason, it is more appropriate to use the (parametric) Spearman rank correlation coefficient.

	Min	Q_1	Median	Mean	Q_3	Max
native	1.00	4.00	5.00	18.83	31.00	1491.00
`RNAfold -C`	1.00	4.00	5.00	22.12	32.00	1491.00
`RNAfold`	1.00	4.00	5.00	12.47	13.00	315.00

Table 2 Statistics on the $5' - 3'$ distance for the STRAND [1], a collection of 4666 RNA secondary structures. Computation of minimum, first quartile (Q_1), median, mean, third quartile (Q_3), and maximum for *(i)* native structures, i.e. the secondary structures given in STRAND database, *(ii)* for the output of `RNAfold` using the STRAND structure as a constraint, and *(iii)* simply using `RNAfold` on the sequence from the STRAND database. We also computed statistics after removing outlier sequences from the STRAND database, i.e. those for which more than 75% of the nucleotides were unpaired. Statistics (shown in supplementary data) are comparable to those before removal of outliers.

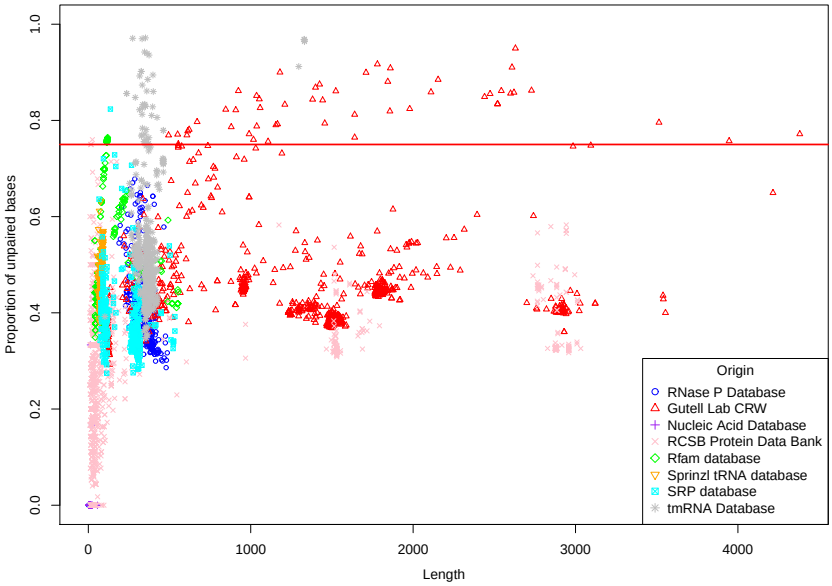


Fig. 4 Scatter plot of the proportion of unpaired nucleotides as as a function of RNA sequence length, for the STRAND database [1]. The average $5' - 3'$ distance for RNA structures from the STRAND is larger than predicted by asymptotic limits, both before and after the removal of outliers – those sequences in which more than 75% of the nucleotides are unpaired. See Figure 5 for the scatter plot of average $5' - 3'$ distance of RNA structures from the STRAND database, after removal of outliers.

correlation between sequence length and $5' - 3'$ distance of all (pseudoknotted) structures[4], we obtain a mean of 16.17, a standard deviation of 38.64, a maximum of 1478, a minimum of 1, and a Spearman correlation of 0.656986. Thus, regardless of whether pseudoknot base pairs are considered when computing $5' - 3'$ distance, there appears to be a correlation between RNA sequence length and $5' - 3'$ distance of *real* secondary structures, rather than *computed* secondary structures. Figure 5 depicts the scatter plot for $5' - 3'$ distance of (pseudoknotted) structures as a function of RNA sequence length. Table 2 presents statistics for the $5' - 3'$ distance for secondary structures from the STRAND database.

[4] For all base pairs (i, j), remove positions $i+1, \ldots, j-1$, then count the number of positions remaining, and return that value plus 1.

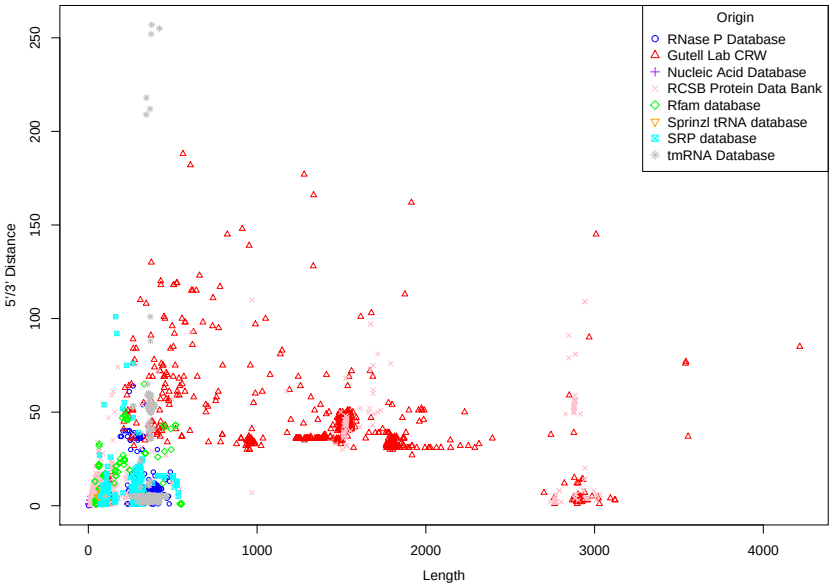


Fig. 5 Scatter plot of $5' - 3'$ distance, as a function of RNA sequence length, for the STRAND database [1]. Since some of the 4666 structures in STRAND are only partly determined, structures were removed if than 75% of their bases were unpaired. This resulted in 4541 remaining structures. Since many of the structures (about 1 out of 4) contain a pseudoknot, Dijkstra's shortest path algorithm was used to determine $5' - 3'$ distance. (The resulting distance is generally less than the $5' - 3'$ distance when no pseudoknots are considered.) Spearman correlation coefficient between sequence length and (pseudoknotted) $5' - 3'$ distance is 0.65.

7 Conclusion

In this paper, we have given a rigorous, mathematical framework to study the expected distance from $5'$ to $3'$ ends of an RNA sequence. We have presented recurrence relations that precisely define the expected distance from $5'$ to $3'$ ends of an RNA sequence, both for the Turner nearest neighbor energy model, as well as for a simple homopolymer model first defined by Stein and Waterman. We have implemented programs in C and Python, with source code made available at our web site, to *compute* (rather than *approximate* by repeated minimum free energy computations) the *expected distance* between $5'$ and $3'$ ends of a given RNA sequence, with respect to the Turner energy model. Using methods of analytical combinatorics, that depend on complex analysis, we have rigorously proven that expected $5' - 3'$ distance is asymptotically bounded by a constant, independent of sequence length. We have determined numerical values for the asymptotic $5' - 3'$ distance for various values of θ as well as for stickiness parameter p.

See `http://bioinformatics.bc.edu/clotelab/Expected5to3distance/` for C and Python programs for the dynamic programming algorithms described in Section 4, for Maple code for Section 5, and Mathematica code for work described in the supplement.

Acknowledgements We would like to than an anonymous referee, who kindly pointed out the paper of Gerland et al. [9], as well as W.A. Lorenz for some discussions concerning a

preliminary version of this paper. In particular, W.A. Lorenz pointed out a simplification in obtaining equation (3). Research of P. Clote and J.-M. Steyaert was supported by the Digiteo Foundation. Additional funding was provided by the National Science Foundation grants DMS-0817971 and DBI-0543506 to PC. 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.

References

1. Andronescu, M., Bereg, V., Hoos, H.H., Condon, A.: RNA STRAND: the RNA secondary structure and statistical analysis database. BMC. Bioinformatics **9**, 340 (2008)
2. Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C.: The Protein Data Bank. Acta Crystallogr. D. Biol. Crystallogr. **58**(Pt), 899–907 (2002)
3. Cormen, T., Leiserson, C., Rivest, R.: Algorithms. McGraw-Hill (1990). 1028 pages
4. Corver, J., Lenches, E., Smith, K., Robison, R.A., Sando, T., Strauss, E.G., Strauss, J.H.: Fine mapping of a cis-acting sequence element in yellow fever virus RNA that is required for RNA replication and cyclization. J. Virol. **77**(3), 2265–2270 (2003)
5. Darty, K., Denise, A., Ponty, Y.: VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics **25**(15), 1974–1975 (2009)
6. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University (2009). ISBN-13: 9780521898065
7. Gallie, D.R.: The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. Genes Dev. **5**(11), 2108–2116 (1991)
8. Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R., Bateman, A.: Rfam: updates to the RNA families database. Nucleic. Acids. Res. **37**(Database), D136–D140 (2009)
9. Gerland, U., Bundschuh, R., Hwa, T.: Force-induced denaturation of RNA. Biophys. J. **81**, 1324–1332 (2001)
10. Gutell, R., Lee, J., Cannone, J.: The accuracy of ribosomal RNA comparative structure models. Current Opinion in Structural Biology **12**, 301–310 (2005)
11. HM, B., J, W., Z, F., L, I., B, S., C, Z.: The nucleic acid database. Methods Biochem Anal. **44**, 199–216 (2003)
12. Hofacker, I.: Vienna RNA secondary structure server. Nucleic Acids Res **31**(13), 3429–3431 (2003)
13. Hofacker, I.L., Schuster, P., Stadler, P.F.: Combinatorics of RNA secondary structures. Discr. Appl. Math. **88**, 207–237 (1998). URL `citeseer.nj.nec.com/1454.html`
14. Hopcroft, J.E., Ullman, J.D.: Formal languages and their relation to automata. Addison-Wesley (1969)
15. Hsu, M.T., Parvin, J.D., Gupta, S., Krystal, M., Palese, P.: Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle. Proc. Natl. Acad. Sci. U.S.A. **84**(22), 8140–8144 (1987)
16. Kneller, E.L., Rakotondrafara, A.M., Miller, W.A.: Cap-independent translation of plant viral RNAs. Virus. Res. **119**(1), 63–75 (2006)
17. Lorenz, W.A., Ponty, Y., Clote, P.: Asymptotics of RNA shapes. J. Comput. Biol. **15**(1), 31–63 (2008)
18. McCaskill, J.: The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers **29**, 1105–1119 (1990)
19. Miller, W.A., White, K.A.: Long-distance RNA-RNA interactions in plant virus gene expression and replication. Annu. Rev. Phytopathol. **44**, 447–467 (2006)
20. Nussinov, R., Jacobson, A.B.: Fast algorithm for predicting the secondary structure of single stranded RNA. Proceedings of the National Academy of Sciences, USA **77**(11), 6309–6313 (1980)
21. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., Steinberg, S.: Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res. **26**, 148–153 (1998)
22. Stein, P.R., Waterman, M.S.: On some new sequences generalizing the Catalan and Motzkin numbers. Discrete Mathematics **26**, 261–272 (1978)

23. Xia, T., J. SantaLucia, J., Burkard, M., Kierzek, R., Schroeder, S., Jiao, X., Cox, C., Turner, D.: Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry **37**, 14,719–35 (1999)
24. Yoffe, A.M., Prinsen, P., Gelbart, W.M., Ben-Shaul, A.: The ends of a large RNA molecule are necessarily close. Nucleic Acids Res. **0**(O), O (2010)
25. Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31(13)**, 3406–3415 (2003)