On the Combinatorics of RNA Secondary Structures in a Polymer-Zeta Model

Markus E. Nebel

based on joint work with **Emma Yu Jin**

CanaDAM 2013

Newfoundland, Canada

Plan of Talk

1. RNA Secondary Structure
   - basic definitions
   - enumeration
   - polymer-zeta model (motivation and definition)
2. Enumeration in the Polymer-Zeta Model
   - fundamentals
   - average number of hairpins
3. Overview of Results and Discussion

RNA Secondary Structure

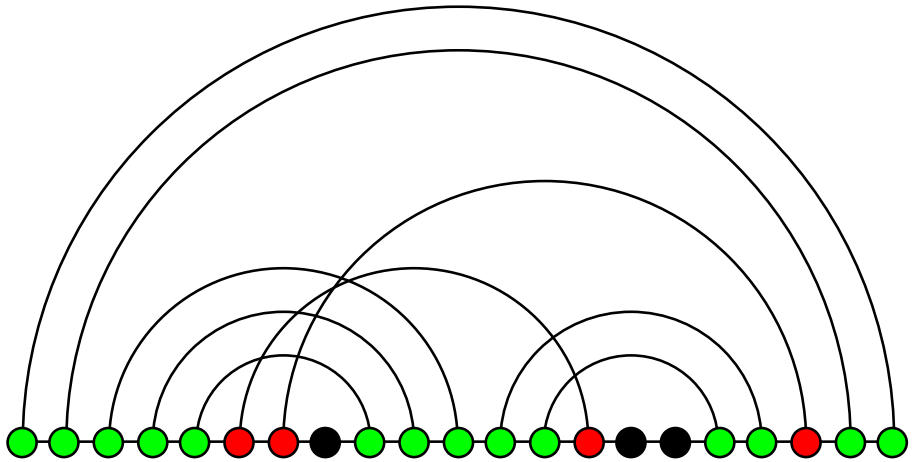
From an abstract point of view, RNA molecules of size n consist of

1. a linear chain of n nodes ($\equiv$ nucleotides) labeled $\{a, c, g, u\}$
   $\rightsquigarrow$ string $s \in \{a, c, g, u\}^n$ called **RNA *sequence***.

RNA Secondary Structure

From an abstract point of view, RNA molecules of size n consist of

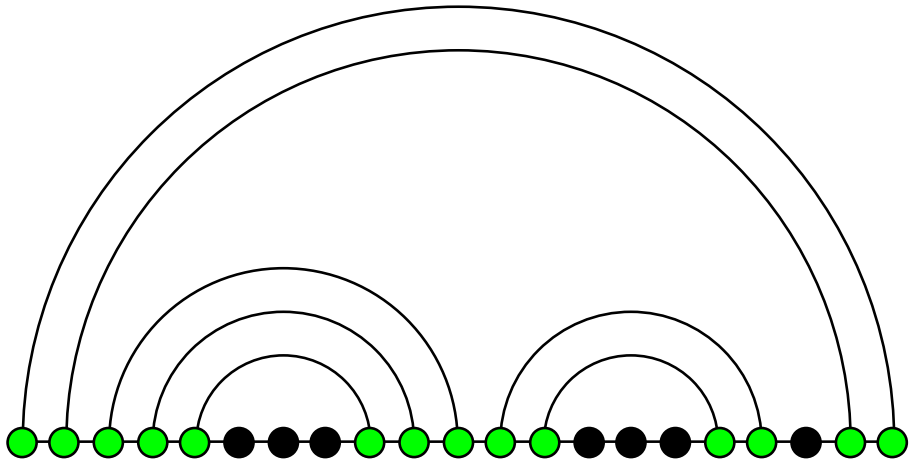
1. a linear chain of n nodes ($\equiv$ nucleotides) labeled $\{a, c, g, u\}$
   $\rightsquigarrow$ string $s \in \{a, c, g, u\}^n$ called **RNA *sequence***.
2. which may be part of **at most one edge** connecting nodes of distance (in the chain) at least 2 (counted by hops).



RNA Secondary Structure

From an abstract point of view, RNA molecules of size n consist of

1. a linear chain of n nodes ($\equiv$ nucleotides) labeled $\{a, c, g, u\}$
   $\rightsquigarrow$ string $s \in \{a, c, g, u\}^n$ called **RNA *sequence***.
2. which may be part of **at most one edge** connecting nodes of distance (in the chain) at least 2 (counted by hops).

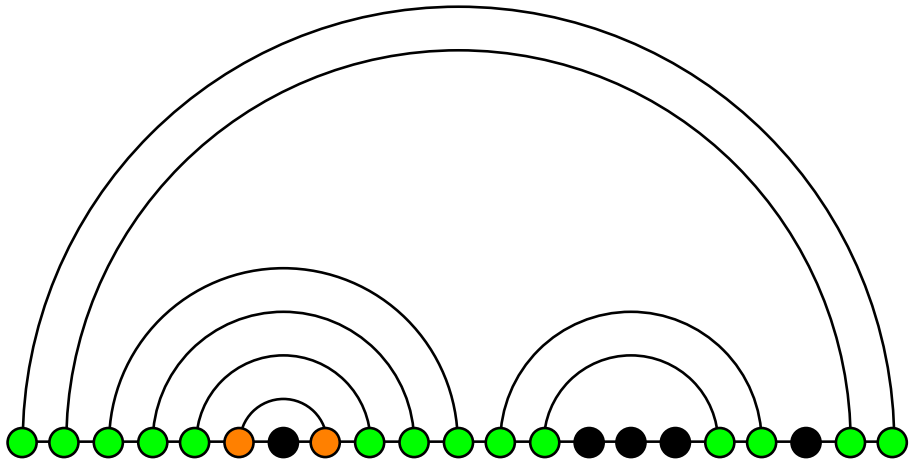


Secondary structure: Edges (arcs) are **not** allowed to **cross**.

RNA Secondary Structure

From an abstract point of view, RNA molecules of size n consist of

1. a linear chain of n nodes ($\equiv$ nucleotides) labeled $\{a, c, g, u\}$
   $\rightsquigarrow$ string $s \in \{a, c, g, u\}^n$ called **RNA *sequence***.
2. which may be part of **at most one edge** connecting nodes of distance (in the chain) at least 2 (counted by hops).



Minimal distance: Edge connecting orange nodes allowed.

Enumeration

Enumerating secondary structures is easy; their number is given by the following recurrence relation:

$$r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1).$$

If we want to take sequence information into account, we can work with

$$r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1) \eta(k+1, n+1) \qquad (1)$$

where $\eta(i,j)$ is the indicator which is 1 iff s_i and s_j are complementary.

Enumeration

Enumerating secondary structures is easy; their number is given by the following recurrence relation:

$$r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1).$$

If we want to take sequence information into account, we can work with

$$r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1) \eta(k+1, n+1) \qquad (1)$$

where $\eta(i,j)$ is the indicator which is 1 iff s_i and s_j are complementary.

Random sequence: Taking expectation of eq. (1); $\eta(i,j) \rightsquigarrow$ so-called *stickiness* p (the expectation of η) corresponding to the probability for two random nucleotides to be complementary.

Enumeration

Enumerating secondary structures is easy; their number is given by the following recurrence relation:

$$r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1).$$

If we want to take sequence information into account, we can work with

$$e(n+1) = e(n) + \sum_{0 \leqslant k \leqslant n-2} e(k) e(n-k-1) \times p \qquad (1)$$

where $\eta(i,j)$ is the indicator which is 1 iff s_i and s_j are complementary.

Random sequence: Taking expectation of eq. (1); $\eta(i,j) \rightsquigarrow$ so-called *stickiness* p (the expectation of η) corresponding to the probability for two random nucleotides to be complementary.

Algorithmic challenge

Input: RNA sequence (cheap with today's lab techniques).

Output: (Predicted) RNA secondary structure (considered a good approximation of 3D conformation).

Algorithmic challenge

Input: RNA sequence (cheap with today's lab techniques).

Output: (Predicted) RNA secondary structure (considered a good approximation of 3D conformation).

Prominent approach: Dynamic programming, i.e. table filling algorithm:

1. Processing input sequence $s_1 s_2 \cdots s_n$,
2. $V(i,j)$ represents the minimal energy possible for a folding of subsequence $s_i \cdots s_j$ subject to the i-th and j-th nucleotide being paired to each other;
3. $W(i,j)$ gives the corresponding minimum without that restriction.

$\rightsquigarrow n^3$ runtime algorithms
(quadratic number of entries each giving rise to linear time).

Motivation for Polymer-Zeta Model

Observation: While computing optimal folding for subsequence $s_i \cdots s_j$, a pairing of s_i and s_k only needs to be considered if pairing of s_i and s_k **already implied a minimum** while considering $s_i \cdots s_{j'}$, $j' < j$.

Speedup: Bookkeeping (candidate list) of s_k observed in minimal pairings for smaller subsequences may reduce the number of combinations to be considered for each entry.

Motivation for Polymer-Zeta Model

Observation: While computing optimal folding for subsequence $s_i \cdots s_j$, a pairing of s_i and s_k only needs to be considered if pairing of s_i and s_k **already implied a minimum** while considering $s_i \cdots s_{j'}$, $j' < j$.

Speedup: Bookkeeping (candidate list) of s_k observed in minimal pairings for smaller subsequences may reduce the number of combinations to be considered for each entry.

Polymer-zeta property: probability for the i-th and j-th nucleotides at distance $d = j - i + 1$ to form a pair is given by $p_d = \frac{b}{d^c}$ (for some constants $b > 0, c > 0$).

$\rightsquigarrow$ candidate list of (expected) constant length and thus expected quadratic run time algorithm.

Question addressed here

For certain classes of RNA (especially mRNA) it is justified to assume the polymer-zeta property.

Question: Is it appropriate in general?

Question addressed here

For certain classes of RNA (especially mRNA) it is justified to assume the polymer-zeta property.

Question: Is it appropriate in general?

Approach: We

- compute the **average shape** of secondary structures (considered a combinatorial object thus no nucleotides, just size)
- assuming the polymer-zeta property
- using methods from **enumerative combinatorics**
- and compare it to statistics derived from native foldings (databases).

Enumeration in the Polymer-Zeta Model

Model: Study $r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1) \times p_{n-k}$ which – in analogy to Bernoulli model – is the **expected number** of structures of size n denoted $\mathbb{E}_{\#}^{c,b}(\mathcal{S}_n)$.

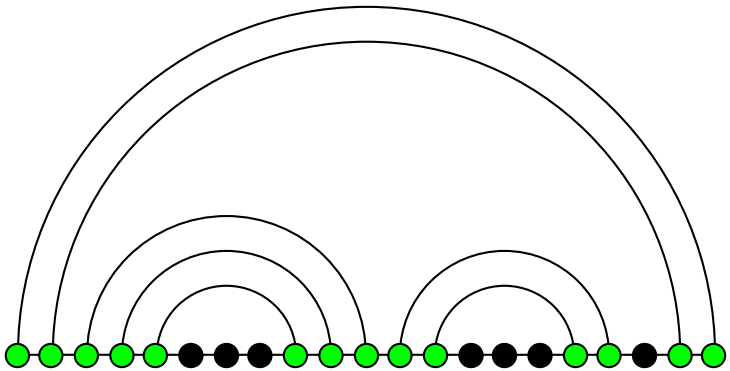
Enumeration in the Polymer-Zeta Model

Model: Study $r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1) \times p_{n-k}$ which – in analogy to Bernoulli model – is the **expected number** of structures of size n denoted $\mathbb{E}_{\#}^{c,b}(\mathcal{S}_n)$.

If we additionally compute the expected number of structures with **parameter value** k (e.g. number of so-called hairpins) $\mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k})$, then

$$\overline{X}_n^{c,b} = \sum_{k \geqslant 1} k \cdot \frac{\mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k})}{\mathbb{E}_{\#}^{c,b}(\mathcal{S}_n)}$$

is the **averaged behavior** of the parameter in consideration.



Enumeration in the Polymer-Zeta Model

Model: Study $r(n+1) = r(n) + \sum_{0 \leqslant k \leqslant n-2} r(k) r(n-k-1) \times p_{n-k}$ which – in analogy to Bernoulli model – is the **expected number** of structures of size n denoted $\mathbb{E}_{\#}^{c,b}(\mathcal{S}_n)$.

We considered $p_d = \frac{b}{d^c}$ for $(c, b) \in \{1, 2\}^2$ (theoretical considerations imply $b = 1$, $c = 1.5$, fitting to mRNA data yields $c = 1.47$).

Reason: Our approach only allows integer values for c since p_d is introduced into our equations by the following *trick* on generating functions: Consider the operator $\Theta = \Theta(z) = z\frac{\partial}{\partial z}$. Then

- For $c = 1$, $\Theta \frac{b}{(n+1)^c} z^n = b z^n$;
- for $c = 2$, $\Theta^2 \frac{b}{(n+1)^c} z^n = b z^n$.

This way, we can derive appropriate differential equations for generating functions.

Average Number of Hairpins

Theorem

Under the assumption of the (c, b)-polymer-zeta model, $c \in \{1, 2\}$, the **average number of hairpins** *in a secondary structure of size n is asymptotically given by*

$$\overline{X}_n^{1,b} = x_{1,b}\, n(1 + \mathcal{O}(n^{-\frac{1}{2}}))$$
$$\overline{X}_n^{2,b} = x_{2,b}\, n(1 + \mathcal{O}((\log n)^{-1}))$$

where $x_{c,b} > 0$ is a constant and for $b \in \{1, 2\}$ we have

$$x_{1,1} \approx 0.1326 \quad x_{1,2} \approx 0.1476$$
$$x_{2,1} \approx 0.1238 \quad x_{2,2} \approx 0.1489$$

Average Number of Hairpins

We start with

$$S_c(z,w) = \sum_{n \geqslant 3} \sum_{k \geqslant 1} \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}) w^k z^n + \sum_{n \geqslant 0} z^n.$$

Representation? Consider class $\mathcal{T}_{n+2,k}$ of so-called **irreducible structures** given by those structures from $\mathcal{S}_{n+2,k}$ with the first and the last base paired. We have for $k \geqslant 2$,

$$\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,k}) = \frac{b}{(n+1)^c} \cdot \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}), \tag{2}$$

and in the case $k = 1$,

$$\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,1}) = \frac{b}{(n+1)^c}(1 + \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,1}))$$

holds.

Average Number of Hairpins

Let $T_c(z,w)$ be the double generating function of $\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,k})$ ($n \geqslant 3$, $k \geqslant 1$). Based on eq. (2), we find

$$T_c(z,w) = \tag{3}$$
$$b \sum_{n \geqslant 3} \sum_{k \geqslant 1} \frac{1}{(n+1)^c} \cdot \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}) w^k z^{n+2} + \sum_{n \geqslant 1} \frac{b}{(n+1)^c} w z^{n+2}.$$

On the other hand, each $\mathsf{S}_{n,k}$-structure can be considered **a sequence of $\mathsf{T}_{i,j}$-structures** with leading, intermediate and trailing run of unpaired bases. In terms of generating functions, we thus have

$$S_c(z,w) = \frac{1}{1 - (T_c(z,w) + z)}. \tag{4}$$

Average Number of Hairpins

Let $T_c(z,w)$ be the double generating function of $\mathbb{E}_{\#}^{c,b}(\mathcal{T}_{n+2,k})$ ($n \geqslant 3$, $k \geqslant 1$). Based on eq. (2), we find

$$T_c(z,w) = \tag{3}$$
$$b \sum_{n \geqslant 3} \sum_{k \geqslant 1} \frac{1}{(n+1)^c} \cdot \mathbb{E}_{\#}^{c,b}(\mathcal{S}_{n,k}) w^k z^{n+2} + \sum_{n \geqslant 1} \frac{b}{(n+1)^c} w z^{n+2}.$$

On the other hand, each $\mathsf{S}_{n,k}$-structure can be considered **a sequence of $\mathsf{T}_{i,j}$-structures** with leading, intermediate and trailing run of unpaired bases. In terms of generating functions, we thus have

$$S_c(z,w) = \frac{1}{1 - (T_c(z,w) + z)}. \tag{4}$$

Average Number of Hairpins

We consider $c = 1$. Dividing by z and **taking the partial derivative in** z (denoted by index z) on both sides of eq. (3), we obtain

$$\left(\frac{T_1(z,w)}{z}\right)_z = b\sum_{n\geqslant 3}\sum_{k\geqslant 1}\mathbb{E}^{1,b}_{\#}(\mathcal{S}_{n,k})w^k z^n + \sum_{n\geqslant 1} bwz^n = bS_1(z,w) + \frac{b(wz-1)}{1-z}$$

and thus get rid of denominator $(n+1)^c$. In combination of eq. (4), we find the functional identity for $S_1 = S_1(z,w)$, given by

$$S_{1,z} = -\frac{1}{z}S_1 + \left[\frac{1}{z} + \frac{bz(wz-1)}{1-z}\right]S_1^2 + zbS_1^3. \tag{5}$$

Now, from eq. (5) we can determine

$$\overline{X}_n^{1,b} = \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)} = \frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)},$$

using methods from singularity analysis.

Average Number of Hairpins

We consider $c = 1$. Dividing by z and **taking the partial derivative in** z (denoted by index z) on both sides of eq. (3), we obtain

$$\left(\frac{T_1(z,w)}{z}\right)_z = b\sum_{n\geqslant 3}\sum_{k\geqslant 1}\mathbb{E}_{\#}^{1,b}(\mathcal{S}_{n,k})w^k z^n + \sum_{n\geqslant 1} bwz^n = bS_1(z,w) + \frac{b(wz-1)}{1-z}$$

and thus get rid of denominator $(n+1)^c$. In combination of eq. (4), we find the functional identity for $S_1 = S_1(z,w)$, given by

$$S_{1,z} = -\frac{1}{z}S_1 + \left[\frac{1}{z} + \frac{bz(wz-1)}{1-z}\right]S_1^2 + zbS_1^3. \tag{5}$$

Now, from eq. (5) we can determine

$$\overline{X}_n^{1,b} = \frac{[z^n]\frac{\partial S_1(z,w)}{\partial w}\Big|_{w=1}}{[z^n]S_1(z,1)} = \frac{[z^n]S_{1,w}(z,1)}{[z^n]S_1(z,1)},$$

using methods from singularity analysis.

Average Number of Hairpins

Singularity Analysis: We proceed along the following steps

- set $w = 1$ (immediate or after taking $\partial/\partial w$);
- determine (unique) **dominant singularity** of resulting generating function,
- among the **fixed** (fixed by equation itself) **and movable** (depending in initial conditions) singularities of the resulting differential equation;
- derive **series expansion** of generating function at dominant singularity (using knowledge on type);
- apply **transfer theorem** (singularity $\leftrightarrow$ exponential rate of grows, series expansion $\leftrightarrow$ subexponential contribution and constants) which yields precise asymptotic for n-th coefficient ($n \to \infty$).

$\rightsquigarrow$ **Asymptotic given in theorem.**

Average Number of Hairpins

Singularity Analysis: We proceed along the following steps

- set $w = 1$ (immediate or after taking $\partial/\partial w$);
- determine (unique) **dominant singularity** of resulting generating function,
- among the **fixed** (fixed by equation itself) **and movable** (depending in initial conditions) singularities of the resulting differential equation;
- derive **series expansion** of generating function at dominant singularity (using knowledge on type);
- apply **transfer theorem** (singularity $\leftrightarrow$ exponential rate of grows, series expansion $\leftrightarrow$ subexponential contribution and constants) which yields precise asymptotic for n-th coefficient ($n \to \infty$).

$\rightsquigarrow$ **Asymptotic given in theorem.**

Average Number of Hairpins

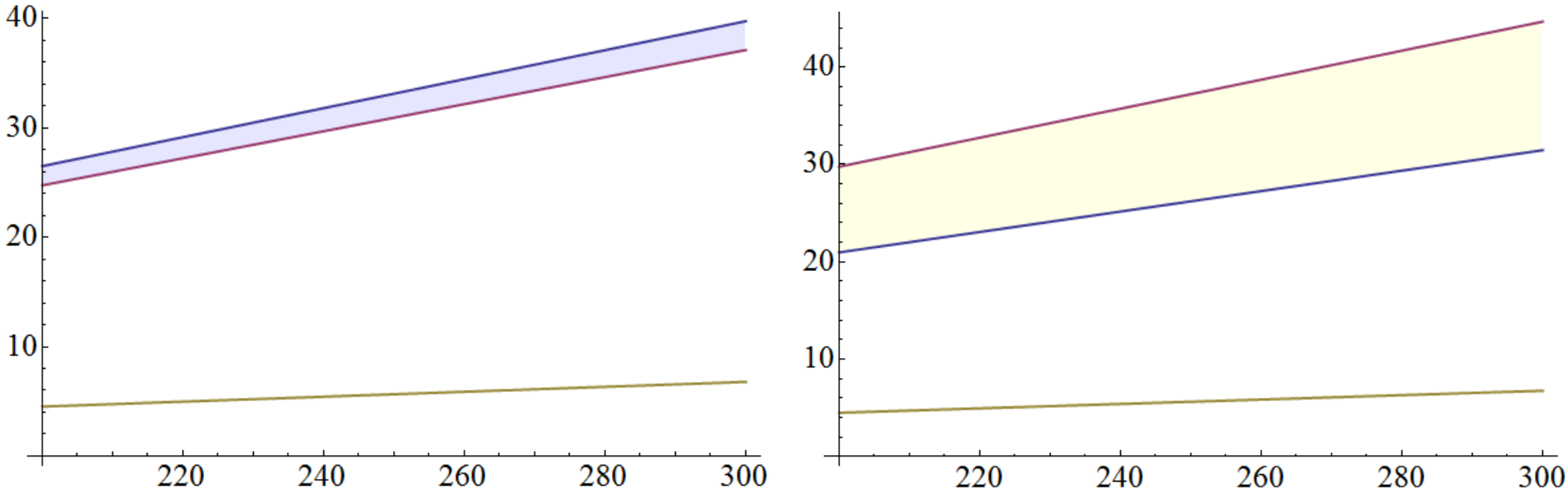


Figure : Plots of the **average number of hairpins** as a function of the structure's size n within our polymer-zeta model ($b = 1$ left, $b = 2$ right). The blue (resp. red) line corresponds to case $c = 1$ (resp. $c = 2$), the greenish line shows the behavior of native RNA secondary structures (as derived from databases.

Results

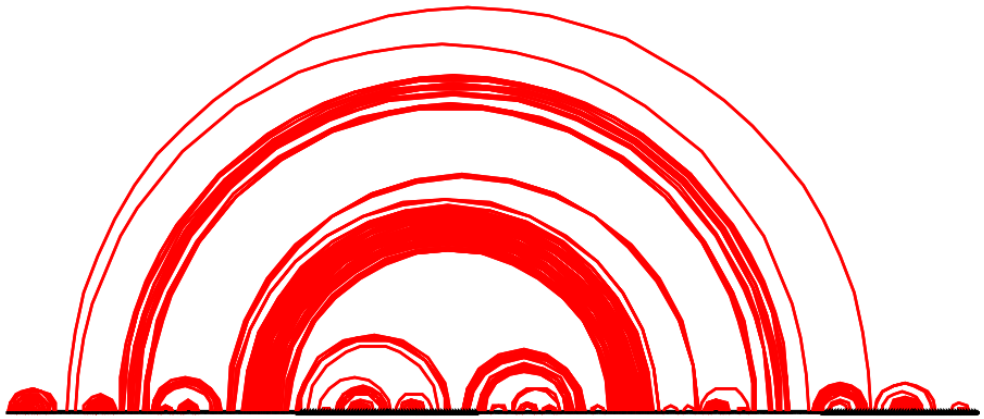
parameter	expectation				
		$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
Number of hairpins	$0.0226n$	$0.1326n$	$0.1049n$	$0.1238n$	$0.1489n$
Length of a hairpin-loop	7.3766	1.7262	2.7636	1.7367	1.5467
Number of bulges	$0.0095n$	$0.0210n$	$0.0277n$	$0.0076n$	$0.0113n$
Length of a bulge	1.5949	2.0476	2.0217	2.4079	2.0265
Number of interior loops	$0.0164n$	$0.0110n$	$0.0141n$	$0.0055n$	$0.0059n$
Total Length of both loops within an interior loop	7.7870	4.2364	4.1560	5.3455	4.4068
Number of multiloop	$0.0106n$	$0.0266n$	$0.0252n$	$0.0064n$	$0.0097n$
Degree of a multiloop	4.1311	3.9774	3.7063	3.8125	3.9278

Discussion

1. The **symbolic method** and **analytic combinatorics** (see Bob Sedgewick's talk on Tuesday) are well-suited to deal with the polymer zeta model of RNA;

Discussion

1. The **symbolic method** and **analytic combinatorics** (see Bob Sedgewick's talk on Tuesday) are well-suited to deal with the polymer zeta model of RNA;
2. We proved various structural parameters to behave realistic, others (e.g. hairpins or *exterior loops*) to behave **rather unrealistic** in that model;



Discussion

1. The **symbolic method** and **analytic combinatorics** (see Bob Sedgewick's talk on Tuesday) are well-suited to deal with the polymer zeta model of RNA;
2. We proved various structural parameters to behave realistic, others (e.g. hairpins or *exterior loops*) to behave **rather unrealistic** in that model;
3. As a consequence, we cannot conclude a speedup of structure prediction by sparsification for arbitrary classes of RNA.

Thank you for your attention!