Small-world networks and RNA secondary structures

Defne Surujon, Yann Ponty, Peter Clote

Small-world networks and RNA secondary structures

Defne Surujon
Biology Department
Boston College
Chestnut Hill, MA 02467
defne.surujon@bc.edu

Yann Ponty
Laboratoire d'Informatiques(LIX)
Ecole Polytechnique
91128 Palaiseau Cedex - France
ponty@lix.polytechnique.fr

Peter Clote
Biology Department
Boston College
Chestnut Hill, MA 02467
Corresponding author: clote@bc.edu

***Abstract*—Let $\mathcal{S}_n$ denote the network of all RNA secondary structures of length n, in which undirected edges exist between structures s, t such that t is obtained from s by the addition, removal or shift of a single base pair. Using context-free grammars, generating functions and complex analysis, we show that the asymptotic average degree is $O(n)$ and that the asymptotic clustering coefficient is $O(1/n)$, from which it follows that the family $\mathcal{S}_n$, $n = 1, 2, 3, \ldots$ of secondary structure networks is not small-world.**

1. Introduction

In this section, we define notions of RNA secondary structure, move sets MS_1, MS_2, and small-world networks. An RNA secondary structure of length n, subsequently called length n structure, is defined to be a set s of ordered pairs (i, j), with $1 \leq i < j \leq n$, such that: (1) There are no base triples; i.e. if $(i, j), (k, \ell) \in s$ and $\{i, j\} \cap \{k, \ell\} \neq \emptyset$, then $i = k$ and $j = \ell$. (2) There are no pseudoknots; i.e. if $(i, j), (k, \ell) \in s$, then it is not the case that $i < k < j < \ell$. (3) There are at least $\theta = 3$ unpaired bases in a hairpin loop; i.e. if $(i, j) \in s$, then $j - i > \theta = 3$. Note that base pairs are *not* required to be Watson-Crick or wobble pairs, as is the case for RNA molecules, such as that depicted in Figure 1a. This definition, sometime called *homopolymer* secondary structure, permits the combinatorial analysis we employ to show that RNA networks are not small-world.

Let $\mathcal{S}_n$ denote the set of all length n structures. The move sets MS_1 and MS_2, defined in [8] for RNA secondary folding kinetics, describe elementary moves that transform a structure s into another structure t. Move set MS_1 [resp. MS_2] consists of either removing or adding [resp. removing, adding or shifting] a single base pair, provided the resulting set of base pairs constitutes a valid structure, where shift moves are depicted in Figure 2. We overload the notation $\mathcal{S}_n$ to also denote the MS_1 network [resp. MS_2 network], whose nodes are the length n structures, where an undirected edge between structures s, t exists when t is obtained from s by a single move from MS_1 [resp. MS_2]. Figure 1b shows the MS_1 network (8 red edges) [resp. MS_2 network (8 red and 8 blue edges)] for length 7 structures, where there are 8 nodes, MS_1 degree $\frac{16}{8} = 2$ and MS_2 degree $\frac{32}{8} = 4$. See [3], [4] for dynamic programming algorithms that compute, respectively, the MS_1 and MS_2 degree for the network of secondary structures of a given RNA sequence.

Small-world networks [12], ubiquitous in biology, sociology, and technology, satisfy two conditions: (1) on average, the minimum path length between any two nodes is small, (2) neighbors of a node tend to be connected to each other. The *global clustering coefficient*, defined in equation (77) of [11], is given by

$$\mathfrak{C}_g(G) = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} \qquad (1)$$

where a *triangle* is a set $\{x, y, z\}$ of nodes, each of which is connected by an edge, and a (connected) triple is a set $\{x, y, z\}$ of nodes, such that there is an edge from x to y and an edge from x to z. Following [5], the family $\{\mathcal{S}_n, n = 1, 2, 3, \ldots\}$ of RNA networks is small-world if the following conditions hold. (1) There is a constant $c_1 \geq 0$, such that the minimum path length between any two nodes of $\mathcal{S}_n$ is bounded above by $c_1 \ln n$. (2) There is a constant $c_2 \geq 0$, such that the average network degree of $\mathcal{S}_n$ is bounded above by $c_2 \ln n$. (3) The global clustering coefficient is bounded away from zero. By Theorem 2, the network size of $\mathcal{S}_n$ is exponential in n. Since there are at most $n/2$ base pairs in any length n structure, condition (1) is satisfied for both the MS_1 and MS_2 networks of RNA structures. It is easy to see that the clustering coefficient of the MS_1 network of RNA structures is zero, so in the remainder of the paper, we concentrate on conditions (2) and (3) for the MS_2 RNA network.

The overall method used is as follows: (1) Give a context-free grammar that generates the set of all secondary structures, possibly containing a specific motif. (2) Use Table 1 to derive and then solve a functional relation for the complex generating function $S(z)$, with the property that the nth Taylor coefficient of $S(z)$, denoted $[z^n]S(z)$, is equal to the number of length n structures, possibly containing a specific motif. (3) Determine the dominant singularity and apply complex analysis [6] to obtain the asymptotic value of $[z^n]S(z)$. For step (3), we use the Flajolet-Odlyzko Theorem, stated as Corollary 2, part (i) on page 224 of [6]. Before stating the theorem, we define the *dominant singularity* of complex function $f(z)$ to be the complex number ρ having smallest absolute value (or modulus) at which $f(z)$ is not differentiable.

Theorem 1 (Flajolet and Odlyzko). Assume that $f(z)$ has a dominant singularity at $z = \rho > 0$, is analytic for $z \neq \rho$ satisfying $|z| \leq |\rho|$, and that

$$\lim_{z \to \rho} f(z) = K(1 - z/\rho)^{\alpha}. \tag{2}$$

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

$$f_n = [z^n]f(z) \sim \frac{K}{\Gamma(-\alpha)} \cdot n^{-\alpha-1} \cdot \rho^{-n}$$

where $\sim$ denotes asymptotic equality and Γ denotes the Gamma function.

The plan of the paper is now as follows. In Section 2, we show that the average MS_2 degree of $\mathcal{S}_n$ is $O(n)$. In Section 3.1 [resp. 3.2] we prove that the average number of triangles [resp. triples] per structure is $O(n)$ [resp. $O(n^2)$], which implies that the asymptotic global clustering coefficient is $O(1/n)$, hence not bounded away from zero. It follows that the family of RNA secondary structure networks is not small-world.

2. Expected network degree

Due to space constraints, details for the computation of the asymptotic number of secondary structures as well as for MS_1 expected degree for homopolymers cannot be given in this paper. Nevertheless, these computations can be found in [2], from which we take the following results. Recalling the notation $\sim$ for asymptotic equality, we have

Theorem 2. If $S(z)$ is the generating function for the number of secondary structures for a homopolymer, then

$$[z^n]S(z) \sim 0.713121 \cdot n^{-3/2} \cdot 2.28879^n$$

If $MS_1\text{degree}(n)$ denotes the MS_1 expected network degree for a homopolymer, then

$$MS_1\text{degree}(n) \sim 0.473475 \cdot n$$

Define the grammar G to consist of the terminal symbols $(, \bullet,)$, $\langle, \star, \rangle$, nonterminal symbols $\widehat{S}, \widehat{T}, S, R, \theta$, with start symbol $\widehat{S}$. Shift moves are represented in the grammar by one of the three expressions: $\star\rangle\,\rangle$, $\langle\,\langle\star$, $\langle\,\star\,\rangle$, as depicted in Figure 2. In particular, $\star\rangle\,\rangle$ represents the right shift depicted in Figure 2a (ignoring possible intervening structure), where base pair (x, y) is transformed to (x, y') for $x < y' < y$; alternatively, the $\star\rangle\,\rangle$ can represent the shift (x, y) to (x, y') for $x < y < y'$, as depicted in Figure 2b. The expression $\langle\,\langle\star$ can represent the left shift depicted in Figure 2c, where base pair (x, y) is transformed to (x', y) for $x < x' < y$; alternatively, $\langle\,\langle\star$ can represent the shift (x, y) to (x', y) for $x' < x < y$, as depicted in Figure 2d. The expression $\langle\,\star\,\rangle$ can represent the right-to-left shift depicted in Figure 2e, where base pair (x, y) is transformed to (y', x) for $y' < x < y$; alternatively, $\langle\,\star\,\rangle$ can represent the shift (x, y) to (y, x') for $x < y < x'$, as depicted in Figure 2f. The grammar G allows us to count the number of secondary structures, that additionally contain a unique occurrence of exactly one of the three expressions: $\star\rangle\,\rangle$, $\langle\,\langle\star$, $\langle\,\star\,\rangle$. Since two shift moves correspond to each of the previous three expressions, it follows that the total number of $MS_2 - MS_1$ (shift-only) moves, summed over all structures for a homopolymer of length n with $\theta = 1$, is equal to $2[z^n]S^{\dagger}(z)$.

The production rules of grammar G are as follows:

$$\begin{aligned}
&\widehat{S} \to \widehat{S}\,\bullet \mid (\,\widehat{S}\,) \mid S\,(\,\widehat{S}\,) \mid \widehat{S}\,(\,R\,) \mid \widehat{T}\\
&\widehat{T} \to \star R\rangle\,\rangle \mid S \star R\rangle\,\rangle \mid \star R\rangle\, S\rangle \mid S \star R\rangle\, S\rangle \mid\\
&\quad \langle\,\langle R\star \mid S\langle\,\langle R\star \mid \langle S\langle R\star \mid S\langle S\langle R\star \mid\\
&\quad \langle R\star R\rangle \mid S\langle R\star R\rangle\\
&S \to \bullet \mid S\,\bullet \mid (\,R\,) \mid S\,(\,R\,)\\
&R \to \theta \mid R\,\bullet \mid (\,R\,) \mid S\,(\,R\,)\\
&\theta \to \bullet\bullet\bullet
\end{aligned} \tag{3}$$

The nonterminal S is responsible for generating all secondary structures of length greater than or equal to 1. In contrast, the nonterminal $\widehat{S}$ is responsible for generating all well-balanced expressions of length greater than or equal to 1, that involve exactly one of the three expressions: $\star\rangle\,\rangle$, $\langle\,\langle\star$, $\langle\,\star\,\rangle$. To that end, the nonterminal $\widehat{T}$ is responsible for generating all such expressions, in which the rightmost symbol is either $\rangle$ or $\star$, but not $\bullet$ or $)$. By induction on length of sequence generated, one can show that G is an nonambiguous context-free grammar that generates

all secondary structures having a unique occurrence of one of $\star\rangle\,\rangle$, $\langle\,\langle\star$, $\langle\,\star\,\rangle$. As mentioned before, 2 times the number of such expressions of length n is equal to the number of $MS_2 - MS_1$ edges in the network of secondary structures.

As explained in [10] and [7], it is possible to automatically transform the previous production rules into equations that relate the corresponding generating functions, where we denote generating functions of $\widehat{S}(z)$, $\widehat{T}$, $S(z)$, $R(z)$ by the same symbols used for the corresponding nonterminals $\widehat{S}$, $\widehat{T}$, S, R. This technique is known in the literature as DSV methodology [10], or as the *symbolic method* [7] – see Table 1. In this fashion, we obtain the following:

$$\begin{aligned}
\widehat{S} &= z\widehat{S} + z^2\widehat{S} + z^2S\widehat{S} + z^2R\widehat{S} + \widehat{T}\\
\widehat{T} &= 2z^3R + 4z^3RS + 2z^3RS^2 + z^3R^2 + z^3SR^2\\
S &= z + zS + z^2R + z^2RS\\
R &= \theta + zR + z^2R + z^2RS\\
\theta &= z^3
\end{aligned}$$

and by eliminating all variables except $\widehat{S}$ and z, we use Mathematica to obtain the quadratic equation in $\widehat{S}$ having two solutions, for which the only solution analytic at 0 is the following:

$$\widehat{S}(z) = \widehat{S} = \frac{A + B\sqrt{P}}{C} \tag{4}$$

where

$$\begin{aligned}
P &= 1 - 2z - z^2 + z^4 + 3z^6 + 2z^7 + z^8\\
A &= 3 - 15z + 23z^2 - 9z^3 - z^4 - 9z^5 +\\
&\quad 23z^6 - 25z^7 + 7z^8 - z^9 + 6z^{10} -\\
&\quad 8z^{11} + 2z^{12} + 2z^{13} + 2z^{14}\\
B &= -3 + 12z - 14z^2 + 4z^3 + 5z^5 - 10z^6 +\\
&\quad 8z^7 - 2z^{10}\\
C &= 2(-z^3 + 3z^4 - z^5 - z^6 - z^7 + z^8 -\\
&\quad 3z^9 + z^{10} + z^{11} + z^{12})
\end{aligned}$$

The *dominant singularity* ρ of $\widehat{S}(z)$ in equation (4) is the complex number having smallest absolute value (or modulus) at which $\widehat{S}(z)$ is not differentiable. For the functions in this paper, the dominant singularity will always be the (complex) root of polynomial P under the radical, having smallest modulus – since the square root function is not differentiable over the complex numbers at zero.

Letting $\widehat{F}(z) = \frac{B\sqrt{P}}{C}$ and noting that the dominant singularity $\rho = 0.436911$, a calculation shows that

$$\begin{aligned}
\lim_{z\to\rho}\widehat{F}(z) &= \lim_{z\to\rho}\frac{B\cdot\sqrt{P'}\cdot(1-z/\rho)}{C'\cdot(1-z/\rho)}\\
P' &= \frac{P}{1-z/\rho}\\
&= 1 + 0.288795z - 0.339007z^2 -\\
&\quad 0.775919z^3 - 0.775919z^4 - 1.775919z^5 -\\
&\quad 1.064714z^6 - 0.436911z^7\\
C' &= \frac{C}{1-z/\rho}\\
&= -2z^3 + 1.422410z^4 + 1.255605z^5 +\\
&\quad 0.873822z^6 + 2z^8 - 1.422410z^9 -\\
&\quad 1.255605z^10 - 0.873822z^11
\end{aligned}$$

and so

$$\begin{aligned}
\lim_{z\to\rho}\widehat{F}(z) &= 0.684877\cdot\lim_{z\to\rho}(1-z/\rho)^{-1/2}\\
&= 0.684877\cdot\lim_{z\to\rho}(1-z/0.436911)^{-1/2}
\end{aligned}$$

Taking $\alpha = -1/2$ in the Flajolet-Odlyzko Theorem [6], we obtain:

$$\begin{aligned}
[z^n]\widehat{F}(z) &\sim \frac{0.684877}{\Gamma(1/2)}\cdot n^{-1/2}\cdot\left(\frac{1}{\rho}\right)^n\\
&= 0.3864\cdot n^{-1/2}\cdot 2.28879^n
\end{aligned}$$

By Theorem 2 the asymptotic number of secondary structures for a homopolymer when $\theta = 3$ is $0.713121\cdot n^{-3/2}\cdot 2.28879^n$, and so we have the following result.

Theorem 3. The asymptotic $MS_2 - MS_1$ degree of $\mathcal{S}_n$ is

$$\begin{aligned}
\frac{2[z^n]\widehat{F}(z)}{[z^n]S(z)} &\sim \frac{0.772801\cdot n^{-1/2}\cdot 2.28879^n}{0.713121\cdot n^{-3/2}\cdot 2.28879^n}\\
&= 1.083688\cdot n
\end{aligned}$$

Adding the asymptotic values from Theorem 2 and Theorem 3, we determine the MS_2 degree.

Corollary 4. The asymptotic MS_2 degree for the network $\mathcal{S}_n$ of RNA structures is $1.557164\cdot n$.

Using a Taylor series expansion at zero for the functions used to determine both the MS_1 and $MS_2 - MS_1$ degree, we have verified that the numerical results for $\mathcal{S}_n$ are identical with those independently computed by the dynamic programming C-implementations described in [3] and [4]. We also nthat the current approach is *much* simpler than the program in [4], although the latter is more general, since it computes the MS_2 degree for any user-specified RNA sequence.

3. Asymptotic MS_2 clustering coefficient

Subsection 3.1 describes a grammar to count the number of triangles for $\mathcal{S}_n$ with respect to MS_2 moves, while Subsection 3.2 describes a grammar to count two particular triples.

3.1 Counting triangles. Let G be the grammar with terminal symbols $($, $)$, $\bullet$, $\star$, nonterminal symbols $S^{\triangle}, S_1, \ldots, S_8, S, R, X, \theta$, start symbol $S^{\triangle}$ and the following production rules:

$$\begin{aligned} S^{\triangle} &\to S_1 \,|\, S_2 \,|\, S_3 \,|\, S_4 \,|\, S_5 \,|\, S_6 \,|\, S_7 \,|\, S_8 \\ S &\to \bullet \,|\, S \bullet \,|\, (\, R\,) \,|\, S\,(\, R\,) \\ R &\to \theta \,|\, R \bullet \,|\, (\, R\,) \,|\, S\,(\, R\,) \\ X &\to \lambda \,|\, R \\ \theta &\to \bullet\bullet\bullet \end{aligned}$$

where λ denotes the empty word, and $S_1, \ldots, S_8$ are specified in the following 8 exhaustive and mutually exclusive cases. Note that $S_1, \ldots, S_3$ generate structures containing type A triangles, while $S_4, \ldots, S_8$ generate structures containing type B triangles.

Rule 1 $\langle\ \star\ \rangle$**.** The following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, y)\}$ and $s \cup \{(y, z)\}$ are also secondary structures, hence form a triangle:

$$S_1 \to S_1 \bullet \,|\, (\, S_1\,) \,|\, S\,(\, S_1\,) \,|\, S_1\,(\, R\,) \,|\, X \langle\, R \star R \rangle$$

with corresponding DSV equations

$$S_1 = zS_1 + z^2S_1 + z^2SS_1 + z^2RS_1 + Xz^3R^2$$

Rule 2 $\star\rangle\ \rangle$**.** The following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, y)\}$ and $s \cup \{(x, z)\}$ are also secondary structures, hence form a triangle:

$$S_2 \to S_2 \bullet \,|\, (\, S_2\,) \,|\, S\,(\, S_2\,) \,|\, S_2\,(\, R\,) \,|\, X \star R \rangle\, X\, \rangle$$

with corresponding DSV equations

$$S_2 = zS_2 + z^2S_2 + z^2SS_2 + z^2RS_2 + X^2z^3R$$

Rule 3 $\langle\ \langle \star$**.** The following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, z)\}$ and $s \cup \{(y, z)\}$ are also secondary structures, hence form a triangle:

$$S_3 \to S_3 \bullet \,|\, (\, S_3\,) \,|\, S\,(\, S_3\,) \,|\, S_3\,(\, R\,) \,|\, X \langle\, X \langle\, R \star$$

with corresponding DSV equations

$$S_3 = zS_3 + z^2S_3 + z^2SS_3 + z^2RS_3 + X^2z^3R$$

Rule 4 $\star\rangle\ \rangle\ \rangle$**.** The following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, y)\}$, $s \cup \{(x, z)\}$ and $s \cup \{(x, w)\}$ are also secondary structures, hence the latter form a triangle:

$$S_4 \to S_4 \bullet \,|\, (\, S_4\,) \,|\, S\,(\, S_4\,) \,|\, S_4\,(\, R\,) \,|\, X \star R \rangle\, X \rangle\, X \rangle$$

with corresponding DSV equations

$$S_4 = zS_4 + z^2S_4 + z^2SS_4 + z^2RS_4 + X^3Rz^4$$

Rule 5 $\langle\ \langle\ \langle \star$**.** For $x < y < z < w$, let $s_1 = (x, w)$, $s_2 = (y, w)$, $s_3 = (z, w)$. The following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, w)\}$, $s \cup \{(y, w)\}$ and $s \cup \{(z, w)\}$ are also secondary structures, hence the latter form a triangle:

$$S_5 \to S_5 \bullet \,|\, (\, S_5\,) \,|\, S\,(\, S_5\,) \,|\, S_5\,(\, R\,) \,|\, X \langle\, X \langle\, X \langle\, R \star$$

with corresponding DSV equations

$$S_5 = zS_5 + z^2S_5 + z^2SS_5 + z^2RS_5 + X^3z^4R$$

Rule 6 $\langle\ \star\ \rangle\ \rangle$**.** For $x < y < z < w$, the following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, y)\}$, $s \cup \{(y, z)\}$ and $s \cup \{(y, w)\}$ are also secondary structures, hence the latter form a triangle:

$$S_6 \to S_6 \bullet \,|\, (\, S_6\,) \,|\, S\,(\, S_6\,) \,|\, S_6\,(\, R\,) \,|\, X \langle\, X \langle\, R \star R \rangle$$

with corresponding DSV equations

$$S_6 = zS_6 + z^2S_6 + z^2SS_6 + z^2RS_6 + X^2z^4R^2$$

Rule 7 $\langle\ \langle\ \star\ \rangle$**.** For $x < y < z < w$, the following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, z)\}$, $s \cup \{(y, z)\}$ and $s \cup \{(z, w)\}$ are also secondary structures, hence the latter form a triangle:

$$S_7 \to S_7 \bullet \,|\, (\, S_7\,) \,|\, S\,(\, S_7\,) \,|\, S_7\,(\, R\,) \,|\, X \langle\, R \star R \rangle\, X \rangle$$

with corresponding DSV equations

$$S_7 = zS_7 + z^2S_7 + z^2SS_7 + z^2RS_7 + X^2z^4R^2$$

Rule 8 $\langle\ \star\ \rangle$ **bis.** The following productions generate all secondary structures s, such that for $x < y < z$, it is the case that $s \cup \{(x, z)\}$, $s \cup \{(x, y)\}$ and $s \cup \{(y, z)\}$ are also secondary structures, hence the latter form a triangle. This grammar is identical to that in rule 1 above, with the exception that S_1 is replaced by S_8.

Let $S^{\triangle}(z)$ denote the generating function for the number of structures containing a unique triangle motif, where $triA(z)$ [resp. $triB(z)$] is the generating function for the collection of structures containing a unique

occurrence of type A [type B] triangle, as treated in rules 1-3 [resp. rules 4-8]. We obtain the following compact form for the DSV equations for the grammar G that generates all structures containing a triangle:

$$\begin{aligned} S^{\triangle} &= triA + triB \\ triA &= triA \cdot z + X \cdot z \cdot triA \cdot z + \\ &\quad triA \cdot z \cdot R \cdot z + X \cdot z \cdot R \cdot z \cdot R \cdot z + \\ &\quad X \cdot z \cdot R \cdot z \cdot X \cdot z + X \cdot z \cdot X \cdot z \cdot R \cdot z \\ triB &= triB \cdot z + X \cdot z \cdot triB \cdot z + \\ &\quad triB \cdot z \cdot R \cdot z + X^3 z^4 R + X^3 z^4 R + \\ &\quad X^2 z^4 R^2 + X^2 z^4 R^2 + X z^3 R^2 \end{aligned}$$

Using Mathematica, we determine the following.

$$[z^n]S^{\triangle}(z) = 0.870311 \cdot 2.28879^n \cdot n^{-1/2}$$

By Theorem 2, the asymptotic number of secondary structures is $0.713121 \cdot n^{-3/2} \cdot 2.28879^n$, and so we have the following result.

Theorem 5. The asymptotic average number of triangles per structure is

$$\begin{aligned} \frac{[z^n]S^{\triangle}(z)}{[z^n]S(z)} &\sim \frac{0.870331 \cdot n^{-1/2} \cdot 2.28879^n}{0.713121 \cdot n^{-3/2} \cdot 2.28879^n} \\ &\sim 1.220453 \cdot n \end{aligned}$$

3.2 Counting triples. In this subsection, we describe a grammar for two particular triples. Let G be the grammar having terminal symbols $\bullet$, `(` , `)` , `[` , `]` , nonterminal symbols $S^{\ddagger}$, $S^{\dagger}$, S, R, X, θ, start symbol $S^{\ddagger}$, and productions given in equation (5) below together with the following:

$$\begin{aligned} S &\to \bullet \mid S \bullet \mid X \texttt{(} R \texttt{)} \\ R &\to \theta \mid R \bullet \mid X \texttt{(} R \texttt{)} \\ X &\to \lambda \mid R \\ \theta &\to \bullet \bullet \bullet \end{aligned}$$

Triple with motif `[ ] [ ]` or `[ [ ] ]`. The following grammar generates all secondary structures s that have two special base pairs (i, j) and (x, y), designated by `[ ]`, which are either sequential or nested. For each structure s, which contains a unique occurrence of the sequential motif `[ ] [ ]` or of the nested motif `[ [ ] ]`, we must count four possible triples: (1) $\{s_1, s_2, s_3\}$, where $s_1 = s - \{(i, j), (x, y)\}$, $s_2 = s - \{(i, j)\}$, $s_3 = s - \{(x, y)\}$. (2) $\{s_1, s_2, s_3\}$, where $s_1 = s$, $s_2 = s - \{(i, j)\}$, $s_3 = s - \{(x, y)\}$. (3) $\{s_1, s_2, s_3\}$, where $s_1 = s - \{(i, j)\}$, $s_2 = s - \{(i, j), (x, y)\}$, $s_3 = s$. (4) $\{s_1, s_2, s_3\}$, where $s_1 = s - \{(x, y)\}$, $s_2 = s - \{(i, j), (x, y)\}$, $s_3 = s$. For this reason, we multiply by 4 the asymptotic number of structures generated by the following grammar G. The grammar G has terminal symbols $\bullet$, `(` , `)` , `[` , `]` , nonterminal symbols $S^{\ddagger}$, $S^{\dagger}$, S, R, X, θ, start symbol $S^{\ddagger}$, and the following production rules.

$$\begin{aligned} S^{\ddagger} &\to S^{\ddagger} \bullet \mid \texttt{(} S^{\ddagger} \texttt{)} \mid S \texttt{(} S^{\ddagger} \texttt{)} \mid S^{\ddagger} \texttt{(} R \texttt{)} \mid \\ &\quad \texttt{[} S^{\dagger} \texttt{]} \mid S \texttt{[} S^{\dagger} \texttt{]} \mid S^{\dagger} \texttt{[} R \texttt{]} \mid S^{\dagger} \texttt{(} S^{\dagger} \texttt{)} \\ S^{\dagger} &\to S^{\dagger} \bullet \mid \texttt{(} S^{\dagger} \texttt{)} \mid S \texttt{(} S^{\dagger} \texttt{)} \mid S^{\dagger} \texttt{(} R \texttt{)} \mid \\ &\quad \texttt{[} R \texttt{]} \mid S \texttt{[} R \texttt{]} \end{aligned} \tag{5}$$

When applying the Flajolet-Odlyzko Theorem in the current case, we have $\rho = 0.436911$ and $\alpha = -3/2$. A computation shows that

$$\begin{aligned} \lim_{z \to \rho} S^{\ddagger}(z) &= 0.0177098\,(1 - z/\rho))^{-3/2} \\ [z^n]S^{\ddagger}(z) &\sim 0.0199834 \cdot n^{1/2} \cdot 2.28879^n \\ \frac{[z^n]S^{\ddagger}(z)}{[z^n]S(z)} &\sim \frac{0.0199834 \cdot n^{1/2} \cdot 2.28879^n}{0.713121 \cdot n^{-3/2} \cdot 2.28879^n} \\ &\sim 0.0280225 \cdot n^2 \end{aligned}$$

As mentioned, the number of triples contributed in the current case is *4 times* the last value. Thus the expected number of triples involving a structure containing `[ ] [ ]` or `[ [ ] ]` is $4 \cdot 0.0280224 \cdot n^2 = 0.1120896 \cdot n^2$.

Theorem 6. The asymptotic average number of triples per structure, for the triples described in this section, is

$$\frac{4[z^n]S^{\ddagger}(z)}{[z^n]S(z)} \sim 0.11209 \cdot n^2$$

From Theorems 5 and 6, we obtain an upper bound for the global clustering coefficient, defined in equation (1).

Theorem 7 (Bound on global clustering coefficient).

$$\mathfrak{C}_g(G) = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}} = O\left(\frac{1}{n}\right)$$

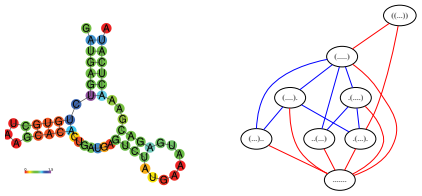
and hence the family $\mathcal{S}_n$, $n = 1, 2, 3, \ldots$ of RNA secondary structures is not small-world.

4. Discussion

In this paper, we have used methods from algebraic combinatorics [7] to determine the asymptotic average degree and asymptotic clustering coefficient of the MS_2 network $\mathcal{S}_n$ of RNA secondary structures. Since the clustering coefficient is not bounded away from zero, it follows that the family $\mathcal{S}_n$, $n = 1, 2, 3, \ldots$, of networks is not small-world. Our rigorous result differs from computer simulations involving a low energy ensemble of structures as studied in [1], [13], etc. In the journal version of this paper, we discuss the relation between

Type of nonterminal	Generating function
$A \to B \mid C$	$A(z) = B(z) + C(z)$
$A \to B\,C$	$A(z) = B(z)C(z)$
$A \to t$	$A(z) = z$
$A \to \varepsilon$	$A(z) = 1$

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



(a) PLMVd (b) RNA network

Figure 1: *(a)* Consensus secondary structure of the type III hammerhead ribozyme from Peach Latent Mosaic Viroid (PLMVd) AJ005312.1/282-335 (isolate LS35, variant ls16b), taken from Rfam [9] family RF00008. *(b)* Network for size 7 homopolymer with $\theta = 3$, having 8 nodes and 8 red MS_1 edges (base pair addition or removal), 8 blue $MS_2 - MS_1$ edges (base pair shift), hence a total of 16 MS_2 edges. It follows that MS_1 degree is $\frac{16}{8} = 2$, while MS2 is $\frac{32}{8} = 4$.

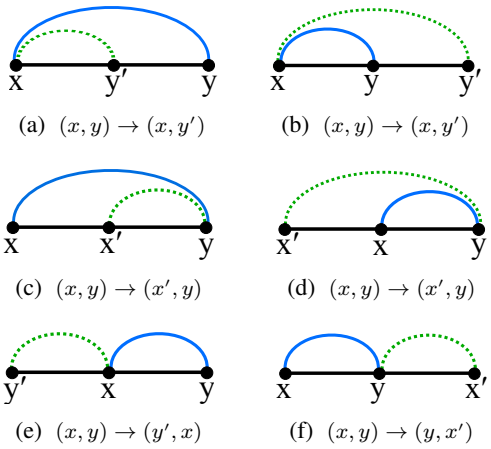
our result and such simulation results, we compute the exact clustering coefficient for $\mathcal{S}_n$, which involves 40 types of triples, and we extend results to a more general model in which the user can stipulate the probability that any two positions can form a base pair.

Acknowledgments

This research was supported in part by National Science Foundation grant DBI-1262439 to PC and the French/Austrian RNALands project (ANR-14-CE34-0011 and FWF-I-1804-N28) to YP. 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] G. R. Bowman and V. S. Pande. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U.S.A.*, 107(24):10890–10895, June 2010.



(a) $(x, y) \to (x, y')$ (b) $(x, y) \to (x, y')$

(c) $(x, y) \to (x', y)$ (d) $(x, y) \to (x', y)$

(e) $(x, y) \to (y', x)$ (f) $(x, y) \to (y, x')$

Figure 2: Illustration of possible shift moves, where each subcaption indicates the terminal symbols involved in the corresponding production rule.

[2] P. Clote. Asymptotic connectivity for the network of RNA secondary structures. arXiv:1508.03815 [q-bio.BM], August 2015.

[3] P Clote. Expected degree for RNA secondary structure networks. *J Comp Chem*, 36(2):103–17, Jan 2015.

[4] P. Clote and A. Bayegan. Network Properties of the Ensemble of RNA Structures. *PLoS. One.*, 10(10):e0139476, 2015.

[5] R. Cont and E. Tanimura. Small-world graphs: characterization and alternative constructions. *Adv. in Appl. Probab.*, 40(4):939–965, 2008.

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

[7] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University, 2009. ISBN-13: 9780521898065.

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

[9] 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.

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

[11] M. E. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, August 2001.

[12] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[13] S. Wuchty. Small worlds in RNA structures. *Nucleic. Acids. Res.*, 31(3):1108–1117, February 2003.