Small-world networks and RNA secondary structures

Defne Surujon, Yann Ponty, Peter Clote

Small-world networks and RNA secondary structures

Defne Surujon[1] Yann Ponty[2] Peter Clote[1*]

1: Biology Department, Boston College, USA. Chestnut Hill, MA 02467
2: Laboratoire d'Informatiques, Ecole Polytechnique, 91128 Palaiseau Cedex - France.

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

Small-world networks, first introduced in Watts and Strogatz (1998), satisfy two properties: (1) the *shortest path distance* between any two nodes is "small" (intuitively, there are six degrees of separation Guare (1990) between any two persons), and (2) the average *clustering coefficient* is large (intuitively, friends of a person tend to be friends of each other). Small-world networks appear to be ubiquitous in biology, sociology, and information technology; indeed, examples include the neural network of *Caenorhabditis elegans* Watts and Strogatz (1998), metabolic networks of 43 organisms representing all three domains of life Jeong et al. (2000), the gene co-expression in *S. cerevisiae* Van Noort et al. (2004), protein folding networks, where nodes correspond to conformations (self-avoiding walks on a 2D lattice) and edges exist between nodes that are connected by an elementary move Scala et al. (2001), and Markov state models of protein folding networks inferred by the software `MSMBuilder` Bowman et al. (2009) from molecular dynamics folding trajectories for the protein villin Bowman and Pande (2010), etc. Additionally, Wuchty (2003) showed by exhaustive enumeration of low energy RNA secondary structures of *E. coli* phe-tRNA, that the corresponding network architecture displays *small-world* properties. For additional examples, see the excellent review of Albert and Barabási Albert and Barabási (2002).

*Corresponding author: `clote@bc.edu`

In this paper, we investigate asymptotic properties of degree and clustering coefficient for the ensemble of all RNA secondary structures, by using methods from algebraic combinatorics. In particular, we rigorously prove that the network of RNA secondary structures is asymptotically *not* small-world, although it displays strong differences from random networks.

2 Preliminaries

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 Flamm et al. (2000) 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 Clote (2015b) and Clote and Bayegan (2015) 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 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 Newman et al. (2001), is given by

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

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 Cont and Tanimura (2008), 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.

Specific properties of RNA networks critically depend on the chosen definition of neighborhood of a structure, leading to the investigation of various move sets in the RNA kinetics literature. In addition to the move sets MS_1 and MS_2 Flamm et al. (2000), which latter also models *defect diffusion* Pörschke (1974), other groups have considered more general move sets that allow helix formation and disassociations Isambert and Siggia (2000).

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 Flajolet and Odlyzko (1990) 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 Flajolet and Odlyzko (1990). 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 3, we show that the average MS_2 degree of $\mathcal{S}_n$ is $O(n)$. In Section 4.2 [resp. 4.2] 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.

3 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 Clote (2015a), 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 $MS_1 degree(n)$ denotes the MS_1 expected network degree for a homopolymer, then

$$MS_1 degree(n) \sim 0.473475 \cdot n$$

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 Lorenz et al. (2008) and Flajolet and Sedgewick (2009), 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 Lorenz et al. (2008), or as the *symbolic method* Flajolet and Sedgewick (2009) – see Table 1. In this fashion, we obtain the following:

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

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

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

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 Flajolet and Odlyzko (1990), 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 Clote (2015b) and Clote and Bayegan (2015). We also note that the current approach is *much* simpler than the program in Clote and Bayegan (2015), although the latter is more general, since it computes the MS_2 degree for any user-specified RNA sequence. Using well-known methods, these asymptotic results can be extended from homopolymers to RNA sequences with Watson-Crick and wobble base pairs by using a "stickiness model", which stipulates the probability p that any two positions can form a base pair, defined by

$$p = 2\,(p_A\,p_U + p_G\,p_U + p_G\,p_C) \tag{5}$$

where p_A, p_C, p_G, p_U are user-specified nucleotide relative frequencies. Since we consider shifts, we need an additional stickiness parameter q, which specifies the probability that a shift can occur between three randomly selected positions in which one position is fixed, defined by

$$q = p_A\,p_U^2 + p_C\,p_G^2 + p_G\left(p_C^2 + p_U^2 + 2p_C\,p_U\right) + p_U\left(p_A^2 + p_G^2 + 2p_A\,p_G\right) \tag{6}$$

By including stickiness parameters into our computations, we obtained values presented in Table 2, which shows asymptotic MS_1 and MS_2 degrees for a number of classes of RNA.

4 Asymptotic MS_2 clustering coefficient

Section 4.1 describes a grammar to count the number of triangles for $\mathcal{S}_n$ with respect to MS_2 moves, while Section 4.2 describes a grammar to count two particular triples.

4.1 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:

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

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:

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

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:

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

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:

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

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:

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

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:

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

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:

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

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*

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

4.2 Counting triples

In this section, 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 (7) 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.

$$\begin{aligned} S^{\ddagger} &\to S^{\ddagger}\,\bullet \mid \texttt{(}\,S^{\ddagger}\,\texttt{)} \mid S\,\texttt{(}\,S^{\ddagger}\,\texttt{)} \mid S^{\ddagger}\,\texttt{(}\,R\,\texttt{)} \mid \\ &\quad \texttt{[}\,S^{\dagger}\,\texttt{]} \mid S\,\texttt{[}\,S^{\dagger}\,\texttt{]} \mid S^{\dagger}\,\texttt{[}\,R\,\texttt{]} \mid S^{\dagger}\,\texttt{(}\,S^{\dagger}\,\texttt{)} \\ S^{\dagger} &\to S^{\dagger}\,\bullet \mid \texttt{(}\,S^{\dagger}\,\texttt{)} \mid S\,\texttt{(}\,S^{\dagger}\,\texttt{)} \mid S^{\dagger}\,\texttt{(}\,R\,\texttt{)} \mid \\ &\quad \texttt{[}\,R\,\texttt{]} \mid S\,\texttt{[}\,R\,\texttt{]} \end{aligned} \tag{7}$$

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)

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

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

5 Discussion

In this paper, we have used methods from algebraic combinatorics Flajolet and Sedgewick (2009) 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 conclusions drawn from computer simulations of Bowman and Pande (2010); Wuchty (2003), which suggest that molecular folding networks are small-world. However, this paradoxical result is possible, since the notion of *finite* small-world network is not precisely defined due to absence of an exact bound for both average path length between any two nodes and for the clustering coefficient.

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

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern Physics*, 74:47–97.

Bowman, G. R., Huang, X., and Pande, V. S. (2009). Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods*, 49(2):197–201.

Bowman, G. R. and Pande, V. S. (2010). Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U.S.A.*, 107(24):10890–10895.

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

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

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

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

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

Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University. ISBN-13: 9780521898065.

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

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

Guare, J. (1990). *Six degrees of separation: A play.* Vintage Books, New York.

Isambert, H. and Siggia, E. D. (2000). Modeling RNA folding paths with pseudoknots: application to hepatitis *delta* virus ribozyme. *Proc. Natl. Acad. Sci. U.S.A.*, 97(12):6515–6520.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654.

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

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2014). Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, 0(O):O.

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

Pörschke, D. (1974). Model calculations on the kinetics of oligonucleotide double-helix coil transitions: Evidence for a fast chain sliding reaction. *Biophys Chem*, 2(2):83–96.

Scala, A., Nunes Amaral, L., and Barthélémy, M. (2001). Small-world networks and the conformation space of a short lattice polymer chain. *Europhys. Lett.*, 55(4):594–600.

Van Noort, V., Snel, B., and Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.*, 5(3):280–284.

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

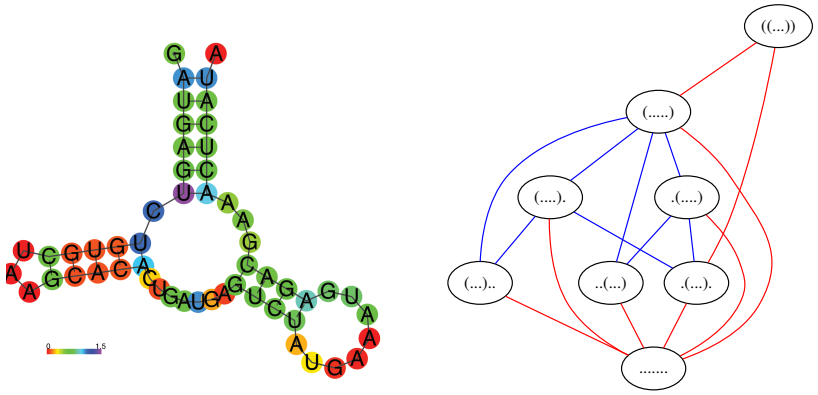
Wuchty, S. (2003). Small worlds in RNA structures. *Nucleic. Acids. Res.*, 31(3):1108–1117.

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 $A(z)$, $B(z)$, $C(z)$.

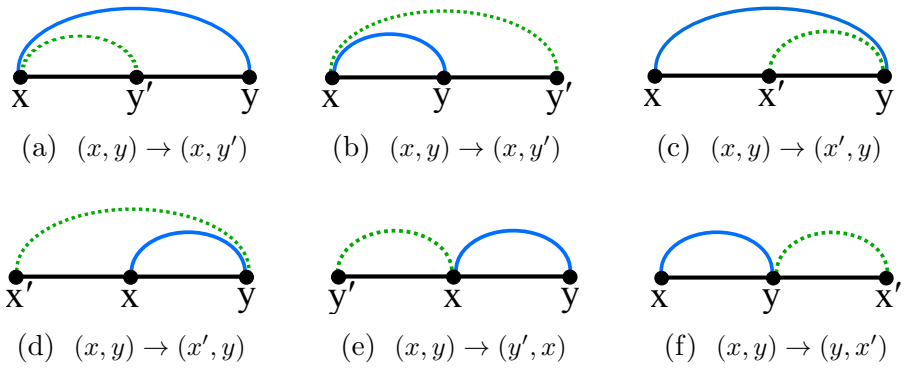
Move set	θ	hp	wc	wcw	wc tRNA	wcw tRNA
MS_1	1	0.55279	0.42265	0.46158	0.42157	0.46021
MS_1	3	0.47348	0.35130	0.38531	0.35038	0.38408
$\mathsf{MS}_2 - \mathsf{MS}_1$	1	1.44721	0.57735	0.84796	0.57985	0.85119
$\mathsf{MS}_2 - \mathsf{MS}_1$	3	1.08369	0.42908	0.31409	0.21550	0.63055
MS_2	1	2.00000	1.00000	1.30954	1.00142	1.31140
MS_2	3	1.55717	0.43527	0.32336	0.22162	0.63971

Table 2: The asymptotic expected degree of the network $\mathcal{S}_n^{\theta,p}$ of secondary structures for move sets MS_1 and MS_2 for different values of threshold θ and base pair and triple stickiness parameters p and q, defined in Equations (5) and (6) respectively. Five models are considered: `hp` – homopolymer model with $p = q = 1$; `wc` – Watson-Crick pairing model with uniform compositional frequency $p_A = p_C = p_G = p_U = \frac{1}{4}$, hence $p = \frac{1}{4}$, and $q = 0.0625$); `wcw` – Watson-Crick and wobble pairing model with uniform compositional frequency, hence $p = \frac{3}{8}$, and $q = 0.15625$; `wc tRNA` – Watson-Crick base pairing model based on the compositional frequencies ($p_A = \frac{1288}{4534}$, $p_U = \frac{1029}{4534}$, $p_G = \frac{1223}{4534}$, $p_U = \frac{994}{4534}$) observed in family RF00005 of 4,534 tRNAs in the Rfam 12.0 database Nawrocki et al. (2014), hence $p = 0.259427$, $q = 0.066394$; `wcw tRNA` – Watson-Crick and Wobble base pairing model using compositional frequency of RF00005, so that $p = 0.377699$, $q = 0.158476$.



(a) PLMVd (b) RNA network

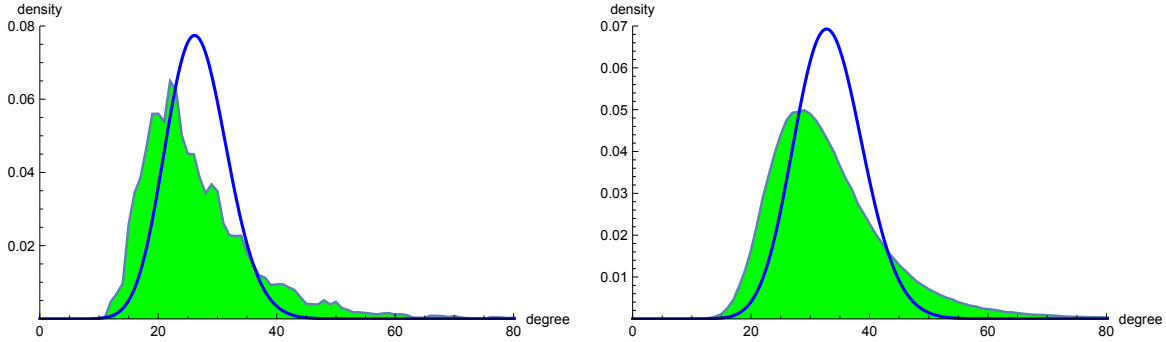
Figure 1: *(a)* Consensus secondary structure of the type III hammerhead ribozyme from Peach Latent Mosaic Viroid (PLMVd) AJ005312.1/282-335 (isolate LS35, variant ls16b), taken from Rfam Gardner et al. (2011) family RF00008. *(b)* Network for size 7 homopolymer with $\theta = 3$, having 8 nodes and 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 $\frac{32}{8} = 4$.



(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.



(a) MS_2 degree distribution (b) log-log plot

Figure 3: (a) MS_2-degree distribution for the 106,633 secondary structures for a 20-nt homopolymer with $\theta = 3$ (green shaded curve), with Poisson distribution of the same mean. (b) MS_2-degree distribution for the 32 nt selenocystein insertion (SECIS) element fruA with sequence CCUCGAGGGG AACCCGAAAG GGACCCGAGA GG.

A Supplement: Complete grammar

In this supplement, we list the complete grammar for all triangles and connected triples, together with the corresponding functional equations. We do not provide any detailed explanation for the complete listing of all possible triples or the complete grammar, since our intent is to provide suggestive supplementary figures and a general orientation for the reader wishing to work through the Mathematica code, available upon request, for our computation of the asymptotic clustering coefficient value $20.7728/n$.

A.1 Triangles and type A triples

Nonterminal S generates all non-empty secondary structures; R [resp. $S2$] generates all secondary structures of length at least 3 [resp. 2]. The non-terminal $S^\dagger$ [resp. $S^\ddagger$] generates all secondary structures containing a unique occurrence of a motif that has exactly 1 [resp. 2] extended connected components. Thus $S^\ddagger$ generates all structures having a unique occurrence of motif 1 or 2 for type A triple, as shown in Figure 7.

$$
\begin{aligned}
S &\to \bullet \mid S\,\bullet \mid X\,(\,R\,)\\
R &\to \theta \mid R\,\bullet \mid X\,(\,R\,)\\
S2 &\to \bullet\bullet \mid R\\
\theta &\to \bullet\bullet\bullet\\
X &\to \lambda \mid S\\
S^\dagger &\to S^\dagger\,\bullet \mid X\,(\,S^\dagger\,) \mid S^\dagger\,(\,R\,) \mid X\,[\,R\,]\\
S^\ddagger &\to S^\ddagger\,\bullet \mid X\,(\,S^\ddagger\,) \mid S^\ddagger\,(\,R\,) \mid X\,[\,S^\dagger\,] \mid S^\dagger\,[\,R\,] \mid S^\dagger\,(\,S^\dagger\,) \qquad \text{(Note: } \times 4\text{)}\\
PK &\to PK\,\bullet \mid X\,(\,PK\,) \mid PK\,(\,R\,) \mid X\,[\,\{\,S2\,]\,\} \mid X\,[\,S2\,\{\,]\,S2\,\} \mid\\
&\quad X\,[\,\{\,S2\,]\,S\,\} \mid X\,[\,S\,\{\,S2\,]\,\} \mid X\,[\,S\,\{\,S\,]\,S\,\}\\
\Lambda_A &\to PK \mid S^\ddagger \qquad \text{(Type A triples)}\\
\Delta_A &\to \Delta_A\,\bullet \mid X\,(\,\Delta_A\,) \mid \Delta_A\,(\,R\,) \mid X\,\langle\,R\bigstar R\,\rangle \mid X\bigstar R\,\rangle\,X\,\rangle \mid X\,\langle\,X\,\langle\,R\bigstar \qquad \text{(type A triangles)}\\
\Delta_B &\to \Delta_B\,\bullet \mid X\,(\,\Delta_B\,) \mid \Delta_B\,(\,R\,) \mid X\bigstar R\,\rangle\,X\,\rangle\,X\,\rangle \mid X\,\langle\,X\,\langle\,X\,\langle\,R\bigstar \mid\\
&\quad X\,\langle\,X\,\langle\,R\bigstar R\,\rangle \mid X\,\langle\,R\bigstar R\,\rangle\,X\,\rangle \mid X\bigstar R\bigstar R\bigstar \qquad \text{(type B triangles)}\\
\Delta &\to \Delta_A \mid \Delta_B \qquad \text{(all triangles)}
\end{aligned}
$$

A.2 Type B triples

The nonterminal symbol Λ_B generates all secondary structures containing a unique occurrence of one of the 12 type B triple motifs, as depicted in Figure 8. Nonterminals $R_1, \ldots, R_{12}$ generate respectively the collection of secondary structures containing a unique occurrence

of motif $1, \ldots, 12$. Note that $S2$ is not the same as R_2 – the former nonterminal $S2$ simply generates all secondary structures of length 2 or greater.

$$
\begin{aligned}
\Lambda_B &\to \Lambda_B \bullet \mid X \texttt{(} \Lambda_B \texttt{)} \mid \Lambda_B \texttt{(} R \texttt{)} \mid R_1 \mid \cdots \mid R_{12} \qquad \text{(type B triples)} \\
R_1 &\to X\bigstar R\bigstar R \rangle\, X \rangle \\
R_2 &\to X\bigstar X \langle\, R\bigstar X \rangle \\
R_3 &\to X\bigstar R\bigstar S2 \rangle \rangle \mid X\bigstar R\bigstar \rangle\, S2 \mid X\bigstar R\bigstar S \rangle\, S \rangle \\
R_4 &\to X\bigstar \langle\, S2 \rangle \bigstar \mid X\bigstar S2 \langle \rangle\, S2\bigstar \mid X\bigstar \langle\, S2 \rangle\, S\bigstar \mid X\bigstar S \langle\, S2 \rangle \bigstar \mid X\bigstar S \langle\, S \rangle\, S\bigstar \\
R_5 &\to X \langle\, X\bigstar R\bigstar X \rangle \\
R_6 &\to X \langle\, X\bigstar R \rangle\, X\bigstar \\
R_7 &\to X\bigstar R \rangle\, X\bigstar R \rangle \\
R_8 &\to X\bigstar R \rangle\, X \langle\, R\bigstar \\
R_9 &\to X \langle\, R\bigstar R\bigstar R \rangle \\
R_{10} &\to X \langle\, X \langle\, R\bigstar R\bigstar \\
R_{11} &\to X \langle\, R\bigstar X \langle\, R\bigstar \\
R_{12} &\to X \langle\, S2 \langle\, \bigstar R\bigstar \mid X \langle\, S \langle\, S\bigstar R\bigstar \mid X \langle\, \langle\, S2\bigstar R\bigstar
\end{aligned}
$$

A.3 Type C and D triples

The nonterminal symbol *triples* generates all secondary structures containing a unique occurrence of one of the 25 type C or D triple motifs, as depicted in Figure 9. Nonterminals Λ_{C1}, Λ_{C2}, Λ_{C367}, Λ_{C4}, Λ_{C5}, Λ_{C8}, Λ_{C9}, Λ_{C10}, Λ_{C11}, Λ_{C12}, Λ_{C13}, Λ_{C14}, Λ_{C15}, $\Lambda_{C\text{disc}}$, Λ_{C20}, Λ_{C21}, Λ_{C22}, Λ_{C23} generate respectively the collection of secondary structures containing a unique occurrence of motif $1, \ldots, 25$, whereby nonterminal Λ_{C367} is the same rule for motif 3,6 and 7, and nonterminal $\Lambda_{C\text{disc}}$ is the same rule for all *disconnected* <u>*successive*</u> motifs – i.e. the 6 type C and D triple motifs 16, 17, 18, 19, 24, 25 which have exactly 2 extended connected components of successive form `[ ] [ ]`.

$$
\begin{aligned}
\Lambda_C &\to \Lambda_C \bullet \mid X\,(\,\Lambda_C\,) \mid \Lambda_C\,(\,R\,) \mid \Lambda_{C1} \mid \Lambda_{C2} \mid \Lambda_{C367} \mid \\
&\quad \Lambda_{C4} \mid \Lambda_{C5} \mid \Lambda_{C8} \mid \Lambda_{C9} \mid \Lambda_{C10} \mid \Lambda_{C11} \mid \Lambda_{C12} \mid \\
&\quad \Lambda_{C13} \mid \Lambda_{C14} \mid \Lambda_{C15} \mid \Lambda_{C_{\rm disc}} \mid \Lambda_{C20} \mid \Lambda_{C21} \mid \Lambda_{C22} \mid \Lambda_{C23} \\
\Lambda_{C1} &\to X\bigstar S^\dagger \rangle\, X \rangle \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C2} &\to X\bigstar \langle\, S2 \rangle\, \rangle\, X \rangle \mid X\bigstar S2 \langle\, \rangle\, S2 \rangle\, X \rangle \mid X\bigstar \langle\, S2 \rangle\, S \rangle\, X \rangle \mid \\
&\quad X\bigstar S \langle\, S2 \rangle\, \rangle\, X \rangle \mid X\bigstar S \langle\, S \rangle\, S \rangle\, X \rangle \\
\Lambda_{C367} &\to X\,[\,\Delta_A\,] \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C4} &\to X \langle\, X\bigstar R \rangle\, X \rangle\, X \rangle \\
\Lambda_{C5} &\to X\bigstar R \rangle\, S^\dagger \rangle \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C8} &\to X \langle\, R\bigstar S^\dagger \rangle \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C9} &\to X \langle\, \langle\, S2\bigstar R \rangle\, X \rangle \mid X \langle\, S2 \langle\, \bigstar R \rangle\, X \rangle \mid X \langle\, S \langle\, S\bigstar R \rangle\, X \rangle \\
\Lambda_{C10} &\to X \langle\, X \langle\, X \langle\, R\bigstar X \rangle \\
\Lambda_{C11} &\to X \langle\, X \langle\, S^\dagger\bigstar \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C12} &\to X\bigstar R \rangle\, X \langle\, \rangle\, S2 \rangle \mid X\bigstar R \rangle\, X \langle\, S2 \rangle\, \rangle \mid X\bigstar R \rangle\, X \langle\, S \rangle\, S \rangle \\
\Lambda_{C13} &\to X \langle\, X \langle\, R\bigstar \rangle\, S2 \rangle \mid X \langle\, X \langle\, R\bigstar S2 \rangle\, \rangle \mid X \langle\, X \langle\, R\bigstar S \rangle\, S \rangle \\
\Lambda_{C14} &\to X \langle\, R\bigstar \langle\, S2 \rangle\, \rangle \mid X \langle\, R\bigstar S2 \langle\, \rangle\, S2 \rangle \mid \\
&\quad X \langle\, R\bigstar \langle\, S2 \rangle\, S \rangle \mid X \langle\, R\bigstar S \langle\, S2 \rangle\, \rangle \mid X \langle\, R\bigstar S \langle\, S \rangle\, S \rangle \\
\Lambda_{C15} &\to X \langle\, X \langle\, \langle\, S2 \rangle\, \bigstar \mid X \langle\, X \langle\, S2 \langle\, \rangle\, S2\bigstar X \langle\, X \langle\, \langle\, S2 \rangle\, S\bigstar \mid \\
&\quad X \langle\, X \langle\, S \langle\, S2 \rangle\, \bigstar \mid X \langle\, X \langle\, S \langle\, S \rangle\, S\bigstar \\
\Lambda_{C_{\rm disc}} &\to \Delta_A\,[\,R\,] \mid \Delta_A\,(\,S^\dagger\,) \mid S^\dagger \langle\, R\bigstar R \rangle \mid S^\dagger \langle\, X \langle\, R\bigstar \mid S^\dagger\bigstar R \rangle\, X \rangle \mid \\
&\quad S^\dagger\,(\,X \langle\, R\bigstar R \rangle\, X\,) \mid S^\dagger\,(\,X\bigstar R \rangle\, X \rangle\, X\,) \mid S^\dagger\,(\,X \langle\, X \langle\, R\bigstar X\,) \\
\Lambda_{C20} &\to X \langle\, S^\dagger\bigstar R \rangle \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C21} &\to X \langle\, S^\dagger \langle\, R\bigstar \qquad \text{(Note: } \times 4\text{)} \\
\Lambda_{C22} &\to X \langle\, \langle\, S2 \rangle\, \bigstar R \rangle \mid X \langle\, S2 \langle\, \rangle\, S2\bigstar R \rangle \mid X \langle\, \langle\, S2 \rangle\, S\bigstar R \rangle \mid \\
&\quad X \langle\, S \langle\, S2 \rangle\, \bigstar R \rangle \mid X \langle\, S \langle\, S \rangle\, S\bigstar R \rangle \\
\Lambda_{C23} &\to X \langle\, S2 \langle\, \rangle\, X \langle\, R\bigstar \mid X \langle\, \langle\, S2 \rangle\, X \langle\, R\bigstar \mid X \langle\, S \langle\, S \rangle\, X \langle\, R\bigstar
\end{aligned}
$$

Finally, the collection of all secondary structures having a unique motif for a connected triple (both non-triangular triples of types A,B,C,D or deriving from triangles) is generated by the rule

$$\Lambda \to \Lambda_A \mid \Lambda_B \mid \Lambda_C \mid \mathit{triangles} \qquad \text{(Note: triangles must be multiplied by 3)}$$

This gives rise to the following functional equations for *all* connected triples, both non-triangular triples, as well as 3 triples associated with each triangle. Since rules $R_1, \ldots, R_{12}$

are written as $R1, \ldots, R12$, we write `S2` in place of $R2$, which latter had been defined by $R2 \to \bullet\bullet \mid R$.

$$
\begin{aligned}
S^{\ddagger} &= S^{\ddagger} \cdot z + X \cdot z \cdot S^{\ddagger} \cdot z + S^{\ddagger} \cdot z \cdot R \cdot z + X \cdot z \cdot S^{\dagger} \cdot z + S^{\dagger} \cdot z \cdot R \cdot z + S^{\dagger} \cdot z \cdot S^{\dagger} \cdot z \\
S^{\dagger} &= S^{\dagger} \cdot z + X \cdot z \cdot S^{\dagger} \cdot z + S^{\dagger} \cdot z \cdot R \cdot z + X \cdot z \cdot R \cdot z \\
S &= z + S \cdot z + X \cdot z \cdot R \cdot z \\
R &= \theta + R \cdot z + X \cdot z \cdot R \cdot z \\
\theta &= z \cdot z \cdot z \\
X &= 1 + S \\
S2 &= z \cdot z + R \\
PK &= PK \cdot z + X \cdot z \cdot PK \cdot z + PK \cdot z \cdot R \cdot z + X \cdot z^4 (S2 + S2^2 + 2SS2 + S^3) \\
\Lambda_A &= PK + 4 \cdot S^{\ddagger} \\
R1 &= X^2 z^4 R^2 \\
R2 &= X^3 z^4 R \\
R3 &= X z^4 R (2S2 + S^2) \\
R4 &= X z^4 (S2 + S2^2 + 2SS2 + S^3) \\
R5 &= X^3 z^4 R \\
R6 &= X^3 z^4 R \\
R7 &= X^2 z^4 R^2 \\
R8 &= X^2 z^4 R^2 \\
R9 &= X z^4 R^3 \\
R10 &= X^2 z^4 R^2 \\
R11 &= X^2 z^4 R^2 \\
R12 &= X z^4 (2S2 + S^2)
\end{aligned}
$$

together with the following, where we write Λ_C in place of Λ_{CD} for reasons of brevity

$$
\begin{aligned}
\Lambda_B &= R1 + R2 + R3 + R4 + R5 + R6 + R7 + R8 + R9 + R10 + R11 + R12 + \\
&\quad \Lambda_B z + X z^2 \Lambda_B + \Lambda_B z^2 R \\
\Lambda_{C1} &= X z S^\dagger z X z \\
\Lambda_{C2} &= X z^5 (S2 + S2^2 + 2SS2 + S^3) \\
\Lambda_{C367} &= X z \Delta_A z \\
\Lambda_{C4} &= X^4 z^5 R \\
\Lambda_{C5} &= X z R z S^\dagger z \\
\Lambda_{C8} &= X z R z S^\dagger z \\
\Lambda_{C9} &= X^2 z^5 (2S2 + S^2) \\
\Lambda_{C10} &= X^4 z^5 R \\
\Lambda_{C11} &= X^2 z^3 S^\dagger \\
\Lambda_{C12} &= X^2 z^5 R (2S2 + S^2) \\
\Lambda_{C13} &= X^2 z^5 R (2S2 + S^2) \\
\Lambda_{C14} &= X z^5 R (S2 + S2^2 + 2SS2 + S^3) \\
\Lambda_{C15} &= X^2 z^5 (S2 + S2^2 + 2SS2 + S^3) \\
\Lambda_{C\,disc} &= \Delta_A z R z + \Delta_A z S^\dagger z + (S^\dagger z R z R z + S^\dagger z X z R z + S^\dagger z R z X z)(1 + X^2 z^2) \\
\Lambda_{C20} &= X z^3 R S^\dagger \\
\Lambda_{C21} &= X z^3 R S^\dagger \\
\Lambda_{C22} &= X z^5 R (S2 + S2^2 + 2SS2 + S^3) \\
\Lambda_{C23} &= X^2 z^5 R (2S2 + S^2) \\
\Lambda_C &= \Lambda_C \cdot z + X \cdot z \cdot \Lambda_C \cdot z + \Lambda_C \cdot z \cdot R \cdot z + 4 \cdot \Lambda_{C1} + \\
&\quad \Lambda_{C2} + 4 \cdot \Lambda_{C367} + \Lambda_{C4} + 4 \cdot \Lambda_{C5} + 4 \cdot \Lambda_{C8} + \Lambda_{C9} + \\
&\quad \Lambda_{C10} + 4 \cdot \Lambda_{C11} + \Lambda_{C12} + \Lambda_{C13} + \Lambda_{C14} + \Lambda_{C15} + \\
&\quad \Lambda_{C\,disc} + 4 \cdot \Lambda_{C20} + 4 \cdot \Lambda_{C21} + \Lambda_{C22} + \Lambda_{C23} \\
\Delta_A &= \Delta_A \cdot z + X \cdot z \cdot \Delta_A \cdot z + \Delta_A \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 \\
\Delta_B &= \Delta_B \cdot z + X \cdot z \cdot \Delta_B \cdot z + \Delta_B \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 \\
\Delta &= \Delta_A + \Delta_B \\
\Lambda &= \Lambda_A + 4 \cdot \Lambda_B + \Lambda_C + 3 \cdot \Delta
\end{aligned}
$$

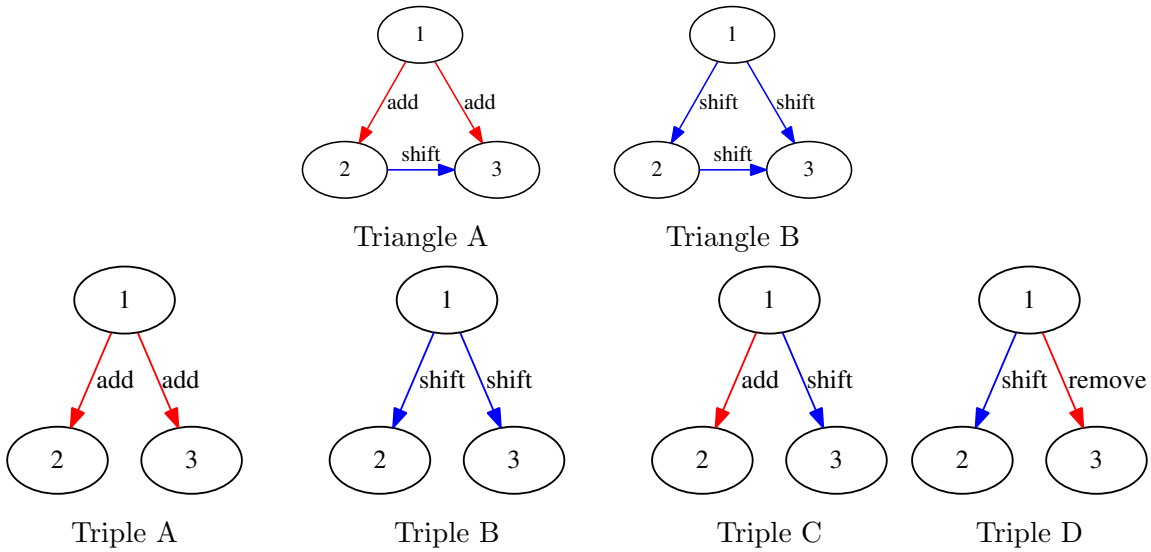


Figure 4: Complete listing of all possible moves for triangles and non-triangular connected triples with a designated first structure, up to equivalence. Consider, for instance, the triangle T (not shown) in which $s_1 \to s_2$ by a shift, $s_1 \to s_3$ by base pair removal, and $s_3 \to s_2$ by a base pair addition. Then T is equivalent to a triangle of type A, where node 1 is occupied by s_3, node 2 resp 3 is occupied by s_1 resp. s_2. Other instances of triangles the reader may consider are analogously equivalent to a triangle of type A or B.

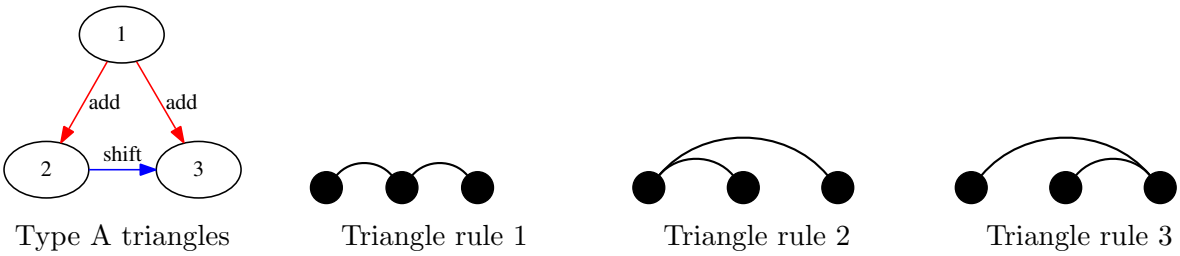


Figure 5: Type A triangles are constituted by structures s_1, s_2, s_3, where s_2, s_3 are obtained from s_1 by adding a base pair with the property that there is a shift move from s_2 to s_3. This type of triangle is described in rules 1,2,3 in Section 4.1. For instance, the motif for triangle rule 1 indicates that there is a base pair $(x, y) \in s_2$ that can be shifted to the base pair $(y, z) \in s_3$; similarly, the motif for rule 2 [resp. 3] indicates that there is a base pair $(x, y) \in s_2$ [resp. $(x, z) \in s_2$] that can be shifted to the base pair $(x, z) \in s_3$ [resp. $(y, z) \in s_3$].

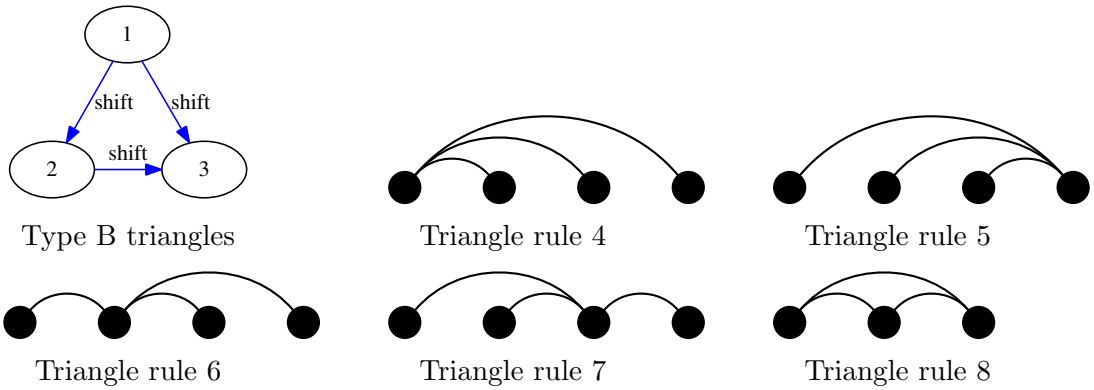


Figure 6: Type B triangles are constituted by structures s_1, s_2, s_3, where s_2, s_3 are obtained from s_1 by a shift with the property that there is a shift move from s_2 to s_3. This type of triangle is described in rules 4-8 in Section 4.1. For instance, the motif in triangle rule 4 indicates that there is a base pair $(x, y) \in s_1$ can be shifted to base pair $(x, z) \in s_2$ and $(x, w) \in s_3$. The other panels have analogous meanings.

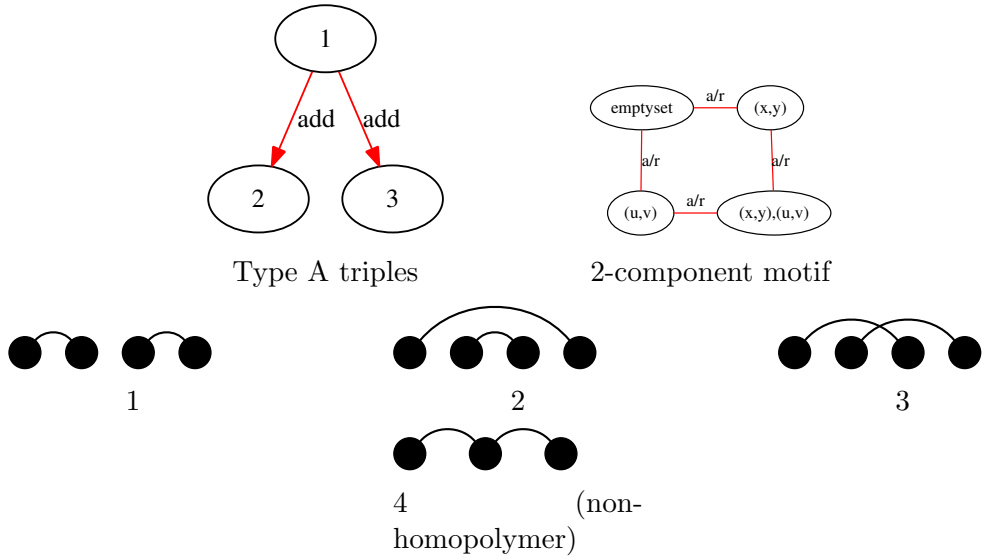


Figure 7: Type A triples are constituted by structures s_1, s_2, s_3, where s_2, s_3 are obtained from s_1 by a base pair addition with the property that there is a no MS_2 move between s_2 and s_3. Type A triples are given in rules 1-3 in Section 4.2, are are represented in panels 1,2,3,4 of this figure. Panel 1 [resp. 2] indicates that structures s_2, s_3 can be obtained from structure s_1 by the addition of *disjoint* base pairs (u, v) (x, y) which are *not nested*, i.e. `( ) ( )`, [resp. which are *nested*, i.e. `( ( ) )`]. Panel 3 indicates that structures s_2, s_3 can be obtained from structure s_1 by the addition of *disjoint* base pairs (u, v) (x, y) which would form a pseudoknot if added simultaneously to s_1, i.e. `( [ ) ]`. Panel 4, which is identical to panel 1 of Figure 5, represents a non-triangular connected triple *only* in the non-polymer case. This panel indicates that structures s_2, s_3 can be obtained from structure s_1 by the addition of *non-disjoint* base pairs (u, v) (v, w) which share a base. To each triple motif that has 2 *extended connected components* (see text) there corresponds a quadrilateral, as shown in the panel with label *2-component motif*, where edges are labeled by *a/r* for base pair addition/removal. To each corner of the quadrilateral, there corresponds a unique triple – thus, panels 1 and 2 actually each represent 4 triples. It follows that the average number of triples per structure for type A(1) and type A(2) triples must be multiplied by 4. The same remark holds for type C,D triples in in Figure 9, which have 2 extended connected components.

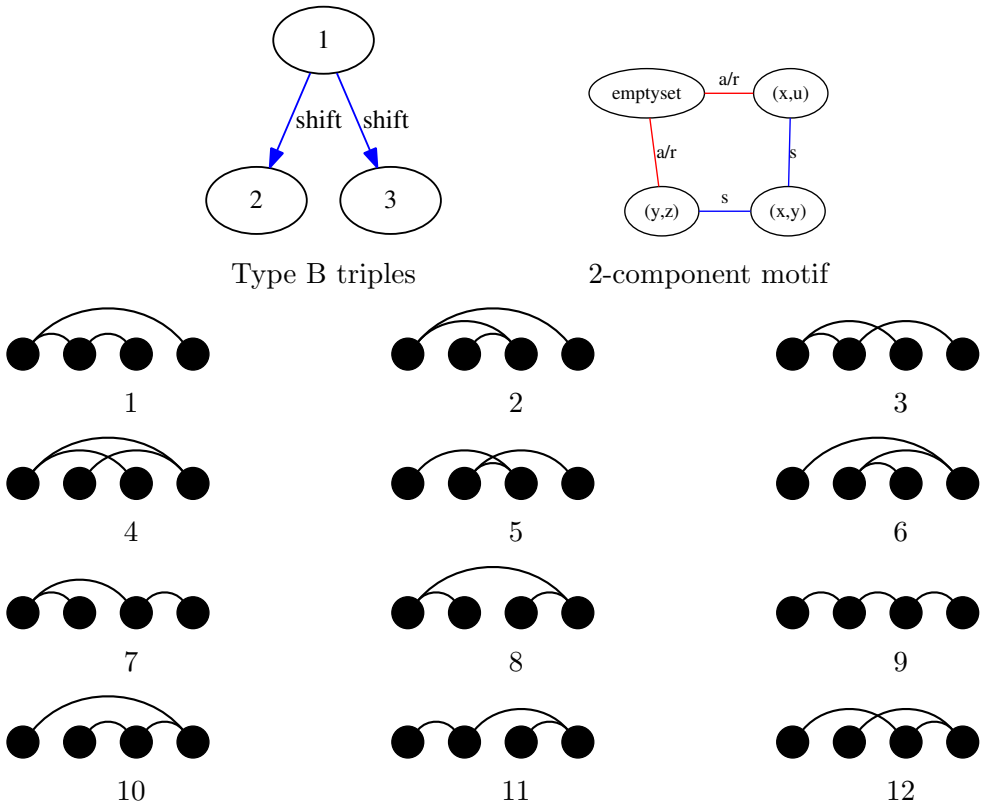


Figure 8: Type B triples are of the form s_1, s_2, s_3, where s_1, s_2 are connected by a shift, as are s_1, s_3, but s_2, s_3 are not connected by any MS_2 move. Note that all motifs are connected. To each type B triple motif, there corresponds a quadrilateral, shown in panel with label *2-component motif*, where edges are labeled by *a/r* for base pair addition/removal or by *s* for base pair shift. For each of these 12 motifs, there is a unique base pair that can shift to the remaining two base pairs – for instance, in motif 1, the base pairs are (x, y), (y, z) and (x, u), for $x < y < z < u$, where (x, y) can be shifted to each of (y, z) and (x, u). This uniquely defined base pair should be located in the quadrilateral diagonally opposite the corner labeled by *emptyset*. Since each corner of the quadrilateral corresponds to one of 4 triples associated with the motif, it follows that the average number of type B triples per structure must be multiplied by 4.

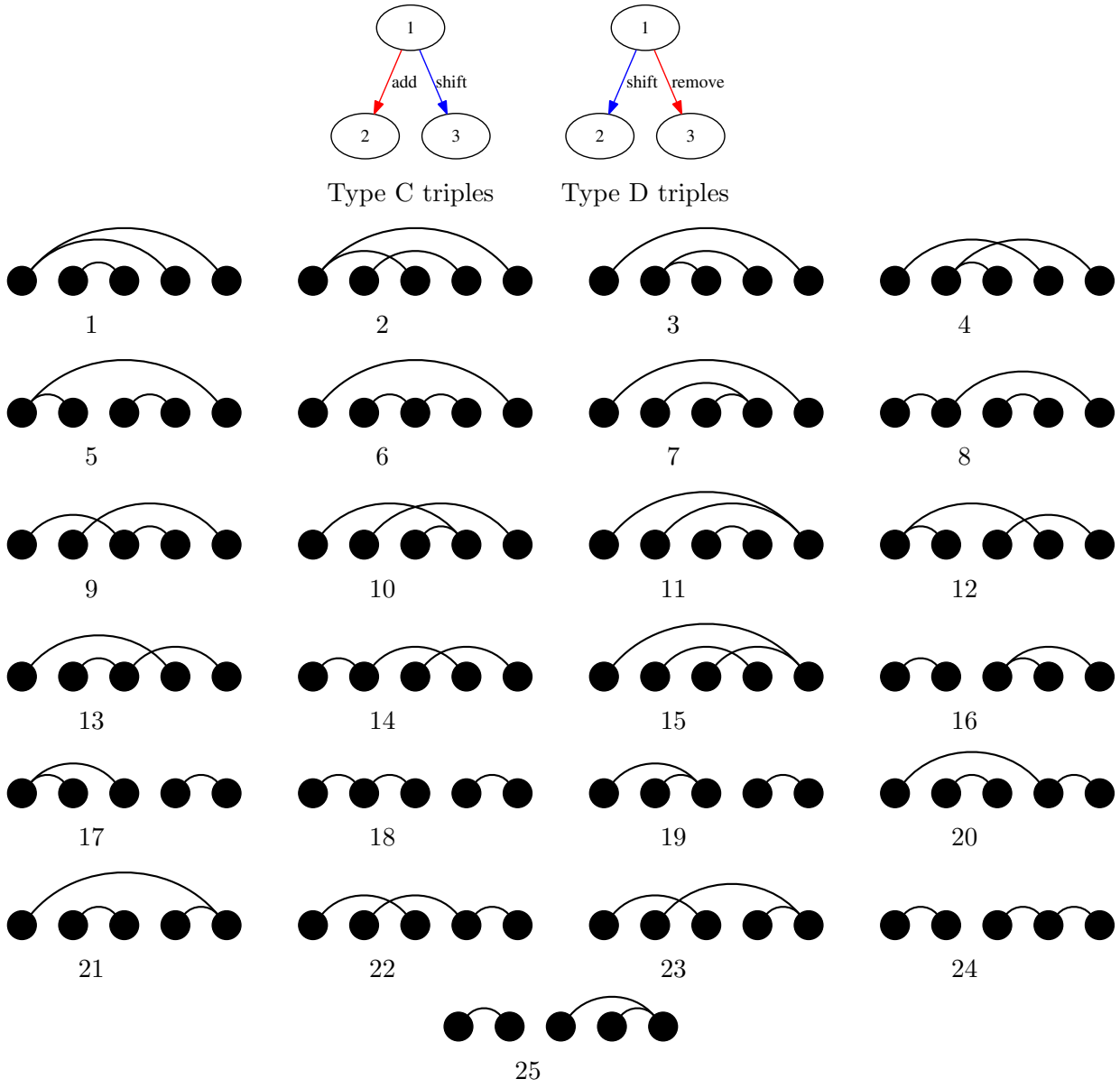


Figure 9: Type C and D triples are of the form s_1, s_2, s_3, where s_1, s_2 are connected by a shift, and s_3 is obtained from s_1 by a base pair addition, but s_2, s_3 are not connected by any MS_2 move. Triples may have either one or two *extended connected components* (see text); in the former case, the asymptotic expected number of triples is $O(n)$, while in the latter case the expected expected number of triples is $O(n^2)$, as is the case for the type A triple motifs in panels 1 and 2 of Figure 7. The 10 motifs having 1 extended connected component are: 2, 4, 9, 10, 12, 13, 14, 15, 22, 23. The 15 motifs having 2 extended connected component are: 1, 3, 5, 6, 7, 8, 11, 16, 17, 18, 19, 20, 21, 24, 25. Of the latter, the six motifs 16,17,18,19,24,25 are *disconnected successive*, and the nine motifs 1,3,5,6,7,8,11,20,21 are *disconnected nested*. To each motif, which has 2 extended connected components, there actually correspond 4 triples, as explained in the caption to Figure 7. For instance, to the motif 16, given by undirected edges $(1,2),(3,4),(3,5)$, there correspond four structures s_1, s_2, s_3, s_4, where $(3,4) \in s_1$, $(3,5) \in s_2$, $(1,2),(3,4) \in s_3$, $(1,2),(3,5) \in s_4$ with the following four triples: (1) $s_1 \to s_2$, $s_1 \to s_3$ (type C triple); (2) $s_2 \to s_3$, $s_2 \to s_1$ (type C triple); (3) $s_3 \to s_4$, $s_3 \to s_1$ (type D triple); (4) $s_4 \to s_3$, $s_4 \to s_2$ (type D triple). This is analogous to the situation summarized in the panel in Figure 7 with label *2-component motif*. Note that motifs 3,6,7 have a type A triangle contained within an outer designated base pair `[ ]`.