Combinatorics of RNA-RNA interaction

Thomas J. X. Li · Christian M. Reidys

Abstract RNA-RNA binding is an important phenomenon observed in various classes of non-coding RNAs and plays a crucial role in a number of regulatory processes. Recently (Alkan et al. 2000; Pervouchine 2004; Huang et al. 2010; Chitsaz et al. 2009) MFE folding algorithms for predicting the joint structure of two interacting RNA molecules have been proposed. Here joint structure means that the intramolecular structures of each partner is pseudoknot-free, that the intermolecular binding pairs are noncrossing, and that there is no so-called “zig-zag” configuration. This paper presents the combinatorics of these interaction structures including their generating function, singularity analysis as well as explicit recurrence relations. In particular, our results imply simple formulas for the asymptotic number of joint structures.

Keywords RNA-RNA interaction · Joint structure · Shape · Symbolic enumeration · Singularity analysis

Mathematics Subject Classification (2000) 05A16 · 92E10
1 Introduction

RNA-RNA binding is an important phenomenon observed in various classes of non-coding RNAs and plays a crucial role in a number of regulatory processes. Examples include the regulation of translation in both prokaryotes (Narberhaus and Vogel, 2007) and eukaryotes (McManus and Sharp, 2002; Bacherie and Slack, 2002), the targeting of chemical modifications (Bachellerie et al., 2002), insertion editing (Benne, 1992), and transcriptional control (Kugel and Goodrich, 2007).

More and more evidence suggests that RNA-RNA interactions also play a role for the functionality of long mRNA-like ncRNAs. A common theme in many RNA classes, including miRNAs, snRNAs, gRNAs, snoRNAs, and in particular many of the proaryotic small RNAs, is the formation of RNA-RNA interaction structures that are much more complex than simple complementary sense-antisense interactions. The interaction between two RNAs is governed by the same physical principles that determine RNA folding: the formation of specific base pairs patterns whose energy is largely determined by base pair stacking and loop strains. Therefore, secondary structures are an appropriate level of description to quantitatively understand the thermodynamics of RNA-RNA binding.

By restricting the space of allowed configurations in ways that are similar to pseudoknot folding algorithms (Rivas and Eddy, 1999), polynomial-time algorithms on secondary structure level can be derived. (Pervouchine, 2004) and (Alkan et al., 2006) proposed MFE folding algorithms for predicting the joint structure of two interacting RNA molecules. In this model, “joint structure” means that the intramolecular structures of each partner is pseudoknot-free, that the intermolecular binding pairs are noncrossing, and that there is no so-called “zig-zag” configuration (see below for details). This structure class seems to include all major interaction complexes. The optimal joint structure can be computed in \(O(N^6)\) time and \(O(N^4)\) space by means of dynamic programming (Pervouchine, 2004). More recently, extensions involving the partition function were proposed by (Chitsaz et al., 2009) (piRNA) and (Huang et al., 2009) (rip), see Fig. 1.

In contrast to the situation for RNA secondary structures, little is known about the joint structures that are the folding targets of rip (Huang et al., 2010). This paper closes this gap and introduces the combinatorics of interaction structures. We present the generating function of joint structures, its singularity analysis as well as explicit recurrence relations. In particular, our results imply simple formulas for the asymptotic number of joint structures.

The paper is organized as follows: in Section 2 we provide several basic facts and context. In Section 3 we introduce joint structures along the lines of (Huang et al., 2009). Section 4 is new: here we are inspired by the ideas of (Reidys et al., 2010) and consider shapes of joint structures. In Section 5 we use shapes in order to compute the generating function of joint structures and Section 6 deals with the singularity analysis. We then integrate our results in Section 7. Finally we present additional results in Section 8.
2 Some basic facts

2.1 Singularity analysis

Let \( f(z) = \sum_{n=0}^{\infty} a_n z^n \) be a generating function with nonnegative coefficients and a radius of convergence \( R > 0 \). In light of the fact that explicit formulas for the coefficients \( a_n \) can be very complicated or even impossible to obtain, we switch over to investigate the estimation of \( a_n \) in terms of the exponential factor \( \gamma \) and the subexponential factor \( P(n) \), that is, \( a_n \sim P(n) \gamma^n \). The derivation of exponential growth rate and subexponential factor is mainly based on singularity analysis. Singularity analysis is a framework that allows to analyze the asymptotics of these coefficients. The key to obtain the asymptotic information about the coefficients of a generating function is its dominant singularities, which raises the question on how to locate them. In the particular case of power series with nonnegative coefficients and a radius of convergence \( R > 0 \), a theorem of Pringsheim guarantees a positive real dominant singularity at \( z = R \). As we are dealing here with combinatorial generating functions we always have this dominant singularity. Furthermore for all our generating functions it is the unique dominant singularity. The class of theorems that deal with the deduction of information about coefficients from the generating function are called transfer-theorems.

To be precise, we say a function \( f(z) \) is \( \Delta_\rho \) analytic at its dominant singularity \( z = \rho \), if it analytic in some domain \( \Delta_\rho(\phi, r) = \{ z \mid |z| < r, z \neq \rho, |\arg(z - \rho)| > \phi \} \), for some \( \phi, r \), where \( r > |\rho| \) and \( 0 < \phi < \frac{\pi}{2} \). We use the notation

\[
(f(z) = \Theta(g(z)) \text{ as } z \to \rho) \iff (f(z)/g(z) \to c \text{ as } z \to \rho),
\]

where \( c \) is some constant. Let \( [z^n]f(z) \) denote the coefficient of \( z^n \) in the power series expansion of \( f(z) \) at \( z = 0 \). Since the Taylor coefficients have the
property
\[ \forall \gamma \in \mathbb{C} \setminus 0; \quad [z^n]f(z) = \gamma^n[z^n]f\left(\frac{z}{\gamma}\right), \]

We can, without loss of generality, reduce our analysis to the case where \( z = 1 \) is the unique dominant singularity. The next theorem transfers the asymptotic expansion of a function around its unique dominant singularity to the asymptotic of the function’s coefficients.

**Theorem 1** \cite{Flajolet2007} Let \( f(z) \) be a \( \Delta_1 \) analytic function at its unique dominant singularity \( z = 1 \). Let
\[ g(z) = (1 - z)^\alpha \log^\beta \left(\frac{1}{1 - z}\right), \quad \alpha, \beta \in \mathbb{R}. \]

That is we have in the intersection of a neighborhood of 1
\[ f(z) = \Theta(g(z)) \quad \text{for } z \to 1. \tag{1} \]

Then we have
\[ [z^n]f(z) = \Theta([z^n]g(z)). \tag{2} \]

**Theorem 2** \cite{Flajolet2007} Suppose \( f(z) = (1 - z)^{-\alpha}, \quad \alpha \in \mathbb{C} \setminus \mathbb{Z}_{\leq 0} \), then
\[
\begin{align*}
  f(z) &\sim \frac{n^{\alpha-1}}{\Gamma(\alpha)} \left[ 1 + \frac{\alpha(\alpha - 1)}{2n} + \frac{\alpha(\alpha - 1)(\alpha - 2)(3\alpha - 1)}{24n^2} + \frac{\alpha^2(\alpha - 1)^2(\alpha - 2)(\alpha - 3)}{48n^3} + O\left(\frac{1}{n^4}\right) \right].
\end{align*}
\tag{3}
\]

### 2.2 Symbolic Enumeration

Symbolic enumeration \cite{Flajolet2007} plays an important role in the following computations. We first introduce the notion of a combinatorial class. Let \( z = (z_1, \ldots, z_d) \) be a vector of \( d \) formal variables and \( k = (k_1, \ldots, k_d) \) be a vector of integers of the same dimension. We use the simplified notation
\[ z^k := z_1^{k_1} \cdots z_d^{k_d}. \]

**Definition 1** A combinatorial class of \( d \) dimension, or simply a class, is an ordered pair \((\mathcal{A}, w_A)\) where \( \mathcal{A} \) is a finite or denumerable set and a size-function \( w_A : \mathcal{A} \to \mathbb{Z}_d^{\geq 0} \) satisfies that \( w_A^{-1}(n) \) is finite for any \( n \in \mathbb{Z}_d^{\geq 0} \).

Given a class \((\mathcal{A}, w_A)\), the size of an element \( a \in \mathcal{A} \) is denoted by \( w_A(a) \), or simply \( w(a) \). We consistently denote by \( \mathcal{A}_n \) the set of elements in \( \mathcal{A} \) that have size \( n \) and use the same group of letters for the cardinality \( A_n = |A_n| \).
The sequence \( \{A_n\} \) is called the counting sequence of class \( \mathcal{A} \). The generating function of a class \( (\mathcal{A}, w_\mathcal{A}) \) is given by

\[
A(z) = \sum_{a \in \mathcal{A}} z^{w_\mathcal{A}(a)} = \sum_n A_n z^n.
\]

There are two special classes: \( \mathcal{E} \) and \( \mathcal{Z}_i \) which contain only one element of size 0 and \( e_i \), respectively. In particular, the generating functions of the classes \( \mathcal{E} \) and \( \mathcal{Z}_i \) are

\[
E(z) = 1 \quad \text{and} \quad Z_i(z) = z^i.
\]

We adhere in the following to a systematic naming convention: classes, their counting sequences, and their generating functions are systematically denoted by the same groups of letters: for instance, \( \mathcal{C} \) for a class, \( \{C_n\} \) for the counting sequence, and \( C(z) \) for its generating function.

Next, we introduce a few fundamental constructions to define a broad variety of combinatorial structures. Let \( \mathcal{A} \) and \( \mathcal{B} \) be combinatorial classes of dimension. Suppose \( \mathcal{A}_i \) are combinatorial classes of 1 dimension. We define

\(-\) \( (\mathcal{A}_1, \mathcal{A}_2) := \{c = (a_1, a_2) \mid a_i \in \mathcal{A}_i\} \) and for \( c = (a_1, a_2) \in (\mathcal{A}_1, \mathcal{A}_2) \)

\[
w_{(\mathcal{A}_1, \mathcal{A}_2)}(c) = (w_{\mathcal{A}_1}(a_1), w_{\mathcal{A}_2}(a_2)).
\]

\(-\) \( \mathcal{A} + \mathcal{B} := \mathcal{A} \cup \mathcal{B} \), if \( \mathcal{A} \cap \mathcal{B} = \emptyset \) and for \( c \in \mathcal{A} + \mathcal{B} \),

\[
w_{\mathcal{A} + \mathcal{B}}(c) = \begin{cases} w_{\mathcal{A}}(c) & \text{if } c \in \mathcal{A} \\ w_{\mathcal{B}}(c) & \text{if } c \in \mathcal{B}. \end{cases}
\]

\(-\) \( \mathcal{A} \times \mathcal{B} := \{c = (a, b) \mid a \in \mathcal{A}, b \in \mathcal{B}\} \) and for \( c \in \mathcal{A} \times \mathcal{B} \),

\[
w_{\mathcal{A} \times \mathcal{B}}(c) = w_{\mathcal{A}}(a) + w_{\mathcal{B}}(b).
\]

\(-\) \( \text{SEQ}(\mathcal{A}) := \mathcal{E} + \mathcal{A} + (\mathcal{A} \times \mathcal{A}) + (\mathcal{A} \times \mathcal{A} \times \mathcal{A}) + \cdots. \)

Plainly, \( \text{SEQ}(\mathcal{A}) \) defines a proper combinatorial class if and only if \( \mathcal{A} \) contains no element of size 0. We immediately observe

**Proposition 1** Suppose \( \mathcal{A}, \mathcal{B} \) and \( \mathcal{C} \) are combinatorial classes of dimension with generating functions \( A(z), B(z) \) and \( C(z) \). Let \( \mathcal{A}_i \) be combinatorial classes of 1 dimension with generating functions \( A_i(z) \). Then

(a) \( \mathcal{C} = (\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_d) \implies C(z) = A_1(z) A_2(z) \cdots A_d(z) \)

(b) \( \mathcal{C} = \mathcal{A} + \mathcal{B} \implies C(z) = A(z) + B(z) \)

(c) \( \mathcal{C} = \mathcal{A} \times \mathcal{B} \implies C(z) = A(z) \cdot B(z) \)

(d) \( \mathcal{C} = \text{SEQ}(\mathcal{A}) \implies C(z) = \frac{1}{1 - A(z)}. \)
2.3 Secondary structures

Let \( f(n) \) denote the number of all noncrossing matchings of \( n \) arcs with generating function \( F(z) = \sum f(n) z^n \). From recursions we can easily derive

\[
z F(z)^2 - F(z) + 1 = 0.
\]

That is we have

\[
F(z) = \frac{1 - \sqrt{1 - 4z^2}}{2z}.
\]

Let \( \mathcal{T}_\sigma \) denote the combinatorial class of \( \sigma \)-canonical secondary structures having arc-length \( \geq 2 \) and \( T_\sigma(n) \) denote the number of all \( \sigma \)-canonical secondary structures with \( n \) vertices having arc-length \( \geq 2 \) and

\[
T_\sigma(z) = \sum T_\sigma(n) z^n
\]

be the generating function.

**Theorem 3** Suppose \( \sigma \in \mathbb{N}, \sigma \geq 1 \) and \( u_\sigma(z) = \frac{(z^2)^{\sigma-1}}{2^{\sigma-1} z^{2\sigma-1} + 1} \). Then

\[
T_\sigma(z) = \frac{1}{u_\sigma(z) z^2 - z + 1} F \left( \sqrt{ \frac{u_\sigma(z) z}{(u_\sigma(z) z^2 - z + 1)} } \right)^2.
\]

where

\[
F(z) = \frac{1 - \sqrt{1 - 4z^2}}{2z}.
\]

Theorem 3 implies that \( T_\sigma(z) \) is an algebraic function for any specified \( \sigma \), since \( F(z) \) is algebraic and \( u_\sigma(z) \) is a rational function.

3 Joint Structures

Given two RNA sequences \( R \) and \( S \) with \( n \) and \( m \) vertices, we index the vertices such that \( R_1 \) is the 5′ end of \( R \) and \( S_1 \) is the 3′ end of \( S \). We refer to the \( i \)th vertex in \( R \) by \( R_i \) and the subgraph induced by \( \{R_i, \ldots, R_j\} \) by \( R[i, j] \).

The intramolecular base pair can be represented by an arc, which is called an interior arc, with its two endpoints contained in either \( R \) or \( S \). Similarly, the extramolecular base pair can be represented by an arc, which is called an exterior arc, with one of its endpoints contained in \( R \) and the other in \( S \). A pre-structure, \( G(R, S, I) \), is a graph consisting of two secondary structures \( R \) and \( S \) with a set \( I \) of noncrossing exterior arcs.

The subgraph \( R[i, j] \) (\( S[i', j'] \)) is called secondary segment if there is no exterior arc \( R_k S_{k'} \) such that \( i \leq k \leq j \) (\( i' \leq k' \leq j' \)), see Fig. 2 (A). An interior arc \( R_i R_j \) is an \( R \)-ancestor of the exterior arc \( R_k S_{k'} \) if \( i < k < j \). Analogously, \( S_{i'} S_{j'} \) is an \( S \)-ancestor of \( R_k S_{k'} \) if \( i' < k' < j' \). We also refer to \( R_k S_{k'} \) as a descendant of \( R_i R_j \) and \( S_{i'} S_{j'} \) in this situation, see Fig. 2 (A). Furthermore, we call \( R_i R_j \) and \( S_{i'} S_{j'} \) dependent if they have a common descendant and
Fig. 2 (A): The joint structure \( J(R, S, I) \) with arc-length \( \geq 3 \), interior stack-length \( \geq 2 \), exterior stack-length \( \geq 3 \). Secondary segments (red): the subgraphs \( R[16, 21] \) and \( S[10, 15] \). Ancestors and descendants: for the exterior arc \( R_5S_5 \), we have the following sets of \( R \)-ancestors and \( S \)-ancestors of \( R_5S_5 \): \{ \( R_1R_{15}, R_3R_{14}, R_4R_6, R_4R_8 \) \} and \( \{ S_1S_{21}, S_2S_{20}, S_3S_9, S_4S_8 \} \). The exterior arc \( R_5S_5 \) is a common descendant of \( R_1R_{15} \) and \( S_1S_9 \), while \( R_{10}S_{17} \) is not. Subsumed arcs: \( R_1R_{15} \) subsumes \( S_3S_9 \) and \( S_1S_{21} \).

We next specify some notation

- an interior arc (or simply arc) of length \( \lambda \) is an arc \( R[i, j] \) (\( S[i', j'] \)) where \( j - i = \lambda \) (\( j' - i' = \lambda \)),
- an interior stack (or simply stack) of length \( \sigma \) is a maximal sequence of “parallel” interior arcs,
  \[
  (R_iR_j, R_{i+1}R_{j-1}, \ldots, R_{i+\sigma-1}R_{j-\sigma+1})
  \]
  or
  \[
  (S_iS_j, S_{i+1}S_{j-1}, \ldots, S_{i+\sigma-1}S_{j-\sigma+1})
  \]
- an exterior stack of length \( \tau \) is a maximal sequence of “parallel” exterior arcs,
  \[
  (R_iS_{i'}, R_{i+1}S_{i'+1}, \ldots, R_{i+\tau-1}S_{i'+\tau-1})
  \]
Let $J_{n,m,h}^{[\lambda,\tau]}$ denote the class of all joint structures with arc-length $\geq \lambda$, interior stack-length $\geq \sigma$, exterior stack-length $\geq \tau$. Similarly, we can define its counting sequence $J_{n,m,h}^{[\lambda,\tau]}(x,y,z)$ and generating function $J_{n,m,h}^{[\lambda,\tau]}(x,y,z)$. In case of $\lambda = 2$, we omit $\lambda$ in the notation. If there is no restriction on the interior and exterior stack-length, we also omit further indices. In the particular case $\sigma = \tau$, we just write $\sigma$ in the notation and omit $\tau$. In Fig. 2 (A), we give an example of joint structure with arc-length $\geq 3$, interior stack-length $\geq 2$ and exterior stack-length $\geq 3$.

We denote the subgraph of a joint structure $J(R,S,I)$ induced by a pair of subsequences $\{R_i, R_{i+1}, \ldots, R_j\}$ and $\{S_{i'}, S_{i'+1}, \ldots, S_{j'}\}$ the block $J_{i,j,i',j'}$. Given a joint structure $J(R,S,I)$, a tight structure of $J(R,S,I)$ is the minimal block $J_{i,j,i',j'}$ containing all the $R$-ancestors and $S$-ancestors of any exterior arc in $J_{i,j,i',j'}$ and all the descendants of any interior arc in $J_{i,j,i',j'}$. In the following, a tight structure is denoted by $J_{i,j,i',j'}^{T}$. In particular, we denote the joint structure $J(R,S,I)$ by $J_{i,j,i',j'}^{T}$ if $J(R,S,I)$ is a tight structure of itself. For any joint structure, there are only four types of tight structures $J_{i,j,i',j'}^{T}$, that is $\{\circ, \triangledown, \triangle, \Box\}$, denoted by $J_{i,j,i',j'}^{T[\circ,\triangledown,\triangle,\Box]}$, respectively. The four types of tight structures $J_{i,j,i',j'}^{T[\circ,\triangledown,\triangle,\Box]}$ are defined as follows:

- $\circ$: $\{R_iS_{i'}, S_{i'} = J_{i,j,i',j'}^{T}\}$ and $i = j'$, $i' = j$;
- $\triangledown$: $R_iR_j \in J_{i,j,i',j'}^{T}$ and $S_{i'}S_{j'} \notin J_{i,j,i',j'}^{T}$;
- $\triangle$: $S_{i'}S_{j'} \in J_{i,j,i',j'}^{T}$ and $R_iR_j \notin J_{i,j,i',j'}^{T}$;
- $\Box$: $\{R_iR_j, S_{i'}S_{j'}\} \in J_{i,j,i',j'}^{T}$.

The key function of tight structures is that they are the building blocks for the decomposition of joint structures.

**Proposition 2** [Huang et al., 2009] Let $J(R,S,I)$ be a joint structure. Then

1. any exterior arc $R_iS_{i'}$ in $J(R,S,I)$ is contained in a unique tight structure.
2. $J(R,S,I)$ decomposes into a unique collection of tight structures and maximal secondary segments.

### 4 Shapes

**Definition 2 (Shape)** A shape is a joint structure containing no secondary segments in which each interior stack and each exterior stack have length exactly one.

Let $\mathcal{S}$ denote the combinatorial class of shapes. Given a joint structure, we can obtain its shape by first removing all secondary segments and second collapsing any stacks into a single arc. That is, we have a map $\varphi: J \rightarrow \mathcal{S}$, see Fig. 3. Let $G(t_1, t_2, h)$ denote the number of shapes having $t_1$ arcs in the top sequence, $t_2$ arcs in the bottom and $h$ exterior arcs having the generating function

$$G(u,v,z) = \sum G(t_1, t_2, h) u^{t_1} v^{t_2} z^h.$$
We next introduce tight shapes, double tight shapes, interaction segments, closed shapes and right closed shapes:

- A tight shape is tight as a structure. Let $G^T$ denote the class of tight shapes by and $G^T(t_1, t_2, h)$ denote the number of tight shapes having $t_1$ arcs in the top sequence, $t_2$ arcs in the bottom and $h$ exterior arcs with generating function

$$G^T(u, v, z) = \sum G^T(t_1, t_2, h) u^{t_1} v^{t_2} z^h.$$ 

Any tight shape, comes as exactly one of the four types $\{\odot, \triangle, \Box\}$. The corresponding classes and generating functions are defined accordingly, $G^{\{\odot, \triangle, \Box\}}(x, y, z)$ respectively.

- A double tight shape is a shape whose leftmost and rightmost blocks are tight structures. Let $G^{DT}$ denote the class of double tight shapes by and $G^{DT}(t_1, t_2, h)$ denote the number of double tight shapes having $t_1$ arcs in the top sequence, $t_2$ arcs in the bottom and $h$ exterior arcs having the generating function

$$G^{DT}(u, v, z) = \sum G^{DT}(t_1, t_2, h) u^{t_1} v^{t_2} z^h.$$ 

- A closed shape is a tight shape of type $\{\triangle, \Box\}$. Let $G^C$ denote the class of closed shapes and $G^C(t_1, t_2, h)$ denote the number of closed shapes having $t_1$ arcs in the top sequence, $t_2$ arcs in the bottom and $h$ exterior arcs with generating function

$$G^C(u, v, z) = \sum G^C(t_1, t_2, h) u^{t_1} v^{t_2} z^h.$$ 

- A right closed shape is a shape whose rightmost block is a closed shape rather than an exterior arc. Let $G^{RC}$ denote the class of right close shapes and $G^{RC}(t_1, t_2, h)$ denote the number of right close shapes having $t_1$ arcs in the top sequence, $t_2$ arcs in the bottom and $h$ exterior arcs with generating function

$$G^{RC}(u, v, z) = \sum G^{RC}(t_1, t_2, h) u^{t_1} v^{t_2} z^h.$$
In a shape, an interaction segment is an empty structure or an tight structure of type $\circ$ (an exterior arc). We denote the class of interaction segment by $I$ and the associated generating function by $I(x, y, z)$. Obviously, $I(x, y, z) = 1 + z$.

**Theorem 4** The generating function $G(u, v, z)$ of shapes satisfies

$$A(u, v, z)G(u, v, z)^2 + B(u, v, z)G(u, v, z) + C(u, v, z) = 0, \quad (4)$$

where

$$A(u, v, z) = (u + v + uv)(z + 1),$$

$$B(u, v, z) = -((u + v + uv)(z + 2) + 1), \quad (5)$$

$$C(u, v, z) = (1 + u)(1 + v)(1 + z).$$

**Proof** Proposition 2 implies that any shape can be decomposed into a unique collection of tight shapes. Furthermore, each shape can be decomposed into a unique collection of close shapes and exterior arcs. We decompose a shape in four steps, see Fig. 4. We translate each decomposition step into the construction of combinatorial classes in the language of symbolic enumeration.

**Step (1):** we decompose a shape into a right closed shape and rightmost interaction segment. We generate $G = G^{RC} \times I + I$.

It follows from Proposition 4 that

$$G(x, y, z) = G^{RC}(x, y, z) \cdot I(x, y, z) + I(x, y, z). \quad (6)$$

**Step (2):** we decompose a right closed shape into the rightmost closed shape and the rest, deriving

$$G^{RC} = G \times G^C,$$

whence

$$G^{RC}(x, y, z) = G(x, y, z) \cdot G^C(x, y, z). \quad (7)$$

**Step (3):** we decompose a closed shape depending on its type. The decomposition operation in this step can be viewed as the "removal" of an interior arc. We derive

$$G^C = G^\triangledown + G^\triangle + G^{\square}$$

$$G^\triangledown = (Z, E, Z) + (Z, E, E) \times G^{DT}$$

$$G^\triangle = (E, Z, Z) + (Z, Z, E) \times G^{DT}$$

$$G^{\square} = (Z, Z, Z) + (Z, Z, E) \times G^{DT}.$$

and obtain the generating functions

$$G^C(x, y, z) = G^\triangledown(x, y, z) + G^\triangle(x, y, z) + G^{\square}(x, y, z)$$

$$G^\triangledown(x, y, z) = x z + x G^{DT}(x, y, z)$$

$$G^\triangle(x, y, z) = y z + y G^{DT}(x, y, z)$$

$$G^{\square}(x, y, z) = x y z + x y G^{DT}(x, y, z). \quad (8)$$
Fig. 4 The shape-grammar. The notations of structural components are shown in the panel below: A: interaction segment; B: arbitrary shape \( G(R, S, I) \); C: right close shape \( G_{RC}(R, S, I) \); D: double tight shape \( G_{DT}(R, S, I) \); E: close shape \( G_C(R, S, I) \); F: type □ tight shape \( G_{□}(R, S, I) \); G: type ▽ tight shape \( G_{▽}(R, S, I) \); H: type △ tight shape \( G_{△}(R, S, I) \); I: type ◦ tight shape \( G_{◦}(R, S, I) \);

**Step (4):** the class of double tight shapes arising from Step (3) can be obtained by excluding the class of interaction segment and the class of closed shapes from the class of shapes. Similarly, we have

\[
G_{DT} = G - I - G_C.
\]

The corresponding generating function accordingly satisfies

\[
G_{DT}(x, y, z) = G(x, y, z) - I(x, y, z) - G_C(x, y, z).
\]

We proceed by solving the set of equations (6)–(9), thereby deriving the functional equation eq. (5) for \( G(x, y, z) \) and the theorem follows.

5 The generating function

We proceed by generating joint structures from shapes via inflation. Let \( \bar{G}_{\sigma,\tau} \) denote the class of joint structures with arc-length \( \geq 2 \), interior stack-length \( \geq \sigma \), exterior stack-length \( \geq \tau \). Let \( J_{\sigma,\tau}(n, m, h) \) denote the number of joint
structures in $\mathcal{J}_{\sigma,\tau}$ having $n$ vertices in the top, $m$ vertices in the bottom and $h$ exterior arcs with its generating function

$$J_{\sigma,\tau}(x, y, z) = \sum J_{\sigma,\tau}(n, m, h) x^n y^m z^h.$$ 

**Theorem 5** For $\sigma \geq 1, \tau \geq 1$, we have

$$J_{\sigma,\tau}(x, y, z) = T_\sigma(x) T_\sigma(y) G(\eta(x), \eta(y), \eta_0),$$

(10)

where

$$\eta(w) = \frac{w^{2\sigma} T_\sigma(w)}{1 - w^2 - w^{2\sigma}(T_\sigma(w)^2 - 1)},$$

$$\eta_0 = \frac{(xyz)^2 T_\sigma(x) T_\sigma(y)}{1 - xyz - \tau(T_\sigma(x) T_\sigma(y) - 1)}.$$ 

**Proof** Let $\mathcal{G}(t_1, t_2, h)$ denote the class of shapes having $t_1$ interior arcs in the top, $t_2$ interior arcs in the bottom and $h$ exterior arcs. For any joint structure, we can obtain a unique shape in $\mathcal{G}$ as follows:

1. Remove all secondary segments.
2. Contract each interior stack into one interior arc and each exterior stack into one exterior arc.

Then we have the surjective map

$$\varphi: \mathcal{J}_{\sigma,\tau} \rightarrow \mathcal{G}.$$ 

Indeed, for any shape $\gamma$ in $\mathcal{G}$, we can construct joint structures with arc-length $\geq 2$, stack-length $\geq \sigma$, exterior stack-length $\geq \tau$. $\varphi: \mathcal{J}_{\sigma,\tau} \rightarrow \mathcal{G}$, induces the partition $\mathcal{J}_{\sigma,\tau} = \bigcup \varphi^{-1}(\gamma)$. Then we have

$$J_{\sigma,\tau}(x, y, z) = \sum_{\gamma \in \mathcal{G}} J_\gamma(x, y, z).$$

(11)

We proceed by computing the generating function $J_\gamma(x, y, z)$. We will construct $J_\gamma(x, y, z)$ via simpler combinatorial classes as building blocks considering $M_\sigma$ (stems), $K_\sigma$ (stacks), $N_\sigma$ (induced stacks), $R$ (interior arcs) and $T_\sigma$ (secondary segments). We inflate a shape $\gamma \in \mathcal{G}(t_1, t_2, h)$ to a joint structure in three steps.

**Step I:** we inflate any interior arc in $\gamma$ to a stack of size at least $\sigma$ and subsequently add additional stacks. The latter are called induced stacks and have to be separated by means of inserting secondary segments, see Fig.\(\text{\ref{fig:joint}}\). Note that during this first inflation step no secondary segments, other than those necessary for separating the nested stacks are inserted. We generate

- secondary segments $T_\sigma$ having stack-length $\geq \sigma$ with generating function $T_\sigma(z)$
- interior arcs $R$ with generating function $R(z) = z^2$
Fig. 5  Step I: a shape (left) is inflated to a joint structure with arc-length $\geq 2$ and interior stack-length $\geq 2$. First, each interior arc in the shape is inflated to a stack of size at least two (middle). Then the shape is inflated to a new joint structure with arc-length $\geq 2$ and interior stack-length $\geq 2$ (right) by adding one stack of size two. Note that there are three ways to insert the secondary segments to separate the induced stacks (red).

- stacks, i.e. pairs consisting of the minimal sequence of arcs $\mathcal{R}^2$ and an arbitrary extension consisting of arcs of arbitrary finite length
  \[
  \mathcal{K}_\sigma = \mathcal{R}^2 \times \text{SEQ} (\mathcal{R})
  \]
  having the generating function
  \[
  K_\sigma(z) = z^{2\sigma} \cdot \frac{1}{1 - z^2},
  \]

- induced stacks, i.e. stacks together with at least one secondary segment on either or both of its sides,
  \[
  \mathcal{N}_\sigma = \mathcal{K}_\sigma \times (T_\sigma^2 - 1),
  \]
  with generating function
  \[
  N_\sigma(z) = \frac{z^{2\sigma}}{1 - z^2} (T_\sigma(z)^2 - 1)
  \]

- stems, that is pairs consisting of stacks $\mathcal{K}_\sigma$ and an arbitrarily long sequence of induced stacks
  \[
  \mathcal{M}_\sigma = \mathcal{K}_\sigma \times \text{SEQ} (\mathcal{N}_\sigma),
  \]
with generating function
\[
M_\sigma(z) = \frac{K_\sigma(z)}{1 - N_\sigma(z)} = \frac{x^{2\sigma}}{1 - z} \frac{1}{1 - z} \frac{1}{1 - z} (T_\sigma(z)^2 - 1).
\]

Note that we inflate both: top as well as bottom sequences. The corresponding generating function is
\[
M_\sigma(x)^{t_1} M_\sigma(y)^{t_2}.
\]

**Step II:** we inflate any exterior arc in \( \gamma \) to an exterior stack of size at least \( \tau \) and subsequently add additional exterior stacks. The latter are called induced exterior stacks and have to be separated by means of inserting secondary segments, see Fig. 6. Note that during this exterior-arc inflation step no secondary segments, other than those necessary for separating the stacks are inserted. We generate
- exterior arc \( R_0 \) with generating function
  \[
  R_0 = xyz.
  \]
- exterior stacks, i.e. pairs consisting of the minimal sequence of exterior arcs \( R_0^t \) and an arbitrary extension consisting of exterior arcs of arbitrary finite length
  \[
  K_\tau = R_0^t \times \text{SEQ}(R_0)
  \]
having the generating function
\[ K'_\tau = (xyz)^r - \frac{1}{1 - xyz}. \]

– induced exterior stacks, i.e. stacks together with at least one secondary segment on either or both its sides,
\[ N'_\tau = K'_\tau \times (J^2_\sigma - 1), \]

with generating function
\[ N'_\tau = \frac{(xyz)^r}{1 - xyz} \left( T_\sigma(x)T_\sigma(y) - 1 \right). \]

– exterior stems, that is pairs consisting of exterior stacks \( K'_\tau \) and an arbitrarily long sequence of induced exterior stacks
\[ M'_\tau = K'_\tau \times \text{SEQ}(N'_\tau), \]

with generating function
\[ M'_\tau = K'_\tau - \frac{(xyz)^r}{1 - xyz} \left( T_\sigma(x)T_\sigma(y) - 1 \right). \]

We inflate all the exterior arcs and the corresponding generating function is
\[ (M'_\tau)^h. \]

**Step III:** here we insert additional secondary segments at the remaining \((2t_1 + h + 1)\) positions in the top and the \((2t_2 + h + 1)\) positions in the bottom, see Fig. 7. Formally, the third inflation is expressed via the combinatorial class
\[ (J^2_\sigma)^{2t_1 + h + 1} (J^2_\sigma)^{2t_2 + h + 1}, \]

where the corresponding generating function is
\[ T_\sigma(x)^{2t_1 + h + 1} T_\sigma(y)^{2t_2 + h + 1}. \]

Combining Step I, Step II and Step III we arrive at
\[ M_\sigma(x)^{t_1} M_\sigma(y)^{t_2} (M'_\tau)^h J_\sigma(x)^{2t_1 + h + 1} J_\sigma(y)^{2t_2 + h + 1}, \]
and accordingly
\[ M_\sigma(x)^{t_1} M_\sigma(y)^{t_2} (M'_\tau)^h T_\sigma(x)^{2t_1 + h + 1} T_\sigma(y)^{2t_2 + h + 1} \]
\[ = T_\sigma(x)T_\sigma(y)(T_\sigma(x)^2)M_\sigma(x))^{t_1}(T_\sigma(y)^2)M_\sigma(y))^{t_2}(M'_\tau T_\sigma(x) T_\sigma(y))^h. \]

Therefore,
\[ J_\gamma(x, y, z) = T_\sigma(x)T_\sigma(y)(T_\sigma(x)^2)M_\sigma(x))^{t_1}(T_\sigma(y)^2)M_\sigma(y))^{t_2}(M'_\tau T_\sigma(x) T_\sigma(y))^h. \]
Fig. 7 Step III: a joint structure (left) obtained in (1) in Fig. 6 is inflated to a new joint structure in $J_2$ (right) by adding secondary segments (red).

Since for any $\gamma, \gamma_1 \in G(t_1, t_2, h)$ we have $J_\gamma(x, y, z) = J_{\gamma_1}(x, y, z)$, we derive

$$J_{\sigma, \tau}(x, y, z) = \sum_{\gamma \in G} J_\gamma(x, y, z) = \sum_{\gamma \in G(t_1, t_2, h)} G(t_1, t_2, h) J_\gamma(x, y, z).$$

Set

$$\eta(w) = \frac{w^{2\sigma} T_\sigma(w)^2}{1 - w^2 - w^{2\sigma}(T_\sigma(w)^2 - 1)}$$

$$\eta_0 = \frac{(xyz)^2 T_\sigma(x) T_\sigma(y)}{1 - xyz - (xyz)^2(T_\sigma(x) T_\sigma(y) - 1)}.$$

According to the generating function

$$G(u, v, z) = \sum G(t_1, t_2, h) u^{t_1} v^{t_2} z^h,$$

we have

$$J_{\sigma, \tau}(x, y, z) = T_\sigma(x) T_\sigma(y) G(\eta(x), \eta(y), \eta_0)$$

and the theorem follows.

6 Asymptotic analysis

6.1 The supercritical paradigm

Suppose $U(z) = G(z, z, z)$. We view $U(z)$ as a generating function, $U(z) = \sum U(l) z^l$, where $U(l)$ denotes the number of shapes having $l$ arcs. It follows from Theorem 4 that $U(z)$ satisfies

$$(z^2 + 2z) U(z)^2 - (z^2 + 3z + 1) U(z) + (1 + z)^2 = 0.$$

Solving this functional equation, we derive

$$U(z) = \frac{1 + 3z + z^2 - \sqrt{1 - 2z - 9z^2 - 10z^3 - 3z^4}}{2z(z + 2)}. \quad (12)$$

It is straightforward to verify that the dominant singularity $\rho$ of $U(z)$ is the minimal and positive real solution of $1 - 2z - 9z^2 - 10z^3 - 3z^4 = 0$ and
Fig. 8 Universality of the square root. We display the dominant singularity of the generating function $U(z)$ of shapes (here at $\rho \approx 0.22144$). All singularities arising from composition of the “outer” function $U(z)$ governed by the supercritical paradigm produce this type of singularity, leading to the subexponential factor $n^{-\frac{3}{2}}$.

$\rho \approx 0.22144$, see Fig. 8.

For our computations the following instance of the supercritical paradigm (Flajolet, 2007) is of central importance: we are given a $D$-finite function, $f(z)$ and an algebraic function $g(u)$ satisfying $g(0) = 0$. Furthermore we suppose that $f(g(u))$ has a unique real valued dominant singularity $\gamma$ and $g$ is regular in a disc with radius slightly larger than $\gamma$. Then the supercritical paradigm stipulates that the subexponential factors of $f(g(u))$ at $u = 0$, given that $g(u)$ satisfies certain conditions, coincide with those of $f(z)$.

**Lemma 1** Let $\vartheta(z)$ be an algebraic, analytic function for $|z| < r$ such that $\vartheta(0) = 0$. In addition suppose $\gamma$ is the unique dominant singularity of $U(\vartheta(z))$ and minimum positive real solution of $\vartheta(z) = \rho$, $|z| < r$, $\vartheta'(z) \neq 0$. Then $U(\vartheta(z))$ has a singular expansion and

$$[z^n]U(\vartheta(z)) \sim c n^{-\frac{3}{2}} (\gamma^{-1})^n,$$

(13)

where $c$ is some constant.

Proof Since $\vartheta(z)$ is an algebraic function such that $\vartheta(0) = 0$ and $U(z)$ is algebraic whence is $D$-finite, we can conclude that the composition $U(\vartheta(z))$ is $D$-finite. In particular $U(\vartheta(z))$ has a singular expansion.

Next, we calculate the singular expansion of the composite function $U(\vartheta(z))$. 
In view of \[ z^n f(z) = \gamma^n z^n f(\gamma z) \] it suffices to analyze the function \( U(\vartheta(\gamma z)) \) and to subsequently rescale in order to obtain the correct exponential factor. For this purpose we set

\[
\tilde{\vartheta}(z) = \vartheta(\gamma z),
\]

where \( \vartheta(z) \) is analytic in \(|z| \leq r\). Consequently \( \tilde{\vartheta}(z) \) is analytic in \(|z| < \tilde{r}\), for some \( 1 < \tilde{r} \). The singular expansion of \( U(z) \), for \( z \to \rho \), is given by

\[
U(z) = u_0 + u_1 (\rho - z)^{\frac{1}{2}} (1 + o(1)).
\]

By construction \( U(\vartheta(\gamma z)) = U(\tilde{\vartheta}(z)) \), \( U(\tilde{\vartheta}(z)) \) has the unique dominant singularity at 1. We have the Taylor expansion of \( \tilde{\vartheta}(z) \) at \( z = 1 \)

\[
\rho - \vartheta(z) = \sum_{n \geq 1} \vartheta_n (1 - z)^n = \tilde{\vartheta}_1 (1 - z)(1 + o(1)).
\]

As for the singular expansion of \( U(\tilde{\vartheta}(z)) \), substituting eq. (14) into the singular expansion of \( U(z) \), for \( z \to 1 \),

\[
U(\tilde{\vartheta}(z)) = u_0 + u_1 \tilde{\vartheta}_1^{\frac{1}{2}} (1 - z)^{\frac{1}{2}} (1 + o(1)),
\]

where \( \tilde{\vartheta}_1 = \tilde{\vartheta}(z)|_{z=1} = \gamma \vartheta(z)|_{z=\gamma} \neq 0 \). By Theorem \( \text{I} \) and Theorem \( \text{II} \) we arrive at

\[
[z^n] U(\tilde{\vartheta}(z)) \sim c n^{-\frac{1}{2}} \quad \text{for some constant } c.
\]

Finally, we use the scaling property of Taylor expansions in order to derive

\[
[z^n] U(\vartheta(z)) = (\gamma^{-1})^n [z^n] U(\tilde{\vartheta}(z))
\]

and the proof is complete.

We remark that Lemma \( \text{I} \) allows under certain conditions to obtain the asymptotics of the coefficients of supercritical compositions of the “outer” function \( U(z) \) and “inner” function \( \vartheta(z) \). The scenario considered here is tailored for asymptotic expressions of \( J_\sigma(s) \).

6.2 Asymptotics of \( J_\sigma(s) \)

In this section we shall assume \( \sigma = \tau \). Let \( J_\sigma(s) \) denote the number of joint structures of total \( s \) vertices having arc-length \( \geq 2 \), stack-length \( \geq \sigma \) and exterior stack-length \( \geq \sigma \) with its generating function

\[
J_\sigma(z) = \sum J_\sigma(s) z^s.
\]

By definition, we have

\[
J_\sigma(z) = J_{\sigma,\sigma}(z, z, 1).
\]
Theorem 6 For $\sigma \geq 1$, we have
\[ J_\sigma(z) = T_\sigma(z)^2 U(\zeta(z)), \]
where
\[ \zeta(z) = \frac{z^{2\sigma} T_\sigma(z)^2}{1 - z^2 - z^{2\sigma}(T_\sigma(z)^2 - 1)}. \]

Furthermore, for $1 \leq \sigma \leq 9$, $J_\sigma(s)$ satisfies
\[ J_\sigma(s) \sim c_\sigma s^{-\frac{2}{3}} (\gamma_\sigma^{-1})^s, \quad \text{for some } c_\sigma, \]
where $\gamma_\sigma$ is the minimal, positive real solution of the equation $\zeta(z) = \rho$, see Table 7. In particular, $c_1 \approx 1.6527921$ and $c_2 \approx 4.3011932$.

Proof By Theorem 5 and the definition, we have
\[ J_\sigma(z) = J_{\sigma,\sigma}(z, z, 1) \]
\[ = T_\sigma(z)^2 G(\zeta(z), \zeta(z), \zeta(z)) \]
\[ = T_\sigma(z)^2 U(\zeta(z)), \]
where
\[ \zeta(z) = \frac{z^{2\sigma} T_\sigma(z)^2}{1 - z^2 - z^{2\sigma}(T_\sigma(z)^2 - 1)}. \]

Since $T_\sigma(z)$ is algebraic, we can conclude that $\zeta(z)$ is algebraic from the closure property of algebraic functions, whence $U(\zeta(z))$ and $J_\sigma(z)$ are $D$-finite. Pringsheim’s Theorem [Titchmarsh, 1939] guarantees that $J_\sigma(z)$ has a dominant real positive singularity $\gamma_\sigma$. We verify that for $1 \leq \sigma \leq 9$, the minimal, positive real solution of the equation $\zeta(z) = \rho$ is strictly smaller than the singularity of $\zeta(z)$, which is actually the singularity of $T_\sigma(z)$. Hence $\gamma_\sigma$ is the unique, minimal, positive real solution of the equation $\zeta(z) = \rho$ and it is straightforward to check that $\zeta'(z)|_{z=\gamma_\sigma} \neq 0$. Therefore the composite function $U(\zeta(z))$ is governed by the supercritical paradigm of Lemma 1. Furthermore $T_\sigma(z)$ is analytic at $\gamma_\sigma$, whence the subexponential factors of $T_\sigma(z)^2 U(\zeta(z))$ coincide with those of the function $U(z)$. Consequently,
\[ J_\sigma(s) \sim c_\sigma s^{-\frac{2}{3}} (\gamma_\sigma^{-1})^s, \quad \text{for some } c_\sigma. \]

The values of $\gamma_\sigma^{-1}$ are listed in Table 1. It remains to calculate the constant coefficient in the asymptotic formula. Setting the singular expansion of $U(z)$ around $\rho$ and the Taylor expansions of $\zeta(z)$ and $T_\sigma(z)^2$ around $\gamma_\sigma$,
\[ U(z) = u_0 + u_1(\rho - z)^{\frac{1}{2}} + O((\rho - z)), \]
\[ \zeta(z) - \rho = g_1(z - \gamma_\sigma) + O((z - \gamma_\sigma)^2), \]
\[ T_\sigma(z)^2 = t_0 + t_1(\gamma_\sigma - z) + O((\gamma_\sigma - z)^2). \]

We proceed by substituting these expansions into $T_\sigma(z)^2 U(\zeta(z))$
\[ J_\sigma(z) = t_0u_0 + t_0u_1g_1^{\frac{1}{2}}(\gamma_\sigma - z)^{\frac{1}{2}} + O(\gamma_\sigma - z). \]
Table 1 Exponential growth rates $\gamma^{-1}_\sigma$ for joint structures with arc-length $\geq 2$, having both stack-length and exterior stack-length $\geq \sigma$.

| $\sigma$ | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $\gamma^{-1}_\sigma$ | 3.48766 | 2.24338 | 1.86724 | 1.67974 | 1.56544 | 1.48763 | 1.43083 | 1.38731 | 1.35276 |

Using Theorem 1 and Theorem 2, we have

$$J_\sigma(s) \sim \frac{\ell_0 u_1(g_1 \gamma \sigma)^{1/2}}{I(-\frac{1}{2})} s^{-\frac{1}{2}} (\gamma \sigma)^{-s}$$

Setting $c_\sigma = \frac{\ell_0 u_1(g_1 \gamma \sigma)^{1/2}}{I(-\frac{1}{2})}$, we compute $c_1 \approx 1.6527921$ and $c_2 \approx 4.3011932$, completing the proof of Theorem 6.

We next observe that eq. (15) allows us to derive a functional equation for $J_\sigma(z)$, which in turn gives a recurrence of $J_\sigma(s)$.

**Corollary 1** For $\sigma \geq 1$, the generating function $J_\sigma(z)$ satisfies the functional equation

$$A(z)J_\sigma(z)^2 + B(z)J_\sigma(z) + C(z) = 0,$$

where

$$A(z) = z^{2\sigma} (2 - 2z + 2z^{2\sigma} - z^{2\sigma}T_\sigma(z)^2),$$

$$B(z) = -\left(1 - 2z^2 + z^4 + (2 + T_\sigma(z)^2)z^{2\sigma} - (2 + T_\sigma(z)^2)z^{4\sigma}\right),$$

$$C(z) = (1 - z^2 + z^{2\sigma})^2.$$ (19)

Furthermore, the number $J_\sigma(s)$ of joint structures with total $s$ vertices satisfies the following recurrence:

$$J_\sigma(s) = c(s) + \sum_{i=1}^{s} b(i) J_\sigma(s - i) + \sum_{i=1}^{s} \sum_{j=0}^{s-i} a(i) J_\sigma(j) J_\sigma(s - i - j),$$

where $a(s)$, $b(s)$ and $c(s)$ are the coefficients of $z^s$ of $A(z)$, $B(z)$ and $C(z)$, respectively.

In Table 2 we list the numbers of joint structures $J_1(s)$ for $s = 1, \ldots, 12$.

*Proof* Substituting $z = \frac{z^{2\sigma} T_\sigma(z)^2}{1 - z^2 - z^{2\sigma} + (2 + T_\sigma(z)^2)z^{2\sigma} - (2 + T_\sigma(z)^2)z^{4\sigma}}$ into eq. (15) and using eq. (16), we obtain eq. (18). Note that $a(0) = 0$ and $b(0) = -1$. Calculating the coefficients of $z^s$ of eq. (18), the recurrence follows immediately.
Table 2 The numbers of joint structures $J_1(s)$ and $J_2(s)$ over a total number of $s = 1, \ldots, 12$ nucleotides.

| $s$ | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|-----|----|----|----|----|----|----|----|----|----|----|----|----|
| $J_1(s)$ | 2  | 4  | 10 | 26 | 70 | 194 | 550 | 1590 | 4674 | 13940 | 42106 | 128610 |
| $J_2(s)$ | 2  | 3  | 4  | 6  | 12 | 26 | 54 | 105 | 200 | 389 | 780 | 1589 |

Fig. 9 Exact enumeration versus asymptotic formula. We plot the number of joint structures with arc-length $\geq 2$ and stack-length $\geq 1$, $(J_1(s))$ versus its asymptotic formula $c_1 s^{-\frac{3}{2}} \cdot 3.48766^s$ (left) and $J_2(s)$ versus $c_2 s^{-\frac{3}{2}} \cdot 2.24338^s$ (right). For representational purposes we separate the curves via setting the respective constants $c_1 = 10^7$.

Having derived exact enumeration as well as the asymptotic expressions for the numbers of joint structures having arc-length $\geq 2$, stack-length as well as exterior stack-length $\geq \sigma$. In Fig. 9 we display that our results work well already for relatively small sequence length. In Fig. 9 we contrast the exact values $J_1(s)$, i.e., the number of joint structures with arc-length $\geq 2$, stack-length $\geq 1$ and exterior stack-length $\geq 1$ and $J_2(s)$, i.e., the number of joint structures with arc-length $\geq 2$, stack-length $\geq 2$ and exterior stack-length $\geq 2$, with the asymptotic formulas given in Theorem 6:

$$J_1(s) \sim c_1 s^{-\frac{3}{2}} \cdot 3.48766^s \quad \text{and} \quad J_2(s) \sim c_2 s^{-\frac{3}{2}} \cdot 2.24338^s.$$

7 Discussion

The discovery of more and more instances of regulatory actions among RNA molecules make evident that RNA-RNA interaction is a problem of central importance. While we can MFE fold joint structures [Huang et al. 2009], our
theoretical understanding of RNA-RNA interaction structures is not very developed. The latter would provide insights in the analysis of folding algorithms in general as well as the derivation of approximation algorithms.

To set the context of our results, let us begin by asking the following question: why a combinatorial theory of interaction structures? After all it should be straightforward to derive their generating function from the (eleven) recursion relations of the original rip-grammar (implied by Proposition 2) presented in Huang et al. (2009). While this statement is in principle correct, such a derivation would fill multiple pages and not be suitable for deriving any asymptotic formulas or allow to deal with specific stack-length conditions. To say the least: the mere extraction of coefficients would present a nontrivial task.

Our theory of joint structures is centered around the concept of shapes. In fact, the key to all results is the simple shape-grammar of Theorem 4. The basic idea here is that the collapsing of stems preserves vital information of the interaction structure. Given a shape a joint structure can be obtained via inflation, see Theorem 5. While there exists a notion of shapes for RNA secondary structures (Giegerich and Meyer, 2002) their combinatorics is not shape-based. Everything is organized around recurrences, which oftentimes hides deeper structural insight and connections. As a result symbolic enumeration has not been employed in order to derive their generating functions. In contrast, RNA pseudoknot structures (Reidys et al., 2010) represent a shape-based structure class (here further complication enters the picture as the generating function of their shapes can only be computed via the reflection principle). The theory of joint structures presented here resembles features of the theory of modular diagrams and is in particular shape-based. However, the shapes of joint structures are governed by simple algebraic generating functions and satisfy a simple recurrence.

We next intend to study the generating function of canonical joint structures having minimum arc-length four. This derivation requires a more detailed look at shapes of joint structures since additional variables has to be introduced. The purpose of these variables is to allow to distinguish specific inflation scenarios.

8 Further results

In this section we generalize our results to joint structures with arc-length $\geq \lambda$, interior stack-length $\geq \sigma$, and exterior stack-length $\geq \tau$. Let $T_{\sigma}^{[\lambda]}$ denote the combinatorial class of $\sigma$-canonical secondary structures having arc-length $\geq \lambda$ and $T_{\sigma}^{[\lambda]}(n)$ denote the number of all $\sigma$-canonical secondary structures with $n$ vertices having arc-length $\geq \lambda$ and

$$T_{\sigma}^{[\lambda]}(z) = \sum T_{\sigma}^{[\lambda]}(n) z^n$$

be the generating function.
**Theorem 7** Let $\sigma \in \mathbb{N}$, $z$ be an indeterminant and let

\[ u_\sigma(z) = \frac{(z^2)^{\sigma-1}}{z^{2\sigma} - z^2 + 1}, \]

\[ v_\lambda(z) = 1 - z + u_\sigma(z) \sum_{h=2}^{\lambda} z^h, \]

then, $T_\sigma^{[\lambda]}(z)$, the generating function of $\sigma$-canonical structures with minimum arc-length $\lambda$ is given by

\[ T_\sigma^{[\lambda]}(z) = \frac{1}{v_\lambda(z)} F \left( \left( \frac{\sqrt{u_\sigma(z)} z}{v_\lambda(z)} \right)^2 \right), \]

where

\[ F(z) = \frac{1 - \sqrt{1 - 4z}}{2z}. \]

Similarly, Theorem 7 implies that $T_\sigma^{[\lambda]}(z)$ is an algebraic function for any specified $\lambda$ and $\sigma$, since $F(z)$ is algebraic and $v_\lambda(z), u_\sigma(z)$ are both rational functions.

We are now in position to establish a generalization of Theorem 5 that allows us to compute the generating function $J_\sigma^{[\lambda]}(x, y, z)$ when $\lambda \leq \tau + 1$.

Let $\mathcal{J}_\sigma^{[\lambda]}(x, y, z)$ denote the class of joint structures with arc-length $\geq \lambda$, interior stack-length $\geq \sigma$, and exterior stack-length $\geq \tau$. Let $J_\sigma^{[\lambda]}(n, m, h)$ denote the number of joint structures in $\mathcal{J}_\sigma^{[\lambda]}$ having $n$ vertices in the top, $m$ vertices in the bottom, $h$ exterior arcs with its generating function

\[ J_\sigma^{[\lambda]}(x, y, z) = \sum J_\sigma^{[\lambda]}(n, m, h) x^n y^m z^h. \]

**Theorem 8** For $\sigma \geq 1$, $\tau \geq 1$, $\lambda \leq \tau + 1$, we have

\[ J_\sigma^{[\lambda]}(x, y, z) = T_\sigma^{[\lambda]}(x) T_\sigma^{[\lambda]}(y) G(\eta(x), \eta(y), \eta_0), \quad (20) \]

where

\[ \eta(w) = \frac{w^{2\sigma} T_\sigma^{[\lambda]}(w)^2}{1 - w^2 - w^{2\sigma} (T_\sigma^{[\lambda]}(w)^2 - 1)} \]

\[ \eta_0 = \frac{(xyz)^\tau T_\sigma^{[\lambda]}(x) T_\sigma^{[\lambda]}(y)}{1 - xyz - (xyz)^\tau (T_\sigma^{[\lambda]}(x) T_\sigma^{[\lambda]}(y) - 1)}. \]
Proof Using the notation and approach of Theorem 5, we arrive at

\[ \mathcal{K}_\sigma = \mathcal{R}^\sigma \times \text{SEQ}(\mathcal{R}) \]
\[ \mathcal{N}_\sigma = \mathcal{K}_\sigma \times \left( (\mathcal{T}_{\sigma})^2 - 1 \right) \]
\[ \mathcal{M}_\sigma = \mathcal{K}_\sigma \times \text{SEQ}(\mathcal{N}_\sigma) \]
\[ \mathcal{K}'_\tau = \mathcal{R}_0^\tau \times \text{SEQ}(\mathcal{R}_0) \]
\[ \mathcal{N}'_\tau = \mathcal{K}'_\tau \times \left( (\mathcal{T}_{\tau})^2 - 1 \right) \]
\[ \mathcal{M}'_\tau = \mathcal{K}'_\tau \times \text{SEQ}(\mathcal{N}'_\tau) \]
\[ \mathcal{J}_{\sigma,\tau}[\lambda] = \mathcal{M}_\sigma(x) \mathcal{T}_{\sigma}(x) \mathcal{G}(\eta(x), \eta(y), \eta_0), \]

where

\[ \eta(w) = \frac{w^{2\sigma} \mathcal{T}_{\sigma}(w)^2}{1 - w^2 - w^{2\sigma} (\mathcal{T}_{\sigma}(w)^2 - 1)} \]
\[ \eta_0 = \frac{(xyz)^2 \mathcal{T}_{\sigma}(x) \mathcal{T}_{\tau}(y)}{1 - xyz - (xyz)^2 (\mathcal{T}_{\sigma}(x) \mathcal{T}_{\tau}(y) - 1)}. \]

The only difference is that \( \mathcal{T}_{\sigma}[\lambda] \) replaces \( \mathcal{T}_{\sigma} \) to make the structure with arc-length \( \geq \lambda \). The key point here is that the restriction \( \lambda \leq \tau + 1 \) guarantees that any 2-arc in \( \gamma \) has after inflation a minimum arc-length of \( \tau + 1 \geq \lambda \).

Therefore, the generating function of class \( \mathcal{J}_{\sigma,\tau}[\lambda] \) satisfies

\[ \mathcal{J}_{\sigma,\tau}[\lambda](x, y, z) = \mathcal{T}_{\sigma}(x) \mathcal{T}_{\tau}(y) \mathcal{G}(\eta(x), \eta(y), \eta_0), \]

where

\[ \eta(w) = \frac{w^{2\sigma} \mathcal{T}_{\sigma}(w)^2}{1 - w^2 - w^{2\sigma} (\mathcal{T}_{\sigma}(w)^2 - 1)} \]
\[ \eta_0 = \frac{(xyz)^2 \mathcal{T}_{\sigma}(x) \mathcal{T}_{\tau}(y)}{1 - xyz - (xyz)^2 (\mathcal{T}_{\sigma}(x) \mathcal{T}_{\tau}(y) - 1)}. \]

We remark that Theorem 8 immediately implies Theorem 5.

Analogously, we have

**Theorem 9** For \( \lambda \leq \sigma + 1 \), we have

\[ \mathcal{J}_{\sigma}[\lambda](z) = \mathcal{T}_{\sigma}(z)^2 \mathcal{U}(\zeta(z)), \]

where

\[ \zeta(z) = \frac{z^{2\sigma} \mathcal{T}_{\sigma}(z)^2}{1 - z^2 - z^{2\sigma} (\mathcal{T}_{\sigma}(z)^2 - 1)}. \]

Furthermore, for \( 1 \leq \sigma \leq 9 \) and \( 1 \leq \lambda \leq 5 \), \( \mathcal{J}_{\sigma}[\lambda](s) \) satisfies

\[ \mathcal{J}_{\sigma}[\lambda](s) \sim c_{\sigma}[\lambda] s^{-\frac{1}{\gamma_{\sigma}[\lambda]}} \left( \frac{1}{\gamma_{\sigma}[\lambda]} \right)^s, \]

for some \( c_{\sigma}[\lambda] \),

where \( \gamma_{\sigma}[\lambda] \) is the minimal, positive real solution of the equation \( \zeta(z) = \rho \), see Table 3. In particular, \( c_1[2] \approx 1.6527921 \), \( c_2[2] \approx 4.3011932 \), and \( c_2[3] \approx 3.8671841 \).

**Acknowledgements** We would like to thank F.W.D. Huang, J. Qin, and R.R. Wang for discussions.
Table 3  Exponential growth rates \((\gamma_{\lambda})^{-1}\) for joint structures with arc-length \(\geq \lambda\), having both stack-length and exterior stack-length \(\geq \sigma\).

| \(\sigma\) | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| \(\lambda = 1\) | 3.77438 | 2.30663 | 1.89559 | 1.69615 | 1.57629 | 1.49541 | 1.43671 | 1.39194 | 1.35651 |
| \(\lambda = 2\) | 3.48766 | 2.24338 | 1.86724 | 1.67974 | 1.56544 | 1.48763 | 1.40383 | 1.38731 | 1.35276 |
| \(\lambda = 3\) | 0.00000 | 2.21090 | 1.84998 | 1.66876 | 1.55773 | 1.48187 | 1.42633 | 1.38368 | 1.34976 |
| \(\lambda = 4\) | 0.00000 | 0.00000 | 1.83971 | 1.66155 | 1.55233 | 1.47764 | 1.42291 | 1.38085 | 1.34737 |
| \(\lambda = 5\) | 0.00000 | 0.00000 | 0.00000 | 1.65691 | 1.54861 | 1.47459 | 1.42036 | 1.37867 | 1.34549 |

References

Alkan, C., Karakoc, E., Nadeau, J., Sahinalp, S. and Zhang, K. (2006) RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, 13, 267–282.

Bachellerie, J., Cavaillé, J. and Hüttenhofer, A. (2002) The expanding snoRNA world. *Biochimie*, 84, 775–790.

Banerjee, D. and Slack, F. (2002) Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays*, 24, 119–129.

Benne, R. (1992) RNA editing in trypanosomes. the use of guide RNAs. *Mol. Biol. Rep.*, 16, 217–227.

Chitsaz, H., Salari, R., Sahinalp, S. and Backofen, R. (2009) A partition function algorithm for interacting nucleic acid strands. In press.

Flajolet, P. and Sedgewick, R. (2007) Analytic combinatorics, Cambridge University Press.

Geissmann, T. and Touati, D. (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J.*, 23, 396–405.

Giegerich, R. and Meyer, C. (2002) *Lecture Notes In Computer Science*, volume 2422, chapter Algebraic Dynamic Programming, pp. 349–364. Springer-Verlag.

Huang, F., Qin, J., Stadler, P. and Reidys, C. (2009) Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*, 25, 2646–2654.

Huang, F., Qin, J., Stadler, P. and Reidys, C. (2010) Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics*, 26, 175–181.

Kugel, J. and Goodrich, J. (2007) An RNA transcriptional regulator templates its own regulatory RNA. *Nat. Struct. Mol. Biol.*, 3, 89–90.

McManus, M. and Sharp, P. (2002) Gene silencing in mammals by small interfering RNAs. *Nature Reviews*, 3, 737–747.

Narberhaus, F. and Vogel, J. (2007) Sensory and regulatory RNAs in prokaryotes: A new german research focus. *RNA Biol.*, 4, 160–164.

Pervouchine, D. (2004) IRIS: Intermolecular RNA interaction search. *Proc. Genome Informatics*, 15, 92–101.
Reidys, C., Wang, R. and Zhao, A. (2010) Modular, \( k \)-noncrossing diagrams. *Electron. J. Combin.*  
Rivas, E. and Eddy, S. (1999) A dynamic programming algorithms for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.  
Titchmarsh, E.C. (1939) The theory of functions, Oxford University Press, Oxford UK.  
Udekwu, K., Darfeuille, F., Vogel, J., Reimegård, J., Holmqvist, E. and Wagner, E. (2005) Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev.*, **19**, 2355–2366.