

Combinatorics of locally optimal RNA secondary structures

Éric Fusy¹

Peter Clote^{1,2,3}

¹ Laboratoire d'Informatique (LIX), Ecole Polytechnique, 91128 Palaiseau, France.

² Laboratoire de Recherches en Informatique (LRI), Université Paris-Sud XI, bat. 490, 91405 Orsay, France

³ Department of Biology, Boston College, Chestnut Hill, MA 02467, USA.
fusy@lix.polytechnique.fr, clote@bc.edu

Abstract

It is a classical result of Stein and Waterman that the asymptotic number of RNA secondary structures is $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$. To provide a better understanding of the kinetics of RNA secondary structure formation, we are interested in determining the asymptotic number of secondary structures that are *locally optimal*, with respect to a particular energy model. In the Nussinov energy model, where each base pair contributes -1 towards the energy of the structure, locally optimal structures are exactly the *saturated* structures, for which we have previously shown that asymptotically, there are $1.07427 \cdot n^{-3/2} \cdot 2.35467^n$ many saturated structures for a sequence of length n . In this paper, we consider the *base stacking energy model*, a mild variant of the Nussinov model, where each stacked base pair contributes -1 toward the energy of the structure. Locally optimal structures with respect to the base stacking energy model are exactly those secondary structures, whose stems cannot be extended. Such structures were first considered by Evers and Giegerich, who described a dynamic programming algorithm to enumerate all locally optimal structures. In this paper, we apply methods from enumerative combinatorics to compute the asymptotic number of such structures.

1 Introduction

Historically, the development of combinatorics for RNA secondary structures [12, 15] has been intimately related to the development of *algorithms* for RNA minimum free energy (MFE) secondary structure [18, 17, 8]. In particular, *counting* the number of secondary structures for sequence of length n is essentially equivalent to computing the Boltzmann partition function, defined by $Z = \sum_S \exp(-E(S)/RT)$, where the sum

is taken over all secondary structures S , the energy of S is denoted by $E(S)$, R is the universal gas constant, and T absolute temperature.¹

In [12], Stein and Waterman proved that the asymptotic number of secondary structures is $1.104366 \cdot n^{-3/2} \cdot 2.618034^n$. Since that time, a number of additional results on the combinatorics of RNA structures have been obtained. In [9], Hofacker et al. derived a number of asymptotic results on the number of structures, expected number of base pairs, etc. for RNA secondary structures. Observing a correspondence with involutions, Haslinger and Stadler

¹If the energy $E(S) = 0$ or if the temperature $T = +\infty$, then the partition function is exactly equal to the number of secondary structures.

[7] provided an upper bound on the number of *bi-secondary* structures, i.e. structures having non-nested pseudoknots that can be displayed on two pages. In [11] Rodland studied the asymptotic number of a number of classes of pseudoknotted structures, while Vernizzi et al. [13] provided recurrence relations for the number of pseudoknotted structures having topological genus g .

In [1], Clote computed the asymptotic number $1.07427 \cdot n^{-3/2} \cdot 2.35467^n$ of *saturated* structures, defined by Zuker [16] as those for which no base pair can be added without violating the definition of secondary structure. In [2], Clote et al. provided another proof for the asymptotic number of saturated structures, which additionally yielded the asymptotic expected number of base pairs $0.337361 \cdot n$ for saturated structures. An overview of methods for RNA enumerative combinatorics is given in Lorenz et al. [10], where additionally it is shown that the asymptotic number of *shapes* of secondary structures for a length n sequence is $2.44251 \cdot n^{-3/2} \cdot 1.32218^n$.² In [9] Hofacker et al. showed that the asymptotic number of *canonical* secondary structures (those having no isolated base pair) is $2.1614 \cdot n^{-3/2} \cdot 1.96798^n$, a result that was confirmed by a different method in Clote et al. [2], where additionally the expected number of base pairs was shown to be $0.31724 \cdot n$.

The interest in combinatorics of saturated structures is that saturated structures are exactly the kinetically trapped structures with respect to the Nussinov energy model, in which each base pair receives a stabilizing energy contribution of -1 . In 2001, Evers and Giegerich [5] developed a dynamic programming algorithm to enumerate those structures in which no stem can be extended. When a strictly positive minimal value is specified for the length of the stems, these structures are saturated in the sense of Zuker [16]. However, as mentioned in [1], when the lengths of stems are not constrained, there are structures that are saturated in the sense of Evers and Giegerich [5], which are not saturated in the sense of Zuker [16]. For clarity of exposition, we will call a secondary structure *G-saturated* if no stem can be extended. In this paper, we give an enumerative framework based on weighted plane trees that allows us to enumerate *G-saturated* structures (as well as recover the enumeration of secondary structures and of saturated structures).

2 Definitions

An RNA secondary structure for a given RNA sequence a_1, \dots, a_n of length n is defined to be a set S of ordered pairs (i, j) , with $1 \leq i < j \leq n$, such that the following conditions are satisfied.

1. *Watson-Crick and wobble pairs*: If $(i, j) \in S$, then $\{a_i, a_j\} \in \{\{A, U\}\{G, C\}\{G, U\}\}$.
2. *No base triples*: If (i, j) and (i, k) belong to S , then $j = k$; if (i, j) and (k, j) belong to S , then $i = k$.
3. *Nonexistence of pseudoknots*: If (i, j) and (k, ℓ) belong to S , then it is not the case that $i < k < j < \ell$.
4. *Threshold requirement for hairpins*: If (i, j) belongs to S , then $j - i > \theta$, for a fixed value $\theta \geq 0$; i.e. there must be at least θ unpaired bases in a hairpin loop.

²The *shape* of a secondary structure was defined by Giegerich [14] to represent its branching topology; for instance, the shape of the well-known clover-leaf structure of tRNA is $[[] [] []]$.

For software, such as MFOLD and RNAfold, that predicts RNA secondary structure, θ is taken to be 3; i.e., for reasons related to steric constraints, every hairpin is required to contain at least three unpaired bases.

A pair $(i, j) \in S$ is called a *link*. An element i is said to be *linked* if it is involved in a link and *free* otherwise. A link (i, j) is said to be *stacked* onto another link (i', j') is $i = i' + 1$ and $j' = j - 1$. A *stem* is a maximal sequence ℓ_0, \dots, ℓ_k of links such that ℓ_{i+1} is stacked onto ℓ_i for $0 \leq i \leq k - 1$; the value k is called the *length* of the stem. In some applications a threshold condition on stems is required:

5. *Threshold requirement for stems:* Each stem has length at least τ , for a fixed value $\tau \geq 0$.

Note that Condition (5) is of no effect for $\tau = 0$.

In this paper, we are concerned with the asymptotic number of locally optimal structures. In order to employ generating functions, we will need to assume the homopolymer model (following a convention established by Stein and Waterman [12]), meaning that any position can pair with any other position (arbitrary base pairs, not only Watson-Crick and wobble pairs). We thus define a secondary structure of a *homopolymer* of length n to be a set S of base pairs (i, j) , where $1 \leq i < j \leq n$, such that the previous conditions (2,3,4,5) are satisfied.

3 Duality: RNA secondary structure \leftrightarrow weighted plane tree

It is well known that secondary structures have a tree shape, and there are several ways to formulate it. Here we find convenient to associate in a bijective way to a secondary structure (in the homopolymer formulation) a rooted plane tree with nonnegative integers (weights) at the corners and at the edges. The transformation is shown in Figure 1. Start with a secondary structure S of length n , the elements in the sequence being ranked from 1 to n . Call *segment* of S a sequence $i, i + 1, \dots, j$ such that $i < j$ and: (i) either $i = 0$, or $1 \leq i \leq n$ and the element i is linked, (ii) either $j = n + 1$, or $1 \leq j \leq n$ and the element j is linked, (iii) all elements in $i + 1, \dots, j - 1$ are free. Note that there are $j - i - 1$ free elements in the segment. Then perform two reduction operations on S :

Stem-reduction Replace each stem ℓ_0, \dots, ℓ_k by a single link.

Segment-reduction Replace each segment by a unit segment (with no free element on it).

Call R the reduced structure (which has no free element). Given the standard plane representation of R , draw a vertex, called a *dual vertex* in each region, and for each link of R , draw a *dual edge* connecting the vertices in the regions on each side of the link. The obtained figure (keeping the dual vertices and dual edges only) is a rooted plane tree T . Note that each edge of T corresponds to a link of R (hence corresponds to a stem of S), and each corner of T corresponds to a segment of S . We *weight* T by giving to each of its corners a weight corresponding to the number of free elements in the corresponding segment, and giving to each of its edges a weight corresponding to the length of the corresponding stem. Several parameters are in correspondence through the bijection (we use the standard terminology for parameters of secondary structures, a node of a tree is called a *leaf* if its arity 0 and an *inner node* if it has positive arity):

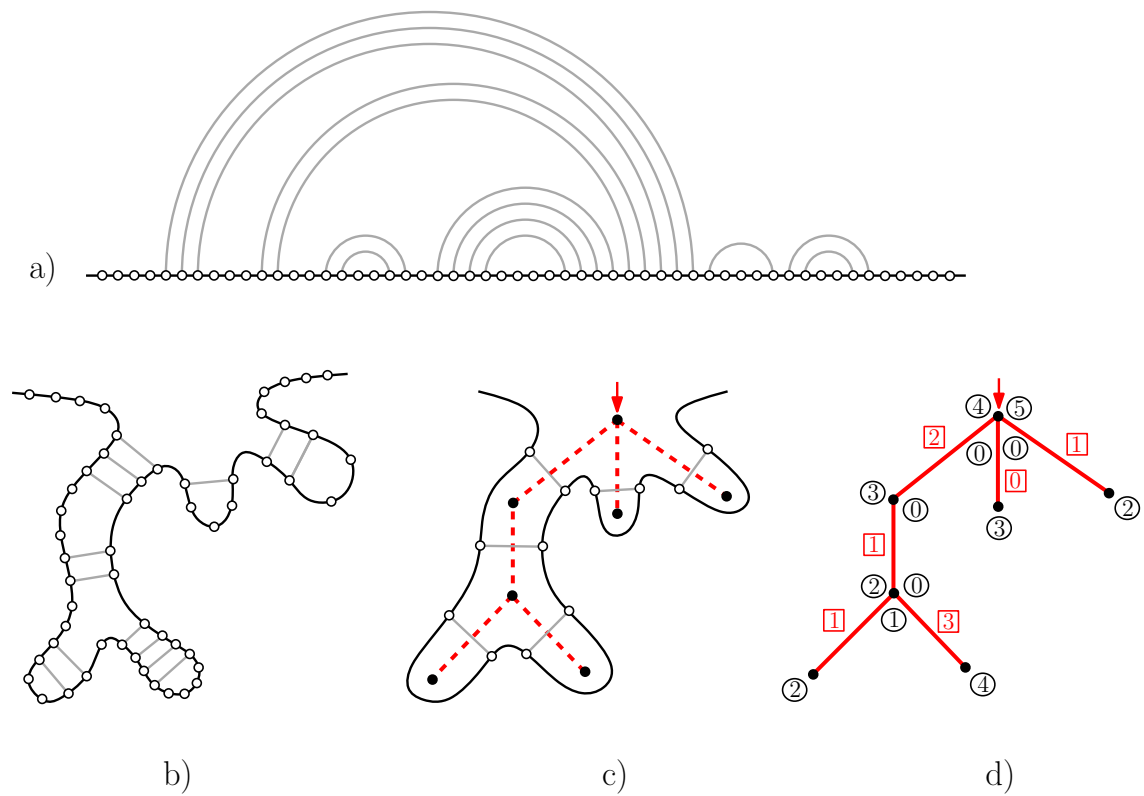


Figure 1: (a) A (homopolymer) secondary structure, (b) deformed into a tree-like shape, (c) the reduced structure superimposed with the dual rooted plane tree (in dashed lines, with the root indicated by an ingoing arrow), (d) the rooted plane tree with weights at corners (surrounded by circles) to indicate segment lengths, and weights at edges (surrounded by squares) to indicate stem lengths.

sequence S	\leftrightarrow	weighted tree T
hairpin	\leftrightarrow	leaf
bulk	\leftrightarrow	inner node with one child
multiloop	\leftrightarrow	inner node with several children
segment with L free elements	\leftrightarrow	corner of weight L
stem of length k	\leftrightarrow	edge of weight k

Note also that the number of links of S is the number $|E|$ of edges plus the total weight W_e over all edges, and that the number of free elements of S is the total weight W_c over all corners, hence the length n of S satisfies $n = 2|E| + 2W_e + W_c$.

A weighted rooted plane tree with at least one edge is called *admissible* if it corresponds to a valid secondary structure (which has at least one link), i.e., if the weights satisfy the following conditions:

1. Each non-root node with one child has at least one of its two incident corners of positive weight (otherwise the stem-reduction would not have been complete).
2. Each corner at a leaf has weight at least θ (to satisfy the θ -threshold condition).
3. Each edge has weight at least τ (to satisfy the τ -threshold condition).

4 Enumeration of locally optimal secondary structures

4.1 Generating functions

For $r \geq 1$, a *weighted combinatorial class indexed by r parameters* is a set \mathcal{A} together with a *weight-function* W from \mathcal{A} to \mathbb{R} and r *parameter-functions* P_1, \dots, P_r (one for each parameter) from \mathcal{A} to \mathbb{N} such that for any fixed integers n_1, \dots, n_r , the set of structures $\gamma \in \mathcal{A}$ such that $P_1(\gamma) = n_1, \dots, P_r(\gamma) = n_r$ is finite. This set is denoted $\mathcal{A}[n_1, \dots, n_r]$. The corresponding multivariate generating function is

$$A(x_1, \dots, x_r) := \sum_{\gamma \in \mathcal{A}} x_1^{P_1(\gamma)} \dots x_r^{P_r(\gamma)} W(\gamma). \quad (1)$$

We say that variable x_i *marks* the parameter P_i , for $1 \leq i \leq r$. We also use the notation

$$[x_1^{n_1} \dots x_r^{n_r}] A(x_1, \dots, x_r) := \sum_{\gamma \in \mathcal{A}[n_1, \dots, n_r]} W(\gamma).$$

In general we consider *enumerative* generating functions, where $W(\cdot)$ assigns weight 1 to each structure. However we allow ourselves to weight these structures, e.g., to weight each secondary structure by $p^{\#(\text{links})}$, with p a so-called *stickiness parameter*. The variables x_i are a priori considered as formal, but one can also evaluate a generating function at given values, provided the sum converges. The *convergence domain* of $A(x_1, \dots, x_r)$ is the set of r -tuples (x_1, \dots, x_r) of nonnegative real values such that $A(x_1, \dots, x_r)$ converges.

As a first example, we briefly recall here how to enumerate (homopolymer) secondary structures, via the dual representation by weighted rooted plane trees and using generating functions. Let \mathcal{F} be the family of rooted plane trees, possibly reduced to a single vertex, with

some marked corners (to be occupied by positive weights later on) incident to inner nodes such that each node with one child has at least one marked corner. Let $F \equiv F(u, v, x)$ be the generating function of \mathcal{F} where u marks the number of leaves, v marks the number of marked corners, and x marks the number of edges. When the root-node v has arity 1, exactly one of its two corners is marked, hence the generating function for trees in \mathcal{F} whose root-node has arity 1 is $2vxF$. When the root-node v has arity $k \geq 2$, there are $(k+1)$ corners incident to v , and each of these can be marked (independently). Hence the generating function for trees in \mathcal{F} where the root-node has arity k is $(1+v)^{k+1}x^kF^k$. Consequently, F satisfies

$$F = u + (2v + v^2)xF + \sum_{k \geq 2} x^k(1+v)^{k+1}F^k = u + \frac{x(1+v)^2F}{1-x(1+v)F} - xF. \quad (2)$$

Let \mathcal{G} be the family of rooted plane trees with at least one edge and with some marked corners (to be occupied by positive weights later on) incident to inner nodes such that each non-root node with one child has at least one marked corner. Let $G \equiv G(u, v, x)$ be the generating function of \mathcal{G} where u marks the number of leaves, v marks the number of marked corners, and x marks the number of edges. Again by decomposing at the root, we get

$$G = \sum_{k \geq 1} x^k(1+v)^{k+1}F^k = \frac{x(1+v)^2F}{1-x(1+v)F}. \quad (3)$$

Let $g(t, s)$ be the series counting secondary structures with at least one link, where t marks the number of free elements, and s marks the number of links. Note that $g(t, s)$ is also the generating function of admissible rooted weighted plane trees where t marks the total weight over corners, and s marks the number of edges plus the total weight over edges. Such a tree is uniquely obtained from a tree in \mathcal{G} where each corner at a leaf is assigned a weight of value at least θ , each non-marked corner at an inner node is assigned weight 0, each marked corner is assigned a positive weight, and each edge is assigned a weight of value at least τ . Hence we have $g(t, s) = G(U, V, X)$, where

$$U := \sum_{i \geq \theta} t^i = \frac{t^\theta}{1-t}, \quad V = \frac{t}{1-t}, \quad X := s \sum_{i \geq \tau} s^i = \frac{s^{\tau+1}}{1-s}.$$

To summarize, we have an expression (written as a system of two equations) for the generating function $g(t, s)$ enumerating secondary structures with at least one link, where t marks the number of free elements and s marks the number of links (the generating function of all secondary structures, including the ones with no link, is clearly $g(t, s) + t + t^2 + \dots = g(t, s) + \frac{t}{1-t}$). Indeed, if we define $f(t, s) := F(U, V, X)$, then we easily see (since the substitutions of variables are rational expressions whose series-expansion have nonnegative coefficients) that there is a one-line equation specifying $f(t, s)$, of the form $f(t, s) = Q(t, s, f(t, s))$, with $Q \equiv Q(t, s, y)$ a rational expression whose series-expansion (in s, t, y) has nonnegative coefficients. And there is a rational expression $R \equiv R(t, s, y)$ whose series-expansion has nonnegative coefficients and such that $g(t, s) = R(t, s, f(t, s))$. Precisely

$$Q = \text{substitute} \left(u = \frac{t^\theta}{1-t}, v = \frac{t}{1-t}, x = \frac{s^{\tau+1}}{1-s} \right) \text{ into } u + \frac{x(1+v)^2y}{1-x(1+v)y} - xy,$$

$$R = \text{substitute} \left(v = \frac{t}{1-t}, x = \frac{s^{\tau+1}}{1-s} \right) \text{ into } \frac{x(1+v)^2y}{1-x(1+v)y}.$$

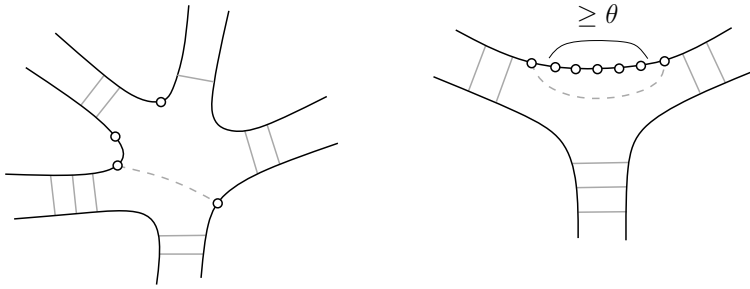


Figure 2: Situations where it is possible to add a link to a secondary structure.

This allows us to extract the counting coefficients. Let $g_p(t)$ be the weighted generating function of secondary structures where t marks the length, and where each structure has weight $p^{\#(\text{links})}$: $g_p(t) = g(t, pt^2) + t/(1-t)$ (the term $t/(1-t)$ gathers secondary structures with no link); for instance for $\theta = 1$ and $\tau = 0$ we find

$$g_p(t) = t + t^2 + (1+p)t^3 + (1+3p)t^4 + (1+6p+p^2)t^5 + (1+10p+6p^2)t^6 + (1+15p+20p^2+p^3)t^7 + \dots$$

4.2 Counting saturated structures

The Nussinov energy $E(S)$ of a secondary structure S is defined as $E(S) = -L(S)$, with $L(S)$ the number of links in S . A secondary structure S is called *saturated* (or *locally optimal* for the Nussinov energy) if it is not possible to add a link to S (i.e., decrease the energy by 1) while keeping a valid secondary structure.

Lemma 1. *Assume $\tau = 0$ (no restriction on the lengths of stems). Saturated secondary structures with at least one link correspond to admissible weighted rooted plane trees such that:*

- all corners have weight at most $\theta + 1$,
- at each node there is at most one corner of strictly positive weight.

Proof. As shown in Figure 2, if there are two positive corners at the same inner node, then it is possible to add a link. Also, if there is a corner with weight at least $\theta + 2$ then one can link the first and last free elements in the corresponding segment. Hence the weight of each corner is at most $\theta + 1$. And these are the only two situations where it is possible to add a link without breaking planarity nor breaking the θ -threshold condition. \square

Call \mathcal{F} the family of rooted plane trees with some marked corners incident to inner nodes (these marked corners are to be occupied by positive weights later on) such that: (i) each node with one child has exactly one marked corner, (ii) each node with several children has at most one marked corner. Let $F \equiv F(u, v, x)$ be the generating function of \mathcal{F} where u marks the number of leaves, v marks the number of marked corners, and x marks the number of edges. When the root-node v has arity 1, exactly one of its two corners is marked, hence the generating function for trees in \mathcal{F} whose root-node has arity 1 is $2vxF$. When the root-vertex v has arity $k \geq 2$, there are $(k+1)$ corners incident to v , and at most one of these corners

has positive weight. Hence the generating function for trees in \mathcal{F} where the root-vertex has arity k is $(1 + (k + 1)v)x^k F^k$. Consequently, F satisfies

$$F = u + 2vxF + \sum_{k \geq 2} (1 + (k + 1)v)x^k F^k,$$

Hence, using the identity $\sum_{k \geq 0} (k + 1)A^k = 1/(1 - A)^2$, F satisfies

$$F = u + \frac{x^2 F^2}{1 - xF} + \frac{v}{(1 - xF)^2} - v. \quad (4)$$

Now let \mathcal{G} be the family of rooted plane trees with at least one edge, and with marked corners incident to inner nodes such that: (i') each node v with one child has exactly one marked corner if v is different from the root-node, and has *at most* one marked corner if v is the root-node, (ii) each node with several children has at most one marked corner. Let $G \equiv G(u, v, x)$ be the generating function of \mathcal{G} where u, v, x mark respectively the numbers of leaves, marked corners, and edges. Decomposing again at the root, we get

$$G = \sum_{k \geq 1} (1 + (k + 1)v)x^k F^k = \frac{xF}{1 - xF} + \frac{v}{(1 - xF)^2} - v. \quad (5)$$

We take here $\tau = 0$ (no restriction on the lengths of stems). Let $g(t, s)$ be the generating function of saturated secondary structures with at least one link, where t marks the number of free elements and s marks the number of links. Then Lemma 1 ensures that $g(t, s) = G(U, V, X)$, where

$$U = t^\theta(1 + t), \quad V = t + \dots + t^{\theta+1} = \frac{t - t^{\theta+2}}{1 - t}, \quad X = \frac{s}{1 - s}.$$

To summarize (in a similar way as for general structures), we have an expression (written as a system of two equations) for the generating function $g(t, s)$ enumerating *saturated* secondary structures with at least one link, where t marks the number of free elements and s marks the number of links (the generating function of all saturated secondary structures, including the ones with no link, is $g(t, s) + t + \dots + t^{\theta+1} = g(t, s) + \frac{t - t^{\theta+2}}{1 - t}$). Indeed, if we define $f(t, s) := F(U, V, X)$, then there is a one-line equation specifying $f(t, s)$, of the form $f(t, s) = Q(t, s, f(t, s))$, with $Q(t, s, y)$ a rational expression whose series-expansion (in s, t, y) has nonnegative coefficients. And there is a rational expression $R(t, s, y)$ whose series-expansion has nonnegative coefficients and such that $g(t, s) = R(t, s, f(t, s))$. Precisely

$$Q = \text{substitute} \left(u = t^\theta(1 + t), v = \frac{t - t^{\theta+2}}{1 - t}, x = \frac{s}{1 - s} \right) \text{ into } u + \frac{x^2 y^2}{1 - xy} + \frac{v}{(1 - xy)^2} - v,$$

$$R = \text{substitute} \left(v = \frac{t - t^{\theta+2}}{1 - t}, x = \frac{s}{1 - s} \right) \text{ into } \frac{xy}{1 - xy} + \frac{v}{(1 - xy)^2} - v.$$

Again this allows us to extract the counting coefficients. Let $g_p(t)$ be the weighted generating function of saturated secondary structures where t marks the length, and where each structure has weight $p^{\#(\text{links})}$: $g_p(t) = g(t, pt^2) + t + \dots + t^{\theta+1}$; for $\theta = 1$ and $\tau = 0$ we find

$$g_p(t) = t + t^2 + pt^3 + 3pt^4 + (4p + p^2)t^5 + (2p + 6p^2)t^6 + (17p^2 + p^3)t^7 + \dots$$

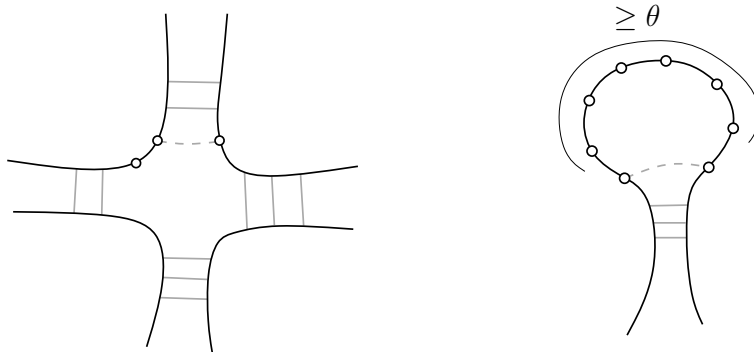


Figure 3: Situations where it is possible to extend a stem of a secondary structure.

4.3 Counting G-saturated structures

The *base stacking energy* $E(S)$ of a secondary structure S is defined as $E(S) := -T(S)$, with $T(S)$ the sum of sizes of all stems of S . A (homopolymer) secondary structure is called *G-saturated* (locally optimal for the base stacking energy) if it is not possible to add a link so as to extend a stem (i.e., decrease by 1 the base stacking energy). In general, the addition of a link to a secondary structure either creates a new stem of length 0 or extends an already existing stem. Hence, in a G-saturated structure a valid link addition always creates a new stem of length 0. In case $\tau > 0$, creating a stem of length 0 is not a valid link addition (since the stems must have positive length), hence no valid link addition to a G-saturated is possible for $\tau > 0$. In other words, the concepts of saturated and of G-saturated structures coincide when $\tau > 0$ (whereas for $\tau = 0$ the class of saturated structures is strictly contained in the class of G-saturated structures). In this section we enumerate G-saturated structures according to the number of free elements and the number of links, for any given values of the threshold parameters τ and θ .

Again we formulate the conditions on the dual representation. For this purpose we define adjacency of corners. Two corners c and c' of a rooted plane tree T are called *adjacent* if they are incident to the same vertex v of T and there is an edge e incident to v such that c and c' are the corners incident to v on each side of e . Note that the two corners on each side of the root (the root is represented as an ingoing arrow in Figure 1) are considered as adjacent only when the root-node v has arity 1 (in which case they are adjacent through the unique edge incident to v).

Lemma 2. *The G-saturated secondary structures with at least one link correspond to admissible weighted rooted plane trees such that:*

- *the corners at leaves have weight at most $\theta + 1$,*
- *any two adjacent corners can not both have strictly positive weight.*

Proof. As shown in Figure 3, if there are two adjacent positive corners, then it is possible to add a link so as to extend an existing stem. Also, if there is a corner of weight at least $\theta + 2$ at a leaf ℓ , then one can link the first and last free elements in the corresponding segment and thus extend the stem associated to the edge leading to ℓ . Hence the weight of a corner at

a leaf is at most $\theta + 1$. And these are the only two situations where it is possible to extend a stem without breaking planarity nor breaking the θ -threshold and τ -threshold condition. \square

Call \mathcal{F} the family of rooted plane trees with some marked corners incident to inner nodes (again these marked corners are to be occupied by positive weights later on) such that: (i) each inner node with one child has exactly one marked corner, (ii) two corners can not both be marked if they are adjacent or if they are the two corners on each side of the root (the root is indicated by an ingoing arrow in Figure 1). Let $F \equiv F(u, v, x)$ be the generating function of \mathcal{F} where u marks the number of leaves, v marks the number of marked corners, and x marks the number of edges. Finding an equation satisfied by F is a little more involved than for saturated structures. At first we need a preliminary study on independent sets (i.e., sets containing only pairwise non-adjacent elements) on a k -sequence or on a k -cycle.

For $k > 0$ and $m \leq k$, let $c_{k,m}$ (resp. $s_{k,m}$) be the number of ways of choosing m marked elements on the oriented cycle $(1, 2, \dots, k)$ (resp. sequence $1, 2, \dots, k$) of k elements such that no two consecutive elements are marked, and let $C_k = C_k(v) := \sum_m c_{k,m} v^m$ (resp. $S_k = S_k(v) := \sum_m s_{k,m} v^m$) be the corresponding (polynomial) generating function. The polynomials S_k are well-known to be the *Fibonacci polynomials* and satisfy an easy recurrence which we briefly recompute here. We take the convention $S_0 = 1$. Let $k \geq 2$. If an independent set on the k -sequence starts with a marked element, then the next element is forbidden and the remaining $(k - 2)$ -sequence might be occupied by any independent set; this gives a contribution vS_{k-2} in S_k , where the factor v takes account of the first element being marked. If an independent set on the k -sequence starts with a non-marked element, then the remaining $(k - 1)$ -sequence might be occupied by any independent set; this gives a contribution S_{k-1} in S_k . Therefore

$$S_k = vS_{k-2} + S_{k-1} \quad \text{for } k \geq 2, \quad S_0 = 1, \quad S_1 = 1 + x.$$

Now define $S \equiv S(v, z) := \sum_{k \geq 0} S_k(v) z^k$. The recurrence on S_k above multiplied by z^k and summed over $k \geq 2$ yields $S - S_0 - zS_1 = vz^2S + z(S - S_0)$. With $S_0 = 1$ and $S_1 = 1 + v$ we obtain

$$S = \frac{1 + vz}{1 - z - vz^2}.$$

Let us go back to independent sets on the k -cycle $(1, \dots, k)$, for $k \geq 3$. In such a set, either 1 is occupied, in which case the adjacent elements 2 and k are unoccupied and the remaining segment $3, \dots, k - 1$ might be occupied by any independent set. This gives contribution vS_{k-3} to C_k . If 1 is unoccupied, then the remaining segment $2, \dots, k$ might be occupied by any independent set; this gives contribution S_{k-1} to C_k . Consequently

$$C_k = vS_{k-3} + S_{k-1} \quad \text{for } k \geq 3.$$

If the root-node v of a tree in \mathcal{F} has arity 1 then exactly one of its two incident corners is marked (by definition of \mathcal{F}), thus the generating function of trees in \mathcal{F} whose root-node has arity 1 is $2vx F$; if v has arity $k \geq 2$ then the marked corners around v form an independent set (no two consecutive corners are marked). Thus, for $k \geq 2$, the generating function of trees in \mathcal{F} whose root-node has arity k is $C_{k+1}(v)x^k F^k$ (since there are $k + 1$ corners incident to

the root-node). Consequently F satisfies

$$\begin{aligned}
F &= u + 2vxF + \sum_{k \geq 2} C_{k+1}(v)x^k F^k \\
&= u + 2vxF + \sum_{k \geq 2} [vS_{k-2} + S_k]x^k F^k \\
&= u + 2vxF + vx^2 F^2 S(v, xF) + (S(v, xF) - 1 - (1+v)xF).
\end{aligned}$$

Using the rational expression of S above and rearranging, we obtain

$$F = u + vxF + \frac{(1 + vx^2 F^2)(1 + vxF)}{1 - xF - vx^2 F^2} - 1 - xF. \quad (6)$$

Now let \mathcal{G} be the family of rooted plane trees with at least one edge and where some corners at inner nodes are marked such that (i) each non-root inner node of arity 1 has exactly one marked corner, (ii) two adjacent corners can not both be marked. And let $G \equiv G(u, v, x)$ be the generating function of \mathcal{G} where u marks the number of leaves, v marks the number of marked corners, and x marks the number of edges. The difference between \mathcal{G} and \mathcal{F} is at the root-vertex: in \mathcal{G} the two corners on each side of the root are allowed to be both marked when the root-vertex has more than one child, and are allowed to be both unmarked when the root-vertex has one child. So we have

$$G = \sum_{k \geq 1} S_{k+1}(v)x^k F^k.$$

Using the rational expression of S above and rearranging, we obtain the following expression for G in terms of F :

$$G = \frac{xvF(1 + 2v + (1+v)vxF)}{1 - xF - vx^2 F^2}. \quad (7)$$

Now let $g(t, s)$ be the generating function of G -saturated structures with at least one link, where t marks the number of free elements and s marks the number of links. By Lemma 2,

$$g(t, s) = G(U, V, X), \quad (8)$$

where

$$U = t^\theta(1 + t), \quad V = \frac{t}{1 - t}, \quad X = \frac{s^{\tau+1}}{1 - s}.$$

The conclusion is similar to the other two cases (general structures, saturated structures): we have an expression (written as a system of two equations) for the generating function $g(t, s)$ enumerating G -saturated secondary structures with at least one link, where t marks the number of free elements and s marks the number of links (the generating function of all G -saturated secondary structures, including the ones with no link, is $g(t, s) + t + t^2 + \dots = g(t, s) + \frac{t}{1-t}$). Indeed, if we define $f(t, s) := F(U, V, X)$, then there is a one-line equation specifying $f(t, s)$, of the form $f(t, s) = Q(t, s, f(t, s))$, with $Q(t, s, y)$ a rational expression whose series-expansion (in s, t, y) has nonnegative coefficients. And there is a rational expression $R(t, s, y)$ whose series-expansion has nonnegative coefficients and such that $g(t, s) = R(t, s, f(t, s))$. Precisely

$$Q = \text{substitute} \left(u = t^\theta(1 + t), v = \frac{t}{1 - t}, x = \frac{s^{\tau+1}}{1 - s} \right) \text{ into } u + vxy + \frac{(1 + vx^2 y^2)(1 + vxy)}{1 - xy - vx^2 y^2} - 1 - xy,$$

$$R = \text{substitute} \left(v = \frac{t}{1-t}, x = \frac{s^{\tau+1}}{1-s} \right) \text{ into } \frac{xy(1+2v+(1+v)vxy)}{1-xy-vx^2y^2}.$$

Again this allows us to extract the counting coefficients. Let $g_p(t)$ be the weighted generating function of G -saturated secondary structures where t marks the length, and where each structure has weight $p^{\#(\text{links})}$: $g_p(t) = g(t, pt^2) + t/(1-t)$; for $\theta = 1$ and $\tau = 0$ we find

$$g_p(t) = t + t^2 + (1+p)t^3 + (1+3p)t^4 + (1+4p+p^2)t^5 + (1+4p+6p^2)t^6 + (1+4p+17p^2+p^3)t^7 + \dots$$

5 Asymptotic results

5.1 Asymptotic enumeration

We show here that the number of structures of length n follows a universal asymptotic behaviour in $c \gamma^n n^{-3/2}$ (with c and γ explicit positive constants), which is typical of tree-structures. The proof classically relies on the Drmota-Lalley-Woods theorem [6, VII.6], which we recall at first. Consider an equation of the form

$$a(t) = \Phi(t, a(t)), \tag{9}$$

where $\Phi(t, y)$ is a rational expression in t and y . Such an equation is called *admissible* if the following conditions are satisfied:

- the rational expression $\Phi(t, y)$ has a series-expansion in t and y with nonnegative coefficients, is nonaffine in y , and satisfies ³ $\Phi(0, 0) = 0$ and $\Phi_y(0, 0) = 0$,
- the unique generating function $y = a(t)$ solution of (9) is aperiodic, i.e., can not be written as $a(t) = t^q \tilde{a}(t^p)$ for some integers p, q with $p \geq 2$.

There is an easy criterion to check the aperiodicity condition: it suffices to prove that there is some n_0 such that $[t^n]a(t) > 0$ for $n \geq n_0$.

Theorem 3 (Drmota-Lalley-Wood). *Let $y = a(t)$ be the generating function that is the unique solution of an admissible equation $y = \Phi(t, y)$. Then*

$$[t^n]a(t) \sim c \gamma^n n^{-3/2},$$

where $\gamma = 1/t_0$, with (t_0, y_0) the unique pair in the convergence domain of $\Phi(t, y)$ that is solution of the singularity system:

$$y = \Phi(t, y), \quad \Phi_y(t, y) = 1;$$

and where

$$c = \sqrt{t_0 \Phi_t(t_0, y_0) / (2\pi \Phi_{y,y}(t_0, y_0))}.$$

Moreover, if $\Psi(t, y)$ is a rational expression not constant in y , that has a series-expansion with nonnegative coefficients, and such that the convergence domain of $\Psi(t, y)$ is contained in the convergence domain of $\Phi(t, y)$, then the coefficients of the generating function $b(t) := \Psi(t, a(t))$ behave as

$$[t^n]b(t) \sim d \gamma^n n^{-3/2},$$

where $d = c \cdot \Psi_y(t_0, y_0)$.

³We use the subscript notation for partial derivatives.

Remark 4. *The Drmota-Lalley-Wood theorem is classically proved (e.g. in [6, VII.6]) for polynomial systems (i.e., for $\Phi(t, y)$ a polynomial). But one easily checks that, more generally, if $\Phi(t, y)$ is a bivariate series that diverges at all its singularities, then the conclusions remain the same.*

From the Drmota-Lalley-Wood theorem we obtain

Proposition 5. *Let p be a fixed positive real value (stickiness parameter). Let $g_p(t)$ be the univariate generating function of general (resp. saturated, resp. G-saturated) homopolymer secondary structures, where t marks the length of the sequence and where each structure has weight $p^{\#(\text{links})}$.*

Then, for any values of the threshold-parameters θ and τ ($\tau = 0$ if one considers saturated structures), there are computable constants $c > 0$ and $\gamma > 1$ (depending on τ , θ , p , and in which setting: general, saturated, or G-saturated) such that

$$[t^n]g_p(t) \sim c \gamma^n n^{-3/2}.$$

Proof. Recall that, in each of the three settings (general, saturated, G-saturated), $g(t, s)$ denotes the generating function of secondary structures with at least one link, where t marks the number of free elements and s marks the number of links. We have seen that, in each of the three settings, there are two rational expressions $Q(t, s, y)$ and $R(t, s, y)$ that have nonnegative coefficients (in the series-expansion), and there is an adjoint generating function $f(t, s)$ such that $f(t, s) = Q(t, s, f(t, s))$ and $g(t, s) = R(t, s, f(t, s))$. In addition, the convergence domain of $Q(t, s, y)$ is clearly the same as the convergence domain of $R(t, s, y)$; for instance, for G-saturated structures, the convergence domain is the set of nonnegative triples (t, s, y) such that $t < 1$, $s < 1$, and $xy + vx^2y^2 < 1$, where $v = t/(1-t)$ and $x = s^{\tau+1}/(1-s)$. Note that in all three settings, $f(0, 0) = 1$ for $\theta = 0$ and $f(0, 0) = 0$ for $\theta > 0$. If we set $a(t) := f(t, pt^2) - \mathbf{1}_{\theta=0}$ (with θ the threshold parameter) and $b(t) := g(t, pt^2)$, then we are in the conditions of the Drmota-Lalley-Wood theorem, with $\Phi(t, y) := Q(t, pt^2, y + \mathbf{1}_{\theta=0}) - \mathbf{1}_{\theta=0}$ and $\Psi(t, y) := R(t, pt^2, y + \mathbf{1}_{\theta=0})$. The conditions for Φ and Ψ are readily checked, we show now the aperiodicity of $a(t) := f(t, pt^2)$ (proving that the n th coefficient is strictly positive for n large enough). Note that it is enough to consider $p = 1$ (the strict positivity of $[t^n]f(t)$ does not depend on $p > 0$). In each of the three settings (general, saturated, G-saturated), $a(t)$ is the enumerative generating function of some explicit class of rooted weighted plane trees. For instance, for saturated structures, $a(t)$ counts admissible rooted weighted plane trees with all corners of weight at most $\theta + 1$, with at most one positive corner per node, and where each node of arity 1 has exactly one positive corner. For $i \geq \tau$, consider the weighted rooted plane tree T_i made of one edge e leading to a leaf ℓ , with weight 1 (resp. 0) at the corner to the left (resp. right) of the root, with weight i on e and weight θ on ℓ . And consider the tree T'_i defined exactly as T_i except that ℓ has weight $\theta + 1$. Note that T_i contributes to $[t^{2i+\theta+3}]a(t)$ and T'_i contributes to $[t^{2i+\theta+4}]a(t)$. Hence $[t^n]a(t) > 0$ for all $n \geq 2\tau + \theta + 3$, so $a(t)$ is aperiodic. In exactly the same way, $a(t)$ is aperiodic in the general setting and in the G-saturated setting.

Theorem 3 ensures that there are $c > 0$ and $\gamma > 0$ such that $[t^n]g(t, pt^2) \sim c\gamma^n n^{-3/2}$; actually we have $\gamma > 1$ since (according to Theorem 3) there is some y_0 such that $(1/\gamma, y_0)$ is in the convergence domain of $\Phi(t, y)$, and since clearly any (t_0, y_0) in the convergence domain of $\Phi(t, y)$ satisfies $t_0 < 1$ (indeed $Q(t, s, y)$ involves the quantity $1/(1-t)$, in each of the three settings). The generating function $g_p(t)$ (which includes also secondary structures with

	$p = 1 \quad \theta = 1 \quad \tau = 0$	$p = 3/8 \quad \theta = 1 \quad \tau = 0$
General	$1.104366 \cdot n^{-3/2} \cdot 2.618034^n$	$1.637405 \cdot n^{-3/2} \cdot 2.041013^n$
Saturated	$1.074271 \cdot n^{-3/2} \cdot 2.354674^n$	$1.527438 \cdot n^{-3/2} \cdot 1.705128^n$
G-saturated	$1.088582 \cdot n^{-3/2} \cdot 2.436901^n$	$1.632293 \cdot n^{-3/2} \cdot 1.826929^n$

Table 1: Asymptotic behaviour of the n th coefficient of the generating function $g_p(t)$ counting secondary structures (general, saturated, or G-saturated) with weight p on each link.

no link, as opposed to $g(t, s)$) satisfies $g_p(t) = g(t, pt^2) + t/(1-t)$ for secondary and for G-saturated structures, and satisfies $g_p(t) = g(t, pt^2) + t + \dots + t^{\theta+1}$ for saturated structures. So the additional term gathering saturated structures with no link has negligible asymptotic contribution in all cases. \square

For $p = 1$, $g_p(t)$ is the enumerative generating function of homopolymer structures. Another value of interest is $p = 3/8$. Indeed, if we want to count RNA secondary structures (each base is labelled by a letter in $\{A, G, C, U\}$) instead of homopolymers, this corresponds to giving weight 4 to each free element (because there are 4 possible labels) and giving weight 6 to each pair of linked elements (because there are 6 allowed labellings out of $4^2 = 16$, due to the Watson-Crick and wobble pairs). Therefore the corresponding enumerative generating function is $g(4t, 6t^2)$. We have

$$[t^n]g(4t, 6t^2) = 4^n [t^n]g(t, 3t^2/8) = 4^n [t^n]g_{3/8}(t).$$

In other words, $[t^n]g_{3/8}$ is the *expected number* of RNA secondary structures with the desired properties (general, saturated, or G-saturated) on a random sequence of size n (random word in $\{A, G, C, U\}^n$).

Table 1 shows the asymptotic behaviour of $[t^n]g_p(t)$ for $p = 1$ and $p = 3/8$ in the three settings. (The methodology to compute γ for saturated structures using computer algebra tools is detailed in [2].) An interesting observation is for $\{\theta = 1, \tau = 2\}$. For a random word X in $\{A, G, C, U\}^n$, let $N_{\text{gen}}(X)$ be the number of general secondary structures with X as underlying sequence, and let $N_{\text{G-sat}}(X)$ be the number of G-saturated structures with X as underlying sequence. It was observed experimentally in [] that $N_{\text{G-sat}}(X) \approx \sqrt{N_{\text{gen}}(X)}$. As mentioned above, for general (resp. G-saturated) structures, $[t^n]g_{3/8}$ is the expectation of $N_{\text{gen}}(X)$ (resp. of $N_{\text{G-sat}}(X)$). We indeed find that, for $\theta = 1$ and $\tau = 2$, the exponential growth rate of $[t^n]g_{3/8}$ is 1.291717 for general structures and is 1.150295 for G-saturated structures, which is only 1.2% away from $\sqrt{1.291717}$.

5.2 Limit law for the number of links

Using a theorem of Drmota [3] (closely related to the Drmota-Lalley-Wood theorem) we show that the number of links in a random secondary structure (general, saturated, or G-saturated) of length n is asymptotically a gaussian law with $\Theta(n)$ expectation and $\Theta(\sqrt{n})$ standard deviation.

Consider an equation of the form

$$a(t, u) = \Phi(t, u, a(t, u)), \tag{10}$$

	$p = 1 \quad \theta = 1 \quad \tau = 0$	$p = 3/8 \quad \theta = 1 \quad \tau = 0$
General	$0.276393 \cdot n + 0.211474 \cdot \sqrt{n} \cdot \mathcal{N}$	$0.230789 \cdot n + 0.218613 \cdot \sqrt{n} \cdot \mathcal{N}$
Saturated	$0.337361 \cdot n + 0.132800 \cdot \sqrt{n} \cdot \mathcal{N}$	$0.321153 \cdot n + 0.123935 \cdot \sqrt{n} \cdot \mathcal{N}$
G-saturated	$0.311958 \cdot n + 0.185032 \cdot \sqrt{n} \cdot \mathcal{N}$	$0.273773 \cdot n + 0.211618 \cdot \sqrt{n} \cdot \mathcal{N}$

Table 2: Asymptotic behaviour of the number of links (\mathcal{N} denotes a normal gaussian law).

where $\Phi(t, u, y)$ is a rational expression in t , u and y . Such an equation is called *admissible* if $\Phi(t, u, y)$ is nonconstant in u , has a series-expansion (in t , u , y) with nonnegative coefficients, the equation $y = \Phi(t, 1, y)$ is admissible (in the sense of Section 5.1), and there is a 3×3 -matrix $m[i, j]$ with integer coefficients and nonzero determinant such that $[t^{m[i,1]}u^{m[i,2]}y^{m[i,3]}]\Phi(t, u, y) > 0$ for all $i \in \{1, 2, 3\}$.

Theorem 6 (Drmotá [3]). *Let $y = a(t, u)$ be a generating function that is the unique solution of an admissible equation $y = \Phi(t, u, y)$. Assume that the generating function $b(t, u) = \sum_{\gamma \in \mathcal{G}} t^{|\gamma|} u^{\chi(\gamma)} W(\gamma)$ of a weighted combinatorial class \mathcal{G} is given by $b(t, u) = \Psi(t, u, a(t, u))$, with $\Psi(t, u, y)$ a rational expression with nonnegative coefficients (in the series-expansion), nonconstant in y , and such that the convergence domain of $\Psi(t, 1, y)$ is included in the one of $\Phi(t, 1, y)$. For $n \geq 0$ let $\mathcal{G}_n := \{\gamma \in \mathcal{G}, |\gamma| = n\}$, and define the random variable X_n as $\chi(\gamma)$, with γ a random structure in \mathcal{G}_n under the distribution*

$$P(\gamma) = \frac{W(\gamma)}{\sum_{\gamma \in \mathcal{G}_n} W(\gamma)}.$$

For $u > 0$ in a neighbourhood of 1, denote by $\rho(u)$ the radius of convergence of $y : t \rightarrow a(t, u)$, and let

$$\mu = -\frac{\rho'(1)}{\rho(1)}, \quad \sigma^2 = -\frac{\rho''(1)}{\rho(1)} - \frac{\rho'(1)}{\rho(1)} + \left(\frac{\rho'(1)}{\rho(1)}\right)^2.$$

Then μ and σ are strictly positive and $\frac{X_n - \mu \cdot n}{\sigma \sqrt{n}}$ converges as a random variable to a normal (gaussian) law.

Remark 7. *Again the theorem was originally proved for polynomial systems, but the arguments of the proof hold more generally when Φ is rational. The role of the condition involving the existence of a nonsingular 3×3 matrix is to grant the strict positivity of σ , as recently proved in [4].*

Proposition 8. *Let $p > 0$. For $n \geq 1$, let X_n be the number of links in a general (resp. saturated, resp. G-saturated) secondary structure of length n taken at random with weight proportional to $p^{\#(\text{links})}$ (uniformly at random when $p = 1$). Then there are computable strictly positive constants μ and σ (depending on p , θ , τ , and on which setting: general, saturated, or G-saturated) such that $(X_n - \mu \cdot n)/\sqrt{n}$ converges as a random variable to a normal (gaussian) law.*

Proof. In each of the three settings (general, saturated, G-saturated), we have called $g(t, s)$ the enumerative generating function of secondary structures with at least one link. We have

seen that there are two rational expressions $Q(t, s, y)$ and $R(t, s, y)$ that have nonnegative coefficients (in the series-expansion), and there is an adjoint generating function $f(t, s)$ such that $f(t, s) = Q(t, s, f(t, s))$ and $g(t, s) = R(t, s, f(t, s))$; and the convergence domain of $Q(t, s, y)$ is the same as the convergence domain of $R(t, s, y)$. Note that the bivariate series $g(t, put^2)$ (with variables t and u) is the weighted generating function of secondary structures (with at least one link) where t marks the length, u marks the number of links, and where each structure has weight $p^{\#(\text{links})}$. It is easily checked that, if we set $a(t, u) := f(t, put^2) - \mathbf{1}_{\theta=0}$ (with θ the threshold parameter) and $b(t) := g(t, put^2)$, then we are in the conditions of Theorem 6, with $\Phi(t, u, y) := Q(t, put^2, y + \mathbf{1}_{\theta=0}) - \mathbf{1}_{\theta=0}$ and $\Psi(t, u, y) := R(t, put^2, y + \mathbf{1}_{\theta=0})$. Indeed the 3×3 matrix condition is readily checked, and for $u = 1$ we get the equation of Proposition 5, where we have already checked that the conditions are satisfied. \square

Table 2 shows the asymptotic behaviour for some standard parameter values. (The methodology to compute μ for saturated structures using computer algebra tools is detailed in [2].) The case $p = 1$ corresponds to a homopolymer of length n taken uniformly at random, while the case $p = 3/8$ corresponds to a (uniformly) random secondary structure on a word in $\{A, G, C, U\}^n$. As expected, saturated structures tend to have more links than G-saturated structures, which tend to have more links than general structures.

Acknowledgements

É. Fusy is supported by the European project ExploreMaps —ERC StG 208471. P. Clote is supported by the National Science Foundation under grants DBI-0543506 and DMS-0817971, and by Digiteo Foundation project *RNAomics*. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- [1] P. Clote. Combinatorics of saturated secondary structures of RNA. *J. Comput. Biol.*, 13(9):1640–1657, November 2006.
- [2] P. Clote, E. Kranakis, D. Krizanc, and B. Salvy. Asymptotics of canonical and saturated RNA secondary structures. *J. Bioinform. Comput. Biol.*, 7(5):869–893, October 2009.
- [3] M. Drmota. Systems of functional equations. *Random Structures Algorithms*, 10(1-2):103–124, 1997.
- [4] M. Drmota, É. Fusy, J. Jué, M. Kang, and V. Kraus. Asymptotic study of subcritical graph classes, 2010. arXiv:1003.4699.
- [5] D. J. Evers and R. Giegerich. Reducing the conformation space in RNA structure prediction. In *German Conference on Bioinformatics (GCB'01)*, pages 1–6, 2001.
- [6] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

- [7] Christian Haslinger and Peter F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorial, and statistical properties. *Bulletin of Mathematical Biology*, 61(3):437–467, May 1999.
- [8] I. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431, 2003.
- [9] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998.
- [10] W. A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *J. Comput. Biol.*, 15(1):31–63, 2008.
- [11] E.A. Rodland. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *J Comput Biol*, 13(6):1197–1213, 2006.
- [12] P. R. Stein and M. S. Waterman. On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Mathematics*, 26:261–272, 1978.
- [13] G. Vernizzi, H. Orland, and A. Zee. Enumeration of RNA structures by matrix models. *Phys. Rev. Lett.*, 94(16):168103, April 2005.
- [14] B. Voss, R. Giegerich, and M. Rehmsmeier. Complete probabilistic analysis of RNA shapes. *BMC Biol.*, 4(5), 2006.
- [15] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Studies in Foundations and Combinatorics, Advances in Mathematics Supplementary Studies*, 1:167–212, 1978.
- [16] M. Zuker. RNA folding prediction: The continued need for interaction between biologists and mathematicians. *Lectures on Mathematics in the Life Sciences*, 17:87–124, 1986.
- [17] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, 2003.
- [18] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.