Counting RNA pseudoknotted structures

Cédric Saule, Mireille Regnier, Jean-Marc Steyaert, Alain Denise

Counting RNA pseudoknotted structures

Cédric Saule [*†] Mireille Régnier [†‡] Jean-Marc Steyaert [‡†]

Alain Denise[*§†¶]

May 19, 2010

Abstract

In 2004, Condon and coauthors gave a hierarchical classification of exact RNA structure prediction algorithms according to the generality of structure classes that they handle. We complete this classification by adding two recent prediction algorithms. More importantly, we precisely quantify the hierarchy by giving closed or asymptotic formulas for the theoretical number of structures of given size n in all the classes but one. This allows to assess the tradeoff between the expressiveness and the computational complexity of RNA structure prediction algorithms.

[*]LRI, Université Paris-Sud and CNRS. Bat 490, 91405 Orsay cedex, France

[†]INRIA Saclay, Parc Orsay Université, 4 rue Jacques Monod, 91893 Orsay cedex, France

[‡]LIX, Ecole Polytechnique and CNRS, 91128 Palaiseau cedex, France

[§]IGM, Université Paris-Sud and CNRS, bât. 400, 91405 Orsay cedex, France

[¶]Corresponding author

1 Introduction

The *ab initio* RNA structure prediction problem consists, given a RNA sequence, in finding a conformation that the molecule is likely to take in the cell. In [3], Condon and coauthors classified RNA structure prediction algorithms according to the inclusion relations between their *classes of structures*. The class of structures of a given algorithm is the set of structures that can be, in theory, returned by the algorithm. Condon *et al.* focused only on *exact* algorithms, that is algorithms that guarantee to give an optimal solution to the structure prediction problem, stated as an optimisation problem. They considered the class of pseudoknot-free structures [11, 23] (PKF) , and the following classes for pseudoknotted structures: Lyngsø and Pedersen (L&P) [9], Dirks and Pierce (D&P) [4], Akutsu and Uemura (A&U) [1, 18], and Rivas and Eddy (R&E) [14]. They notably proved the following inclusion relations: $PKF \subset L\&P \subset D\&P \subset A\&U \subset R\&E$. Since then, two other exact prediction algorithms have been developed, involving new classes: Reeder and Giegerich (R&G) [13] and Cao and Chen (C&C) [2] algorithms.

In this paper, we aim to quantify the tradeoff between the computational complexity and the expressiveness of all these algorithms. For this purpose, we compare them from the double point of view of their computational complexities and the cardinalities of their classes of structures, for a given size n. And we give closed or asymptotic formulas for the theoretical number of structures of given size n except for the class $R\&E$. More precisely, we establish that, except for the $L\&P$ class whose asymptotic formula is simpler, the number of structures of size n is, asymptotically, $\frac{\alpha}{2\sqrt{\pi}\, n^{3/2}}\omega^n$, where α and ω

are two constants which depend of the class. Table 5 summarizes our results.

[Table 1 about here.]

Additionally, we place the two new classes, R&G and C&C, in Condon *et al*'s hierarchy.

A number of works have been done on combinatorial enumeration of RNA structures without pseudoknots, see e.g. [22, 19, 5, 10, 8] or, more recently, with pseudoknots, as in [20, 15, 6, 7] for instance. Our purpose is different, as our classes of structures are not defined *per se*, but correspond to given exact prediction algorithms.

The paper is organised as follows. In Section 2, we give some notation and definitions. In Section 3, we present a bijection between the L&P class and a class of combinatorial planar maps, leading to a closed formula for the L&P class. In Section 4, we establish that each of the classes D&P, A&U, R&G, C&C, and L&P can be encoded by a context-free language. For each of them, we derive an equation for the generating function, leading to an asymptotic formula for the number of structures of size n. In Section 5, we conclude by giving some remarks on the expressiveness of the structure prediction algorithms compared to their complexity.

2 Definitions and notation.

An RNA secondary structure (possibly with pseudoknots) is given by a sequence of integers $(1, 2, \ldots, n)$ and a list of pairs (i, j), called *basepairs* or *arcs*, where $i < j$ and each

number in $\{1, 2, \ldots, n\}$ appears exactly in one pair. Such a structure can be represented as in Figure 1, where each basepair (i, j) is represented by an edge between i and j. In real RNA structures there are unpaired bases, but we do not consider them.

Definition 1 (Crossing arcs)**.** *Let (i, j) and (k, l) two arcs such that $i < k$. We say that (i, j) and (k, l) are* crossing *if $i < k < j < l$.*

Definition 2 (Crossing graph)**.** *The* crossing graph *of an RNA structure is a graph G defined as follows: the vertices of G are the arcs of the structure, and two vertices are connected by an edge if and only if their two corresponding arcs are crossing.*

Definition 3 (Pseudoknot)**.** *A* pseudoknot *is a set of arcs that is not a singleton and that corresponds to a maximal connected component in the crossing graph.*

Definition 4 (Simple pseudoknot [1])**.** *A pseudoknot P is* simple *if there exist two numbers j_1 and j_2, with $j_1 < j_2$, such that:*

- *each arc (i, j) in P satisfies either $i < j_1 < j \leq j_2$ or $j_1 \leq i < j_2 < j$,*
- *and if two arcs (i, j) and (i', j') satisfy $i < i' < j_1$ or $j_1 \leq i < i'$, then $j > j'$.*

The first property ensures that, for each arc of P, one of its ends exactly is between j_1 and j_2. And the arcs are divided in two sets: those having their other end smaller than j_1, and those having their other end greater than j_2. We call these two sets, respectively, the *left part* and the *right part* of the pseudoknot. The second property of the definition ensures that two arcs in the same set cannot intersect each other. Figure 1 shows a simple pseudoknot.

[Figure 1 about here.]

Definition 5 (H-type Pseudoknot)**.** *A* H-type pseudoknot *is a simple pseudoknot having the following additional property: each arc in one of the two above sets crosses all the arcs of the other set.*

3 A bijection between the L&P structures and a class of planar maps.

The Lyngsø-Pedersen (L&P) class is the simplest class of pseudoknotted structures. According to [9] and [3], a structure is in the L&P class if and only if it contains either no pseudoknot or a unique H-type pseudoknot, and this pseudoknot is not embedded under any arc. (Fig. 2).

[Figure 2 about here.]

Between any two consecutive ends of the arcs of the pseudoknots, there can be a nested structure. Theorem 1, and its straightforward Corollary 1, give the closed formula and the asymptotic formula for the number of such structures, respectively.

Theorem 1. *The number of* L&P structures with n arcs is:

$$LP(n) = \frac{1}{2} \cdot 4^n - \binom{2n+1}{n} + \binom{2n-1}{n-1} + \frac{1}{n+1}\binom{2n}{n}.$$

Corollary 1.

$$LP(n) \sim \frac{1}{2} \cdot 4^n.$$

Proof of Theorem 1. The proof is bijective: we establish a bijection between the set of L&P structures of any size n and the set of rooted isthmusless planar maps with n edges and one or two vertices. The first three terms of the formula count the number of such maps with two vertices [16, 21], while the last term, a Catalan number, counts the number of such maps with one vertex [17]. Hence the theorem.

A *planar map* is a proper embedding of a connected planar graph. It is said *isthmusless* if the deletion of any edge does not split the graph. A *rooted* planar map is a planar map where a vertex and an edge adjacent to it are distinguished.

A permutation of a given finite set of integer numbers is a bijection from this set to itself. A permutation σ can be represented by its set of *cycles*, that is the cycles of numbers $(n_1, n_2, \ldots, n_k)$ such that $\sigma(n_i) = n_{i+1}$ for any i between 1 and $k-1$, and $\sigma(n_k) = n_1$.

Any planar map with n edges can be represented by two permutations σ and τ on $\{+1, -1, +2, -2, \ldots, +(n-1), -(n-1), +n, -n\}$, in the following way: the edges of the map are numbered from 1 to n. Then, for any edge i, one labels its extremities with $+i$ and $-i$, respectively. By convention, the root edge is labelled with $+1$ and -1, in such a way that -1 labels the extremity adjacent to the root vertex. Now, the two permutations are as follows:

- the permutation σ is an involution without fixed points that represents the edges of the map. Each cycle of σ is of size two and contains both ends of one edge: $\sigma = (+1, -1), (+2, -2), \ldots, (+n, -n)$.

- the permutation τ has as many cycles as vertices in the map. Each cycle is given by the sequence of labellings around the corresponding vertex, clockwise.

Figure 3 shows a planar map and two permutations that represent it. By convention, the drawing is such that the root edge separates the external face from an internal face.

[Figure 3 about here.]

Let us consider a L&P structure S with n edges, and let us label the left extremities of its arcs with $+1, +2, \ldots, +n$ from left to right, and give to each right foot the label $-i$ if the corresponding left foot has label $+i$. Let $w = [w_1, w_2, \ldots w_{2n}]$ be the sequence of labels of S, from left to right. From any w we can now construct two permutations σ and τ that represent an isthmussless rooted planar map with one or two vertices. Regarding σ, we just set $\sigma = (+1, -1) \ldots (+n, -n)$.

Let us first consider the simple case where there is no crossing in the structure. It is known for a long time that such nested structures are counted by Catalan numbers. This can be established, for example by a folkloric bijection with planar maps having one vertex, by setting σ as above, and $\tau = (w)$. See Figure 4 for an illustration.

[Figure 4 about here.]

Now suppose that there is a pseudoknot in the structure, and let us present a bijection between the set of such structures and the set of rooted ithmusless planar maps with two vertices. Start from w. Since τ must have two cycles, we have to split w in two parts that will be the two cycles. Let us define the left set (resp. the right set) of arcs of

the pseudoknot, respectively, as the set of arcs whose left (resp. right) extremities are in the left (resp. right) part of the pseudoknot, where left and right parts are defined as in Section 2. There are two cases:

Case 1. There is only one arc in the right set. In this case, let ℓ be the position of the first right extremity of an arc in the left set. We cut w between positions $\ell - 1$ and ℓ. Each part corresponds to a cycle of τ: $\tau = (w_1, \ldots, w_{\ell-1})(w_\ell, \ldots, w_{2n})$. See Figure 5 for an illustration.

[Figure 5 about here.]

Case 2. There are at least two arcs in the right set. We cut w just before the first right extremity of an arc in the right set. See Figure 6.

[Figure 6 about here.]

Let us show that, in both cases, the resulting map is planar and isthmusless. At first, remark that if the map is not planar or has an isthmus, necessarily it comes from arcs that are involved in the pseudoknot. Indeed, by construction, non crossing arcs in the structure give non crossing loops in the map. So, without loss of generality, we can consider only structures where all the arcs are involved in the pseudoknot. Consider such a structure with n arcs. In the case 1, we have $w = [+1, +2, \ldots, +(n-1), +n, -(n-1), -(n-2), \ldots, -1, -n]$, hence $\tau = (+1, +2, \ldots, +n)(-(n-1), -(n-2), \ldots, -1, -n)$. Clearly, this gives a planar map, since the two cycles of τ are in opposite order. And there is no isthmus because all edges go from one vertex to the other. In the case 2, we have

$w = [+1, +2, \ldots, +(\ell-1), +\ell, +(\ell+1), \ldots, +n, -(\ell-1), \ldots, -2, -1, -n, -(n-1), \ldots, -\ell]$, hence $\tau = (+1, +2, \ldots, +(\ell-1), +\ell, +(\ell+1), \ldots, +n, -(\ell-1), \ldots, -2, -1)(-n, -(n-1), \ldots, -\ell)$. Again, this gives a planar map: edges $1, 2, \ldots, \ell-1$ are nested loops, and edges $\ell, \ldots, n$ go from one vertex to the other, without any crossing. And there is no isthmus because the number of edges going from one vertex to the other, $n-\ell+1$, is greater or equal to 2.

Now let us present the converse transformation. Consider an isthmusless rooted planar map with two vertices, given by $\sigma = (+1, -1), (+2, -2), \ldots, (+n, -n)$ and τ having two cycles. We aim to construct the sequence w that represents the corresponding pseudoknotted structure. Let us consider the cycle of τ which contains 1, and write it in such a way that it begins with 1. Let us call u this sequence of labels. This gives the first part of the sequence w. We are now searching for the second part of w, that is the sequence v such that $uv = w$. For that purpose, consider the set of *isolated labels*, that is the labels in u that have not their opposite label in u. We have the two following cases:

Case 1. There is no pair $(+i, -i)$ in u such that the isolated labels are located between $+i$ and $-i$. Let $+j$ the penultimate isolated label in u. Write the second cycle of τ in such a way that it begins with $-j$. This gives v, and there is exactly one edge in the second part of the pseudoknot.

Case 2. There is a pair of labels $(+i, -i)$ in u such that all isolated labels are located between $+i$ and $-i$. Let $+j$ the last isolated label in u. Write the second cycle of τ in such a way that it begins with $-j$. This gives v. In this case, there are at least two

edges in the second part of the pseudoknot.

□

4 Asymptotic enumeration of pseudoknotted structures.

4.1 A context-free encoding for simple and H-type pseudoknots

As will be seen farther, all the classes that are involved in exact prediction algorithms but one involve either H-type pseudoknots or simple pseudoknots. The only exception is the R&E class. Here we define a transformation that allow to encode any class of pseudoknotted structures where all pseudoknots are simple by a context-free language.

Let us first recall some definitions. Let L be a language on a given alphabet A, and $w = w_1 w_2 \dots w_n$ a word of L, where the w_i's are the letters of w. A word v is a *subword* of w if $v = w_{i_1} w_{i_2} \dots w_{i_k}$, where $1 \leq i_1 < i_2 < \dots < i_k \leq n$. The *projection* of w onto an alphabet $A' \in A$ is the subword w' obtained by erasing in w all letters that do not belong to A'. The projection of L onto A' is the set of projections of the words of L onto A'. Finally, let us recall that the Dyck language on any two-letter alphabet $\{d, \bar{d}\}$ is the language of balanced parentheses strings, where d and $\bar{d}$ stand, respectively, for opening and closing parentheses. Now we can state twe two following straightforward lemmas:

Lemma 1. *Any class of pseudoknotted structures where all pseudoknots are simple can be encoded by the words of a language L on the alphabet $\{d, \bar{d}, x, \bar{x}, y, \bar{y}\}$ where*

- *d and $\bar{d}$ encode, respectively, the left and right ends of arcs that are not involved in pseudoknots;*
- *x and $\bar{x}$ encode, respectively, the left and right ends of arcs that are involved in the left parts of pseudoknots;*
- *y and $\bar{y}$ encode, respectively, the left and right ends of arcs that are involved in the right parts of pseudoknots.*

Additionally, the projection of the language to the alphabet $\{d, \bar{d}\}$ (resp. $\{x, \bar{x}\}$, $\{y, \bar{y}\}$) is a sublanguage of the Dyck language on the same alphabet.

Lemma 2. *Let S be a pseudoknotted structure, and w be the word on $\{d, \bar{d}, x, \bar{x}, y, \bar{y}\}$ that encodes S. Then every simple pseudoknot in S is encoded by a subword v of w, such that*

$$v = x^n\, y^{m_1}\, \bar{x}^{n_1}\, y^{m_2}\, \bar{x}^{n_2}\ \ldots\ y^{m_k}\, \bar{x}^{n_k}\, \bar{y}^m\,,$$

where $n_1 + n_2 + \ldots + n_k = n$ and $m_1 + m_2 + \ldots + m_k = m$.

Remark that a H-type pseudoknot is a simple pseudoknot where $k = 1$. Thus every H-type pseudoknot in S is encoded by a subword $v = x^n\, y^m\, \bar{x}^n\, \bar{y}^m$. Finally, the following Proposition gives a way to encode any pseudoknotted structure where all pseudoknots are simple by a subset of the Dyck language with three kinds of pairs of parentheses, that is on the alphabet $\{d, \bar{d}, x, \bar{x}, y, \bar{y}\}$.

Proposition 1. *Let S be a pseudoknotted structure, and w be the word on $\{d, \bar{d}, x, \bar{x}, y, \bar{y}\}$ that encodes S. Then w can be encoded by a word where every subword $v = x^n\, y^{m_1}\, \bar{x}^{n_1}\, y^{m_2}\, \bar{x}^{n_2}\ \ldots\ y^{m_k}\, \bar{x}^{n_k}\, \bar{y}^m\,$, corresponding to a H-type pseudoknot is replaced with $v' = x^n\, y^{m_1} \bar{y}^{m_1}\, \bar{x}^{n_1}\, y^{m_2} \bar{y}^{m_2}\, \bar{x}^{n_2}\ \ldots\ y^{m_k} \bar{y}^{m_k}\, \bar{x}^{n_k}$.*

In particular , every subword $v = x^n\, y^m\, \bar{x}^n\, \bar{y}^m$ corresponding to a simple pseudoknot is replaced with $v' = x^n\, y^m\, \bar{y}^m\, \bar{x}^n$.

Proof. The proof is straightforward, as there is an immediate one-to-one correspondance between the two kinds of words below. The transformation is illustrated in Figure 7(a) and Figure 7(b), respectively, for simple pseudoknots and for the particular case of H-type pseudoknots. □

[Figure 7 about here.]

4.2 Asymptotic results.

For each of the D&P, A&U, R&G, and C&C classes, we give an asymptotic equivalent for the number of structures of size n. In each case, the proof is in three steps:

1. We design an unambiguous context-free grammar which generates the language that encodes the considered structures, according to Proposition 1.

2. From the grammar, we deduce an algebraic equation satisfied by the ordinary generating function (o.g.f.) of the language.

3. From this equation, we compute an asymptotic formula for the number of structures of size n.

For any class $X\&Y$, we write $X\&Y(n)$ for its number of structures having n arcs.

4.2.1 The Akutsu-Uemura class (A&U).

Following [1, 3], the A&U structures are composed of non crossing edges and of any number of simple pseudoknots. As these pseudoknot can embed other substructures which can be pseudoknotted in turn, they are said to be *recursive* [1].

Theorem 2.

$$A\&U(n) \;=\; \frac{\alpha_1}{2\sqrt{\pi}}{\omega_1}^n n^{-3/2}(1+O(1/n)),$$

where $\alpha_1 = 0.6575407644...,\ \omega_1 = 7.547308334...$, *are algebraic constants.*

Proof. Let $L_{A\&U}$ be the language that encodes the A&U class, according to Proposition 1. The following unambiguous context-free grammar generates $L_{A\&U}$:

$$\begin{aligned} S &\rightarrow dS\bar{d}S|P \\ P &\rightarrow xSX\bar{x}S|\epsilon \\ X &\rightarrow xSX\bar{x}SY|yYS\bar{y}S \\ Y &\rightarrow ySY\bar{y}S|\epsilon \end{aligned}$$

The two rules in the first line allow to generate non crossing arcs and to place pseudoknots anywhere. The other rules generate words which correspond to the code for a simple pseudoknot as showed in Figure 8.

[Figure 8 about here.]

Given the grammar, we obtain the set of recursive equations for the o.g.f. of the various sets defined in the 1-to-1 encoding. Letting the formal symbol z denote an arc,

we thus have through a straightforward translation:

$$\begin{aligned} S(z) &= zS^2(z) + P(z) \\ P(z) &= zS^2(z)X(z) + 1 \\ X(z) &= zS^2(z)Y(z)(X(z) + 1) \\ Y(z) &= zS^2(z)Y(z) + 1. \end{aligned}$$

By iterated bottom-up substitutions, we ultimately get that the o.g.f. $S(z)$ is solution of the algebraic equation

$$F(z, S) = z^2S^4 - 2zS^3 + zS^2 + S - 1 = 0, \tag{1}$$

from which we can derive the number of structures of size n.

For this proof we present in some details the main steps of the computations that have to be performed in order to get the asymptotics for an o.g.f. given by the algebraic implicit equation $F(z, S) = 0$ satisfied by the o.g.f. $S(z)$. The foundations can be found in [**?**] (Chapter 7).

Since $\partial F/\partial z|_{z=0,S=1} = 1$ is defined and $\partial F/\partial S|_{z=0,S=1} = 1$ is non vanishing, $z = 0$ is not a singular point for S; by the implicit function theorem, $S(z)$ exists as a regular function in a circular neighborhood of $z = 0$ where $\partial F/\partial S$ is non-zero. The degree in S of this bivariate equation being 4, and the coefficient $a(z)$ of S^4 satisfying $a(0) = a'(0) = 0$, this bivariate equation defines two folds $z = \zeta(S)$.

The radius of convergence ρ_1 of the o.g.f. $S(z)$ is thus a solution of the system

$\{F(z,S)=0, \partial F/\partial S(z,S)=0\}$. At such a point the local holomorphic solution $z=\zeta(S)$ is no longer invertible, which implies that this point is a singular point for the o.g.f. $S(z)$.

Let $(z=\rho_1, S=\sigma_1)$ be the point of the Riemann surface of the solution located on the fold issued from $(z=0, S=1)$, that satisfies $\partial F/\partial S=0$ and that has the smallest modulus. This point is unique and located on the positive real axis, since the o.g.f. is indeed a function of z with all coefficients being positive. Since the first derivative, $\frac{dz}{dS}=-\frac{\partial F/\partial S}{\partial F/\partial z}$ vanishes at $(z=\rho_1, S=\sigma_1)$ and the second derivative $\frac{d^2z}{dS^2}=-\frac{\partial^2 F/\partial S^2}{\partial F/\partial z}$ is strictly positive, $(z-\rho_1)^{1/2}$ is well defined in a neighborhood of $S=\sigma_1$. At this point, the local expansion of z with respect to S writes:

$$z=\rho_1+\frac{1}{2}\frac{d^2z}{dS^2}(S-\sigma_1)^2+\frac{1}{3!}\frac{d^3z}{dS^3}(S-\sigma_1)^3+\dots, \tag{2}$$

and we get the Taylor expansion at $S=\sigma_1$:

$$\sqrt{1-z/\rho_1}=\beta_1(S-\sigma_1)+\beta_2(S-\sigma_1)^2+\dots, \tag{3}$$

with $\beta_1=-\sqrt{\frac{1}{2}\frac{\partial^2 F/\partial S^2}{\partial F/\partial z}}$. This equation can now be inverted locally which yields:

$$S=\sigma_1-\sqrt{\frac{2\rho_1\partial F/\partial z|_{z=\rho_1,S=\sigma_1}}{\partial^2 F/\partial S^2|_{z=\rho_1,S=\sigma_1}}}\sqrt{1-z/\rho_1}+O(1-z/\rho_1)\ . \tag{4}$$

This expansion can be calculated at any order, so that we obtain for the coefficients $A\&U(n)$ an infinite asymptotic development. The dominant term is given by the first square root in the previous expansion. Since it is well-known that $[z^n]\sqrt{1-z/\rho}=\frac{1}{2\sqrt{\pi}}\rho^{-n}n^{-3/2}(1+O(1/n))$:

$$[z^n]S(z)=\sqrt{\frac{2\rho_1\partial F/\partial z|_{z=\rho_1,S=\sigma_1}}{\partial^2 F/\partial S^2|_{z=\rho_1,S=\sigma_1}}}\frac{1}{2\sqrt{\pi}}\rho_1^{-n}n^{-3/2}(1+O(1/n)). \tag{5}$$

We thus get the general form of the solution, as stated in the theorem, with $\alpha_1 = \sqrt{\frac{2\rho_1 \partial F/\partial z|_{z=\rho_1,S=\sigma_1}}{\partial^2 F/\partial S^2|_{z=\rho_1,S=\sigma_1}}}$ and $\omega_1 = 1/\rho_1$. In order to get the values for the constants in the expansions and for the radius of convergence, we used Maple. From Equation 1, we compute the partial derivatives $\partial F/\partial z = 2zS^4 - 2S^3 + S^2$ and $\partial F/\partial S = 4 * z^2 * S^3 - 6 * z * S^2 + 2 * z * S + 1$. The system is too complex to be solved formally; so we lower the degree in S by considering the combination $R = 4F - S\partial F/\partial S = -2zS^3 + 2zS^2 + 3S - 4$ which has to vanish at the points where F and $\partial F/\partial S$ do. Since R is of degree 1 in z, it is easy to get an expression for z that we substitute into $\partial F/\partial S$, obtaining that $8S^3 - 31S^2 + 42S - 20$ should equivalently be zero. Hence we obtain 3 possible algebraic roots, one being real σ_1 and the other two conjugate complex numbers. Only $\sigma_1 = 1.403556586...$ and the associated real value of z for which $F(z, S) = 0$ — $\rho_1 = 0.1324975681...$ — are of interest. A direct approximate solution using the floating point solver of Maple confirms this situation and a more involved study or the Riemann surface also yields $\rho_1 = 0.1324975681...$ to be the radius of convergence of the series. Further computations provide all the constants encountered in the proof and stated in the theorem. □

4.2.2 The Dirks and Pierce class (D&P).

Structures of D&P class are characterized by the presence of non crossing edges and any number of H-type pseudoknots [4, 3].

Theorem 3.

$$D\&P(n) = \frac{\alpha_2}{2\sqrt{\pi}} {\omega_2}^n n^{-3/2}(1 + O(1/n)),$$

where $\alpha_2 = 0.7534777262...,\ \omega_2 = 7.3148684640...$, *are algebraic constants.*

Proof. The following unambiguous grammar generates the language that encodes the D&P structures, according to Proposition 1:

$$\begin{aligned} S &\rightarrow dS\bar{d}S|P \\ P &\rightarrow xXS\bar{x}S|\epsilon \\ X &\rightarrow xSX\bar{x}S|ySY\bar{y}S \\ Y &\rightarrow ySY\bar{y}S|\epsilon \end{aligned}$$

The first line allows to generate structures without pseudoknots and to place pseudoknots, by symbol P, anywhere in the sequence. The last three lines generate words which correspond to the code for H-Type pseudoknot. P generates the first arc of the left set. Other arcs in the left set can be generated by X. The symbol Y generates arcs of the right part.

From this grammar, we get the following algebraic equation:

$$F(z,S) = z^3S^6 - z^2S^5 + 2zS^3 - zS^2 - S + 1 = 0 \tag{6}$$

which is very similar to the equation satisfied by the o.g.f. for the $A\&U$ family. We solve it in the same way, and find out the dominant singularity in $z = \rho_2 = 0.1367078581...,\ S = \sigma_2 = 1.439796009...$, with the same local behaviour, implying similar asymptotics for the coefficients. The only problem encountered in finding this dominant singularity comes from the fact that there exists another singularity closer to the origin in $z = \mu =$

$0.08794976637...,\ S = \tau = 7.169944393...$, but which is not on the same fold of the Riemann surface and which therefore does not have to be taken into consideration. □

4.2.3 The Reeder ang Giegerich class (R&G).

The R&G class which corresponds to the structures handled by Reeder and Giegerich's algorithms [13]. It has a $\mathcal{O}(n^4)$ time complexity.

Theorem 4.

$$R\&G(n) \ = \ \frac{\alpha_3}{2\sqrt{\pi}}{\omega_3}^n n^{-3/2}(1+O(1/n)),$$

where $\alpha_3 = 1.165192913...,\ \omega_3 = 6.576040092...$, *are algebraic constants.*

Proof. In [13], the following grammar is given (we removed the unpaired bases):

$$S \rightarrow SS|dSd\bar{}|x^kSy^lS\bar{x}^kS\bar{y}^l|\epsilon.$$

This grammar is not context-free. However, we remark that the pseudoknot defined here is a particular case of a H-Type pseudoknot. So by applying Proposition 1 again, we define the following context free grammar :

$$\begin{aligned} S &\rightarrow dS\bar{d}S|P \\ P &\rightarrow xX\bar{x}S|\epsilon \\ X &\rightarrow xX\bar{x}|SyY\bar{y}S \\ Y &\rightarrow yY\bar{y}|S \end{aligned}$$

The related algebraic equation

$$F(z,S) = z^2S^4 + z(z-1)^2S^2 - (z-1)^2S + (z-1)^2 = 0 \qquad (7)$$

is again very similar to the equation satisfied by the o.g.f. for the $A\&U$ family. We solve it in the same way, and find out the dominant singularity in $z = \rho_3 = 0.1520671994...$, $S = \sigma_3 = 1.589450164...$, with the same local behaviour, implying similar asymptotics for the coefficients. □

Additionally, the following theorem places this new class into Condon *et al.*'s classification.

Theorem 5. *$R\&G \subset D\&P$, $L\&P \cap R\&G \neq \emptyset$ and $R\&G \not\subset L\&P$*

Proof. The grammar wich describes the pseudoknots in R&G is less general than the grammar for H-type pseudoknots. So $R\&G \subset D\&P$ and $L\&P \cap R\&G \neq \emptyset$. As R&G structures can contain several pseudoknots, we have $L\&P \cap R\&G \neq L\&P$. □

4.2.4 The Cao and Chen class (C&C).

The $C\&C$ class corresponds to the structures handled by Cao and Chen's algorithm [2], whose complexity is $\mathcal{O}(()n^6)$.

Theorem 6.

$$C\&C(n) \;=\; \frac{\alpha_4}{2\sqrt{\pi}}{\omega_4}^n n^{-3/2}(1 + O(1/n)),$$

where $\alpha_4 = 1.665071176...$, $\omega_4 = 5.856765093...$, are algebraic constants.

Proof. The following non context-free grammar generates the C&C structures:

$$S \rightarrow SS|dS\bar{d}|x^k Sy^l \bar{x}^k S\bar{y}^l|\epsilon.$$

It can be translated into a context-free grammar which is a restriction of the R&G grammar:

$$\begin{aligned} S &\rightarrow dS\bar{d}S|P \\ P &\rightarrow xX\bar{x}S|\epsilon \\ X &\rightarrow xX\bar{x}|SyY\bar{y}S \\ Y &\rightarrow yY\bar{y}|\epsilon \end{aligned}$$

Now the following algebraic holds for the o.g.f of C&C structures:

$$F(z,S) = z^2S^3 + z(z-1)^2S^2 - (z-1)^2S + (z-1)^2 = 0. \quad (8)$$

Again, it is very similar to the equation satisfied by the o.g.f. for the $A\&U$ class. We solve it in the same way, and find out the dominant singularity in $z = \rho_4 = 0.1707427197...$, $S = \sigma_4 = 1.7663614360...$, with the same local behaviour, implying similar asymptotics for the coefficients. □

Additionally, we easily state that

Theorem 7. *$C\&C \subset D\&P$, $L\&P \cap C\&C \neq 0$, $C\&C \not\subset L\&P$ and $C\&C \subset R\&G$*

4.2.5 The Lyngsø and Pedersen class (L&P).

We already gave a closed formula and an asymptotic equivalent for this class in Section 3. We briefly outline below another way to prove Theorem 1: we prove that any L&P structure can be encoded by a word of a non ambiguous context-free language.

Further standard computations lead to the generating function, then to the closed formula.

Theorem 8. *The number of* L&P structures of size n, L&P (n) satisfies the following asymptotics formula when n tends to infinity :

$$L\&P(n) \;=\; \frac{1}{2}\,4^n(1+O(n^{-1/2})).$$

Proof. Any L&P structure of size n can be encoded by a word of length n of the context-free language generated by the following nonambiguous grammar:

$$\begin{aligned}
S &\rightarrow dD\bar{d}S|P\\
D &\rightarrow dD\bar{d}D|\epsilon\\
P &\rightarrow xDX\bar{x}D|\epsilon\\
X &\rightarrow xDX\bar{x}D|yDY\bar{y}D\\
Y &\rightarrow yDY\bar{y}D|\epsilon
\end{aligned}$$

The system of equations which the o.g.f. $S(z) = \Sigma_n L\&P(n)z^n$ satisfies, where n is the number of base pairs in contact deduces from the grammar:

$$\begin{aligned} S(z) &= zS(z)D(z) + P(z) \\ D(z) &= zD(z)^2(z) + 1 \\ P(z) &= zD^2(z)X(z) + 1 \\ X(z) &= zD^2(z)(X(z) + Y(z)) \\ Y(z) &= zY(z)D^2(z) + 1 \end{aligned}$$

The series $D(z)$ is readily identified to be the o.g.f. for the Dyck language: $D(z) = (1 - \sqrt{1-4z})/2z$. Contrarily to what we encountered previously this system can now be solved explicitely, since all the other equations are linear and the system is clearly trigonal; so we get successively $Y(z),\ X(z),\ P(z)$ and $S(z)$, using repeatedly the fact that $zD^2(z) = D(z) - 1$. Ultimately we find:

$$S(z) = \frac{8(1 - 5z + 5z^2 + (3z-1)\sqrt{1-4z})}{(1 - 4z - \sqrt{1-4z})^2(1 + \sqrt{1-4z})}.$$

The denominator vanishes for $z = 0$ and $z = 1/4$, but $S(z)$ is not singular at the origin, since it has a Taylor development: $S(z) = 1 + z + 3z^2 + 12z^3 + 51z^4 + 218z^5 + 926z^6 + 3902z^7 + O(z^8)$. Hence $S(z)$ has its dominant singularity in $z = \rho_5 = 1/4$ where it admits the following expansion in $\sqrt{1-4z}$:

$$S(z) \ = \ \frac{1}{2}\frac{1}{1-4z} - \frac{3}{2}\frac{1}{\sqrt{1-4z}} + 8 - 4\sqrt{1-4z} + O(1-4z).$$

Consequently, the coefficients of $S(z)$ have the following asymptotic expansion:

$$L\&P(n) = \frac{1}{2}\,4^n - \frac{3}{2\sqrt{\pi}}\,4^n n^{-1/2} + \frac{4}{2\sqrt{\pi}} 4^n n^{-3/2}(1 + O(1/n)).$$

□

5 Conclusion

We proved that most classes of pseudoknotted structures that can be predicted by exact algorithms (all but R&E for which the problem remains open) can be encoded by context-free languages. We extended Condon *et al.*'s hierarchy by adding two more classes, and we computed closed or asymptotic formulas for the cardinality of all classes but one.

These results, summarized in Table 5, allow us to quantify the relationship between the complexity of each algorithm and the generality of the class that it can handle.

Notably, from a strict quantitative point of view, the growth of complexity by a factor n^2 between the PKF and L&P classes seems not to be justified compared to the very small increase in cardinality.

At a first glance, the situation seems to be even worse for the C&C class, whose related algorithm has a stronger complexity than the R&G one, while $C\&C \subset R\&G$ and the ratio of their cardinalities is exponential. However, the C&C algorithm computes the partition function with an elaborated thermodynamic model, and the R&G algorithm does not.

On the other hand, A&U and D&P have the same complexity whereas the A&U class is exponentially larger than the D&P one. But D&P computes the partition function and the increase of cardinality of A&U does not allow to find known biological structures in that class [12].

Finally, the linear increasing between PKF andR&G complexities seems very reasonable compared to the exponential increase of the cardinality.

Acknowledgements

This research was supported in part by the ANR project BRASERO ANR-06-BLAN-0045, and by the Digiteo project "RNAomics".

References

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104:45–62, 2000.

[2] S. Cao and S-J Chen. Predicting structured and stabilities for h-type pseudoknots with interhelix loop. *RNA*, 15:696–706, 2009.

[3] A. Condon, B. Davy, B. Rastegari, S. Zhao, and F. Tarrant. Classifying RNA pseudknotted structures. *Theorical computer science*, 320:35–50, 2004.

[4] N.A. Dirks, R.M. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J Comput Chem*, 24:1664–1677, 2003.

[5] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math*, 89, 1996.

[6] F. W. D. Huang and M. Reidys. Statistics of canonical RNA pseudoknot structures. *Journal of Theoretical Biology*, 253(3):570–578, 2008.

[7] E. Y. Jin and C. M. Reidys. RNA pseudoknot structures with arc-length ≥ 3 and stack-length $\geq \sigma$. *Discrete Appl. Math.*, 158(1):25–36, 2010.

[8] W.A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *Journal of Computational Biology*, 15(1):31–63, Jan–Feb 2008.

[9] R. B. Lyngsø and Pedersen C. N. RNA pseudoknot prediction in energy based models. *Journal of computational biology*, 7:409–428, 2000.

[10] M. E. Nebel. Combinatorial properties of RNA secondary structures. *Journal of Computational Biology*, 9(3):541–574, 2003.

[11] R. Nussinov, G. Pieczenik, J. R. Griggs, and Kleitman D. J. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35:68–82, 1978.

[12] B. Condon A. Rastegari. Parsing nucleic acid pseudoknotted secondary structure: algorithm and applications. *Journal of computational biology*, 14:16–32, 2007.

[13] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:104, 2004.

[14] E. Rivas and S. R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285:2053–2068, 1999.

[15] E. A. Rødland. Pseudoknots in RNA secondary structures: representation, enumeration, and prevalence. *Journal of Computational Biology*, 13(6):1197–1213, 2006.

[16] N. J. A. Sloane and Simon Plouffe. *The Encyclopedia of Integer Sequences*. Academic Press, 1995.

[17] W.T. Tutte. A census of planar maps. *Canadian Journal of Mathematics*, 15:249–271, 1963.

[18] Y. Uemura, A. Hasegawa, S. Kobayashi, and T. Yokomori. Tree adjoining grammars for RNA structures prediction. *Theorical computer science*, 210:277–303, 1999.

[19] M. Vauchaussade de Chaumont and X.G. Viennot. Enumeration of RNA's secondary structures by complexity. In V. Capasso, E. Grosso, and S.L. Paven-Fontana, editors, *Mathematics in Medecine and Biology*, volume 57 of *Lecture Notes in Biomathematics*, pages 360–365, 1985.

[20] G. Vernizzi, H. Orland, and A. Zee. Enumeration of RNA structures by matrix models. *Phys. Rev. Lett.*, 94:168103, 2005.

[21] T. R. S. Walsh and A. B. Lehman. Counting rooted maps by genus. iii: Nonseparable maps. *J. Combinatorial Theory Ser. B*, 18:222–259, 1975.

[22] M. S. Waterman. Secondary structure of single-stranded nucleic acids. *Advances in Mathematics Supplementary Studies*, 1(1):167–212, 1978.

[23] M. Zucker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acid Research*, 9:133–148, 1981.

List of Figures

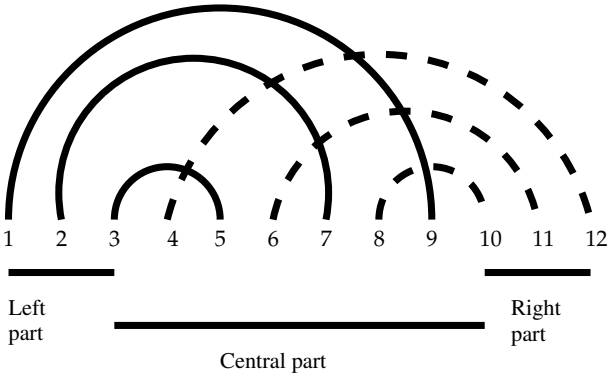


Figure 1: A pseudoknot given by the sequence $(1, 2, \ldots, 12)$ and the arcs $(1, 9)$, $(2, 7)$, $(3, 5)$, $(4, 12)$, $(6, 11)$, $(8, 10)$. This pseudoknot is simple, with $j_1 = 4$ and $j_2 = 9$.

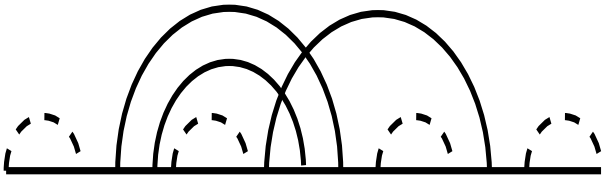
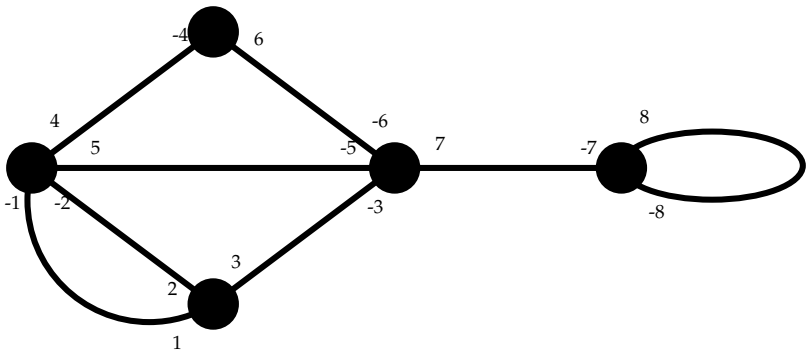


Figure 2: A structure from the L&P class.



$\sigma = (1,-1)(2,-2)(3,-3)(4,-4)(5,-5)(6,-6)(7,-7)(8,-8)$

$\tau = (1,2,3)(-1,4,5,-2)(-4,6)(-6,7,-3,-5)(-7,8,-8)$

Figure 3: A planar map and its two associated permutations σ and τ.

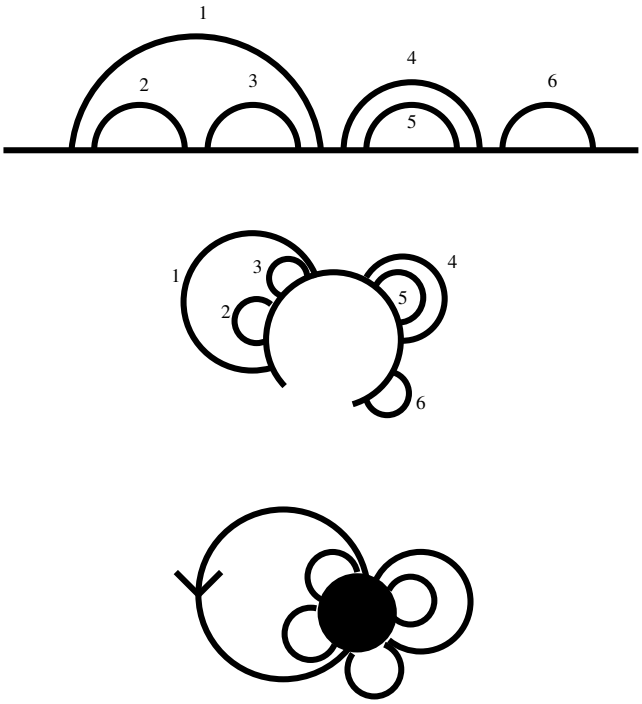


Figure 4: An illustration of the straightforward bijection between nested structures and planar maps with one vertex.

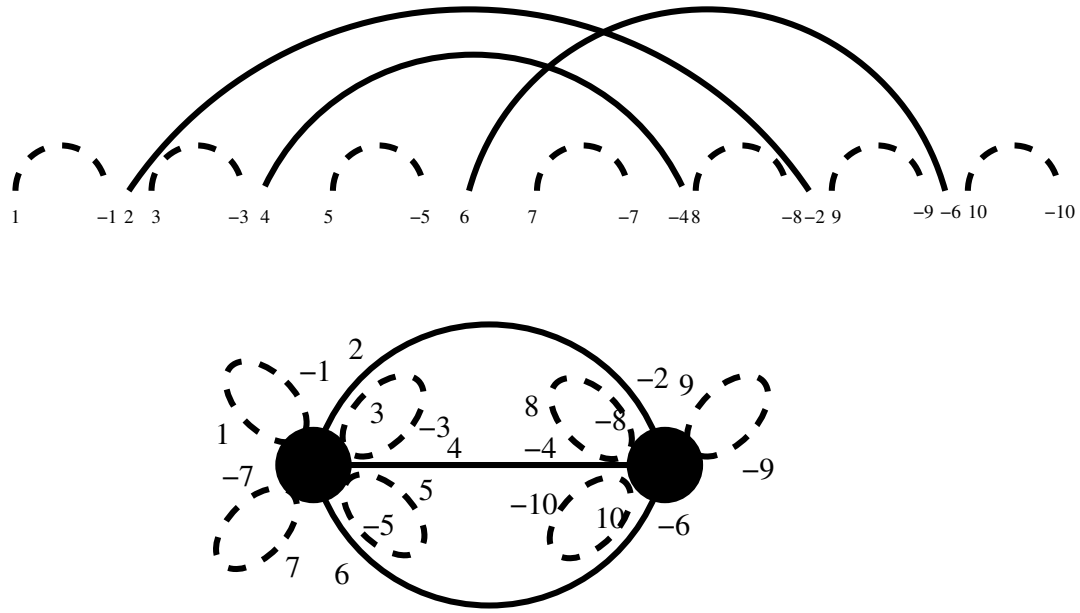


Figure 5: Top, a L&P structure corresponding to case 1. Bottom, the corresponding planar map. Arcs not involved in the pseudoknot are drawn in dotted lines.

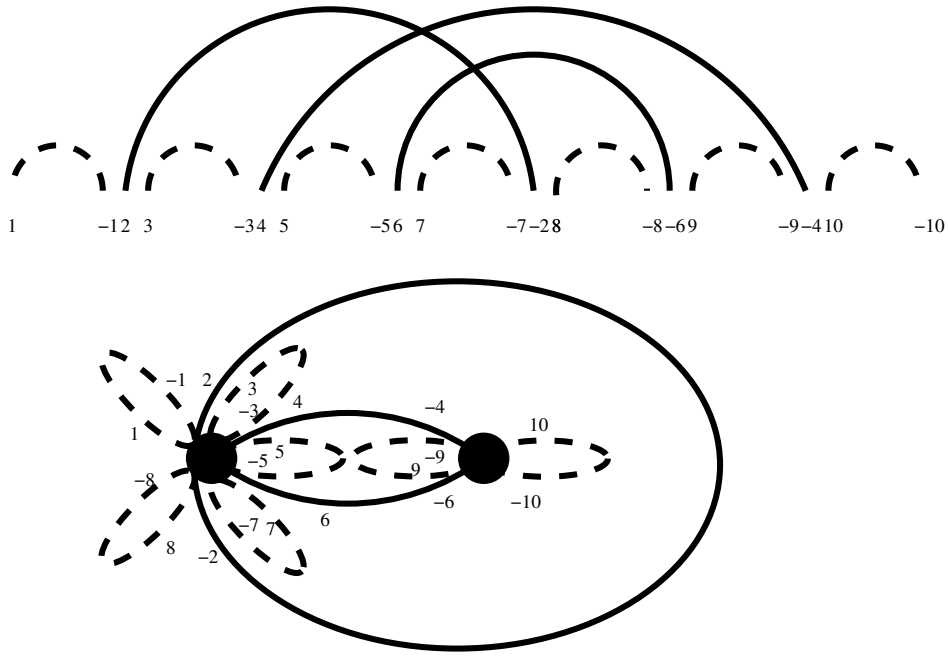
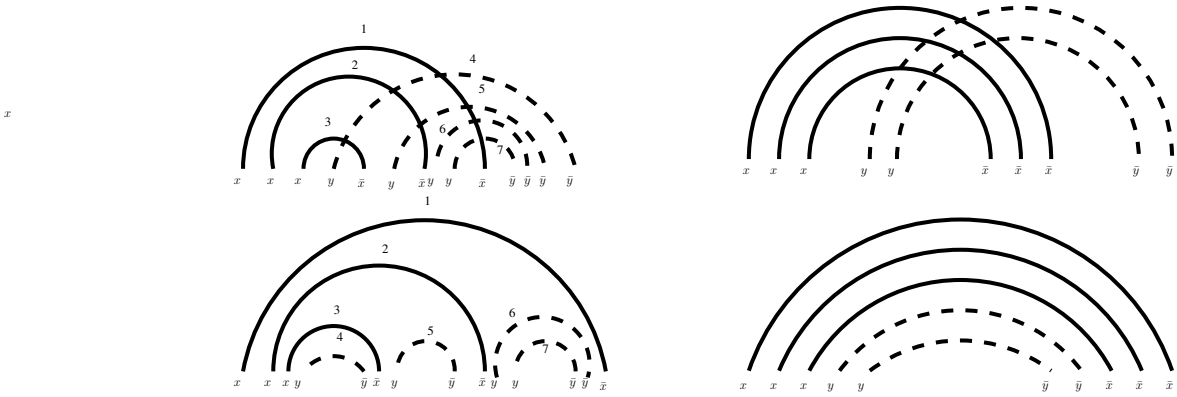


Figure 6: Top, a L&P structure corresponding to case 2. Bottom, the corresponding planar map. Arcs not involved in the pseudoknot are drawn in dotted lines.



(a) Simple pseudoknot (b) H-type pseudoknot

Figure 7: Top: two pseudoknots and their encodings v. Bottom: the corresponding nested structures and their encodings v' given by Proposition 1. Full lines represent x and $\bar{x}$, dotted lines represent y and by.

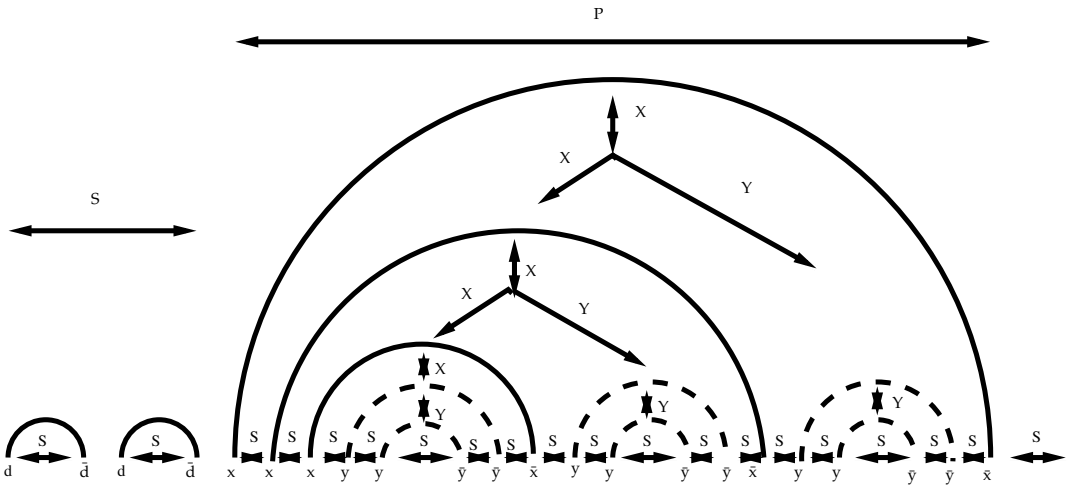


Figure 8: Building a structure with the grammar of $L_{A\&U}$.

List of Tables

Class	asympt.	α	ω	Compl.	Remark
PKF	$\frac{\alpha}{2\sqrt{\pi}n^{3/2}}\omega^n$	2	4	$\mathcal{O}(n^3)$	Catalan numbers
L&P *	$\frac{1}{2}\omega^n$	-	4	$\mathcal{O}(n^5)$	Closed formula
C&C *	$\frac{\alpha}{2\sqrt{\pi}n^{3/2}}\omega^n$	1,6651	5,857	$\mathcal{O}(n^6)$	
R&G *	$\frac{\alpha}{2\sqrt{\pi}n^{3/2}}\omega^n$	0,1651	6,576	$\mathcal{O}(n^4)$	
D&P *	$\frac{\alpha}{2\sqrt{\pi}n^{3/2}}\omega^n$	0,7535	7,315	$\mathcal{O}(n^5)$	
A&U *	$\frac{\alpha}{2\sqrt{\pi}n^{3/2}}\omega^n$	0,6575	7,547	$\mathcal{O}(n^5)$	
R&E	open	-	-	$\mathcal{O}(n^6)$	
All	$\sqrt{2}\cdot 2^n\cdot\left(\frac{n}{e}\right)^n$	-	-	NPC	Involutions with no fixed points

Table 1: Counting and complexity results. We indicate by "*" the classes that had not been counted before. The class "All" denotes the whole set of pseudoknotted structures. The row "Compl" gives the complexity of each algorithm.