Digital search trees and chaos game representation

Peggy Cénac, Brigitte Chauvin, Stéphane Ginouillac, Nicolas Pouyanne

ESAIM: Probability and Statistics Will be set by the publisher
URL: http://www.emath.fr/ps/

DIGITAL SEARCH TREES AND CHAOS GAME REPRESENTATION

Peggy Cénac[1], Brigitte Chauvin[2], Stéphane Ginouillac[2] and Nicolas Pouyanne[2]

Abstract. In this paper, we consider a possible representation of a DNA sequence in a quaternary tree, in which one can visualize repetitions of subwords (seen as suffixes of subsequences). The CGR-tree turns a sequence of letters into a Digital Search Tree (DST), obtained from the suffixes of the reversed sequence. Several results are known concerning the height, the insertion depth for DST built from independent successive random sequences having the same distribution. Here the successive inserted words are strongly dependent. We give the asymptotic behaviour of the insertion depth and the length of branches for the CGR-tree obtained from the suffixes of a reversed i.i.d. or Markovian sequence. This behaviour turns out to be at first order the same one as in the case of independent words. As a by-product, asymptotic results on the length of longest runs in a Markovian sequence are obtained.

Résumé. La représentation définie ici est une représentation possible de séquence d'ADN dans un arbre quaternaire dont la construction permet de visualiser les répétitions de suffixes. À partir d'une séquence de lettres, on construit un arbre digital de recherche (*Digital Search Tree*) sur l'ensemble des suffixes de la séquence inversée. Des résultats sur la hauteur et la profondeur d'insertion ont été établis lorsque les séquences à placer dans l'arbre sont indépendantes les unes des autres. Ici les mots à insérer sont fortement dépendants. On donne le comportement asymptotique de la profondeur d'insertion et de la longueur des branches pour un arbre obtenu à partir des suffixes d'une séquence i.i.d. ou markovienne retournée. Au premier ordre, cette asymptotique est la même que dans le cas où les mots insérés sont indépendants. De plus, certains résultats peuvent aussi s'interpréter comme des résultats de convergence sur les longueurs de plus longues répétitions d'une lettre dans une séquence Markovienne.

1991 Mathematics Subject Classification. Primary: 60C05, 68R15. Secondary: 92D20, 05D40.

The dates will be set by the publisher.

1. Introduction

In the last years, DNA has been represented by means of several methods in order to make pattern visualization easier and to detect local or global similarities (see for instance Roy et al. [27]). The *Chaos Game Representation* (CGR) provides both a graphical representation and a storage tool. From

Keywords and phrases: Random tree, Digital Search Tree, CGR, lengths of the paths, height, insertion depth, asymptotic growth, strong convergence

[1] INRIA Rocquencourt and Université Paul Sabatier (Toulouse III) – INRIA Domaine de Voluceau B.P.105 78 153 Le Chesnay Cedex (France)

[2] LAMA, UMR CNRS 8100, Bâtiment Fermat, Université de Versailles - Saint-Quentin F-78035 Versailles

a sequence in a finite alphabet, CGR defines a trajectory in a bounded subset of $\mathbb{R}^d$ that keeps all statistical properties of the sequence. Jeffrey [16] was the first to apply this iterative method to DNA sequences. Cénac [5], Cénac et al. [6] study the CGR with an extension of word-counting based methods of analysis. In this context, sequences are made of 4 nucleotides named A (adenine), C (cytosine), G (guanine) and T (thymine).

The CGR of a sequence $U_1 \dots U_n \dots$ of letters U_n from a finite alphabet $\mathcal{A}$ is the sequence $(\mathcal{X}_n)_{n\geq 0}$ of points in an appropriate compact subset S of $\mathbb{R}^d$ defined by

$$\begin{cases} \mathcal{X}_0 \in S \\ \mathcal{X}_{n+1} = \theta\big(\mathcal{X}_n + \ell_{U_{n+1}}\big), \end{cases}$$

where θ is a real parameter $(0 < \theta < 1)$, each letter $u \in \mathcal{A}$ being assigned to a given point $\ell_u \in S$. In the particular case of Jeffrey's representation, $\mathcal{A} = \{A, C, G, T\}$ is the set of nucleotides, $S = [0,1]^2$ is the unit square. Each letter is placed at a vertex as follows:

$$\ell_A = (0,0), \quad \ell_C = (0,1), \quad \ell_G = (1,1), \quad \ell_T = (1,0),$$

$\theta = \frac{1}{2}$ and the first point $\mathcal{X}_0$ is the center of the square. Then, iteratively, the point $\mathcal{X}_{n+1}$ is the middle of the segment between $\mathcal{X}_n$ and the square's vertex $\ell_{U_{n+1}}$:

$$\mathcal{X}_{n+1} = \frac{\mathcal{X}_n + \ell_{U_{n+1}}}{2},$$

or, equivalently,

$$\mathcal{X}_n = \sum_{k=1}^{n} \frac{\ell_{U_k}}{2^{n-k+1}} + \frac{\mathcal{X}_0}{2^n}.$$

Figure 1 represents the construction of the word ATGCGAGTGT.

With each deterministic word $w = u_1 \dots u_n$, we associate the half-opened subsquare Sw defined by the formula

$$Sw \stackrel{\text{def}}{=} \sum_{k=1}^{n} \frac{\ell_{u_k}}{2^{n-k+1}} + \frac{1}{2^n}[0,1[^2;$$

it has center $\sum_{k=1}^{n} \ell_{u_k}/2^{n-k+1} + \mathcal{X}_0/2^n$ and side $1/2^n$. For a given random or deterministic sequence $U_1 \dots U_n \dots$, for any word w and any $n \geq |w|$ (the notation $|w|$ stands for the number of letters in w), counting the number of points $(\mathcal{X}_i)_{1\leq i\leq n}$ that belong to the subsquare Sw is tantamount to counting the number of occurences of w as a subword of $U_1 \dots U_n$. Indeed, all successive words from the sequence having w as a suffix are represented in Sw. See Figure 1 for an example with three-letter subwords. This provides tables of word frequencies (see Goldman [14]). One can generalize it to any subdivision of the unit square; when the number of subsquares is not a power of 4, the table of word frequencies defines a counting of words with noninteger length (see Almeida et al. [2]).

The following property of the CGR is important: *the value of any $\mathcal{X}_n$ contains the historical information of the whole sequence $\mathcal{X}_1, \dots \mathcal{X}_n$.* Indeed, notice first that, by construction, $\mathcal{X}_n \in Su$ with $U_n = u$; the whole sequence is now given by the inductive formula $\mathcal{X}_{n-1} = 2\mathcal{X}_n - \ell_{U_n}$.

We define a representation of a random DNA sequence $U = (U_n)_{n\geq 1}$ as a random quaternary tree, the *CGR-tree*, in which one can visualize repetitions of subwords. We adopt the classical order (A, C, G, T) on letters. Let $\mathcal{T}$ be the complete infinite 4-ary tree; each node of $\mathcal{T}$ has four branches corresponding to letters (A, C, G, T) that are ordered in the same way. The CGR-tree of U is an

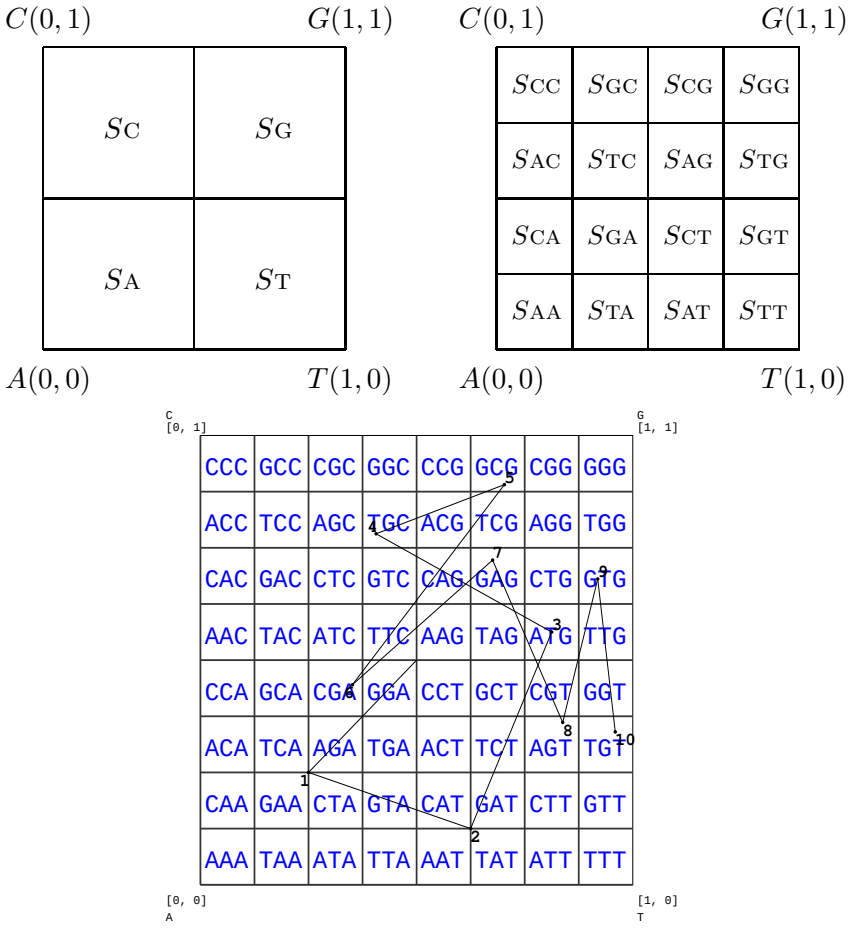

FIGURE 1. Chaos Game Representation of the first 10 nucleotides of the *E. Coli* threonine gene thrA: ATGCGAGTGT. The coordinates for each nucleotide are calculated recursively using $(0.5, 0.5)$ as starting position. The sequence is read from left to right. Point number 3 corresponds to the first 3-letter word ATG. It is located in the corresponding quadrant. The second 3-letter word TGC corresponds to point 4 and so on.

increasing sequence $\mathcal{T}_1 \subset \mathcal{T}_2 \ldots \subset \mathcal{T}_n \subset \ldots$ of finite subtrees of $\mathcal{T}$, each $\mathcal{T}_n$ having n nodes. The $\mathcal{T}_n$'s are built by successively inserting the *reversed prefixes*

$$W(n) = U_n \ldots U_1 \tag{1}$$

as follows in the complete infinite tree. First letter $W(1) = U_1$ is inserted in the complete infinite tree at level 1, *i.e.* just under the root, at the node that corresponds to the letter U_1. Inductively, the insertion of the word $W(n) = U_n \ldots U_1$ is made as follows: try to insert it at level 1 at the node $\mathcal{N}$ that corresponds to the letter U_n. If this node $\mathcal{N}$ is vacant, insert $W(n)$ at $\mathcal{N}$; if $\mathcal{N}$ is not vacant, try to insert $W(n)$ in the subtree having $\mathcal{N}$ as a root, at the node that corresponds to the letter U_{n-1}, and so on. One repeats this operation until the node at level k that corresponds to letter U_{n-k+1} is vacant; word $W(n)$ is then inserted at that node.

We complete our construction by labelling the n-th inserted node with the word $W(n)$. One readily obtains this way the process of a digital search tree (DST), as stated in the following proposition.

Figure 2 shows the very first steps of construction of the tree that corresponds to any sequence that begins with $GAGCACAGTGGAAGGG$. The insertion of this complete 16-letter prefix is represented in Figure 3. In these figures, each node has been labelled by its order of insertion to make the example more readable.

Proposition 1.1. *The CGR-tree of a random sequence $U = U_1U_2\ldots$ is a digital search tree, obtained by insertion in a quaternary tree of the successive reversed prefixes U_1, U_2U_1, $U_3U_2U_1$, ... of the sequence.*

The main results of our paper are the following convergence results, the random sequence U being supposed to be Markovian. If ℓ_n and $\mathcal{L}_n$ denote respectively the length of the shortest and of the longest branch of the CGR-tree, then $\ell_n/\ln n$ *and* $\mathcal{L}_n/\ln n$ *converge almost surely to some constants* (Theorem 3.1). Moreover, if D_n denotes the insertion depth and if M_n is the length of a uniformly chosen random path, then $D_n/\ln n$ *and* $M_n/\ln n$ *converge in probability to a common constant* (Theorem 4.1).

Remark 1.2. A given CGR-tree without its labels (i.e. a given shape of tree) is equivalent to a list of words in the sequence without their order. More precisely, one can associate with a shape of CGR-tree, a representation in the unit square as described below. With any node of the tree (which is in bijection with a word $w = W_1\ldots W_d$), we associate the center of the corresponding square Sw,

$$\mathcal{X}_w \stackrel{\text{def}}{=} \sum_{k=1}^{d} \frac{\ell_{W_k}}{2^{d-k+1}} + \frac{\mathcal{X}_0}{2^d}.$$

For example, Figure 3 shows this "*historyless representation*" for the word $GAGCACAGTGGAAGGG$. Moreover Figure 4 enables us to qualitatively compare the original and the historyless representations on an example.

Several results are known (see chap. 6 in Mahmoud [19]), concerning the height, the insertion depth and the profile for DST obtained from *independent* successive sequences, having the same distribution. It is far from our situation where the successive inserted words are strongly dependent from each other. Various results concerning the so-called Bernoulli model (binary trees, independent sequences and the two letters have the same probability 1/2 of appearance) can be found in Mahmoud [19]. Aldous and Shields [1] prove by embedding in continuous time, that the height satisfies $H_n - \log_2 n \to 0$ in probability. Also Drmota [7] proves that the height of such DSTs is concentrated: $\mathbb{E}[H_n - \mathbb{E}(H_n)]^L$ is asymptotically bounded for any $L > 0$.

For DST constructed from independent sequences on an m-letter alphabet with nonsymmetric (i.e. non equal probabilities on the letters) i.i.d or Markovian sources, Pittel [22] gets several results on the insertion depth and on the height. Despite the independence of the sequences, Pittel's work seems to be the closest to ours, and some parts of our proofs are inspired by it.

Some proofs in the sequel use classical results on the distribution of word occurences in a random sequence of letters (independent or Markovian sequences). Blom and Thorburn [4] give the generating function of the first occurence of a word for i.i.d. sequences, based on a recurrence relation on the probabilities. This result is extended to Markovian sequences by Robin and Daudin [26]. Several studies in this domain are based on generating functions, for example Régnier [24], Reinert et al. [25], Stefanov and Pakes [29]. Nonetheless, other approaches are considered: one of the more general

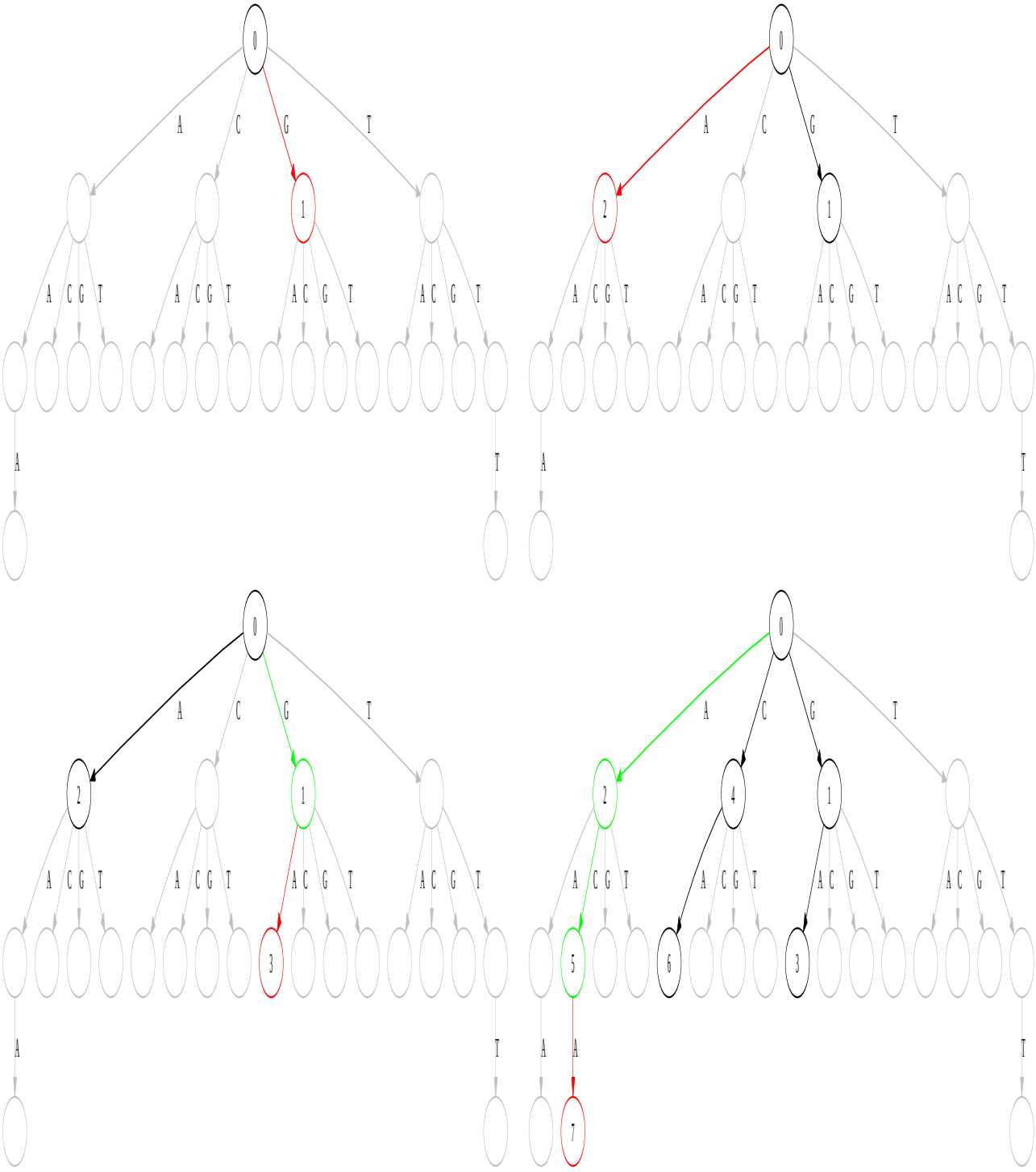

FIGURE 2. Insertion of a sequence $GAGCACAGTGGAAGGG...$ in its CGR-tree: first, second, third and seventh steps.

techniques is the Markov chain embedding method introduced by Fu [11] and further developped by Fu and Koutras [12], Koutras [17]. A martingale approach (see Gerber and Li [13], Li [18], Williams [30]) is an alternative to the Markov chain embedding method to solve problems around Penney [20] Game. These two approaches are compared in Pozdnyakov et al. [23]. Whatever method one uses, the distribution of the first occurence of a word strongly depends on its overlapping structure. This dependence is at the core of our proofs.

As a by-product, our results yield asymptotic properties on the length of the longest run, which is a natural object of study. In i.i.d. and symmetric sequences, Erdős and Révész [9] establish almost sure results about the growth of the longest run. These results are extended to Markov chains in Samarova [28], and Gordon et al. [15] show that the probabilistic behaviour of the length of the longest run is closely approximated by that of the maximum of some i.i.d. exponential random variables.

The paper is organized as follows. In Section 2 we establish the assumptions and notations we use throughout. Section 3 is devoted to almost sure convergence of the shortest and the longest branches in CGR-trees. In Section 4 asymptotic behaviour of the insertion depth is studied. An appendix deals separately with the domain of definition of the generating function of a certain waiting time related to the overlapping structure of words.

2. Assumptions and notations

In all the sequel, the sequence $U = U_1 \dots U_n \dots$ is supposed to be a Markov chain of order 1, with transition matrix Q and invariant measure p as initial distribution.

For any deterministic infinite sequence s, let us denote by $s^{(n)}$ the word formed by the n first letters of s, that is to say $s^{(n)} \stackrel{\text{def}}{=} s_1 \dots s_n$, where s_i is the i-th letter of s. The measure p is extended to reversed words the following way: $p(s^{(n)}) \stackrel{\text{def}}{=} \mathbb{P}(U_1 = s_n, \dots, U_n = s_1)$. The need for reversing the word $s^{(n)}$ comes from the construction of the CGR-tree which is based on reversed sequences (1).

We define the constants

$$\begin{aligned} h_+ &\stackrel{\text{def}}{=} \lim_{n\to+\infty} \frac{1}{n} \max\Big\{\ln\Big(\frac{1}{p\big(s^{(n)}\big)}\Big),\ p\big(s^{(n)}\big) > 0\Big\},\\ h_- &\stackrel{\text{def}}{=} \lim_{n\to+\infty} \frac{1}{n} \min\Big\{\ln\Big(\frac{1}{p\big(s^{(n)}\big)}\Big),\ p\big(s^{(n)}\big) > 0\Big\},\\ h &\stackrel{\text{def}}{=} \lim_{n\to+\infty} \frac{1}{n} \mathbb{E}\Big[\ln\Big(\frac{1}{p\big(U^{(n)}\big)}\Big)\Big]. \end{aligned}$$

Due to an argument of sub-additivity (see Pittel [22]), these limits are well defined (in fact, in a more general than Markovian sequences framework). Moreover, Pittel proves the existence of two infinite sequences denoted here by s_+ and s_- such that

$$h_+ = \lim_{n\to\infty} \frac{1}{n} \ln\Big(\frac{1}{p\big(s_+^{(n)}\big)}\Big), \quad \text{and} \quad h_- = \lim_{n\to\infty} \frac{1}{n} \ln\Big(\frac{1}{p\big(s_-^{(n)}\big)}\Big). \tag{2}$$

For any $n \geq 1$, the notation $\mathcal{T}_n \stackrel{\text{def}}{=} \mathcal{T}_n(W)$ stands for the finite tree with n nodes (without counting the root), built from the first n sequences $W(1), \dots, W(n)$, which are the successive reversed prefixes of the sequence $(U_n)_n$, as defined by (1). $\mathcal{T}_0$ denotes the tree reduced to the root. In particular, the random trees are increasing: $\mathcal{T}_0 \subset \mathcal{T}_1 \dots \subset \mathcal{T}_n \subset \dots \subset \mathcal{T}$.

Let us define ℓ_n (resp. $\mathcal{L}_n$) as the length of the shortest (resp. the longest) path from the root to a feasible external node of the tree $\mathcal{T}_n(w)$. Moreover, D_n denotes the insertion depth of $W(n)$ in $\mathcal{T}_{n-1}$ to build $\mathcal{T}_n$. Finally M_n is the length of a path of $\mathcal{T}_n$, randomly and uniformly chosen in the n possible paths.

The following random variables play a key role in the proofs. For the sake of precision, let us recall that s is deterministic, the randomness is uniquely due to the generation of the sequence U. First we define for any infinite sequence s and for any $n \geq 0$,

$$X_n(s) \stackrel{\text{def}}{=} \begin{cases} 0 \text{ if } s_1 \text{ is not in } \mathcal{T}_n \\ \max\{k \text{ such that } s^{(k)} \text{ is already inserted in } \mathcal{T}_n\}. \end{cases} \tag{3}$$

Notice that $X_0(s) = 0$. Every infinite sequence corresponds to a branch of the infinite tree $\mathcal{T}$ (root at level 0, node that corresponds to s_1 at level 1, node that corresponds to s_2 at level 2, *etc.*); the random variable $X_n(s)$ is the length of the branch associated with s in the tree $\mathcal{T}_n$. For any $k \geq 0$, $T_k(s)$ denotes the size of the first tree where $s^{(k)}$ is inserted:

$$T_k(s) \stackrel{\text{def}}{=} \min\{n,\ X_n(s) = k\}$$

(notice that $T_0(s) = 0$).

These two variables are in duality in the following sense: one has equality of the events

$$\{X_n(s) \geq k\} = \{T_k(s) \leq n\} \tag{4}$$

and consequently, $\{T_k(s) = n\} \subset \{X_n(s) = k\}$ since $X_n(s) - X_{n-1}(s) \in \{0, 1\}$.

In our example of Figures 2 and 3, the drawn random sequence is $GAGCACAGTGGAAGGG\ldots$ If one takes a deterministic sequence s such that $s^{(3)} = ACA$, then $X_0(s) = X_1(s) = 0$, $X_2(s) = X_3(s) = X_4(s) = 1$, $X_5(s) = X_6(s) = 2$ and $X_k(s) = 3$ for $7 \leq k \leq 18$. The first three values of $T_k(s)$ are consequently $T_1(s) = 2$, $T_2(s) = 5$, $T_3(s) = 7$.

Moreover, the random variable $T_k(s)$ can be decomposed as follows,

$$T_k(s) = \sum_{r=1}^{k} Z_r(s), \tag{5}$$

where $Z_r(s) \stackrel{\text{def}}{=} T_r(s) - T_{r-1}(s)$ is the number of letters to read before the branch that corresponds to s increases by 1. In what follows, $Z_r(s)$ can be viewed as the waiting time n of the first occurence of $s^{(r)}$ in the sequence

$$\ldots U_{n+T_{r-1}(s)} U_{n-1+T_{r-1}(s)} \ldots U_{1+T_{r-1}(s)} s^{(r-1)},$$

i.e. $Z_r(s)$ can also be defined as

$$Z_r(s) = \min\{n \geq 1,\ U_{n+T_{r-1}(s)} \ldots U_{n+T_{r-1}(s)-r+1} = s_1 \ldots s_r\}.$$

Because of the Markovianity of the model, the random variables $Z_r(s)$ are independent.

Let us then introduce $Y_r(s)$ as being the waiting time of the first occurence of $s^{(r)}$ in the sequence

$$\ldots U_{n+T_{r-1}(s)} U_{n-1+T_{r-1}(s)} \ldots U_{1+T_{r-1}(s)},$$

that is to say

$$Y_r(s) = \min\{n \geq r,\ U_{n+T_{r-1}(s)} \dots U_{n+T_{r-1}(s)-r+1} = s_1 \dots s_r\}.$$

One has readily the inequality $Z_r(s) \leq Y_r(s)$. More precisely, if the word $s^{(r)}$ is inserted in the sequence before time $T_{r-1}(s) + r$, there is some overlapping between prefixes of $s^{(r-1)}$ and suffixes of $s^{(r)}$. See Figure 5 for an example where $r = 6$ and $s_1s_2s_3 = s_4s_5s_6$. Actually, variables $Z_r(s)$ and $Y_r(s)$ are related by

$$Z_r(s) = \mathbb{1}_{\{Z_r(s)<r\}} Z_r(s) + \mathbb{1}_{\{Z_r(s)\geq r\}} Y_r(s).$$

Since the sequence $(U_n)_{n\geq 1}$ is stationary, the conditional distribution of $Y_r(s)$ given $T_{r-1}(s)$ is the distribution of the first occurence of the word $s^{(r)}$ in the realization of a Markov chain of order 1, whose transition matrix is Q and whose initial distribution is its invariant measure. In particular the conditional distribution of $Y_r(s)$ given $T_{r-1}(s)$ is independent of $T_{r-1}(s)$.

The generating function $\Phi(s^{(r)}, t) \stackrel{\text{def}}{=} \mathbb{E}[t^{Y_r(s)}]$ is given by Robin and Daudin [26]:

$$\Phi(s^{(r)}, t) = \Big(\gamma_r(t) + (1-t)\delta_r(t^{-1})\Big)^{-1}, \tag{6}$$

where the functions γ and δ are respectively defined as

$$\gamma_r(t) \stackrel{\text{def}}{=} \frac{1-t}{tp\big(s_r\big)} \sum_{m\geq 1} Q^m(s_1, s_r) t^m, \quad \delta_r(t^{-1}) \stackrel{\text{def}}{=} \sum_{m=1}^{r} \frac{\mathbb{1}_{\{s_r \dots s_{r-m+1} = s_m \dots s_1\}}}{t^m p\big(s^{(m)}\big)}, \tag{7}$$

and where $Q^m(u, v)$ denotes the transition probability from u to v in m steps.

Remark 2.1. In the particular case when the sequence of nucleotides $(U_n)_{n\geq 1}$ is supposed to be independent and identically distributed according to the non degenerated law (p_A, p_C, p_G, p_T), the transition probability $Q^m(s_1, s_r)$ is equal to $p(s_r)$, and hence $\gamma_r(t) = 1$.

Proposition 2.2. (i) *The generating function of $Y_r(s)$ defined by (6) has a ray of convergence $\geq 1 + \kappa p\big(s^{(r)}\big)$ where κ is a positive constant independent of r and s.*

(ii) *Let γ denote the second largest eigenvalue of the transition matrix Q. For all $t \in]-\gamma^{-1}, \gamma^{-1}[$,*

$$\big|\gamma_r(t) - 1\big| \leq \frac{|1-t|}{1-\gamma|t|}\kappa', \tag{8}$$

where κ' is some positive constant independent of r and s (if $\gamma = 0$ or if the sequence is i.i.d., we adopt the convention $\gamma^{-1} = +\infty$ so that the result remains valid).

Proof. The proof of Proposition 2.2 is given in Appendix A. □

3. Length of the branches

In this section we are concerned with the asymptotic behaviour of the length ℓ_n (resp. $\mathcal{L}_n$) of the shortest (resp. longest) branch of the CGR-tree.

Theorem 3.1.

$$\frac{\ell_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_+}, \quad \textit{and} \quad \frac{\mathcal{L}_n}{\ln n} \xrightarrow[n\to\infty]{\text{a.s.}} \frac{1}{h_-}.$$

According to the definition of $X_n(s)$, the lengths ℓ_n and $\mathcal{L}_n$ are functions of X_n:

$$\ell_n = \min_{s\in\mathcal{A}^{\mathbb{N}}} X_n(s), \quad \text{and} \quad \mathcal{L}_n = \max_{s\in\mathcal{A}^{\mathbb{N}}} X_n(s). \tag{9}$$

The following key lemma gives an asymptotic result on $X_n(s)$, under suitable assumptions on s. Our proof of Theorem 3.1 is based on it.

Lemma 3.2. *Let s be such that there exists*

$$\lim_{n\longrightarrow+\infty} \frac{1}{n} \ln\left(\frac{1}{p\left(s^{(n)}\right)}\right) \stackrel{\text{def}}{=} h(s) > 0. \tag{10}$$

Then

$$\frac{X_n(s)}{\ln n} \xrightarrow[\text{n}\to\infty]{\text{a.s.}} \frac{1}{h(s)}.$$

Remark 3.3. Let $\tilde{v} \stackrel{\text{def}}{=} vv\ldots$ consist of repetitions of a letter v. Then $X_n(\tilde{v})$ is the length of the branch associated with $\tilde{v}$ in $\mathcal{T}_n$. For such a sequence (and exclusively for them) the random variable $Y_k(\tilde{v})$ is equal to $T_k(\tilde{v})$. Consequently $X_n(\tilde{v})$ is the length of the longest run of 'v' in $U_1 \ldots U_n$. When $(U_n)_{n\geq 1}$ is a sequence of i.i.d. trials, Erdős and Révész [9], Erdős and Révész [10], Petrov [21] showed that

$$\frac{X_n(\tilde{v})}{\ln n} \xrightarrow[\text{n}\to\infty]{\text{a.s.}} \frac{1}{\ln \frac{1}{p}},$$

where $p \stackrel{\text{def}}{=} \mathbb{P}(U_i = v)$. This convergence result is a particular case of Lemma 3.2.

Simulations. In a first set of computations, two random sequences whose letters are i.i.d. were generated. On Figure 6, in the first graph, letters are equally-likely drawn; in the second one, they are drawn with respective probabilities $(p_A, p_C, p_G, p_T) = (0.4, 0.3, 0.2, 0.1)$. On can visualize the dynamic convergence of $\mathcal{L}_n/\ln n$, $\ell_n/\ln n$ and of the normalized insertion depth $D_n/\ln n$ (see section 4) to their respective constant limits.

Figure 7 is made from simulations of 2,000 random sequences of length 100,000 with i.i.d. letters under the distribution $(p_A, p_C, p_G, p_T) = (0.6, 0.1, 0.1, 0.2)$. On the x-axis, respectively, lengths of the shortest branches, insertion depth of the last inserted word, lengths of the longest branches. On the y-axis, number of occurences (histograms).

Proof of Lemma 3.2. Since $X_n(s) = k$ for $n = T_k(s)$ (see Equation (4)), by monotonicity arguments, it is sufficient to prove that

$$\frac{\ln T_k(s)}{k} \xrightarrow[\text{k}\to\infty]{\text{a.s.}} h(s).$$

Let $\varepsilon_r(s) \stackrel{\text{def}}{=} Z_r(s) - \mathbb{E}\left[Z_r(s)\right]$, so that $T_k(s)$ admits the decomposition

$$T_k(s) = \mathbb{E}\left[T_k(s)\right] + \sum_{r=1}^{k} \varepsilon_r(s).$$

If $(M_k(s))_k$ is the martingale defined by

$$M_k(s) \stackrel{\text{def}}{=} \sum_{r=1}^{k} \varepsilon_r(s),$$

taking the logarithm in the preceding equation leads to

$$\ln T_k(s) = \ln \mathbb{E}\left[T_k(s)\right] + \ln\left(1 + \frac{M_k(s)}{\mathbb{E}\left[T_k(s)\right]}\right). \tag{11}$$

• It is shown in Robin and Daudin [26] that $\mathbb{E}\left[Z_n(s)\right] = 1/p\left(s^{(n)}\right)$ so that the sequence $\frac{1}{n}\ln \mathbb{E}\left[Z_n(s)\right]$ converges to $h(s)$ as n tends to infinity ($h(s)$ is defined by (10)). Since $\mathbb{E}\left[T_k(s)\right] = \sum_{r=1}^{k} \mathbb{E}\left[Z_r(s)\right]$ (see (5)), the equality

$$\lim_{k\to\infty} \frac{1}{k} \ln \mathbb{E}\left[T_k(s)\right] = h(s)$$

is a straightforward consequence of the following elementary result: if $(x_k)_k$ is a sequence of positive numbers such that $\lim_{k\to\infty} \frac{1}{k}\ln\left(x_k\right) = h > 0$, then $\lim_{k\to\infty} \frac{1}{k}\ln\left(\sum_{r=1}^{k} x_r\right) = h$.

• The martingale $(M_k(s))_k$ is square integrable; its increasing process is denoted by $\left(\langle M(s)\rangle_k\right)_k$. Robin and Daudin [26] have shown that the variance of $Z_r(s)$ satisfies $\mathbb{V}\left[Z_r(s)\right] \le 4r/p\left(s^{(r)}\right)^2$, so that

$$\langle M(s)\rangle_k = O\left(ke^{2kh(s)}\right).$$

One can thus apply the Law of Large Numbers for martingales (see Duflo [8] for a reference on the subject): for any $\alpha > 0$,

$$M_k(s) = O\left(\langle M(s)\rangle_k^{1/2}\left(\ln\langle M(s)\rangle_k\right)^{\frac{1+\alpha}{2}}\right) \quad a.s.$$

Consequently,

$$\frac{M_k(s)}{\mathbb{E}\left[T_k(s)\right]} = O\left(k^{1+\alpha/2}\right) \quad a.s.$$

which completes the proof of Lemma 3.2. □

Proof of Theorem 3.1. It is inspired from Pittel [22]. Clearly the definition given in Equation (9) yields

$$\ell_n \le X_n(s_+) \quad \text{and} \quad \mathcal{L}_n \ge X_n(s_-)$$

(definitions of s_+ and s_- were given in (2)). Hence, by Lemma 3.2

$$\limsup_{n\to\infty} \frac{\ell_n}{\ln n} \le \frac{1}{h_+}, \quad \liminf_{n\to\infty} \frac{\mathcal{L}_n}{\ln n} \ge \frac{1}{h_-} \quad \text{a.s.}$$

• *Proof for* ℓ_n
For any integer r,

$$\mathbb{P}(\ell_n \le r-1) \le \sum_{s^{(r)}\in\mathcal{A}^r} \mathbb{P}(X_n(s) \le r-1) \le \sum_{s^{(r)}\in\mathcal{A}^r} \mathbb{P}(T_r(s) \ge n), \tag{12}$$

where the above sums are taken over the set $\mathcal{A}^r$ of words with length r (for a proper meaning of this formula, one should replace s by any infinite word having $s^{(r)}$ as prefix, in both occurences). We abuse of this notation from now on. Since the generating functions $\Phi(s^j, t)$ are defined for any

$1 \le t < \min\{\gamma^{-1}, 1+\kappa p(s^{(r)})\}$ and $j \le r$ (see Assertion i) in Proposition 2.2), each term of the sum (12) can be controlled by

$$\mathbb{P}(T_r(s) \ge n) \le t^{-n}\mathbb{E}[t^{T_r(s)}] \le t^{-n}\prod_{j=1}^{r}\Phi(s^{(j)},t).$$

In particular, bounding above all the overlapping functions $\mathbb{1}_{\{s_j\dots s_1 = s_r\dots s_{r-j+1}\}}$ by 1 in (7), we deduce from (6) and from Assertion ii) of Proposition 2.2 that

$$\mathbb{P}(T_r(s) \ge n) \le t^{-n}\prod_{j=1}^{r}\biggl(1+(1-t)\Bigl(\sum_{\nu=1}^{j}\frac{1}{t^\nu p(s^{(\nu)})}+\frac{\kappa'}{1-\gamma t}\Bigr)\biggr)^{-1}.$$

Let $0<\varepsilon<1$. There exists a constant $c_2 \in]0,1[$ depending only on ε such that

$$p(s^{(j)}) > c_2\alpha^j, \quad \text{with} \quad \alpha \stackrel{\text{def}}{=} \exp(-(1+\varepsilon^2)h_+)$$

(for the sake of brevity c, c_1 and c_2 denote different constants all along the text). We then have

$$\mathbb{P}(T_r(s) \ge n) \le t^{-n}\prod_{j=1}^{r}\biggl(1+(1-t)\Bigl(\frac{1-(\alpha t)^{-j}}{c_2(\alpha t-1)}+\frac{\kappa'}{1-\gamma t}\Bigr)\biggr)^{-1}.$$

Choosing $t = 1+c_2\kappa\alpha^r$, Inequality (8) is valid if r is large enough, so that

$$\mathbb{P}(T_r(s) \ge n) \le ct^{-n}\prod_{j=1}^{r}\biggl(1-\kappa\alpha^{r-j}\frac{\alpha^j-(1+c_2\kappa\alpha^r)^{-j}}{\alpha(1+c_2\kappa\alpha^r)-1}-\frac{\alpha^r c_2\kappa\kappa'}{1-\gamma(1+c_2\kappa\alpha^r)}\biggr)^{-1}.$$

Moreover since obvioulsy

$$\lim_{j\to\infty}\frac{\alpha^j-(1+c_2\kappa\alpha^r)^{-j}}{\alpha(1+c_2\kappa\alpha^r)-1} = \frac{1}{1-\alpha},$$

and $c_2\kappa\kappa'/\bigl(1-\gamma(1+c_2\kappa\alpha^r)\bigr)$ is uniformly bounded in r, there exist two positive constants λ and L independent of j and r such that

$$\mathbb{P}(T_r(s) \ge n) \le (1+c_2\kappa\alpha^r)^{-n}L\prod_{j=1}^{r}\Bigl(1-\lambda\alpha^{r-j}\Bigr)^{-1}.$$

In addition, the product can be bounded above by

$$\prod_{j=1}^{r}\Bigl(1-\lambda\alpha^{r-j}\Bigr)^{-1} \le \prod_{j=0}^{\infty}\Bigl(1-\lambda\alpha^{j}\Bigr)^{-1} = R < \infty.$$

Consequently,

$$\mathbb{P}(T_r(s) \ge n) \le LR(1+c_2\kappa\alpha^r)^{-n}.$$

For $r = \lfloor (1-\varepsilon)\frac{\ln n}{h_+} \rfloor$ and ε small enough, there exists a constant R' such that

$$\mathbb{P}(T_r(s) > n) \leq R' \exp(-c_2 \kappa n^\theta),$$

where $\theta = \varepsilon - \varepsilon^2 + \varepsilon^3 > 0$. We then deduce from (12) that

$$\mathbb{P}(\ell_n \leq r-1) \leq 4^r R' \exp(-c_2 \kappa n^\theta),$$

which is the general term of a convergent series. Borel-Cantelli Lemma applies so that

$$\liminf_{n\to\infty} \frac{\ell_n}{\ln n} \geq \frac{1}{h_+} \quad \text{a.s.}$$

- *Proof for $\mathcal{L}_n$*
  To complete the proof, one needs to show that

$$\limsup_{n\to\infty} \frac{\mathcal{L}_n}{\ln n} \leq \frac{1}{h_-} \quad \text{a.s.}$$

Again, since $X_n(s) = k$ for $n = T_k(s)$, by monotonicity arguments it suffices to show that

$$\liminf_{k\to\infty} \min_{s^{(k)}\in\mathcal{A}^k} \frac{\ln T_k(s)}{k} \geq h_- \quad \text{a.s.}$$

(notations of (12)).

Let $0 < \varepsilon < 1$. As in the previous proof for the shortest branches, it suffices to bound above

$$\mathbb{P}\left(\min_{s^{(k)}\in\mathcal{A}^k} T_k(s) < e^{kh_-(1-\varepsilon)} \right)$$

by the general term of a convergent series to apply Borel-Cantelli Lemma. Obviously,

$$\mathbb{P}\left(\min_{s^{(k)}\in\mathcal{A}^k} T_k(s) < e^{kh_-(1-\varepsilon)} \right) \leq \sum_{s^{(k)}\in\mathcal{A}^k} \mathbb{P}\left(T_k(s) < e^{kh_-(1-\varepsilon)} \right).$$

If t is any real number in $]0,1[$ and if $n \stackrel{\text{def}}{=} \exp(kh_-(1-\varepsilon))$,

$$\mathbb{P}\left(T_k(s) < e^{kh_-(1-\varepsilon)} \right) = \mathbb{P}\left(t^{T_k(s)} > t^n \right)$$

and the decomposition (5), together with the independence of the $Z_r(s)$ for $1 \leq r \leq k$, yield

$$\mathbb{P}\left(t^{T_k(s)} > t^n \right) \leq t^{-n} \prod_{r=1}^{k} \mathbb{E}\left[t^{Z_r(s)} \right].$$

The proof consists in bounding above

$$\sum_{s^{(k)}\in\mathcal{A}^k} t^{-n} \prod_{r=1}^{k} \mathbb{E}\left[t^{Z_r(s)} \right] \tag{13}$$

by the general term of a convergent series, taking t of the form

$$t \stackrel{\text{def}}{=} (1 + c/n)^{-1}$$

so that the sequence $(t^n)_n$ is bounded.

The generating function of $Z_r(s)$ is given by Robin and Daudin [26] and strongly depends on the overlapping structure of the word $s^{(r)}$. As $0 < t < 1$, this function is well defined at t and is given by (see Assertion i) of Proposition 2.2)

$$\mathbb{E}\big[t^{Z_r(s)}\big] = 1 - \frac{(1-t)}{t^r p\big(s^{(r)}\big)\big(\gamma_r(t) + (1-t)\delta_r(t^{-1})\big)}, \tag{14}$$

where $\gamma_r(t)$ and $\delta_r(t)$ are defined in (7). Moreover, from Assertion ii) of Proposition 2.2, it is obvious that there exists a constant θ independent of r and s such that,

$$\gamma_r(t) \leq 1 + \theta(1-t). \tag{15}$$

Besides, by elementary change of variable, one has successively

$$\begin{aligned} t^r p\big(s^{(r)}\big)\delta_r(t^{-1}) &= \sum_{m=1}^{r} \mathbb{1}_{\{s_r \ldots s_{r-m+1} = s_m \ldots s_1\}} \frac{t^r p\big(s^{(r)}\big)}{t^m p\big(s^{(m)}\big)} \\ &= \sum_{m=1}^{r} \mathbb{1}_{\{s_r \ldots s_m = s_{r-m+1} \ldots s_1\}} t^{m-1} \frac{p\big(s^{(r)}\big)}{p\big(s^{(r-m+1)}\big)} \\ &= \sum_{m=1}^{r} \mathbb{1}_{\{s_r \ldots s_m = s_{r-m+1} \ldots s_1\}} t^{m-1} \frac{p\big(s^{(m)}\big)}{p\big(s_m\big)}. \end{aligned}$$

When m is large enough, h_-'s definition implies that

$$p\big(s^{(m)}\big) \leq \beta^m, \quad \text{where} \quad \beta \stackrel{\text{def}}{=} \exp(-(1-\varepsilon^2)h_-),$$

so that there exists positive constants ρ and c such that, for any r,

$$p\big(s^{(r)}\big) \leq c\beta^r \quad \text{and} \quad t^r p\big(s^{(r)}\big)\delta_r(t^{-1}) \leq 1 + \rho \sum_{m=2}^{r} \mathbb{1}_{\{s_r \ldots s_m = s_{r-m+1} \ldots s_1\}} \beta^m. \tag{16}$$

Thus Formula (14) with inequalities (15) and (16) yield, for any $r \leq k$,

$$\mathbb{E}[t^{Z_r(s)}] \leq 1 - \frac{1}{c\beta^r\left(\frac{1}{1-t} + \theta\right) + 1 + q_k(s)}, \tag{17}$$

where $q_k(s)$, that depends on the overlapping structure of $s^{(k)}$, is defined by

$$q_k(s) \stackrel{\text{def}}{=} \rho \max_{2 \leq r \leq k} \sum_{m=2}^{r} \mathbb{1}_{\{s_r \ldots s_m = s_{r-m+1} \ldots s_1\}} \beta^m.$$

Note that whatever the overlapping structure is, $q_k(s)$ is controlled by

$$0 \le q_k(s) \le \frac{\rho}{1-\beta}. \tag{18}$$

Thus,

$$\prod_{r=1}^{k} \mathbb{E}[t^{Z_r(s)}] \le \exp\Big[-\sum_{r=1}^{k} \ln\Big(1 - \frac{1}{c\beta^r\big((1-t)^{-1}+\theta\big) + 1 + q_k(s)}\Big)^{-1}\Big].$$

Since the function $x \mapsto \ln 1/(1-x)$ is increasing, comparing this sum with an integral and after the change of variable $y = c\beta^x\big((1-t)^{-1}+\theta\big)$, one obtains

$$\prod_{r=1}^{k} \mathbb{E}[t^{Z_r(s)}] \le \exp\Big[-\frac{1}{\ln \beta^{-1}} \int_{c\beta^k\left((1-t)^{-1}+\theta\right)}^{c((1-t)^{-1}+\theta)} \ln\Big(1 - \frac{1}{y+1+q_k(s)}\Big)^{-1} \frac{dy}{y}\Big].$$

This integral is convergent in a neighbourhood of $+\infty$, hence there exists a constant C, independent of k and s such that

$$\prod_{r=1}^{k} \mathbb{E}[t^{Z_r(s)}] \le C \exp\Big[-\frac{1}{\ln \beta^{-1}} \int_{c\beta^k\left((1-t)^{-1}+\theta\right)}^{+\infty} \ln\Big(1 - \frac{1}{y+1+q_k(s)}\Big)^{-1} \frac{dy}{y}\Big]. \tag{19}$$

The classical dilogarithm $\mathrm{Li}_2(z) = \sum_{k\ge1} z^k/k^2$, analytically continued to the complex plane slit along the ray $[1, +\infty[$, satisfies $\frac{d}{dy}\mathrm{Li}_2(-\frac{v}{y}) = \frac{1}{y}\log(1+v/y)$. This leads to the formula

$$\int_{a_k}^{+\infty} \ln\left(1 - \frac{1}{y+1+q_k(s)}\right)^{-1} \frac{dy}{y} = \mathrm{Li}_2\left(-\frac{q_k(s)}{a_k}\right) - \mathrm{Li}_2\left(-\frac{1+q_k(s)}{a_k}\right)$$

with the notation $a_k = c\beta^k\big((1-t)^{-1}+\theta\big)$. Choosing $t = (1+c/n)^{-1}$ yields readily

$$a_k \underset{k\to+\infty}{\sim} \exp(-kh_-(\varepsilon - \varepsilon^2)). \tag{20}$$

Moreover, in a neighbourhood of $-\infty$,

$$\mathrm{Li}_2(x) = -\frac{1}{2}\ln^2(-x) - \zeta(2) + O(\frac{1}{x}), \tag{21}$$

and the function $\mathrm{Li}_2(x) + \frac{1}{2}\ln^2(-x)$ is non-decreasing on $]-\infty, 0[$, so that

$$\begin{cases} \mathrm{Li}_2(x) \ge -\dfrac{1}{2}\ln^2(-x) - \zeta(2) & (x < 0) \\ \mathrm{Li}_2(x) \le -\dfrac{1}{2}\ln^2(-x) - \dfrac{\zeta(2)}{2} & (x < -1), \end{cases} \tag{22}$$

noting that $\mathrm{Li}_2(-1) = -\dfrac{\zeta(2)}{2}$. Hence, if k is such that $a_k < 1$,

$$\int_{a_k}^{+\infty} \ln\left(1 - \frac{1}{y+1+q_k(s)}\right)^{-1} \frac{dy}{y} \ge \mathrm{Li}_2\left(-\frac{q_k(s)}{a_k}\right) + \frac{1}{2}\ln^2(a_k) + \frac{\zeta(2)}{2} \tag{23}$$

with $\ln a_k$ being asymptotically proportional to k because of (20). Thus, the behaviour of the integral in (19) as k tends to $+\infty$ depends on the asymptotics of $q_k(s)$.

Let $z_k \stackrel{\text{def}}{=} \exp\bigl(-\sqrt{k}\bigr)$. The end of the proof consists, for a given k, in splitting the sum (13) into prefixes $s^{(k)}$ that respectively satisfy $q_k(s) < \exp\bigl(-\sqrt{k}\bigr)$ or $q_k(s) \geq \exp\bigl(-\sqrt{k}\bigr)$. These two cases correspond to words that respectly have few or many overlapping patterns. The choice $z_k = \exp\bigl(-\sqrt{k}\bigr)$ is arbitrary and many other sequence could have been taken provided that they converge to zero with a speed of the form $\exp[-o(k)]$.

First let us consider the case of prefixes $s^{(k)}$ such that $q_k(s) < \exp\bigl(-\sqrt{k}\bigr)$. For such words, (22) and (23) imply that

$$\int_{a_k}^{+\infty} \ln\left(1 - \frac{1}{y+1+q_k(s)}\right)^{-1} \frac{dy}{y} \geq -\frac{1}{2}\ln^2\left(\frac{z_k}{a_k}\right) + \frac{1}{2}\ln^2(a_k) - \frac{\zeta(2)}{2},$$

the second member of this inequality being, as k tends to infinity, of the form

$$k\sqrt{k}h_-(\varepsilon - \varepsilon^2) + O(k).$$

Consequently,

$$\prod_{r=1}^{k} \mathbb{E}[t^{Z_r(s)}] \leq \exp\left[-\frac{\varepsilon}{1+\varepsilon}k^{3/2} + O(k)\right].$$

There are 4^k words of length k, hence very roughly, by taking the sum over the prefixes $s^{(k)}$ such that $q_k(s) < z_k$, and since t^{-n} is bounded, the contribution of these prefixes to the sum (13) satisfies

$$\sum_{s^{(k)}\in\mathcal{A}^k,\ q_k(s)<z_k} t^{-n} \prod_{r=1}^{k} \mathbb{E}\bigl[t^{Z_r(s)}\bigr] \leq 4^k \exp\left[-\frac{\varepsilon}{1+\varepsilon}k^{3/2} + O(k)\right],$$

which is the general term of a convergent series.

It remains to study the case $q_k(s) \geq z_k$. For such words, let us only consider the inequalities (18) and (19) that lead to

$$\prod_{r=1}^{k} \mathbb{E}[t^{Z_r(s)}] \leq C \exp\left[-\frac{1}{\ln\beta^{-1}} \int_{a_k}^{+\infty} \ln\Bigl(1 - \frac{1}{y+1+\rho(1-\beta)^{-1}}\Bigr)^{-1} \frac{dy}{y}\right].$$

Since $x \leq \log(1-x)^{-1}$, after some work of integration,

$$\prod_{r=1}^{k} \mathbb{E}[t^{Z_r(s)}] \leq \exp\Bigl(-\frac{\varepsilon}{1+\varepsilon}k + o(k)\Bigr). \tag{24}$$

The natural question arising now is: how many words $s^{(k)}$ are there, such that $q_k(s) \geq z_k$? Let us define

$$E_k \stackrel{\text{def}}{=} \Bigl\{s^{(k)},\ q_k(s) \geq e^{-\sqrt{k}}\Bigr\}.$$

The definition of $q_k(s)$ implies clearly that

$$E_k \subseteq \Big\{ s^{(k)},\ \exists r \leq k,\ \rho \sum_{m=2}^{r} \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m \geq e^{-\sqrt{k}} \Big\}.$$

For any $r \leq k$ and $x > 0$, let us define the set

$$S_r(x) \stackrel{\text{def}}{=} \Big\{ s^{(k)},\ \sum_{m=2}^{r} \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} \beta^m < x \Big\}.$$

For any $l \in \{2, \dots, r\}$, one has the following inclusion

$$\bigcap_{m=2}^{\ell} \Big\{ s^{(k)},\ \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} = 0 \Big\} \subset S_r\Big(\frac{\beta^{\ell+1}}{1-\beta}\Big).$$

If the notation B^c denotes the complementary set of B in $\mathcal{A}^k$,

$$S_r^c\Big(\frac{\beta^{\ell+1}}{1-\beta}\Big) \subset \bigcup_{m=2}^{\ell} \Big\{ s^{(k)},\ \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} = 1 \Big\}$$

Since $e^{-\sqrt{k}} = \rho \beta^{\ell+1} (1-\beta)^{-1}$ for $\ell \stackrel{\text{def}}{=} \sqrt{k} / \ln\left(\beta^{-1}\right) + \ln\left(\rho^{-1}(1-\beta)\right) / \ln \beta$,

$$E_k \subset \bigcup_{r=1}^{k} \bigcup_{m=2}^{\lfloor \ell \rfloor + 1} \Big\{ s^{(k)},\ \mathbb{1}_{\{s_r \dots s_m = s_{r-m+1} \dots s_1\}} = 1 \Big\},$$

so that the number of words $s^{(k)}$ such that $q_k(s) \geq z_k$ is bounded above by

$$\# E_k \leq \sum_{r=1}^{k} \sum_{m=2}^{\lfloor \ell \rfloor + 1} 4^{m-1} \in O(k 4^{\sqrt{k} / \ln\left(\beta^{-1}\right)}. \tag{25}$$

Putting (24) and (25) together is sufficient to show that the contribution of prefixes $s^{(k)}$ such that $q_k(s) \geq z_k$ to the sum (13), namely

$$\sum_{s^{(k)} \in \mathcal{A}^k,\ q_k(s) \geq z_k} t^{-n} \prod_{r=1}^{k} \mathbb{E}\big[t^{Z_r(s)}\big],$$

is the general term of a convergent series too.

Finally, the whole sum (13) is the general term of a convergent series, which completes the proof of the inequality

$$\limsup_{n \to \infty} \frac{\mathcal{L}_n}{\ln n} \leq \frac{1}{h_-} \quad \text{a.s.}$$

□

4. Insertion depth

This section is devoted to the asymptotic behaviour of the insertion depth denoted by D_n and to the length of a path randomly and uniformly chosen denoted by M_n (see section 2). D_n is defined as the length of the path leading to the node where $W(n)$ is inserted. In other words, D_n is the amount of digits to be checked before the position of $W(n)$ is found. Theorem 3.1 immediately implies a first asymptotic result on D_n. Indeed, $D_n = \ell_n$ whenever $\ell_{n+1} > \ell_n$, which happens infinitely often a.s., since $\lim_{n\to\infty} \ell_n = \infty$ a.s. Hence,

$$\liminf_{n\to\infty} \frac{D_n}{\ln n} = \liminf_{n\to\infty} \frac{\ell_n}{\ln n} = \frac{1}{h_+} \quad \text{a.s.}$$

Similarly, $D_n = \mathcal{L}_n$ whenever $\mathcal{L}_{n+1} > \mathcal{L}_n$, and hence

$$\limsup_{n\to\infty} \frac{D_n}{\ln n} = \limsup_{n\to\infty} \frac{\mathcal{L}_n}{\ln n} = \frac{1}{h_-} \quad \text{a.s.}$$

Theorem 4.1 states full convergence in probability of these random variables to the constant $1/h$.

Theorem 4.1.

$$\frac{D_n}{\ln n} \xrightarrow[n\to\infty]{\text{P}} \frac{1}{h} \quad \textit{and} \quad \lim_{n\to\infty} \frac{M_n}{\ln n} \xrightarrow[n\to\infty]{\text{P}} \frac{1}{h}.$$

Remark 4.2. For an i.i.d. sequence $U = U_1U_2\ldots$, in the case when the random variables U_i are not uniformly distributed in $\{A, C, G, T\}$, Theorem 4.1 implies that $\frac{D_n}{\ln n}$ does not converge a.s. because

$$\limsup_{n\to\infty} \frac{D_n}{\ln n} \geq \frac{1}{h} > \frac{1}{h_+} = \liminf_{n\to\infty} \frac{D_n}{\ln n}.$$

Proof of Theorem 4.1. It suffices to consider D_n since, by definition of M_n,

$$\mathbb{P}(M_n = r) = \frac{1}{n} \sum_{\nu=1}^{n} \mathbb{P}(D_\nu = r).$$

Let $\varepsilon > 0$. To prove Theorem 4.1, we get the convergence $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$, where

$$A_n \stackrel{\text{def}}{=} \Big\{ U \in \mathcal{A}^{\mathbb{N}},\ \Big|\frac{D_n}{\ln n} - \frac{1}{h}\Big| \geq \frac{\varepsilon}{h} \Big\},$$

by using the obvious decomposition

$$\mathbb{P}(A_n) = \mathbb{P}\left(\frac{D_n}{\ln n} \geq \frac{1+\varepsilon}{h}\right) + \mathbb{P}\left(\frac{D_n}{\ln n} \leq \frac{1-\varepsilon}{h}\right).$$

• Because of X_n's definition (3),

$$D_n = X_{n-1}\big(W(n)\big) + 1$$

so that the duality (4) between $X_n(s)$ and $T_k(s)$ implies that

$$\mathbb{P}\left(\frac{D_n}{\ln n} \geq \frac{1+\varepsilon}{h}\right) \leq \mathbb{P}\Big(X_{n-1}\big(W(n)\big) \geq k-1\Big) \leq \mathbb{P}\Big(T_{k-1}\big(W(n)\big) \leq n-1\Big) \tag{26}$$

with $k \stackrel{\text{def}}{=} \lfloor \frac{1+\varepsilon}{h} \ln n \rfloor$. Furthermore,

$$\mathbb{P}\Big(T_{k-1}\big(W(n)\big) \leq n-1\Big) \leq \mathbb{P}\Big(\big\{T_{k-1}\big(W(n)\big) \leq n-1\big\} \cap B_{n,k_0}\Big) + \mathbb{P}\Big(B^c_{n,k_0}\Big)$$

where B_{n,k_0} is defined, for any $k_0 \leq n$, by

$$B_{n,k_0} \stackrel{\text{def}}{=} \bigcap_{k_0 \leq j \leq n} \Big\{U \in \mathcal{A}^{\mathbb{N}},\ \Big|\frac{1}{j}\ln\Big(\frac{1}{p\big(W(n)^{(j)}\big)}\Big) - h\Big| \leq \varepsilon^2 h\Big\}.$$

Since the sequence U is stationary, $\mathbb{P}\big(W(n)^{(j)}\big) = \mathbb{P}\big(U^{(j)}\big)$ so that Ergodic Theorem implies

$$\lim_{j\to\infty} \frac{1}{j} \ln\left(\frac{1}{p\big(W(n)^{(j)}\big)}\right) = h \qquad \text{a.s.}$$

which leads to $\mathbb{P}\big(B_{n,k_0}\big) = 1$ when both k_0 and n are large enough. If $\mathcal{S}_{n,k_0}$ denotes the set of words

$$\mathcal{S}_{n,k_0} \stackrel{\text{def}}{=} \left\{s^{(n)} \in \mathcal{A}^n,\ \forall j \in \{k_0,\dots,n\}\ \Big|\frac{1}{j}\ln\left(\frac{1}{p\big(s^{(j)}\big)}\right) - h\Big| \leq \varepsilon^2 h\right\},$$

when k_0 and n are large enough,

$$\begin{aligned}\mathbb{P}\Big(T_{k-1}\big(W(n)\big) \leq n-1\Big) &\leq \sum_{s^{(n)}\in\mathcal{S}_{n,k_0}} \mathbb{P}\Big(W(n)^{(n)} = s^{(n)},\ T_{k-1}(s) \leq n-1\Big)\\ &\leq \sum_{s^{(n)}\in\mathcal{S}_{n,k_0}} \mathbb{P}\Big(T_{k-1}(s) \leq n-1\Big).\end{aligned}$$

Such a probability has already been bounded above at the end of Theorem 3.1's proof; similarly,

$$\sum_{s^{(n)}\in\mathcal{S}_{n,k_0}} \mathbb{P}\Big(T_{k-1}(s) \leq n-1\Big) = O\left(n\exp\Big(-\frac{\varepsilon}{1+\varepsilon}n + \frac{\ln 4}{(1-\varepsilon^2)h}\sqrt{n}\Big)\right) \tag{27}$$

so that (26) and (27) show that $\mathbb{P}\left(\frac{D_n}{\ln n} \geq \frac{1+\varepsilon}{h}\right)$ tends to zero when n goes off to infinity.

• Our argument showing that $\mathbb{P}\left(\frac{D_n}{\ln n} \leq \frac{1-\varepsilon}{h}\right)$ tends to zero when n tends to infinity is similar. If now $k \stackrel{\text{def}}{=} \lfloor \frac{1-\varepsilon}{h} \ln n \rfloor$,

$$\mathbb{P}\left(\frac{D_n}{\ln n} \leq \frac{1-\varepsilon}{h}\right) \leq \mathbb{P}\Big(X_{n-1}\big(W(n)\big) \leq k-1\Big) = \mathbb{P}\Big(T_k\big(W(n)\big) \geq n\Big),$$

so that

$$\mathbb{P}\left(\frac{D_n}{\ln n} \leq \frac{1-\varepsilon}{h}\right) \leq \mathbb{P}\Big(\big\{T_k\big(W(n)\big) \geq n\big\} \cap B_{n,k_0}\Big) + \mathbb{P}\big(B^c_{n,k_0}\big).$$

As before, $\mathbb{P}\big(B^c_{n,k_0}\big) = 0$ when k_0 and n are large enough and

$$\begin{aligned}\mathbb{P}\Big(T_k\big(W(n)\big) \geq n\Big) &\leq \sum_{s^{(n)} \in \mathcal{S}_{n,k_0}} \mathbb{P}\Big(W(n)^{(n)} = s^{(n)},\ T_k(s) \geq n\Big) \\ &\leq \sum_{s^{(n)} \in \mathcal{S}_{n,k_0}} \mathbb{P}\Big(T_k(s) \geq n\Big).\end{aligned}$$

Like in the proof of Theorem 3.1, on shows that

$$\sum_{s^{(n)} \in \mathcal{S}_{n,k_0}} \mathbb{P}\Big(T_k(s) \geq n\Big) = O\left(4^n \exp\left(-\kappa n^\theta/2\right)\right)$$

which implies that $\mathbb{P}\left(\frac{D_n}{\ln n} \leq \frac{1-\varepsilon}{h}\right)$ tends to zero when n tends to infinity. The proof of Theorem 4.1 is complete. □

Appendix A. Domain of definition of the generating function $\Phi(s^{(r)}, t)$

A.1. **Proof of Assertion ii)**

There exists a function $K(s_1, s_r, m)$ uniformly bounded by the constant

$$K \stackrel{\text{def}}{=} \sup_{s_1, s_r, m} |K(s_1, s_r, m))|$$

such that

$$Q^m(s_1, s_r) - p\big(s_r\big) = K(s_1, s_r, m)\gamma^m, \tag{28}$$

where γ is the second eigenvalue of the transition matrix. Consequently,

$$\begin{aligned} |\gamma_r(t) - 1| &= \Big|\frac{1-t}{tp\big(s_r\big)} \sum_{m\geq 1} K(s_1, s_r, m)(\gamma t)^m\Big| \\ &\leq \frac{\gamma K}{\min_u p(u)} \frac{|1-t|}{1-\gamma|t|}. \end{aligned}$$

Hence Assertion ii) holds with $\kappa' \stackrel{\text{def}}{=} \gamma K / \min_u p(u)$.

A.2. **Proof of Assertion i)**

On the unit disc $|t| < 1$, the series

$$S(t) \stackrel{\text{def}}{=} \frac{1}{t} \sum_{m\geq 1} Q^m(s_1, s_r) t^m \tag{29}$$

is convergent and one has the decomposition

$$\frac{1-t}{p\big(s_r\big)t} \sum_{m\geq 1} Q^m(s_1, s_r) t^m = 1 + \frac{1-t}{p\big(s_r\big)t} \sum_{m\geq 1} \big[Q^m(s_1, s_r) - p\big(s_r\big)\big] t^m.$$

The function

$$\sum_{m\geq 1} \big[Q^m(s_1, s_r) - p\big(s_r\big)\big] t^m$$

is analytically continuable to the domain $\gamma|t| < 1$, and then the series

$$\frac{1-t}{tp\big(s_r\big)} \sum_{m\geq 1} Q^m(s_1, s_r) t^m$$

converges on the same domain. One has to determine the zeroes of

$$\begin{aligned} D(t) \stackrel{\text{def}}{=} p\big(s^{(r)}\big)t^r &+ \frac{(1-t)p\big(s^{(r)}\big)t^r}{p\big(s_r\big)t} \sum_{z\geq 1} t^z \big[Q^z(s_1, s_r) - p\big(s_r\big)\big] \\ &+ (1-t)\big[1 + \sum_{j=2}^{r} t^{j-1} \frac{p\big(s^{(j)}\big)}{p(s_j)} \mathbb{1}_{\{s_r \ldots s_j = s_{r-j+1} \ldots s_1\}}\big]. \end{aligned}$$

Assuming that some $0 < t < 1$ were a real root of $D(t)$, then

$$\begin{aligned} 0 &< \frac{(1-t)p\big(s^{(r)}\big)t^r}{p\big(s_r\big)t}\sum_{z\geq 1} t^z Q^z(s_1, s_r) \\ &= (t-1)\big[1+\sum_{j=2}^{r} t^{j-1}\frac{p\big(s^{(j)}\big)}{p(s_j)}\mathbb{1}_{\{s_r\ldots s_j = s_{r-j+1}\ldots s_1\}}\big] < 0. \end{aligned}$$

It is thus obvious that there are no real root of $D(t)$ in $]0,1[$. Moreover, one can readily check that 0 and 1 are not zeroes of $D(t)$. We now look for a root of the form $t = 1+\varepsilon$ with $\varepsilon > 0$. Such an ε satisfies

$$\varepsilon = \frac{(1+\varepsilon)^r p\big(s^{(r)}\big)\Big(1 - \dfrac{\varepsilon}{p(s_r)(1+\varepsilon)}\displaystyle\sum_{z\geq 1} t^z[Q^z(s_1,s_r) - p(s_r)]\Big)}{1+\displaystyle\sum_{j=2}^{r}(1+\varepsilon)^{j-1}\frac{p\big(s^{(j)}\big)}{p(s_j)}\mathbb{1}_{\{s_r\ldots s_j = s_{r-j+1}\ldots s_1\}}},$$

so that

$$\varepsilon \geq \kappa p\big(s^{(r)}\big).$$

This implies that $\Phi(s^{(r)}, t)$ is at least defined on $\big[0, 1+\kappa p\big(s^{(r)}\big)\big[$. This implies the result.

References

[1] D. Aldous and P. Shields. A diffusion limit for a class of randomly-growing binary search trees. *Probab. Theory Related Fields*, 79:509–542, 1998.

[2] J.S. Almeida, J.A. Carriço, A. Maretzek, P.A. Noble, and Fletcher M. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, 17(5):429–437, 2001.

[3] Patrick Billingsley. *Ergodic theory and information*. John Wiley & Sons Inc., New York, 1965.

[4] G. Blom and D. Thorburn. How many random digits are required until given sequences are obtained ? *Journal of Applied Probabilities*, 19:518–531, 1982.

[5] P Cénac. Test on the structure of biological sequences via chaos game representation. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 27, 36 pp. (electronic), 2005. ISSN 1544-6115.

[6] P. Cénac, G. Fayolle, and J.M. Lasgouttes. Dynamical systems in the analysis of biological sequences. Technical Report 5351, INRIA, october 2004.

[7] M. Drmota. The variance of the height of digital search trees. *Acta Informatica*, 38:261–276, 2002.

[8] Marie Duflo. *Random Iterative Models*. Springer, 1997.

[9] P. Erdős and P. Révész. On the length of the longest head run. In I. Csizàr and P. Elias, editors, *Topics in Information Theory*, volume 16, pages 219–228, North-Holland, Amsterdam, 1975. Colloq. Math. Soc. Jànos Bolyai.

[10] P. Erdős and P. Révész. On the length of the longest head-run. In *Topics in information theory (Second Colloq., Keszthely, 1975)*, pages 219–228. Colloq. Math. Soc. János Bolyai, Vol. 16. North-Holland, Amsterdam, 1977.

[11] J.C. Fu. Bounds for reliability of large consecutive-k-out-of-n:f system. *IEEE trans. Reliability*, (35):316–319, 1986.

[12] J.C. Fu and M.V. Koutras. Distribution theory of runs: a markov chain approach. *J. Amer. Statist. Soc.*, (89):1050–1058, 1994.

[13] H. Gerber and S. Li. The occurence of sequence patterns in repeated experiments and hitting times in a markov chain. *Stochastic Processes and their Applications*, (11):101–108, 1981.
[14] N. Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res.*, 21(10):2487–2491, 1993.
[15] L. Gordon, M.F. Schilling, and M.S. Waterman. An extreme value theory for long head runs. *Probability Theory and rlated Fields*, (72):279–287, 1986.
[16] H.J. Jeffrey. Chaos Game Representation of gene structure. *Nucleic Acid. Res*, 18:2163–2170, 1990.
[17] Markos V. Koutras. Waiting times and number of appearances of events in a sequence of discrete random variables. In *Advances in combinatorial methods and applications to probability and statistics*, Stat. Ind. Technol., pages 363–384. Birkhäuser Boston, Boston, MA, 1997.
[18] Shuo-Yen Robert Li. A martingale approach to the study of occurrence of sequence patterns in repeated experiments. *Ann. Probab.*, 8(6):1171–1176, 1980. ISSN 0091-1798.
[19] Hosam M. Mahmoud. *Evolution of random search trees*. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons Inc., New York, 1992. ISBN 0-471-53228-2. A Wiley-Interscience Publication.
[20] W Penney. Problem: Penney-ante. *J. Recreational Math.*, 2:241, 1969.
[21] V. Petrov. On the probabilities of large deviations for sums of independent random variables. *Theory Prob. Appl.*, (10):287–298, 1965.
[22] B. Pittel. Asymptotic growth of a class of random trees. *Annals Probab.*, 13:414–427, 1985.
[23] Vladimir Pozdnyakov, Joseph Glaz, Martin Kulldorff, and J. Michael Steele. A martingale approach to scan statistics. *Ann. Inst. Statist. Math.*, 57(1):21–37, 2005. ISSN 0020-3157.
[24] M. Régnier. A unified approach to word occurence probabilities. *Discrete Applied Mathematics*, 104:259–280, 2000.
[25] G. Reinert, S. Schbath, and M.S. Waterman. Probabilistic and statistical properties of words: An overview. *Journal of Computational Biology*, 7(1/2):1–46, 2000.
[26] S. Robin and J.J. Daudin. Exact distribution of word occurences in a random sequence of letters. *J. Appl. Prob.*, 36:179–193, 1999.
[27] A. Roy, C. Raychaudhury, and A. Nandy. Novel techniques of graphical representation and analysis of DNA sequences - A review. *J. Biosci.*, 23(1):55–71, 1998.
[28] S.S. Samarova. On the length of the longest head-run for a markov chain with two states. *Theory of probability and its applications*, 26(3):498–509, 1981.
[29] V. Stefanov and Anthony G Pakes. Explicit distributional results in pattern formation. *Annals of Applied Probabilities*, 7:666–678, 1997.
[30] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X; 0-521-40605-6.

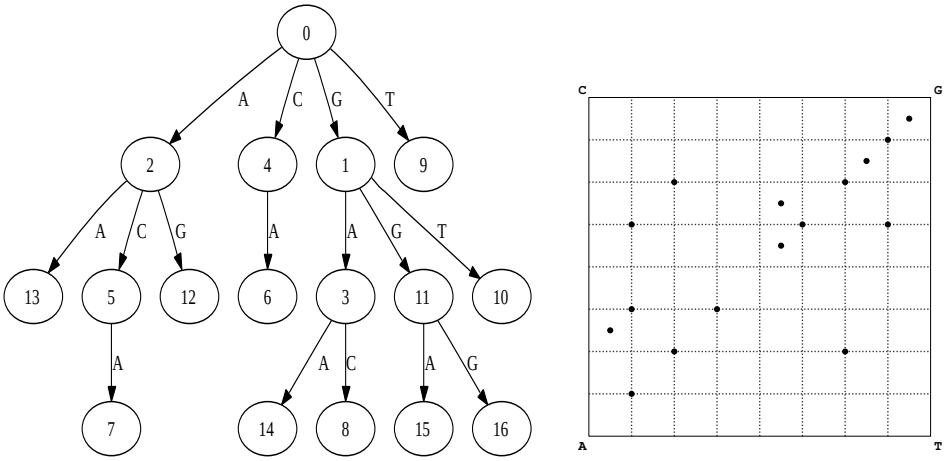

FIGURE 3. Representation of 16 nucleotides of *Mus Musculus* GAGCACAGTG-GAAGGG in the CGR-tree (on the left) and in the "historyless representation" (on the right).

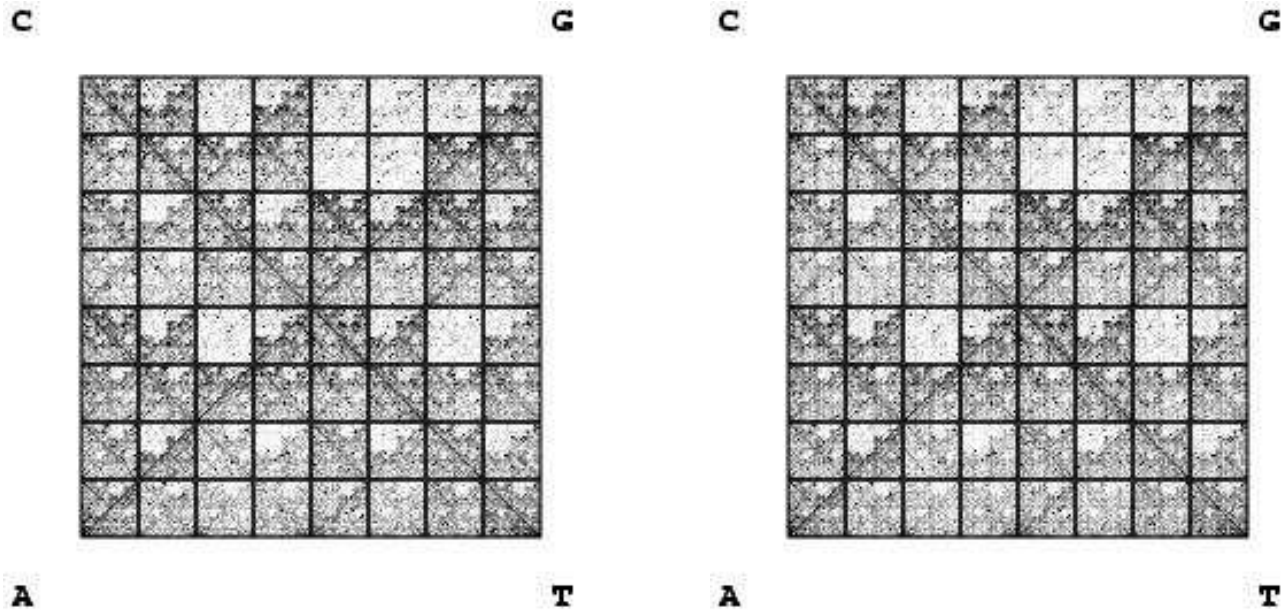

FIGURE 4. Chaos Game Representation (on the left) and historyless representation (on the right) of the first 400000 nucleotides of Chromosome 2 of *Homo Sapiens*.

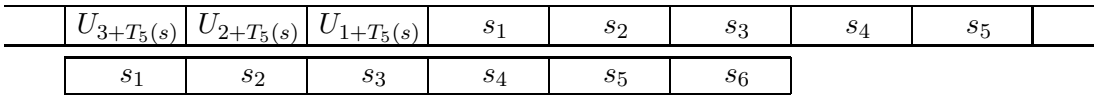

FIGURE 5. How overlapping intervenes in $Z_r(s)$' definition. In this example, one takes $r = 6$. In the random sequence, prefix $s^{(6)}$ can occur starting from $U_{3+T_5(s)}$ only if $s_1s_2s_3 = s_4s_5s_6$.

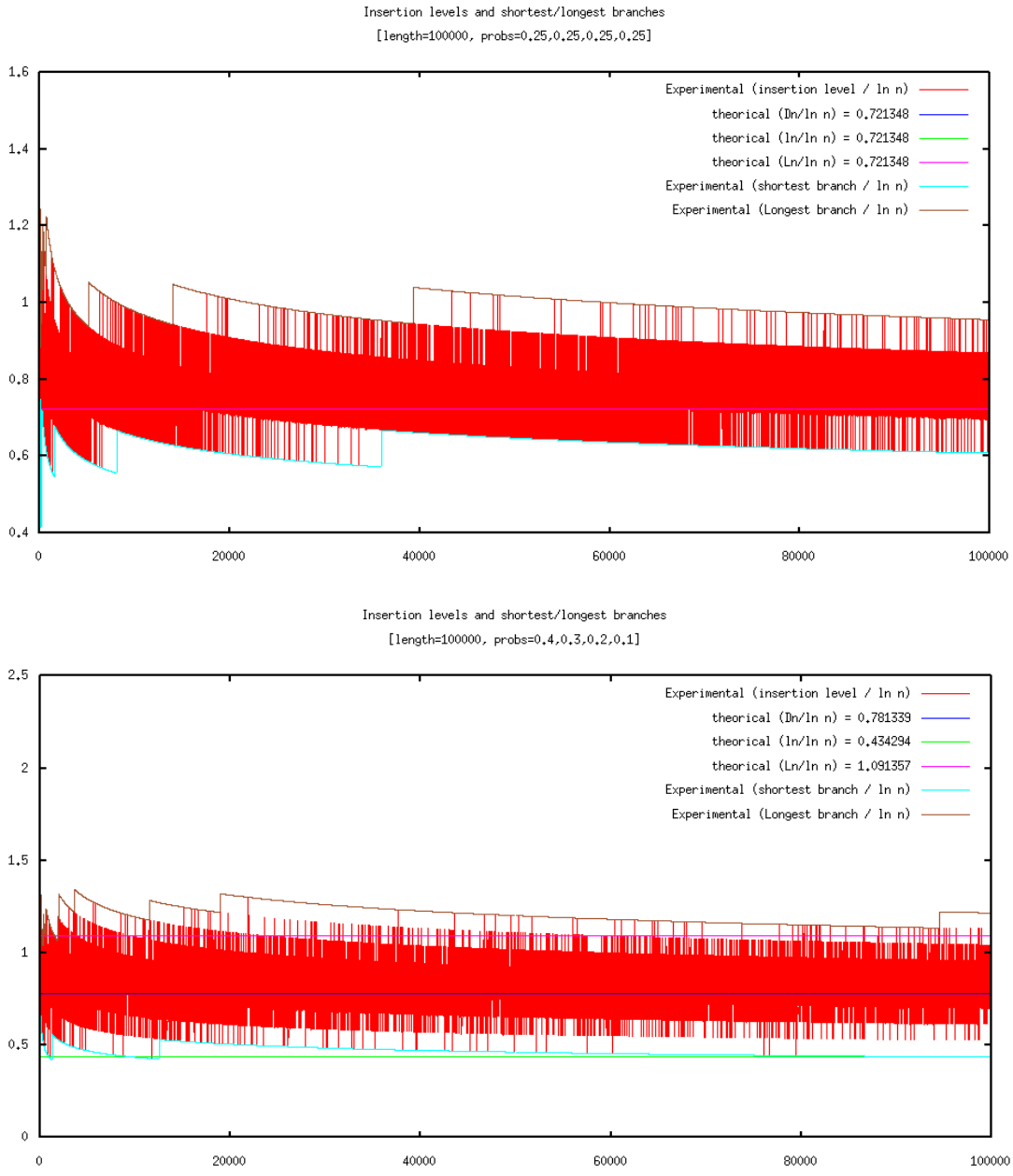

FIGURE 6. Simulations of two random sequences. On the first graphic, letters of the sequence are i.i.d. and equally likely distributed; on the second one, i.i.d. letters have probabilities $(p_A, p_C, p_G, p_T) = (0.4, 0.3, 0.2, 0.1)$. On the x-axis, number n of inserted letters; on the y-axis, normalized insertion depth $D_n / \ln n$ (oscillating curve), lengths of the shortest and of the longest branch (regular "under" and "upper envelops"). The horizontal lines correspond to the constant limits of these three random variables (on the first graph, these three limits have the same value).

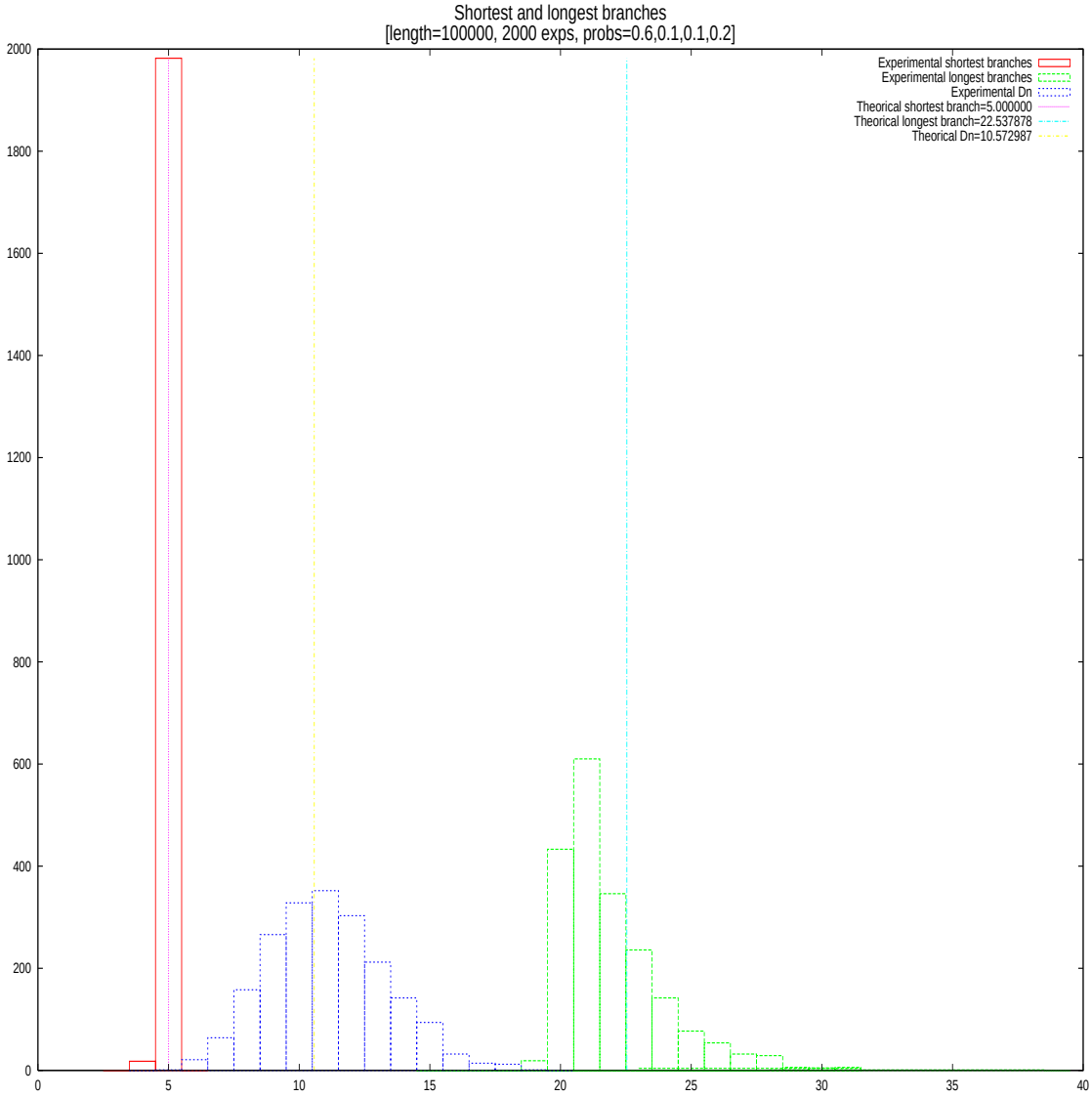

FIGURE 7. Simulations of 2000 sequences of 100,000 i.i.d. letters. On the left, histogram of shortest branches; in the middle, histogram of insertion depth of the last inserted word; on the right, histogram of longest branches. Vertical lines are their expected values, namely $\ln(10^5) \times \ell$ where ℓ respectively equals the limit of $\ell_n/\ln n$, $D_n/\ln n$ and $\mathcal{L}_n/\ln n$.