Digital Search Trees and Chaos Game Representation

Peggy Cénac, Brigitte Chauvin, Stéphane Ginouillac, Nicolas Pouyanne

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Digital Search Trees and Chaos Game Representation

Peggy Cénac — Brigitte Chauvin — Stéphane Ginouillac — Nicolas Pouyanne

N° 5856

March 2006

Thème BIO

ISSN 0249-6399 ISRN INRIA/RR--5856--FR+ENG

Digital Search Trees and Chaos Game Representation

Peggy Cénac*, Brigitte Chauvin†, Stéphane Ginouillac†, Nicolas Pouyanne†

Thème BIO — Systèmes biologiques
Projet Preval

Rapport de recherche n° 5856 — March 2006 — 27 pages

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 length of branches for the CGR-tree obtained from the suffixes of a reversed i.i.d. or Markovian sequence As a by-product, asymptotic results on the length of longest runs in a Markovian sequence are obtained

Key-words: Random tree, Digital Search Tree, CGR, lengths of the paths, height, insertion depth, asymptotic growth, strong convergence

* INRIA Domaine de Voluceau B.P.105 78 153 Le Chesnay Cedex (France)
† LAMA Bâtiment Fermat, Université de Versailles F-78035 Versailles

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 —Télécopie : +33 1 39 63 53 30

Arbres digitaux de recherche et systèmes dynamiques pour l'étude de séquences biologiques

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 des propriétés asymptotiques sur la profondeur d'insertion et les longueurs des branches, pour un arbre obtenu à partir des suffixes d'une séquence i.i.d. ou markovienne retournée 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.

Mots-clés : Arbre aéatoire, arbre de recherche digital, longueurs des branches, hauteur, profondeur d'insertion, croissance asymptotique, convergence forte

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. [26]). The *Chaos Game Representation* (CGR) provides both a graphical representation and a storage tool. From 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. Cénac [5], Cénac et al. [6] study the CGR with an extension of word-counting based methods of analysis. Jeffrey [15] was the first to apply this iterative method to DNA sequences. 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 $(\mathcal{X}_i)_{i\geq 0}$ of points in an appropriate compact subset S of $\mathbb{R}^d$ defined by

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

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 $\mathcal{X}_{i+1}$ is the middle of the segment between $\mathcal{X}_i$ and the square's vertex $\ell_{U_{i+1}}$:

$$\mathcal{X}_{i+1} = \frac{\mathcal{X}_i + \ell_{U_{i+1}}}{2},$$

or, equivalently,

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

Figure 1 represents the construction of the word ATGCGAGTGT.

With each deterministic word $w = u_1 \dots u_n$, we associate the half-opened sub-square 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 $n \geq |w|$, 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, Sw contains all the successive words from the sequence having w as a suffix. See Figure 1 for an example with three-letter subwords. This provides tables of word frequencies (see Goldman [13]). 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 $\mathcal{X}_i$ contains the historical information of the whole sequence $\mathcal{X}_1, \dots \mathcal{X}_i$.* Indeed, notice first that, by construction, $\mathcal{X}_i \in Su$ with $U_i = u$; the whole sequence is now given by the inductive formula $\mathcal{X}_{i-1} = 2\mathcal{X}_i - \ell_{U_i}$.

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

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

as follows in the complete infinite tree. The DNA sequence U is read from left to right. 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 $W(n) = U_n \dots U_1$ 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 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.

Proposition 1.1. *The CGR-tree of a random sequence $U = U_1 U_2 \dots$ is a digital search tree, obtained by insertion in a quaternary tree of the successive reversed prefixes U_1, $U_2 U_1$, $U_3 U_2 U_1$, ... of the sequence.*

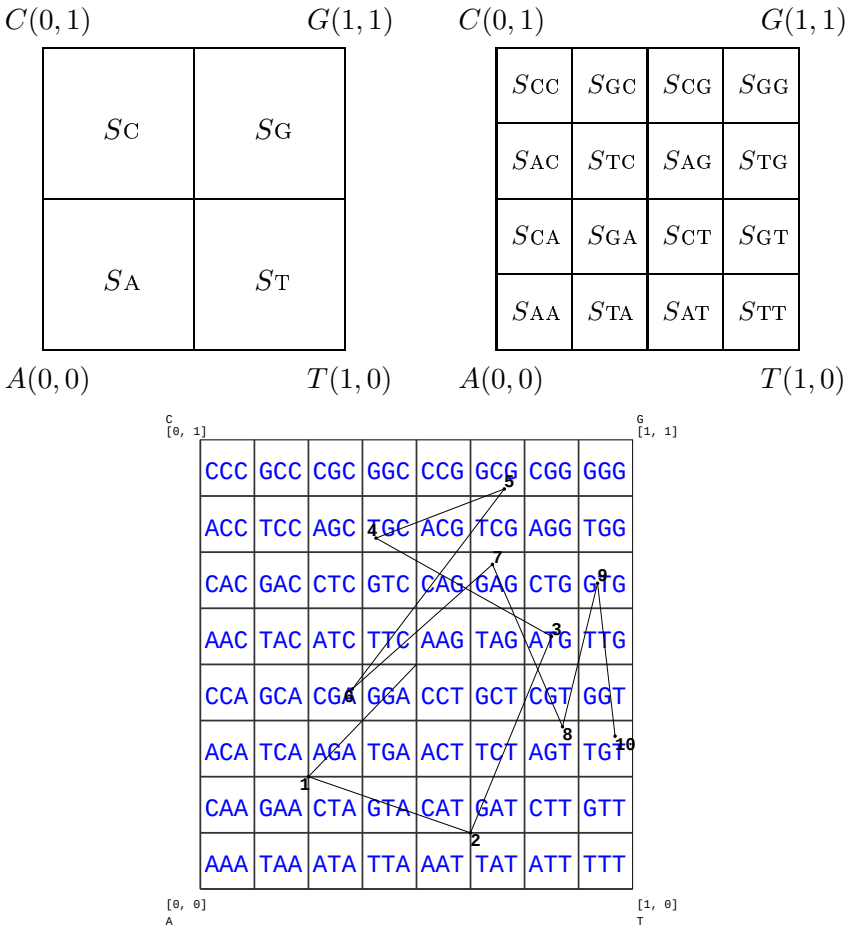

Fig. 1: Chaos Game Representation of the first 10 nucleotides of the *E. Coli* threonine gene thrA: ATGCGAGTGT. The coordinates for each nucleotide are calculated recursively using $(0.5, 0.5)$ as starting position. The sequence is read from left to right. Point number 3 corresponds to the first 3-letter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

Figure 2 represents for instance the beginning of the construction of the CGR-tree of a sequence $GAGCACAGTGGAAGGG$. In this figure, each node has been labelled by its order of insertion to make the example more readable.

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).

Remarque 1. 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 $w = U_1 \dots U_d$), we associate the center of the corresponding square Sw,

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

For example, Figure 2 shows this so-called "*normalized CGR representation*" for the particular word $GAGCACAGTGGAAGGG$. Moreover Figure 3 enables us to qualitatively compare the original and the normalized CGR representations on an example.

Several results are known (see chap. 6 in Mahmoud [18]), 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 [18]. 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 alphabet with m letters, with nonsymmetric (i.e. non equal probabilities on the letters) i.i.d or Markovian sources, Pittel [21] gets several results on the insertion depth and on the height.

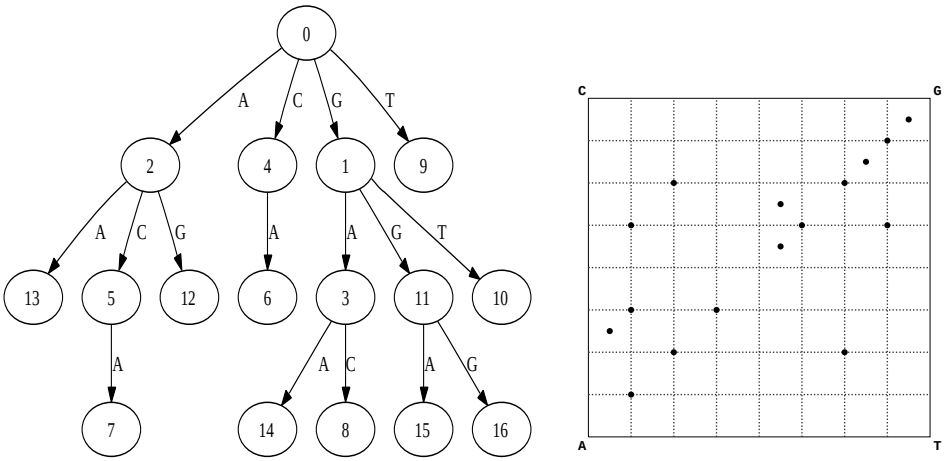

Fig. 2: Representation of 16 nucleotides of *Mus Musculus* GAGCACAGTG-GAAGGG in the CGR-tree (on the left) and in the "normalized" CGR (on the right).

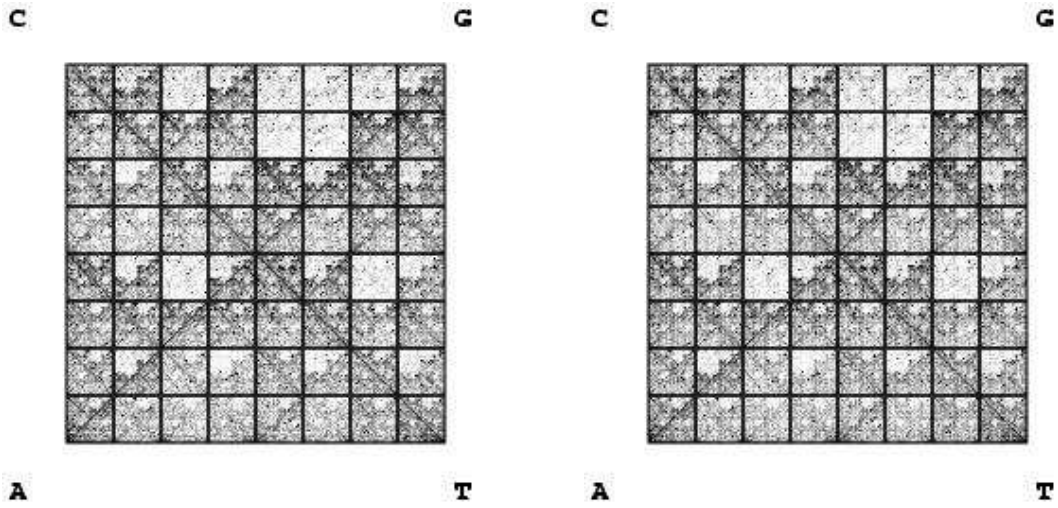

Fig. 3: Chaos Game Representation (on the left) and "normalized" CGR (on the right) of the first 400000 nucleotides of Chromosome 2 of *Homo Sapiens*.

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 [25]. Several studies in this domain are based on generating functions, for example Régnier [23], Reinert et al. [24], Stefanov and Pakes [28]. Nonetheless, other approaches are considered: one of the more general techniques is the Markov chain embedding method introduced by Fu [10] and further developped by Fu and Koutras [11], Koutras [16]. A martingale approach (see Gerber and Li [12], Li [17], Williams [29]) is an alternative to the Markov chain embedding method to solve problems around Penney [19] Game. These two approaches are compared in Pozdnyakov et al. [22]. 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.

Moreover, 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 [8] establish almost sure results about the growth of the longest run. These results are extended to Markov chains in Samarova [27], and Gordon et al. [14] 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 ${}^tQ$ (where ${}^tQ$ denotes the transposed matrix of Q) and invariant measure as initial distribution.

For any deterministic infinite sequence s, let us denote by $s^{(j)}$ the word formed by the j first letters of s, that is to say $s^{(j)} \stackrel{\text{def}}{=} s_1 \dots s_j$, where s_i denotes the i^{th} letter of s. Let $p(s^{(j)})$ denote $p(s^{(j)}) \stackrel{\text{def}}{=} \mathbb{P}(U_1 = s_j, \dots, U_j = s_1)$. The need for reversing the

word $s^{(j)}$ comes from the construction of the CGR-tree, based on reversed sequences (1). One can remark that for a single-letter word u, $p(u)$ denotes the invariant probability of the letter u.

Moreover, we define the constants

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

Due to an argument of sub-additivity (see Pittel [21]), these limits are well defined (in fact, in a more general than Markovian sequences framework). Moreover, Pittel shows that there exist 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 $j \geq 1$, the notation $\mathcal{T}_j \stackrel{\text{def}}{=} \mathcal{T}_j(W)$ denotes the finite tree with j nodes (without counting the root), built from the first j sequences $W(1), \ldots, W(j)$, which are the successive reversed prefixes of the sequence U_n, as defined in (1). $\mathcal{T}_0$ denotes the tree reduced to the root. In particular, the random trees are increasing: $\mathcal{T}_0 \subset \mathcal{T}_1 \ldots \subset \mathcal{T}_j \subset \ldots \subset \mathcal{T}$.

Let us define ℓ_n (resp. $\mathcal{L}_n$) as the length of the shortest path (resp. the longest) from the root to a feasible external node of the tree $\mathcal{T}_{n-1}(w)$. 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 $j \geq 0$,

$$X_j(s) \stackrel{\text{def}}{=} \begin{cases} 0 \text{ if } s_1 \text{ is not in } \mathcal{T}_j \\ \max\{k : \text{word } s^{(k)} \text{ is already inserted in } \mathcal{T}_j\} \end{cases}$$

(notice that $X_0(s) = 0$). For any $k \geq 0$, $T_k(s)$ denotes the size of the first tree where $s^{(k)}$ is inserted:

$$T_k(s) \stackrel{\text{def}}{=} \min\{j : X_j(s) = k\}$$

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

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

$$\{X_j(s) \geq k\} = \{T_k(s) \leq j\} \tag{3}$$

and consequently, $\{T_k(s) = j\} \subset \{X_j(s) = k\}$. Moreover, the random variable $T_k(s)$ can be decomposed as follows,

$$T_k(s) = \sum_{r=1}^{k} Z_r(s), \tag{4}$$

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 j of the first occurence of $s^{(r)}$ in the sequence

$$\dots U_{j+T_{r-1}(s)} U_{j-1+T_{r-1}(s)} \dots U_{1+T_{r-1}(s)} s^{(r-1)},$$

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

$$Z_r(s) = \min\{j \geq 1 \big| U_{j+T_{r-1}(s)} \dots U_{j+T_{r-1}(s)-r+1} = s_1 \dots s_r\}.$$

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

$$\dots U_{j+T_{r-1}(s)} U_{j-1+T_{r-1}(s)} \dots U_{1+T_{r-1}(s)},$$

that is to say

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

Due to the possible overlapping between the prefixes of $s^{(r-1)}$ and the suffixes of $s^{(r)}$, we have the inequality $Z_r(s) \leq Y_r(s)$. 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, with transition matrix ${}^tQ$ and with initial distribution the 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 [25]:

$$\Phi(s^{(r)}, t) = \Big(\gamma_r(t) + (1-t)\delta_r(t^{-1})\Big)^{-1}, \tag{5}$$

where the functions γ and δ are respectively defined as

$$\gamma_r(t) \stackrel{\text{def}}{=} \frac{1-t}{tp(s_r)} \sum_{m\geq 1} Q^m(s_1, s_r)t^m, \quad \delta_r(t^{-1}) \stackrel{\text{def}}{=} \sum_{m=1}^{r} \frac{\mathbb{1}_{\{s_r \ldots s_{r-m+1}=s_m \ldots s_1\}}}{t^m p(s^{(m)})}, \qquad (6)$$

and where $Q^m(u,v)$ denotes the transition probability from u to v in m steps. In Appendix A we study the domain of definition of $\Phi(s^{(r)}, t)$ when t is in a neighbourhood of 1.

Remarque 2. 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 2.1. *(i) The generating function of $Y_r(s)$ is at least defined on $[0, 1+\kappa p(s^{(r)})[$, where κ is a constant independent of r and s.*

(ii) Let γ denote the second largest eigenvalue of the transition matrix Q. For all $t \in [0, \gamma^{-1}[$,

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

where κ' is another constant independent of r and s.

(iii)

$$\mathbb{E}\big[Y_r(s)\big] \leq \sum_{j=1}^{r} \frac{\mathbb{1}_{\{s_r \ldots s_{r-j+1}=s_j \ldots s_1\}}}{p(s^{(j)})} + M,$$

where M is independent of r and s.

Proof. The proof of Proposition 2.1 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.

Théorème 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:

$$\ell_n = 1 + \min_{s\in\mathcal{A}^\infty} X_{n-1}(s), \quad \text{and} \quad \mathcal{L}_n = 1 + \max_{s\in\mathcal{A}^\infty} X_{n-1}(s). \tag{7}$$

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.

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

$$\lim_{n\longrightarrow+\infty} \frac{1}{n} \ln\left(\frac{1}{p\big(s^{(n)}\big)}\right) \stackrel{\text{def}}{=} h(s) > 0. \tag{8}$$

Then

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

Remarque 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 [8], Erdős and Révész [9], Petrov [20] 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.

Proof of Lemma 3.2. Since $X_n(s) = k$ for $n = T_k(s)$ (see Remark 3), by monotonicity arguments, it is sufficient to prove that

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

Obviously from the definition of $T_k(s)$, we have $\mathbb{P}(T_k(s) = j) \leq p\big(s^{(k)}\big)$. Let $0 < \varepsilon < 1$. Roughly,

$$\begin{aligned}\mathbb{P}\big(\ln T_k(s) < (1-\varepsilon)kh(s)\big) &= \mathbb{P}\big(T_k(s) < \exp((1-\varepsilon)kh(s))\big)\\ &\leq \exp\big((1-\varepsilon)kh(s)\big)p\big(s^{(k)}\big)\end{aligned}$$

Moreover, assumption (8) implies that there exists a constant c_1 depending on ε such that, for all $k \geq 1$,

$$p(s^{(k)}) \leq c_1 \exp\big(-(1-\varepsilon^2)kh(s)\big)$$

and therefore

$$\mathbb{P}\big(\ln T_k(s) < (1-\varepsilon)kh(s)\big) \leq c_1 \exp\big(-kh(s)(-\varepsilon^2+\varepsilon)\big),$$

which is the general term of a convergent series when ε is small enough. We deduce from Borel-Cantelli Lemma that almost surely

$$\liminf_{k\to\infty} \frac{\ln T_k(s)}{k} \geq (1-\varepsilon)h(s).$$

Conversely, Markov inequality yields

$$\mathbb{P}\big(\ln T_k(s) > (1+\varepsilon)kh(s)\big) \leq e^{-(1+\varepsilon)kh(s)}\mathbb{E}[T_k(s)].$$

By the decomposition (4) and the inequality $Z_r(s) \leq Y_r(s)$, one has

$$\mathbb{E}[T_k(s)] \leq \sum_{r=1}^{k} \mathbb{E}[Y_r(s)].$$

>From assertion iii) of Proposition 2.1, and since $p(s^{(j)}) \geq p(s^{(r)})$, for any $j \leq r$,

$$\mathbb{E}[T_k(s)] \leq \left(\sum_{r=1}^{k} \frac{r}{p(s^{(r)})}\right) + kM \leq \frac{1}{p\big(s^{(k)}\big)}\frac{k(k+1)}{2} + kM,$$

and finally,

$$\mathbb{P}\Big(\ln T_k(s) > (1+\varepsilon)kh(s)\Big) \leq \left(\frac{1}{2}\frac{k(k+1)}{p\big(s^{(k)}\big)} + kM\right)e^{-(1+\varepsilon)kh(s)}.$$

>From the assumption (8), there exists a constant c_2 depending on ε such that for all $k \geq 1$,

$$p(s^{(k)}) \geq c_2 \exp\big(-(1+\varepsilon^2)kh(s)\big),$$

which leads to

$$\mathbb{P}\Big(\ln T_k(s) > (1+\varepsilon)kh(s)\Big) \leq M'k^2 e^{-kh(s)\left(\varepsilon-\varepsilon^2\right)}.$$

It is the general term of a convergent series as soon as $\varepsilon < 1$. Consequently, due to Borel-Cantelli Lemma,

$$\limsup_{k\to\infty} \frac{\ln T_k(s)}{k} \leq (1+\varepsilon)h(s),$$

which concludes the proof of lemma 3.2. □

Proof of Theorem 3.1. It is inspired from Pittel [21]. Clearly the definition given in Equation (7) yields

$$\ell_n \leq 1 + X_{n-1}(s_+) \quad \text{and} \quad \mathcal{L}_n \geq 1 + X_{n-1}(s_-)$$

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

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

- *Proof for ℓ_n:*
  For any integer r, we have the inequalities

$$\mathbb{P}(\ell_n \leq r) \leq \sum_{s^{(r)}\in\mathcal{A}^r} \mathbb{P}(X_{n-1}(s) \leq r-1) = \sum_{s^{(r)}\in\mathcal{A}^r} \mathbb{P}(T_r(s) \geq n), \tag{9}$$

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 for $t \in [1, 1+\kappa p(s^{(r)})]$, the generating functions $\Phi(s^j, t)$ are defined for any $j \leq r$ (see Assertion i) in Proposition 2.1), each term of the sum (9) can be controlled by

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

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 (6), we deduce from (5) and from Assertion ii) of Proposition 2.1 that

$$\mathbb{P}(T_r(s) \geq n) \leq t^{-n}\prod_{j=1}^{r}\left(1 + (1-t)\left(\sum_{\nu=1}^{j} \frac{1}{t^\nu p(s^{(\nu)})} + \frac{\kappa'}{1-\gamma t}\right)\right)^{-1}.$$

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

$$\mathbb{P}(T_r(s) \geq n) \leq t^{-n}\prod_{j=1}^{r}\left(1 + (1-t)\left(\frac{1-(\alpha t)^{-j}}{c_2(\alpha t - 1)} + \frac{\kappa'}{1-\gamma t}\right)\right)^{-1}.$$

Choosing $t = 1 + c_2\kappa\alpha^r$, where κ is the constant defined in Assertion i) of Proposition 2.1, one gets

$$\mathbb{P}(T_r(s) \geq n) \leq t^{-n} \prod_{j=1}^{r} \left(1 - \kappa\alpha^{r-j} \frac{\alpha^j - (1 + c_2\kappa\alpha^r)^{-j}}{\alpha(1 + c_2\kappa\alpha^r) - 1} - \frac{\alpha^r c_2\kappa\kappa'}{1 - \gamma(1 + c_2\kappa\alpha^r)}\right)^{-1}.$$

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

$$\mathbb{P}(T_r(s) \geq n) \leq (1 + c_2\kappa\alpha^r)^{-n} L \prod_{j=1}^{r} \Big(1 - \lambda\alpha^{r-j}\Big)^{-1}.$$

In addition, the product can be bounded above by

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

Consequently,

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

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 (9) that

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

which is the general term of a convergent series. Once again, Borel-Cantelli Lemma applies and

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

- *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

$$\liminf_{k\to\infty} \min_s \frac{\ln T_k(s)}{k} \geq h_- \quad \text{a.s.}$$

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

$$\mathbb{P}\big(\min_s T_k(s) < \exp^{kh_-(1-\varepsilon)}\big)$$

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

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

Choosing $t \in [0,1]$, this implies that

$$\mathbb{P}\big(T_k(s) < \exp^{kh_-(1-\varepsilon)}\big) \leq \mathbb{P}\big(t^{T_k(s)} > t^n\big),$$

where $n \stackrel{\text{def}}{=} \exp(kh_-(1-\varepsilon))$. The decomposition (4), together with the independence of the $Z_j(s)$ for $1 \leq j \leq k$, yield

$$\mathbb{P}\big(t^{T_k(s)} > t^n\big) \leq t^{-n} \prod_{j=1}^{k} \mathbb{E}\big[t^{Z_j(s)}\big].$$

In order to bound above the term t^n (with $0 < t < 1$), let $t \stackrel{\text{def}}{=} (1+c/n)^{-1}$ be chosen, where c is any constant such that, for all $j \geq 1$,

$$p\big(s^{(j)}\big) \leq c\beta^j, \quad \text{where} \quad \beta \stackrel{\text{def}}{=} \exp(-(1-\varepsilon^2)h_-). \tag{10}$$

The generating function of $Z_j(s)$ is given by Robin and Daudin [25] and strongly depends on the overlapping structure of the word $s^{(j)}$. For $0 < t < 1$, this function is well defined and is given by (see Assertion i) of Proposition 2.1)

$$\mathbb{E}\big[t^{Z_j(s)}\big] = 1 - \frac{(1-t)}{t^j p\big(s_j\big)\big(\gamma_j(t) + (1-t)\delta_j(t^{-1})\big)},$$

where $\gamma_j(t)$ and $\delta_j(t)$ are defined in (6). Moreover, from Assertion ii) of Proposition 2.1, it is obvious that there exists a constant θ independent of j and s such that, for $t = (1+1/n)^{-1}$,

$$\gamma_j(t) \leq 1 + \theta(1-t).$$

Since the function $x \mapsto (A+x)/(B+x)$ increases when $B \geq A$, and since inequality (10) holds, the generating function satisfies

$$\mathbb{E}[t^{Z_j(s)}] \leq 1 - \frac{c^{-1}}{\beta^j\Big(\frac{1}{1-t} + \theta\Big) + c^{-1} + q_k(s)}, \tag{11}$$

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

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

(definition of δ_j were given in 6). Whatever the overlapping structure is, $q_k(s)$ is controlled by

$$0 \leq q_k(s) \leq \frac{\beta}{1-\beta}. \tag{12}$$

Thus,

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

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

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

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

$$\prod_{j=1}^{k} \mathbb{E}[t^{Z_j(s)}] \leq C \exp\Big[-\frac{1}{\ln \beta^{-1}} \int_{\beta^k\left((1-t)^{-1}+\theta\right)}^{+\infty} \ln\Big(1 - \frac{c^{-1}}{y + c^{-1} + q_k(s)}\Big)^{-1} \frac{dy}{y}\Big]. \tag{13}$$

If $\mathrm{Li}_2(z) = \sum_{k\geq 1} z^k/k^2$ denotes the classical dilogarithm, one has $\frac{d}{dy}\mathrm{Li}_2(-\frac{v}{y}) = \frac{1}{y}\log(1+v/y)$, which leads to the formula

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

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

$$\mathrm{Li}_2(x) = -\frac{1}{2}\ln^2(-x) + O(1), \tag{14}$$

which yields, under the assumptions $a_k \to 0$ and $q_k(s) \to 0$ when k tends to infinity,

$$\int_{a_k}^{+\infty} \ln\left(1-\frac{c^{-1}}{y+c^{-1}+q_k(s)}\right)^{-1}\frac{dy}{y} = \mathrm{Li}_2\left(-\frac{q_k(s)}{a_k}\right) + \frac{1}{2}\ln^2 a_k + O(\ln a_k).$$

Thus, the behaviour of the integral in (13) depends on the asymptotics of $q_k(s)$.

First let us consider the case of the words $s^{(k)}$ such that $q_k(s) < \exp\bigl(-\sqrt{k}\bigr)$ and let $z_k \stackrel{\text{def}}{=} \exp\bigl(-\sqrt{k}\bigr)$. For such words, the above equality implies

$$\int_{a_k}^{+\infty} \ln\left(1-\frac{c^{-1}}{y+c^{-1}+q_k(s)}\right)^{-1}\frac{dy}{y} \leq \ln a_k \ln z_k - \frac{1}{2}\ln^2 z_k + O(\ln z_k).$$

Consequently, when $a_k = \beta^k\bigl((1-t)^{-1}+\theta\bigr) \sim \exp(-kh_-(\varepsilon-\varepsilon^2))$, one gets

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

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

$$\sum_{s^{(k)}\in\mathcal{A}_k |\ q_k(s)<z_k} \mathbb{P}\bigl(T_k(s) < \exp^{kh_-(1-\varepsilon)}\bigr) \leq 4^k \exp\left[-\frac{\varepsilon}{2(1+\varepsilon)}k^{3/2} + O(k)\right],$$

which is the general term of a convergent series.

It remains to study the case when $q_k(s) \geq z_k$. For these words, let us only consider the inequality (12) and then

$$\prod_{j=1}^{k} \mathbb{E}[t^{Z_j(s)}] \leq C'\exp\left[-\frac{1}{\ln\beta^{-1}}\int_{\beta^k\left((1-t)^{-1}+\kappa\right)}^{+\infty} \ln\Bigl(1-\frac{c^{-1}}{y+c^{-1}+\beta(1-\beta)^{-1}}\Bigr)^{-1}\frac{dy}{y}\right].$$

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

$$\prod_{j=1}^{k} \mathbb{E}[t^{Z_j(s)}] \leq \exp\Big(-\frac{c^{-1}}{\ln \beta^{-1}} k h_- \varepsilon(1-\varepsilon) + o(k)\Big).$$

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

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

Moreover, clearly from the definition of $q_k(s)$,

$$|E_k| \leq \Big| \Big\{ s^{(k)} \Big| \exists j \leq k : \sum_{m=1}^{j-1} \mathbb{1}_{\{s_m \ldots s_1 = s_j \ldots s_{j-m+1}\}} \beta^{j-m} \geq e^{-\sqrt{k}} \Big\} \Big|.$$

Let us define the set, for $j \leq k$

$$S_j(t) \stackrel{\text{def}}{=} \Big\{ s^{(k)} \Big| \sum_{m=1}^{j-1} \mathbb{1}_{\{s_m \ldots s_1 = s_j \ldots s_{j-m+1}\}} \beta^{j-m} \geq t \Big\}.$$

One has the following inclusion,

$$\bigcap_{m=\ell}^{j-1} \Big\{ s^{(k)} \Big| \mathbb{1}_{\{s_m \ldots s_1 = s_j \ldots s_{j-m+1}\}} = 0 \Big\} \subset S_j^c\Big(\frac{\beta^{j-\ell}}{1-\beta}\Big),$$

where the notation B^c denotes the complementary set of B in $\mathcal{A}^k$. Since $e^{-\sqrt{k}} = \frac{\beta^{j-\ell}}{1-\beta}$ for $\ell \stackrel{\text{def}}{=} j - \sqrt{k}/\ln\big(\beta^{-1}\big)$,

$$E_k \subset \bigcup_{j=1}^{k} \bigcup_{m=\ell}^{j-1} \Big\{ s^{(k)} \Big| \mathbb{1}_{\{s_m \ldots s_1 = s_j \ldots s_{j-m+1}\}} = 1. \Big\}.$$

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

$$\big|E_k\big| \leq \sum_{j=1}^{k} \sum_{m=\ell}^{j-1} 4^{j-m} \leq \frac{4}{3} k 4^{\sqrt{k}/\ln\left(\beta^{-1}\right)}.$$

It remains to apply Borel-Cantelli Lemma and

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

□

Remarque 4. For $\tilde{v}$ consisting of repetitions of a single letter v, with $p(v) \stackrel{\text{def}}{=} p$, by use of the same method, one can establish the following inequalities :

$$\begin{aligned}\mathbb{P}(X_n(\tilde{v}) < k) &\leq (1+p^k)^{-(n+1)}\Big(1+p^k\frac{(t_1p)^{-k}-1}{1-t_1p}\Big)^{-1},\\ \mathbb{P}(X_n(\tilde{v}) \geq k) &\leq e\Big(1+\frac{1}{n+1}\frac{(t_2p)^{-k}-1}{1-t_2p}\Big)^{-1},\end{aligned}$$

where $t_1 \stackrel{\text{def}}{=} 1+p^k$ and $t_2 \stackrel{\text{def}}{=} (1+1/n)^{-1}$, for k being such that the generating function $\Phi(s^{(k)}, t_1)$ is defined. Hence one can derive asymptotic properties for longest runs. Moreover, since the maximum G_n of the four longest runs in $U_1 \dots U_n$ is smaller than the height of the CGR-tree $\mathcal{T}_n$, Theorem 3.1 yields

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

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 random 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

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

The following theorem completes the asymptotic behaviour.

Théorème 4.1.

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

Remarque 5. 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 fix $\varepsilon > 0$ and a positive integer j_0. Let us introduce some complementary notations. We prove Theorem 4.1 getting the convergence $\lim_{n\to\infty} \mathbb{P}(A_n) = 0$, where

$$A_n \stackrel{\text{def}}{=} \Big\{U \in \mathcal{A}^\infty \quad : \Big|\frac{D_n}{\ln n} - \frac{1}{h}\Big| \geq \frac{\varepsilon}{h}\Big\}.$$

The set A_n is then decomposed by intersection with the set B_{n,j_0} defined below and its complementary,

$$B_{n,j_0} \stackrel{\text{def}}{=} \bigcap_{j\geq j_0} \Big\{U \in \mathcal{A}^\infty \quad : \Big|\frac{1}{j}\ln\Big(\frac{1}{p\big(W(n)^{(j)}\big)}\Big) - h\Big| \leq \varepsilon^2 h\Big\},$$

and consequently

$$\mathbb{P}\big(A_n\big) \leq \mathbb{P}(A_n \cap B_{n,j_0}) + \mathbb{P}(R_{j_0,\varepsilon}), \tag{15}$$

where

$$\mathbb{P}(R_{j_0,\varepsilon}) = \mathbb{P}\Big(\sup_{j\geq j_0} \Big|\frac{1}{j}\ln\Big(\frac{1}{p\big(U^{(j)}\big)}\Big) - h\Big| \geq \varepsilon^2 h\Big),$$

and $U = U_1, U_2, \ldots$ is the generic process (remember that this process is stationary). By Shannon-McMillan-Breiman Theorem (see for example Billingsley [3]), the second term of (15) is bounded above, since

$$\lim_{j_0\to\infty} \mathbb{P}(R_{j_0,\varepsilon}) = 0 \quad \text{for each } \varepsilon > 0.$$

Moreover, by definition of X_n, the insertion depth can be written

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

Since $X_n(s)$ for $T_k(s)$ are in duality,

$$\mathbb{P}(A_n \cap B_{n,j_0}) \leq \sum_r P_{n,r} \sum_{U^{(r)}} \mathbb{P}(T_r(U) = n),$$

where the first sum ranges over integers r such that $|r/\ln n - 1/h| \geq \varepsilon/h$, and the second sum ranges over prefixes $U^{(r)}$ such that

$$\max\Big\{\Big|\frac{1}{j}\ln\Big(\frac{1}{p\big(U^{(j)}\big)}\Big) - h\Big| : j_0 \leq j \leq r\Big\} \leq \varepsilon^2 h.$$

Each word $U^{(r)}$ of this second sum satisfies, for $1 \leq j \leq r$, and for some constants $c_1 = c_1$ and $c_2 = c_2$ only depending on ε,

$$c_2\alpha_2^j \leq p\big(U^{(j)}\big) \leq c_1\alpha_1^j, \quad \alpha_1 \stackrel{\text{def}}{=} \exp(-(1-\varepsilon^2)h), \quad \alpha_2 \stackrel{\text{def}}{=} \exp(-(1+\varepsilon^2)h).$$

Using the lower (resp. upper) estimate of $p\big(U^{(j)}\big)$ for $r \leq r_2 \stackrel{\text{def}}{=} \lfloor(1-\varepsilon)\frac{\ln n}{h}\rfloor$ (resp. for $r > r_1 \stackrel{\text{def}}{=} \lfloor(1+\varepsilon)\frac{\ln n}{h}\rfloor$) and arguing exactly as in Section 3, one gets

$$\mathbb{P}(A_n \cap B_{n,j_0}) = \mathcal{O}(1)\Big(\exp\big(-Mn^{\varepsilon(1-\varepsilon+\varepsilon^2)}\big) + n^{-\frac{c_1^{-1}}{\ln\alpha_1^{-1}}\varepsilon+o(1)} + e^{-(M'+o(1))(\ln n)^{3/2}}\Big)$$

where M and M' are two positive constants independent on n. □

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

$$Q^m(s_1, s_r) - p\big(s_r\big) = K(s_1, s_r, m)\gamma^m, \tag{16}$$

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

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

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

A.2 Proof of assertion iii)

The derivation of the generating function $\Phi(s^{(j)}, t)$ leads to

$$\mathbb{E}[Y_r(s)] = \sum_{j=1}^{r} \frac{\mathbb{1}_{\{s_r \ldots s_{r-j+1}\}}}{p\big(s^{(j)}\big)} - \gamma_r'(1).$$

Assertion is proved with $M \stackrel{\text{def}}{=} \gamma K/(1-\gamma) \min_u p(u)$.

A.3 Proof of assertion i)

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

$$S(t) \stackrel{\text{def}}{=} \frac{1}{t} \sum_{m\geq 1} Q^m(s_1, s_r) t^m \tag{17}$$

is well defined 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 sum

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

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

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

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

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

It is thus obvious that there are no real root of $D(t)$ in $]0, 1[$. Moreover, we can readily check that 0 and 1 are not zeros of $D(t)$. We now look for a root of the form $t = 1 + \varepsilon$ with $\varepsilon > 0$, that is to say such that

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

We have then the inequality

$$\begin{aligned} \varepsilon &\geq \frac{c p\big(s^{(r)}\big)}{1 + \sum_{k=1}^{r-1} p\big(s^{(r-k)}\big)} \\ &\geq \kappa p\big(s^{(r)}\big). \end{aligned}$$

Finally, the generating function of $Y_r(s)$ denoted by $\Phi(s^{(r)}, t)$ is at least defined on $\big[0, 1 + \kappa p\big(s^{(r)}\big)\big[$.

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] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] H.J. Jeffrey. Chaos Game Representation of gene structure. *Nucleic Acid. Res*, 18:2163–2170, 1990.

[16] 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.

[17] 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.

[18] 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.

[19] W Penney. Problem: Penney-ante. *J. Recreational Math.*, 2:241, 1969.

[20] V. Petrov. On the probabilities of large deviations for sums of independent random variables. *Theory Prob. Appl.*, (10):287–298, 1965.

[21] B. Pittel. Asymptotic growth of a class of random trees. *Annals Probab.*, 13: 414–427, 1985.

[22] 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.

[23] M. Régnier. A unified approach to word occurence probabilities. *Discrete Applied Mathematics*, 104:259–280, 2000.

[24] 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.

[25] S. Robin and J.J. Daudin. Exact distribution of word occurences in a random sequence of letters. *J. Appl. Prob.*, 36:179–193, 1999.

[26] 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.

[27] 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.

[28] V. Stefanov and Anthony G Pakes. Explicit distributional results in pattern formation. *Annals of Applied Probabilities*, 7:666–678, 1997.

[29] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X; 0-521-40605-6.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399