# Accurate prediction of the statistics of repetitions in random sequences: a case study in Archaea genomes Mireille Regnier, Philippe Chassignet 

## To cite this version:

Mireille Regnier, Philippe Chassignet. Accurate prediction of the statistics of repetitions in random sequences: a case study in Archaea genomes. Frontiers in Bioengineering and Biotechnology, 2016, 4, 10.3389/fbioe.2016.00035 . hal-01304366

HAL Id: hal-01304366
https://inria.hal.science/hal-01304366
Submitted on 19 Apr 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Accurate prediction of the statistics of repetitions in random sequences: a case study in Archaea genomes 

Mireille Régnier ${ }^{1,2, *}$ and Philippe Chassignet ${ }^{2}$<br>${ }^{1}$ Inria, France<br>${ }^{2}$ LIX, Ecole Polytechnique, France<br>Correspondence*:<br>Mireille Régnier<br>AMIB, Inria, 1 rue d'Estienne d'Orves, Palaiseau, 91 120, France, mireille.regnier@inria.fr

Repetitive structures in biological sequences: algorithms and applications

## 00ABSTRACT

Repetitive patterns in genomic sequences have a great biological significance and also algorithmic implications. Analytic combinatorics allow to derive formula for the expected length of repetitions in a random sequence. Asymptotic results, that generalize previous works on a binary alphabet, are easily computable. Simulations on random sequences show their accuracy. As an application, the sample case of Archaea genomes illustrates how biological sequences may differ from random sequences.

Keywords: $K$-mers; combinatorics; probability

## 1 INTRODUCTION

This paper provides combinatorial tools to distinguish biologically significant events from random repetitions in sequences. This is a key issue in several genomic problems as many repetitive structures can be found in genomes. One may cite microsatellites, retrotransposons, DNA transposons, long terminal repeats (LTR), long interspersed nuclear elements (LINE), ribosomal DNA, short interspersed nuclear elements (SINE). In Treangen and Salzberg (2012), it is claimed that half of the genome consists of different types of repeats. Knowledge about the length of a maximal repeat is a key issue for assembly, notably the design of algorithms that rely upon de Bruijn graphs. In re-sequencing, it is a common assumption for aligners that any sequenced "read" should map to a single position in a genome : in the ideal case where no sequencing error occurs, this implies that the length of the reads is larger than the length of the maximal repetition. Average lengths of the repeats are given in Gu et al. (2000). Recently, heuristics have been proposed and implemented Devillers and Schbath (2012); Rizk et al. (2013); Chikhi and Medvedev (2014).
It was observed in Jacquet and Szpankowski (1994) that the average length of maximal common prefixes in a random set of $n$ words is asymptotically equivalent to the average length of maximal repetitions in a random sequence of length $n$. The first model is easier to address; therefore, sets of words are considered below in the theoretical analysis. A comparison with the distribution of maximal repetitions in random sequences or real Archaea genomic sequences is presented in Section 3.

## 2 MAIN RESULTS

It is assumed throughout this study that sequences and words are randomly generated according to a biased Bernoulli model on an alphabet of size $V$. Let $p_{1}, \cdots p_{V}$ denote the probabilities of the $V$ characters $\chi_{1}, \cdots, \chi_{V}$.

DEFINITION 2.1. For any $i$ in $\{1, \cdots, V\}$, one notes

$$
\beta_{i}=\log \frac{1}{p_{i}}
$$

Additionnally

$$
\begin{align*}
& p_{\min }=\min \left\{p_{i} ; 1 \leq i \leq V\right\} \quad \text { and } \quad \alpha_{\min }=\frac{1}{\log \frac{1}{p_{\min }}}=\frac{1}{\max \left(\beta_{i}\right)}  \tag{1}\\
& p_{\max }=\max \left\{p_{i} ; 1 \leq i \leq V\right\} \quad \text { and } \quad \alpha_{\max }=\frac{1}{\log \frac{1}{p_{\max }}}=\frac{1}{\min \left(\beta_{i}\right)} \tag{2}
\end{align*}
$$

The two values $\min \left(\beta_{i}\right)$ and $\max \left(\beta_{i}\right)$ are different when the Bernoulli model is non uniform.

### 2.1 ENUMERATION

DEFINITION 2.2. Given $U$ a set of words and an integer $k, k \geq 2$, a unique $k$-mer in $U$ is a word $w \chi_{i}$
where

$$
\begin{align*}
\phi\left(k_{1}, \cdots, k_{V}\right) & =p_{1}^{k_{1}} \cdots p_{V}^{k_{V}}  \tag{4}\\
\psi_{n}\left(k_{1}, \cdots, k_{V}\right) & =\sum_{i=1}^{V} p_{i}\left[\left(1-\phi\left(k_{1}, \cdots, k_{V}\right) p_{i}\right)^{n-1}-\left(1-\phi\left(k_{1}, \cdots, k_{V}\right)\right)^{n-1}\right] \tag{5}
\end{align*}
$$

Proof. A word $w \chi_{i}$ is a unique $(k+1)$-mer iff (i) $w$ has length $k$ and is the prefix of at least two words, including $w \chi_{i}$; (ii) $w \chi_{i}$ is not repeated.

Event (i) has probability $n \phi\left(k_{1}, \cdots, k_{V}\right) p_{i}\left[1-\left(1-\phi\left(k_{1}, \cdots, k_{V}\right)\right)^{n-1}\right]$.
Event (ii), that is a sub-event of (i), has probability $n \phi\left(k_{1}, \cdots, k_{V}\right) p_{i}\left[1-\left(1-\phi\left(k_{1}, \cdots, k_{V}\right) p_{i}\right)^{n-1}\right]$.

### 2.2 A COMBINATORIAL EXPRESSION

DEfinition 2.4. Given a $k$-mer $w$, let $\alpha$ denote $\frac{k}{\log n}$ and $k_{i}$ denote the number of occurrences of character $\chi_{i}$ in $w$. The objective function is

$$
\begin{equation*}
\rho\left(k_{1}, \cdots, k_{V}\right)=\sum_{i=1}^{V} \frac{k_{i}}{k} \beta_{i}-\frac{1}{\alpha} . \tag{6}
\end{equation*}
$$

The character distribution $\left(k_{1}, \cdots, k_{V}\right)$ of a $k$-mer may be viewed as barycentric coordinates for a point $\beta\left(k_{1}, \cdots, k_{V}\right)=\sum_{i=1}^{V} \frac{k_{i}}{k} \beta_{i}$ that lies in $\left[\min \left(\beta_{i}\right) ; \max \left(\beta_{i}\right)\right]=\left[\frac{1}{\alpha_{\max }} ; \frac{1}{\alpha_{\min }}\right]$. The order of $\beta$ points on that interval allows for a classification of $k$-mers that is a key to this study.

Definition 2.5. A k-mer w is said

- a common $k$-mer if $\rho\left(k_{1}, \cdots, k_{V}\right)<0$;
- a transition $k$-mer if $\rho\left(k_{1}, \cdots, k_{V}\right) \geq 0$ and its ancestor is a common $k$-mer;
- $a$ rare $k$-mer , otherwise.

Remark: If $\rho\left(k_{1}, \cdots, k_{V}\right)=0$, the condition on the ancestor is trivially satisfied.
Definition 2.6. Given a set $U$ of $n$ words and an integer $k$, let $D k(n)$ denote the set of character distributions $\left(k_{1}, \cdots, k_{V}\right)$ for rare and transition $k$-mers. Let $E_{k}(n)$ denote the set of character distributions for common $k$-mers.

The set $D_{k}(n)$ is the empty set if $k<\alpha_{\text {min }} \log n$ and is the set of character distributions $\left(k_{1}, \cdots, k_{V}\right)$ if $k>\alpha_{\max } \log n$. Computation of (3) is split among the two sets $D_{k}(n)$ and $E_{k}(n)$. Computations show that the main contribution arises from transition $k$-mers. A probabilistic interpretation will be discussed in 2.4.

Notation: Let $S(k)$ and $T(k)$ be

$$
\begin{align*}
S(k) & =n \sum_{D_{k}(n)}\binom{k}{k_{1} \cdots k_{V}} \phi\left(k_{1}, \cdots, k_{V}\right) \psi_{n}\left(k_{1}, \cdots, k_{V}\right)  \tag{7}\\
T(k) & =n \sum_{E_{k}(n)}\binom{k}{k_{1} \cdots k_{V}} \phi\left(k_{1}, \cdots, k_{V}\right) \psi_{n}\left(k_{1}, \cdots, k_{V}\right) \tag{8}
\end{align*}
$$

So $\mu(n, k)$ rewrites

$$
\begin{equation*}
\mu(n, k)=S(k)+T(k) . \tag{9}
\end{equation*}
$$

These sums $S(k)$ and $T(k)$ can be efficiently computed for moderate $k$, up to a few hundred, approximately. In practice, $\alpha_{\max } \log n$ is below this threshold for the sizes of actual genomes and for their ordinary GC content value. The simulations in Section 3 show that this estimation is rather tight. Behaviour and asymptotic estimates are derived and discussed in the next section.

### 2.3 ASYMPTOTIC ESTIMATES

113 In this section, asymptotic estimates for (3) are derived. First, some characteristic functions are intro114 duced. Then, a Large Deviation Principle applies, that allows to compute asymptotics for the dominating 115 term. Amortized terms are also computed. It is shown in Section 3 that this second order term cannot be neglected in the finite range.

117 Notations For general alphabets, asymptotic behaviour is a function of the solution of an equation and depends on domains whose bounds are defined below.

119 Definition 2.7. Let $\left(p_{i}\right)_{1 \leq i \leq V}$ be a Bernoulli probability distribution. Let $\sigma_{2}$ denote $\sum_{i=1}^{V} p_{i}^{2}$.
120 The fundamental ratio, noted $\tilde{\alpha}$, is $\left(\sum_{i} p_{i} \log \frac{1}{p_{i}}\right)^{-1}$.
121 The transition ratio, noted $\bar{\alpha}$, is $\sigma_{2}\left(\sum_{i} p_{i}^{2} \log \frac{1}{p_{i}}\right)^{-1}$.

Let $\psi$ be the function defined in $\left[\alpha_{\text {min }}, \alpha_{\text {ext }}\right]$ as

$$
\begin{aligned}
\alpha_{\min } \leq \alpha \leq \bar{\alpha} & : \quad \psi(\alpha)=\tau_{\alpha}+\alpha \log \left(\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right) \\
\bar{\alpha} \leq \alpha \quad & : \quad \psi(\alpha)=2-\alpha \log \frac{1}{\sigma_{2}}
\end{aligned}
$$

Proposition 2.1. The following property holds

$$
\alpha_{\min } \leq \tilde{\alpha} \leq \bar{\alpha} \leq \alpha_{\max } \leq \alpha_{e x t}
$$

126 Function $\psi$ increases on $\left[\alpha_{\text {min }}, \tilde{\alpha}\right]$ and decreases on $[\tilde{\alpha}, \infty]$. It satisfies

$$
\begin{equation*}
\psi\left(\alpha_{\min }\right)=\psi\left(\alpha_{e x t}\right)=0 \text { and } \psi(\tilde{\alpha})=1 . \tag{11}
\end{equation*}
$$

127 Remark: Uniqueness of $\tau_{\alpha}$ is shown in Section 4.2. As $\tau_{\tilde{\alpha}}=2, \psi$ is continuous at $\alpha=\tilde{\alpha}$, with $\psi(\bar{\alpha})=$ $2-\bar{\alpha} \log \frac{1}{\sigma_{2}}$.

Asymptotic results

THEOREM 2.2. Given a length $\alpha \log n$, when $n$ tends to $\infty$ the ratio $\frac{\log \mu(n, \alpha \log n)}{\log n}$ satisfies :

$$
\begin{align*}
& 0 \leq \alpha \leq \alpha_{\min } \text { or } \alpha_{e x t} \leq \alpha \quad: \frac{\log \mu(n, \alpha \log n)}{\log n} \leq 0  \tag{12}\\
& \alpha_{\min } \leq \alpha \leq \alpha_{e x t} \quad: \quad \frac{\log \mu(n, \alpha \log n)}{\log n} \sim \psi(\alpha) \tag{13}
\end{align*}
$$

131 Moreover, let $\xi$ be the function defined in $\left[\alpha_{\text {min }}, \alpha_{\text {ext }}\right]$ as $\xi(\alpha)=\frac{\mu(n, \alpha \log n)}{\log n}-\psi(\alpha)$. It satisfies

$$
\begin{gather*}
\alpha_{\min } \leq \alpha \leq \bar{\alpha}: \xi(\alpha) \quad \sim-\frac{V-1}{2} \frac{\log (\alpha \log n)}{\log n}  \tag{14}\\
\bar{\alpha} \leq \alpha \leq \alpha_{e x t}: \xi(\alpha) \sim \frac{\log \left(1-\sigma_{2}\right)}{\log n} \tag{15}
\end{gather*}
$$

$$
\begin{equation*}
\frac{\log T \tilde{(k)}}{k} \sim \max \left\{-\sum_{i=1}^{V} \frac{k_{i}}{k} \log \frac{k_{i}}{k} ; \rho\left(k_{1}, \cdots, k_{V}\right)=0\right\} \tag{16}
\end{equation*}
$$

134 The maximization problem rewrites as

$$
\begin{equation*}
\max \left\{\sum_{i=1}^{V} \theta_{i} \log \frac{1}{\theta_{i}} ; \sum_{i=1}^{V} \theta_{i}=1 ; \sum_{i=1}^{V} \beta_{i} \theta_{i}=\frac{1}{\alpha} ; 0 \leq \theta_{i} \leq 1\right\} \tag{17}
\end{equation*}
$$

The maximum value is $\tau_{\alpha}+\alpha \log \left(\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right)$ that is reached for the $V$-tuple $\left(\theta_{i}=\frac{e^{-\beta_{i} \tau_{\alpha}}}{\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}}\right)_{1 \leq i \leq V}$. $S(k)$ satisfies again a Large Deviation Principle when $\alpha<\bar{\alpha}$, which yields the asymptotic result in this range. For larger $\alpha, S(k)$ is approximately $\left(1-\sigma_{2}\right) n^{1-\alpha \log \frac{1}{\sigma_{2}}}$ that dominates $T(k)$.
Details for the proof, including the short and long lengths, are provided in Section 4.
PROOF. The key to the proof when $\alpha$ ranges in $\left[\alpha_{\min }, \alpha_{\max }\right]$ is that $\psi_{n}\left(k_{1}, \cdots, k_{V}\right)$ is maximal when $\rho\left(k_{1}, \cdots, k_{V}\right)$ is close to 0 . Sum $T(k)$ satisfies a Large Deviation Principle.

Remark: The discussion will depend of the ratio $\alpha=\frac{k}{\log n}$. Possible values for $\alpha$ range over a discrete set as they are constrained to be the ratio of an integer by the log of an integer. An interesting property is that, for any real $\alpha$, the set $T=\{n \in N ; \alpha \log n \in N\}$ is either empty or infinite. Indeed, when $T$ is non-empty, it contains all values $n(\alpha)^{p}$ where $n(\alpha)$ denotes the minimum value of $T$. It is beyond the scope of this paper to establish the number of other possible solutions.

Domains Different domains arise from this Theorem, that were observed in Park et al. (2009). Equalities $\psi\left(\alpha_{\min }\right)=0$ and $\psi(\bar{\alpha})=2-\bar{\alpha} \log \frac{1}{\sigma_{2}}$ show that there is a continuity between domains.
When $\alpha$ lies inside the domain $\left[\alpha_{\min }, \alpha_{e x t}\right]$, the ratio $\frac{\log \mu(n, \alpha \log n)}{\log n}$ is positive and parameters $\mu(n, \alpha \log n)$ are sub-linear in the size $n$ of the text: some $k$-mers -mostly transition $k$-mers- are unique $k$-mers. Observe that the maximum value for $\psi(\alpha)$ is 1 . When the Bernoulli model is uniform, this central domain is empty.

When the length is smaller than the completion length $\alpha_{\min } \log n$ or greater than the extinction length $\alpha_{\text {ext }} \log n$, the ratio $\frac{\log \mu(n, \alpha \log n)}{\log n}$ is negative.

### 2.4 A PROBABILISTIC INTERPRETATION

166 The main contribution to $\mu(n, k)$ arises from $k$-mers with an objective function close to 0 , i.e. transition
Oscillations Parameters $\left(k_{1}, \cdots, k_{V}\right)$ in the combinatorial sums are integers. As the optimum values $\left(k \theta_{i}\right)_{1 \leq i \leq V}$ may not be integers, the practical maximum is a close point on the lattice $\left(k_{1}, \cdots, k_{V}\right)$. The difference introduces a multiplicative factor that ranges in $\left[-\log \frac{p_{\max }}{p_{\text {min }}}, \log \frac{p_{\text {max }}}{p_{\text {min }}}\right]$. This leads to a small oscillation of $\log \mu(n, k)$. For large $n$, this contribution to $\frac{\log \mu(n, k)}{\log n}$ becomes negligible. As mentioned above, the set of lengths $n$ that are admissible for a given $\alpha$ is very sparse. Nevertheless, an approximate value may be used : for instance, for an integer $k^{\prime}, \frac{1}{k^{\prime}} \log \left[n(\alpha)^{\frac{k^{\prime}}{k}}\right\rceil$ is very close to $\alpha$. This oscillation phenomenon was first observed in Nicodème (2005).

Binary alphabets Results for binary alphabets in Park et al. (2009) steadily follow from Theorem 2.2. A rewriting of $\psi$ leads to alternative expression (18). This explicit expression points out the dependency to the distances to $\alpha_{\min }$ and $\alpha_{\max }$, and the behaviour around these points.

Corollary 2.1. Assume the alphabet is binary. Then

$$
\begin{equation*}
\psi(\alpha)=\frac{\alpha}{\log \frac{p_{\max }}{p_{\min }}} \log \left[s_{\alpha^{\frac{1}{\alpha}-\frac{1}{\alpha_{\min }}}+s_{\left.\alpha^{\frac{1}{\alpha}-\frac{1}{\alpha_{\max }}}\right]} . \frac{1}{} .}\right. \tag{18}
\end{equation*}
$$

where

$$
\begin{equation*}
s_{\alpha}=\frac{\alpha_{\min }}{\alpha_{\max }} \cdot \frac{\alpha-\alpha_{\min }}{\alpha_{\max }-\alpha} . \tag{19}
\end{equation*}
$$

A similar result holds for DNA sequences when the alphabet is 4 -ary and the probability distribution satisfies $p_{A}=p_{T}$ and $p_{C}=p_{G}$. Such a distribution is defined by its GC-content $p_{G}+p_{C}$. $k$-mers. Such $k$-mers exist in the transition phase $\left[\alpha_{\min } \log n, \alpha_{\max } \log n\right]$ where they coexist with rare or common $k$-mers. Observe that this phase is shrinked when the Bernoulli model is uniform, as $p_{\min }=$ $p_{\max }$ and $\alpha_{\min }=\alpha_{\max }$. Therefore, most unique $k$-mers are concentrated on the two lengths $\left\lfloor\alpha_{\min } \log n\right\rfloor$ and $\left\lceil\alpha_{\text {min }} \log n\right\rceil$, as observed initially in Fagin et al. (1979b).

Let $k$ be some integer in the transition phase. First, the relative contribution of $S(k)$ and $T(k)$ to $\mu(n, k)$ varies with the length $k$. For lengths close to $\alpha_{\text {min }} \log n$, most words are common and $T(k)$ dominates $S(k)$. When $k$ increases, the proportion of common words decreases and the relative contribution of $T(k)$ decreases.
Second, the dominating term in $\mu(n, k)$ arises from transition $k$-mers. Let $w$ be a word of length $k$, the character distribution in $w$ be $\left(k_{1}, \cdots, k_{V}\right)$ and $\chi_{i}$ be some character. The number of words that admit $w$ or $w \chi_{i}$ as a prefix fluctuates around the expectations $n \phi\left(k_{1}, \cdots, k_{V}\right)$ and $n \phi\left(k_{1}, \cdots, k_{V}\right) p_{i}$, respectively. On the one hand, when word $w \chi_{i}$ is a rare word, $n \phi\left(k_{1}, \cdots, k_{V}\right)$ is smaller than 1 . The smallest $n \phi\left(k_{1}, \cdots, k_{V}\right)$ is, the less likely the actual number of occurrences of $w$ is greater than 2 and the smallest the contribution of $w \chi_{i}$ to $S(k)$, and $\mu(n, k)$, is. On the other hand, let $w \chi_{i}$ be a common $k+1$-mer; $w$ is a common $k$-mer and then $n \phi\left(k_{1}, \cdots, k_{V}\right)$ is greater than 1 . The largest $n \phi\left(k_{1}, \cdots, k_{V}\right)$ is, the more likely the word $w \chi_{i}$ is repeated and the smallest the contribution to $T(k)$, and $\mu(n, k)$, is.

For a short length, i.e. $k$ smaller than the completion length $k_{\text {min }}$, all words are common. In a given sequence, most $k$-mers are repeated at least twice and there is (almost) no unique $k$-mers.
For a large length $k$, i.e. $k$ greater than $k_{\max }$, all words are rare. Nevertheless the number of unique $k$-mers remains sublinear in $n$ in the range $\left[\alpha_{\max } \log n, \alpha_{e x t} \log n\right]$ : the sum of small contributions arising from a large number of possible words is significant.

## 3 EXPERIMENTS AND ANALYSIS

191 Simulations are presented for random and real data. For each simulation, a suffix tree Ukkonen (1995)
A folk theorem Szpankowski (2001); Jacquet and Szpankowski (2015)claims that the objective function is concentrated around $\frac{1}{\tilde{\alpha}}-\frac{1}{\alpha}$. Consequently, when $\alpha=\tilde{\alpha}$, most $k$-mers are transition $k$-mers and the exponent, the $\psi$ function, is maximal. is built, where each leaf represents a unique $k$-mer. For random cases, the Ukkonen's insertion step is iterated until a tree with exactly $n$ leaves is build. This requires $n+k_{i n s}$ insertions of symbols, where $k_{i n s}>0$ is relatively small (there is a value of a few dozen in practice for considered $n$ ). One can observe that the event of having $n$ leaves after $n+k-1$ insertions corresponds to the fact that the trailing $k$-mer is unique in the sequence of length $n+k-1$.
Even if a statistical bias exists, with respect to the case of a set of N random words analyzed in previous sections, this bias for respective values on $k$ and $n$ is below the numeric precision used for tables below.
Then, one simulation that is related to the case of a set of $n$ random words, requires the generation of the order of N random symbols from a small alphabet, following a Bernoulli scheme. For this range of $n$, and even in the case of a hundred consecutive simulations, this corresponds to a regular use of a common random number generator Knuth (1998).
A first set of simulation deals with the case of random sequences over a binary alphabet, since the results can be compared with previous work. A second set addresses the case of random sequences over a quaternary alphabet $\{A, C, G, T\}$ with a constrained distribution such that probabilities $p_{A} \approx p_{T}$ and $p_{C} \approx p_{G}$ as it is the case for DNA sequences (where the sum $p_{C}+p_{G}$ is also known as the GC-content). Results on such random sequences are then compared with the sample biological sequence of an Archaea (Haloferax volcanii)

An implementation with a suffix array Manber and Myers (1993) allows for a compact representation and an efficient counting Beller et al. (2013).

### 3.1 RANDOM DATA

A hundred binary sequences were randomly generated. The number of leaves in each tree was fixed to $n=5000000$ and the Bernoulli parameter was $p_{\max }=0.7000$. Therefore, $p_{\min }=0.3000, \tilde{p}=$ 0.5429 and $\log n=15.4249$. The thresholds for $\alpha$ and the corresponding lengths $\alpha \log n$ are :

| $\alpha_{\text {min }}=0.8306$ | $\tilde{\tilde{\alpha}}=1.6370$ | $\bar{\alpha}=2.0484$ | $\alpha_{\max }=2.8035$ | $\alpha_{\text {ext }}=3.6714$ |
| :--- | :--- | :--- | :--- | :--- |
| $k_{\text {min }}=12.81$ | $\tilde{k}=25.25$ | $\bar{k}=31.60$ | $k_{\max }=43.24$ | $k_{\text {ext }}=56.63$ |

Statistical behaviour on random sets Throughout experiments, every sample profile for a given sequence fluctuates very little around the expectation, as mentioned in 2.1.
Table 3.1 provides experimental results averaged over a hundred binary sequences. Short length with no observed unique $k$-mer are removed. Column 2 gives the mean of $B(k+1$ ), i.e. the mean number of observed leaves at depth $k+1$, over the set of a hundred simulations. Columns 3 to 5 give the computed values for $S(k), T(k)$ and $\mu(k)$, using the expressions (7), (8) and (9).
The actual number of leaves $B(n, k+1)$ is very close to the average value $\mu(n, k)$, and simulations show this is the general case when (only) a hundred simulations are performed : $\mu(n, k)$ is a very good prediction.
Observed lengths of extinction also show very little variations. In array below, each column gives $n_{k}$, the number of sequences out of the one hundred sample set for which the longest repetition had length $k$.

| $k$ | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 |
| ---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $n_{k}$ | 10 | 16 | 13 | 19 | 14 | 14 | 6 | 1 | 1 | 2 | 1 | 1 | 0 | 2 |

Distribution of the extinction level for 100 random binary sequences.
$p_{\max }$ is 0.7 .

In the binary case, the predicted extinction length is between 56 and 57. It is noticeable that, in most cases, the observed depth is slightly smaller than this value. In Table 3.1, value 0.04 for $\mu(n, 61)$ means that one expects a total of 4 leaves at depth 60 over one hundred sequences. In that run, exists a total amount of 8 .

## Quality of estimates

1. Tightness of the asymptotic estimates. Asymptotic estimates (13) given in Column 7 significantly overestimate the observed values in Column 6 that is computed directly from column 2 and $n$. A first conclusion is that first order asymptotics provide a poor prediction : next term is $O\left(\frac{1}{\log n}\right)$ that goes slowly to 0 .
2. Tightness of the second order asymptotics. Second term for the asymptotic $\xi(\alpha)$ ensures a much better approximation in Column 8.
3. Growth of asymptotic estimates. Observed values increase with length until $k=\tilde{k}$ and then decrease. This is consistent with the variation of asymptotic values $\psi(\alpha)$.

Dependency to probability bias Thresholds were computed for a given sequence length $n$ and various probabilities. The more $p_{\max }$ departs from 0.5 , the value for the uniform model, the largest the extinction length is. The completion length $k_{\text {min }}$, slightly decreases while the extinction length significantly increases. Nevertheless, this effect is limited when the largest probability $p_{\max }$ remains in the range [0.5; 0.7].


Table 3.1 : Mean profile for 100 random binary sequences. $\left(p_{\max } ; p_{\text {min }}\right)=(0.7 ; 0.3)$.

| $p_{\max }$ | $k_{\min }$ | $\tilde{k}$ | $\bar{k}$ | $k_{\max }$ | $k_{\text {ext }}$ |
| :---: | ---: | :---: | :---: | ---: | :---: |
| 0.50 | 22.25 | 22.25 | 22.25 | 22.25 | 44.51 |
| 0.55 | 19.32 | 22.42 | 22.74 | 25.80 | 45.16 |
| 0.60 | 16.83 | 22.92 | 24.27 | 30.20 | 47.18 |
| 0.65 | 14.69 | 23.82 | 27.06 | 35.81 | 50.83 |
| 0.70 | 12.81 | 25.25 | 31.60 | 43.25 | 56.63 |
| 0.75 | 11.13 | 27.43 | 38.80 | 53.62 | 65.64 |
| 0.80 | 9.58 | 30.83 | 50.63 | 69.13 | 79.99 |
| 0.85 | 8.13 | 36.49 | 71.78 | 94.91 | 104.80 |
| 0.90 | 6.70 | 47.45 | 116.72 | 146.40 | 155.45 |
| 0.95 | 5.15 | 77.70 | 259.56 | 300.72 | 309.05 |

Dependency of thresholds to $p_{\max }$ for binary alphabets. $n=5000000$

### 3.2 LONG REPETITIONS IN ARCHAEA GENOMES

244 The experimental data set is the sequence from Haloferax volcanii DS2 chromosome, complete genome 245 Hartman et al. (2010). The alphabet is quaternary. Profile results are shown in Table 3.2 (a).

| $k$ | observed $B(k+1)$ | $S(k)$ | predicted $T(k)$ | $\mu(n, k)$ |
| :---: | :---: | :---: | :---: | :---: |
| 6 | 4 | 0 | 0.05 | 0.05 |
| 7 | 1975 | 0 | $4 \mathrm{e}+02$ | $4 \mathrm{e}+02$ |
| 8 | 41349 | 0 | $2 \mathrm{e}+04$ | $2 \mathrm{e}+04$ |
| 9 | 178523 | 781.2 | 213568.8 | 214350.1 |
| 10 | 382032 | 66858.4 | 617279.6 | 684137.9 |
| 11 | 542386 | 171711.2 | 742379.1 | 914090.3 |
| -12 | $\overline{5} 70499$ | $\overline{4} 079 \overline{9} \overline{6} . \overline{5}$ | $\underline{2} 15 \overline{5} 4 \overline{2} . \overline{2}$ | $\overline{6} 2 \overline{39} 18 . \overline{7}$ |
| 13 | 459330 | $2 \overline{59} \overline{86} 0.7$ | 6512.5 | $26 \overline{63} \overline{73} .2$ |
| 14 | 305002 | 87488.6 | 0 | 87488.6 |
| 15 | 169317 | 25704.4 | 0 | 25704.4 |
| 16 | 86379 | 7264.7 | 0 | 7264.7 |
| 17 | 40391 | 2028.2 | 0 | 2028.2 |
| 18 | 17432 | 564.1 | 0 | 564.1 |
| 19 | 7866 | 156.7 | 0 | 156.7 |
| 20 | 3830 | 43.5 | 0 | 43.5 |
| 21 | 1957 | 12.1 | 0 | 12.1 |
| 22 | 1229 | 3.4 | 0 | 3.4 |
| 23 | 910 | 0.9 | 0 | 0.9 |
| 24 | 733 | 0.3 | 0 | 0.3 |
| 25 | 617 | 0.07 | 0 | 0.07 |
| 26 | 561 | 0.02 | 0 | 0.02 |
| 27 | 492 | 0.006 | 0 | 0.006 |
| 28 | 446 | 0.002 | 0 | 0.002 |
| 29 | 436 | 0.0005 | 0 | 0.0005 |
| 30 | 397 | 0.0001 | 0 | 0.0001 |
| 31 | 374 | $1 \mathrm{e}-05$ | 0 | $1 \mathrm{e}-05$ |
| 32 | 359 | 2e-06 | 0 | 2e-06 |
| 33 | 322 | $2 \mathrm{e}-08$ | 0 | $2 \mathrm{e}-08$ |
| ... | truncated | ... | truncated | ... |

Table 3.2 (a) : Profile for the sequence from Haloferax volcanii DS2 chromosome, complete genome

Sequence length is $n=2847757$. The observed symbol frequencies are $p_{A}=0.1655 ; p_{C}=$ $0.3334 ; p_{G}=0.3330 ; p_{T}=0.1681$. Therefore, observed $G C$-content is 0.6664 . Parameters for an approximate degenerated quaternary model are $p_{A}=p_{T}=p_{\min }=0.1668 ; p_{C}=p_{G}=p_{\max }=0.3332 ; \tilde{p}=$ 0.2645 ; and $\log n=14.8620$. The thresholds for the domain are

| $\alpha_{\text {min }}=0.5584$ | $\tilde{\tilde{\alpha}}=0.7520$ | $\bar{\alpha}=0.8079$ | $\alpha_{\text {max }}=0.9099$ | $\alpha_{\text {ext }}=1.5609$ |
| :--- | :--- | :--- | :--- | :--- |
| $k_{\text {min }}=8.30$ | $\tilde{k}=11.18$ | $\bar{k}=12.01$ | $k_{\max }=13.52$ | $k_{\text {ext }}=23.20$ |

Statistics on one hundred random sequences with same parameters are shown on figure 3.2(b). GCcontent is 0.6664 . Extinction level is provided in Table 3.2. Observe first a good match between the observed values, the predicted values for $\mu(n, k)$ and the asymptotic values for random data. As shown for binary alphabets, the observed extinction level for random sequences departs very little from the predicted $k_{\text {ext }}$ level.

| $k$ | 21 | 22 | 23 | 24 | 25 |
| ---: | :--- | :--- | :--- | :---: | :---: |
| $n_{k}$ | 26 | 42 | 18 | 7 | 7 |

Table 3.2 Distribution of the extinction level for 100 random degenerated quaternary sequences. $G C$-content is 0.6664 .

Numerous differences with random data can be observed on real genomes.
Interestingly, the behaviour for short lengths and in the transition phase is similar to the random behaviour. Observation and prediction have the same order of magnitude. In particular, the number of unique $k$-mers is maximum for length $\tilde{k}$ where observation and prediction coincide. For a real genome and a length $k$ smaller than $k_{\text {min }}$, observed $B(n, k+1)$ is larger than predicted $\mu(n, k)$. This indicates,


Table 3.2 (b) : Mean profile for 100 random degenerated quaternary sequences. GC-content is 0.6664
at a level $k+1$ where completion is expected, more leaves in the real trie, more missing words at level $k+2$. Simultaneously, less internal nodes occur at level $k+1$ because the total sum is constant and equal to $V^{k+1}$.

The effect of (non-random) repetitions is more sensible in the decreasing domain. First, the number of unique $k$-mers decreases much more slowly than expected for lengths larger than $k_{\text {max }}$. A significant gap can be observed around extinction level $k_{\text {ext }}$. The decrease rate, that was around $0.02-0.04$ drops to 0.007 and then becomes even lower. Finally, the extinction level is much larger than the predicted value 23 : the largest repetition is 1395 bp long.

To evaluate the contribution of long repetitions, one may erase the longest ones. When a word $w$ is repeated, any proper suffix of $w$ is also repeated. Consequently, once the longest repeated word is erased, one unique $k$-mer (only) disappears for each length larger than the length of the second largest subsequence (here, 935). The profile remains far from the random profile. This observation is still true if the 10 longest subsequences are erased.

## 4 COMBINATORIAL AND ANALYTIC DERIVATION

### 4.1 LAGRANGE MULTIPLIERS

274 Lagrange multipliers method allows to maximize an expression under constraints. To compute (17), one sets

$$
\begin{align*}
& F=\sum_{i=1}^{V} \theta_{i} \log \theta_{i} ;  \tag{20}\\
& G=\sum_{i=1}^{V} \theta_{i}  \tag{21}\\
& H=\sum_{i=1}^{V} \theta_{i} \beta_{i} \tag{22}
\end{align*}
$$

276 Two constraints are given :

$$
G=1 \text { and } H=\frac{1}{\alpha}
$$

277 An intermediary function $\phi_{\alpha}\left(\tau_{1}, \cdots, \tau_{V}\right)$ is defined

$$
\begin{equation*}
\phi_{\alpha}=F+\lambda_{\alpha} G+\tau_{\alpha} H \tag{23}
\end{equation*}
$$

278 In order to maximize $\phi$ under these two constraints, $\phi$ function is derived with respect to each random
279 variable $\tau_{i}$. This yields $V$ equations

$$
\begin{equation*}
1+\log \theta_{i}+\lambda_{\alpha}+\tau_{\alpha} \beta_{i}=0 \tag{24}
\end{equation*}
$$

280 Two indices $i_{\min }$ and $i_{\max }$ are chosen that satisfy $\beta_{i_{\min }} \neq \beta_{i_{\max }}$. For instance

$$
\begin{aligned}
\beta_{i_{\min }} & =\min \left(\beta_{i}\right)_{1 \leq i \leq V}=\log \frac{1}{p_{\max }} \\
\beta_{i_{\max }} & =\max \left(\beta_{i}\right)_{1 \leq i \leq V}=\log \frac{1}{p_{\min }}
\end{aligned}
$$

281 Solving Equations (24) with indices $i_{\min }$ and $i_{\max }$ yields

$$
\begin{aligned}
\tau_{\alpha} & =\frac{\log \theta_{i_{\min }}-\log \theta_{i_{\max }}}{\beta_{i_{\max }}-\beta_{i_{\min }}}=\log \frac{\theta_{i_{\min }}}{\theta_{i_{\max }}}{ }^{\frac{1}{\beta_{m_{\max }-\beta_{i_{\min }}}}} ; \\
1+\lambda_{\alpha} & =\frac{\beta_{i_{\min } \log \theta_{i_{\max }}-\beta_{i_{\max }} \log \theta_{i_{\min }}}^{\beta_{i_{\max }}-\beta_{i_{\min }}}}{} .
\end{aligned}
$$

282 Remaining equations rewrite :

$$
\begin{equation*}
\log \theta_{i}=\log \theta_{i_{\text {min }}}+\tau_{\alpha}\left(\beta_{i_{\text {min }}}-\beta_{i}\right) \tag{25}
\end{equation*}
$$

283 The constraint $\sum_{i=1}^{V} \theta_{i}=1$ yields

$$
\theta_{i_{\min }} e^{\beta_{1} \tau_{\alpha}} \sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}=1
$$

284 and Equation 25 rewrites, for each index $i$ :

$$
\begin{equation*}
\theta_{i}=\frac{e^{-\beta_{i} \tau_{\alpha}}}{\sum_{i=1}^{V} \beta_{i} e^{-\beta_{i} \tau_{\alpha}}} \tag{26}
\end{equation*}
$$

285 Finally, Equation $\sum_{i=1}^{V} \theta_{i} \beta_{i}=\frac{1}{\alpha}$ yields Equation (10).

$$
\frac{1}{\alpha}=\frac{\sum_{i=1}^{V} \beta_{i} e^{-\beta_{i} \tau_{\alpha}}}{\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}}
$$

286 For this $V$-tuple

$$
\sum_{i=1}^{V} \theta_{i} \log \theta_{i}=-\left(\sum_{i=1}^{V} \theta_{i} \beta_{i}\right) \tau_{\alpha}-\left(\sum_{i=1}^{V} \theta_{i}\right) \log \left(\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right)=-\frac{\tau_{\alpha}}{\alpha}-\log \left(\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right)
$$

### 4.2 APPROXIMATION ORDERS

287 Derivating the RHS of (10) yields $\frac{\sum_{i \neq j}\left(\beta_{i}+\beta_{j}\right)^{2} e^{-\left(\beta_{i}+\beta_{j}\right) \tau}}{\left(\sum_{i} e^{-\beta_{i} \tau}\right)^{2}}$ that is positive. Therefore, for any $\alpha$, the solution 288 to (10) is unique. Moreover, $\tau_{\alpha}$ increases with $\alpha$. Let

$$
\begin{align*}
& \psi_{1}(\alpha)=\tau_{\alpha}+\alpha \log \left(\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right)  \tag{27}\\
& \psi_{2}(\alpha)=2-\alpha \log \frac{1}{\sigma_{2}} \tag{28}
\end{align*}
$$

289 Notably, the solutions $\tau_{\alpha}$ of (10) associated to the four increasing values of $\alpha:\left(\alpha_{\min }, \tilde{\alpha}, \bar{\alpha}, \alpha_{\max }\right)$ are $290(-\infty, 1,2,+\infty)$. Computing $\psi$ for these values yields (11) and Equality $\psi_{1}(\tilde{\alpha})=\psi_{2}(\tilde{\alpha})$.

Derivating both expressions yields

$$
\begin{align*}
\frac{\partial \psi_{1}}{\partial \alpha}(\alpha) & =\log \left(\sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right)  \tag{29}\\
\frac{\partial \psi_{1}}{\partial \alpha}(\alpha)-\frac{\partial \psi_{2}}{\partial \alpha}(\alpha) & =\log \left(\sigma_{2} \sum_{i=1}^{V} e^{-\beta_{i} \tau_{\alpha}}\right) \tag{30}
\end{align*}
$$

292 Both derivatives are monotone functions of $\tau_{\alpha}$. In (30), derivative is 0 when $\alpha=\bar{\alpha}$. Therefore, $\psi$ is the 293 maximum of the two values $\psi_{1}$ and $\psi_{2}$ over the interval [ $\alpha_{\min }, \alpha_{\max }$ ]. The former equation is 0 if $\alpha=\tilde{\alpha}$. Therefore, $\psi$ is maximum when $\alpha=\tilde{\alpha}$.

### 4.3 APPROXIMATIONS

295 Short lengths Assume that $k \leq \alpha_{\min } \log n$. Each term $\phi\left(k_{1}, \cdots, k_{V}\right)$ is lower bounded by $p_{\text {min }}^{k}=$ $296 n^{\alpha \log p_{\text {min }}}=n^{-\frac{\alpha}{\alpha_{\text {min }}}}$. Each term $\psi_{n}\left(k_{1}, \cdots, k_{V}\right)$ is trivially bounded by $e^{-n^{1-\frac{\alpha}{\alpha_{m i n}}}}$ that is upper bounded 297 by 1 and $n \psi_{n}\left(k_{1}, \cdots, k_{V}\right)$ tends to 0 when $n$ goes to $\infty$. As $\sum\binom{k}{k_{1} \cdots k_{V}} \phi\left(k_{1}, \cdots, k_{V}\right)=1$, the 298 ratio $\frac{\log \mu(n, k)}{\log n}$ is negative.

299 Moderate and large lengths For a length $k$ in the transition domain $\left[\alpha_{\min } \log n, \alpha_{\max } \log n\right]$, the objective 300 function may be either positive or negative. When $k>\alpha_{\max } \log n$, set $E_{k}(n)$ is empty and $\mu(n, k)$ reduces 301 to $S(k)$.
302 The maximum $M$ among the terms $e^{k\left(-\sum_{i} \frac{k_{i}}{k} \log \frac{k_{i}}{k}-\frac{1}{k} \log n \phi\left(k_{1}, \cdots, k_{V}\right)\right)}$ in $T(k)$ is reached when $303 \rho\left(k_{1}, \cdots, k_{V}\right)$ is 0 . Due to the exponential decrease of $e^{-n \phi\left(k_{1}, \cdots, k_{V}\right)}$ when $n \phi\left(k_{1}, \cdots, k_{V}\right) \geq 1, \frac{T(k)}{k}$ is $306 S(k)$ rewrites $\sigma_{2}{ }^{k} \tilde{S}(k)+\left(S(k)-\sigma_{2}{ }^{k} \tilde{S}(k)\right)$ where $\tilde{S}(k)=\sum_{\rho\left(k_{1}, \cdots, k_{V}\right) \leq 0}\binom{k}{k_{i}}\left(\frac{p_{1}^{2}}{\sigma_{2}}\right)^{k_{1}} \cdots\left(\frac{p_{V}^{2}}{\sigma_{2}}\right)^{k_{V}}$. 307 This sum satisfies a Large Deviation Principle when $\rho\left(k_{1}, \cdots, k_{V}\right)+\frac{1}{\alpha} \geq \frac{1}{\tilde{\alpha}}$, or $\alpha<\tilde{\alpha}$. In this range, $308 \frac{\tilde{S}(k)}{k} \sim \max \left\{-\sum_{i=1}^{V} \frac{k_{i}}{k} \log \frac{k_{i}}{k}\right\}$, that was shown to be $\psi(\alpha)$.
309 When $\alpha>\tilde{\alpha}$, sum $\tilde{S}(k)$ rewrites $1-\bar{S}(k)$ where

$$
\bar{S}(k)=\sum_{\rho\left(k_{1}, \cdots, k_{V}\right)+\frac{1}{\alpha}<\frac{1}{\bar{\alpha}}}\binom{k}{k_{i}}\left(\frac{p_{1}^{2}}{\sigma_{2}}\right)^{k_{1}} \cdots\left(\frac{p_{V}^{2}}{\sigma_{2}}\right)^{j_{V}}
$$

310 This sum satisfies a Large Deviation Principle and $\frac{\bar{S}(k)}{k} \sim \max \left\{-\sum_{i=1}^{V} \frac{k_{i}}{k} \log \frac{k_{i}}{k}+\sum_{i=1}^{V} \frac{k_{i}}{k} \log \frac{p_{i}^{2}}{\sigma_{2}}\right\}$. As $311 \sum_{i=1}^{V} \frac{k_{i}}{k} \log \frac{p_{i}^{2}}{\sigma_{2}}=-\frac{2}{\alpha}+\log \frac{1}{\sigma_{2}}$, this maximum is

$$
-\frac{1}{\alpha}\left[2-\alpha \log \frac{1}{\sigma_{2}}-\psi(\alpha)\right]
$$

312 that is negative.

### 4.4 BINARY CASE

$s_{\alpha}{ }^{-\left(\frac{1}{\alpha_{\min }}-\frac{1}{\alpha}\right)}$ and $e^{-\left(\beta_{2}-\frac{1}{\alpha}\right) \tau_{\alpha}}=s_{\alpha}^{-\left(\frac{1}{\alpha_{\max }}-\frac{1}{\alpha}\right)}$.

## 5 CONCLUSION

319 This paper describes the behaviour of the number of unique or repeated $k$-mers in a random sequence, approximate repetitions.

## DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## ACKNOWLEDGEMENT

334 Funding: Inria-Cnrs-Poncelet grant Carnage.

## REFERENCES

Beller, T., Gog, S., Ohlebusch, E., and Schnattinger, T. (2013), Computing the longest common prefix array based on the burrows-wheeler transform, Journal of Discrete Algorithms, 18, 22-31
Chikhi, R. and Medvedev, P. (2014), Informed and automated k-mer size selection for genome assembly., Bioinformatics, 30, 1, 31-37, doi:10.1093/bioinformatics/btt310
Devillers, H. and Schbath, S. (2012), Separating significant matches from spurious matches in dna sequences, Journal of Computational Biology, 19, 1, 1-12, doi:10.1089/cmb.2011.0070
Fagin, R., Nievergelt, J., Pippenger, N., and Strong, H. R. (1979a), Extendible hashinga fast access method for dynamic files, ACM Transactions on Database Systems (TODS), 4, 3, 315-344
Fagin, R., Nievergelt, J., Pippenger, N., and Strong, R. (1979b), Extendible hashing: A fast access method for dynamic files, A.C.M. Trans. Database Syst., 4, 315-344
Flajolet, P., Kirschenhofer, P., and Tichy, R. F. (1988), Deviations from uniformity in random strings, Probability Theory and Related Fields, 80, 139-150
Gu, Z., Wang, H., Nekrutenko, A., and Li, W. H. (2000), Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence., Gene, 259, 1-2, 81-88
Hartman, A. L., Norais, C., Badger, J. H., Delmas, S., Haldenby, S., Madupu, R., et al. (2010), The complete genome sequence of haloferax volcanii ds2, a model archaeon., PLoS One, 5, 3, e9605, doi:10.1371/journal.pone. 0009605
Jacquet, P. and Szpankowski, W. (1994), Autocorrelation on words and its applications: Analysis of suffix trees by string-ruler approach, Journal of Combinatorial Theory, Series A, 66, 2, 237 - 269, doi:http://dx.doi.org/10.1016/0097-3165(94)90065-5
Jacquet, P. and Szpankowski, W. (2015), Analytic Pattern Matching: From DNA to Twitter (Cambridge University Press)
Janson, S., Lonardi, S., and Szpankowski, W. (2004), On the average sequence complexity, in Combinatorial Pattern Matching (Springer), 74-88
Knuth, D. (1998), The art of computer programming, volume two, seminumerical algorithms
Magner, A., Knessl, C., and Szpankowski, W. (2014), Expected external profile of patricia tries, in Proceedings of the Meeting on Analytic Algorithmics and Combinatorics (Society for Industrial and Applied Mathematics), 16-24
Mahmoud, H. (1992), Evolution of Random Search Trees (John Wiley \& Sons, New York)
Manber, U. and Myers, G. (1993), Suffix arrays: a new method for on-line string searches, siam Journal on Computing, 22, 5, 935-948
Nicodème, P. (2005), Average profiles, from tries to suffix-trees, in C. Martìnez, ed., 2005 International Conference on Analysis of Algorithms, volume AD of DMTCS Proceedings (Discrete Mathematics and Theoretical Computer Science, Barcelona, Spain), volume AD of DMTCS Proceedings, 257-266
Park, G., Hwang, H.-K., Nicodeme, P., and Szpankowski, W. (2009), Profile of trie, SIAM Journal on Computing, 38, 5, 1821-1880
Rizk, G., Lavenier, D., and Chikhi, R. (2013), Dsk: k-mer counting with very low memory usage., Bioinformatics, 29, 5, 652-653, doi:10.1093/bioinformatics/btt020
Sedgewick, R. and Flajolet, P. (2009), Analytic combinatorics, Cambridge University
Szpankowski, W. (2001), Average Case Analysis of Algorithms on Sequences (John Wiley and Sons, New York)
Treangen, T. J. and Salzberg, S. L. (2012), Repetitive dna and next-generation sequencing: computational challenges and solutions., Nat Rev Genet, 13, 1, 36-46, doi:10.1038/nrg3117
Ukkonen, E. (1995), On-line construction of suffix trees, Algorithmica, 14, 3, 249-260

