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-01304366v1
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

1 Inria, France
2 LIX, Ecole Polytechnique, France

Correspondence*: Mireille Régnier
AMIB, Inria, 1 rue d’Estienne d’Orves, Palaiseau, 91 120, France, mireille.regnier@inria.fr

ABSTRACT
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.
The prediction of the length of maximal common prefixes for words in a random set is a problem that has been extensively studied. Typical parameters are the background probability model, the size $V$ of the alphabet, the length $n$ of the sequence, ... Deviation from uniformity was studied for a uniform model as early as 1988 Flajolet et al. (1988). A complexity index that captures the richness of the language is addressed in Janson et al. (2004). A distribution model, valid for binary alphabets and biased distributions, was introduced in Park et al. (2009), the so-called trie profile and extended to Patricia tries in Magner et al. (2014). The authors pointed out different “regimes” of randomness and a phase transition, by means of analytic combinatorics Sedgewick and Flajolet (2009).

Our first goal is to extend results of Park et al. (2009) to the case of a general $V$-alphabet, including the special case \{A, C, G, T\} where $V$ is 4. A second goal is to compare the results consistency with random data and real genomic data in the finite range.

To achieve the first goal, we rely on an alternative, and simpler, probabilistic and combinatorial approach that is interesting per se. It avoids generating functions and the Poissonization-dePoissonization cycle that is used in Park et al. (2009) and it extends to non-binary alphabets. In that case, there is no closed formula for the asymptotic behavior. Nevertheless, the Lagrange multipliers allow to derive it as the solution of an equation that can be computed numerically.

Explicit and computable bounds for the profile of a random set of $n$ words are provided. Three domains can be observed. A first domain is defined by a threshold $k$ for the length, called the completion length: any prefix with a length smaller than this threshold occurs at least twice. This threshold is extremely stable over the data sets and it is highly predictable. A similar phenomenon was observed for a uniform model in Fagin et al. (1979a) and a biased model Mahmoud (1992); Park et al. (2009). For larger lengths, some prefixes occur only once. In a second domain, called the transition phase, the number of maximal common prefixes is sublinear in the size $n$ of the sequence: increasing first, then decreasing slowly and, finally, dropping rapidly. In the third domain, for a length larger than some extinction length, almost no common prefix of that length occurs. Despite the fact these bounds are asymptotic, a good convergence is shown in practice for random texts when a second order term is known.

Differences between the model and the observation are studied on the special case of Archaea genomes. A dependency to the GC-content, that is a characteristic of each genome, is exhibited. Regimes and transitions are studied on these genomic data and theoretical results are confirmed, with a drift in the values of transition thresholds. Notably, the length of the largest repetitions is much larger than expected. This difference between the model and the observation arises from the occurrences of long repeated regions.

Section 2 is devoted to Main Results, to be proved in Section 4. First, some notations are introduced; then, an algebraic expression for the expectation of the number of maximal common prefixes in a sequence is derived in Theorem 2.1. Second, this expression is split between two sums that are computable in practical ranges. Then, a Large Deviation principle applies, that yields first and second order asymptotic terms, and oscillations, in Theorem 2.2. A comparison between exact, approximate and asymptotic expressions 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

$$\beta_i = \log \frac{1}{p_i}.$$
Sample Prediction of repetitions

69 Additionally

\[ p_{\min} = \min\{p_i; 1 \leq i \leq V\} \quad \text{and} \quad \alpha_{\min} = \frac{1}{\log \frac{1}{p_{\min}}} = \frac{1}{\max(\beta_i)} ; \]  

\[ p_{\max} = \max\{p_i; 1 \leq i \leq V\} \quad \text{and} \quad \alpha_{\max} = \frac{1}{\log \frac{1}{p_{\max}}} = \frac{1}{\min(\beta_i)} . \]  

70 The two values \( \min(\beta_i) \) and \( \max(\beta_i) \) 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 \( w_{\chi_i} \)
of length \( k \) such that

71 1. \( w \) is a prefix of at least two words in \( U \);
72 2. \( w_{\chi_i} \) is a prefix of a single word.

75 By convention, a unique 1-mer is a character \( \chi_i \) that is a prefix of a single word.

DEFINITION 2.3. Let \( U \) be a set of \( n \) words.

77 For \( k \geq 1 \), one denotes \( B(n, k) \) the number of unique \( k \)-mers in \( U \).

78 One denotes \( \mu(n, k - 1) \) the expectation of \( B(n, k) \) over all sets of \( n \) words.

Remark: It follows from Definition 2.2 that quantity \( B(n, k) \) is upper bounded by \( n \). Observe that, for
each random set \( U \), it is the sum of a large number \( -V^k \) of correlated random variables. Expectation
\( \mu(n, k) \) is studied below and compared in Section 3 with \( B(n, k + 1) \).

82 Profiles of repetitions can be expressed as a combinatorial sum.

THEOREM 2.1. Given a length \( k \), the expectation \( \mu(n, k) \) satisfies :

\[ \mu(n, k) = n \sum_{k_1 + \cdots + k_V = k} \binom{k}{k_1, \cdots, k_V} \phi(k_1, \cdots, k_V) \psi_n(k_1, \cdots, k_V) \]  

84 where

\[ \phi(k_1, \cdots, k_V) = p_1^{k_1} \cdots p_V^{k_V} \]  

\[ \psi_n(k_1, \cdots, k_V) = \sum_{i=1}^{V} p_i[(1 - \phi(k_1, \cdots, k_V)p_i)^{n-1} - (1 - \phi(k_1, \cdots, k_V))^{n-1}] . \]  

85 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.

87 Event (i) has probability \( n\phi(k_1, \cdots, k_V)p_i[1 - (1 - \phi(k_1, \cdots, k_V)p_i)^{n-1}] \).
88 Event (ii), that is a sub-event of (i), has probability \( n\phi(k_1, \cdots, k_V)p_i[1 - (1 - \phi(k_1, \cdots, k_V)p_i)^{n-1}] \).
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

$$\rho(k_1, \cdots, k_V) = \sum_{i=1}^{V} \frac{k_i}{k} \beta_i - \frac{1}{\alpha}. \quad (6)$$

The character distribution $(k_1, \cdots, k_V)$ of a k-mer may be viewed as barycentric coordinates for a point $\beta(k_1, \cdots, k_V) = \sum_{i=1}^{V} \frac{k_i}{k} \beta_i$ that lies in $[\min(\beta_i); \max(\beta_i)] = \left[\frac{1}{\alpha_{\max}}; \frac{1}{\alpha_{\min}}\right]$. 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 $\rho(k_1, \cdots, k_V) < 0$;
- a transition k-mer if $\rho(k_1, \cdots, k_V) \geq 0$ and its ancestor is a common k-mer;
- a rare k-mer, otherwise.

**Remark:** If $\rho(k_1, \cdots, k_V) = 0$, 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 $(k_1, \cdots, k_V)$ 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 $k < \alpha_{\min} \log n$ and is the set of character distributions $(k_1, \cdots, k_V)$ 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

$$S(k) = n \sum_{D_k(n)} \left( \begin{array}{c} k \\ k_1 \cdots k_V \end{array} \right) \phi(k_1, \cdots, k_V) \psi_n(k_1, \cdots, k_V); \quad (7)$$

$$T(k) = n \sum_{E_k(n)} \left( \begin{array}{c} k \\ k_1 \cdots k_V \end{array} \right) \phi(k_1, \cdots, k_V) \psi_n(k_1, \cdots, k_V). \quad (8)$$

So $\mu(n, k)$ rewrites

$$\mu(n, k) = S(k) + T(k). \quad (9)$$

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

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

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

Definition 2.7. Let \((p_i)_{1 \leq i \leq V}\) be a Bernoulli probability distribution. Let \(\sigma^2\) denote \(\sum_{i=1}^{V} p_i^2\).
The fundamental ratio, noted \(\hat{\alpha}\), is \((\sum_i p_i \log \frac{1}{p_i})^{-1}\).
The transition ratio, noted \(\bar{\alpha}\), is \(\sigma^2(\sum_i p_i^2 \log \frac{1}{p_i})^{-1}\).
The extinction ratio, noted \(\alpha_{ext}\), is \(2 \log \frac{1}{\sigma^2}\).

Definition 2.8. Let \(\alpha\) be a real value in \([\alpha_{min}, \alpha_{max}]\). Let \(\tau_\alpha\) be the unique real root of the equation
\[
\frac{1}{\alpha} = \frac{\sum_{i=1}^{V} \beta_i e^{-\beta_i \tau}}{\sum_{i=1}^{V} e^{-\beta_i \tau}}
\]
(10)

Let \(\psi\) be the function defined in \([\alpha_{min}, \alpha_{ext}]\) as
\[
\alpha_{min} \leq \alpha \leq \tilde{\alpha} : \psi(\alpha) = \tau_\alpha + \alpha \log(\sum_{i=1}^{V} e^{-\beta_i \tau}) ;
\]
\[
\tilde{\alpha} \leq \alpha : \psi(\alpha) = 2 - \alpha \log \frac{1}{\sigma^2} .
\]

Proposition 2.1. The following property holds
\[
\alpha_{min} \leq \tilde{\alpha} \leq \alpha_{max} \leq \alpha_{ext} .
\]
Function \(\psi\) increases on \([\alpha_{min}, \tilde{\alpha}]\) and decreases on \([\tilde{\alpha}, \infty]\). It satisfies
\[
\psi(\alpha_{min}) = \psi(\alpha_{ext}) = 0 \text{ and } \psi(\tilde{\alpha}) = 1 .
\]
(11)

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

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:

$$0 \leq \alpha \leq \alpha_{\text{min}} \text{ or } \alpha_{\text{ext}} \leq \alpha : \frac{\log \mu(n, \alpha \log n)}{\log n} \leq 0; \quad (12)$$

$$\alpha_{\text{min}} \leq \alpha \leq \alpha_{\text{ext}} : \frac{\log \mu(n, \alpha \log n)}{\log n} \sim \psi(\alpha). \quad (13)$$

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

$$\alpha_{\text{min}} \leq \alpha \leq \bar{\alpha} : \xi(\alpha) \sim -\frac{V-1}{2} \frac{\log(\log n)}{\log n}; \quad (14)$$

$$\bar{\alpha} \leq \alpha \leq \alpha_{\text{ext}} : \xi(\alpha) \sim \frac{\log(1-\sigma_2)}{\log n}. \quad (15)$$

PROOF. The key to the proof when $\alpha$ ranges in $[\alpha_{\text{min}}, \alpha_{\text{max}}]$ is that $\psi_n(k_1, \ldots, k_V)$ is close to 0. Sum $T(k)$ satisfies a Large Deviation Principle.

$$\frac{\log T(k)}{k} \sim \max\left\{-\sum_{i=1}^V k_i \log \frac{k_i}{k}; \rho(k_1, \ldots, k_V) = 0\right\}. \quad (16)$$

The maximization problem rewrites as

$$\max\left\{\sum_{i=1}^V \theta_i \log \frac{1}{\theta_i}; \sum_{i=1}^V \theta_i = 1; \sum_{i=1}^V \beta_i \theta_i = \frac{1}{\alpha}; 0 \leq \theta_i \leq 1\right\} \quad (17)$$

The maximum value is $\tau_\alpha + \alpha \log(\sum_{i=1}^V e^{-\beta_i \tau_\alpha})$ that is reached for the $V$-tuple $(\theta_i = \frac{e^{-\beta_i \tau_\alpha}}{\sum_{i=1}^V e^{-\beta_i \tau_\alpha}})_{1 \leq i \leq V}$. $S(k)$ 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 $(1-\sigma_2)n^{1-\alpha \log \frac{1}{\sigma_2}}$ that dominates $T(k)$.

Details for the proof, including the short and long lengths, are provided in Section 4.

Remark: The discussion will depend on 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 $\psi(\alpha_{\text{min}}) = 0$ 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 $[\alpha_{\text{min}}, \alpha_{\text{ext}}]$, 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 $\alpha_{\text{min}} \log n$ or greater than the extinction length $\alpha_{\text{ext}} \log n$, the ratio $\frac{\log \mu(n, \alpha \log n)}{\log n}$ is negative.
Oscillations Parameters \((k_1, \ldots, k_V)\) in the combinatorial sums are integers. As the optimum values \((k_\theta)_1 \leq \theta \leq V\) may not be integers, the practical maximum is a close point on the lattice \((k_1, \ldots, k_V)\). The difference introduces a multiplicative factor that ranges in \([- \log \frac{p_{\max}}{p_{\min}}, \log \frac{p_{\max}}{p_{\min}}]\). 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 \(k', \frac{1}{k} \log \left(\frac{n(\alpha)^k}{k'}\right)\) 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
\[
\psi(\alpha) = \frac{\alpha}{\log \frac{p_{\max}}{p_{\min}}} \log \left[ s_\alpha \frac{1}{\alpha_{\max}} - \frac{1}{\alpha_{\min}} + s_\alpha \frac{1}{\alpha_{\min}} - \frac{1}{\alpha_{\max}} \right] \tag{18}
\]
where
\[
s_\alpha = \frac{\alpha_{\min}}{\alpha_{\max}} \cdot \frac{\alpha - \alpha_{\min}}{\alpha_{\max} - \alpha} \tag{19}
\]

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 \(p_G + p_C\).

**2.4 A Probabilistic Interpretation**

The main contribution to \(\mu(n, k)\) arises from \(k\)-mers with an objective function close to 0, i.e. transition \(k\)-mers. Such \(k\)-mers exist in the transition phase \([\alpha_{\min} \log n, \alpha_{\max} \log n]\) 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 \([\alpha_{\min} \log n]\) and \([\alpha_{\min} \log n]\), 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 \(\alpha_{\min} \log n\), 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 \((k_1, \ldots, k_V)\) and \(\chi_i\) be some character. The number of words that admit \(w\) or \(w\chi_i\) as a prefix fluctuates around the expectations \(n\phi(k_1, \ldots, k_V)\) and \(n\phi(k_1, \ldots, k_V)p_i\), respectively. On the one hand, when word \(w\chi_i\) is a rare word, \(n\phi(k_1, \ldots, k_V)\) is smaller than 1. The smallest \(n\phi(k_1, \ldots, k_V)\) 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 \(n\phi(k_1, \ldots, k_V)\) is greater than 1. The largest \(n\phi(k_1, \ldots, k_V)\) 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 \(k_{\min}\), 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 \([\alpha_{\min} \log n, \alpha_{\max} \log n]\): the sum of small contributions arising from a large number of possible words is significant.
A folk theorem Szpankowski (2001); Jacquet and Szpankowski (2015) claims that the objective function is concentrated around $\frac{1}{\alpha} - \frac{1}{\tilde{\alpha}}$. Consequently, when $\alpha = \tilde{\alpha}$, most $k$-mers are transition $k$-mers and the exponent, the $\psi$ function, is maximal.

### 3 EXPERIMENTS AND ANALYSIS

Simulations are presented for random and real data. For each simulation, a suffix tree Ukkonen (1995) 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 $n + k_{ins}$ insertions of symbols, where $k_{ins} > 0$ 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) and 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, $p_{min} = 0.3000$, $\hat{p} = 0.5429$ and $\log n = 15.4249$. The thresholds for $\alpha$ and the corresponding lengths $\alpha \log n$ are:

| $\alpha_{min}$ | $k_{min}$ | $\hat{\alpha}$ | $\hat{k}$ | $\alpha_{max}$ | $k_{max}$ | $\alpha_{ext}$ | $k_{ext}$ |
|---------------|-----------|----------------|----------|---------------|----------|---------------|----------|
| 0.8306        | 12.81     | 1.6370         | 25.25    | 2.0484        | 31.60    | 2.8035        | 43.24    |
| 3000          | 15        | 3000           | 15       | 3000          | 15       | 3000          | 15       |

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 $k$. 

---

This is a provisional file, not the final typeset article
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 $p_{\text{max}}$ 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 $p_{\text{max}}$ remains in the range $[0.5; 0.7]$. 

---

| $k$ | 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. $p_{\text{max}}$ is 0.7.
Sample

### Prediction of repetitions

| $k$ | observed $B(k+1)$ | predicted $S(k)$ | $T(k)$ | $\mu(n, k)$ | observed $\log B(k+1)$ | asymptotic $\psi(\alpha)$ | $\psi(\alpha) + \xi(\alpha)$ |
|-----|-------------------|-----------------|--------|-------------|------------------------|------------------------|------------------------|
| 11  | 0.29              | 0               | 0.3    | 0.3         | -0.0803                | 0.1341                 |
| 12  | 7.91              | 0               | 8.3    | 8.3         |                        |                        |
| 13  | 87.87             | 0.1             | 86.9   | 87.1        | 0.2902                 | 0.0843                 |
| 14  | 552.88            | 1.2             | 550.3  | 551.5       | 0.4094                 | 0.3340                 |
| 15  | 2456.77           | 86.6            | 2366.4 | 2453.0      | 0.5061                 | 0.4962                 |
| 16  | 8269.20           | 209.4           | 8069.1 | 8278.5      | 0.5848                 | 0.6181                 |
| 17  | 22516.20          | 406.1           | 22097.7| 22503.8     | 0.6497                 | 0.7136                 |
| 18  | 51085.15          | 4823.8          | 46267.2| 51091.0     | 0.7028                 | 0.7897                 |
| 19  | 99387.01          | 6636.1          | 92717.6| 99353.7     | 0.7460                 | 0.8504                 |
| 20  | 169303.03         | 37415.5         | 131882.6| 169298.1   | 0.7805                 | 0.8994                 |
| 21  | 256358.10         | 7.91            | 8.3    | 8.3         | 0.1341                 |                        |
| 22  | 87.87             | 0.1             | 86.9   | 87.1        | 0.2902                 | 0.0843                 |
| 23  | 552.88            | 1.2             | 550.3  | 551.5       | 0.4094                 | 0.3340                 |
| 24  | 2456.77           | 86.6            | 2366.4 | 2453.0      | 0.5061                 | 0.4962                 |
| 25  | 8269.20           | 209.4           | 8069.1 | 8278.5      | 0.5848                 | 0.6181                 |
| 26  | 22516.20          | 406.1           | 22097.7| 22503.8     | 0.6497                 | 0.7136                 |
| 27  | 51085.15          | 4823.8          | 46267.2| 51091.0     | 0.7028                 | 0.7897                 |
| 28  | 99387.01          | 6636.1          | 92717.6| 99353.7     | 0.7460                 | 0.8504                 |
| 29  | 169303.03         | 37415.5         | 131882.6| 169298.1   | 0.7805                 | 0.8994                 |
| 30  | 256358.10         | 7.91            | 8.3    | 8.3         | 0.1341                 |                        |

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

This is a provisional file, not the final typeset article
Sample

Prediction of repetitions

| $p_{\text{max}}$ | $k_{\text{min}}$ | $k$ | $\bar{k}$ | $k_{\text{max}}$ | $k_{\text{ext}}$ |
|------------------|------------------|------|----------|------------------|------------------|
| 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 $p_{\text{max}}$ for binary alphabets. $n = 5000000$

### 3.2 LONG REPETITIONS IN ARCHAEA GENOMES

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

| $k$ | observed | $B(k+1)$ | $S(k)$ | $T(k)$ | $\mu(n,k)$ | $k_{\text{min}}$ | $\bar{k}$ | $k_{\text{max}}$ | $k_{\text{ext}}$ |
|-----|----------|----------|--------|--------|------------|------------------|---------|------------------|------------------|
| 6   | 4        | 0        | 0.05   | 0.05   |            |                  |         |                  |                  |
| 7   | 1975     | 0        | 4e+02  | 4e+02  |            |                  |         |                  |                  |
| 8   | 41349    | 0        | 2e+04  | 2e+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  | 570499   | 407976.5 | 215942.2 | 623918.7 |            |                  |         |                  |                  |
| 13  | 459330   | 259860.7 | 6512.5 | 266373.2 |            |                  |         |                  |                  |
| 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      | 1e-05    | 0      | 1e-05    |            |                  |         |                  |                  |
| 32  | 359      | 2e-06    | 0      | 2e-06    |            |                  |         |                  |                  |
| 33  | 322      | 2e-08    | 0      | 2e-08    |            |                  |         |                  |                  |

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 GC-content is 0.6664. Parameters for an approximate degenerated quaternary model are $p_A = p_T = p_{\text{min}} = 0.1680; p_C = p_G = p_{\text{max}} = 0.3332 \ ; \tilde{p} = 0.2645$; and $\log n = 14.8620$. The thresholds for the domain are

$$
\begin{align*}
\alpha_{\text{min}} &= 0.5584 & \tilde{\alpha} &= 0.7520 & \bar{\alpha} &= 0.8079 & \alpha_{\text{max}} &= 0.9099 & \alpha_{\text{ext}} &= 1.5609 \\
k_{\text{min}} &= 8.30 & \tilde{k} &= 11.18 & k &= 12.01 & k_{\text{max}} &= 13.52 & k_{\text{ext}} &= 23.20
\end{align*}
$$

Statistics on one hundred random sequences with same parameters are shown on figure 3.2(b). GC-content 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.

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,

| $k$ | observed $B(n,k+1)$ | predicted $S(k)$ | predicted $T(k)$ | observed $\log B(n,k+1)$ | asymptotic $\log n$ | $\psi(\alpha)$ | $\psi(\alpha) + \xi\bar{\alpha}$ |
|-----|---------------------|------------------|------------------|--------------------------|---------------------|----------------|-------------------------------|
| 6   | 0.03                | 0                | 0.0              | -0.2359                  | 0.7242              | 0.9280         | 0.6956                        |
| 7   | 363.29              | 0                | 363.9            | 0.3967                   | 0.6704              | 0.9235         | 0.7564                        |
| 8   | 21236.17            | 0                | 21252.2          | 0.6704                   | 0.9235              | 0.9985         | 0.7564                        |
| 9   | 214371.12           | 781.6            | 213574.7        | 0.8260                   | 0.7242              | 0.9280         | 0.6956                        |
| 10  | 684344.68           | 66877.4          | 617315.1        | 0.9041                   | 0.9280              | 0.9235         | 0.7564                        |
| 11  | 914013.67           | 171742.8         | 742383.0        | 0.9235                   | 0.9280              | 0.9985         | 0.7564                        |
| 12  | 623870.12           | 407973.4         | 215914.6        | 0.8978                   | 0.9235              | 0.9985         | 0.7147                        |
| 13  | 266366.73           | 259826.1         | 6510.8          | 0.8978                   | 0.9235              | 0.9985         | 0.7147                        |
| 14  | 87424.58            | 87471.6          | 0               | 0.7656                   | 0.9235              | 0.9985         | 0.7147                        |
| 15  | 25704.95            | 25698.5          | 0               | 0.6832                   | 0.9235              | 0.9985         | 0.7147                        |
| 16  | 7253.72             | 7262.9           | 0               | 0.5981                   | 0.9235              | 0.9985         | 0.7147                        |
| 17  | 2025.99             | 2027.6           | 0               | 0.5123                   | 0.9235              | 0.9985         | 0.7147                        |
| 18  | 565.97              | 563.9            | 0               | 0.4265                   | 0.9235              | 0.9985         | 0.7147                        |
| 19  | 155.90              | 156.7            | 0               | 0.3397                   | 0.9235              | 0.9985         | 0.7147                        |
| 20  | 43.52               | 43.5             | 0               | 0.2539                   | 0.9235              | 0.9985         | 0.7147                        |
| 21  | 12.28               | 12.1             | 0               | 0.1688                   | 0.9235              | 0.9985         | 0.7147                        |
| 22  | 3.06                | 3.4              | 0               | 0.0753                   | 0.9235              | 0.9985         | 0.7147                        |
| 23  | 0.80                | 0.9              | 0               | -0.0150                  | 0.9235              | 0.9985         | 0.7147                        |
| 24  | 0.28                | 0.3              | 0               | -0.0857                  | 0.9235              | 0.9985         | 0.7147                        |
| 25  | 0.14                | 0.1              | 0               | -0.1323                  | 0.9235              | 0.9985         | 0.7147                        |

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

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

$$
F = \sum_{i=1}^{V} \theta_i \log \theta_i ;
$$

(20)

$$
G = \sum_{i=1}^{V} \theta_i ;
$$

(21)

$$
H = \sum_{i=1}^{V} \theta_i \beta_i .
$$

(22)

Two constraints are given :

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

An intermediary function $\phi_{\alpha}(\tau_1, \cdots, \tau_V)$ is defined

$$
\phi_{\alpha} = F + \lambda_{\alpha} G + \tau_{\alpha} H
$$

(23)

In order to maximize $\phi$ under these two constraints, $\phi$ function is derived with respect to each random
variable $\tau_i$. This yields $V$ equations

$$
1 + \log \theta_i + \lambda_{\alpha} + \tau_{\alpha} \beta_i = 0 .
$$

(24)

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

$$
\beta_{i_{\text{min}}} = \min(\beta_i)_{1 \leq i \leq V} = \log \frac{1}{p_{\text{max}}} ;
$$

$$
\beta_{i_{\text{max}}} = \max(\beta_i)_{1 \leq i \leq V} = \log \frac{1}{p_{\text{min}}} .
$$
Solving Equations (24) with indices $i_{\text{min}}$ and $i_{\text{max}}$ yields

$$\tau_{\alpha} = \log \frac{\theta_{i_{\text{min}}} - \log \theta_{i_{\text{max}}}}{\beta_{i_{\text{max}}} - \beta_{i_{\text{min}}}} = \log \frac{\theta_{i_{\text{min}}}}{\theta_{i_{\text{max}}}} \frac{1}{\beta_{i_{\text{max}}} - \beta_{i_{\text{min}}}} ;$$

$$1 + \lambda_{\alpha} = \frac{\beta_{i_{\text{min}}} \log \theta_{i_{\text{max}}} - \beta_{i_{\text{max}}} \log \theta_{i_{\text{min}}}}{\beta_{i_{\text{max}}} - \beta_{i_{\text{min}}}} .$$

Remaining equations rewrite:

$$\log \theta_i = \log \theta_{i_{\text{min}}} + \tau_{\alpha} (\beta_{i_{\text{min}}} - \beta_i) . \quad (25)$$

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

$$\theta_{i_{\text{min}}} e^{\beta_1 \tau_{\alpha}} \sum_{i=1}^{V} e^{-\beta_i \tau_{\alpha}} = 1 ,$$

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

$$\theta_i = \frac{e^{-\beta_i \tau_{\alpha}} \sum_{i=1}^{V} e^{-\beta_i \tau_{\alpha}}}{\sum_{i=1}^{V} \beta_i e^{-\beta_i \tau_{\alpha}}} . \quad (26)$$

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

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

For this $V$-tuple

$$\sum_{i=1}^{V} \theta_i \log \theta_i = -(\sum_{i=1}^{V} \theta_i \beta_i) \tau_{\alpha} - (\sum_{i=1}^{V} \theta_i) \log(\sum_{i=1}^{V} e^{-\beta_i \tau_{\alpha}}) = -\frac{\tau_{\alpha}}{\alpha} - \log(\sum_{i=1}^{V} e^{-\beta_i \tau_{\alpha}}) .$$

### 4.2 APPROXIMATION ORDERS

Deriving the RHS of (10) yields $\frac{\sum_{i \neq j} (\beta_i + \beta_j)^2 e^{-(\beta_i + \beta_j)\tau_{\alpha}}}{(\sum_i e^{-\beta_i \tau_{\alpha}})^2}$ that is positive. Therefore, for any $\alpha$, the solution to (10) is unique. Moreover, $\tau_{\alpha}$ increases with $\alpha$. Let

$$\psi_1(\alpha) = \tau_{\alpha} + \alpha \log(\sum_{i=1}^{V} e^{-\beta_i \tau_{\alpha}}) ; \quad (27)$$

$$\psi_2(\alpha) = 2 - \alpha \log \frac{1}{\sigma_2} . \quad (28)$$

Notably, the solutions $\tau_{\alpha}$ of (10) associated to the four increasing values of $\alpha : (\alpha_{\text{min}}, \tilde{\alpha}, \bar{\alpha}, \alpha_{\text{max}})$ are $(-\infty, 1, 2, +\infty)$. Computing $\psi$ for these values yields (11) and Equality $\psi_1(\tilde{\alpha}) = \psi_2(\bar{\alpha})$.
Derivating both expressions yields
\[ \frac{\partial \psi}{\partial \alpha}(\alpha) = \log(\sum_{i=1}^{V} e^{-\beta_i \tau_\alpha}) ; \]  
(29)
\[ \frac{\partial \psi_1}{\partial \alpha}(\alpha) - \frac{\partial \psi_2}{\partial \alpha}(\alpha) = \log(\sigma_2 \sum_{i=1}^{V} e^{-\beta_i \tau_\alpha}) \]  
(30)

Both derivatives are monotone functions of \( \tau_\alpha \). In (30), derivative is 0 when \( \alpha = \tilde{\alpha} \). Therefore, \( \psi \) is the 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

**Short Lengths** Assume that \( k \leq \alpha_{\min} \log n \). Each term \( \phi(k_1, \cdots, k_V) \) is lower bounded by \( p_{\min}^k = n^{\alpha \log p_{\min}} = n^{-\frac{\alpha}{\alpha_{\min}}} \). Each term \( \psi_n(k_1, \cdots, k_V) \) is trivially bounded by \( e^{-n^{1-\frac{\alpha}{\alpha_{\min}}}} \) that is upper bounded by 1 and \( n\psi_n(k_1, \cdots, k_V) \) tends to 0 when \( n \) goes to \( \infty \). As \( \sum \left( \frac{k}{k_1 \cdots k_V} \right) \phi(k_1, \cdots, k_V) = 1 \), the ratio \( \frac{\log \mu(n,k)}{\log n} \) is negative.

**Moderate and Large Lengths** For a length \( k \) in the transition domain \([\alpha_{\min} \log n, \alpha_{\max} \log n]\), the objective function may be either positive or negative. When \( k > \alpha_{\max} \log n \), set \( E_k(n) \) is empty and \( \mu(n,k) \) reduces to \( S(k) \).

The maximum \( M \) among the terms \( e^{k(-\sum_{i} \frac{k_i}{k} \log \frac{k_i}{k} - \frac{1}{\alpha} \log n \phi(k_1, \cdots, k_V))} \) in \( T(k) \) is reached when \( \rho(k_1, \cdots, k_V) \) is 0. Due to the exponential decrease of \( e^{-n^{1-\frac{\alpha}{\alpha_{\min}}}} \) when \( n \phi(k_1, \cdots, k_V) \geq 1 \), \( T(k) \) is upper bounded. Computation of \( \log M \) is done with Lagrange multipliers, as explained above.

Computation of \( S(k) \) relies on the local development of \( \psi_n(k_1, \cdots, k_V) \), that is \( n(1-\sigma_2)\phi(k_1, \cdots, k_V) \).

\( S(k) \) rewrites \( \sigma_2^k \hat{S}(k) + (S(k) - \sigma_2^k \hat{S}(k)) \) where \( \hat{S}(k) = \sum_{\rho(k_1, \cdots, k_V) \leq 0} \left( \frac{k}{k_i} \right) \left( \frac{p_{12}}{\sigma_2} \right)^{k_1} \cdots \left( \frac{p_{12}}{\sigma_2} \right)^{k_V} \).

This sum satisfies a Large Deviation Principle when \( \rho(k_1, \cdots, k_V) + \frac{1}{\alpha} \geq \frac{1}{\alpha} \), or \( \alpha < \tilde{\alpha} \). In this range, \( \frac{\hat{S}(k)}{k} \sim \max\{-\sum_{i=1}^{V} \frac{k_i}{k} \log \frac{k_i}{k}\} \), that was shown to be \( \psi(\alpha) \).

When \( \alpha > \tilde{\alpha} \), sum \( \hat{S}(k) \) rewrites 1 - \( \hat{S}(k) \) where
\[ \hat{S}(k) = \sum_{\rho(k_1, \cdots, k_V) + \frac{1}{\alpha} < \frac{1}{\alpha}} \left( \frac{k}{k_i} \right) \left( \frac{p_{12}}{\sigma_2} \right)^{k_1} \cdots \left( \frac{p_{12}}{\sigma_2} \right)^{k_V} \).

This sum satisfies a Large Deviation Principle and \( \frac{\hat{S}(k)}{k} \sim \max\{-\sum_{i=1}^{V} \frac{k_i}{k} \log \frac{k_i}{k} + \sum_{i=1}^{V} \frac{k_i}{k} \log \frac{p_{12}}{\sigma_2}\} \). As \( \sum_{i=1}^{V} \frac{k_i}{k} \log \frac{p_{12}}{\sigma_2} = -\frac{2}{\alpha} + \log \frac{1}{\sigma_2} \), this maximum is
\[ -\frac{1}{\alpha} \left[ 2 - \alpha \log \frac{1}{\sigma_2} - \psi(\alpha) \right] \]
that is negative.

Frontiers in Bioengineering and Biotechnology
4.4 BINARY CASE

Barycentric coordinates of $\alpha$ are unique. Indeed, (10) reduces to a linear equation on the variable $\tau$.

\[
\frac{1}{\alpha} = \frac{\beta_1 + \beta_2 e^{-(\beta_2 - \beta_1)\tau}}{1 + e^{-(\beta_2 - \beta_1)\tau}}
\]

where $\beta_2 - \beta_1 = \beta_{\min} - \beta_{\max} = \log \frac{p_{\max}}{p_{\min}}$. Therefore, $e^{-(\beta_2 - \beta_1)\tau} = \frac{1 - \alpha \beta_1}{\alpha \beta_2 - 1}$. Finally

\[
\tau_\alpha = \frac{1}{\log \frac{p_{\max}}{p_{\min}}} \log \frac{\alpha \beta_2 - 1}{1 - \alpha \beta_1} = \frac{1}{\log \frac{p_{\max}}{p_{\min}}} \log \frac{\frac{1}{\beta_{\min}} - \frac{1}{\alpha}}{\frac{1}{\beta_{\max}} - \frac{1}{\alpha}}.
\]

Function $\psi$ rewrites, in the binary case:

\[
\psi_\alpha = \tau_\alpha = \alpha \log e^{-\frac{1}{\alpha} \tau_\alpha} \left( e^{-(\beta_1 - \frac{1}{\alpha}) \tau_\alpha} + e^{-(\beta_2 - \frac{1}{\alpha}) \tau_\alpha} \right).
\]

Observing that $e^{-(\beta_2 - \beta_2)\tau_\alpha} = s_\alpha$ and changing variable $\tau_\alpha$ into $(\beta_2 - \beta_1)$ yields $e^{-(\beta_1 - \frac{1}{\alpha}) \tau_\alpha} = s_\alpha^{-(\frac{1}{\beta_{\min}} - \frac{1}{\beta})}$ and $e^{-(\beta_2 - \frac{1}{\alpha}) \tau_\alpha} = s_\alpha^{-(\frac{1}{\beta_{\max}} - \frac{1}{\beta})}$.

5 CONCLUSION

This paper describes the behaviour of the number of unique or repeated $k$-mers in a random sequence, on a general alphabet. Derivation relies on a combination of analytic combinatorics and on Lagrange multipliers. It simplifies an approach provided for binary alphabets and allows to address larger alphabets, including the quaternary alphabets such as DNA alphabet. Precise asymptotic estimates are provided and a probabilistic interpretation is given. They are validated on random simulated data and shown to be valid in the finite range. Therefore, they provide a valuable tool to estimate a suitable read length for assembly purposes and tune parameters for assembly algorithms. Real genomes significantly depart from the random behaviour for long repetitions. The general shape of the trie profile is observed, with a maximum of the number of unique $k$-mers at the expected length. However, for real genomes, a number of very short $k$-mers are missing and, on the contrary, one observes a number of very long repetitions. Besides these events, the behaviours are rather similar.

In the future, it is worth extending the method to generalized Patricia tries, Markov models and 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

Funding: INRIA-CNRS-PONECLET 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 hashing a 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. Martinez, 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., Nicodème, 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/btt200.

Sedgewick, R. and Flajolet, P. (2009), Analytic combinatorics, *Cambridge University Press*.

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.

Frontiers in Bioengineering and Biotechnology