J. Appl. Prob. **45,** 440–455 (2008)
Printed in England
© *Applied Probability Trust* 2008

POISSON APPROXIMATION FOR THE NUMBER OF REPEATS IN A STATIONARY MARKOV CHAIN

NARJISS TOUYAR,* ** *Université de Rouen*
SOPHIE SCHBATH,*** *Institut National de la Recherche Agronomique*
DOMINIQUE CELLIER * **** AND
HÉLÈNE DAUCHEL,* ***** *Université de Rouen*

Abstract

Detection of repeated sequences within complete genomes is a powerful tool to help understanding genome dynamics and species evolutionary history. To distinguish significant repeats from those that can be obtained just by chance, statistical methods have to be developed. In this paper we show that the distribution of the number of long repeats in long sequences generated by stationary Markov chains can be approximated by a Poisson distribution with explicit parameter. Thanks to the Chen–Stein method we provide a bound for the approximation error; this bound converges to 0 as soon as the length n of the sequence tends to ∞ and the length t of the repeats satisfies $n^2\rho^t = \mathrm{O}(1)$ for some $0 < \rho < 1$. Using this Poisson approximation, p-values can then be easily calculated to determine if a given genome is significantly enriched in repeats of length t.

Keywords: Poisson approximation; number of repeats; Chen–Stein method; Markov chain; DNA sequence

2000 Mathematics Subject Classification: Primary 62E17
Secondary 60C05

1. Introduction

Genomes are dynamic and redundant structures: during the life of an organism or over generations in a given species, genomes are regularly subject to various small-scale or large-scale rearrangements such as sequence inversions, insertions, deletions, or duplications. Thus, the current architecture of genomes is the result of the accumulation of numerous past molecular events that lead to DNA remodeling. Among those, the process of duplication, which can concern the genes or the extragenic segments of the genome, has a major contribution. After they occur, genome duplications bring a primary redundancy of the genetic information, but, secondarily, the duplicated genome segments provide potential substrates for the evolution to new genomic functions by other types of mutations. Thus, the duplication events are the essential engines of the genomes' evolution, leading to a functional diversity, genetic

Received 27 October 2006; revision received 14 March 2008.
* Postal address: Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes EA4051, Université de Rouen, 76821 Mont Saint Aignan Cedex, France.
** Email address: narjiss.touyar@univ-rouen.fr
*** Postal address: INRA, UR1077 Mathématique, Informatique et Génome, Domaine de Vilvert, 78352 Jouy-en-Josas, France. Email address: sophie.schbath@jouy.inra.fr
**** Email address: dominique.cellier@univ-rouen.fr
***** Email address: helene.dauchel@univ-rouen.fr

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

innovations, which offer an adaptability to new environments, and the creation of new organisms (for a review, see [10]).

Hence, the detection of the redundant regions (also named repeated sequences) within a fully sequenced current genome, provides biologists with the opportunity to understand the dynamics of genomes and the species history. Several algorithmic methods have been proposed for efficiently detecting tandem or dispersed repeats in DNA sequences; see [3], [4], [5], [6], and [7]. But, a relevant statistical analysis of these repeats (count, length, location) should then be done in order to distinguish the significant repeats from those that can be just obtained 'by chance'. Significant repeats are then candidates for further biological investigations of their nature, dynamic, or functional implications.

Studying the statistical significance of the number of repeats of a given length t observed in a given sequence, denoted by N_t^{obs}, relies on the possibility of evaluating the distribution of the random count N_t in some relevant random sequences. Indeed, it will allow us to evaluate the p-value $\mathrm{P}(N_t \geq N_t^{\text{obs}})$ and then to decide if a sequence is significantly rich or sparse in repeats of size t; this p-value measures how far the observed number of repeats is from the expected count under the chosen model. No result exists on the exact distribution of N_t in random sequences. Arratia *et al.* [1] proposed a Poisson approximation for long repeats in a sequence of independent random letters (Bernoulli model). In this paper we generalize this result by considering Markov models. While Bernoulli models only fit the letter composition of the DNA sequence, Markov chain models allow us to fit the 1-letter up to an $(m+1)$-letter word composition, where m is the order of the Markov model. Markov models are widely used in biological sequence analyses and the question of finding motifs with unexpected frequencies highlighted the interest of using Markov models of order $m \geq 1$. (See [8] and [9].)

The number of repeats of length t is defined by a sum of random Bernoulli variables Y_α which are equal to 1 if a repeat of length t starts at position $\alpha = (i, j)$ and 0 otherwise (Section 2). In this problem the difficulty comes from the fact that the Y_αs are not independent. As in [1], we have used the Chen–Stein method to bound the error while approximating N_t by a Poisson variable with mean $\lambda_t := \mathrm{E}(N_t)$. We first considered the first-order Markov chain model. We show that this error converges to 0 as the length n of the sequence tends to ∞ and the size t of the repeats grows like $\log(n)$ (Section 3). Finally, we generalize the approximation theorem to higher-order Markov chains and we derive the parameter of the limiting Poisson distribution for the number of repeats.

2. Number of repeats

2.1. Random sequences

We consider an infinite random sequence $S^\infty = X_{-\infty}, \ldots, X_1, \ldots, X_n, \ldots, X_{+\infty}$ on the alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ generated by a stationary one-order Markov chain with transition matrix $\mathbf{\Pi} = (\pi(a, b))_{a,b\in\mathcal{A}}$. We assume that $\pi(a, b) > 0$ for all $a, b \in \mathcal{A}$, which is satisfied with long DNA sequences. This model will be referred to as 'model M1' in the remainder. We denote by μ the unique stationary distribution, defined by $\mu(a) = \sum_{b\in\mathcal{A}} \mu(b)\pi(b, a)$ for all $a \in \mathcal{A}$. Moreover, the ℓ-step transition probability between a and b will be denoted by $\pi^{(\ell)}(a, b)$, and we set $\rho := \max_{a,b\in\mathcal{A}} \pi(a, b)$, $0 < \rho < 1$. Therefore, $\pi^{(\ell)}(a, b) < \rho$ for all $\ell \geq 1$.

2.2. Repeat occurrence and probability

We say that a *repeat* of length t (or more) occurs in S^∞ at position $\alpha = (i, j)$, $i < j$, if and only if the word of t letters starting at i is identical to the one starting at j. Let R_α be the

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

random indicator function of a repeat of length t occurring at position $\alpha = (i, j),\ i < j$:

$$R_\alpha = \mathbf{1}\{X_i \cdots X_{i+t-1} = X_j \cdots X_{j+t-1}\}. \tag{2.1}$$

We say that a repeat of length t (or more) *starts* at $\alpha = (i, j)$ if there is a repeat of length t at (i, j) but not at $(i-1, j-1)$. For $\alpha = (i, j),\ i < j$, we define the random indicator function of a repeat of length t starting in S^∞ at position α by

$$Y_\alpha \equiv Y_{(i,j)} = \mathbf{1}\{X_{i-1} \neq X_{j-1},\ X_i \cdots X_{i+t-1} = X_j \cdots X_{j+t-1}\}. \tag{2.2}$$

Because of biological considerations, only disjoint duplications will be considered (i.e. $j \geq i + t$). For the sake of simplicity, we will in fact restrict ourselves to the case where $j > i + t$. Since the position $j - 1$ is part of the event 'a repeat starts at $\alpha = (i, j)$', the condition $j > i + t$ just allows us to obtain disjoint blocks of letters $X_i \cdots X_{i+t-1}$ and $X_{j-1} \cdots X_{j+t-1}$, which will simplify the expression of the probability that a repeat starts at a given position and other probabilities of that kind. In the remainder a repeat starting at $\alpha = (i, j)$ with $j > i + t$ will be called a leftmost non-self-overlapping repeat.

The following lemma gives the probability p_α for a non-self-overlapping repeat of length t to start at position α.

Lemma 2.1. *Under model M1, the probability p_α for a repeat of length t to start at $\alpha = (i, j),\ j > i + t$, i.e. $p_\alpha = \mathrm{E}(Y_\alpha)$, is*

$$p_\alpha = \sum_{b,c \neq b} \sum_{(a_1,\ldots,a_t) \in \mathcal{A}^t} \mu(b)\pi(b, a_1)\pi(c, a_1)(\pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t))^2 \pi^{(j-i-t)}(a_t, c).$$

Proof. To calculate p_α, we sum over all possible values for $X_{i-1}, X_i, \ldots, X_{i+t-1}$ and X_{j-1}, and we use the Markov property, i.e.

$$\begin{aligned}
\mathrm{E}(Y_\alpha) &= \mathrm{P}(X_{i-1} \neq X_{j-1},\ X_i = X_j, \ldots, X_{i+t-1} = X_{j+t-1}) \\
&= \sum_{b,c \neq b} \sum_{(a_1,\ldots,a_t) \in \mathcal{A}^t} \mathrm{P}(X_{i-1} = b,\ X_i = a_1, \ldots, X_{i+t-1} = a_t,\ X_{j-1} = c, \\
&\qquad\qquad X_j = a_1, \ldots, X_{j+t-1} = a_t) \\
&= \sum_{b,c \neq b} \sum_{(a_1,\ldots,a_t) \in \mathcal{A}^t} \mu(b)\pi(b, a_1)\pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t)\pi^{(j-i-t)}(a_t, c) \\
&\qquad\qquad \times \pi(c, a_1)\pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t).
\end{aligned}$$

This probability p_α of the occurrence of a repeat of length t decreases exponentially to 0 as t increases. Indeed, we can show that

$$p_\alpha \leq \rho^t. \tag{2.3}$$

To obtain this inequality, we bound Y_α by R_α, i.e.

$$\begin{aligned}
p_\alpha &\leq \mathrm{E}(R_\alpha) \\
&= \sum_{(a_1,\ldots,a_t) \in \mathcal{A}^t} \mu(a_1)\pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t) \\
&\qquad\qquad \times (\pi^{(j-i-t+1)}(a_t, a_1)\pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t)),
\end{aligned}$$

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

and we bound each of the last t transition probabilities by ρ. The sum of

$$\mu(a_1)\pi(a_1, a_2)\cdots\pi(a_{t-1}, a_t)$$

over $a_1, \ldots, a_t$ will then collapse to 1 leading to $p_\alpha \le \rho^t$.

2.3. Number of leftmost non-self-overlapping repeats

In this paper we are interested in the number of non-self-overlapping repeats of length t or more in a finite sequence $S = X_1, \ldots, X_n$. Let us define the following count:

$$N_t = \sum_{\alpha\in I} Y_\alpha, \tag{2.4}$$

where $I = \{\alpha = (i, j) \mid 1 \le i < i + t < j \le n - t + 1\}$. Because S is finite and X_0 is not observed, N_t is not exactly the number of non-self-overlapping repeats of length t in S but the probability that both counts differ is bounded by $(n - 2t)\rho^t$. Indeed, if they differ then there exists $j \in \{t + 1, \ldots, n - t + 1\}$ such that $X_1 \cdots X_t = X_j \cdots X_{j+t-1}$. Therefore, under the asymptotic framework $n^2\rho^t = \mathrm{O}(1)$ and $t = \mathrm{o}(n)$, both counts have asymptotically the same distribution and we will now focus on the count N_t defined by (2.4) in the infinite sequence S^∞.

Set $\lambda_t := \mathrm{E}(N_t)$, the expected number of leftmost non-self-overlapping repeats of length t. Its expression is given in the next lemma.

Lemma 2.2. *Under model M1, the expected number of leftmost non-self-overlapping repeats of length t is given by*

$$\lambda_t = \sum_{b,c\ne b} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}^t} \mu(b)\pi(b, a_1)\pi(c, a_1)(\pi(a_1, a_2)\cdots\pi(a_{t-1}, a_t))^2 \times \sum_{\ell=0}^{n-2t-1} (n - \ell - 2t)\pi^{(\ell+1)}(a_t, c). \tag{2.5}$$

If $t = o(n)$ then $\lambda_t = \mathrm{O}(n^2\rho^t)$, where $\rho = \max_{a,b}\pi(a, b)$.

Proof. Using the expression of $\mathrm{E}(Y_\alpha)$, Lemma 2.1 leads to

$$\begin{aligned}\lambda_t &= \sum_{\alpha\in I} \mathrm{E}(Y_\alpha)\\ &= \sum_{i=1}^{n-2t} \sum_{j=i+t+1}^{n-t+1} \sum_{b,c\ne b} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}^t} \mu(b)\pi(b, a_1)\pi(c, a_1) \times (\pi(a_1, a_2)\cdots\pi(a_{t-1}, a_t))^2\pi^{(j-i-t)}(a_t, c).\end{aligned}$$

Equation (2.5) holds using the change of variable $\ell := j - i - t - 1$, i.e.

$$\sum_{i=1}^{n-2t} \sum_{j=i+t+1}^{n-t+1} \pi^{(j-i-t)}(a_t, c) = \sum_{\ell=0}^{n-2t-1} (n - \ell - 2t)\pi^{(\ell+1)}(a_t, c).$$

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

To obtain the order of magnitude of λ, we just use (2.3):

$$\begin{aligned}\lambda_t &\le \rho^t \sum_{i=1}^{n-2t} (n-i-2t+1) \\ &\le \rho^t \frac{(n-2t)(n-2t+1)}{2}.\end{aligned}$$

Consequently, if $t = o(n)$ then $\lambda = O(n^2\rho^t)$.

3. Poisson approximation

Our aim is now to prove that N_t can be approximated by a Poisson variable with mean λ_t when n tends to ∞ and $\lambda_t = O(1)$. Note that the asymptotic condition $\lambda_t \asymp 1$ is equivalent to $t \asymp 2\log_{1/\rho}(n)$. For this, we use the Chen–Stein method.

3.1. Chen–Stein method

The Chen–Stein method gives a bound of the total variation distance between the distribution of a sum of non independent and identically distributed Bernoulli variables Y_α, $\alpha \in I$, and the distribution of a Poisson variable with parameter $\lambda = \sum_{\alpha\in I} \mathrm{E}(Y_\alpha)$. Recall that the total variation distance (denoted by d_{TV}) between two positive integer random variables X and Y is defined in the same way as half the maximum over $i \in \mathcal{N}$ of $|\mathrm{P}(X = i) - \mathrm{P}(Y = i)|$.

Theorem 3.1. ([2, Theorem 1.A].) *Let I be an index set. Suppose that, for each $\alpha \in I$, Y_α is a Bernoulli random variable with $p_\alpha = \mathrm{P}(Y_\alpha = 1) > 0$. Let Z_α be independent Poisson variables with mean p_α and take B_α to be a neighborhood of α such that $\alpha \in B_\alpha \subset I$. Let*

$$\begin{aligned}b_1 &= \sum_{\alpha\in I} \sum_{\beta\in B_\alpha} \mathrm{E}(Y_\alpha)\,\mathrm{E}(Y_\beta), \\ b_2 &= \sum_{\alpha\in I} \sum_{\beta\in B_\alpha,\, \beta\neq\alpha} \mathrm{E}(Y_\alpha Y_\beta), \\ b_3 &= \sum_{\alpha\in I} \mathrm{E}\,|\,\mathrm{E}(Y_\alpha - p_\alpha \mid \sigma(Y_\beta : \beta \notin B_\alpha))|.\end{aligned}$$

Thus,

$$d_{\mathrm{TV}}\left(\mathcal{L}\left(\sum_\alpha Y_\alpha\right), \mathcal{L}\left(\sum_\alpha Z_\alpha\right)\right) \le b_1 + b_2 + b_3,$$

where $\mathcal{L}(\cdot)$ denotes the distribution of a random variable.

To prove a Poisson approximation for the number N_t of leftmost non-self-overlapping repeats of length t, we will apply this theorem to the Bernoulli variables Y_α defined by (2.2). We will first choose an appropriate neighborhood B_α, then we will bound the quantities b_1, b_2, and b_3 and show that they converge to 0 when $n \to +\infty$ and $n^2\rho^t = O(1)$.

3.2. Choice of the neighborhoods

The neighborhood B_α has to be interpreted as a neighborhood of strong dependence between Y_α and Y_β, $\beta \in B_\alpha$. Intuitively, we could say that $\beta = (i', j')$ is not a neighbor of $\alpha = (i, j)$ as soon as the occurrences of t-letter words occurring at positions i' and j' are disjoint from the t-letter words occurring at i and j. As we will see later, for b_3 to converge to 0, we need

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

to enlarge the neighborhood such that the occurrences at positions (i, j) are not only disjoint from the occurrences at positions (i', j') but separated apart from t letters. In other words, we take

$$B_\alpha := \{\beta = (i', j') \in I \mid \min\{|i - i'|, |i - j'|, |j - i'|, |j - j'|\} < 2t\}, \qquad \alpha = (i, j) \in I.$$

Let us calculate the size of this neighborhood, or more precisely the number of pairs $(\alpha, \beta) \in I^2$ which are neighbors. Set $G = \{(\alpha, \beta) \in I^2 \mid \beta \in B_\alpha\}$, and let H be its complement in I^2. Therefore, we have

$$|G| = |I|^2 - |H|,$$

where $|\cdot|$ stands for the cardinality and H is given by

$$\begin{aligned} H &= \{(\alpha, \beta) \in I^2 \mid \beta \notin B_\alpha\} \\ &= \{(i, i', j, j') \in \{1, \dots, n - t + 1\}^4 \text{ such that } j - i > t,\ j' - i' > t,\ |i - i'| \geq 2t, \\ &\qquad |j - j'| \geq 2t,\ |i - j'| \geq 2t,\ |i' - j| \geq 2t\}. \end{aligned}$$

The cardinality of I is

$$|I| = \binom{n - 2t + 1}{2} = \frac{(n - 2t + 1)(n - 2t)}{2},$$

and the cardinality of H satisfies

$$6\binom{n - 7t + 4}{4} \leq |H| \leq 6\binom{n - 4t + 4}{4}.$$

Indeed, we have $\binom{4}{2}|J_{2t}| \leq |H| \leq \binom{4}{2}|J_t|$, where

$$J_z = \{(x_1, x_2, x_3, x_4) \in \{1, \dots, n - t + 1\}^4 \text{ such that } x_{i+1} \geq x_i + z \text{ for all } i \in \{1, 2, 3\}\}.$$

Moreover, there exists a one-to-one transformation between J_z and the set of quadruples (y_1, y_2, y_3, y_4) such that $1 \leq y_1 < y_2 < y_3 < y_4 \leq n - t + 1 - 3(z - 1)$: set $y_i = x_i - (i - 1)z$ for all $i \in \{1, 2, 3\}$. Therefore, $|J_z| = \binom{n - t - 3z + 4}{4}$. Finally, we obtain

$$2n^3 t + \mathrm{o}(n^3 t) \leq |G| \leq 5n^3 t + \mathrm{o}(n^3 t),$$

when $t = \mathrm{o}(n)$. Thus, $|G| \asymp n^3 t$ as $n \to +\infty$.

3.3. A bound for b_1

Using (2.3), we have

$$b_1 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \mathrm{E}(Y_\alpha)\,\mathrm{E}(Y_\beta) \leq |G|\rho^{2t}.$$

Since $|G| \asymp n^3 t$, we obtain

$$b_1 = \mathrm{O}\left(\frac{t}{n}(n^2\rho^t)^2\right). \tag{3.1}$$

Under the asymptotic conditions $n^2\rho^t = \mathrm{O}(1)$ and $t = \mathrm{o}(n)$, b_1 will converge to 0 as $n \to \infty$.

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

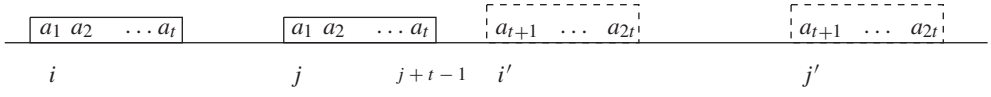

FIGURE 1: Two repeats at (i, j) and (i', j') with no overlap.

3.4. A bound for b_2

Bounding b_2 requires calculating $\mathrm{E}(Y_\alpha Y_\beta)$, i.e. the probability that a repeat starts at position $\alpha = (i, j)$ and another repeat starts at position $\beta = (i', j')$ for $\beta \in B_\alpha$ and $\beta \neq \alpha$. We can establish that $\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}$ (cf. Proposition 3.1, below). It follows that

$$b_2 = \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} \mathrm{E}(Y_\alpha Y_\beta) \leq |G| \rho^{2t}.$$

Since $|G| \asymp n^3 t$, we obtain

$$b_2 = \mathrm{O}\left(\frac{t}{n} (n^2 \rho^t)^2 \right), \tag{3.2}$$

as for b_1.

Proposition 3.1. *The probability that two repeats start at positions $\alpha = (i, j)$ and $\beta = (i', j')$ for $\beta \in B_\alpha$ and $\beta \neq \alpha$ satisfies*

$$\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}.$$

Proof. We will distinguish four cases depending on the number d of overlaps between the four t-letter words occurring at positions i, j, i', and j'. Owing to no self-overlapping repeats, d can only take four values: $d \in \{0, 1, 2, 3\}$.

In most cases we will bound Y_α by R_α defined in (2.1), leading to

$$\begin{aligned} \mathrm{E}(Y_\alpha Y_\beta) &\leq \mathrm{E}(R_\alpha R_\beta) \\ &\leq \mathrm{P}(X_i = X_j, \ldots, X_{i+t-1} = X_{j+t-1} \text{ and } X_{i'} = X_{j'}, \ldots, X_{i'+t-1} = X_{j'+t-1}). \end{aligned} \tag{3.3}$$

Case 1: $d = 0$. In this case the four occurrences at i, j, i', and j' are disjoints. We start from (3.3) and we sum over all possible values $(a_1, \ldots, a_{2t}) \in \mathcal{A}^{2t}$ for $X_i, \ldots, X_{i+t-1}$ and $X_{i'}, \ldots, X_{i'+t-1}$, i.e.

$$\begin{aligned} \mathrm{E}(Y_\alpha Y_\beta) \leq \sum_{a_1, \ldots, a_{2t}} \mathrm{P}(&X_i = a_1, \ldots, X_{i+t-1} = a_t,\ X_j = a_1, \ldots, X_{j+t-1} = a_t, \\ &X_{i'} = a_{t+1}, \ldots, X_{i'+t-1} = a_{2t},\ X_{j'} = a_{t+1}, \ldots, X_{j'+t-1} = a_{2t}). \end{aligned}$$

Without loss of generality, suppose that $i < j < i' < j'$ (see Figure 1). We now use the Markov property and bound two particular ℓ-step transition probabilities between the four occurrences by ρ to obtain

$$\begin{aligned} \mathrm{E}(Y_\alpha Y_\beta) &\leq \rho^2 \sum_{a_1, \ldots, a_{2t}} \mu(a_1) \pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t) \\ &\qquad \times \pi^{(i'-j-t+1)}(a_t, a_{t+1}) \pi(a_{t+1}, a_{t+2}) \cdots \pi(a_{2t-1}, a_{2t}) \\ &\qquad \times \pi(a_1, a_2) \cdots \pi(a_{t-1}, a_t) \pi(a_{t+1}, a_{t+2}) \cdots \pi(a_{2t-1}, a_{2t}) \\ &\leq \rho^{2t}. \end{aligned}$$

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

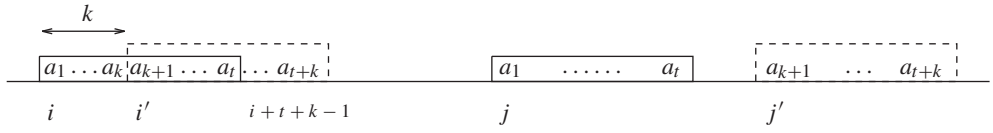

FIGURE 2: Two repeats at (i, j) and (i', j') with one overlap.

The last inequality is obtained by bounding each of the last $(2t - 2)$ transition probabilities by ρ, the remaining sum being equal to 1.

Case 2: $d = 1$. We should distinguish eight symmetrical subcases according to the couple of occurrences which overlap ((i, i'), (i, j'), (i', j), or (j, j')), and, for this couple of overlapping occurrences, which one occurs first? For example, we choose to present the case where occurrences at i and i' overlap and $i < i'$. Let us denote by k the lag between the overlapping occurrences, i.e. here $k = i' - i$ (see Figure 2).

We start from (3.3) and we then sum over all possible $(t + k)$-tuples $(a_1, \ldots, a_{t+k})$ for $X_i, \ldots, X_{i+t+k-1}$; this is sufficient because the t-letter words occurring at positions j and j' will necessarily be $a_1 \cdots a_t$ and $a_{k+1} \cdots a_{t+k}$. It follows that

$$\mathrm{E}(Y_\alpha Y_\beta) \le \sum_{a_1,\ldots,a_{t+k}} \mathrm{P}(X_i = a_1, \ldots, X_{i+t+k-1} = a_{t+k},\ X_j = a_1, \ldots, X_{j+t-1} = a_t,\ X_{j'} = a_{k+1}, \ldots, X_{j'+t-1} = a_{t+k}).$$

As for the case in which $d = 0$, we use the Markov property and we bound the two ℓ-step transition probabilities by ρ to obtain

$$\begin{aligned}\mathrm{E}(Y_\alpha Y_\beta) &\le \rho^2 \sum_{a_1,\ldots,a_{t+k}} \mu(a_1)\pi(a_1, a_2)\cdots\pi(a_{t+k-1}, a_{t+k}) \\ &\qquad \times \pi(a_1, a_2)\cdots\pi(a_{t-1}, a_t)\pi(a_{k+1}, a_{k+2})\cdots\pi(a_{t+k-1}, a_{t+k}) \\ &\le \rho^{2t}.\end{aligned}$$

The last inequality is obtained by bounding each of the last $(2t - 2)$ transition probabilities by ρ, the remaining sum being equal to 1.

Remark 3.1. Note that the exponent of ρ in the bound of $\mathrm{E}(Y_\alpha Y_\beta)$ corresponds to the difference between the total number of transition probabilities multiplied by each other and the number of different letters needed to achieve both repeats plus 1 (due to the first letter). In fact, this remark is general and will be used directly for cases $d = 2$ and $d = 3$; straightforward details will then be omitted.

Case 3: $d = 2$. This is the most complicated case. First of all, we have to consider the following two situations: the two overlaps concern either two pairs of occurrences or three occurrences. In the first case the overlaps are necessarily between occurrences at positions $\{i, i'\}$ and $\{j, j'\}$ ($\{i, j'\}$ and $\{i', j\}$ are not possible since $i < j$ and $i' < j'$). In the second case there are four symmetric cases according to the single occurrence.

Consider the situation where there are two pairs of overlapping occurrences. Let us show that $\mathrm{E}(Y_\alpha Y_\beta) \le \rho^{2t}$ in this case. Set $k = i' - i$ and $k' = j' - j$, the lags between the overlapping occurrences. If $k = k'$ then $Y_\alpha Y_\beta = 0$, owing to the starting condition. We need to distinguish two cases: $kk' > 0$ and $kk' < 0$. Let us start by considering the simplest case when $kk' < 0$.

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

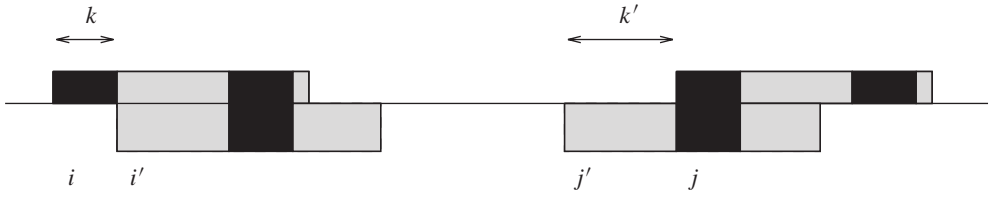

FIGURE 3: Two repeats at (i, j) and (i', j') with two pairs of overlapping occurrences for the special case in which $(i' - i)(j' - j) < 0$. The black rectangles represent the first k letters $a_1, \ldots, a_k$ of the repeat at (i, j) and the gray rectangles represent the first k' letters $b_1, \ldots, b_{k'}$ of the repeat at (i', j') or some prefixes.

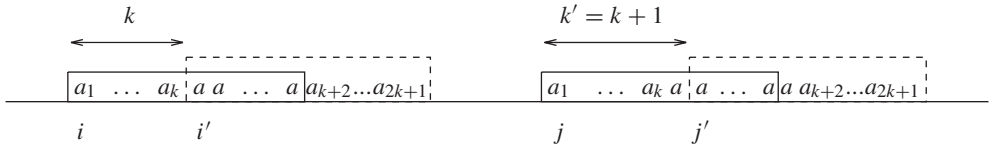

FIGURE 4: Two repeats at (i, j) and (i', j') with two pairs of overlapping occurrences for the special case in which $k' = k + 1$.

- Suppose that $kk' < 0$. By symmetry, suppose that $i < i'$ and $j' < j$, as shown in Figure 3. We can remark that both repeats are completely determined as soon as we have fixed the first k letters $a_1, \ldots, a_k$ of the repeat at (i, j) and the first k' letters $b_1, \ldots, b_{k'}$ of the repeat at (i', j'). Moreover, a product of $2t + k + k' - 1$ transition probabilities will appear when writing down (3.3). If we bound each of them, except $k + k'$ particular ones, by ρ, we will obtain (see Remark 3.1) $\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}$.

- Now suppose that $kk' > 0$. The result will also be $\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}$. But to show this result, we need to distinguish between subcases depending on the value of $1 \leq |k' - k| \leq t - 1$. We will only detail the two representative subcases, $|k' - k| = 1$ and $|k' - k| = t - 2$ (the other subcases are left to the reader). For symmetrical reasons, we can suppose that $k' > k$.

  When $k' = k + 1$ (see Figure 4), the last $t - k$ letters of the repeat at (i, j) and the first $t - k$ letters of the repeat at (i', j') are all equal to $a := a_{k+1}$. Therefore, the repeated t-letter words are $a_1 a_2 \cdots a_k a \cdots a$ and $a \cdots a a_{k+2} \cdots a_{2k+1}$. To calculate $\mathrm{E}(R_\alpha R_\beta)$, we will then have to sum over all possible $(2k + 1)$-tuples $a_1, \ldots, a_{2k+1}$. Since (3.3) contains $2(t + k)$ transition probabilities, we obtain $\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}$.

  When $k' - k = t - 2$, for instance, $k = 1$ and $k' = t - 1$ (see Figure 5), the second letter, say a_2, of the repeat at (i, j) is equal to its last letter but also to the first and penultimate letter of the repeat at (i', j'). Therefore, the repeated t-letter words are $a_1 a_2 a_3 \cdots a_{t-1} a_2$ and $a_2 a_3 \cdots a_{t-1} a_2 a_t$. To calculate $\mathrm{E}(R_\alpha R_\beta)$, we will then have to sum over all possible t-tuples $a_1, \ldots, a_t$. Since (3.3) contains $3t - 1$ transition probabilities, we obtain $\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}$. If $k = 0$ and $k' = t - 2$, it is even simpler.

Now consider the situation where there are three overlapping occurrences. Without loss of generality, suppose that the three overlapping occurrences are the ones starting at i, i', and j

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

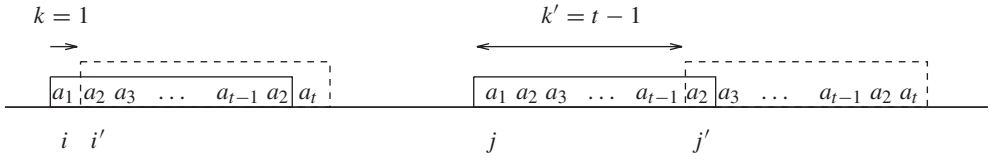

FIGURE 5: Two repeats at (i, j) and (i', j') with two pairs of overlapping occurrences for the special case in which $i' = i + 1$ $(k = 1)$ and $j' = j + t - 1$ $(k' = t - 1)$.

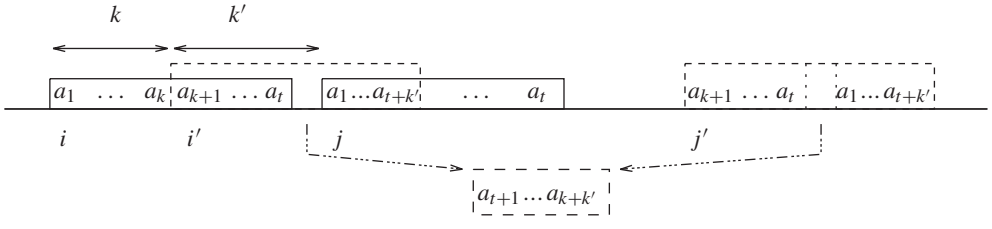

FIGURE 6: Two repeats at (i, j) and (i', j') with three overlapping occurrences.

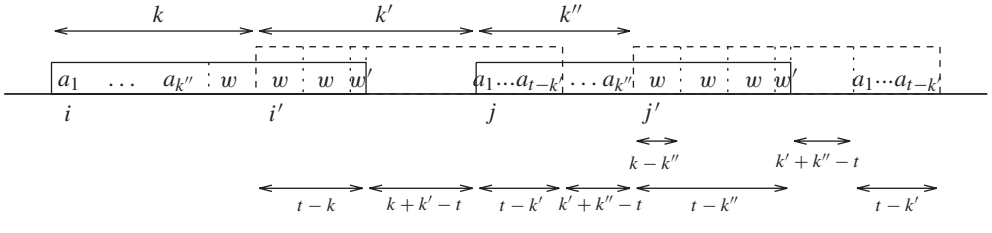

FIGURE 7: Two repeats at (i, j) and (i', j') with three overlaps for the special case in which $i < i' < j < j'$ and $k = i' - i \geq k'' = j' - j$.

(see Figure 6); we necessarily have $i < i' < j$. Set $k = i' - i$ and $k' = j - i'$, the lag between the overlapping occurrences. Then, if the repeated t-letter word at (i, j) is $a_1 \cdots a_t$, the t-letter word repeated at (i', j') starts with the last $(t - k)$ letters $a_{k+1} \cdots a_t$ and ends with the first $(t - k')$ letters $a_1 \cdots a_{t-k'}$. Note that $k' > t - k$ because occurrences starting at i and j have to be separated by at least one letter. Therefore, we need to complete the $(k + k' - t)$ central letters of the repeat at (i', j'), say by $a_{t+1}, \ldots, a_{k+k'}$. Finally, when calculating $\mathrm{E}(R_\alpha R_\beta)$, we sum over $(k + k')$ letters and we have a product of $2t + k + k' - 1$ transition probabilities. It follows that $\mathrm{E}(Y_\alpha Y_\beta) \leq \rho^{2t}$.

Case 4: $d = 3$. Here the four occurrences at i, j, i', and j' overlap together with one of the two following conditions: either $i < i' < j < j'$ or $i' < i < j' < j$. By symmetry, suppose that $i < i' < j < j'$ (see Figure 7), and set $k = i' - i$, $k' = j - i'$, and $k'' = j' - j$. We should distinguish between the two cases $k \geq k''$ and $k \leq k''$, but the technique is similar; thus, we consider the case in which $k \geq k''$.

The crucial point is that the t letters of the repeat at (i, j) are not free (it is the same for the other repeat). Indeed, consider the $(k - k'')$-letter word w composed of the first $(k - k'')$ letter

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

of the repeat starting at (i', j'). From Figure 7 we can see that the repeat at (i, j) is necessarily of the form $a_1 \cdots a_{k''} w \cdots w w'$, where w' is a prefix of w'. Moreover, the last $(t-k')$ letters of the repeat at (i', j') are equal to the first $(t-k')$ letters $a_1 \cdots a_{t-k'}$ of the repeat at (i, j). Finally, it remains to fix the $(k'+k''-t)$ central letters of the repeat at (i', j') to fix. This leads to a sum over $k+k'+k''-t$ letters when developing $\mathrm{E}(R_\alpha R_\beta)$, with a product of $(t+k+k'+k''-1)$ transition probabilities. Finally, we obtain $\mathrm{E}(Y_\alpha Y_\beta) \le \rho^{2t}$.

This completes the proof of Proposition 3.1.

3.5. A bound for b_3

Recall that the quantity b_3 is given by

$$b_3 = \sum_{\alpha \in I} \mathrm{E}\,|\,\mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid \sigma(Y_\beta : \beta \notin B_\alpha))|.$$

Set $\alpha = (i, j)$. Note that if $\beta = (i', j') \notin B_\alpha$ then i' and j' belong to $A_1 \cup A_2 \cup A_3$ with $A_1 = \{1, \ldots, i-2t\}$, $A_2 = \{i+2t, \ldots, j-2t\}$, and $A_3 = \{j+2t, \ldots, n-t+1\}$. We will distinguish between two cases: either A_2 is empty or not.

Case (i): $j < i + 4t$. In this case A_2 is empty and we have

$$\sigma(Y_\beta : \beta \notin B_\alpha) \subset \sigma(X_1, \ldots, X_{i-t-1}, X_{j+2t-1}, \ldots, X_n);$$

cf. Figure 8. Therefore, the Markov property leads to

$$\begin{aligned}
&\mathrm{E}\,|\,\mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid \sigma(Y_\beta : \beta \notin B_\alpha))| \\
&\quad \le \sum_{(x,y)\in\mathcal{A}^2} |\mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid X_{i-t-1} = x,\ X_{j+2t-1} = y)| \qquad (3.4)\\
&\qquad\qquad \times \mathrm{P}(X_{i-t-1} = x,\ X_{j+2t-1} = y) \\
&\quad \le \sum_{(x,y)\in\mathcal{A}^2} \Bigg| \sum_{b\in\mathcal{A}} \sum_{c\ne b} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}^t} \mathrm{P}(X_{i-t-1} = x,\ X_{i-1} = b,\ X_i = a_1, \ldots, \\
&\qquad\qquad\qquad X_{i+t-1} = a_t,\ X_{j-1} = c,\ X_j = a_1, \ldots, \\
&\qquad\qquad\qquad X_{j+t-1} = a_t,\ X_{j+2t-1} = y) \\
&\qquad\qquad\qquad - \mathrm{E}(Y_\alpha)\,\mathrm{P}(X_{i-t-1} = x,\ X_{j+2t-1} = y)| \\
&\quad \le \sum_{(x,y)\in\mathcal{A}^2} \sum_{b\in\mathcal{A}} \sum_{c\ne b} \sum_{(a_1,\ldots,a_t)\in\mathcal{A}^t} \mu(x)\pi(b,a_1)\pi(c,a_1)\pi^{(j-i-t)}(a_t,c) \\
&\qquad\qquad\qquad \times (\pi(a_1,a_2)\cdots\pi(a_{t-1},a_t))^2 \\
&\qquad\qquad\qquad \times |\pi^{(t)}(x,b)\pi^{(t)}(a_t,y) - \mu(b)\pi^{(j-i+3t)}(x,y)|. \qquad (3.5)
\end{aligned}$$

To evaluate the absolute value of the right-hand term in the last inequality of (3.5), we will use a decomposition of the transition matrix based on its diagonalization. Let $(\alpha_t)_{t=1,\ldots,4}$ be the eigenvalues of $\mathbf{\Pi}$ such that $|\alpha_1| \ge |\alpha_2| \ge |\alpha_3| \ge |\alpha_4|$. Because of the Perron–Frobenius theorem, we have $\alpha_1 = 1$ and $|\alpha_2| < 1$. Let $\boldsymbol{D} = \mathrm{diag}(1, \alpha_2, \alpha_3, \alpha_4)$ and let $\boldsymbol{P}$ be the eigenvector matrix such that $\mathbf{\Pi} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$. For all $k \in \{1, \ldots, 4\}$, $\boldsymbol{I}_k$ denotes the 4×4 matrix such that all its entries are equal to 0 except $I_k(k,k) = 1$ and we define $\boldsymbol{Q}_k = \boldsymbol{P}\boldsymbol{I}_k\boldsymbol{P}^{-1}$. Since $(1,1,1,1)^\top$ is a right eigenvector of $\mathbf{\Pi}$ for the eigenvalue 1, we have $Q_1(a,b) = \mu(b)$

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

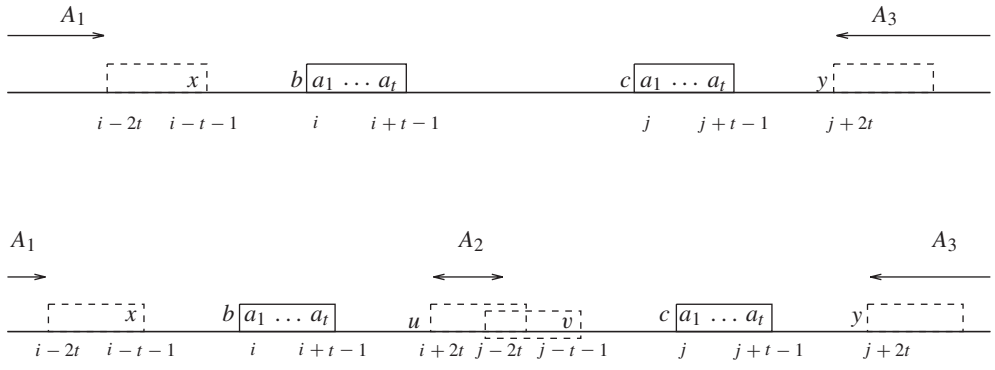

FIGURE 8: Repeat at position $\alpha = (i, j)$ and possible values for $\beta \notin B_\alpha$ when $A_2 = \{i+2t, \ldots, j-2t\}$ is empty (*top*) or not (*bottom*).

for all $a, b \in \mathcal{A}$. Moreover, we have $\boldsymbol{\Pi}^\ell = \boldsymbol{P}\boldsymbol{D}^\ell\boldsymbol{P}^{-1}$ and $\boldsymbol{D}^\ell = \sum_{k=1}^4 \alpha_k^\ell \boldsymbol{I}_k$, leading to $\boldsymbol{\Pi}^\ell = \sum_{k=1}^4 \alpha_k^\ell \boldsymbol{Q}_k$. Thus, we obtain

$$\begin{aligned}
&|\pi^{(t)}(x,b)\pi^{(t)}(a_t,y) - \mu(b)\pi^{(j-i+3t)}(x,y)| \\
&\quad = \Bigg| \sum_{\substack{k,k'=1 \\ (k,k')\neq(1,1)}}^4 \alpha_k^t \alpha_{k'}^t Q_k(x,b) Q_{k'}(a_t,y) - \mu(b) \sum_{k=2}^4 \alpha_k^{(j-i+3t)} Q_k(x,y) \Bigg| \\
&\quad \le \sum_{\substack{k,k'=1 \\ (k,k')\neq(1,1)}}^4 |\alpha_k|^t |\alpha_{k'}|^t Q_k(x,b) Q_{k'}(a_t,y) + \mu(b) \sum_{k=2}^4 |\alpha_k|^{(j-i+3t)} Q_k(x,y) \\
&\quad \le |\alpha_2|^t C(b, a_t, t, x, y),
\end{aligned}$$

where

$$C(b,a_t,t,x,y) = \sum_{(k,k')\neq(1,1)}^4 \frac{|\alpha_k|^t|\alpha_{k'}|^t}{|\alpha_2|^t} Q_k(x,b) Q_{k'}(a_t,y) + \mu(b) \sum_{k=2}^4 \frac{|\alpha_k|^{3t}}{|\alpha_2|^t} Q_k(x,y);$$

note that $C(b, a_t, t, x, y) = \mathrm{O}(1)$ as $t \to \infty$.

Let us return to (3.5). If we abbreviate α_2 by α, we obtain

$$\begin{aligned}
&\mathrm{E}\,|\,\mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid \sigma(Y_\beta : \beta \notin B_\alpha))| \\
&\quad \le \rho|\alpha|^t \sum_b \sum_{c\neq b} \sum_{(a_1,\ldots,a_t)} \pi(b,a_1)\pi(c,a_1)(\pi(a_1,a_2)\cdots\pi(a_{t-1},a_t))^2 \\
&\qquad\qquad \times \sum_{(x,y)} \mu(x) C(b,a_t,t,x,y) \\
&\quad \le \rho^t |\alpha|^t C_1(t),
\end{aligned} \tag{3.6}$$

where $C_1(t)$ is bounded and given by

$$C_1(t) = \sum_b \sum_{c\neq b} \sum_{(a_1,\ldots,a_t)} \pi(b,a_1)\pi(c,a_1)\pi(a_1,a_2)\cdots\pi(a_{t-1},a_t) \sum_{(x,y)} \mu(x) C(b,a_t,t,x,y).$$

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

Case (ii): $j \geq i + 4t$. In this case A_2 is not empty and we have

$$\sigma(Y_\beta : \beta \notin B_\alpha) \subset \sigma(X_1, \ldots, X_{i-t-1}, X_{i+2t-1}, \ldots, X_{j-t-1}, X_{j+2t-1}, \ldots, X_n);$$

cf. Figure 8. From the Markov property, it follows that

$$\mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid \sigma(Y_\beta : \beta \notin B_\alpha)) = \mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid X_{i-t-1}, X_{i+2t-1}, X_{j-t-1}, X_{j+2t-1}).$$

Now we will sum over the values x, u, v, and y that the variables X_{i-t-1}, X_{i+2t-1}, X_{j-t-1}, X_{j+2t-1} can take, and we follow different steps from those used for the first case. We finally obtain

$$\begin{aligned}
&\mathrm{E}\,|\,\mathrm{E}(Y_\alpha - \mathrm{E}(Y_\alpha) \mid \sigma(Y_\beta : \beta \notin B_\alpha))| \\
&\quad \leq \sum_{(x,u,v,y)} \sum_{(b,a_1,\ldots,a_t)} \sum_{c \neq b} \mu(x)\pi(b,a_1)\pi(c,a_1)\pi^{(j-i-3t)}(u,v) \\
&\qquad\qquad \times (\pi(a_1,a_2)\cdots\pi(a_{t-1},a_t))^2 \\
&\qquad\qquad \times |\pi^{(t)}(x,b)\pi^{(t)}(a_t,u)\pi^{(t)}(v,c)\pi^{(t)}(a_t,y) \\
&\qquad\qquad\quad - \mu(b)\pi^{(j-i-t)}(a_t,c)\pi^{(3t)}(x,u)\pi^{(3t)}(v,y)| \\
&\quad \leq \rho^t |\alpha|^t C_2(t), \qquad (3.7)
\end{aligned}$$

where

$$\begin{aligned}
C_2(t) = \sum_b \sum_{c \neq b} \sum_{(a_1,\ldots,a_t)} &\pi(b,a_1)\pi(c,a_1)\pi(a_1,a_2)\cdots\pi(a_{t-1},a_t) \\
&\times \sum_{(x,u,v,y)} \mu(x) C'(b,a_t,t,x,u,v,y)
\end{aligned}$$

and

$$\begin{aligned}
&C'(b,a_t,t,x,u,v,y) \\
&\quad = \sum_{\substack{k_1,k_2,k_3,k_4=1 \\ (k_1,k_2,k_3,k_4)\neq(1,1,1,1)}}^{4} \frac{|\alpha_{k_1}\alpha_{k_2}\alpha_{k_3}\alpha_{k_4}|^t}{|\alpha|^t} Q_{k_1}(x,b) Q_{k_2}(a_t,u) Q_{k_3}(v,c,) Q_{k_4}(a_t,y) \\
&\qquad + \mu(b) \sum_{\substack{k_1,k_2,k_3=1 \\ (k_1,k_2,k_3)\neq(1,1,1)}}^{4} \frac{|\alpha_{k_1}\alpha_{k_2}\alpha_{k_3}|^{3t}}{|\alpha|^t} Q_{k_1}(a_t,c) Q_{k_2}(x,u) Q_{k_3}(v,y).
\end{aligned}$$

Note that $C_2(t)$ is bounded as $t \to \infty$.

Finally, it follows from (3.6), (3.7), and $|I| = \mathrm{O}(n^2)$ that

$$b_3 = \mathrm{O}(n^2 \rho^t |\alpha|^t); \qquad (3.8)$$

b_3 will then converge to 0 when $n^2\rho^t = \mathrm{O}(1)$ and $t \to \infty$.

3.6. Approximation theorems

Thanks to the Chen–Stein theorem and (3.1), (3.2), and (3.8) for the calculation of b_1, b_2, and b_3, respectively, we have proved the following theorem.

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

Theorem 3.2. *There exist two explicit constants K_1 and K_2 which only depend on t such that the total variation distance between the distribution of the number N_t of non-self-overlapping leftmost repeats of length t in a Markov chain of length n and a Poisson distribution with parameter $\lambda_t = \mathrm{E}(N_t)$ given in (2.5) satisfies the following relations.*

- *When $t = \mathrm{o}(n)$, we have*

$$d_{\mathrm{TV}}(\mathcal{L}(N_t), \mathcal{P}_o(\lambda)) \le K_1 \frac{t}{n}(n^2\rho^t)^2 + K_2(n^2\rho^t)|\alpha|^{t+1},$$

  where $|\alpha| < 1$ is the second largest eigenvalue of the transition matrix and ρ is the largest transition probability.

- *Moreover, if $n^2\rho^t = \mathrm{O}(1)$ then*

$$d_{\mathrm{TV}}(\mathcal{L}(N_t), \mathcal{P}_o(\lambda)) = \mathrm{o}(1).$$

A straightforward corollary follows easily to approximate the number $N_t^{(k)}$ of leftmost repeats of length t separated by at least k letters in the sequence. This corollary will be used in the next section for the generalization to Markov chain models of order $m > 1$. As a particular case, $N_t^{(1)} = N_t$. The main differences between the proofs are as follows.

- The index set $I^{(k)}$ of the possible positions $\alpha = (i, j)$ of leftmost repeats of length t separated by at least k letters is

$$I^{(k)} = \{\alpha = (i, j) \mid 1 \le i < i + t + k - 1 < j \le n - t + 1\};$$

  therefore, $N_t^{(k)} = \sum_{\alpha \in I^{(k)}} Y_\alpha$.

- The expected number $\lambda_t^{(k)}$ of leftmost repeats of length t separated by at least k letters is given by

$$\lambda_t^{(k)} = \sum_{a_1,\dots,a_t,b,\, c \ne b} \mu(b)\pi(b, a_1)\pi(c, a_1)(\pi(a_1, a_2)\cdots\pi(a_{t-1}, a_t))^2 \times \sum_{\ell=0}^{n-2t-k} (n - \ell - 2t)\pi^{(\ell+k)}(a_t, c). \tag{3.9}$$

- The neighborhood has to be changed into $B_\alpha^{(k)} = B_\alpha \cap I^{(k)}$ for $\alpha \in I^{(k)}$.

Corollary 3.1. *The total variation distance between the distribution of the number $N_t^{(k)}$ of leftmost repeats of length t separated by at least k letters in a Markov chain of length n and a Poisson distribution with parameter $\lambda_t^{(k)}$ given in (3.9) satisfies*

$$d_{\mathrm{TV}}(\mathcal{L}(N_t^{(k)}), \mathcal{P}_o(\lambda_t^{(k)})) = \mathrm{o}(1) \quad \text{if } n^2\rho^t = \mathrm{O}(1) \text{ and } t = \mathrm{o}(n);$$

ρ is the largest transition probability of the Markov model.

4. Generalization to m-order Markov chain models

To treat the general case of m-order Markov chain models ($m \ge 1$), it will be enough to use the property that an m-order Markov chain on the alphabet $\mathcal{A}$ is a one-order Markov chain on the alphabet $\mathcal{A}^m$.

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

Let us consider a random sequence $S = X_1, \dots, X_n$ generated by an m-order Markov chain with transition matrix $\mathbf{\Pi} = (\pi(a_1 \cdots a_m, b))$ and stationary distribution on $\mathcal{A}^m$ denoted by μ. The sequence $S^* = X_1^*, \dots, X_{n-m+1}^*$ such that $X_i^* = X_i \cdots X_{i+m-1} \in \mathcal{A}^m$ is then a one-order Markov chain on the alphabet $\mathcal{A}^m$ with transition probabilities $\pi^*(a_1 \cdots a_m, b_1 \cdots b_m)$ such that

$$\pi^*(a_1 \cdots a_m, b_1 \cdots b_m) = \begin{cases} \pi(a_1 \cdots a_m, b_m) & \text{if } a_2 \cdots a_m = b_1 \cdots b_{m-1}, \\ 0 & \text{otherwise.} \end{cases}$$

We assume that $\pi(a_1 \cdots a_m, b) > 0$ for all $a_1, \dots, a_m, b \in \mathcal{A}$.

Now the number $N_t(S)$ of non-self-overlapping leftmost repeats of length t in the sequence S is equal to the number $N_{t-m+1}^{(m)}(S^*)$ of leftmost repeats of length $t-m+1$ separated by at least m letters in the sequence S^*. Indeed, we have

$$\begin{aligned} &\mathbf{1}\{X_{i-1} \neq X_{j-1},\ X_i \cdots X_{i+t-1} = X_j \cdots X_{j+t-1}\} \\ &\quad = \mathbf{1}\{X_{i-1}^* \neq X_{j-1}^*,\ X_i^* \cdots X_{i+t-m}^* = X_j^* \cdots X_{j+t-m}^*\}, \end{aligned}$$

leading to

$$N_t(S) = N_{t-m+1}^{(m)}(S^*).$$

From (3.9), the expected number of non-self-overlapping leftmost repeats of length t in S, denoted by γ_t, is given by

$$\begin{aligned} \gamma_t &= \sum_{A_1,\dots,A_{t-m+1},B,C\neq B} \mu(B)\pi^*(B, A_1)\pi^*(C, A_1)(\pi^*(A_1, A_2)\cdots\pi^*(A_{t-m}, A_{t-m+1}))^2 \\ &\qquad \times \sum_{\ell=0}^{n-2(t-m+1)-m} (n-\ell-2(t-m+1))(\pi^*)^{(\ell+m)}(A_{t-m+1}, C) \\ &= \sum_{a_1,\dots,a_t,b,c\neq b} \mu(ba_1\cdots a_{m-1})\pi(ba_1\cdots a_{m-1}, a_m)\pi(ca_1\cdots a_{m-1}, a_m) \\ &\qquad \times (\pi(a_1\cdots a_m, a_{m+1})\cdots\pi(a_{t-m}\cdots a_{t-1}, a_t))^2 \\ &\qquad \times \sum_{\ell=0}^{n-2t+m-2} (n-\ell-2t+2m-2)(\pi^*)^{(\ell+m)}(a_{t-m+1}\cdots a_t, ca_1\cdots a_{m-1}), \end{aligned} \tag{4.1}$$

where $(\pi^*)^{(\ell+m)}(A, B)$, $A, B \in \mathcal{A}^m$ is the $(\ell+m)$-step transition probability between A and B in S^*. Applying Theorem 3.1 to $N_{t-m+1}^{(m)}(S^*)$ gives the following corollary.

Corollary 4.1. *Let S be a random sequence generated by an m-order Markov chain whose largest transition probability is denoted by ρ. We assume that $\rho < 1$. Under the condition $n^2\rho^{t-m+1} = \mathrm{O}(1)$, the number $N_t(S)$ of non-self-overlapping leftmost repeats of length t in S can be approximated by a Poisson variable whose expectation γ_t is given by (4.1):*

$$d_{\mathrm{TV}}(\mathcal{L}(N_t(S)), \mathcal{P}_o(\gamma_t)) = \mathrm{o}(1).$$

5. Conclusion

We have used the Chen–Stein method to bound the total variation distance between the distribution of the number N_t of leftmost repeats of length t separated by at least one letter in a

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press

finite Markov chain S, and the Poisson distribution with parameter $\mathrm{E}(N_t)$. We showed that this bound converges to 0 when the length n of the sequence tends to ∞, $t = \mathrm{o}(n)$, $n^2\rho^t = \mathrm{O}(1)$, and the largest transition probability ρ of the Markov model is strictly less than 1; the third condition means that the Poisson approximation is valid for long enough repeats.

The approximation theorems also hold for leftmost disjoint repeats, i.e. when allowing the position j of the second occurrence to be equal to $i + t$. In this case we just have to be careful about the fact that the first letter a_1 of the repeated pattern has to be different from the letter b that occurs at position $i - 1$. The expression of the repeat probability $p_{(i,i+t)}$ will then slightly differ from the one given by Lemma 2.1 (there will be no more letter c) and an additional term corresponding to $j - i = t$ will appear in the expected count.

Allowing overlapping repeats, i.e. allowing $j \in \{i + 1, \ldots, i + t - 1\}$, is technically more complicated because constraints on the letters of the repeated pattern $a_1a_2 \cdots a_t$ will appear. Since overlapping repeats are not the most interesting from a biological point of view, we have not investigated the general case further.

In practice, the transition probabilities $\pi(a, b)$ are estimated from the observed sequence by their maximum likelihood estimates $\hat{\pi}(a, b) =$ (observed number of ab)/(observed number of a). Since we derived the explicit formula for the mean count λ_t, we can calculate its plug-in estimator $\hat{\lambda}_t$ and show that $|\lambda_t - \hat{\lambda}_t|$ converges to 0 in our asymptotic framework. Since the total variation distance between two Poisson distributions with respective parameters λ_t and $\hat{\lambda}_t$ is bounded by $|\lambda_t - \hat{\lambda}_t|$, we can use the Poisson distribution with parameter $\hat{\lambda}_t$ to approximate the distribution of N_t and then to approximate the p-value $\mathrm{P}(N_t \geq N_t^{\mathrm{obs}})$.

Acknowledgement

N. Touyar would like to thank Joël Alexandre for his strong support.

References

[1] Arratia, R., Martin, D., Reinert, G. and Waterman, M. (1996). Poisson process approximation for sequence repeats and sequencing by hybridization. *J. Comput. Biol.* **3,** 425–463.
[2] Barbour, A., Holst, L. and Janson, S. (1992). *Poisson Approximation*. Clarendon Press, Oxford.
[3] Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580.
[4] Delcher, A. L. *et al.* (1999). Alignment of whole genomes. *Nucleic Acids Res.* **27,** 2369–2376.
[5] Kolpakov, R., Bana, G. and Kucherov, G. (2003). Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* **31,** 3672–3678.
[6] Kurtz, S. *et al.* (2001). Reputer: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* **29,** 4633–4642.
[7] Lefèbvre, A., Lecroq, T., Dauchel, H. and Alexandre, J. (2003). FORRepeats: detects repeats on entire chromosomes and between genomes. *Bioinformatics* **19,** 319–326.
[8] Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview. *J. Comput. Biol.* **7,** 1–46.
[9] Robin, S., Rodolphe, F. and Schbath, S. (2005). *DNA, Words and Models*. Cambridge University Press.
[10] Taylor, J.S. and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Ann. Rev. Genet.* **38,** 615–643.

https://doi.org/10.1239/jap/1214950359 Published online by Cambridge University Press