Discrete Mathematics and Theoretical Computer Science **AC**, 2003, 243–258

q-gram analysis and urn models

Pierre Nicodème

LIX - École polytechnique, 91128 Palaiseau cedex, France
nicodeme@lix.polytechnique.fr

Words of fixed size q are commonly referred to as q-grams. We consider the problem of q-gram filtration, a method commonly used to speed up sequence comparison. We are interested in the statistics of the number of q-grams common to two random texts (where multiplicities are not counted) in the non uniform Bernoulli model. In the exact and dependent model, when omitting border effects, a q-gram in a random sequence depends on the $q-1$ preceding q-grams. In an approximate and independent model, we draw randomly a q-gram at each position, independently of the others positions. Using ball and urn models, we analyze the independent model. Numerical simulations show that this model is an excellent first order approximation to the dependent model. We provide an algorithm to compute the moments.

Keywords: Sequence comparison, Bernoulli model, urn models

1 Introduction

We consider in this article random sequences and some statistics of occurrences of words of a fixed size q that are referred to as q-grams. In the context of pattern matching, when a word of size m matches a text with at most k errors, Jokinen and Ukkonen [JU91] give the lower bound $m+1-(k+1)q$ for the number of common q-grams to the pattern and the text. The QUASAR package (q-gram Alignment based on Suffix Arrays) [BCF$^+$99] uses this result in all against all comparisons of (a large number of) DNA sequences to filter out pairs with low similarity. In this process, if there are N sequences in the database, a sequence is used on the average $N/2$ times as query while the other sequences are seen as targets. For one pair of sequences, the count of common q-grams is done by adding one for each q-gram (with possible repetitions) of the query that is present in the target. Szpankowski and Sutinen [SS98] analyze q-gram filtration in a similar context, but they count with repetitions the q-grams of the target that are present in the query. They study the correlations of the q-grams of the querying word, extending previous works of Guibas and Odlyzko [GO78] and of Régnier and Szpankowski [RS98].

Flajolet *et al.* [FKT88] (in the Bernoulli uniform model) and Szpankowski [Szp93] (in the non uniform model) give an important asymptotic result: in a random text of size n, when $n \to \infty$, every word of length $l < \log n/h$ almost surely occurs; here, h is the Rényi entropy [†] of the alphabet considered. This gives a lower bound on the size of q-grams to consider when comparing two sequences.

The related problem of missing words in a random text has been studied in the Bernoulli uniform case by Rahmann and Rivals [RR00, RR03]. They give the first two moments in the case of 2-grams.

[†] If ω_{min} is the minimum of the probability of the letters of the alphabet, $h = \log 1/\omega_{min}$.

We consider here a slightly different model of QUASAR for common q-grams to two sequences, where repetitions are not counted. We also consider the number of q-grams occurring at least twice in a random sequence. The underlying model for the sequences is the Bernoulli non-uniform model. In this model, the letters of the sequences are drawn independently along the distribution of the letters of the alphabet. Our approach compares the statistics of the q-gram sequence problem, where a q-gram at position i (with $q \leq i \leq n-q+1$) depends on the $q-1$ preceding positions of the sequence (dependent model) and the statistics of the q-gram independent model, where a q-gram is drawn at each position, independently of the preceding positions. We use for this last problem an urn and ball model that we describe in Section 2. We give analytic results for the expectation and variance of repeated (resp. common) q-grams for one (resp. two) sequence(s) in the independent model, and compare these to results obtained by simulations, in both the dependent and independent models.

This article is organized as follows. In Section 2 we provide definitions and give equivalent models to the problem of q-grams repetitions or to this of sharing common q-grams, in the dependent and in the independent settings. Section 3 analyzes the independent model. We give explicit formula for the expectation and the variance of the number of repeated or common q-grams. We conclude Section 3 by applying the Chen-Stein method [Che75, Ste70] to prove Poisson convergence in the independent case. We proceed to experimental comparisons of the dependent and independent models in Section 4. Complexity of the computations and a fast and flexible algorithm to perform these are given in Section 5. We describe additional results about the dependent model in Section 6.

2 Definitions and models

We consider an alphabet $\Sigma = \{l_1, l_2, \ldots, l_s\}$ of size s, where letter l_i has probability ω_i, words of size q over this alphabet, and sequences of length $n+q-1$, corresponding to n q-grams; (the more general case of sequences of different lengths and different letter probability will be considered in Section 3.3). In the independent model, we draw randomly n q-grams for the sequence(s).

We consider the counting as follows: let two sequences S_1 and S_2 be

$$S_1 = aaaaaaabababaaaabbbbaaa \quad \text{and} \quad S_2 = bbbbababababaaaaaaaaa.$$

The *repeated* 4-grams of S_1 are *aaaa*, *abab*, *baba* and *baaa* and the number of repeated 4-grams of S_1 is 4. The *common* 4-grams to S_1 and S_2 are *aaaa*, *abab*, *baba*, *bbbb* and *baaa*, and the number ‡ of common 4-grams is 5.

There are three different but equivalent models to consider the q-gram problem, graph models, tree models and urn models. This models are compared in Table 1. We describe them further.

Graph models. The *de Bruijn graph* [dB46] $B(s,q)$ is defined as a directed graph $G(V,E)$ where the set of vertices V is labeled $V = \{0, 1, 2, \ldots, s^q-1\}$ and there is an edge from vertex v_i to vertex v_j if

$$v_j = s \times v_i \ (\mathrm{mod}\ s^q) + x, \quad x = 0|1|2|\ldots|(s-1).$$

A more intuitive description is as follows: we associate bijectively to each q-gram the vertex labeled with the integer defined by the q-gram in base s. For a q-gram $g = g_1 g_2 \cdots g_q$, shift the letters one position left

‡ In QUASAR, if S_1 is the query and S_2 is the target, the contribution of the q-gram *baaa* to the number of common q-grams is 2 (this q-gram is present twice in S_1).

<table>
<tr><th colspan="2">Repeated q-grams in a sequence</th></tr>
<tr><th>Dependent</th><th>Independent</th></tr>
<tr><td>reading n q-grams in a random sequence of length $n+q-1$</td><td>drawing n independent q-grams</td></tr>
<tr><td colspan="2">number of repeated q-grams?</td></tr>
<tr><td colspan="2">random walk of length n in</td></tr>
<tr><td>de Bruijn graph $B(s,q)$</td><td>complete graph $K(s^q)$</td></tr>
<tr><td colspan="2">number of self-intersections of the walk?</td></tr>
<tr><td>suffix-tree built over a random sequence of length $n+q-1$</td><td>trie with n keys</td></tr>
<tr><td colspan="2">number of internal nodes at depth q?</td></tr>
<tr><td></td><td>throwing independently n balls in s^q urns; number of urns with collisions?</td></tr>
</table>

<table>
<tr><th colspan="2">Common q-grams to two sequences
(with same sizes and letter distributions)</th></tr>
<tr><th>Dependent</th><th>Independent</th></tr>
<tr><td>reading twice n q-grams in two random sequences of size $n+q-1$</td><td>drawing twice n independent q-grams</td></tr>
<tr><td colspan="2">number of common q-grams?</td></tr>
<tr><td colspan="2">2 random walks of length n in</td></tr>
<tr><td>de Bruijn graph $B(s,q)$</td><td>complete graph $K(s^q)$</td></tr>
<tr><td colspan="2">number of intersections of the 2 walks?</td></tr>
<tr><td>2 suffix-trees of different colors built over 2 sequences of length $n+q-1$</td><td>2 tries of different colors with n keys</td></tr>
<tr><td colspan="2">number of bicolor nodes at depth q when superposing the 2 trees ?</td></tr>
<tr><td></td><td>throwing independently n white balls and n black balls in s^q urns; number of urns with bicolor collisions?</td></tr>
</table>

Tab. 1: Equivalent models

and drop the leftmost letter. Putting at the (free) rightmost position any letter l_i of the alphabet gives a q-gram, which is one of the successors of g in the graph $B(s,q)$.

We then have

$$g_1 g_2 g_3 \cdots g_q \xrightarrow{l_i} g_2 g_3 \cdots g_q l_i,$$

and the corresponding edge is weighted by the probability ω_i of letter l_i.

We consider in the independent model the *complete graph* $K(s^q)$ where the vertices are labeled from 0 to s^q-1. As previously we bijectively associate to each q-gram a vertex of the graph. The weight of an edge from vertex v_i to edge v_j is the probability of the q-gram associated to v_j.

The statistics of the number of repeated q-grams in a random sequence of length $n+q-1$ is the same as the statistics of the number of self-intersections of a random walk of length n over the de Bruijn graph $B(s,q)$ (dependent model); the statistics of the number of repeated q-grams when drawing independently n

q-grams is the same as the statistics of self-intersections of a random walk over the complete graph $K(s^q)$. In both cases, the probability of starting the walk at vertex v is the probability of the q-gram associated to v. When considering the number of common q-grams to two sequences, we consider two random walks and the number of intersections of these walks.

Tree models. As an alternative model we can consider suffix-trees [Apo85, Wei73] in the dependent model and tries in the independent model.

The suffix-tree is built on a random sequence of length $n+q-1$ and the trie is built by insertion of n keys of infinite length. The number of internal nodes at depth q in the suffix-tree (resp. the trie) models the number of repeated q grams in one sequence in the dependent (resp. independent model). The trie structure is fundamental to many algorithms; for descriptions see [Knu73, AHU74, Sed88, GBY91]. Jacquet and Szpankowski [JS91a] do the asymptotic analysis of the average depth of a trie in the Markovian case. Clement *et al.* [CVF01] analyze the average depth, height and size of tries in the general case of dynamic sources. A very relevant result to our problem is the analysis (in the Bernoulli model) of the average depth in suffix trees (depth of a key chosen at random) by Jacquet and Szpankowski [JS94]. They prove (and verify experimentally [JS91b]) that the average depth of keys behaves similarly in tries and in suffix trees (the distributions are asymptotically identical). They also prove that, asymptotically, the *e*xpectation of the size of suffix-trees and of tries are the same; however, they give no result for the variance or the distribution of this statistics. See also Fayolle [Fay02] for a direct computation of this expectation.

Urn models. The independent model can be modeled by a ball and urn system. We throw n indistinguishable balls in $m = s^q$ urns which are labeled from 0 to $s^q - 1$. To each q-gram $W = w_0 w_1 \dots w_{q-1}$ is bijectively associated the urn labeled with the value of the integer W in base s. The probability p_j that a ball falls in urn j is the probability $\mathbf{P}(W_j)$ of the q-gram W_j associated to this urn. Urn models have been extensively studied. Classical references are Johnson and Kotz [JK77] or Kolchin *et al.* [KSC78]. See also Flajolet and Sedgewick [SF96] for a combinatorial introduction and Drmota *et al.* [DGG01] and Flajolet *et al.* [FGP03] for recent results.

In the one sequence problem, we are interested in the number of urns containing more than one ball (number of collisions). When considering two sequences B and C with respective lengths $b+q-1$ and $c+q-1$, we throw in the urns b black and c white balls; we now want to compute the statistics of number of urns containing simultaneously black and white balls (number of bicolor collisions); this counts the number of common q-grams.

3 Analysis of the independent model

In this section we give analytic results about the independent model. Our counting tools are multivariate generating functions from which we derive the first two moments of the considered statistics. We also use the Chen-Stein method to prove Poisson convergence. The generating functions could be computed by the methods presented in Kolchin *et al.* [KSC78]; however the Poissonization-Depoissonization method that we present in the next section makes the proofs elementary and elegant.

3.1 Basic Poissonization and Depoissonization

The problem with an urn system is that the urns are not independent of each other. The very powerful Poissonization method provides independence. See Szpankowski [Szp01] for an extensive introduction to Poissonization-Depoissonization.

Let us consider the problem of repeated q-grams in one sequence to illustrate the method. We assume now that the number of balls thrown in the system is no longer exactly n but that it follows a Poisson distribution $\mathcal{P}_z$ of parameter z. For a sequence of functions $\phi_n(u)$ we consider its Poisson transform $\Psi(z,u)$ which is

$$\Psi(z,u) = \sum_{n\geq 0} \phi_n(u) \frac{z^n}{n!} e^{-z}.$$

This implies the basic Depoissonization

$$\phi_n(u) = [z^n] n! e^z \Psi(z,u),$$

where $[z^n]F(z)$ denotes the n-th Taylor coefficient of $F(z)$.

For our concerns, $\phi_n(u)$ and $\Psi(z,u)$ respectively are the distribution generating function of the number of urns without collisions when (1) exactly n balls or (2) a random number of balls distributed as $\mathcal{P}_z$ are thrown in the system. Considering the urn i in the Poisson model, the probability g_{ik} that there are k balls in this urn is

$$g_{ik} = \sum_{n\geq k} e^{-z} \frac{z^n}{n!} \binom{n}{k} p_i^k (1-p_i)^{n-k} = e^{-z} \frac{(p_i z)^k}{k!} \sum_{n\geq k} \frac{\big((1-p_i)z\big)^{n-k}}{(n-k)!} = \frac{(p_i z)^k}{k!} e^{-p_i z}.$$

Therefore the number of balls in urn i follows a Poisson law of parameter $p_i z$ and the urns behave independently of each other.

We will mark by u the number of urns without collisions. The generating function of the urn i is

$$\psi_i(z,u) = e^{-p_i z}\big(e^{p_i z} - 1 - p_i z + u(1+p_i z)\big).$$

By independence of the urns, we obtain by product the generating function of the system:

$$\Psi(z,u) = \prod_{0\leq i\leq s^q-1} \psi_i(z,u) = e^{-z} \prod_{0\leq i\leq s^q-1} \big(e^{p_i z} + (u-1)(1+p_i z)\big).$$

Applying the basic Depoissonization method, we obtain the bivariate exponential generating function [§] $F(z,u)$ counting the number of urns without collisions in the exact model,

$$\phi_n(u) = n![z^n] \prod_{0\leq i\leq s^q-1} \big(e^{p_i z} + (u-1)(1+p_i z)\big) \Rightarrow F(z,u) = \sum_{n\geq 0} \frac{\phi_n(u) z^n}{n!} = \prod_{0\leq i\leq s^q-1} \Big(e^{p_i z} + (u-1)(1+p_i z)\Big) \tag{1}$$

3.2 Repeated q-grams (one sequence)

We derived in the preceding section (Equation 1) the generating function $F(z,u) = \sum_{n,k} f_{nk} u^k z^n / n!$ where f_{nk} is the probability that there are k urns without collisions when one throws n balls in the system. We compute in this section the asymptotic expectation μ_n and variance σ_n^2 of the number of urns without collisions. The expectation γ_n of the number of urns with collisions is $\gamma_n = m - \mu_n$ and the variance is again σ_n^2.

[§] Flajolet *et al.* [FGT92] obtain this generating function by language shuffling; it seems difficult to apply this elegant method to the case of common q-grams.

$\Sigma = \{0,1\}$ $s = 2$ $q = 10$ $m = s^q = 1024$ $n = 300$ $p_0 = \mathbf{P}(0)$

solid lines: theoretical μ_n and σ_n^2 for the independent model; dots: simulations

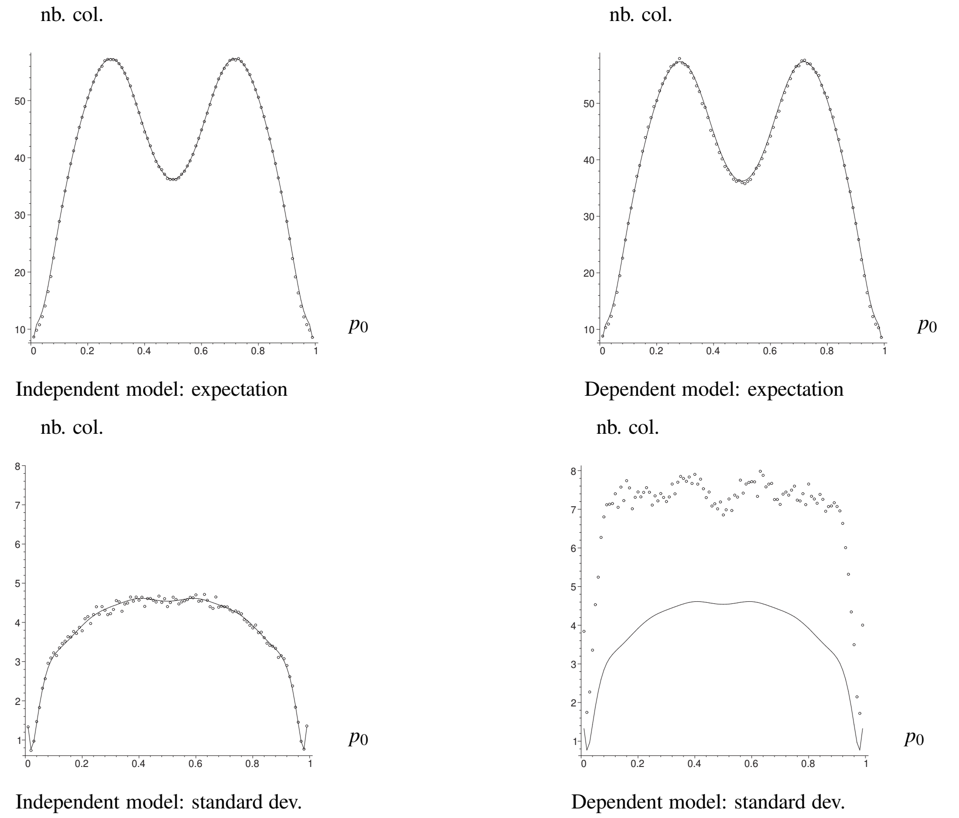

Fig. 1: Repeated q-grams in the independent and dependent models

We use the standard techniques of generating functions to obtain the expectation and the variance of the considered statistics.

We have for the expectation μ_n and for the second moment $m_n^{(2)}$

$$m(z) = \sum_{n\geq 0} \mu_n \frac{z^n}{n!} = \left.\frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \sum_i (1+p_i z) \prod_{j\neq i} e^{p_j z} = \sum_i (1+p_i z) e^{(1-p_i)z} \quad \text{and}$$

$$m^{(2)}(z) = \sum_{n\geq 0} m_n^{(2)} \frac{z^n}{n!} = \left.\frac{\partial}{\partial u} u \frac{\partial F(z,u)}{\partial u}\right|_{u=1} = \sum_i (1+p_i z) e^{(1-p_i)z} + \sum_i \sum_{j\neq i} (1+p_i z)(1+p_j z) e^{(1-(p_i+p_j))z}$$

We obtain μ_n and $m_n^{(2)}$ by extraction of the Taylor coefficient of order n from the corresponding generating functions:

$$\mu_n = n![z^n]m(z) \quad \text{and} \quad m_n^{(2)} = n![z^n]m^{(2)}(z).$$

This gives

$$\mu_n = \sum_i \left((1-p_i)^n + n p_i (1-p_i)^{n-1}\right) \quad \text{and}$$

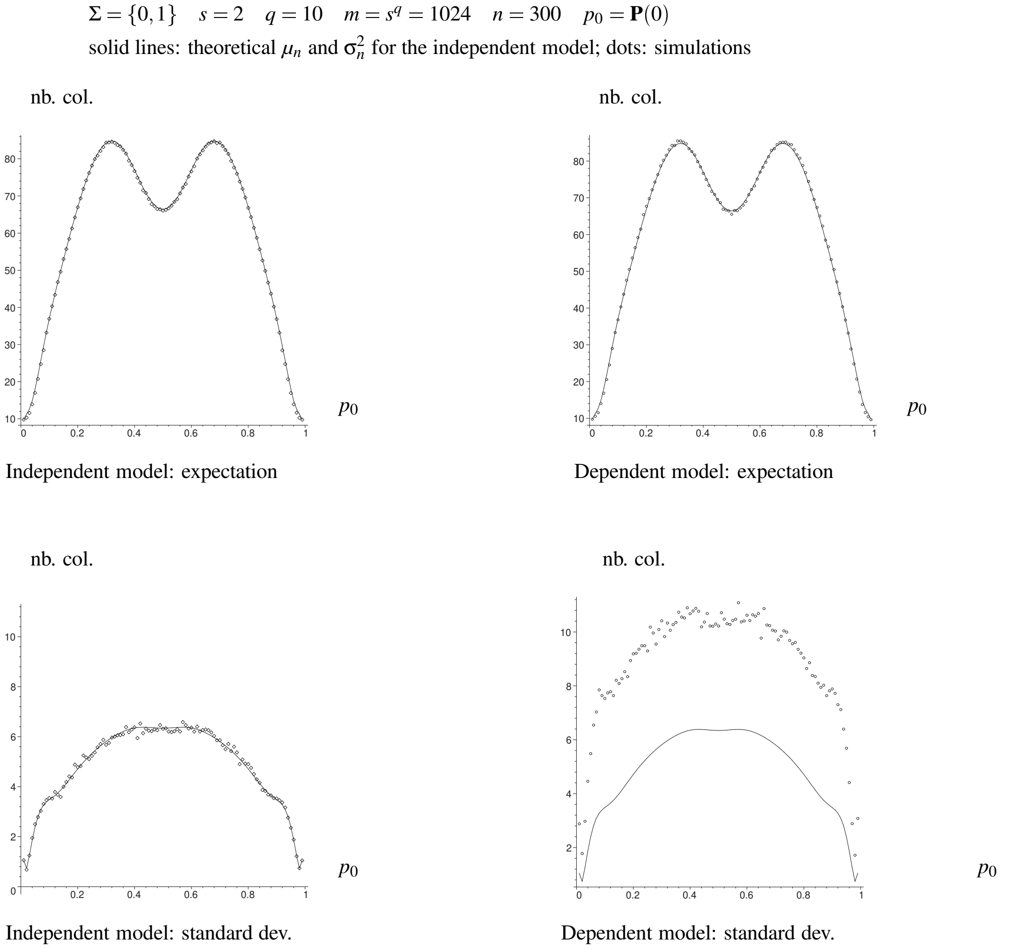

Fig. 2: Common q-grams to two sequences of same size in the independent and dependent models

$$m_n^{(2)} = \sum_i \left((1-p_i)^n + np_i(1-p_i)^{n-1} - (1-2p_i)^n - 2np_i(1-2p_i)^{n-1} - n(n-1)p_i^2(1-2p_i)^{n-2}\right)$$
$$+ \sum_i \sum_j \left((1-p_i-p_j)^n + n(p_i+p_j)(1-p_i-p_j)^{n-1} + n(n-1)p_ip_j(1-p_i-p_j)^{n-2}\right) \qquad (2)$$

We consider asymptotically that $n \to \infty$ in such a way that $n \times p_i \to \theta_i$ (with $0 < \theta_i < \infty$) for all i, which is also $p_i \approx \theta_i/n$. This can be achieved by letting q tend to infinity and taking as before $m = s^q$.

Since the number of terms in a summation is $m = O(n)$, we perform the asymptotic expansions (in negative powers of n) of μ_n and $m_n^{(2)}$. We have

$$\left(1 - \frac{a}{n}\right)^n = e^{-a}\left(1 - \frac{1}{2}\frac{a^2}{n}\right) + O\left(\frac{1}{n^2}\right) \quad \text{as} \quad n \to \infty.$$

This gives

$$\mu_n = \sum_i \left(e^{-\theta_i}(1+\theta_i) + \frac{1}{2n} e^{-\theta_i}\theta_i^2(1-\theta_i) + O\left(\frac{1}{n^2}\right)\right) \quad \text{and} \quad \gamma_n = m - \mu_n. \tag{3}$$

Considering $m_n^{(2)}$ we use once more asymptotic expansions of large powers of binomials. Moreover, when we take the asymptotic development of $m_n^{(2)}$, the double sum is composed of terms like $\sum_i \sum_j a_i b_j$ (where a_i and b_j respectively are functions of θ_i and θ_j). We therefore have

$$\sum_i \sum_j a_i b_j = \sum_i a_i \sum_j b_j = \sum_i a_i \sum_i b_i.$$

This simple remark induces that the complexity of computing the second moment is that of a simple summation.

Putting together the expansions for $m_n^{(2)}$, we obtain

$$\sigma_n^2 = m_n^{(2)} - \mu_n^2 \approx \sum_i e^{-\theta_i}(1+\theta_i)\left(1 - e^{-\theta_i}(1+\theta_i)\right) - \frac{1}{n}\left(\sum_i \theta_i^2 e^{-\theta_i}\right)^2. \tag{4}$$

3.3 Common q-grams to two sequences

We consider now two random sequences B and C with respective lengths $b+q-1$ and $c+q-1$ over an alphabet $\Sigma = \{l_1, l_2, \ldots, l_s\}$. The sequences B and C are generated by two Bernoulli models M_b and M_c where the probability of letter l_i respectively are ω_i and π_i. This induces generally for a q-gram and for the associated urn two different probability with respect to the two models. We denote by p_i (resp. x_i) the probability of urn i in the model M_b (resp. M_c).

As mentioned previously, we associate to the q-grams of sequence B white balls and to the q-grams of sequence C black balls. We are interested in the statistics of the number of bicolor collisions in a system of $m = s^q$ urns.

As in the previous section, we can use Poissonization-Depoissonization. However, this time, we need two poissonian flows into the urn system, one for the white balls and one for the black balls. This implies again that the flows into each urn are Poisson and that the urns behave independently of each other. The (double) Depoissonization follows the same way as in the preceding Section and we skip the corresponding proofs.

The multivariate generating function counting the urns without bicolor collisions is now

$$F(z,t,u) = \prod_{0 \le i \le s^q - 1} \left(e^{p_i z + x_i t} + (u-1)(e^{p_i z} + e^{x_i t} - 1)\right) = \sum f_{kbc} u^k z^b t^c,$$

where f_{kbc} is the probability that there are k urns without bicolor collisions when throwing b white and c black balls in the urns.

Intuitively, throwing v white and w black balls in the urn i has probability $p_i^v x_i^w$; the corresponding generating function is $(p_i z)^v (x_i t)^w / v! w!$; summing up over all the indices gives the generating function $e^{p_i z + x_i t}$ for urn i. The urns without bicolor collisions either contain white balls only $(e^{p_i z} - 1)$ or black balls only $(e^{x_i t} - 1)$ or are empty (1). Marking these urns with the parameter u provides the factor $(e^{p_i z} + e^{x_i t} - 1)$ that has also to be subtracted from the generating function to avoid double counting.

The expectation μ_{bc} and second moment $m_{bc}^{(2)}$ of the statistics of q-grams common to B and C respectively are

$$\mu_{bc} = [z^b t^c] \left. \frac{\partial F(z,t,u)}{\partial u} \right|_{u=1} \quad \text{and} \quad m_{bc}^{(2)} = [z^b t^c] \left. \frac{\partial}{\partial u} u \frac{\partial F(z,t,u)}{\partial u} \right|_{u=1}.$$

We consider an asymptotic model where b and c tend to infinity and where, for all i, we have $b \times p_i \to \theta_i$ and $c \times x_i \to \kappa_i$. This gives

$$\mu_{bc} = \sum_i e^{-\theta_i} + e^{-\kappa_i} - e^{\theta_i} e^{-\kappa_i} - \frac{1}{2b} \theta_i^2 e^{-\theta_i}(1-e^{-\kappa_i}) - \frac{1}{2c} \kappa_i^2 e^{-\kappa_i}(1-e^{-\theta_i}) + o(1) \quad \text{and} \quad \gamma_{bc} = m - \mu_{bc}, \tag{5}$$

where γ_{bc} is the asymptotic expectation of the number of bicolor collisions.

For the variance, we obtain (omitting second order terms),

$$\begin{aligned} \sigma_{bc}^2 = m_{bc}^{(2)} - \mu_{bc}^2 \approx & \sum_i (e^{-\theta_i} + e^{-\kappa_i} - e^{\theta_i} e^{-\kappa_i})(1 - e^{-\theta_i} - e^{-\kappa_i} + e^{\theta_i} e^{-\kappa_i}) \\ & - \frac{1}{b} \left(\left(\sum_i \theta_i e^{-\theta_i} \left(1 - e^{-\kappa_i}\right) \right)^2 - \sum_i \theta_i^2 e^{-2\theta_i} \left(1 - e^{-\kappa_i}\right)^2 \right) \\ & - \frac{1}{c} \left(\left(\sum_i \kappa_i e^{-\kappa_i} \left(1 - e^{-\theta_i}\right) \right)^2 - \sum_i \kappa_i^2 e^{-2\kappa_i} \left(1 - e^{-\theta_i}\right)^2 \right) \end{aligned} \tag{6}$$

When B and C have same length $n+q-1$ and same distribution of letters (with p_i probability of urn i and $n \times p_i \to \theta_i$), these formula reduce to

$$\gamma_n = m - \mu_n = m - \sum_i \left(e^{-\theta_i}(2 - e^{-\theta_i}) - \frac{\theta_i^2 e^{-\theta_i}}{n}(1 - e^{-\theta_i}) \right) + o(1) \quad \text{and} \tag{7}$$

$$\begin{aligned} \sigma_n^2 = m_n^{(2)} - \mu_n^2 \approx & \sum_i e^{-\theta_i}(2 - e^{-\theta_i}) \left(1 - e^{-\theta_i}(2 - e^{-\theta_i})\right) \\ & - \frac{2}{n} \left(\left(\sum_i \theta_i e^{-\theta_i} \left(1 - e^{-\theta_i}\right) \right)^2 - \sum_i \theta_i^2 e^{-2\theta_i} (1 - e^{-\theta_i})^2 \right) \end{aligned} \tag{8}$$

3.4 Poisson convergence

Among several applications of the Chen-Stein method [Ste70, Che75], Barbour and Holst [BH89] consider a sum W of partially dependent variables I_k. Under some technical conditions, they prove (Theorem 2.3) that

$$d(W, \mathcal{P}_\mu) \leq \mathbf{min}(1,\mu) \left(1 - \frac{\mathbf{Var}(W)}{\mu} \right),$$

where $\mathcal{P}_\mu$ is a Poisson distribution of parameter μ and $d(W, \mathcal{P}_\mu)$ is the total variation distance ¶ between W and $\mathcal{P}_\mu$. They apply this result to a system of urns where there is probability p_i that a ball falls in urn i.

¶ The total variation distance between two positive integer random variables of respective probability generating functions $\sum_n f_n z^n$ and $\sum_n g_n z^n$ is the sum $\sum_n |f_n - g_n|$.

(Balls are thrown independently of each other). Using a coupling method, they prove Poisson convergence of the number of empty urns.

We consider bicolor collisions in the system of Section 3.3 ($m = s^q$ urns, b white balls and c black balls, and probability p_i (resp. x_i) that a white (resp. black) ball falls into urn i).

Let $I_k = 1$ if there is a bicolor collision in urn k and $I_k = 0$ otherwise. The random variable W_{bc} counts the number of urns with bicolor collisions and can be represented as

$$W_{bc} = I_1 + \cdots + I_m.$$

We use the following coupling: if there are no white (resp. black) balls in urn k, for each urn i with $i \neq k$ and for each white (resp. black) ball in urn i (if any), throw the ball in urn k with probability p_k and leave it in urn i with probability $1 - p_k$. Repeat this step until there are white and black balls (collision) in urn k. The number of steps done is finite with probability one. Once this done, let $J_{ik} = 1$ if there is a bicolor collision in urn i and $J_{ik} = 0$ otherwise, and $I_{ik} = I_i$ for all k. We have $J_{ik} \leq I_{ik}$ for $i \neq k$ and for each k,

$$\mathcal{L}(J_{1k}, \ldots, J_{mk}) = \mathcal{L}(I_{1k}, \ldots, I_{mk} | I_k = 1).$$

It follows then from Theorem 2.3 of Barbour and Holst [BH89] that

$$d(W_{bc}, \mathcal{P}_{\gamma_{bc}}) \leq \mathbf{min}(1, \gamma_{bc}) \left(1 - \frac{\sigma_{bc}^2}{\gamma_{bc}}\right),$$

where γ_{bc} is given in Equation 5 and $\sigma_{bc}^2 = \mathbf{Var}(W_{bc})$ in Equation 6.

The same technique applies to repeated q-grams in the independent model, proving also Poisson convergence of this statistics.

These results imply Gaussian convergence when the parameters of the Poissons tend to infinity.

4 Experimental comparison between the dependent and the independent model

Figures 1 and 2 compare for repeated and common q-grams the theoretical results of the independent case with the simulations made for the independent and the dependent cases. This shows that the independent approximation is excellent for the expectation. (Similarly, the expectation of number of occurrences of words in random texts is independent of autocorrelation, in contrast to the standard deviation). The second moment differs clearly, which implies different behaviors of the corresponding parameters in suffix-trees and in tries.

When considering two random sequences of length 300 over the DNA alphabet, with the letter distribution of *H. influenzae* and 10-grams, we obtain for the statistics of common q-grams $\gamma_n = 0.1523$ and, for the simulations, 0.1514 (resp. 0.1503) in the independent (resp. dependent) model. We also have $\sigma_n^2 = 0.3888$ and for the simulations 0.3884 (resp. 0.5051). The total variation distance of the distribution obtained by simulations (independent case) and a Poisson of parameter 0.1523 is under 10^{-4}.

We used the algorithm described in the following section to compute the theoretical values of the parameters under consideration here.

The values of q_1 to q_{i-1} have been computed previously when Procedure Calcsum is entered and $d = s - i$.
$s = |\Sigma|$ and q are handled as global constants.

Procedure Calcsum (f, d, n, ϕ)**:**
    $i = s - d$

    $u = \sum_{k=1}^{i-1} q_k$

    If $d > 1$ **Then**
        For j **To** $s - u$ **Do**
            $q_i = j$
            $f =$ **Calcsum**$(f, d-1, n, \phi)$
        End of for
    Else
        $q_s = q - \sum_{k=1}^{s-1} q_k$

        $f = f + \frac{q!}{q_1!q_2!\ldots q_s!}\phi(\theta_{q_1,\ldots,q_s}, n)$

    End of if
    Return (f)
End of procedure

$$\theta_\xi = \theta_{q_1,\ldots q_s} = n \times \omega_1^{q_1}\omega_2^{q_2}\ldots\omega_s^{q_s}$$

$$\phi_1 = \left(e^{-\theta_\xi}(1+\theta_\xi) + \frac{1}{2n}e^{-\theta_\xi}\theta_\xi^2(1-\theta_\xi)\right)$$

$$\gamma_n = m - \textbf{Calcsum}(0, s, n, \phi_1)$$

$$\phi_2 = e^{-\theta_\xi}(1+\theta_\xi)\left(1 - e^{-\theta_\xi}(1+\theta_\xi)\right)$$

$$\phi_3 = \theta_\xi^2 e^{-\theta_\xi}$$

$$\sigma_n^2 = \textbf{Calcsum}(0, s, n, \phi_2) - \tfrac{1}{n}\left(\textbf{Calcsum}(0, s, n, \phi_3)\right)^2$$

Fig. 3: Recursive procedure to compute the expectation γ_n and variance σ_n^2 of repeated q-grams.

5 Algorithms and complexity

When considering Equations such as 3, 4, 7 and 8 we can group together the indices i for urns with same probability (corresponding to words with same distribution of letters). For any two indices i and j in the same group, we have $\theta_i = \theta_j$. Considering the size s of the alphabet and q-grams, let $T_{s,q}$ be the number of groups so defined; this is also the number of terms in the (reduced) summations. The population $P_{s,q}$ of a group whose q-grams are composed of q_1 (resp. $q_2, \dots q_s$) letters l_1 (resp. $l_2, \dots, l_s$) is the multinomial

$$P_{s,q} = \binom{q}{q_1, q_2, \dots, q_s}.$$

Applying this to Equation 7, we get

$$\gamma_n \approx m - \sum_{q_1+q_2+\cdots+q_s=q} \binom{q}{q_1, q_2, \dots, q_s} \left(2e^{-\omega_1^{q_1}\omega_2^{q_2}\dots\omega_s^{q_s} n} - e^{-2\omega_1^{q_1}\omega_2^{q_2}\dots\omega_s^{q_s} n}\right) \text{ as } n \to \infty,$$

where ω_i is the probability of occurrence of letter l_i. The number of summations $T_{s,q}$ is the number of compositions of the integer q with s positive or null summands. This is equal to the number of compositions of $q+s$ with s strictly positive summands. The generating function $C_s(z)$ of compositions with s (strictly positive) summands is

$$C_s(z) = \left(\frac{z}{1-z}\right)^s,$$

and therefore we have

$$T_{s,q} = [z^{q+s}] \left(\frac{z}{1-z}\right)^s = \binom{q+s-1}{s-1}.$$

See Flajolet and Sedgewick [FS] for a combinatorial introduction to compositions of integers.

$T_{s,q}$ is the complexity of computations of the formula giving the expectations and standard deviations of the random variables we considered. This is 286 for 10-grams and DNA and 1540 (resp. 8855) for 3-grams (resp. 4-grams) and amino-acids (20 letters alphabet).

Algorithms. We provide in Figure 3 a recursive algorithm to compute in the independent model the expectation γ_n and the standard deviation σ_n^2 of repeated q-grams (Equations 3 and 4 of Section 3.2). The same algorithm applies to compute the expectation γ_n and standard deviation σ_n^2 of common q-grams or of urns with bicolor collisions (Equations 7 and 8 of Section 3.3).

6 Other results about the dependent model

Dependent model and regular systems for repeated q-grams. We consider again an alphabet of size s and an integer q. We prove in this section that the language of words containing exactly e repeated q-grams is regular, for all e. This construction is however hyperexponential and therefore not applicable to practical needs.

We virtually consider the de Bruijn graph $B(s,q)$ and for each possible vertex v_i we consider three possible labels 0, 1 and 2. We make $3^{(s^q)}-1$ copies of the de Bruijn graph to handle all possible configurations of labels in the copies. A vertex with label 1 corresponds to the first repetition of a q-gram. As soon as such a q-gram is met again, the label of this vertex changes to 2; this means a jump to another copy.

We remark that any of the copies of the de Bruijn graph can be represented by the sequence of its s^q labels; this assigns a number N to this copy, namely $N = \sum_{0 \le v \le s^q-1} m(v) \times 3^v$, where $m(v)$ is the label of v.

Let $\delta(v,l)$ be the transition function of the de Bruijn graph $B(s,q)$ seen as an automaton (all states are terminal and this automaton recognizes all the sequences).

We connect the copies of the de Bruijn graph, making one large automaton $\mathcal{A}$ with transition function Δ. A trie $\mathcal{T}$ whose leaves are all possible q-grams is also connected to the original de Bruijn automaton; the root of this trie is the initial state of $\mathcal{A}$.

A state in automaton $\mathcal{A}$ (without considering those of $\mathcal{T}$) is a sequence of two numbers, the number of a state in the original de Bruijn graph $B(s,q)$ and the number of the copy. We consider a state $V_1 = [v, N_1]$ in $\mathcal{A}$. Let $m(V_1) = m(v) \in \{0,1,2\}$ be the label of the vertex V_1. We remark that $m(V_1)$ is the value of the v-th digit of N_1 in base 3. We have the following cases.

- If $m(V_1) = 0$ or $m(V_1) = 1$ then let $\Delta([v,N_1],l) = [\delta(v,l),N_2] = V_2$ where N_2 is obtained by adding 1 to the v-th digit of N_1 in base 3 (this implies that $m(V_2) = m(V_1) + 1$).
- If $m(V_1) = 2$ then $\Delta([v,N_1],l) = [\delta(v,l),N_1]$.

We mark with a letter u in $\mathcal{A}$ all states whose label is 1. This implies that a mark is emitted at the first repetition of a q-gram. We proceed now as in Nicodème *et al.* [NSF02] and in Nicodème [Nic00] by applying the Chomsky-Schützenberger algorithm [CS63] to the marked automaton $\mathcal{A}$. All the states of the original de Bruijn automaton were terminal, and all the states of $\mathcal{A}$ are terminal. This implies that the marked automaton $\mathcal{A}$ recognizes all the sequences where a mark u (a letter of size zero) is inserted after each first repetition of any q-gram. The Chomsky-Schützenberger theorem about regular languages then asserts that the generating function $F(z,u)$ of the language recognized by the marked automaton $\mathcal{A}$ is rational. Moreover the number of repeated q-grams is bounded by s^q; this implies that $F(z,u)$ is a polynomial in u and that the general form of $F(z,u)$ is

$$F(z,u) = \sum_{0 \le i \le s^q} u^i \frac{g_i(z)}{h_i(z)} = \sum_{0 \le i \le s^q} u^i f_i(z),$$

where $g_i(z)$ and $h_i(z)$ are polynomials and $f_i(z)$ is the generating function of sequences with exactly i repeated q-grams.

We remark that this construction applies to the Bernoulli non-uniform model and also to any Markov model for the sequence.

Limiting distributions. We consider the random variable K counting the number of common q-grams to two sequences of same size and same length. In Figure 4 we compare simulation results of this normalized variable and a standard Gaussian in the uniform ($p_0 = 0.5$) and in the biased case ($p_0 = 0.1$). Although there is a slight shift of the tails to the right, it is reasonable to conjecture that the limiting distribution is Gaussian in both cases; this should also hold when considering sequences with different lengths and letter distributions. Simulation results for repeated q-grams in the uniform and biased models also indicate that the limiting distributions should be Gaussian.

7 Conclusion

We considered in this article the statistics of repeated q-grams in a sequence or of common q-grams to two sequences. We compared this "dependent" model to an independent model where q-grams are drawn independently of each other. We analyzed this independent model by use of urn models and showed experimentally that the first moment of the dependent model behaves like that of the independent model; this holds no more for the second moments but there seems to be a strong correlation between the second

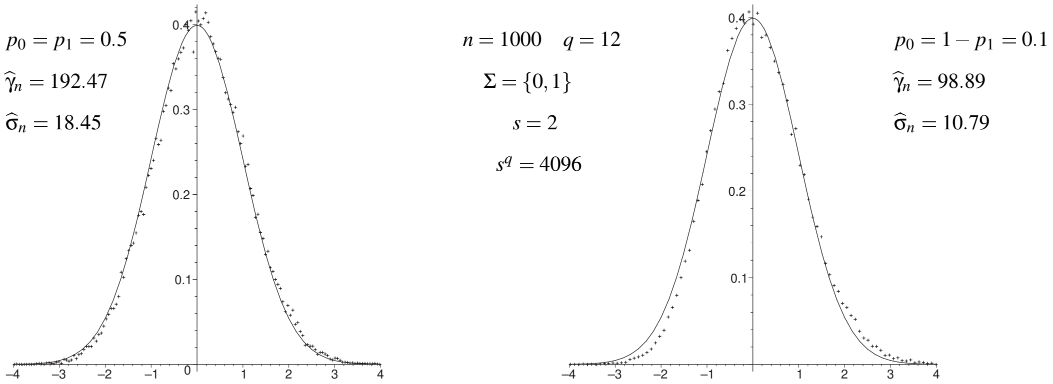

Fig. 4: Distribution of the normalized random variable $(K - \widehat{\gamma}_n)/\widehat{\sigma}_n^2$ where K counts the number of common q-grams to two random sequences of size 1000; $\widehat{\gamma}_n$ and $\widehat{\sigma}_n$ are the empirical expectation and standard deviation of the variable K. The solid lines represents the normal distribution $\mathcal{N}(0,1)$. (Results obtained after 50000 simulations).

moments in the two models. We provide a compact, simple and fast algorithm to compute the two first moments of repeated or common q-grams in the independent model. A direct analysis of the dependent model is challenging; the methods of Jacquet and Szpankowski [JS94], of Rahman and Rivals [RR03] and of Fayolle [Fay02] could lead to a successful approach. How far the present work can be extended to analyze the filtering step of the alignment package BLAST [AGM+90] is an opened question.

Acknowledgement We thank Philippe Jacquet and Bruno Salvy for helpful discussions.

References

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215:403–410, 1990.

[AHU74] A.V. Aho, J.E. Hopcroft, and J.D. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, 1974.

[Apo85] A. Apostolico. The Myriad Virtues of Suffix Trees, *Combinatorial Algorithms on Words*. Springer Verlag, 3I F12, 1985.

[BCF+99] S. Burkhardt, A. Crauser, P. Ferragina, H.-P. Lenhof, E. Rivals, and M. Vingron. Q-gram based database searching using a suffix array (QUASAR). In *Third Annual International Conference on Computational Molecular Biology*, pages 77–83, Lyon, April, 1999.

[BH89] A. D. Barbour and L. Holst. Some applications of the Chen-Stein method for proving Poisson convergence. *Adv. Appl. Prob.*, 21:74–90, 1989.

[Che75] L. H. Chen. Poisson approximation for dependent trials. *Annals of Probability*, 3:534–545, 1975.

[CS63] N. Chomsky and M. P. Schützenberger. The algebraic theory of context-free languages. *Computer Programming and Formal Languages,*, pages 118–161, 1963. P. Braffort and D. Hirschberg, eds, North Holland.

[CVF01] J. Clement, B. Vallée, and P. Flajolet. Dynamical sources in information theory: A general analysis of trie structures. *Algorithmica*, 29:307–369, 2001.

[dB46] N. J. de Bruijn. A combinatorial problem. *Proc. Kon. Ned. Akad. Wet.*, 49(*Part 2*):758–764, 1946.

[DGG01] M. Drmota, D. Gardy, and B. Gittenberger. A unified presentation of some urns models. *Algorithmica*, 29:120–147, 2001.

[Fay02] Julien Fayolle. Paramètres des arbres suffixes dans le cas de sources simples, 2002. Mémoire de DEA, Université Paris VI.

[FGP03] P. Flajolet, J. Gabarró, and H. Pekari. Analytic urns. submitted, 2003.

[FGT92] Philippe Flajolet, Danièle Gardy, and Loÿs Thimonnier. Birthday Paradox, Coupon Collector, Caching Algorithms and Self organizing Search. *Discrete Applied Mathematics*, 39:207–229, 1992.

[FKT88] P. Flajolet, P. Kirschenhofer, and R. F. Tichy. Deviations from uniformity in random strings. *Probability Theory and Related Fields*, 80:139–150, 1988.

[FS] P. Flajolet and R. Sedgewick. Analytic combinatorics. Book in preparation.

[GBY91] G. Gonnet and R. Baeza-Yates. *Hanbook of Algorithms and Data Structures: in Pascal and C,* second ed.. Addison-Wesley, 1991.

[GO78] L. Guibas and A. Odlyzko. Periods in strings. *J. Combin. Theory*, A(30):19–43, 1978.

[JK77] N. L. Johnson and S. Kotz. *Urn Models and their Applications*. John Wiley and Sons, 1977.

[JS91a] P. Jacquet and W. Szpankowski. Analysis of digital tries with markovian dependencies. *IEEE Transactions on Information Theory*, 37(5):1470–1475, 1991.

[JS91b] P. Jacquet and W. Szpankowsli. What can we learn about suffix trees from independent tries? In *1991 Workshop on Algorithms and Data Structures,* Lecture Notes in Computer Science 520, pages 228–229. Springer Verlag, 1991.

[JS94] P. Jacquet and W. Szpankowsli. Autocorrelation on words and its applications. Analysis of Suffix Trees by String Ruler Approach. *J. Combin. Theory*, A(66):237–269, 1994.

[JU91] P. Jokinen and E. Ukkonen. Two algorithms for approximate string matching in static texts. In *Proceedings Mathematical Foundations of Computer Science 1991* (ed. A. Tarlecki)*, Lecture Notes in Computer Science 520*, pages 240–248. Springer Verlag, 1991.

[Knu73] D. Knuth. *The Art of Computer Programming. Sorting and Searching*. Addison-Wesley, 1973.

[KSC78] V. F. Kolchin, B. Sevast'yanov, and V. Chistyakov. *Random Allocations*. Wiley, New-York, 1978.

[Nic00] P. Nicodème. Regexpcount, a symbolic package for counting problems on regular expressions and words. In *German Conference on Bioinformatics GCB, Heidelberg*, pages 63–73. Logos Verlag, Berlin, 2000.

[NSF02] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–618, 2002. Extended version of an article published in the proceedings of 7th Annual European Symposium on Algorithms ESA'99, Prague, July 1999.

[RR00] S. Rahmann and E. Rivals. Exact and efficient computation of the expected number of missing and common words in random texts. In *11th Symposium on Combinatorial Pattern Matching (CPM 2000). Lecture Notes in Computer Science, vol.1848*, pages 375–387. Springer, 2000.

[RR03] S. Rahmann and E. Rivals. The number of missing words in random texts. *Combinatorics, Probability and Computing*, 12:73–87, 2003.

[RS98] M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1998.

[Sed88] R. Sedgewick. *Algorithms in C: Fundamentals, Data Structures, Sorting, Searching : in Pascal and C,* third ed.. Addison-Wesley, reading, Mass., 1988.

[SF96] Robert Sedgewick and Philippe Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley Publishing Company, 1996.

[SS98] W. Szpankowski and E. Sutinen. On the collapse of q-gram filtration. In *Proc. Int. Conf FUN with Algorithms, Elba, Italy*, pages 178–193, 1998.

[Ste70] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *6th Berkeley Symp. Math. Statis. Prob. 2*, pages 583–602, 1970.

[Szp93] W. Szpankowski. A generalized suffix-tree and its (un)expected behaviors. *SIAM Journal of Computing*, 22:1176–1198, 1993.

[Szp01] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley-Interscience, 2001.

[Wei73] P. Weiner. Linear Pattern Matching Algorithms. In *14-th Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.