Combinatorics, Probability and Computing (2001) 10, 293–308. © 2001 Cambridge University Press
DOI 10.1017/S096354830100476X Printed in the United Kingdom

Poisson Approximation for the Non-Overlapping Appearances of Several Words in Markov Chains

OURANIA CHRYSSAPHINOU, STAVROS PAPASTAVRIDIS and EUTICHIA VAGGELATOU

Department of Mathematics, University of Athens, Greece
(e-mail: ocrysaf@math.uoa.gr
spapast@math.uoa.gr
evagel@math.uoa.gr)

Received 14 February 2000; revised 4 October 2000

Let $X_1, \ldots, X_n$ be a sequence of r.v.s produced by a stationary Markov chain with state space an alphabet $\Omega = \{\omega_1, \ldots, \omega_q\}$, $q \geqslant 2$. We consider a set of words $\{A_1, \ldots, A_r\}$, $r \geqslant 2$, with letters from the alphabet Ω. We allow the words to have *self-overlaps* as well as *overlaps* between them. Let $\mathscr{E}$ denote the event of the appearance of a word from the set $\{A_1, \ldots, A_r\}$ at a given position. Moreover, define by N the number of non-overlapping (competing renewal) appearances of $\mathscr{E}$ in the sequence $X_1, \ldots, X_n$. We derive a bound on the total variation distance between the distribution of N and a *Poisson* distribution with parameter $\mathbb{E}N$. The Stein–Chen method and combinatorial arguments concerning the structure of words are employed. As a corollary, we obtain an analogous result for the i.i.d. case. Furthermore, we prove that, under quite general conditions, the r.v. N converges in distribution to a Poisson r.v. A numerical example is presented to illustrate the performance of the bound in the Markov case.

1. Introduction

Problems concerning appearances of words produced under the independent identically distributed (i.i.d.) hypothesis, and more generally under the Markovian hypothesis, are of great interest because of their applications in a variety of fields, such as communication systems, cryptology, and analysis of biological sequences. For general results we refer to the papers by Blom and Thorburn [6], and Guibas and Odlyzko [18, 19], and the references therein. For more specific results, see Arratia, Martin, Reinert and Waterman [3], Biggins and Cannings [5], Breen, Waterman and Zhang [7], Chryssaphinou and Papastavridis [9, 10], Chryssaphinou, Papastavridis and Vaggelatou [11], Fu [14], Gerber and Li [15], Godbole [16], Godbole and Schaffner [17], Fousler and Karlin [13], Karlin and Ost [20], Neuhauser [23], Régnier [24], Reinert and Schbath [25], Robin and Daudin [26], Schbath [27], Schwager [28], Tanushev and Arratia [29] and Waterman [30].

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

During the past decade, the Stein–Chen method has been used by many authors to derive bounds on the error that occurs when approximating r.v.s related to the above problems using Poisson distributions. For general literature on this topic, we refer to Arratia, Goldstein and Gordon [1, 2], Barbour, Holst and Janson [4] and Chen [8].

In order to present our results we need some notation. We consider a doubly infinite sequence $\{X_a\}_{a\in Z}$ of trials produced by an irreducible, aperiodic stationary Markov chain on a finite alphabet $\Omega = \{\omega_1, \ldots, \omega_q\}$, $q \geqslant 2$, with transition matrix $P = \{p(z,w)\}_{z,w\in\Omega}$ and stationary distribution $\pi = \{\pi_z\}_{z\in\Omega}$. Also, let $\{A_1, \ldots, A_r\}$, $r \geqslant 2$, be a set of words with letters from the alphabet Ω, where each word A_i has the form $a_{i,1} \ldots a_{i,k_i}$, $i = 1, \ldots, r$. We define the indicator r.v. $I_a(A_i)$ of the appearance of A_i at a position $a \in Z$ as follows:

$$I_a(A_i) := I[X_{a-k_i+1} = a_{i,1}, \ldots, X_a = a_{i,k_i}]$$

with mean

$$\chi_i := \pi(a_{i,1}) \prod_{t=1}^{k_i-1} p(a_{i,t}, a_{i,t+1}).$$

Let $\mathscr{E}$ denote the event of the appearance of a word from the set $\{A_1, \ldots, A_r\}$ at a given position. We assume that this set is *reduced* in the sense of Guibas and Odlyzko [19]. This means that A_i is never a subword of A_j, for any two words in the above set. Then the indicator r.v. $I_a(\mathscr{E})$ of the occurrence of $\mathscr{E}$ at position $a \in Z$ is given by

$$I_a(\mathscr{E}) = \sum_{i=1}^{r} I_a(A_i)$$

with mean

$$\chi =: \sum_{i=1}^{r} \chi_i.$$

We are interested in the r.v. N which counts the non-overlapping appearances of $\mathscr{E}$ in the finite sequence $\{X_a\}_{1\leqslant a\leqslant n}$. The r.v. N is also called *competing renewal* in the literature (see Tanushev and Arratia [29]). For example, if we have the set $\{A_1, A_2\} = \{ACGA, CGAC\}$, where the words have letters from the DNA alphabet $\{A, C, G, T\}$, then, in the sequence

$$CGT\,\overbrace{ACGA}\,\overbrace{CGAC}\,GTC\,\overbrace{ACGA}\,\overbrace{ACGA}\,CCGT,$$

the r.v. N takes the value 4.

We set $M := \max_{1\leqslant i\leqslant r} k_i$ and $m := \min_{1\leqslant i\leqslant r} k_i$ and define the r.v. N to be

$$N := \sum_{a=m}^{n} N_a,$$

where

$$N_a = \sum_{i=1}^{r} I_a(A_i) \prod_{b=a-k_i+1}^{a-1} (1 - N_b),\ a \in Z.$$

The r.v. N_a is the indicator of a non-overlapping (competing renewal) appearance of $\mathscr{E}$

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

ending at position a. In the above definition, we consider an infinite sequence of N_as in order to neglect the edge effects produced by the finite sequence and therefore to facilitate calculations. In this case, we have $\mathbb{E}N = (n - m + 1)\,\mathbb{E}N_a$.

Our main aim is to find a bound on the total variation distance between the distribution of N and the Poisson distribution with parameter the mean $\mathbb{E}N$ of N. The proof is based on the Stein–Chen method (Arratia, Goldstein and Gordon [1, 2], Barbour, Holst and Janson [4], Chen [8]). This in turn will lead to the convergence of N to a Poisson r.v., under general conditions. The main difficulty of the problem arises from the fact that we have to take into account the possible autocorrelations (self-overlaps), as well as the correlations (overlaps) of the words in the sense of Guibas and Odlyzko [19].

Reinert and Schbath [25] examined the overlapping appearances of several words and proved the convergence of this number to a compound Poisson r.v., under the condition that the amount of overlap between two words should be sufficiently small. Our results concern the non-overlapping appearances of several words and hold for any kind of words, with no restriction on the amount of overlap between the words. The fact that overlaps of any sort are allowed does not facilitate an approach that treats every word separately, as in Reinert and Schbath [25], but necessitates seeing the totality of the words.

Schwager [28] introduced a recursive way to find the distribution of N in the case where the trials are produced by a high-order Markov chain. Later, Fu [14], by embedding the problem to a certain Markov chain, computed the distribution of N for the i.i.d. case, using rather large matrices. Koutras [21] improved the former technique introducing the Markov chain embeddable variables of binomial type (MVB).

In this work, we are looking for closed form bounds that can be evaluated in an easier way.

This paper is organized as follows. In Section 2, we present bounds for the total variation distance between the distribution of N and the Poisson distribution with parameter $\mathbb{E}N$ for the Markov and i.i.d. cases (Theorem 2.1 and Corollary 2.3, respectively). We give a numerical example which shows the behaviour of the derived bound in the Markov case. In Section 3, using the previous results, we prove the convergence of N to a Poisson distribution, under general conditions (Theorems 3.2 and 3.3).

2. Main results

As we mentioned in the previous section, the main purpose of this work is to find a bound on the total variation distance between the distribution $\mathcal{L}(N)$ of N and the Poisson distribution with mean $\mathbb{E}N$. Generally, for any two r.v.s X_1, X_2 defined on the same probability space $(\mathcal{X}, \mathbb{P})$ and with values in Z_+, the total variation distance between their distributions $\mathcal{L}(X_1)$ and $\mathcal{L}(X_2)$ is, by definition,

$$d_{TV}(\mathcal{L}(X_1), \mathcal{L}(X_2)) := \sup_{A \subseteq Z_+} |\mathbb{P}(X_1 \in A) - \mathbb{P}(X_2 \in A)| = \min \mathbb{P}(Y_1 \neq Y_2),$$

where the minimum ranges over all couplings (Y_1, Y_2) of $\mathcal{L}(X_1)$, $\mathcal{L}(X_2)$ (see Barbour, Holst and Janson [4]).

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

The r.v.s N_a, which determine N, have a strong kind of dependence. Therefore, we cannot directly apply the Stein–Chen method to approximate the r.v. N. For that reason we use an auxiliary r.v. denoted by Q and defined by

$$Q := \sum_{a=m}^{n} Q_a,$$

where

$$Q_a := \sum_{i=1}^{r} I_a(A_i) \prod_{b=a-k_i+1}^{a-1} (1 - I_b(\mathscr{E})),\ a \in Z.$$

The indicator r.v. Q_a takes the value 1 if and only if $\mathscr{E}$ occurs at position a and it does not overlap another preceding appearance of $\mathscr{E}$. In other words, Q_a is the indicator r.v. of a *clump* of $\mathscr{E}$ ending at position a. Here, the clumps of $\mathscr{E}$ can be composed of *any* words from the set $\{A_1, \ldots, A_r\}$, whereas Reinert and Schbath [25] considered clumps consisting of the same word.

The reason for using the r.v.s Q_a is that these have a *weaker* kind of dependence than N_a, as we will see later on. So the total variation distance between the distribution of Q and a Poisson distribution can be bounded in a more efficient way than that of N. Moreover, the difference between N and Q becomes negligible in probability under quite general conditions (Lemma 3.1).

Before presenting the first main result of this work, we need to define the following quantities. Let $p^{(t)}(z,w)$ denote the t-order transition probability of the Markov chain $\{X_a\}_{a\in Z}$, and let $p_R^{(t)}(z,w) = \frac{\pi(w)p^{(t)}(w,z)}{\pi(z)}$ denote the t-order transition probability of the *reversed* Markov chain $\{X_a^R\}_{a\in Z}$, for $t \geqslant 1$. Then, using a coupling argument (Lindvall [22, p. 96]), it follows that

$$\max_{z\in\Omega}\left\{\max\left(\sum_{w\in\Omega}\left|p^{(t)}(z,w)-\pi(w)\right|, \sum_{w\in\Omega}\left|p_R^{(t)}(z,w)-\pi(w)\right|\right)\right\} \leqslant 2\varrho^t, \quad \forall\, t \geqslant 1, \tag{2.1}$$

where $\varrho := 1 - \min\{\delta(P), \delta(P_R)\}$, P_R is the transition matrix of the reversed chain and $\delta(H) := \sum_{w\in\Omega} \min_{z\in\Omega} h(z,w)$ for a transition matrix $H = \{h(z,w)\}_{z,w\in\Omega}$. Moreover, we define the quantity

$$\Xi := \sup_{t\geqslant 1} \max_{1\leqslant i,j\leqslant r} \frac{p^{(t)}(a_{j,k_j}, a_{i,1})}{\pi(a_{i,1})}.$$

Now we are ready to state the first result of this work, which concerns a bound for the distance $d_{TV}(\mathscr{L}(N), \mathrm{Po}(\mathbb{E}N))$, where $\mathscr{L}(N)$ denotes the distribution of N and $\mathrm{Po}(\mathbb{E}N)$ is the Poisson distribution with parameter $\mathbb{E}N$.

Theorem 2.1. *If N is the number of non-overlapping appearances of $\mathscr{E}$ in the Markov sequence $X_1, \ldots, X_n$, then*

$$d_{TV}(\mathscr{L}(N), \mathrm{Po}(\mathbb{E}N)) \leqslant B_{2.1}(s) := \left(1 - e^{-\mathbb{E}Q}\right) D_1(s) + \left(1 \wedge \sqrt{\frac{2}{e\mathbb{E}Q}}\right) D_2(s) + D_3,\ s \geqslant 1,$$

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

where

$$\begin{aligned} D_1(s) &:= (4M+2s-5)(\mathbb{E}Q_a + \chi\,\Xi) - \Xi\,(2m-1)\,\chi, \\ D_2(s) &:= 2\,\mathbb{E}Q\,\varrho^s\{2+2\varrho^s+\varrho^{2(M-1)+s}\}, \\ D_3 &:= \left\{1+\left(1\wedge\frac{1}{\sqrt{\mathbb{E}N}}\right)\right\}(\mathbb{E}N-\mathbb{E}Q). \end{aligned}$$

The quantity $s \geqslant 1$ is chosen so as to obtain the $\min_{s\geqslant 1} B_{2.1}(s)$. Besides, in order to apply the above bound, we need to evaluate the means $\mathbb{E}Q$ and $\mathbb{E}N$. We have $\mathbb{E}Q = (n-m+1)\mathbb{E}Q_a$. Moreover, the mean $\mathbb{E}Q_a$ is given by

$$\mathbb{E}Q_a = \sum_{i=1}^{r} \mathbb{P}\left(I_a(A_i)=1, \bigcap_{p\in\mathscr{P}'(A_i)} (I_{a-k_i}(A_i^p)=0), \bigcap_{\substack{j=1\\ j\neq i}}^{r} \bigcap_{\substack{v:b_v\in C_{ji},\\ b_v=1}} \left(I_{a-k_i}\left(A_j^{(k_j-v)}\right)=0\right)\right),$$

where, in general, $I_a(X)$ denotes the indicator of the event that a copy of the word X ends at position a. Moreover, $\mathscr{P}'(A_i)$ is the set of *principal* periods of the word A_i (Schbath [27]), C_{ji} is the correlation between the words A_j and A_i (Guibas and Odlyzko [19]) and A_i^p is the subword which consists of the p first letters of A_i. In particular, an integer p is a period of a word $A = a_1a_2\ldots a_k$ if $a_i = a_{i+p}$ for all $i = 1,\ldots,k-p$. The set $\mathscr{P}'(A)$ consists of the shortest period p_0 of A together with those which are not multiples of p_0. Notice that $\mathscr{P}'(A)$ is a subset of the whole set of periods of A; the latter is denoted by $\mathscr{P}(A)$ in the literature (Schbath [27]). Furthermore, the correlation C_{ji} of two words A_j and A_i with lengths k_j and k_i, respectively, is a vector $(b_{k_j}\cdots b_1)$ with $b_t \in \{0,1\}$, for $t = 1,\ldots,k_j$. If $b_t = 1$, the index t shows that the two words can make an overlap of length t. For example, consider the set of words $\{ACGAC, ACACA\}$. Then $\mathscr{P}'(ACGAC) = \{3\}$, $\mathscr{P}'(ACACA) = \{2\}$, $C_{12} = (00010)$ and $C_{21} = (00001)$.

In order to find a simpler formula for the mean $\mathbb{E}Q_a$, we define the sets

$$\mathscr{P}'(A_j,A_i) := \begin{cases} \mathscr{P}'(A_i), & \text{if } i=j, \\ \{k_j - v \text{ where } v : b_v \in C_{ji} \text{ and } b_v = 1\}, & \text{if } i\neq j, \end{cases}$$

and

$$\mathscr{B}_i := \{B : B = b_1\ldots b_{M-1} \text{ such that } B_p \neq A_j^p, \quad \forall\, p\in\mathscr{P}'(A_j,A_i), \quad \forall\, j=1,\ldots,r\},$$

where B_p denotes the p last letters of the word B (that is, $B_p = b_{M-p}\ldots b_{M-1}$). Now the mean $\mathbb{E}Q_a$ can be written in the form

$$\begin{aligned} \mathbb{E}Q_a &= \sum_{i=1}^{r} \mathbb{P}\left(I_a(A_i)=1, \bigcap_{j=1}^{r} \bigcap_{p\in\mathscr{P}'(A_j,A_i)} (I_{a-k_i}(A_j^p)=0)\right) \\ &= \sum_{i=1}^{r} \sum_{B\in\mathscr{B}_i} \mathbb{P}(I_{a-k_i}(B)=1, I_a(A_i)=1) \\ &= \sum_{i=1}^{r} \sum_{B\in\mathscr{B}_i} \mathbb{E}I_a(BA_i). \end{aligned} \tag{2.2}$$

Relation (2.2) will be helpful in the proof of Theorem 2.1. Notice also that there cannot

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

be found a simpler formula for $\mathbb{E}Q_a$, because the events $\{\bigcup_{p\in\mathscr{P}'(A_j,A_i)}(I_{a-k_i}(A_j^p)=1)\}$ and $\{\bigcup_{p\in\mathscr{P}'(A_l,A_i)}(I_{a-k_i}(A_l^p)=1)\}$, for $j,l=1,\ldots,r$, $j\neq l$, are not always disjoint.

For the case of one word A ($r=1$), formula (2.2) reduces to

$$\mathbb{E}Q_a=\mathbb{E}I_a(A)-\sum_{p\in\mathscr{P}'(A)}\mathbb{E}I_a(A^pA),$$

which was derived by Schbath [27].

Concerning the mean $\mathbb{E}N$, we present a result which is, in a way, an extension of Theorem 3.1 of Blom and Thorburn [6] to the case where the trials are produced by a stationary Markov chain. Following the same notation as in the previous reference, we will use the *leading numbers*, which are defined as follows:

$$\varepsilon_t(j,i)=\begin{cases}1, & \text{if there is an overlap of length } t \text{ between } A_j \text{ and } A_i,\\ 0, & \text{if there is no overlap of length } t \text{ between } A_j \text{ and } A_i,\end{cases}$$

and they concern the appearance of the word A_j before A_i, for $i,j=1,\ldots,r$.

The Markovian hypothesis and the combinatorial structure of the problem require some additional notation. We define

$$e_{ji} := \sum_{t=1}^{k_i-1}\varepsilon_t(j,i)C(j,i,t), \quad i,j=1,\ldots,r \text{ and } i\neq j$$

$$e_{ii} := 1+\sum_{t=1}^{k_i-1}\varepsilon_t(i,i)C(i,i,t), \quad i=1,\ldots,r,$$

where

$$C(j,i,t) := \begin{cases} p(a_{j,k_j},a_{i,t+1}), & \text{if } t=k_i-1,\\ p(a_{j,k_j},a_{i,t+1})\prod_{l=t+1}^{k_i-1}p(a_{i,l},a_{i,l+1}), & \text{if } t<k_i-1.\end{cases}$$

The quantity $C(j,i,t)$ can be thought of as being the probability of observing the k_i-t last letters of A_i successively, given the fact that the last letter of A_j has just occurred.

With the notation as above we have the following result.

Lemma 2.2. *The mean* $\mathbb{E}N$ *is given by*

$$\mathbb{E}N=(n-m+1)\sum_{i=1}^{r}y_i,$$

where $(y_1,\ldots,y_r)$ *is the solution of the* $r\times r$ *linear system of equations*

$$\sum_{j=1}^{r}y_je_{ji}=\chi_i, \quad i=1,\ldots,r.$$

Proof. When a word A_i appears at a given position, this appearance is either non-overlapping or it overlaps a preceding non-overlapping appearance of a word A_j, $j=$

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

$1,\ldots,r$. So, if $y_{a,i}$ denotes the probability of the non-overlapping appearance of a word A_i at a position a (that is, $y_{a,i} = \mathbb{P}(N_a = 1, I_a(A_i) = 1)$), then

$$\chi_i = \mathbb{E}I_a(A_i) = y_{a,i} + \sum_{j=1}^{r} \sum_{b=a-k_i+1}^{a-1} y_{b,j}\varepsilon_{b-a+k_i}(j,i)C(j,i,b-a+k_i), \tag{2.3}$$

where $\varepsilon_{b-a+k_i}(j,i)$ denotes whether there can be an overlap of length $b-a+k_i$, and the product $\varepsilon_{b-a+k_i}(j,i)\,C(j,i,b-a+k_i)$ gives the probability of observing the word A_i at position a conditional on the occurrence of A_j at position b. Because of stationarity, we may set $y_i = y_{a,i}$, $a = m,\ldots,n$, and relation (2.3) becomes

$$\begin{aligned}
\chi_i &= y_i + \sum_{j=1}^{r} y_j \sum_{t=1}^{k_i-1} \varepsilon_t(j,i)C(j,i,t) \\
&= y_i + y_i \sum_{t=1}^{k_i-1} \varepsilon_t(i,i)C(i,i,t) + \sum_{j\neq i}^{r} y_j \sum_{t=1}^{k_i-1} \varepsilon_t(j,i)C(j,i,t) \\
&= y_i \left\{ 1 + \sum_{t=1}^{k_i-1} \varepsilon_t(i,i)C(i,i,t) \right\} + \sum_{j\neq i}^{r} y_j \sum_{t=1}^{k_i-1} \varepsilon_t(j,i)C(j,i,t) \\
&= y_i e_{ii} + \sum_{j\neq i}^{r} y_j e_{ji} \\
&= \sum_{j=1}^{r} y_j e_{ji}, \quad i = 1,\ldots,r.
\end{aligned}$$

□

Remark. We note that Lemma 2.2 gives the asymptotic mean $\mathbb{E}N$, while Theorem 26 of Régnier [24] gives the mean taking into account the edge effects. Besides, the mean $\mathbb{E}N$ can be also calculated recursively using the results obtained by Chryssaphinou and Papastavridis [10] and probabilistic arguments.

Now we proceed to prove Theorem 2.1.

Proof of Theorem 2.1. Applying the triangular inequality, we get

$$\begin{aligned}
d_{TV}(\mathscr{L}(N), \mathrm{Po}(\mathbb{E}N)) \leqslant\ & d_{TV}(\mathscr{L}(N), \mathscr{L}(Q)) + d_{TV}(\mathscr{L}(Q), \mathrm{Po}(\mathbb{E}Q)) \\
& + d_{TV}(\mathrm{Po}(\mathbb{E}Q), \mathrm{Po}(\mathbb{E}N)).
\end{aligned} \tag{2.4}$$

We will find upper bounds for the distances at the right-hand side of this inequality.

A bound for the distance $d_{TV}(\mathrm{Po}(\mathbb{E}Q), \mathrm{Po}(\mathbb{E}N))$ is given by

$$d_{TV}(\mathrm{Po}(\mathbb{E}Q), \mathrm{Po}(\mathbb{E}N)) \leqslant \left(1 \wedge \frac{1}{\sqrt{\mathbb{E}N}}\right)(\mathbb{E}N - \mathbb{E}Q) \tag{2.5}$$

(see Remark 1.1.4 of Barbour, Holst and Janson [4]).

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

Furthermore, a bound for $d_{TV}(\mathscr{L}(N), \mathscr{L}(Q))$ can be obtained as follows:

$$\begin{aligned} d_{TV}(\mathscr{L}(N), \mathscr{L}(Q)) &\leqslant \mathbb{P}(N \neq Q) \leqslant \sum_{a=m}^{n} \mathbb{P}(N_a \neq Q_a) \\ &= \sum_{a=m}^{n} \mathbb{E}(N_a - Q_a) = \mathbb{E}(N - Q), \end{aligned} \tag{2.6}$$

because the r.v. $N_a - Q_a \geqslant 0$ takes only the values 1 and 0, so that $\mathbb{E}(N_a - Q_a) = \mathbb{P}(N_a \neq Q_a)$.

Next, we focus on the bounding of the term $d_{TV}(\mathscr{L}(Q), \mathrm{Po}(\mathbb{E}Q))$, which constitutes the main part of this proof. We will use the Stein–Chen method as it is presented in Arratia, Goldstein and Gordon [2, Theorem 1] or Barbour, Holst and Janson [4, Theorem 1.A]. For that purpose, we set $\Gamma = \{m, \ldots, n\}$ and $\Gamma_a = \Gamma \backslash \{a\}$. Moreover, we define the set Γ_a^s which contains all the indices b, other than a, for which the indicators Q_a and Q_b are 'strongly' dependent. Taking into account that Q_a and Q_b are 'weakly' dependent whenever $|b - a| \geqslant 2M + s - 2$, for some sufficiently large $s \geqslant 1$, we choose

$$\Gamma_a^s = \{b \in \Gamma_a : |b - a| \leqslant 2M + s - 3\}, s \geqslant 1.$$

The quantity s results from the Markovian structure of the problem, which requires a larger neighbourhood than in the i.i.d. case. Obviously, in the latter case, we take $s = 1$. We have

$$\begin{aligned} d_{TV}(\mathscr{L}(Q), \mathrm{Po}(\mathbb{E}Q)) \leqslant\ & \frac{1 - e^{-\mathbb{E}Q}}{\mathbb{E}Q} \sum_{a \in \Gamma} \left\{ (\mathbb{E}Q_a)^2 + \sum_{b \in \Gamma_a^s} \mathbb{E}Q_a \mathbb{E}Q_b + \sum_{b \in \Gamma_a^s} \mathbb{E}(Q_a Q_b) \right\} \\ & + \left(1 \wedge \sqrt{\frac{2}{e\mathbb{E}Q}} \right) \sum_{a \in \Gamma} n_a, \end{aligned} \tag{2.7}$$

where $n_a = \mathbb{E}|\mathbb{E}\{Q_a | \sigma(Q_b : b \in \Gamma_a^w)\} - \mathbb{E}Q_a|$ and $\Gamma_a^w = \Gamma_a \backslash \Gamma_a^s$. Since $|\Gamma_a^s| \leqslant 2(2M + s - 3)$, we have $\sum_{b \in \Gamma_a^s} \mathbb{E}Q_a \mathbb{E}Q_b \leqslant 2(2M + s - 3)(\mathbb{E}Q_a)^2$. Next we treat the term $\sum_{b \in \Gamma_a^s} \mathbb{E}(Q_a\, Q_b)$. We partition the set Γ_a^s in the following way:

$$\begin{aligned} \Gamma_a^s =\ & \{a - 2M - s + 3, \ldots, a - m\} \cup \{a - m + 1, \ldots, a - 1, a + 1, \ldots, a + m - 1\} \\ & \cup \{a + m, \ldots, a + 2M + s - 3\}. \end{aligned}$$

The mean $\mathbb{E}(Q_a Q_b)$ is zero whenever the appearances of two words from the reduced set $\{A_1, \ldots, A_r\}$ at positions a and b overlap. This takes place when $b \in \{a - m + 1, \ldots, a - 1, a + 1, \ldots, a + m - 1\}$. On the other hand, if $b \in \{a + m, \ldots, a + 2M + s - 3\}$, using relation (2.2) we get

$$\begin{aligned} \mathbb{E}(Q_a Q_b) &\leqslant \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{B \in \mathscr{B}_i} \mathbb{P}(I_{a-k_i}(B) = 1, I_a(A_i) = 1, I_b(A_j) = 1) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} \sum_{B \in \mathscr{B}_i} \mathbb{P}(I_{a-k_i}(B) = 1, I_a(A_i) = 1) \frac{p^{(b-a-k_j+1)}(a_{i,k_i}, a_{j,1})}{\pi(a_{j,1})} \chi_j \\ &\leqslant \Xi\, \mathbb{E}Q_a\, \chi. \end{aligned}$$

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

Because of symmetry, if $b \in \{a-2M-s+3,\dots,a-m\}$, then

$$\mathbb{E}(Q_a Q_b) \leqslant \Xi\, \mathbb{E}Q_b\, \chi,$$

so that

$$\sum_{b\in\Gamma_a^s} \mathbb{E}(Q_a Q_b) \leqslant 2\,\Xi\,(s+2M-m-2)\,\mathbb{E}Q_a\,\chi. \tag{2.8}$$

Next, we pay attention to the term $\sum_{a\in\Gamma} n_a$. The r.v. Q_a is measurable with respect to $\sigma(X_{a-2M+2},\dots,X_a)$, and it holds true that

$$\sigma(Q_1,\dots,Q_{a-s-2M+2},Q_{a+s+2M-2},\dots,Q_n) \subseteq \sigma(X_1,\dots,X_{a-s-2M+2},X_{a+s},\dots,X_n),$$

which, in combination with the Markov property and relation (2.2), gives

$$\begin{aligned}
n_a \;\leqslant\; & \sum_{z,w\in\Omega} \big|\mathbb{E}(Q_a|X_{a-s-2M+2}=z,X_{a+s}=w)-\mathbb{E}Q_a\big| \\
& \times \mathbb{P}(X_{a-s-2M+2}=z,X_{a+s}=w) \\
= \;& \sum_{z,w\in\Omega} \big|\mathbb{P}(Q_a=1,X_{a-s-2M+2}=z,X_{a+s}=w) \\
& -\mathbb{E}Q_a\mathbb{P}(X_{a-s-2M+2}=z,X_{a+s}=w)\big| \\
\leqslant \;& \sum_{z,w\in\Omega}\sum_{i=1}^{r}\sum_{B\in\mathscr{B}_i} \big|\mathbb{P}(X_{a-s-2M+2}=z,I_{a-k_i}(B)=1,I_a(A_i)=1,X_{a+s}=w) \\
& -\mathbb{E}I_a(BA_i)\mathbb{P}(X_{a-s-2M+2}=z,X_{a+s}=w)\big| \\
= \;& \sum_{z,w\in\Omega}\sum_{i=1}^{r}\sum_{B\in\mathscr{B}_i} \Big|\pi(z)\frac{p^{(s+M-k_i)}(z,b_1)}{\pi(b_1)}\mathbb{E}I_a(BA_i)p^{(s)}(a_{i,k_i},w) \\
& -\pi(z)\mathbb{E}I_a(BA_i)p^{(2s+2M-2)}(z,w)\Big| \\
= \;& \sum_{z,w\in\Omega}\sum_{i=1}^{r}\sum_{B\in\mathscr{B}_i} \big|\mathbb{E}I_a(BA_i)\big|\, p_R^{(s+M-k_i)}(b_1,z)p^{(s)}(a_{i,k_i},w) \\
& -\pi(z)p^{(2s+2M-2)}(z,w)\big|,
\end{aligned} \tag{2.9}$$

where b_1 is the first letter of the word B. Furthermore, if we set

$$\eta^{(t)}(z,w) := |p^{(t)}(z,w)-\pi(w)| \quad\text{and}\quad \eta_R^{(t)}(z,w) := |p_R^{(t)}(z,w)-\pi(w)|,$$

then

$$\begin{aligned}
&\big|\, p_R^{(s+M-k_i)}(b_1,z)p^{(s)}(a_{i,k_i},w)-\pi(z)p^{(2s+2M-2)}(z,w)\big| \\
&\quad\leqslant\; \eta_R^{(s+M-k_i)}(b_1,z)\eta^{(s)}(a_{i,k_i},w)+\eta_R^{(s+M-k_i)}(b_1,z)\pi(w) \\
&\qquad +\eta^{(s)}(a_{i,k_i},w)\pi(z)+\eta^{(2s+2M-2)}(z,w)\pi(z).
\end{aligned}$$

Combining the former inequality with relation (2.9), we obtain

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

$$
\begin{aligned}
n_a \;\leqslant\; & \sum_{i=1}^{r} \sum_{B\in\mathscr{B}_i} \mathbb{E}I_a(BA_i) \sum_{z,w\in\Omega} \big\{ \eta_R^{(s+M-k_i)}(b_1,z)\eta^{(s)}(a_{i,k_i},w) \\
& + \eta_R^{(s+M-k_i)}(b_1,z)\pi(w) + \eta^{(s)}(a_{i,k_i},w)\pi(z) + \eta^{(2s+2M-2)}(z,w)\pi(z) \big\} \\
= \; & \sum_{i=1}^{r} \sum_{B\in\mathscr{B}_i} \mathbb{E}I_a(BA_i) \Bigg\{ \sum_{z,w\in\Omega} \eta_R^{(s+M-k_i)}(b_1,z)\eta^{(s)}(a_{i,k_i},w) \\
& + \sum_{z\in\Omega} \eta_R^{(s+M-k_i)}(b_1,z) + \sum_{w\in\Omega} \eta^{(s)}(a_{i,k_i},w) + \sum_{z,w\in\Omega} \eta^{(2s+2M-2)}(z,w)\pi(z) \Bigg\}.
\end{aligned}
$$

Using inequality (2.1), we find

$$
\sum_{w\in\Omega} \eta_R^{(t)}(z,w) \leqslant 2\varrho^t \quad \text{and} \quad \sum_{w\in\Omega} \eta^{(t)}(z,w) \leqslant 2\varrho^t, \quad \forall\, z\in\Omega.
$$

Therefore,

$$
n_a \leqslant 2\varrho^s \sum_{i=1}^{r} \sum_{B\in\mathscr{B}_i} \mathbb{E}I_a(BA_i)\big\{2 + 2\varrho^s + \varrho^{2(M-1)+s}\big\},
$$

since $M + s - k_i \geqslant s$. Finally,

$$
n_a \leqslant 2\,\varrho^s\, \mathbb{E}Q_a \left\{2 + 2\varrho^s + \varrho^{2(M-1)+s}\right\}. \tag{2.10}
$$

Relations (2.7), (2.8) and (2.10) give us

$$
\begin{aligned}
d_{TV}(\mathscr{L}(Q), \mathrm{Po}(\mathbb{E}Q)) \;\leqslant\; & \left(1 - e^{-\mathbb{E}Q}\right) \{(4M + 2s - 5)(\mathbb{E}Q_a + \chi\,\Xi) - \Xi\,(2m-1)\,\chi\} \\
& + 2\,\mathbb{E}Q \left(1 \wedge \sqrt{\frac{2}{e\mathbb{E}Q}}\right) \varrho^s \left\{2 + 2\varrho^s + \varrho^{2(M-1)+s}\right\}. \qquad (2.11)
\end{aligned}
$$

Finally, inequality (2.4) in conjunction with (2.5), (2.6) and (2.11) yields the required result. □

Next, we treat the case where the sequence $\{X_a\}_{a\in Z}$ consists of independent and identical r.v.s, which take values in the alphabet $\Omega = \{\omega_1, \ldots, \omega_q\}$ with probabilities $p_1, \ldots, p_q$, respectively. For $a = m, \ldots, n$, relation (2.2) reduces to

$$
\mathbb{E}Q_a = \sum_{i=1}^{r} \chi_i\, U_i,
$$

where $U_i = 1 - \mathbb{P}(\bigcup_{j=1}^{r} \bigcup_{p\in\mathscr{P}'(A_j,A_i)} (I_{a-k_i}(A_j^p) = 1)), \chi_i = \prod_{t=1}^{k_i} \mathbb{P}(a_{i,t})$ and $\mathbb{P}(a_{i,t}) = p_j$ if $a_{i,t} = \omega_j$. The mean $\mathbb{E}N$ is again given by Lemma 2.2, where the quantities e_{ji} now take the form

$$
\begin{aligned}
e_{ji} \;&:=\; \sum_{t=1}^{k_i-1} \varepsilon_t(j,i)\mathbb{P}(a_{i,t+1}) \cdots \mathbb{P}(a_{i,k_i}), \quad i,j = 1,\ldots,r \;\text{ and }\; i \neq j, \\
e_{ii} \;&:=\; 1 + \sum_{t=1}^{k_i-1} \varepsilon_t(i,i)\mathbb{P}(a_{i,t+1}) \cdots \mathbb{P}(a_{i,k_i}), \quad i = 1,\ldots,r.
\end{aligned}
$$

The above relations are again a generalization of the formulas given by Theorem 3.1 of

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

Blom and Thorburn [6] to the case where the words do not have the same length and the letters of the alphabet do not have equal probabilities. We also note that these relations are in agreement with Theorem 3 of Breen, Waterman and Zhang [7] or with formula (15) of Tanushev and Arratia [29].

Now the relevant bound for the i.i.d case is an immediate consequence of Theorem 2.1, if we take $s = 1$, $\Xi = 1$ and $\varrho = 0$.

Corollary 2.3. *Let N be the number of non-overlapping appearances of $\mathscr{E}$ in the i.i.d. sequence $X_1, \ldots, X_n$. Then*

$$d_{TV}(\mathscr{L}(N), \mathrm{Po}(\mathbb{E}N)) \leqslant (1 - e^{-\mathbb{E}Q})\{(4M-3)\mathbb{E}Q_a + 2(2M-m-1)\chi\} + \left\{1 + \left(1 \wedge \frac{1}{\sqrt{\mathbb{E}N}}\right)\right\}(\mathbb{E}N - \mathbb{E}Q).$$

When the trials are independent, the number N of non-overlapping appearances of $\mathscr{E}$ is equal to the number of appearances of $\mathscr{E}$ as *renewal* in the sense of Feller [12]. In Markov chains this fact does not hold generally, except for the case of studying appearances of one word only.

In Table 1, we present bounds for the $d_{TV}(\mathscr{L}(N), \mathrm{Po}(\mathbb{E}N))$ computed according to Theorem 2.1. For the calculations we used specific words with letters from the DNA alphabet $\Omega = \{A, C, G, T\}$, and considered the transition matrix P below, which corresponds to the genome of the bacteriophage *Lambda* (Reinert and Schbath [25]):

$$P = \begin{pmatrix} 0.2994 & 0.2086 & 0.2215 & 0.2705 \\ 0.2830 & 0.2198 & 0.2740 & 0.2232 \\ 0.2540 & 0.2820 & 0.2480 & 0.2160 \\ 0.1813 & 0.2232 & 0.3164 & 0.2791 \end{pmatrix}.$$

The matrix P has stationary distribution $(0.254374, 0.234228, 0.264254, 0.247145)$ and $\varrho = 0.1726$.

We observe that, if the mean $\mathbb{E}N$ is fixed and the lengths of the words increase at a rate slower than n, then the bound vanishes to zero.

3. Convergence to the Poisson distribution

Let $\{X_{n,a}\}_{1 \leqslant a \leqslant n, n \geqslant 1}$ be a triangular array of r.v.s produced by an irreducible, aperiodic stationary Markov chain on the alphabet Ω with transition matrix $P = \{p(z,w)\}_{z,w \in \Omega}$, such that $p_{\max} := \max_{z,w \in \Omega} p(z,w) < 1$ and stationary distribution $\pi = \{\pi_z\}_{z \in \Omega}$. Consider also a sequence of words $\{A_{1,n}, \ldots, A_{r,n}\}_{n \geqslant 1}$ with lengths $\{k_{1,n}, \ldots, k_{r,n}\}_{n \geqslant 1}$, respectively. Set $M_n = \max_{1 \leqslant i \leqslant r} k_{i,n}$, $m_n = \min_{1 \leqslant i \leqslant r} k_{i,n}$ and assume that $\lim_n m_n = +\infty$. Furthermore, let $\mathscr{E}_n$ denote the event of the appearance of a word from the set $\{A_{1,n}, \ldots, A_{r,n}\}$ and let N_n denote the number of non-overlapping appearances of $\mathscr{E}_n$ in the sequence $X_{n,1}, \ldots, X_{n,n}$. We will prove that, under general conditions, the r.v. N_n converges in distribution to a Poisson r.v. (Theorems 3.2 and 3.3).

Before proceeding to the limit results of this work, we present a lemma essential for the proof of Theorem 3.2.

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

Table 1 Numerical values for the estimate $B_{2.1}$ of Theorem 2.1

A_1	A_2	Ξ	s	n	$\mathbb{E}N$	$B_{2.1}$
$ACGAC$	$ACACA$	1.17701	4	321		0.027776
$ACGACG$	$ACACA$	1.17701	4	484		0.022242
$ACGACG$	$ACACAC$	1.11254	5	1316	0.5	0.008124
$ACGACGA$	$ACACAC$	1.17701	5	2211		0.005910
$ACGACGA$	$ACACACA$	1.17701	6	4906		0.002556
$ACGACGAC$	$ACACACA$	1.17701	6	8305		0.001768
$ACGAC$	$ACACA$	1.17701	4	638		0.045211
$ACGACG$	$ACACA$	1.17701	4	964		0.036026
$ACGACG$	$ACACAC$	1.11254	5	2627	1	0.013214
$ACGACGA$	$ACACAC$	1.17701	5	4417		0.009612
$ACGACGA$	$ACACACA$	1.17701	6	9805		0.004123
$ACGACGAC$	$ACACACA$	1.17701	6	16603		0.002857
$ACGAC$	$ACACA$	1.17701	4	1271		0.062596
$ACGACG$	$ACACA$	1.17701	4	1924		0.049597
$ACGACG$	$ACACAC$	1.11254	5	5249	2	0.018354
$ACGACGA$	$ACACAC$	1.17701	5	8829		0.013352
$ACGACGA$	$ACACACA$	1.17701	6	19603		0.005657
$ACGACGAC$	$ACACACA$	1.17701	6	33200		0.003926
$ACGAC$	$ACACA$	1.17701	4	1904		0.070299
$ACGACG$	$ACACA$	1.17701	4	2884		0.055305
$ACGACG$	$ACACAC$	1.11254	5	7871	3	0.020592
$ACGACG$	$ACACAC$	1.17701	5	13241		0.014989
$ACGACGA$	$ACACACA$	1.17701	6	29402		0.006253
$ACGACGAC$	$ACACACA$	1.17701	6	49796		0.004363

We also note that, in what follows, we will insert the index n to the quantities introduced in Section 2 that depend on n.

Lemma 3.1. *If the following assumptions are valid,*

(1) $\lim_n \mathbb{E}N_n = \lambda < +\infty$,
(2) $\lim_n \frac{M_n}{n} = 0$,

then the expected number $(n - m_n + 1)\chi_n$ *of appearances of* $\mathscr{E}_n$ *is bounded, and* $\lim_n \mathbb{E}Q_n = \lambda$.

Proof. We have

$$\begin{aligned}
\chi_n &= \mathbb{P}(I_a(\mathscr{E}_n) = 1, N_{n,a} = 1) + \mathbb{P}(I_a(\mathscr{E}_n) = 1, N_{n,a} = 0) \\
&= \mathbb{E}N_{n,a} + \sum_{i=1}^{r} \mathbb{P}\left(\bigcup_{b=a-k_{i,n}+1}^{a-1} (N_{n,b} = 1),\, I_a(A_{i,n}) = 1 \right) \\
&\leqslant \mathbb{E}N_{n,a} + \sum_{i=1}^{r} \sum_{b=a-k_{i,n}+1}^{a-1} \mathbb{P}(N_{n,b} = 1,\, I_a(A_{i,n}) = 1). \qquad (3.12)
\end{aligned}$$

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

It can be verified that the probability $\mathbb{P}(N_{n,b} = 1, I_a(A_{i,n}) = 1)$ is bounded by

$$\mathbb{E}N_{n,b}\, p_{\max}^{(a-b)},$$

so that summing inequality (3.12) for $a = m_n, \ldots, n$, we get

$$\begin{aligned}(n - m_n + 1)\chi_n &\leqslant \mathbb{E}N_n + \sum_{i=1}^{r}\sum_{a=m_n}^{n}\sum_{b=a-k_{i,n}+1}^{a-1}\mathbb{E}N_{n,b}\, p_{\max}^{(a-b)} \\ &= \mathbb{E}N_n + \sum_{i=1}^{r}\sum_{b=m_n}^{n}\mathbb{E}N_{n,b}\sum_{a=b+1}^{b+k_{i,n}-1} p_{\max}^{(a-b)} \\ &\leqslant \mathbb{E}N_n + r\frac{p_{\max}}{(1-p_{\max})}\mathbb{E}N_n,\end{aligned}$$

which is bounded under the assumption $\lim_n \mathbb{E}N_n = \lambda$. As we have seen in the proof of Theorem 2.1 (see relation (2.6)), it holds true that

$$\mathbb{E}(N_n - Q_n) = \sum_{a=m_n}^{n}\mathbb{P}(N_{n,a} \neq Q_{n,a}).$$

The event $\{N_{n,a} \neq Q_{n,a}\} = \{N_{n,a} = 1, Q_{n,a} = 0\}$ takes place when the following happen:

(i) $\mathscr{E}_n$ occurs at position a and its appearance is non-overlapping,
(ii) another occurrence of $\mathscr{E}_n$ has taken place at a preceding position b, which overlaps the appearance at a, and
(iii) a third appearance of $\mathscr{E}_n$ has occurred at a point c before b, which overlaps the one at position b but it does not overlap the occurrence at a.

Thus

$$\begin{aligned}&\mathbb{P}(N_{n,a} \neq Q_{n,a}) \\ &\quad= \mathbb{P}\left(N_{n,a} = 1, \bigcup_{b=a-M_n+1}^{a-1}\left\{I_b(\mathscr{E}_n) = 1, N_{n,b} = 0, \bigcup_{c=b-M_n+1}^{a-M_n}(N_{n,c} = 1)\right\}\right) \\ &\quad\leqslant \mathbb{P}\left(N_{n,a} = 1, \bigcup_{c=a-2M_n+2}^{a-M_n}(N_{n,c} = 1)\right) \\ &\quad\leqslant \sum_{c=a-2M_n+2}^{a-M_n}\sum_{i=1}^{r}\sum_{j=1}^{r}\chi_{j,n}\frac{p^{(a-k_i+1-c)}(a_{j,k_j}, a_{i,1})}{\pi(a_{i,1})}\chi_{i,n} \\ &\quad\leqslant \Xi_n \sum_{c=a-2M_n+2}^{a-M_n}\chi_n^2 \\ &\quad= \Xi_n(M_n - 1)\,\chi_n^2.\end{aligned}$$

We observe that $\lim_{t\to+\infty}\frac{p^{(t)}(z,w)}{\pi(w)} = 1$, and therefore the sequence $\left\{\frac{p^{(t)}(z,w)}{\pi(w)}\right\}_{t\geqslant 1}$, is bounded for all $z, w \in \Omega$. Hence

$$\Xi_n \leqslant \sup_{t\geqslant 1}\max_{z,w\in\Omega}\frac{p^{(t)}(z,w)}{\pi(w)} < +\infty, \quad \forall\, n.$$

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

Finally, we obtain

$$\mathbb{E}(N_n - Q_n) \leqslant \Xi_n (n - m_n + 1)(M_n - 1)\chi_n^2,$$

which tends to zero under our assumptions. □

Now we are ready to present the main limit result of this work.

Theorem 3.2. *If the following assumptions hold,*

(1) $\lim_n \mathbb{E}N_n = \lambda < +\infty$,
(2) $\lim_n \frac{M_n}{n} = 0$,

then N_n converges in distribution to a Poisson r.v. with parameter λ.

Proof. The proof is based on showing that the distance $d_{TV}(\mathscr{L}(N_n), \mathrm{Po}(\mathbb{E}N_n)) \to 0$, as $n \to +\infty$. Since $d_{TV}(\mathscr{L}(N_n), \mathrm{Po}(\mathbb{E}N_n)) \leqslant B_{2.1}(s)$, for all $s \geqslant 1$ (Theorem 2.1), it suffices to take $s = M_n$ and prove that $B_{2.1}(M_n) \to 0$, as n tends to infinity.

Using Lemma 3.1, it is easy to verify that the bound $B_{2.1}(M_n)$ is of order $O\left(\frac{M_n}{n}\right)$, which leads us immediately to the desired conclusion. □

In the above results, we assumed that the quantity r is a fixed positive integer. If r depends on n ($r = r_n$) in such a way that $\{r_n\}_{n\geqslant 1}$ is a bounded sequence, Lemma 3.1 and Theorem 3.2 are again valid. In the special case where $\lim_n r_n = +\infty$, in order to obtain the analogous results, we need an additional assumption so that $n\chi_n$ stays bounded. Hence, Theorem 3.2 can be restated as follows.

Theorem 3.3. *Under the following conditions,*

(1) $\lim_n \mathbb{E}N_n = \lambda < +\infty$,
(2) $\lim_n \frac{M_n}{n} = 0$
(3) $\log_{1/p_{\max}}(nr_n) = m_n + O(1)$,

the r.v. N_n converges in distribution to a Poisson r.v. with parameter λ.

Acknowledgements

We would like to thank Professor A. D. Barbour for the valuable discussions we had with him, which led us to study the above problem under the Markovian hypothesis. We would also like to thank the referee for thoughtful commentary that led to improvement of the exposition.

References

[1] Arratia, R., Goldstein, L. and Gordon, L. (1989) Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* **17** 9–25.
[2] Arratia, R., Goldstein, L., and Gordon, L. (1990) Poisson approximation and the Chen–Stein method. *Statistical Science* **5** 403–434.

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

[3] Arratia, R., Martin, D., Reinert, G. and Waterman, M. (1996) Poisson process approximation for sequence repeats, and sequencing by hybridization. *J. Comp. Biology* **3** 425–463.
[4] Barbour, A. D., Holst, L. and Janson, S. (1992) *Poisson Approximation*, Oxford University Press.
[5] Biggins, J. D. and Cannings, C. (1987) Markov renewal processes, counters and repeated sequences in Markov chains. *Adv. Appl. Probab.* **19** 521–545.
[6] Blom, G. and Thorburn, D. (1982) How many random digits are required until given sequences are obtained? *J. Appl. Probab.* **19** 518–531.
[7] Breen, S., Waterman, M. S. and Zhang, N. (1985) Renewal theory for several patterns. *J. Appl. Probab.* **22** 228–234.
[8] Chen, L. H. Y. (1975) Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545.
[9] Chryssaphinou, O. and Papastavridis, S. (1988) A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *J. Appl. Probab.* **25** 428–431.
[10] Chryssaphinou, O. and Papastavridis, S. (1990) The occurrence of sequence patterns in repeated dependent experiments. *Th. Probab. Appl.* **35** 145–152.
[11] Chryssaphinou, O., Papastavridis, S. and Vaggelatou, E. (1999) On the number of appearances of a word in a sequence of i.i.d. trials. *Meth. Comp. Appl. Probab.* **1** 329–348.
[12] Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, Vol I, 3rd edn, Wiley, New York.
[13] Fousler, D. E. and Karlin, S. (1987) Maximal success durations for a semi-Markov process. *Stoch. Proc. Appl.* **24** 203–224.
[14] Fu, J. C. (1996) Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Stat. Sinica* **6** 957–974.
[15] Gerber, H. U. and Li, S.-Y. R. (1981) The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stoch. Proc. Appl.* **11** 101–108.
[16] Godbole, A. P. (1991) Poisson approximations for runs and patterns of rare events. *Adv. Appl. Probab.* **23** 851–865.
[17] Godbole, A. P. and Schaffner, A. A. (1993) Improved Poisson approximations for word patterns. *Adv. Appl. Probab.* **25** 334–347.
[18] Guibas, L. and Odlyzko, A. (1981) Periods in strings. *J. Combin. Theory Ser. A* **30** 19–42.
[19] Guibas, L. and Odlyzko, A. (1981) String overlaps, pattern matching, and nontransitive games. *J. Combin. Theory Ser. A* **30** 183–208.
[20] Karlin, S. and Ost, F. (1987) Counts of long aligned word matches among random letter sequences. *Ann. Probab.* **19** 293–351.
[21] Koutras, M. V. (2000) Applications of Markov chains to the distribution theory of runs and patterns. In *Handbook of Statistics* (C. R. Rao and D. N. Shanbhag, eds), Vol. 20 of *Stochastic Processes: Modelling and Simulation.*
[22] Lindvall, T. (1992) *Lectures on the Coupling Method*, Wiley, New York.
[23] Neuhauser, C. (1996) A phase transition for the distribution of matching blocks. *Combin. Probab. Comput.* **5** 139–159.
[24] Régnier, M. (2000) A unified approach to word occurrence probabilities. *D. Appl. Math.* **104** 259–280.
[25] Reinert, G. and Schbath, S. (1998) Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biology* **5** 223–253.
[26] Robin, S. and Daudin, J. J. (1999) Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.* **36** 179–193.
[27] Schbath, S. (1995) Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics* **1** 1–16.
[28] Schwager, S. J. (1983) Run probabilities in sequences of Markov dependent trials. *J. Amer. Statist. Assoc.* **78** 168–175.

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press

[29] Tanushev, M. and Arratia, R. (1997) Central limit theorem for renewal theory for several patterns. *J. Comp. Biology* **4** 35–44.
[30] Waterman, M. S. (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman and Hall, London.

https://doi.org/10.1017/S096354830100476X Published online by Cambridge University Press