

# On the length of the longest exact position match in a Markov sequence <sup>\*</sup> <sup>†</sup>

G. Reinert<sup>‡</sup> and Michael S. Waterman<sup>§</sup>

## Abstract

A mixed Poisson approximation and a Poisson approximation for the length of the longest exact match of a Markov sequence across another sequence are provided, where the match is required to start at position 1 in the first sequence. This problem arises when looking for suitable anchors in whole genome alignments.

---

<sup>\*</sup>AMS 2000 subject classifications. 62E17, 92D20.

<sup>†</sup>Key words and phrases: Poisson approximation, mixed Poisson approximation, length of longest match, Markov chain, Chen-Stein method

<sup>‡</sup>Corresponding author. Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK; e-mail *reinert@stats.ox.ac.uk*; phone 0044 1865 288251; fax ++44 1865 272595. GR was supported in part by MMCOMNET grant no. FP6-2003-BEST-Path-012999.

<sup>§</sup>Molecular and Computational Biology, University of Southern California, 835 W 37th Street, SHS 172, Los Angeles, California 90089-1340, USA; e-mail *msw@usc.edu*; phone 001 213 7407439; fax ++1 213 7402437. MSW was supported by NIH grant no. P50 HG 002790.

# 1 Introduction

When aligning whole genomes, often a seed-and-extend technique is used. Starting from exact or near-exact matches, reliable ones among these matches are selected as anchors, and then the remaining stretches are filled in using local and global alignment. See *Lippert et al.* [1] for a discussion of genome alignment methods using anchors. To select a match that is both sensitive and specific, [1] introduce a score based on the length,  $R_n$ , of the longest exact match of a random sequence across another sequence, where shifts are not allowed. For  $R_n$  and the associated scores, [1] find that their approach based on a mixed Poisson approximation, although valid, is computationally not feasible if the distribution of the random letters making up the random sequences is not uniform, as the mixing takes place over too many terms; the authors resort to a Monte Carlo method. Here we provide a Poisson approximation for the number of matches of fixed length, along with bounds provided by the Chen-Stein method, and we obtain an approximate expression for the cumulative distribution function of  $R_n$  that is easy to compute. The bound on the error in the approximation turns out to be small, thus making our suggestion a useful approach.

In [1] an i.i.d. model is used as a null model; [2] derive a Poisson approximation for the length of the longest exact position match in an i.i.d. sequence. Here we extend the results of [2] to a Markov sequence; most of the main ideas can also be found in [2]. The set-up for our problem is as follows. Let  $\mathbf{A} = A_1A_2 \dots A_n$  and  $\mathbf{B} = B_1B_2 \dots B_n$  be two independent sequences with letters from a finite alphabet  $\mathcal{A}$  with  $d$  elements. As in [6], for example, we assume that  $\mathbf{A}$  is part of an infinite sequence  $\dots, A_{-1}, A_0, A_1, A_2, \dots$ , generated by a stationary first-order Markov chain with transition matrix  $\Pi = (\pi(a, b))_{a, b \in \mathcal{A}}$ ; similarly,  $\mathbf{B}$

is part of an infinite sequence  $\dots, B_{-1}, B_0, B_1, B_2, \dots$ , generated by a stationary first-order Markov chain with the same transition matrix  $\Pi$ . We assume that  $\pi(a, b) > 0$  for all  $a$  and  $b$ , and that

$$\rho = \max_{a, b \in \mathcal{A}} \pi(a, b) < 1. \quad (1)$$

Denote by  $\mu$  the unique stationary distribution for the chain, and by  $\pi^{(\ell)}(a, b)$  the  $\ell$ -step transition probability between  $a$  and  $b$ . Let  $\mu^* = \max_{a \in \mathcal{A}} \mu(a)$  be the maximum of the stationary probabilities. We put

$$R_n = \max_m \{A_k = B_{j+k}, k = 1, \dots, m, \text{ for some } 0 \leq j \leq n - m\};$$

thus  $R_n$  denotes the length of the longest exact match of a random sequence across another sequence, where shifts are not allowed.

Note that if the match in sequence  $\mathbf{A}$  was not required to start at position 1, the problem would reduce to the distribution of the well understood

$$H_n = \max_m \{A_{i+k} = B_{j+k}, k = 1, \dots, m, \text{ for some } 0 \leq i, j \leq n - m\},$$

see for example *Waterman* [7]. Our problem differs from the study of  $H_n$  by requiring an exact match beginning at a fixed position in the first sequence.

To reveal the Poisson-type structure in the problem, we use a standard duality argument as follows. If  $R_n < m$  then there are no matches of length  $m$  (or longer) in the sequence. Ignoring end effects, this means that there are no occurrences of  $A_1 \dots A_m$  in  $\mathbf{B}$ . Let  $W_m$  denote the number of (clumps of) matches of length  $m$  (or longer) in the sequence, so that  $P(R_n < m) \approx P(W_m = 0)$ .

In Section 2 we shall give a mixed Poisson approximation for  $P(W_m = 0)$ . Section 3 derives the Poisson approximation for  $P(W_m = 0)$  and applies it to obtain an approximation, with bound, for  $P(R_n < m)$ . We shall also give a numerical example.

## 2 A mixed Poisson approximation

For Poisson and mixed Poisson approximation it is useful to think in terms of clumps of occurrences, see [4] or [3], because de-clumping disentangles the dependence arising from self-overlap of words. We say that a *clump* of a word  $\omega = \omega_1\omega_2 \dots \omega_m$  starts at position  $i$  in  $\mathbf{B}$  if there is an occurrence of  $\omega$  at position  $i$ , and there is no (overlapping) occurrence of  $\omega$  at positions  $i - m + 1, \dots, i - 1$ .

Thus when ignoring end effects the study of  $R_n$  is equivalent to the study of

$$W_m = \sum_{i=1}^{\bar{n}} \mathbf{1}(\text{a clump of } A_1 \dots A_m \text{ starts at position } i \text{ in } \mathbf{B}),$$

where we abbreviate  $\bar{n} = n - m + 1$ . End effects only arise from the possibility that, when embedded in an infinite sequence, the sequence  $\mathbf{B} = B_1B_2 \dots B_n$  starts within a clump in the infinite sequence.

Assume that  $\mathbf{B}_\infty = \dots B_{-1}B_0B_1 \dots B_nB_{n+1} \dots$  is an infinite sequence for now, so that we can ignore end effects. Then we have

$$R_n < m \iff W_m = 0.$$

If  $m$  is large enough, then a fixed word  $\omega$  of length  $m$  will rarely occur at a given position  $i$  in the random sequence  $\mathbf{B}$ . When using clumps in order to account for the strong dependence between neighbouring occurrences in the case that  $\omega$  has a large amount of self-overlap, it

is plausible and indeed established that the number of clumps of  $\omega$  in  $\mathbf{B}$  is approximately Poisson distributed, Proposition 1 below. For any fixed  $\omega$ , we let

$$W_m(\omega) = \sum_{i=1}^{\bar{n}} \mathbf{1}(\text{a clump of } \omega \text{ starts at position } i \text{ in } \mathbf{B}).$$

In what follows we shall always assume that  $\omega = w_1 \cdots w_m \in \mathcal{A}^m$ , so that

$$\mu(\omega) = \mu(w_1) \prod_{i=1}^{m-1} \pi(w_i, w_{i+1})$$

is the probability of a random word of length  $m$  equals  $\omega$ . If there is a  $p$  such that  $w_i = w_{i+p}$ ,  $i = 1, \dots, m - p$ , then  $p$  is called a *period* of  $\omega$ . A period is a *principal* period if it is not a strict multiple of the minimal period. An occurrence of  $\omega$  starting at position  $i$  is a clump if and only if for none of the periods  $p$  of  $\omega$ , the truncated word  $\omega^{(p)} = w_1 \cdots w_p$  starts at position  $i - p$ . It is easy to see that it suffices to consider all principal periods. The probability that a clump of  $\omega$  starts at a given position in the sequence is then given by

$$\tilde{\mu}(w) = \mu(w) - \sum_{p \in \mathcal{P}'(\omega)} \mu(w^{(p)}w), \quad (2)$$

where  $\omega^{(p)}\omega = w_1 \cdots w_p w_1 \cdots w_m$  is the concatenated word, and  $\mathcal{P}'(\omega)$  is the set of principal periods of  $\omega$ . In particular,

$$EW_m(\omega) = \tilde{\lambda}(\omega) := \bar{n} \tilde{\mu}(\omega).$$

To describe the distance between the distributions of non-negative integer valued random variables  $X$  and  $Y$  we use the total variation distance, defined by

$$d_{TV}(X, Y) = \sup_{B \subset \{0, 1, \dots\}} |P(X \in B) - P(Y \in B)|.$$

We shall need some more notation, see [3]. For a 1-order Markov chain we diagonalize the transition matrix as follows. Let  $(\alpha_t)_{t=1, \dots, d}$  be the eigenvalues of  $\Pi$  such that  $|\alpha_1| \geq$

$|\alpha_2| \geq \dots \geq |\alpha_d|$ . It follows from the Perron-Frobenius Theorem that  $|\alpha_2| < 1$ ; we put

$$\alpha := \alpha_2. \quad (3)$$

Let  $D = \text{Diag}(1, \alpha, \alpha_3, \dots, \alpha_d)$ . We decompose  $\Pi = PDP^{-1}$  such that the first column of  $P$  is  $(1, 1, \dots, 1)^T$ ; then the first row of  $P^{-1}$  is the vector of stationary distribution  $(\mu(a), a \in \mathcal{A})$ .

For all  $t \in \{1, \dots, d\}$ ,  $I_t$  denotes the  $d \times d$  matrix such that all its entries are equal to 0 except  $I_t(t, t) = 1$ , and we define

$$Q_t := PI_tP^{-1}. \quad (4)$$

Then we may decompose the  $\ell$ -step transition matrix  $\Pi^\ell$  as

$$\Pi^\ell = \sum_{t=1}^d \alpha_t^\ell Q_t. \quad (5)$$

Furthermore we put

$$\gamma(m) = \max_{a, b \in \mathcal{A}} \sum_{x, y \in \mathcal{A}} \mu(x) \left| \frac{1}{\mu(b)} \sum_{(t, t') \neq (1, 1)} \frac{\alpha_t^m \alpha_{t'}^m}{\alpha^m} Q_t(x, b) Q_{t'}(a, y) - \sum_{t=2}^d \frac{\alpha_t^{4m-2}}{\alpha^m} Q_t(x, y) \right|.$$

Corollary 6.4.6. in [3] immediately gives the following proposition (see also [5] for the original source).

**Proposition 1** *Let  $\tilde{Z}(\omega) \sim Po(\tilde{\lambda}(\omega))$  be Poisson distributed with mean  $\tilde{\lambda}(\omega)$ . Then*

$$\begin{aligned} d_{TV}(\mathcal{L}(W_m(\omega)), Po(\tilde{\lambda}(\omega))) &\leq \bar{n} \tilde{\mu}(\omega) \left\{ (6m - 5) \tilde{\mu}(\omega) + \gamma(m) |\alpha|^m \right. \\ &\quad \left. + \frac{2}{\mu(w_1)} \mu(\omega) \sum_{s=1}^{2m-2} \Pi^s(w_m, w_1) \right\} \\ &\quad + (m - 1)(\mu(\omega) - \tilde{\mu}(\omega)). \end{aligned}$$

Proposition 1 only counts the number of occurrences of a fixed word, whereas in our problem, the first  $m$  letters of the sequence  $\mathbf{A}$ , namely  $A_1 \dots A_m$ , constitute a random word.

Thus we need to condition on the words  $\omega$  that  $A_1 \dots A_m$  take on, and using the rule of total probability, we obtain a mixed Poisson approximation.

**Theorem 1** *Assume that*

$$0 < \mu_* = \min_{a \in \mathcal{A}} \mu(a) \leq \mu^* < 1.$$

*With the above notation,*

$$\begin{aligned} & |P(W_m = 0) - \sum_{\omega} \mu(\omega) P(Po(\tilde{\lambda}(\omega)) = 0)| \\ & \leq \bar{n} \mu^* \rho^{m-1} \left\{ \left( (6m - 5) + \frac{2\rho}{\mu_*(1 - \rho)} \right) \mu^* \rho^{m-1} + \gamma(m) |\alpha|^m \right\} + (m - 1) Rem_0 \\ & =: Rem_1. \end{aligned}$$

*Here,  $Rem_0$  is given in Lemma 1 below.*

**Remark 1** 1. *A Poisson approximation should be valid in the regime that  $\lambda(\omega) \asymp 1$ .*

*If  $\bar{n} \mu^* \rho^{m-1} = O(1)$  then the relevant regime for our approximations is  $\mu^* \rho^{m-1} = O(n^{-1})$ ; and if  $m$  is at least as large as  $\log_{|\alpha|^{-1}} n$ , then  $R_0$ ,  $R_1$  and  $R_2$  will be of order  $O(n^{-1} \log n)$ .*

2. *In the i.i.d. case, [2] obtain the simpler bound*

$$|P(W_m = 0) - \sum_{\omega} \mu(\omega) P(Po(\tilde{\lambda}(\omega)) = 0)| \leq Rem_1 \leq (8m - 7) \bar{n} \pi_3^m.$$

For the proof of Theorem 1, we shall employ the following lemma.

**Lemma 1** *With  $\rho$  given in (1), we have that*

$$\sum_{\omega} \mu(\omega) (\mu(\omega) - \tilde{\mu}(\omega)) \leq Rem_0,$$

where for  $\rho \neq \frac{1}{d}$ ,

$$Rem_0 \leq \mu^* \rho^{m-1} \frac{(d\rho)^m - 1}{d\rho - 1},$$

and for  $\rho = \frac{1}{d}$ ,

$$Rem_0 \leq (m-1)\mu^* \rho^{m-1}.$$

In general,  $\rho d \geq 1$ . However, if the letter distribution is close to uniform, and if  $m$  is relatively large, then  $\rho^2 d < 1$ , and the above bound will be small.

We note that  $\rho = \frac{1}{d}$  implies that the maximal transition probability is  $\frac{1}{d}$ . As there for each starting point there are  $d$  possible transitions, their probabilities summing to 1, it follows that all transitions are equally likely;  $\rho = \frac{1}{d}$  thus corresponds to the i.i.d. case.

**Proof of Lemma 1.** From (2),

$$\mu(\omega) - \tilde{\mu}(\omega) = \sum_{p \in \mathcal{P}'(\omega)} \mu(\omega^{(p)}\omega).$$

To bound the sum  $\sum_{p \in \mathcal{P}'(\omega)} \mu(\omega^{(p)}\omega)$  we consider the cases that  $p \leq \lfloor \frac{m}{2} \rfloor$  and  $p \geq \lfloor \frac{m}{2} \rfloor + 1$  separately.

For  $p \geq \lfloor \frac{m}{2} \rfloor + 1$  we note that  $2p + 1 \geq m$ , and writing out the period yields

$$\begin{aligned} & \sum_{\omega} \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(\omega^{(p)}\omega) \\ &= \sum_{\omega} \mathbf{1}(p \in \mathcal{P}'(\omega)) P(A_1 \dots A_{m+p} = \omega^{(p)}\omega) \\ &= \sum_{\omega} \mathbf{1}(p \in \mathcal{P}'(\omega)) P(A_1 \dots A_{m+p} = \omega^{(p)}\omega; A_1 = A_{p+1} = A_{2p+1}, \dots, \\ & \quad A_{m-p} = A_m = A_{m+p}; A_{m-p+1} = A_{m+1}, \dots, A_p = A_{2p}) \\ &\leq P(A_{\ell} = A_{\ell+p} = A_{\ell+2p}, \ell = 1, \dots, m-p; A_{\ell} = A_{\ell+p}, \ell = m-p+1, \dots, p). \end{aligned}$$

This probability involves  $m - p$  possibly distinct letters which occur three times each, and  $2p - m$  possibly distinct letters which occur twice each. So in total there are  $m - p + 2p - m = p$  possibly distinct letters in each word, leading to the total number of possible words to be bounded by  $d^p$ . As each sequence of length  $m + p$  has probability at most  $\mu^* \rho^{m+p-1}$ , we may bound

$$\sum_{\omega} \sum_{p=\lfloor \frac{m}{2} \rfloor + 1}^{m-1} \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(w^{(p)} w) \leq \sum_{p=1}^{\lfloor \frac{m}{2} \rfloor} d^p \mu^* \rho^{m+p-1}.$$

For  $p \leq \lfloor \frac{m}{2} \rfloor$  we note that if a word  $\omega$  has period  $p \leq \lfloor \frac{m}{2} \rfloor$ , then the letters  $w_{p+1}, \dots, w_m$  are uniquely determined. Therefore any word can possess at most one principal period  $p \leq \lfloor \frac{m}{2} \rfloor$ , and can involve at most  $p$  distinct letters. Using again the bound  $\mu(w^{(p)} w) \leq \mu^* \rho^{m+p-1}$ , we bound

$$\sum_{p=1}^{\lfloor \frac{m}{2} \rfloor} \sum_{\omega} \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(w^{(p)} w) \leq \sum_{p=1}^{\lfloor \frac{m}{2} \rfloor} d^p \mu^* (\rho)^{m+p-1}.$$

Combining both cases we obtain that

$$\sum_{p=1}^{m-1} \sum_{\omega} \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(w^{(p)} w) \leq \sum_{p=1}^{m-1} d^p \mu^* (\rho)^{m+p-1}.$$

Application of the geometric series now gives the result.

Now we proceed to the proof of Theorem 1.

**Proof of Theorem 1.** Writing out the different sequences that  $A_1 A_2 \dots A_m$  can take on, we have

$$\begin{aligned} P(W_m = 0) &= \sum_{\omega} \mu(\omega) P(W_m(\omega) = 0) \\ &= \sum_{\omega} \mu(\omega) P(Po(\tilde{\lambda}(\omega)) = 0) + \sum_{\omega} \mu(\omega) \epsilon_1(\omega), \end{aligned}$$

where, by Proposition 1,

$$\begin{aligned}
\left| \sum_{\omega} \mu(\omega) \epsilon_{1\omega} \right| &\leq \bar{n} \sum_{\omega} \mu(\omega) \tilde{\mu}(\omega) \left\{ (6m-5) \tilde{\mu}(\omega) + \gamma(m) |\alpha|^m \right. \\
&\quad \left. + \frac{2}{\mu_*} \mu(\omega) \sum_{s=1}^{2m-2} \Pi^s(w_m, w_1) \right\} \\
&\quad + (m-1) \sum_{\omega} \mu(\omega) (\mu(\omega) - \tilde{\mu}(\omega)) \\
&=: Rem_1.
\end{aligned}$$

For  $Rem_1$  we employ Lemma 1 and the crude bound  $\mu(\omega) \leq \mu^* \rho^{m-1}$  to bound

$$\begin{aligned}
Rem_1 &\leq \bar{n} \sum_{\omega} \mu(\omega) \tilde{\mu}(\omega) \left\{ (6m-5) \tilde{\mu}(\omega) + \gamma(m) |\alpha|^m + \frac{2}{\mu_*} \mu(\omega) \sum_{s=1}^{2m-2} \rho^s \right\} + (m-1) Rem_0 \\
&\leq \bar{n} \left( (6m-5) + \frac{2\rho}{\mu_*(1-\rho)} \right) \sum_{\omega} (\mu(\omega))^3 + \bar{n} \mu^* \gamma(m) |\alpha|^m \rho^{m-1} + (m-1) Rem_0 \\
&\leq \bar{n} \mu^* \rho^{m-1} \left\{ \left( (6m-5) + \frac{2\rho}{\mu_*(1-\rho)} \right) \mu^* \rho^{m-1} + \gamma(m) |\alpha|^m \right\} + (m-1) Rem_0.
\end{aligned}$$

This finishes the proof.

### 3 Poisson approximation to the mixed Poisson approximation

Although Theorem 1 is valid, the probability  $\sum_{\omega} \mu(\omega) P(Po(\tilde{\lambda}(\omega)) = 0)$  is difficult to evaluate, the sum growing exponentially with alphabet size. As much of the computational difficulty lies in accounting for the different periods in all words  $\omega \in \mathcal{A}^m$ , our idea is to approximate  $P(Po(\tilde{\lambda}(\omega)) = 0)$  by the simpler expression  $P(Po(\lambda(\omega)) = 0)$ , where

$$\lambda(\omega) := \bar{n} \mu(\omega).$$

Thus we ignore the period correction in the Poisson parameter. While this may much distort the limiting distribution for words  $\omega$  with a large amount of self-overlap, there are not too

many such words in  $\mathcal{A}^m$ ; indeed we provide a bound on the error in this approximation in the next theorem.

**Theorem 2** *For  $\omega \in \mathcal{A}^m$ , let  $\tilde{Z}(\omega)$  have Poisson distribution with mean  $\tilde{\lambda}(\omega)$ , and let  $Z(\omega)$  have Poisson distribution with mean  $\lambda(\omega)$ . Then*

$$\left| \sum_{\omega} \mu(\omega) P(\tilde{Z}(\omega) = 0) - \sum_{\omega} \mu(\omega) P(Z(\omega) = 0) \right| \leq (1 - e^{-\bar{n}\mu^* \rho^{m-1}}) \text{Rem}_0,$$

with  $\text{Rem}_0$  given in Lemma 1.

**Proof of Theorem 2.** Firstly we wish to bound

$$|P(\tilde{Z}(\omega) = 0) - P(Z(\omega) = 0)| = |e^{-\tilde{\lambda}(\omega)} - e^{-\lambda(\omega)}|.$$

By series expansion it is easy to see that for any  $0 \leq \nu \leq \lambda < \infty$ ,

$$\lambda(e^{\lambda-\nu} - 1) \leq (\lambda - \nu)(e^{\lambda} - 1)$$

and direct manipulation yields

$$e^{-\nu} - e^{-\lambda} \leq (\lambda - \nu) \frac{1 - e^{-\lambda}}{\lambda}.$$

Applying this bound with  $\lambda = \lambda(\omega)$  and  $\nu = \tilde{\lambda}(\omega)$ , we have for each fixed  $\omega$  that

$$|P(\tilde{Z}(\omega) = 0) - P(Z(\omega) = 0)| \leq \frac{1 - e^{-\lambda(\omega)}}{\lambda(\omega)} \bar{n} \sum_{p \in \mathcal{P}'(\omega)} \mu(w^{(p)} w).$$

Thus

$$\left| \sum_{\omega} \mu(\omega) P(\tilde{Z}(\omega) = 0) - \sum_{\omega} \mu(\omega) P(Z(\omega) = 0) \right| = \text{Rem}_2,$$

where

$$\begin{aligned} |\text{Rem}_2| &\leq \sum_{\omega} \mu(\omega) \frac{1 - e^{-\lambda(\omega)}}{\lambda(\omega)} \bar{n} \sum_{p \in \mathcal{P}'(\omega)} \mu(w^{(p)} w) \\ &\leq (1 - e^{-\bar{n}\mu^* \rho^{m-1}}) \sum_{p=1}^{m-1} \sum_{\omega} \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(w^{(p)} w). \end{aligned} \quad (6)$$

In the last step we used the uniform bound  $\mu(\omega) \leq \mu^* \rho^{m-1}$  for all  $\omega$ . Apply Lemma 1 to obtain the result.

Now we apply our results to the original problem, the cumulative distribution function of  $R_n$ , the length of the longest exact position match.

**Corollary 1** *For  $\omega \in \mathcal{A}^m$ , as in Theorem 2 let  $Z(\omega)$  have Poisson distribution with mean  $\lambda(\omega)$ . Then*

$$|P(R_n < m) - \sum_{\omega \in \mathcal{A}^m} P(Z(\omega) = 0)| \leq Rem_3,$$

where

$$Rem_3 = Rem_1 + \left\{ (m-1)\mu^* \rho^{m-1} + \left(1 - e^{-\bar{n}\mu^* \rho^{m-1}}\right) \right\} Rem_0,$$

with  $Rem_1$  given in Theorem 1 and  $Rem_0$  given in Lemma 1.

**Proof of Corollary 1.** In view of Theorem 1 and Theorem 2, all that is required is to bound the end effects, resulting from  $\mathbf{B}$  having been idealized as just a part of an infinite sequence when it came to counting clumps. To bound the end effects, note that (see, e.g. [3], Equation (6.4.10) )

$$\begin{aligned} P\{\mathbf{1}(R_n > m) \neq \mathbf{1}(W_m = 0)\} &\leq (m-1) \sum_{\omega} \mu(\omega)(\mu(\omega) - \tilde{\mu}(\omega)) \\ &\leq (m-1)\mu^* \rho^{m-1} \sum_{\omega} \sum_p \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(\omega^{(p)}\omega) \\ &\leq (m-1)\mu^* \rho^{m-1} Rem_0. \end{aligned}$$

Applying Theorem 1 and Theorem 2 finishes the proof.

In the i.i.d. case, [2] obtain a stronger theorem, making use of the combinatorics from requiring matches in independent sequences.

**Theorem 3** Let  $\pi^* = \max_{a \in \mathcal{A}} \pi(a)$ . For  $\omega \in \mathcal{A}^m$ , let  $\tilde{Z}(\omega)$  have Poisson distribution with mean  $\tilde{\lambda}(\omega)$ , and let  $Z(\omega)$  have Poisson distribution with mean  $\lambda(\omega)$ . Abbreviate  $f = \frac{\pi_2^2}{\pi_3}$ .

Then

$$\left| \sum_{\omega} \mu(\omega) P(\tilde{Z}(\omega) = 0) - \sum_{\omega} \mu(\omega) P(Z(\omega) = 0) \right| = \text{Rem}_2,$$

where

$$\text{Rem}_2 \leq (1 - e^{-\bar{n}(\pi^*)^m}) \left\{ (\pi_2)^m \frac{f^{-m + \lfloor \frac{m}{2} \rfloor + 1}}{1 - f} + \lfloor \frac{m}{2} \rfloor (\pi^*)^m \right\}.$$

Here,  $\lfloor x \rfloor$  denotes the integer part of  $x$ .

**Example (from [2]).** Suppose as in [1] that the sequences are i.i.d., and that  $n = 5.74 \times 10^9$ , the estimated length of the human genome, NCBI build 28 and build 34, with alphabet  $\mathcal{A} = \{A, C, G, T\}$  of size  $d = 4$ , and base-composition in a non-repeat region estimated as  $p_A = p_T = 0.29$  and  $p_C = p_G = 0.21$ , so that  $\pi^* = 0.29$  and  $p = 2p_A = 0.58$ . Then, truncating after the first four digits,  $\pi_2 = 0.2564$ ,  $\pi_2^2 = 0.0657$ ,  $\pi_3 = 0.0673$ ,  $f = 0.9768$ , and we can calculate the mean of  $\lambda(A_1 A_2 \cdots A_m)$  using that

$$E\lambda(A_1 A_2 \cdots A_m) = \bar{n} \sum_{\omega} \mu(\omega)^2 = \bar{n} 2^{-m} \sum_{k=0}^m \binom{m}{k} p^{2k} (1-p)^{2(m-k)}.$$

Table 1 gives the expected Poisson parameter for  $m = 15, \dots, 22$ .

$m$	15	16	17	18	19	20	21	22
$E\lambda$	7.8105	2.0026	0.5134	0.1316	0.0337	0.0086	0.0022	0.0005

Table 1: Expected Poisson parameter

Thus for  $m = 15$  we would expect  $W_{15} = 0$ , and hence  $R_n < 15$ , with low probability, whereas for  $m = 22$  we would expect  $W_{22} = 0$  with high probability, hence  $R_n < 22$  with high probability.

Table 2 below gives a summary of the estimated probability  $\rho(m) = \sum_{\omega \in \mathcal{A}^m} P(Z(\omega) = 0)$  for  $P(R_n \geq m) \approx 1 - P(W_m = 0)$  obtained in Corollary 1, for  $m = 15, 16, \dots, 22$ , along with the Monte-Carlo estimates  $\hat{\rho}(m)$  from [1]; we note that Table 8 in [1] indeed gives estimates for  $P(R_n \geq m)$  instead of  $P(R_n < m)$  as written *ibid*. We add our bound from Corollary 1 along with the estimated standard deviation  $\sqrt{\text{Var}\hat{\rho}(m)}$  from [1], and the separate remainder terms contributing to our bound; recall that our bound for  $Rem_2$  is given in (6).

$m$	$\rho(m)$	$\hat{\rho}(m)$	bound	$\sqrt{\text{Var}\hat{\rho}(m)}$	$Rem_1$	$Rem_2$
15	0.981	0.977	1.83 e-06	9.46 e-05	1.70 e-06	1.29 e-07
16	0.772	0.787	1.60 e-07	3.14 e-04	1.23 e-07	3.77 e-08
17	0.369	0.410	1.92 e-08	3.60 e-04	8.82 e-09	1.03 e-08
18	0.119	0.144	2.79 e-09	1.78 e-04	6.30 e-10	2.16 e-09
19	0.0328	0.0414	2.99 e-10	6.29 e-05	4.49 e-11	2.54 e-10
20	0.00859	0.0111	2.80 e-11	1.76 e-05	3.19 e-12	2.48 e-11
21	0.00221	0.00289	2.32 e-12	4.86 e-06	2.25 e-13	2.10 e-12
22	0.000568	0.000753	2.01 e-13	1.34 e-06	1.59 e-14	1.85 e-13

Table 2: Estimated probabilities, bounds, and remainder terms

Our approximated probabilities are similar to the Monte-Carlo estimates in [1]. However, whereas [1] can only conclude that, say, an approximate 95% confidence interval for the true

probability  $P(R_n \geq m)$  is given by  $\hat{\rho} \pm 1.96\sqrt{\text{Var}\hat{\rho}(m)}$ , we indeed proved that the true probability will lie within  $\rho(m) \pm \text{bound}$ , which is a shorter interval for all values of  $m$  considered in this example.

Also we see that both remainder terms  $Rem_1$  and  $Rem_2$  contribute in similar magnitude to the bound  $Rem_3$ , indicating that the bound on the error made in replacing the mixed Poisson approximation by the Poisson approximation is not much larger than the bound on the error made by the mixed Poisson approximation in the first place.

## References

- [1] LIPPERT, R.A., ZHAO, X., FLOREA, L., MOBARRY, C., AND ISTRAIL, S. (2004). Finding Anchors for Genomic Sequence Comparison. In *Proceedings of the 8th Annual International Conference on Research in Computational Biology (RECOMB '04)*, ACM Press, 233–241. Also in *J. Comp. Biol.* **12**, 762–776 (2005).
- [2] REINERT, G. AND WATERMAN, M.S. (2007). On the length of the longest exact position match in a random sequence. *Transactions on Computational Biology and Bioinformatics* **4**(1), 2007, 153-156.
- [3] REINERT, G., SCHBATH, S., AND WATERMAN, M.S. (2005). Statistics on words with applications to biological sequences. In *Lothaire: Applied Combinatorics on Words* J. Berstel and D. Perrin, eds., Cambridge University Press, 251–328.
- [4] ROBIN, S., RODOLPHE, F., AND SCHBATH, S. (2005). *DNA, Words and Models. Statistics of Exceptional Words*. Cambridge University Press.

- [5] SCHBATH, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*. **1** 1–16.
- [6] TOUYAR, N., SCHBATH, S., CELLIER, D. AND DAUCHEL, H. (2008). Poisson approximation for the number of repeats in a Markov chain model. *Submitted*.
- [7] WATERMAN, M.S. (1995). *Introduction to Computational Biology*. Chapman and Hall.