On the length of the longest exact position match in a Markov sequence

Gesine Reinert[1] and Michael S. Waterman[2]

[1] Department of Statistics, University of Oxford,
1 South Parks Road, Oxford OX1 3TG, UK
(e-mail: reinert@stats.ox.ac.uk)

[2] 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)

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.
Keywords. Poisson approximation, mixed Poisson approximation, length of longest match, Markov sequence Chen-Stein method.

1 Introduction

hen 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. (2004)* for a discussion of genome alignment methods using anchors. To select a match that is both sensitive and specific, *Lippert et al. (2004)* 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, *Lippert et al. (2004)* 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 *Lippert et al. (2004)* an i.i.d. model is used as a null model; *Reinert and Waterman (2007)* 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 *Reinert and Waterman (2007)* to a Markov sequence; most of the main ideas can also be found in *Reinert and Waterman (2007)* . The set-up for our problem is as follows. Let $\mathbf{A} = A_1 A_2 \dots A_n$ and $\mathbf{B} = B_1 B_2 \dots B_n$ be two independent sequences with letters from a finite alphabet $\mathcal{A}$ with d elements. As in *Touyar et al. (2008)*, 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 $\Pi = (\pi(a,b))_{a,b,\in\mathcal{A}}$. We assume that $\pi(a,b) > 0$ for all a and b, and that

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

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

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

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

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

see for example *Waterman* [6]. 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 this paper we shall first give a mixed Poisson approximation for $P(W_m = 0)$, then derives the Poisson approximation for $P(W_m = 0)$, and finally apply it to obtain an approximation, with bound, for $P(R_n < m)$.

2 A mixed Poisson approximation

For Poisson and mixed Poisson approximation it is useful to think in terms of clumps of occurrences, see *Robin et al. (2005)* or *Reinert et al. (2005)*, 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 ω at position i, and there is no (overlapping) occurrence of ω at positions $i-m+1, \dots, i-1$.

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

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

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

Assume that $\mathbf{B}_\infty = \cdots B_{-1}B_0B_1\cdots B_nB_{n+1}\cdots$ 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 ω 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 ω has a large amount of self-overlap, it is plausible and indeed established that the number of clumps of ω in $\mathbf{B}$ is approximately Poisson distributed, Proposition 1 below. For any fixed ω, 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 ω. If there is a p such that $w_i = w_{i+p}, i = 1, \ldots, m-p$, then p is called a *period* of ω. A period is a *principal* period if it is not a strict multiple of the minimal period. An occurrence of ω starting at postion i is a clump if and only if for none of the periods p of ω, 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 ω starts at a given position in the sequence is then given by

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

where $\omega^{(p)}\omega = w_1\cdots w_pw_1\cdots w_m$ is the concatendequationted word, and $\mathcal{P}'(\omega)$ is the set of principal periods of ω. In particular,

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

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

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

We shall need some more notation, see *Reinert et al. (2005)*. For a 1-order Markov chain we diagonalize the transition matrix as follows. Let $(\alpha_t)_{t=1,\dots,d}$ be the eigenvalues of Π such that $|\alpha_1| \geq |\alpha_2| \geq \cdots \geq |\alpha_d|$.

$$\alpha := \alpha_2 < 1.$$

Let $D = \mathrm{Diag}(1, \alpha, \alpha_3, \cdots, \alpha_d)$. 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 $Q_t := PI_tP^{-1}$. Then we may decompose the ℓ-step transition matrix Π^ℓ as

$$\Pi^\ell = \sum_{t=1}^{d} \alpha_t^h Q_t.$$

Furthermore we put

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

Corollary 6.4.6. in *Reinert et al. (2005)* immediately gives the following proposition.

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

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

While Proposition 1 only counts the number of occurrences of a fixed word, in our problem, the first m letters $A_1 \dots A_m$ of the sequence $\mathbf{A}$ constitute a random word. Thus we need to condition on the words ω 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.

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

Lemma 1. *With ρ 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 $\rho = \frac{1}{d}$ corresponds to the uniformly distributed case.

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 ω 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*

$$\left| \sum_\omega \mu(\omega) P(\tilde{Z}(\omega) = 0) - \sum_\omega \mu(\omega) P(Z(\omega) = 0) \right| \leq (1 - e^{-\bar{n}\mu^* \rho^{m-1}}) Rem_0,$$

with Rem_0 given in Lemma 1.

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.

We note that in the i.i.d. case, *Reinert and Waterman (2007)* obtain a stronger theorem, making use of the combinatorics from requiring matches in independent sequences.

Remark 1. Lippert et al. (2004) introduce as Z-score

$$Z_{i,n} = \max_m \{ A_{i+k} = A_{j+k}, k = 0, \dots, m-1; 1 \leq i \neq j \leq \bar{n} \}.$$

This is similar to R_n but allows self-overlap. *Lippert et al.* show that the probability $P\{\prod_{i=1}^L \mathbf{1}(Z_{i,n} \geq k)\}$ that the scores $Z_{i,n}$ exceed k consecutively across L positions can be expressed by probabilities involving only R_n, so Corollary 1 can be applied to approximate the distribution of the scores.

References

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).

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.

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.

Robin, S., Rodolphe, F., and Schbath, S. (2005). *DNA, Words and Models. Statistics of Exceptional Words.* Cambridge University Press.

Touyar, N., Schbath, S., Cellier, D. and Dauchel, H. (2008). Poisson approximation for the number of repeats in a Markov chain model. *Submitted.*

Waterman, M.S. (1995). *Introduction to Computational Biology.* Chapman and Hall.