This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

On the Length of the Longest Exact Position Match in a Random Sequence

Gesine Reinert and Michael S. Waterman

Abstract—A mixed Poisson approximation and a Poisson approximation for the length of the longest exact match of a random 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.

Index Terms—Poisson approximation, mixed Poisson approximation, length of longest match, Chen-Stein method.

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, they 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, they 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.

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 i.i.d. letters from a finite alphabet $\mathcal{A}$ with d elements. Let $\pi(a)$ be the probability that a random letter takes on the letter a, and let $\pi^* = \max_{a\in\mathcal{A}} \pi(a)$ be the maximum of these probabilities. The letter distribution is not necessarily uniform. 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 Waterman [3]. Our problem differs from the study of H_n by requiring an exact match beginning at a fixed position in the first sequence.

- *G. Reinert is with Keble College and the Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. E-mail: reinert@stats.ox.ac.uk.*
- *M.S. Waterman is with the Molecular and Computational Biology Department, University of Southern California, 835 West 37th Street, SHS 172 Los Angeles, CA 90089-1340. E-mail: msw@usc.edu.*

Manuscript received 16 Aug. 2005; accepted 10 Jan. 2006; published online 31 July 2006.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0089-0805.

Digital Object Indentifier 10.1109/TCBB.2007.1023
1545-5963/07/$25.00 © 2007 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

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 $P(R_n < m)$. Finally, in Section 4, we illustrate that the approximation for $P(R_n < m)$ is indeed easily computable.

2 A Mixed Poisson Approximation

For Poisson and mixed Poisson approximation, it is useful to think in terms of clumps of occurrences, see Reinert et al. [2], because declumping 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

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

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

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

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 neighboring 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

$$\mu(\omega) = \prod_{i=1}^{m} \pi(w_i)$$

is the probability that a random word of length m equals ω. If there is a p such that $w_i = w_{i+p}$, $i = 1, \dots, m-p$, 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 position 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

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

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 ω. In particular,

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

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

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

It will be convenient to abbreviate, for $r = 1, 2, 3, \dots$,

$$\pi_r = \sum_{a \in \mathcal{A}} (\pi(a))^r,$$

the probability that r random letters match.

Corollary 6.4.6. in [2], together with the independence of the letters, 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,*

$$\begin{aligned} &d_{TV}\big(\mathcal{L}(W_m(\omega)), Po(\tilde{\lambda}(\omega))\big) \\ &\quad \le (n-m+1)\tilde{\mu}(\omega)\{(6m-5)\tilde{\mu}(\omega) + 2(m-1)\mu(\omega)\}. \end{aligned}$$

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 ω that $A_1 \dots A_m$ take on and, using the rule of total probability, we obtain a mixed Poisson approximation.

Theorem 1. *With the above notation,*

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

Remark 1. Recalling that $\pi^* = \max_{a \in \mathcal{A}} \pi(a)$, we note that $\tilde{\lambda}(\omega) \le n(\pi^*)^m$. If we consider the regime that $n(\pi^*)^m$ is approximately constant, with m fixed, then $(\pi^*)^m = O(n^{-1})$ and, using the bound $\pi_3 \le (\pi^*)^2 \sum_{a \in \mathcal{A}} \pi(a) = (\pi^*)^2$, we obtain that $Rem_1 = O(n^{-1})$, thus indicating that the bound in Theorem 1 is of useful order.

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,

$$\begin{aligned} \Big|\sum_\omega \mu(\omega)\epsilon_1 \omega\Big| &\le \bar{n} \sum_\omega \mu(\omega)\tilde{\mu}(\omega)\{(6m-5)\tilde{\mu}(\omega) + 2(m-1)\mu(\omega)\} \\ &=: Rem_1. \end{aligned}$$

For Rem_1, we use that $\tilde{\mu}(\omega) \le \mu(\omega)$ to bound

$$Rem_1 \le (8m-7)\bar{n} \sum_\omega (\mu(\omega))^3.$$

Now, if $A_1 \cdots A_m$, $B_1 \cdots B_m$, and $C_1 \cdots C_m$ are three independent random words, then

$$\begin{aligned} &\sum_\omega (\mu(\omega))^3 \\ &= \sum_\omega P(A_1 \cdots A_m = \omega) P(B_1 \cdots B_m = \omega) P(C_1 \cdots C_m = \omega) \\ &= P(A_1 \cdots A_m = B_1 \cdots B_m = C_1 \cdots C_m) \\ &= (\pi_3)^m, \end{aligned}$$

using that the letters are independent so that

$$Rem_1 \le (8m-7)\bar{n}(\pi_3)^m,$$

as claimed. □

3 Poisson Approximation to the Mixed Poisson Approximation

Although Theorem 1 is valid, $\sum_\omega \mu(\omega) P(Po(\tilde{\lambda}(\omega)) = 0)$ 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. Recall that $\pi^* = \max_{a \in \mathcal{A}} \pi(a)$.

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 $\lambda(\omega)$. Abbreviate $f = \frac{\pi_2^2}{\pi_3}$. Then,*

$$\Big|\sum_\omega \mu(\omega) P(\tilde{Z}(\omega) = 0) - \sum_\omega \mu(\omega) P(Z(\omega) = 0)\Big| = Rem_2,$$

where

$$Rem_2 \le (1 - e^{-\bar{n}(\pi^*)^m}) \Big\{ (\pi_2)^m \frac{f^{-m+\lfloor \frac{m}{2} \rfloor + 1}}{1-f} + \lfloor \frac{m}{2} \rfloor (\pi^*)^m \Big\}. \tag{1}$$

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

Remark 2. In the regime considered in Remark 1, that $n(\pi^*)^m$ is approximately constant, it follows that $Rem_2 = O(n^{-1})$, indicating that the bound in Theorem 2 is usable.

Proof of Theorem 2. First, 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 $0 \le \nu \le \lambda < \infty$,

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

and direct manipulation yields

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

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

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

Thus,

$$\Big|\sum_\omega \mu(\omega) P(\tilde{Z}(\omega) = 0) - \sum_\omega \mu(\omega) P(Z(\omega) = 0)\Big| = Rem_2,$$

where

$$\begin{aligned} |Rem_2| &\le \sum_\omega \mu(\omega) \frac{1 - e^{-\lambda(\omega)}}{\lambda(\omega)} \bar{n} \sum_{p \in \mathcal{P}'(\omega)} \mu(w^{(p)} w) \\ &\le (1 - e^{-\bar{n}(\pi^*)^m}) \sum_{p=1}^{m-1} \sum_\omega \mathbf{1}(p \in \mathcal{P}'(\omega)) \mu(w^{(p)} w). \end{aligned} \tag{2}$$

In the last step, we used the uniform bound $\mu(\omega) \le (\pi^*)^m$ for all ω. To bound the sum $\sum_{p \in \mathcal{P}'(\omega)} \mu(w^{(p)} w)$, we consider the cases that $p \le \lfloor \frac{m}{2} \rfloor$ and $p \ge \lfloor \frac{m}{2} \rfloor + 1$ separately.

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

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

But this probability can be expressed by the probability π_3 that three random letters match and the probability π_2 that two random letters match. We have $m-p$ equations forcing the matching of three random letters each, and $2p-m$ equations forcing the matching of two random letters each. As the letters are independent, the probabilities are easy to calculate;

$$\begin{aligned}&P(A_\ell=A_{\ell+p}=A_{\ell+2p},\ell=1,\dots,m-p;\\&\quad A_\ell=A_{\ell+p},\ell=m-p+1,\dots,p)\\&\quad=\pi_3^{m-p}\pi_2^{2p-m}=\left(\frac{\pi_3}{\pi_2}\right)^m\left(\frac{\pi_2^2}{\pi_3}\right)^p.\end{aligned}$$

Thus, with $f=\frac{\pi_2^2}{\pi_3}$, which is less or equal to 1,

$$\begin{aligned}&\sum_{p=\lfloor\frac{m}{2}\rfloor+1}^{m-1}\sum_\omega \mathbf{1}(p\in\mathcal{P}'(\omega))\mu(w^{(p)}w)\le\\&\sum_{p=\lfloor\frac{m}{2}\rfloor+1}^{m-1}\pi_3^{m-p}\pi_2^{2p-m}=(\pi_2)^m\frac{f^{-m+\lfloor\frac{m}{2}\rfloor+1}-1}{1-f}.\end{aligned}\tag{3}$$

For $p\le\lfloor\frac{m}{2}\rfloor$, we note that, if a word ω has period $p\le\lfloor\frac{m}{2}\rfloor$, then the letters $w_{p+1},\dots,w_m$ are uniquely determined. Therefore, any word can possess at most one prinicpal period $p\le\lfloor\frac{m}{2}\rfloor$.

Again spelling out the periodicity, we obtain that

$$\begin{aligned}&\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}\sum_\omega \mathbf{1}(p\in\mathcal{P}'(\omega))\mu(w^{(p)}w)\\&\le\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}P\Big(A_i=A_{i+p}=\dots=A_{i+(\lfloor\frac{m+p}{p}\rfloor)p},i=1,\dots,m(modp);\\&\quad A_j=A_{j+p}=\dots=A_{j+(\lfloor\frac{m}{p}\rfloor)p},j=m(modp)+1,\dots,p\Big)\\&=\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}\left(\pi_{\lfloor\frac{m+p}{p}\rfloor}\right)^{p-m(mod\ p)}\left(\pi_{\lfloor\frac{m+p}{p}\rfloor+1}\right)^{m(mod\ p)}.\end{aligned}\tag{4}$$

Expression (4) can be bounded further by using that $\pi_r\le\pi^*\pi_{r-1}\le(\pi^*)^{r-1}$, giving

$$\begin{aligned}\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}\left(\pi_{\lfloor\frac{m+p}{p}\rfloor}\right)^{p-m(mod\ p)}\left(\pi_{\lfloor\frac{m+p}{p}\rfloor+1}\right)^{m(mod\ p)}&\le\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}(\pi^*)^{m(mod\ p)}\left(\pi_{\lfloor\frac{m+p}{p}\rfloor}\right)^p\\&=\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}(\pi^*)^{m-\lfloor\frac{m}{p}\rfloor p}\left(\pi_{\lfloor\frac{m+p}{p}\rfloor}\right)^p\\&\le(\pi^*)^m\lfloor\frac{m}{2}\rfloor.\end{aligned}\tag{5}$$

Summarizing, we obtain from (3) and (5) that

$$\sum_{p=1}^{m-1}\sum_\omega \mathbf{1}(p\in\mathcal{P}'(\omega))\mu(w^{(p)}w)\le(\pi_2)^m\frac{f^{-m+\lfloor\frac{m}{2}\rfloor+1}-1}{1-f}+\lfloor\frac{m}{2}\rfloor(\pi^*)^m.\tag{6}$$

Substituting in (2) gives the stated result. □

Remark 3. As $\pi^*=\max_{a\in\mathcal{A}}\pi(a)$, the bound $\mu(w^{(p)}w)\le(\pi^*)^{m+p}$ is immediate. Using that any word of length m that has a principal period $p\le\lfloor\frac{m}{2}\rfloor$ is completely determined by its first p letters, instead of using (4), a "quick and dirty" bound is

$$\begin{aligned}\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}\sum_\omega \mathbf{1}(p\in\mathcal{P}'(\omega))\mu(w^{(p)}w)&\le\sum_{p=1}^{\lfloor\frac{m}{2}\rfloor}d^p(\pi^*)^{m+p}\\&=(\pi^*)^{m+1}d\frac{(\pi^*d)^{\lfloor\frac{m}{2}\rfloor}-1}{\pi^*d-1}.\end{aligned}$$

As $\sum_{a\in\mathcal{A}}\pi_a=1$, we have that $\pi^*d\ge1$. However, if the letter distribution is close to uniform and if m is relatively large, then the above bound will be small.

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,*

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

where

$$Rem_3=Rem_1+\left(1+(m-1)(\pi^*)^m(1-e^{-\bar{n}(\pi^*)^m})^{-1}\right)Rem_2,$$

with Rem_1 given in Theorem 1 and Rem_2 given in Theorem 2.

Remark 4. In the regime that $n(\pi^*)^m$ is approximately constant, we have already seen in Remark 1 and in Remark 2 that $Rem_1=O(n^{-1})$ and $Rem_2=O(n^{-1})$ and, so, also $Rem_3=O(n^{-1})$, providing a useful bound.

Proof of Corollary 1. In view of Theorem 1 and Theorem 2, all that is required is to bound the end effects, resulting from **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., [2, Equation (6.4.10)])

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

We now use (6), giving that

$$\begin{aligned}&P\{\mathbf{1}(R_n>m)\ne\mathbf{1}(W_m=0)\}\\&\quad\le(m-1)(\pi^*)^m\left\{(\pi_2)^m\frac{1-f^{\lfloor\frac{m}{2}\rfloor+1-m}}{f-1}+(\pi^*)^m\lfloor\frac{m}{2}\rfloor\right\}.\end{aligned}$$

Applying Theorem 1 and Theorem 2 finishes the proof. □

Remark 5. Lippert et al. [1] introduce as the Z-score

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

This is similar to R_n, but allows self-overlap. Lippert et al. [1] show that the probability $P\{\prod_{i=1}^L\mathbf{1}(Z_{i,n}\ge k)\}$ 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.

4 Numerical Illustration

A counting argument shows that $\sum_\omega\mu(\omega)P(Z(\omega)=0)=\sum_\omega\mu(\omega)e^{-\lambda(\omega)}$ is not as difficult to evaluate as the expression $\sum_\omega\mu(\omega)P(\tilde{Z}(\omega)=0)$, as follows: Let $n_a(\omega)$ denote the number

TABLE 1
Expected Poisson Parameter

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

of times that letter $a \in \mathcal{A}$ appears in ω. Then, as the letters are independent,

$$\mu(\omega) = \prod_{a \in \mathcal{A}} (\pi(a))^{n_a(\omega)}$$

and, hence, we obtain the multinomial expression

$$\sum_{\omega} \mu(\omega) P(Z(\omega) = 0)$$
$$= \sum_{\substack{(n_a, a \in \mathcal{A}): n_a \in \{0,1,\dots,m\}, \\ \sum_{a \in \mathcal{A}} n_a = m}} \binom{m}{(n_a, a \in \mathcal{A})} \left\{ \prod_{a \in \mathcal{A}} \pi(a)^{n_a} \right\} \exp\left\{ -\bar{n} \prod_{a \in \mathcal{A}} \pi(a)^{n_a} \right\}. \tag{7}$$

While there does not appear to exist a simplifying expression in general, we note that (7) is a polynomial problem in m; indeed, we only need to evaluate $O(m^{d-1})$ summands instead of $O(d^m)$ summands. As we consider d typically much smaller than m, this is a considerable reduction in complexity.

In particular, if $\mathcal{A} = \{A, C, G, T\}$ and if $\pi_A = \pi_T, \pi_C = \pi_G$, as may be reasonable to assume when considering both a DNA sequence and its reverse-complement, then denoting the base-pair probabilities by $p = 2\pi_A = 1 - 2\pi_C$, (7) simplifies to a binomial expectation,

$$\sum_{\omega} \mu(\omega) P(Z(\omega) = 0) =$$
$$\sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \exp\left\{ -\bar{n} 2^{-m} p^k (1-p)^{m-k} \right\},$$

where k stands for the sum $n_A + n_T$.

Example. Suppose as in [1] 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 nonrepeat 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 $m = 15, \dots, 22$.

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 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 $\sqrt{Var\hat{\rho}(m)}$ from [1] and the separate remainder terms contributing to our bound; recall that Rem_2 is given in (1).

Our approximated probabilities are similar to the Monte-Carlo estimates in [1]. However, whereas [1] can only conclude that, say,

TABLE 2
Estimated Probabilities, Bounds, and Remainder Terms

m	$\rho(m)$	$\hat{\rho}(m)$	bound	$\sqrt{Var\hat{\rho}(m)}$	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

an approximate 95 percent confidence interval for the true probability $P(R_n \geq m)$ is given by $\hat{\rho} \pm 1.96\sqrt{Var\hat{\rho}(m)}$, we indeed proved that the true probability will lie within $\rho(m) \pm bound$, 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.

ACKNOWLEDGMENTS

The authors would like to thank an anonymous referee for very helpful comments. G. Reinert was supported in part by EPSRC grant no. GR/R52183/01. M.S. Waterman was supported by US National Institutes of Health grant no. P50 HG 002790.

REFERENCES

[1] R.A. Lippert, X. Zhao, L. Florea, C. Mobarry, and S. Istrail, "Finding Anchors for Genomic Sequence Comparison," *Proc. Eighth Ann. Int'l Conf. Research in Computational Biology*, pp. 233-241, 2004, also in *J. Compuational Bioliogy*, vol. 12, pp. 762-776, 2005.

[2] G. Reinert, S. Schbath, and M.S. Waterman, "Statistics on Words with Applications to Biological Sequences," *Lothaire: Applied Combinatorics on Words*, J. Berstel and D. Perrin, eds., pp. 251-328, Cambridge Univ. Press, 2005.

[3] M.S. Waterman, *Introduction to Computational Biology*. Chapman and Hall, 1995.