Word Match Counts Between Markovian Biological Sequences

Conrad Burden[1], Paul Leopardi[1], and Sylvain Forêt[2]

[1] Mathematical Sciences Institute, Australian National University, Canberra, Australia
conrad.burden@anu.edu.au,
WWW home page: http://wwwmaths.anu.edu.au/~burden
[2] Research School of Biology, Australian National University, Canberra, Australia

Abstract. The D_2 statistic, which counts the number of word matches between two given sequences, has long been proposed as a measure of similarity for biological sequences. Much of the mathematically rigorous work carried out to date on the properties of the D_2 statistic has been restricted to the case of 'Bernoulli' sequences composed of identically and independently distributed letters. Here the properties of the distribution of this statistic for the biologically more realistic case of Markovian sequences is studied. The approach is novel in that Markovian dependency is defined for sequences with periodic boundary conditions, and this enables exact analytic formulae for the mean and variance to be derived. The formulae are confirmed using numerical simulations, and asymptotic approximations to the full distribution are tested.

Keywords: word matches, biological sequence comparison.

1 Introduction

The D_2 statistic is defined as the number of exact word matches of pre-specified length k between two sequences of letters from a finite alphabet $\mathcal{A}$. This statistic [13], and its many variants [19, 17, 9, 10] have been proposed as a measures of similarity between biological sequences in cases where the more commonly used alignment methods may not be appropriate. The distributional properties of the D_2 statistic under the null hypothesis of sequences composed of independently and identically distributed (i.i.d.) letters have been studied extensively [13, 6, 11, 8, 7, 4].

Analysis of the k-mer spectra of the genomes of several species provides strong evidence that genomic sequences are more appropriately modelled as having a Markovian dependence [5]. In the current work existing exact analytic results results for the mean, variance and an empirical distribution of D_2 for i.i.d. sequences is extended to the case of Markovian sequences.

A previous study of this problem, with some approximations, has been carried out by Kantorovitz et al. [12] in the process of developing a method for detecting regulatory modules in genomic sequences. The current study differs in that we

consider sequences with periodic boundary conditions (PBCs), for which we introduce a new definition of Markovian sequences. The restriction to periodic sequences simplifies calculations of the mean and variance, enabling an exact analytic formula for the variance for first order Markovian sequences which is rapidly computable to double precision accuracy for arbitrary sequence lengths. In biological aplications of the analogous results for i.i.d. sequences [7, 4] we have found generally that the PBCs are not an impediment, as they can simply be imposed on the sequences prior to calculating D_2 without without seriously affecting its efficacy as a measure of sequence similarity.

2 Definitions

Definition 1. *Consider a sequence* $\mathbf{x} = x_1, x_2 \ldots$ *of letters from an alphabet* $\mathcal{A}$ *of size* d*. We say that* $\mathbf{x}$ *has periodic boundary conditions (PBCs) and is of length* m *if* $x_{i+m} = x_i$ *for all* $i = 1, 2, \ldots$.

A sequence $\mathbf{X} = X_1, X_2 \ldots$ *of random letters has an* θ*-th order Markovian dependence if*

$$\begin{aligned} \text{Prob}\left((X_{i+\theta} = b | (X_i, \ldots, X_{i+\theta-1} = (a_1, \ldots, a_\theta)\right) \\ = M(a_1, \ldots, a_\theta; b) \ , \end{aligned} \tag{1}$$

for a specified $d^\theta \times d$ *matrix* M *satisfying*

$$0 \le M(a_1, \ldots, a_\theta; b) \le 1 \ ; \quad \sum_{b \in \mathcal{A}} M(a_1, \ldots, a_\theta; b) = 1 \ , \tag{2}$$

for all $a_1, \ldots, a_\theta, b \in \mathcal{A}$.

As a shorthand notation, we will write a string of length θ in bold italics:

$$\boldsymbol{x} = (x_1, \ldots x_\theta) \ , \tag{3}$$

and write any substring of $\mathbf{X}$ of length θ in a similar fashion, labelled by the index of the first element:

$$\boldsymbol{X}_i = (X_i, \ldots X_{i+\theta-1}) \ , \tag{4}$$

Thus (1) is written more compactly as

$$\text{Prob}\left(X_{i+\theta} = b | \boldsymbol{X}_i = \boldsymbol{a}\right) = M(\boldsymbol{a}; b) \ . \tag{5}$$

Following the notation of ref. [18], define a $d^\theta \times d^\theta$ square matrix $\mathbb{M}$ as

$$\mathbb{M}(\boldsymbol{a}, \boldsymbol{b}) = \begin{cases} M(\boldsymbol{a}; b_\theta) & \text{if } (a_2, \ldots, a_\theta) = (b_1, \ldots b_{\theta-1}), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Then the Markovian dependency can be written as a first order Markovian dependency as

$$\text{Prob}\left(\boldsymbol{X}_{i+1} = \boldsymbol{b} | \boldsymbol{X}_i = \boldsymbol{a}\right) = \mathbb{M}(\boldsymbol{a}, \boldsymbol{b}) \ . \tag{7}$$

2.1 Markov Sequences with PBCs

Definition 2. *Given an order θ Markov transition matrix M, we define a Markov sequence with PBCs of length n to be a random sequence $\mathbf{X} = X_1, X_2 \ldots, X_n$ for which the probability of observing the sequence $\mathbf{x} = x_1, x_2 \ldots, x_n \in \mathcal{A}^n$ is*

$$\text{Prob}\,(\mathbf{X} = \mathbf{x}) = \frac{\mathbb{M}(\boldsymbol{x}_1, \boldsymbol{x}_2)\mathbb{M}(\boldsymbol{x_2}, \boldsymbol{x_3}) \ldots \mathbb{M}(\boldsymbol{x}_m, \boldsymbol{x}_1)}{\text{tr}\,(\mathbb{M}^m)} \quad , \tag{8}$$

where $\mathbb{M}$ is the equivalent first order transition matrix defined by (6) [15].

It is shown in ref. [4] that the following algorithm gives a practical way of generating such a sequence

Algorithm 1
Step 1: *Generate $\boldsymbol{X}_1 = X_1, \ldots X_\theta$ from the uniform distribution* $\text{Prob}\,(\boldsymbol{X}_1 = \boldsymbol{x}) = 1/d^\theta$ *for all $\boldsymbol{x} \in \mathcal{A}^\theta$.*
Step 2: *Generate $X_{\theta+1}, \ldots, X_{\theta+n}$ using (5).*
Step 3: *If $\boldsymbol{X}_{n+1} = \boldsymbol{X}_1$, accept the sequence $\mathbf{X} = X_1, X_2 \ldots, X_n$, otherwise repeat from Step 1 until an accepted sequence is obtained.*

Note that, counter-intuitively, it is important that the initial θ-mer is chosen from a uniform distribution and not the stationary distribution of the Markov model in order to generate the correct distribution.

3 Strand Symmetry and the Transition Matrix M

To model genomic DNA sequences, the transition matrix M is generally estimated from observed word counts in a genome or part of a genome via the asymptotic maximum likelihood estimator for infinitely long sequences

$$\hat{M}(\boldsymbol{a}; b) = \frac{N_{\boldsymbol{a}b}}{N_{\boldsymbol{a}}} \tag{9}$$

where $N_{\boldsymbol{a}b}$ is the number of occurrences of the $(\theta+1)$-mer $(a_1 \ldots a_\theta b)$ and $N_{\boldsymbol{a}} = \sum_{c \in \mathcal{A}} N_{\boldsymbol{a}c}$ is the number of occurences of the θ-mer $\boldsymbol{a}$ [18].

Most genomic sequences, when examined on a sufficiently large scale, are observed to have the property of strand symmetry [1]. That is, the number of occurrences of any given k-mer is, to a good approximation, equal to the number of occurrences of its reverse complement. In the interests of reducing the number of free parameters one would like to build this property into genomic Markov models.

To give a more mathematical framework to this statement, let us assume that the alphabet size d is even and each letter $a \in \mathcal{A}$ has a complement $\bar{a}$ such that $\bar{\bar{a}} = a$ and $\bar{a} \neq a$. In general the alphabet splits into $d/2$ 'purines' and $d/2$ 'pyrimidines'. For the usual nucleotide alphabet, A and G are purines, C and T are pyrimidines and $\bar{A} = T$, $\bar{G} = C$. We wish to determine what

practical restrictions are placed on the estimate $\hat{M}$ of the transition matrix by the strand-symmetry restriction

$$\text{Prob}\left((X_1,\ldots,X_n)=(x_1,\ldots,x_n)\right)=\text{Prob}\left((X_1,\ldots,X_n)=(\bar{x}_n,\ldots,\bar{x}_1)\right)\ . \tag{10}$$

To this purpose it is more convenient to work with the square matrix $\mathbb{M}$ defined by (6). Define a matrix $\mathbb{N}_{\boldsymbol{ab}}$, $\boldsymbol{a},\boldsymbol{b}\in\mathcal{A}^\theta$, from the count matrix $N(\boldsymbol{a},b)$ in a manner analogous to (6). Then $\mathbb{M}$ is estimated by normalising the rows of $\mathbb{N}$ to add to 1:

$$\hat{\mathbb{M}}(\boldsymbol{a},\boldsymbol{b})=\frac{\mathbb{N}_{\boldsymbol{ab}}}{\sum_{c\in\mathcal{A}^\theta}\mathbb{N}_{\boldsymbol{ac}}}\ . \tag{11}$$

Now rearrange the order of columns of $\mathbb{N}$ to form a new matrix Q so that if the rows are labelled by the complete set of θ-mers $\boldsymbol{w}_1,\boldsymbol{w}_2,\ldots\boldsymbol{w}_{d^\theta}$ the columns are labelled in the order $\bar{\boldsymbol{w}}_1,\bar{\boldsymbol{w}}_2,\ldots\bar{\boldsymbol{w}}_{d^\theta}$, where $\bar{\boldsymbol{w}}_i=\overline{(w_{i1}\ldots w_{i\theta})}=(\bar{w}_{i\theta}\ldots\bar{w}_{i1})$ is the reverse complement of the ith θ-mer $\boldsymbol{w}_i$. Strand symmetry implies that Q will be symmetric because the probability of making the transition $(a_1\ldots a_\theta)\to(b_1\ldots b_\theta)$ will be the same as the probability of making the transition $(\bar{b}_\theta\ldots\bar{b}_1)\to(\bar{a}_\theta\ldots\bar{a}_1)$.

The problem of determining the most general form of transition matrix is equivalent to answering the following question: how many independent non-zero elements does the matrix Q have, given the restrictions

1. $Q_{\boldsymbol{ab}}=Q_{\bar{\boldsymbol{b}}\bar{\boldsymbol{a}}}$
2. $Q_{\boldsymbol{ab}}$ is zero unless $(a_2\ldots a_\theta)=(\bar{b}_\theta\ldots\bar{b}_2)$?

Consider first any diagonal element $Q_{\boldsymbol{aa}}$. It will be zero unless $(a_2\ldots a_\theta)=(\bar{a}_\theta\ldots\bar{a}_2)$. If θ is even, this requires $a_{\theta/2+1}=\bar{a}_{\theta/2+1}$ which is impossible since no letter of the alphabet is its own complement. If θ is odd this condition can be satisfied by independently specifiying the letters $a_1,\ldots,a_{(\theta+1)/2}$. Thus the number of non-zero diagonal elements of Q is

$$\begin{array}{ll}0 & \text{if } \theta \text{ is even}\ ,\\ d^{(\theta+1)/2} & \text{if } \theta \text{ is odd}\ .\end{array} \tag{12}$$

Now consider the off-diagonal elements of Q. The total number of non-zero off-diagonal elements is

$$\begin{array}{ll}d^{\theta+1}-0 & \text{if } \theta \text{ is even}\ ,\\ d^{\theta+1}-d^{(\theta+1)/2} & \text{if } \theta \text{ is odd}\ .\end{array} \tag{13}$$

Since the matrix is symmetric, exactly half of these are independent:

$$\begin{array}{ll}\frac{1}{2}d^{\theta+1} & \text{if } \theta \text{ is even}\ ,\\ \frac{1}{2}d^{(\theta+1)/2}(d^{(\theta+1)/2}-1) & \text{if } \theta \text{ is odd}\ .\end{array} \tag{14}$$

Finally, adding the number of independent diagonal elements gives the number of independent elements of M as

$$\begin{array}{ll}\frac{1}{2}d^{\theta+1} & \text{if } \theta \text{ is even}\ ,\\ \frac{1}{2}d^{(\theta+1)/2}(d^{(\theta+1)/2}+1) & \text{if } \theta \text{ is odd}\ .\end{array} \tag{15}$$

The number of independent elements of Q for an alphabet of size $d = 4$ is listed in Table 1.

In order to estimate a strand-symmetric transition matrix from a given observed sequence, it is sufficient to extend the definition of the count matrix $N_{\boldsymbol{a}b}$ to include a count of the number of occurrences of then $(\theta+1)$-mer $(a_1 \dots a_\theta b)$ in the sequence plus the number of occurrences of the same $(\theta+1)$-mer in its reverse complement. Such a matrix will automatically have the same number of independent elements as the corresponding Q, and the matrix $\hat{M}(\boldsymbol{a}, b)$ constructed according to (9) will then have the required symmetry.

Table 1. The number of independent $\theta + 1$-mer word counts $N_{\boldsymbol{a}b}$ needed to estimate a θ-order strand-symmetric transition matrix for an alphabet of $d = 4$ letters.

θ	# of independent elements of Q	# of non-zero elements of $\mathbb{M}$
1	10	16
2	32	64
3	136	256
4	512	1024
5	2080	4096
6	8192	16384

4 The D_2 Statistic

We now consider statistical properties of the alignment-free sequence similarity measure known as the D_2 statistic. The distributional properties of this statistic presented here apply to any higher order Markov model with or without the constraint of strand symmetry.

Definition 3. *Given two sequences* $\mathbf{X}$ *and* $\mathbf{Y}$ *with PBCs of length* m *and* n *respectively, the* D_2 *statistic is defined as the number of* k*-word matches, including overlaps, between* $\mathbf{X}$ *and* $\mathbf{Y}$*:*

$$D_2(k, M) = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} \ , \tag{16}$$

where I_{ij} *is the word match indicator random variable for words length* k *positioned at site* i *in sequence* $\mathbf{X}$ *and site* j *in sequence* $\mathbf{Y}$*:*

$$I_{ij} = \begin{cases} 1 & \text{if } (X_i, \dots, X_{i+k-1}) = (Y_j, \dots, Y_{j+k-1}) \ , \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

More specifically, we are interested in the case where the two sequences are Markovian. From (6) and (7) it is easy to see that any ensemble of pairs of random sequences $(\mathbf{X}, \mathbf{Y})$ generated by an θ-order transition matrix M is in one-to-one correspondence with an ensemble of pairs of random sequences $(\mathbb{X}, \mathbb{Y})$ of letters from a d^θ-letter alphabet generated by the equivalent sparse $d^\theta \times d^\theta$ matrix $\mathbb{M}$. Furthermore, any k-mer match between $\mathbf{X}$ and $\mathbf{Y}$ corresponds to a $(k-\theta+1)$-mer match between $\mathbb{X}$ and $\mathbb{Y}$. It follows that the distributional properties of D_2 for Markovian sequences can be determined in terms of the properties of D_2 for an equivalent first order system. In particular, for the mean and variance:

$$\begin{aligned} E(D_2(k,M)) &= E(D_2(k-\theta+1,\mathbb{M})) \ , \\ \mathrm{Var}\,(D_2(k,M)) &= \mathrm{Var}\,(D_2(k-\theta+1,\mathbb{M})) \ , \end{aligned} \qquad k \geq \theta \ . \tag{18}$$

Therefore, to calculate $E(D_2(k,M))$ and $\mathrm{Var}\,(D_2(k,M))$ for any $k \geq \theta$ it is sufficient to derive formulae for a first order Markov model. These formulae are given below. An R implementation [16] of the mean and variance for $k \geq \theta$ and a formula for the mean when $k < \theta$ will be published elsewhere [4].

4.1 D_2 Mean for $\theta = 1$

The mean of D_2 for $\theta = 1$ is

$$E(D_2) = \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \mathrm{tr}\,[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1}] \ , \tag{19}$$

where the Hadamard product $A \circ B$ of two matrices A and B is defined as the matrix whose (α, β)-th element is

$$(A \circ B)_{\alpha\beta} = A_{\alpha\beta} B_{\alpha\beta} \ . \tag{20}$$

Proof. We have that

$$E(D_2) = \sum_{i=1}^{m} \sum_{j=1}^{n} E(I_{ij}) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathrm{Prob}\,(I_{ij} = 1) \ , \tag{21}$$

where

$$\mathrm{Prob}\,(I_{ij} = 1) = \sum_{w \in \mathcal{A}^k} \mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w) \mathrm{Prob}\,(Y_j \ldots Y_{j+k-1} = w) \ . \tag{22}$$

To calculate $\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w)$ we sum (8) over all sequences $\mathbf{x}$ such that $(x_i \ldots x_{i+k-1}) = w$. Thus

$$\begin{aligned} &\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w) = \\ &\qquad \frac{M^{m-k+1}(w_k, w_1) M(w_1, w_2) \ldots M(w_{k-1}, w_k)}{\mathrm{tr}\,(M^m)} \ , \end{aligned} \tag{23}$$

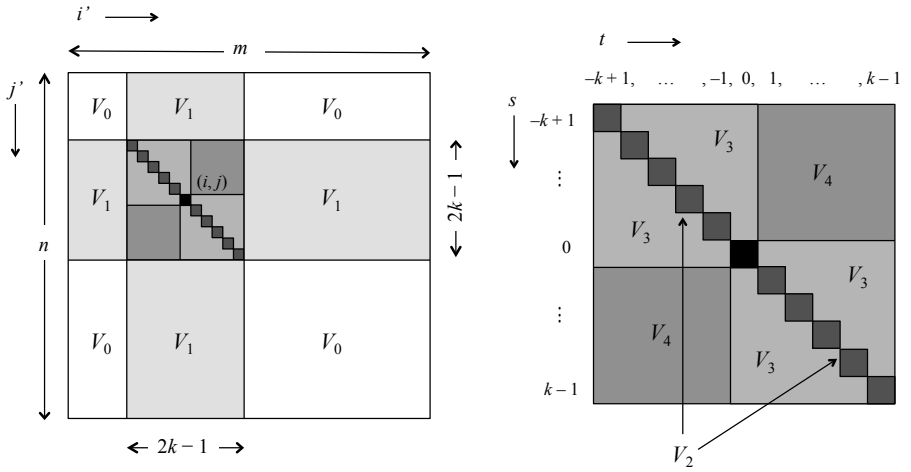

Fig. 1. Contributions to $\mathrm{Var}\,(D_2)$ via the sum in (27). The left-hand diagram shows the (i', j')-plane for a fixed value of (i, j), shown as the black square. The right-hand diagram is an expanded view of the 'accordion' region $-k+1 \le s, t \le k-1$, where $t = i' - i$ and $s = j' - j$ up to PBCs.

where the factor $M^{m-k+1}(w_k, w_1)$ arises from summing over the letters $x_1, \ldots, x_{i-1}, x_{i+k}, \ldots, x_m$. Similarly we have

$$\mathrm{Prob}\,(Y_j \ldots Y_{j+k-1} = w) = \frac{M^{n-k+1}(w_k, w_1) M(w_1, w_2) \ldots M(w_{k-1}, w_k)}{\mathrm{tr}\,(M^n)} \quad . \tag{24}$$

Substituting (23) and (24) into (22) gives

$$\mathrm{Prob}\,(I_{ij} = 1) = \frac{\mathrm{tr}\,[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1}]}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \quad . \tag{25}$$

Equation (21) then gives the required result. □

4.2 D_2 Variance for $\theta = 1$

The exact variance of D_2 for Markovian sequences with PBCs requires an extensive calculation. Here we give a summary of the result, which is valid for $m, n \ge 2k$. Full technical details of the derivation will be published elsewhere [4].

We have

$$\mathrm{Var}\,(D_2) = E({D_2}^2) - E(D_2)^2 \quad . \tag{26}$$

The second term can be calculated from (19). The first term is a sum of contributions obtained from (16) by partitioning a sum over words beginning at

positions i and i' in sequence $\mathbf{X}$ and beginning at j and j' in sequence $\mathbf{Y}$,

$$\begin{aligned} E({D_2}^2) &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} E(I_{ij} I_{i'j'}) \\ &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} \text{Prob}\,(I_{ij} = 1, I_{i'j'} = 1) \\ &= V_0 + V_1 + V_2 + V_3 + V_4 \ . \end{aligned} \tag{27}$$

The partitioning reflects the degree of overlap between words in each of the two sequences, and is illustrated in Fig. 1. We assume $m, n \geq 2k$, which will almost certainly be the case in any biological application.

We will write a Hadamard product of q factors, $M \circ \ldots \circ M$, using the shorthand notation $M^{\circ q}$. With this notation, the contributions to the variance are:

$$\begin{aligned} V_0 = & \frac{mn}{\text{tr}\,(M^m)\text{tr}\,(M^n)} \times \\ & \sum_{r=0}^{m-2k} \sum_{s=0}^{n-2k} \text{tr}\,\left[(M^{r+1} \circ M^{s+1})(M \circ M)^{k-1} \ \times \right. \\ & \left. (M^{m-2k-r+1} \circ M^{n-2k-s+1})(M \circ M)^{k-1} \right] \ , \end{aligned} \tag{28}$$

$$\begin{aligned} V_1 = & \frac{mn}{\text{tr}\,(M^m)\text{tr}\,(M^n)} \times \\ & \left\{ \sum_{s=0}^{n-2k} \left[\text{tr}\,\{ [(M \circ M \circ M)^{k-1} \circ (M^{s+1})^T] \times \right. \right. \\ & \qquad \left. (M^{m-k+1} \circ M^{n-2k-s+1}) \} \right. \\ & + 2 \sum_{r=1}^{k-1} \text{tr}\,\{ (M \circ M)^r \times \\ & \qquad [(M \circ M \circ M)^{k-r-1} \circ (M^{s+1})^T] \times \\ & \qquad \left. \left. (M \circ M)^r (M^{m-k-r+1} \circ M^{n-2k-s+1}) \} \right] \right. \\ & \left. + \text{ the same with } m \text{ and } n \text{ interchanged.} \right\} \ , \end{aligned} \tag{29}$$

$$V_2 = \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \left\{ \mathrm{tr}\,[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1}] + 2\sum_{t=1}^{k-1} \mathrm{tr}\,[(M^{m-k-t+1} \circ M^{n-k-t+1}) \times (M \circ M)^{k+t-1}] \right\} \ , \tag{30}$$

$$V_3 = \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \sum_{t=1}^{k-1}\sum_{s=0}^{t-1} \mathrm{tr}\left[(M \circ M)^s Q (M \circ M)^s \times (M^{m-k-t+1} \circ M^{n-k-s+1} + M^{n-k-t+1} \circ M^{m-k-s+1})\right] \ , \tag{31}$$

where

$$Q = \begin{cases} (M^{\circ(2\nu+3)})^{\rho-1} \circ [(M^{\circ(2\nu+1)})^{t-s-\rho+1}]^T & \\ & \text{if } \rho > 0 \ , \\ (M^{\circ(2\nu+1)})^{t-s-1} \circ (M^{\circ(2\nu-1)})^T & \\ & \text{if } \rho = 0 \ , \end{cases} \tag{32}$$

and

$$\nu = \left\lfloor \frac{k-s}{t-s} \right\rfloor \ , \qquad \rho = (k-s) \bmod (t-s) \ . \tag{33}$$

Finally,

$$V_4 = \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \sum_{r,t=1}^{k-1} \mathrm{tr}\,U \ , \tag{34}$$

where

$$
U = \begin{cases}
\left\{(M^{\circ(2\nu+1)})^{t-1} \circ (M^{m-k-t+1})^T\right\} M^{\circ 2\nu} \times \\
\quad \left\{(M^{\circ(2\nu+1)})^{r-1} \circ (M^{n-k-r+1})^T\right\} M^{\circ 2\nu} \\
\qquad\qquad \text{if } \zeta = 0 \ , \\
\left\{(M^{\circ(2\nu+1)})^{r-\zeta+1} \circ M^{m-k-t+1}\right\} \times \\
\quad (M^{\circ(2\nu+2)})^{\zeta-1} \times \\
\qquad \left\{(M^{\circ(2\nu+1)})^{t-\zeta+1} \circ M^{n-k-r+1}\right\} \times \\
\qquad\quad (M^{\circ(2\nu+2)})^{\zeta-1} \\
\qquad\qquad\qquad \text{if } 0 < \zeta \le r, t \ , \\
\left\{(M^{\circ(2\nu+3)})^{\zeta-r-1} \circ (M^{m-k-t+1})^T\right\} \times \\
\quad (M^{\circ(2\nu+2)})^{r} \left\{(M^{\circ(2\nu+1)})^{t-\zeta+1} \circ M^{n-k-r+1}\right\} \\
\qquad \times (M^{\circ(2\nu+2)})^{r} \\
\qquad\qquad\qquad \text{if } r < \zeta \le t \ , \\
\{\text{as above with } m \text{ and } n \text{ interchanged} \\
\quad \text{and } r \text{ and } t \text{ interchanged}\} \\
\qquad\qquad\qquad \text{if } t < \zeta \le r \ , \\
\left\{(M^{\circ(2\nu+3)})^{\zeta-r-1} \circ (M^{m-k-t+1})^T\right\} \times \\
\quad (M^{\circ(2\nu+2)})^{t+r-\zeta+1} \times \\
\qquad \left\{(M^{\circ(2\nu+3)})^{\zeta-t-1} \circ (M^{n-k-r+1})^T\right\} \times \\
\qquad\quad (M^{\circ(2\nu+2)})^{t+r-\zeta+1} \\
\qquad\qquad\qquad \text{if } r, t < \zeta \ ,
\end{cases}
$$

and

$$
\nu = \left\lfloor \frac{k}{r+t} \right\rfloor \ , \qquad \zeta = k \bmod (r+t) \ . \tag{35}
$$

5 Numerical Simulations

For short sequences and small alphabets the distribution of the D_2 statistic can be computed by enumerating all possible sequences. We have confirmed the accuracy of the formulae for the mean and variance given in Sect. 4 to 11 significant figures by generating the complete distribution of D_2 using double precision arithmetic for sequences up to length $m = n = 9$ for $k = 3$, $d = 2$ and up to length $m = n = 7$ for $k = 2$, $D = 3$. The Markov matrices M are generated randomly by choosing each element from a uniform distribution on the interval $[0, 1]$ and then normalising each row sum to 1. Two examples of the exact D_2 distribution are shown in Fig. 2. Note that the introduction of random Markov matrices is to enable an efficient check of the above formulae for a range of M, and is not intended to have any biological meaning. Maximum likelihood estimates of Markov transition matrices from various genomes have

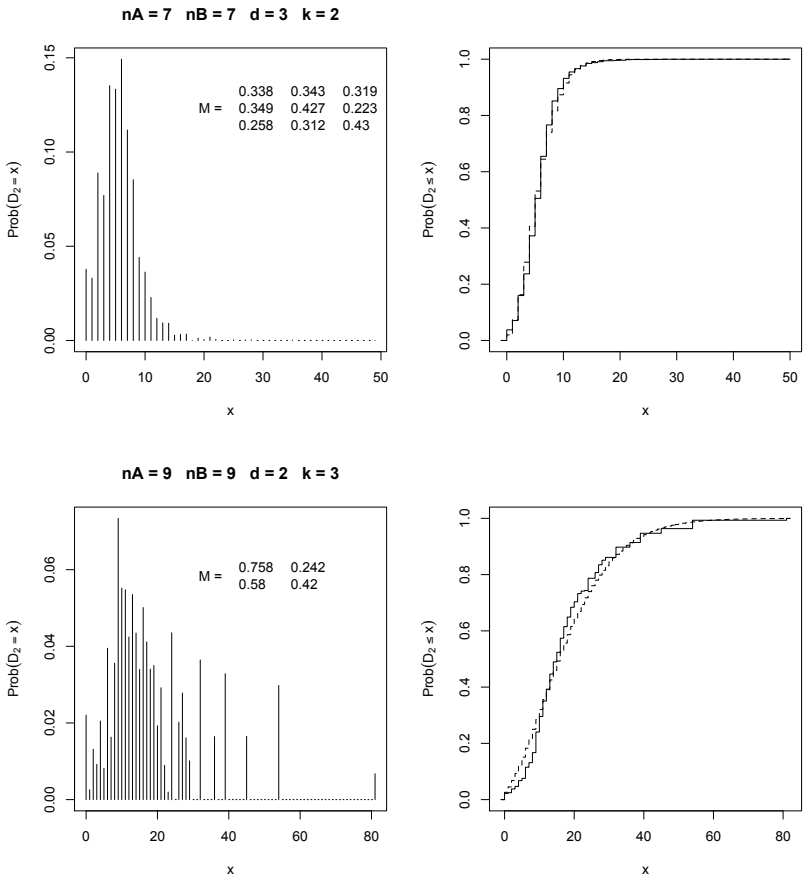

Fig. 2. The exact distiribution of the D_2 statistic for short sequences of length n_A, n_B and words of length k from an alphabet of size d. The Markov matrix M has been generated randomly in each case. Also shown (dashed curve) is the cumulative distribution of the Pólya-Aeppli distribution with mean and variance set to the theoretical values using the formulae of Sect. 4.

been published, for instance, by Chor et al. [5], which can be used in biological applications. We note that the Chor estimates are close to satisfying the strand-symmetry condition restrictions of Sect. 3 (data not shown).

For longer sequences of realistic biological length, the distribution of D_2 can be estimated from a Monte Carlo ensemble of random sequences generated from the algorithm described in Sect. 2.1. Examples of cumulative distribution functions for $d = 4$, $k = 4$ estimated from ensembles of $10,000$ pairs of independently generated random sequences of length $m = n = 100$ and 400 are shown in Figs. 3

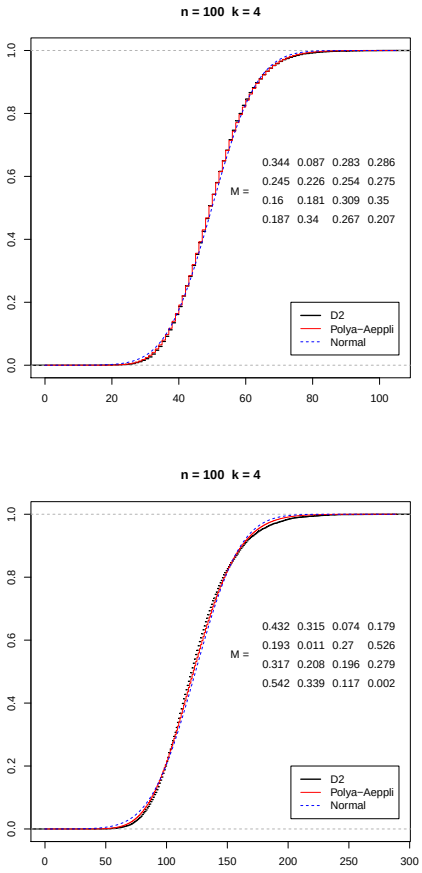

Fig. 3. Two examples of empirical cumulative distribution of the D_2 statistic estimated from $10,000$ independently generated random sequences of length $m = n = 100$ for words of length $k = 4$ and an alphabet of size $d = 4$. The Markov matrix M has been generated randomly in each case. Also shown are the cumulative distribution of the normal and Pólya-Aeppli distributions with mean and variance set to the theoretical values using the formulae of Sect. 4.

and 4 respectively. The Markov matrix is again generated randomly, and it is interesting to note that the mean of the distribution can vary considerably with M. We have made a number of simulations, and find that in roughly the expected proportion of times the mean and variance calculated from the formulae of Sect. 4 lie within the 95% confidence intervals computed from the ensemble.

For the case of sequences composed of i.i.d. letters certain rigorous results are known for the asymptotic distribution of D_2 as the sequence lengths $m, n \to \infty$.

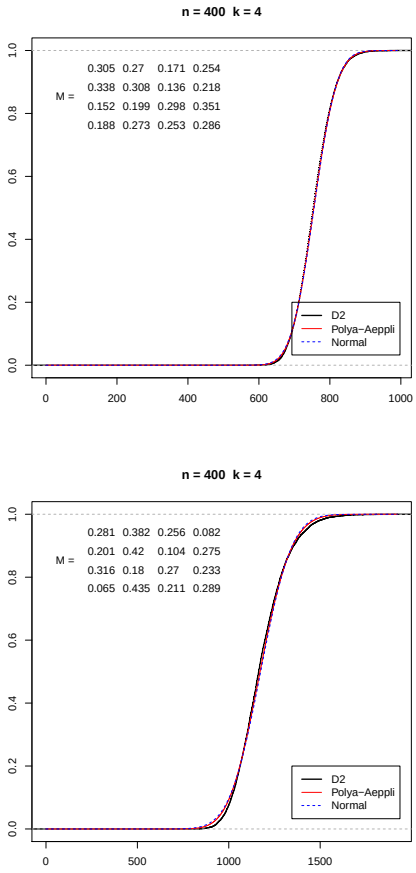

Fig. 4. The same as Figure 3, except $m = n = 400$.

For $m = n$, it has been shown that the limiting distribution is normal in the regime $k < 1/2 \log_b n + \text{const.}$ [3] and Pólya-Aeppli in the regime $k > 2 \log_b n + \text{const.}$ [13]. Here $b = 1/\sum_{a \in \mathcal{A}} p_a^2$ where p_a is the probability of occurrence of letter a. A Pólya-Aeppli random variable is the sum of a Poisson number of geometric random variables, and is therefore an example of a compound Poisson random variable. It often arises in the study of random word counts as a Poisson number of clumps of overlapping words, each clump containing a geometric number of k-words [18]. Although the asymptotic results for D_2 are not proved for Markovian sequences, it is a reasonable experiment to compare our numerical simulations with these distributions as they may potentially provide an accurate estimate of p-values in biological applications.

One would not expect the asymptotic distributions to be an accurate fit for the short sequences considered in Fig. 2. Nevertheless we have included the Pólya-Aeppli distribution function and find it to be surprisingly close for the $d = 3$ case. Disagreement arises in the tail of the distribution because, for combinatoric reasons, certain values of D_2 within the range 0 to mn do not occur, whereas the Pólya-Aeppli has support over the whole range (and also out to ∞, albeit with very low probability).

If one were dealing with i.i.d. sequences with a uniform letter distribution, then the parameters $m = n = 100$ or 400, $k = 4$ used for the simulations in Figs. 3 and 4 would inhabit the region between the normal and Pólya-Aeppli asymptotic regimes described above. Both asymptotic distributions are superimposed on the empirical distribution functions in Figs. 3 and 4. We observe that the normal and Pólya-Aeppli do not differ greatly from one another, though the Pólya-Aeppli does appear to give a better fit, particularly in the important tail of the distribution relevant to estimating p-values.

6 Conclusions

This paper introduces the concept of periodic boundary conditions for Markovian sequences as an elegant mathematical construct which avoids the inconvenience of boundary effects in analytic calculations. We have demonstrated that the mean and variance of the D_2 word match statistic can be calculated analytically and readily computed to any desired accuracy through formulae involving only traces of products of matrices. Calculation of the mean and variance is fast as powers of Hadamard products need only be calculated once for a given Markovian model, and only need to be calculated up to the point of convergence. For biological applications such as measuring sequence similarity or identifying regions of regulatory motifs, sequences lengths tend to be of at least a few hundred letters. In these cases loss of information about boundary effects is unlikely to be a serious impediment. For instance, in previous studies of a database of cis-regulatory modelled as a set of i.i.d. sequences was successfully studied using the D_2 statistics simply by imposing PBCs on the sequences prior to calculating the D_2 [7, 2].

The current work is a preliminary study designed to illustrate the computational effectiveness of imposing periodic boundary conditions when calculating the D_2 statistic. In ongoing work we are testing the agreement between the theoretical Markovian distributions studied herein and empirical distributions from genomic DNA. In general, we find that the empirical distribution tends to have heavier left and right tails, suggesting the existence of a subset of k-mers which are over- or under-represented within the genomes studied [4].

Further work also needs to be done on extending the results to more viable variants of the D_2 statistic. It has been argued that a potential shortcoming of the D_2 statistic is that the signal of sequence similarity one is trying to detect maybe hidden by its variability due to noise in each of the single sequences, and that to overcome this problem one should instead calculate a 'centred' version of D_2 in

which word count vectors are replaced with those centred about their mean [13, 17]. There also exist 'standardised' versions of D_2 [14, 9] designed to account for biases arising from the fact that some words are naturally over-represented, and 'weighted' versions [10] designed to account for higher substitution rates of chemically similar amino acids in protein sequences. Extension of the mathematical formalisms developed herein to these D_2 variants, as well as a more compete study of the accuracy of approximating p-values with asymptotic distributions, will be the subject of future work.

Acknowledgement

This work was supported in part by ARC Discovery Grant DP120101422 and NHMRC grant 525453.

References

1. Baisnée, P.F., Hampson, S., Baldi, P.: Why are complementary DNA strands symmetric? Bioinformatics **18**(8) (2002) 1021–1033
2. Burden, C.J., Jing, J., Wilson, S.R.: Alignment-free sequence comparison for biologically realistic sequences of moderate length. Statistical Applications in Genetics and Molecular Biology **11**(1) (2012) Article 3
3. Burden, C.J., Kantorovitz, M.R., Wilson, S.R.: Approximate word matches between two random sequences. Annals of Applied Probability **18**(1) (2008) 1–21
4. Burden, C.J., Leopardi, P., Forêt, S.: The distribution of word matches between Markovian sequences with periodic boundary conditions. (in preparation) (2013)
5. Chor, B., Horn, D., Goldman, N., Levy, Y., Massingham, T.: Genomic DNA k-mer spectra: models and modalities. Genome Biology **10** (2009) R108
6. Forêt, S., Kantorovitz, M.R., Burden, C.J.: Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. BMC Bioinformatics **7 Suppl 5** (2006) S21
7. Forêt, S., Wilson, S.R., Burden, C.J.: Characterizing the $D2$ statistic: Word matches in biological sequences. Stat. Appl. Genet. Mo. B. **8**(1) (2009) Article 43
8. Forêt, S., Wilson, S.R., Burden, C.J.: Empirical distribution of k-word matches in biological sequences. Pattern Recogn. **42** (2009) 539–548
9. Göke, J., Schulz, M., Lasserre, J., Vingron, M.: Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. Bioinformatics **28**(5) (2012) 656–663
10. Jing, J., Wilson, S.R., Burden, C.J.: Weighted k-word matches: A sequence comparison tool for proteins. ANZIAM J. **52 (CTAC2010)** (2011) 172–189
11. Kantorovitz, M.R., Booth, H.S., Burden, C.J., Wilson, S.R.: Asymptotic behavior of k-word matches between two uniformly distributed sequences. J. Appl. Probab. **44** (2006) 788–805
12. Kantorovitz, M.R., Robinson, G.E., Sinha, S.: A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics **23**(13) (2007) i249–55
13. Lippert, R.A., Huang, H., Waterman, M.S.: Distributional regimes for the number of k-word matches between two random sequences. Proc. Natl. Acad. Sci. USA **99**(22) (2002) 13980–9

14. Liu, X., Wan, L., Li, J., Reinert, G., Waterman, M.S., Sun, F.: New powerful statistics for alignment-free sequence comparison under a pattern transfer model. J. Theoret. Biol. **284** (2011) 106–116
15. Percus, J., Percus, O.: The statistics of words on rings. Communications on pure and applied mathematics **59** (2006) 145–160
16. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2012)
17. Reinert, G., Chew, D., Sun, F., Waterman, M.S.: Alignment-free sequence comparison (I): statistics and power. J. Comput. Biol. **16**(12) (2009) 1615–1634
18. Reinert, G., Schbath, S., Waterman, M.: Statistics on words with applications to biological sequences. In Lothaire, M., ed.: Applied Combinatorics on Words. Cambridge University Press (2005)
19. Vinga, S., Almeida, J.: Alignment-free sequence comparison-a review. Bioinformatics **19**(4) (2003) 513–23