The Distribution of Short Word Match Counts Between Markovian Sequences

Conrad J. Burden[1], Paul Leopardi[1] and Sylvain Forêt[2]
[1]*Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia*
[2]*Research School of Biology, Australian National University, Canberra, ACT 0200, Australia*
{*conrad.burden, paul.leopardi, sylvain.foret*}*@anu.edu.au*

Keywords: Word Matches: Biological Sequence Comparison.

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.

1 INTRODUCTION

The D_2 statistic, defined as the number of short word matches of pre-specified length k between two sequences of letters from a finite alphabet $\mathcal{A}$. This statistic (Lippert et al., 2002), and its many variants (Vinga and Almeida, 2003; Reinert et al., 2009; Göke et al., 2012; Jing et al., 2011) 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 (Lippert et al., 2002; Forêt et al., 2006; Kantorovitz et al., 2006; Forêt et al., 2009b; Forêt et al., 2009a; Burden et al., 2012).

Analysis of the k-mer spectra of the genomes of several species provides strong evidence that genomic sequences are more approptriately modelled as having a Markovian dependence (Chor et al., 2009). 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. (Kantorovitz et al., 2007) 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 (Forêt et al., 2009a; Burden et al., 2012) 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

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 ω-*th order Markovian dependence* if

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

for a specified $d^\omega \times d$ matrix M satisfying

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

for all $a_1,\dots,a_\omega,b \in \mathcal{A}$. As a shorthand notation, we will write a string of length ω with an arrow above:

$$\vec{x} = (x_1,\dots x_\omega), \tag{3}$$

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

$$\vec{X}_i = (X_i,\dots X_{i+\omega-1}), \tag{4}$$

Thus Eq.(1) is written more compactly as

$$\text{Prob}\,(X_{i+\omega} = b|\vec{X}_i = \vec{a}) = M(\vec{a};b). \tag{5}$$

Following the notation of ref. (Reinert et al., 2005), define a $d^\omega \times d^\omega$ square matrix $\mathbb{M}$ as

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

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

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

2.1 Markovian Sequences with PBCs

Given an order ω Markovian matrix $\mathbb{M}$, we first attempt to define a periodic random sequence $\mathbf{X} = X_1,X_2\dots,X_n$ of length n via the following algorithm:

Algorithm 1.

Step 0: Choose a probability distribution on the set of strings of length ω:
$\text{Prob}\,(\vec{X}_1 = \vec{x}) = \pi(\vec{x})$, where $0 \le \pi(\vec{x}) \le 1$ and $\sum_{\vec{x}\in\mathcal{A}^\omega} \pi(\vec{x}) = 1$.

Step 1: Generate $\vec{X}_1 = X_1,\dots X_\omega$ from this distribution.

Step 2: Generate $X_{\omega+1},\dots,X_{\omega+n}$ using Eq. (7).

Step 3: If $\vec{X}_{n+1} = \vec{X}_1$, accept the sequence $\mathbf{X} = X_1,X_2\dots,X_n$, otherwise repeat from Step 1 until an accepted sequence is obtained.

Clearly this algorithm entails that

$$\text{Prob}\,(\mathbf{X} = \mathbf{x}) = \frac{\pi(\vec{x}_1)\mathbb{M}(\vec{x}_1,\vec{x}_2),\mathbb{M}(\vec{x}_2,\vec{x}_3)\dots\mathbb{M}(\vec{x}_n,\vec{x}_1)}{\sum_{\vec{u}_1,\dots,\vec{u}_n\in\mathcal{A}^\omega} \pi(\vec{u}_1)\mathbb{M}(\vec{u}_1,\vec{u}_2),\mathbb{M}(\vec{u}_2,\vec{u}_3)\dots\mathbb{M}(\vec{u}_n,\vec{u}_1)}. \tag{8}$$

Now the idea behind PBCs is that there should be no privileged position along the sequence from which to begin numbering. Thus we further impose a condition that the sequence should have no privileged starting point, that is, for each $i = 1,\dots,n$,

$$\text{Prob}\,(\mathbf{X} = x_{i+1}x_{i+2}\dots x_n x_1\dots x_i) = \text{Prob}\,(\mathbf{X} = \mathbf{x}). \tag{9}$$

Eqs. (8) and (9) imply that $\pi(\vec{x}_{i+1}) = \pi(\vec{x}_1)$ for each i and for very sequence $\mathbf{x} \in \mathcal{A}^n$, which can only happen if

$$\pi(\vec{x}) = \frac{1}{d^\omega} \qquad \forall \vec{x} \in \mathcal{A}^\omega. \tag{10}$$

This leads to the following definition:

Definition 2.1. Given a Markovian matrix $\mathbb{M}$ of order ω, a random Markovian sequence with PBCs of length n is one generated by Algorithm 1 with the initial distribution π in Step 0 equal to the uniform distribution Eq. (10).

It follows from Eq. (8) that for a random Markovian sequence $\mathbf{X}$ of length n, the probability of the configuration $\mathbf{x} = (x_1\dots,x_m)$ occuring is

$$\text{Prob}\,(\mathbf{X} = \mathbf{x}) = \frac{\mathbb{M}(\vec{x}_1,\vec{x}_2)\mathbb{M}(\vec{x}_2,\vec{x}_3)\dots\mathbb{M}(\vec{x}_m,\vec{x}_1)}{\text{tr}\,(\mathbb{M}^m)}. \tag{11}$$

3 THE D_2 STATISTIC

Definition 3.1. Given random Markov 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 $\mathbf{Y}$:

$$D_2 = \sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}, \tag{12}$$

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 $\mathbf{Y}$:

$$I_{ij} = \begin{cases} 1 & \text{if } (X_i,\dots,X_{i+k-1}) = (Y_j,\dots,Y_{j+k-1}), \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

3.1 D_2 Mean for Arbitrary ω

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

$$(\mathbb{A}\circ\mathbb{B})_{\alpha\beta} = \mathbb{A}_{\alpha\beta}\mathbb{B}_{\alpha\beta}. \tag{14}$$

The mean of D_2 is

$$E(D_2) = \begin{cases} \dfrac{mn}{\mathrm{tr}\,(\mathbb{M}^m)\mathrm{tr}\,(\mathbb{M}^n)} \times \\ \qquad \mathrm{tr}\,[(\mathbb{M}^{m-k+\omega} \circ \mathbb{M}^{n-k+\omega})(\mathbb{M} \circ \mathbb{M})^{k-\omega}] \\ \qquad\qquad\qquad\qquad \text{if } k \geq \omega, \\ \dfrac{mn}{\mathrm{tr}\,(\mathbb{M}^m)\mathrm{tr}\,(\mathbb{M}^n)} \times \\ \qquad \sum_{u,v \in \mathcal{A}^{k-\omega}} \sum_{w \in \mathcal{A}^k} \mathbb{M}^m((wu),(wu))\mathbb{M}^n((wv),(wv)) \\ \qquad\qquad\qquad\qquad \text{if } k < \omega, \end{cases} \tag{15}$$

where (wu) means the ω-tuple $(w_1 \ldots w_k u_1 \ldots u_{\omega-k})$ and similarly for (wv).

Proof. We have that

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

where

$$\mathrm{Prob}\,(I_{ij} = 1) = \sum_{w \in \mathcal{A}^k} \mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w) \times \mathrm{Prob}\,(Y_j \ldots Y_{j+k-1} = w). \tag{17}$$

To calculate $\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w)$ we must consider separately the cases $k \geq \omega$ and $k < \omega$

Consider first the case where $k \geq \omega$. The required probability is calculated by summing Eq.(11) over all sequences $\mathbf{x}$ such that $(x_i \ldots x_{i+k-1}) = w$. The definition of the matrix $\mathbb{M}$, Eq.(6), ensures that it is sufficient to restrict only those ω-tuples $\vec{x}_i$ located within the word w, since contributions to the sum from any partially overlapping ω-tuples will be zero unless the overlap letters match those of w (see Fig. 1(a)). Thus

$$\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w) = \frac{\mathbb{M}^{m-k+\omega}(\vec{w}_{k-\omega+1}, \vec{w}_1)\mathbb{M}(\vec{w}_1, \vec{w}_2) \ldots \mathbb{M}(\vec{w}_{k-\omega}, \vec{w}_{k-\omega+1})}{\mathrm{tr}\,(\mathbb{M}^m)} \tag{18}$$

where the ω-tuples $\vec{x}_1, \ldots, \vec{x}_{i-1}, \vec{x}_{i+k-\omega+1}, \ldots, \vec{x}_m$ have been summed over. Similarly we have

$$\mathrm{Prob}\,(Y_j \ldots Y_{j+k-1} = w) = \frac{\mathbb{M}^{n-k+\omega}(\vec{w}_{k-\omega+1}, \vec{w}_1)\mathbb{M}(\vec{w}_1, \vec{w}_2) \ldots \mathbb{M}(\vec{w}_{k-\omega}, \vec{w}_{k-\omega+1})}{\mathrm{tr}\,(\mathbb{M}^n)}. \tag{19}$$

The definition of the matrix $\mathbb{M}$ ensures that the sum over the k-word w in Eq. (17) is equivalent to a sum over a set of independent ω-tuples $\vec{w}_1, \ldots, \vec{w}_{k-\omega+1}$.

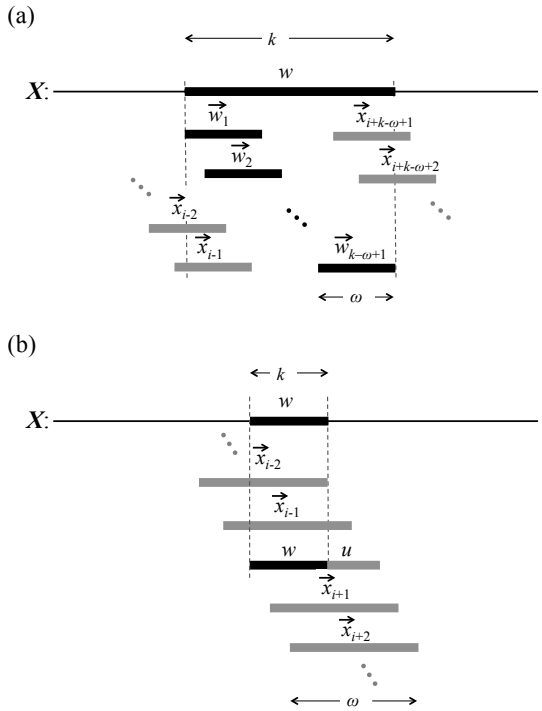

Figure 1: Covering of the sequence $\mathbf{X}$ with ω-tuples for the calculation of $\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w)$ (a) in the case where $k \geq \omega$, and (b) in the case where $k < \omega$.

Thus substituting Eqs. (18) and (19) into Eq. (17) gives

$$\mathrm{Prob}\,(I_{ij} = 1) = \frac{\mathrm{tr}\,[(\mathbb{M}^{m-k+\omega} \circ \mathbb{M}^{n-k+\omega})(\mathbb{M} \circ \mathbb{M})^{k-\omega}]}{\mathrm{tr}\,(\mathbb{M}^m)\mathrm{tr}\,(\mathbb{M}^n)} \tag{20}$$

Eq. (16) then gives the required result for the case $k \geq \omega$.

For the case $k < \omega$, the $\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w)$ is again calculated by summing Eq.(11) over all sequences $\mathbf{x}$ such that $(x_i \ldots x_{i+k-1}) = w$. In this case it is sufficient to restrict any one of ω-tuples overlapping w to equal w on the overlap, and the structure of $\mathbb{M}$ will ensure that only terms in which the other overlapping ω-tuples match w will contribute to the sum. Accordingly set $\vec{x}_i = (w_1 \ldots w_k u_1 \ldots u_{\omega-k})$, where the $u_1 \ldots u_{\omega-k}$ are not fixed (see Fig. 1(b)). Then

$$\mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w) = \frac{1}{\mathrm{tr}\,(\mathbb{M}^m)} \sum_{u \in \mathcal{A}^{\omega-k}} \mathbb{M}^m((wu),(wu)), \tag{21}$$

and similarly

$$\mathrm{Prob}\,(Y_j \ldots Y_{j+k-1} = w) = \frac{1}{\mathrm{tr}\,(\mathbb{M}^n)} \sum_{v \in \mathcal{A}^{\omega-k}} \mathbb{M}^n((wv),(wv)). \tag{22}$$

Substituting these two probabilities into Eqs.(17) and (16) gives the required result. □

3.2 D_2 Variance for $\omega = 1$

The exact variance of D_2 for Markovian sequences with PBCs requires an extensive calculation. The case $\omega > 1$ remains intractable, essentially for the same reason that it was necessary to treat the $k < \omega$ case separately for the above derivation of the mean; namely that probabilities must be calculated for sequence configurations that cannot be covered with ω-tuples conveniently lying within certain specified segments. Nevertheless, the variance can be calculated for the $\omega = 1$ case. Here we give a summary of the result, which is valid for $m, n \geq 2k$. Full technical details of the derivation will be published elsewhere.

We have

$$\mathrm{Var}\,(D_2) = E({D_2}^2) - E(D_2)^2. \tag{23}$$

The second term can be calculated from Eq.(15). The first term is a sum of contributions obtained from Eq.(12) by partitioning a sum over words beginning at positions i and i' in sequence $\mathbf{X}$ and beginning at j and j' in sequence $\mathbf{Y}$,

$$\begin{aligned} E({D_2}^2) &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} E(I_{ij} I_{i'j'}) \\ &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} \mathrm{Prob}\,(I_{ij} = 1, I_{i'j'} = 1) \\ &= V_0 + V_1 + V_2 + V_3 + V_4. \end{aligned} \tag{24}$$

The partitioning reflects the degree of overlap between words in each of the two sequences, and is illustrated in Fig. 2. 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:

$$\begin{aligned} V_0 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ & \sum_{r=0}^{m-2k} \sum_{s=0}^{n-2k} \mathrm{tr}\,\Big[(M^{r+1} \circ M^{s+1})(M \circ M)^{k-1} \times \\ & (M^{m-2k-r+1} \circ M^{n-2k-s+1})(M \circ M)^{k-1}\Big], \end{aligned} \tag{25}$$

$$\begin{aligned} V_1 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ & \Bigg\{ \sum_{s=0}^{n-2k} \Big[\mathrm{tr}\,\{[(M \circ M \circ M)^{k-1} \circ (M^{s+1})^T] \times \\ & (M^{m-k+1} \circ M^{n-2k-s+1})\} \\ & + 2 \sum_{r=1}^{k-1} \mathrm{tr}\,\{(M \circ M)^r \times \\ & [(M \circ M \circ M)^{k-r-1} \circ (M^{s+1})^T] \times \\ & (M \circ M)^r (M^{m-k-r+1} \circ M^{n-2k-s+1})\}\Big] \\ & + \text{the same with } m \text{ and } n \text{ interchanged.} \Bigg\}, \end{aligned} \tag{26}$$

$$\begin{aligned} V_2 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ & \Bigg\{ \mathrm{tr}\,[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1}] \\ & + 2 \sum_{t=1}^{k-1} \mathrm{tr}\,[(M^{m-k-t+1} \circ M^{n-k-t+1}) \times \\ & (M \circ M)^{k+t-1}] \Bigg\}, \end{aligned} \tag{27}$$

$$\begin{aligned} V_3 = {} & \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ & \sum_{t=1}^{k-1} \sum_{s=0}^{t-1} \mathrm{tr}\,\big[(M \circ M)^s Q (M \circ M)^s \times \\ & (M^{m-k-t+1} \circ M^{n-k-s+1} + \\ & M^{n-k-t+1} \circ M^{m-k-s+1})\big], \end{aligned} \tag{28}$$

where

$$Q = \begin{cases} (M^{\circ(2\nu+3)})^{\rho-1} \circ [(M^{\circ(2\nu+1)})^{t-s-\rho+1}]^T & \text{if } \rho > 0, \\ (M^{\circ(2\nu+1)})^{t-s-1} \circ (M^{\circ(2\nu-1)})^T & \text{if } \rho = 0, \end{cases} \tag{29}$$

and

$$\nu = \left\lfloor \frac{k-s}{t-s} \right\rfloor, \qquad \rho = (k-s) \bmod (t-s). \tag{30}$$

Finally,

$$V_4 = \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \sum_{r,t=1}^{k-1} \mathrm{tr}\,U, \tag{31}$$

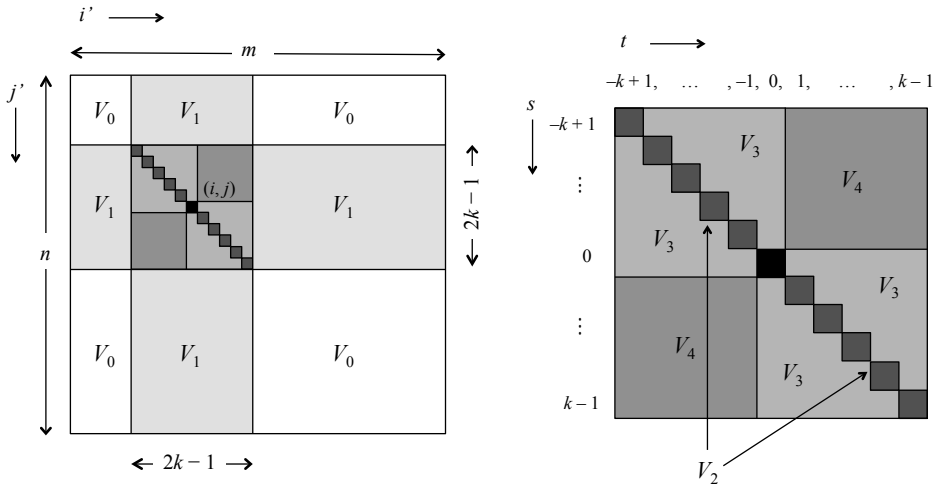

Figure 2: Contributions to Var (D_2) via the sum in Eq. (24). The left-hand diagram shows the (i', j')-plane for a fixed value of (i, j), shown as the black square. The right-hand diagram is an expanded view of the 'accordion' region $-k+1 \le s,t \le k-1$, where $t = i' - i$ and $s = j' - j$ up to PBCs.

where

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

and

$$
\nu = \left\lfloor \frac{k}{r+t} \right\rfloor, \qquad \zeta = k \bmod (r+t). \tag{32}
$$

4 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 Section 3 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 Figure 3. 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 matrices from various genomes have been published, for instance, by Chor et al. (Chor et al., 2009), which can be used in biological applications.

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 Section 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 Figures 4 and 5 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

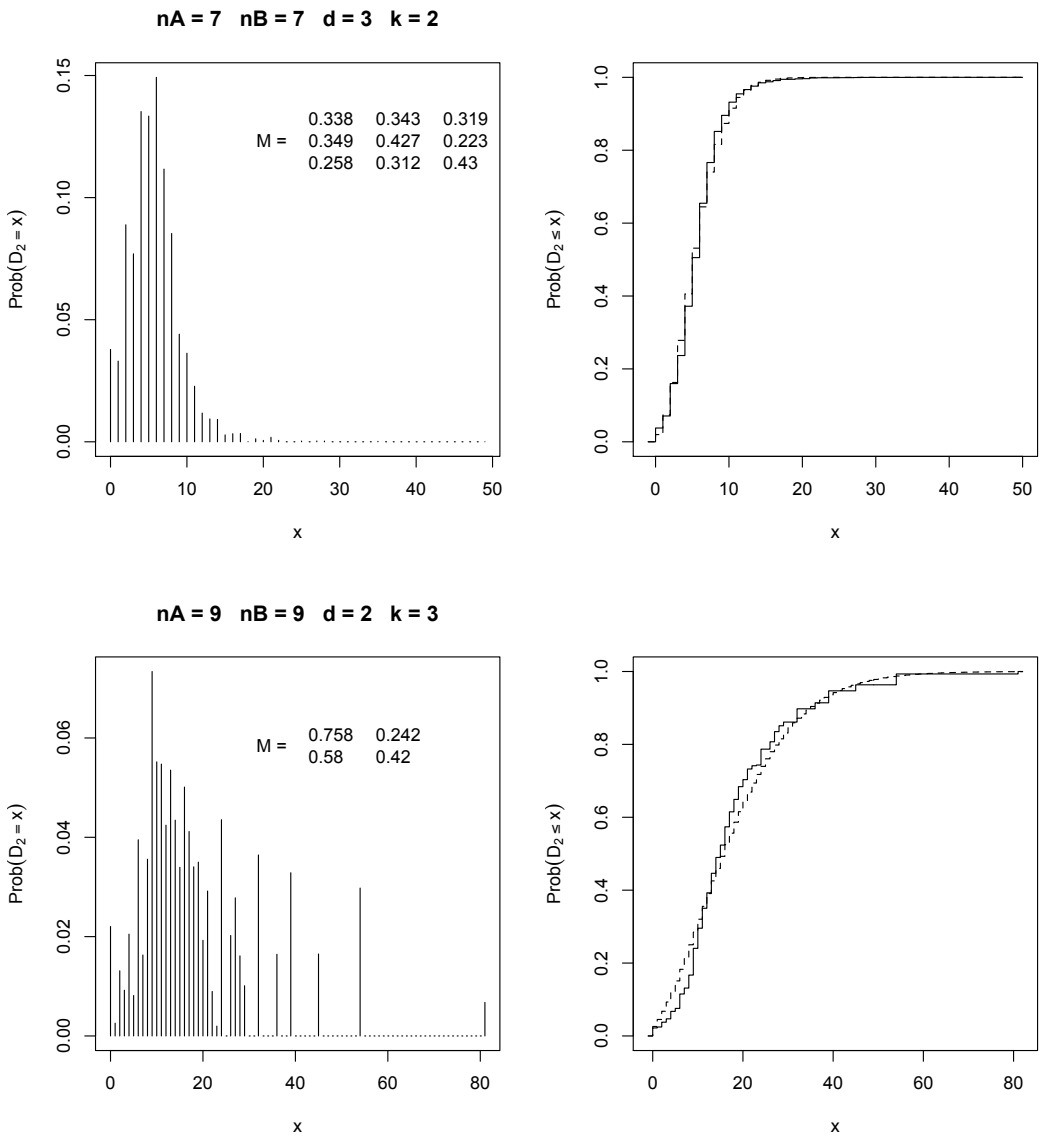

Figure 3: The exact distiribution of the D_2 statistic for short sequences of length m, n and words of length k from an alphabet of size d. The Markov matrix M has been generated randomly in each case. Also shown (dashed curve) is the cumulative distribution of the Pólya-Aeppli distribution with mean and variance set to the theoretical values using the formulae of Section 3.

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 Section 3 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 $m, n \to \infty$. For $m = n$, it has been shown that the limiting distribution is normal in the regime $k < 1/2 \log_b n +$ const. (Burden et al., 2008) and Pólya-Aeppli in the regime $k > 2 \log_b n +$ const. (Lippert et al., 2002). 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 (Reinert et al., 2005).

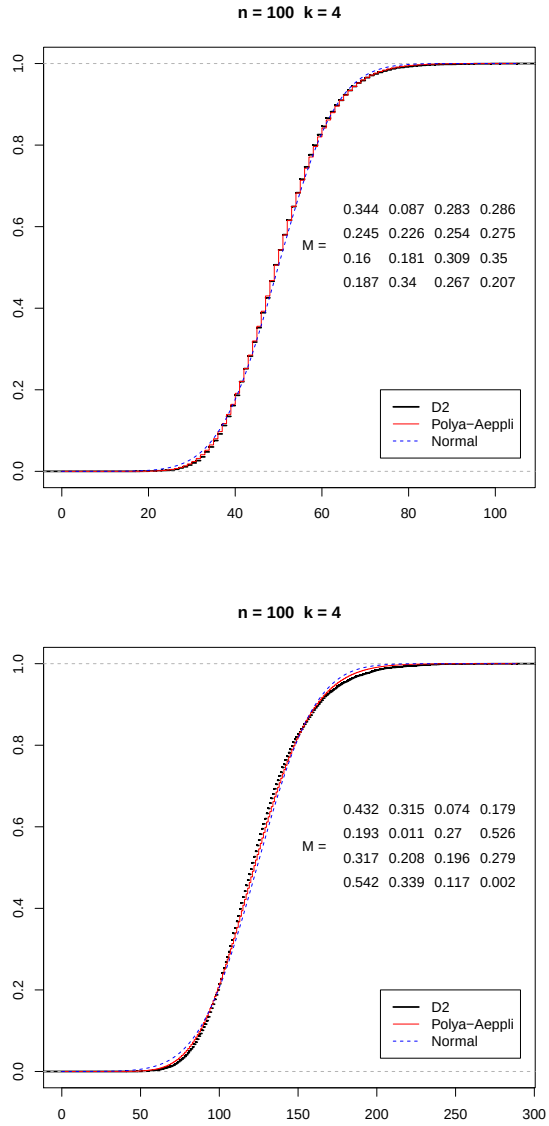

Figure 4: Two examples of empirical cumulative distribution of the D_2 statistic estimated from 10,000 independently generated random sequences of length $m = n = 100$ for words of length $k = 4$ and an alphabet of size $d = 4$. The Markov matrix M has been generated randomly in each case. Also shown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

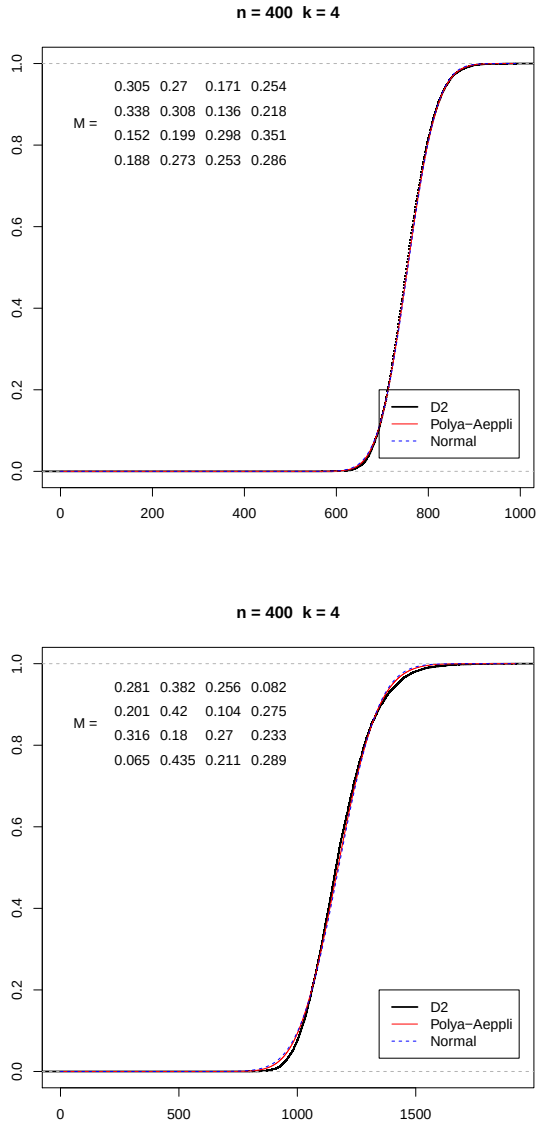

Figure 5: The same as Figure 4, except $m = n = 400$.

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 Figure 3. 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 Fig-

ures 4 and 5 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 Figures 4 and 5. 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.

5 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 (Forêt et al., 2009a; Burden et al., 2012).

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.

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 (Lippert et al., 2002; Reinert et al., 2009). There also exist 'standardised' versions of D_2 (Liu et al., 2011; Göke et al., 2012) designed to account for biases arising from the fact that some words are naturally over-represented, and 'weighted' versions (Jing et al., 2011) 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.

ACKNOWLEDGEMENTS

This work was funded in part by ARC Discovery grants DP0987298 and DP120101422.

REFERENCES

Burden, C. J., Jing, J., and Wilson, S. R. (2012). Alignment-free sequence comparison for biologically realistic sequences of moderate length. *Statistical Applications in Genetics and Molecular Biology*, 11(1):Article 3.

Burden, C. J., Kantorovitz, M. R., and Wilson, S. R. (2008). Approximate word matches between two random sequences. *Annals of Applied Probability*, 18(1):1–21.

Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic DNA k-mer spectra: models and modalities. *Genome Biology*, 10:R108.

Forêt, S., Kantorovitz, M. R., and Burden, C. J. (2006). Asymptotic behaviour and optimal word size for exact and approximate word matches between random sequences. *BMC Bioinformatics*, 7 Suppl 5:S21.

Forêt, S., Wilson, S. R., and Burden, C. J. (2009a). Characterizing the $D2$ statistic: Word matches in biological sequences. *Stat. Appl. Genet. Mo. B.*, 8(1):Article 43.

Forêt, S., Wilson, S. R., and Burden, C. J. (2009b). Empirical distribution of k-word matches in biological sequences. *Pattern Recogn.*, 42:539–548.

Göke, J., Schulz, M., Lasserre, J., and Vingron, M. (2012). Estimation of pairwise sequence similarity of mammalian enhancers with word neighbourhood counts. *Bioinformatics*, 28(5):656–663.

Jing, J., Wilson, S. R., and Burden, C. J. (2011). Weighted k-word matches: A sequence comparison tool for proteins. *ANZIAM J.*, page To appear.

Kantorovitz, M. R., Booth, H. S., Burden, C. J., and Wilson, S. R. (2006). Asymptotic behavior of k-word matches between two uniformly distributed sequences. *J. Appl. Probab.*, 44:788–805.

Kantorovitz, M. R., Robinson, G. E., and Sinha, S. (2007). A statistical method for alignment-free comparison of

regulatory sequences. *Bioinformatics*, 23(13):i249–55.

Lippert, R. A., Huang, H., and Waterman, M. S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA*, 99(22):13980–9.

Liu, X., Wan, L., Li, J., Reinert, G., Waterman, M. S., and Sun, F. (2011). New powerful statistics for alignment-free sequence comparison under a pattern transfer model. *J. Theoret. Biol.*, 284:106–116.

Reinert, G., Chew, D., Sun, F., and Waterman, M. S. (2009). Alignment-free sequence comparison (i): statistics and power. *J. Comput. Biol.*, 16(12):1615–1634.

Reinert, G., Schbath, S., and Waterman, M. (2005). Statistics on words with applications to biological sequences. In Lothaire, M., editor, *Applied Combinatorics on Words*, chapter 6. Cambridge University Press.

Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics*, 19(4):513–23.