# The distribution of word matches between Markovian sequences with periodic boundary conditions

Conrad J. Burden[1], Paul Leopardi[1] and Sylvan Forêt[2]

**Abstract**

Word match counts have traditionally been proposed as an alignment-free measure of similarity for biological sequences. The $D_2$ statistic, which simply counts the number of exact word matches between two sequences, is a useful test bed for developing rigorous mathematical results, which can then be extended to more biologically useful measures. The distributional properties of the $D_2$ statistic under the null hypothesis of identically and independently distributed letters have been studied extensively, but no comprehensive study of the $D_2$ distribution for biologically more realistic higher-order Markovian sequences exists. Here we derive exact formulae for the mean and variance of the $D_2$ statistic for Markovian sequences of any order, and demonstrate through Monte Carlo simulations that the entire distribution is accurately characterised by a Pólya-Aeppli distribution for sequence lengths of biological interest. 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. We also carry out an preliminary comparison between the theoretical Markovian $D_2$ distribution and an empirical $D_2$ distribution from the human genome.

[1] Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia
[2] Research School of Biology, Australian National University, Canberra, ACT 0200, Australia

# 1 Introduction

The $D_2$ statistic is defined as the number of short word matches of a given prespecified length $k$ between two sequences of letters from a finite alphabet $\mathcal{A}$. This statistic was first analysed in the precise form studied below by Lippert, Huang and Waterman [17]. It was motivated by more general statistics based on word counts proposed by Blaisdell [1], and by a statistic defined as a sum over word lengths of weighted inner products of word counts, known as $d^2$ [25, 12]. Such statistics have been proposed as a measures of similarity between biological sequences in cases where the more commonly used alignment methods may not be appropriate. A review of word-based alignment-free sequence comparison measures in existence at or about the time of the Lippert, Huang and Waterman paper (including angle metrics [23, 24] which bear considerable similarity to $D_2$) can be found in ref. [26].

In subsequent developments a number of variants of the $D_2$ statistic have been studied and analysed. A shortcoming of the $D_2$ statistic, first noted in ref. [17], is that the signal of biological sequence similarity one is trying to detect, namely simultaneous over-representation of certain words in both sequences, is masked by the natural variability of word counts in each of the two sequences. This is most likely to be a problem for longer sequences, though perhaps not for sequences of short to moderate length [3]. To address this problem, Reinert et al. [20] introduced centred and standardised statistics, which were demonstrated to have higher power to detect sequence similarity [27]. Other variations on the $D_2$ statistic include allowing word matches up to a certain number of mismatches [4] for detecting regulatory modules [8, 9, 11], and the introduction of weighting factors to acknowledge chemically similar amino acids when studying protein sequences [13, 2].

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. Rigorous results for limiting asymptotic distributions are known for $D_2$ [17, 14] and for $D_2$ with mismatches [4]. Exact analytic formulae exist for the mean [28] and variance [14, 10] of $D_2$, and of the weighted [13] and centred [3] versions of $D_2$. Accurate approximations to distribution of $D_2$ and its variants in terms of gamma and Pólya-Aeppli (or compound Poisson) distributions have been demonstrated via Monte Carlo simulations [10, 9, 13, 3], allowing for fast and practical calculations of approximate p-values under the i.i.d. null hypothesis.

However, 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], possibly up to fifth order. In the current work we extend previous exact analytic results results for the mean, variance and an empirical distribution of $D_2$ for i.i.d. sequences to the case of Markovian sequences. A previous study of this problem, with some approximations, has been carried out by Kantorovitz et al. [15] 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. For i.i.d. sequences we have found imposition of PBCs to be an approximation which works well for biologically realistic sequences [10]. The restriction to periodic sequences simplifies calculations of the mean and variance, enabling an exact analytic formula for the variance for Markovian sequences which is computable to double precision accuracy for arbitrary sequence lengths.

The layout of the paper is as follows. In Section 2 we define Markovian sequences of arbitrary order with periodic boundary conditions in terms of an algorithm for generating such sequences. In Section 3 we define the $D_2$ statistic and derive exact analytic formulae for its mean and variance for Markovian sequences. In Section 4 the accuracy of the mean and variance formulae are checked numerically, and hypothesised asymptotic distributions are demonstrated to provide accurate representations of the complete $D_2$ distribution. These distributions are compared with empirical distributions of $D_2$ from the human genome. Conclusions are drawn in Section 5. Technical details of the derivation of $\mathrm{Var}\,(D_2)$ are given in an appendix, and computer codes for evaluating the mean and variance are given in the the Supplementary Material.

# 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 *$\omega$-th order Markovian dependence* if

$$\mathrm{Prob}\,((X_{i+\omega} = b | (X_i, \ldots, X_{i+\omega-1} = (a_1, \ldots, a_\omega)) = M(a_1, \ldots, a_\omega; b), \qquad (1)$$

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

$$0 \leq M(a_1, \ldots, a_\omega; b) \leq 1; \qquad \sum_{b \in \mathcal{A}} M(a_1, \ldots, a_\omega; b) = 1, \qquad (2)$$

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

$$\vec{x} = (x_1, \ldots x_\omega), \qquad (3)$$

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

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

Thus Eq.(1) is written more compactly as

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

Following the notation of ref. [21], 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 $\omega$ 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 $\omega$: 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

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

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 every 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 $\omega$, a random Markovian sequence with PBCs of length $n$ is one generated by Algorithm 1 with the inital distribution $\pi$ 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 $\mathbf{x} = (x_1 \ldots, x_m)$ occuring is

$$\text{Prob}\,(\mathbf{X} = \mathbf{x}) = \frac{\mathbb{M}(\vec{x_1}, \vec{x_2})\mathbb{M}(\vec{x_2}, \vec{x_3}) \ldots \mathbb{M}(\vec{x}_m, \vec{x}_1)}{\text{tr}\,(\mathbb{M}^m)}. \tag{11}$$

The distribution Eq. (11) has also been proposed by Percus and Percus [18], who made an extensive study of the probability distribution of words on periodic sequences, which they refer to as rings. Our approach is novel in that it gives an algorithm which can be implemented in practice to generate an ensemble of such sequences.

# 3 The $D_2$ statistic

## 3.1 Definition of $D_2$

**Definition 3.1.** Given two random 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

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

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

Two Markovian sequences $\mathbf{X}$ and $\mathbf{Y}$ of order $\omega$ generated by the $d^\omega \times d$ matrix $M$ define a random variable $D_2(k, M)$. By Eq. (7), an equivalent specification of this situation is a pair of first order Markovian sequences $\mathbb{X}$ and $\mathbb{Y}$ consisting of letters of an alphabet of size $d^\omega$ generated by the square matrix $\mathbb{M}$ defined by Eq. (6). The sparse structure of $\mathbb{M}$ ensures that the set of possible sequence pairs $(\mathbb{X}, \mathbb{Y})$ is in one-to-one correspondence with the set of possible sequence pairs $(\mathbf{X}, \mathbf{Y})$, and furthermore, for $k \geq \omega$, a word match of length $k$ between $\mathbf{X}$ and $\mathbf{Y}$ is equivalent to a word match of length $k - \omega + 1$ 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:

$$D_2(k, M) \equiv D_2(k - \omega + 1, \mathbb{M}), \qquad k \geq \omega. \tag{14}$$

### 3.2 $D_2$ mean for arbitrary $\omega$

Below we derive an exact formula for $E(D_2)$ for arbitrary order Markovian sequences. In principle, the mean for any $k \geq \omega$ case can be derived in terms of an equivalent $\omega = 1$ case. However here we give an *ab initio* proof for any $\omega$, noting that, for $k \geq \omega$, the result is consistent with Eq. (14).

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

$$(\mathbb{A} \circ \mathbb{B})_{\alpha\beta} = \mathbb{A}_{\alpha\beta}\mathbb{B}_{\alpha\beta}. \tag{15}$$

The mean of $D_2$ is

$$E(D_2(k, M)) = \begin{cases} \frac{mn}{\mathrm{tr}\,(\mathbb{M}^m)\mathrm{tr}\,(\mathbb{M}^n)} \mathrm{tr}\,[(\mathbb{M}^{m-k+\omega} \circ \mathbb{M}^{n-k+\omega})(\mathbb{M} \circ \mathbb{M})^{k-\omega}] & \text{if } k \geq \omega, \\ \frac{mn}{\mathrm{tr}\,(\mathbb{M}^m)\mathrm{tr}\,(\mathbb{M}^n)} \sum_{u,v \in \mathcal{A}^{\omega-k}} \sum_{w \in \mathcal{A}^k} \mathbb{M}^m((wu),(wu))\mathbb{M}^n((wv),(wv)) & \text{if } k < \omega, \end{cases} \tag{16}$$

where $\mathbb{M}$ is defined by Eq. (6), $(wu)$ means the $\omega$-tuple $(w_1 \dots w_k u_1 \dots u_{\omega-k})$, and similarly for $(wv)$.

*Proof.* We have that

$$E(D_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} E(I_{ij}) = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{Prob}\,(I_{ij} = 1), \tag{17}$$

where

$$\mathrm{Prob}\,(I_{ij} = 1) = \sum_{w \in \mathcal{A}^k} \mathrm{Prob}\,(X_i \dots X_{i+k-1} = w)\mathrm{Prob}\,(Y_j \dots Y_{j+k-1} = w). \tag{18}$$

To calculate $\mathrm{Prob}\,(X_i \dots X_{i+k-1} = w)$ 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 $\mathbf{x}$ subject to the restriction 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 $\omega$-tuples $\vec{x}_i$ located within the word $w$, since contributions to the sum from any partially overlapping $\omega$-tupes will be zero unless the overlap letters letters match those of $w$ (see Fig. 1(a)). Thus

$$\text{Prob}\,(X_i \ldots X_{i+k-1} = w) = \frac{\mathbb{M}^{m-k+\omega}(\vec{w}_{k-\omega+1}, \vec{w}_1)\mathbb{M}(\vec{w}_1, \vec{w}_2) \ldots \mathbb{M}(\vec{w}_{k-\omega}, \vec{w}_{k-\omega+1})}{\text{tr}\,(\mathbb{M}^m)}. \tag{19}$$

where the $\omega$-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

$$\text{Prob}\,(Y_j \ldots Y_{j+k-1} = w) = \frac{\mathbb{M}^{n-k+\omega}(\vec{w}_{k-\omega+1}, \vec{w}_1)\mathbb{M}(\vec{w}_1, \vec{w}_2) \ldots \mathbb{M}(\vec{w}_{k-\omega}, \vec{w}_{k-\omega+1})}{\text{tr}\,(\mathbb{M}^n)}. \tag{20}$$

The definition Eq. (6) of the matrix $\mathbb{M}$ ensures that the sum over the $k$-word $w$ in Eq. (18) is equivalent to a sum over a set of independent $\omega$-tuples $\vec{w}_1, \ldots, \vec{w}_{k-\omega+1}$. Thus subsitituting Eqs. (19) and (20) into Eq. (18) gives

$$\text{Prob}\,(I_{ij} = 1) = \frac{\text{tr}\,[(\mathbb{M}^{m-k+\omega} \circ \mathbb{M}^{n-k+\omega})(\mathbb{M} \circ \mathbb{M})^{k-\omega}]}{\text{tr}\,(\mathbb{M}^m)\text{tr}\,(\mathbb{M}^n)} \tag{21}$$

Eq. (17) then gives the required result for the case $k \geq \omega$.

For the case $k < \omega$, the 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 the $\omega$-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 $\omega$-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

$$\text{Prob}\,(X_i \ldots X_{i+k-1} = w) = \frac{1}{\text{tr}\,(\mathbb{M}^m)} \sum_{u \in \mathcal{A}^{\omega-k}} \mathbb{M}^m((wu), (wu)), \tag{22}$$

and similarly

$$\text{Prob}\,(Y_j \ldots Y_{j+k-1} = w) = \frac{1}{\text{tr}\,(\mathbb{M}^n)} \sum_{v \in \mathcal{A}^{\omega-k}} \mathbb{M}^n((wv), (wv)). \tag{23}$$

Substituting these two probabilities into Eqs.(18) and (17) gives the required result. □

## 3.3 $D_2$ variance for $k \geq \omega$

For $k \geq \omega$, Eq. (14) ensures that any $\omega > 1$ case can be reduced to an equivalent $\omega = 1$ case via the relation

$$\text{Var}\,(D_2(k, M)) = \text{Var}\,(D_2(k - \omega + 1, \mathbb{M})), \qquad k \geq \omega, \tag{24}$$

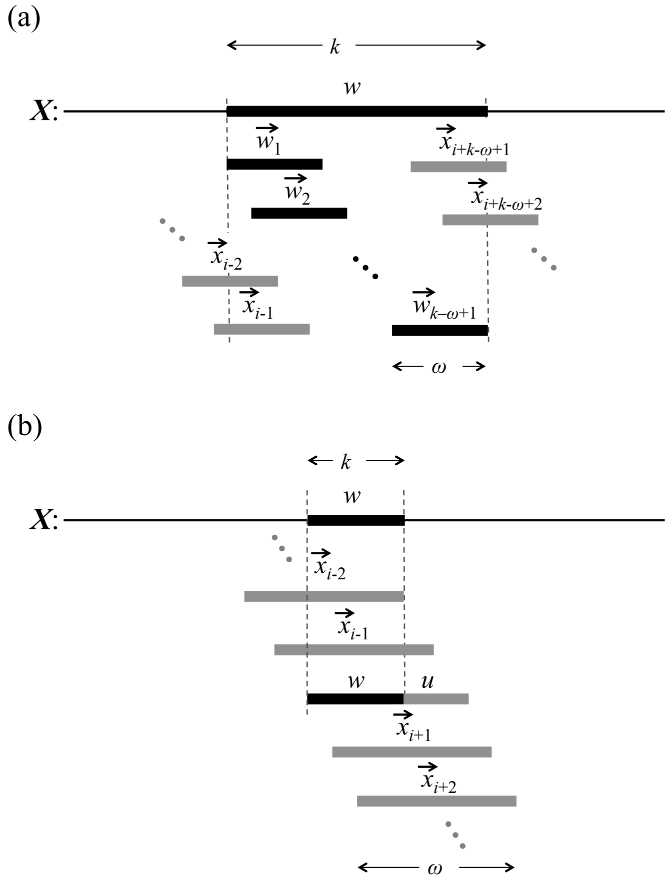

Figure 1: Covering of the sequence $\mathbf{X}$ with $\omega$-mers for the calculation of $\text{Prob}\,(X_i \ldots X_{i+k-1} = w)$ (a) in the case where $k \geq \omega$, and (b) in the case where $k < \omega$.

where $\mathbb{M}$ is a square first order Markov matrix. Even for $\omega = 1$ the exact variance of $D_2$ for Markovian sequences with PBCs requires an extensive calculation. Here we give a summary of the $\omega = 1$ result, and leave the technical details of the derivation to the Appendix. The case $k < \omega$ remains intractable.

For the remainder of this section we take $M$ to be a square $d \times d$ first order Markov matrix. We have

$$\text{Var}\,(D_2) = E({D_2}^2) - E(D_2)^2. \qquad (25)$$

The second term can be calculated from Eq.(16). The first term is a sum of contributions obtained from Eq.(12) by partitioning a sum over words beginning

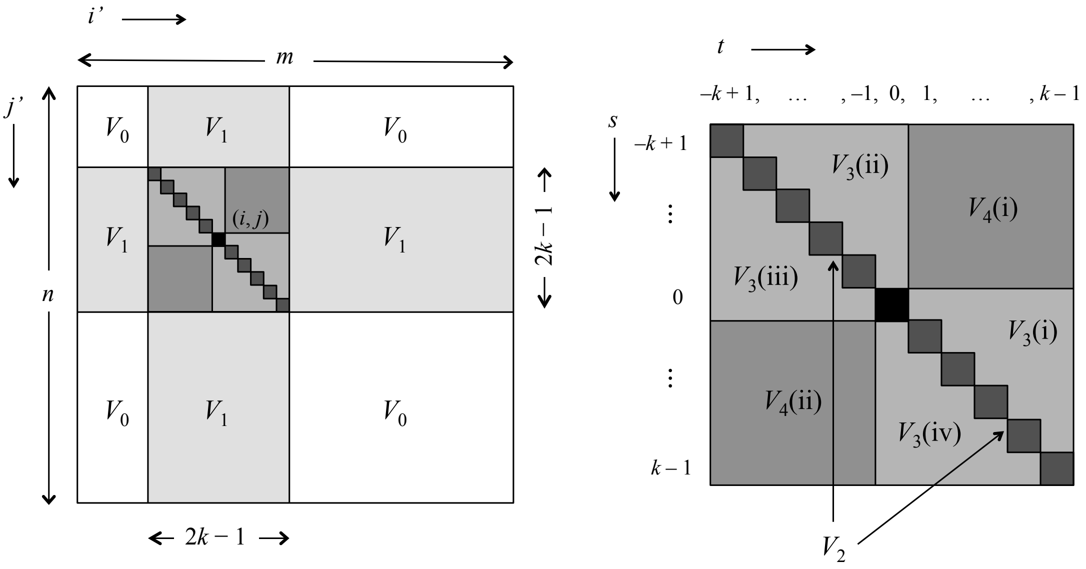

Figure 2: Contributions to Var $(D_2)$ via the sum in Eq. (26). The left-hand diagram shows the $(i', j')$-plane for a fixed value of $(i, j)$, shown as the black square. The right-hand diagram is an expanded view of the ‘accordion’ region $-k+1 \leq s, t \leq k-1$, where $t = i' - i$ and $s = j' - j$ up to PBCs (see Eqs. (40) and (41)).

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

$$\begin{aligned} E({D_2}^2) &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} E(I_{ij} I_{i'j'}) \\ &= \sum_{i,i'=1}^{m} \sum_{j,j'=1}^{n} \text{Prob}\,(I_{ij} = 1, I_{i'j'} = 1) \\ &= V_0 + V_1 + V_2 + V_3 + V_4. \end{aligned} \tag{26}$$

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 Hadmard product of $q$ factors, $M \circ \ldots \circ M$, using the shorthand notation $M^{\circ q}$. With this notation, the contributions to $E({D_2}^2)$ are:

$$V_0 = \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \sum_{r=0}^{m-2k} \sum_{s=0}^{n-2k} \mathrm{tr}\,\left[(M^{r+1} \circ M^{s+1})(M \circ M)^{k-1}(M^{m-2k-r+1} \circ M^{n-2k-s+1})(M \circ M)^{k-1}\right], \tag{27}$$

$$\begin{aligned} V_1 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ & \left\{ \sum_{s=0}^{n-2k} \Big[ \mathrm{tr}\,\{[(M \circ M \circ M)^{k-1} \circ (M^{s+1})^T](M^{m-k+1} \circ M^{n-2k-s+1})\} \right. \\ & \quad + 2\sum_{r=1}^{k-1} \mathrm{tr}\,\{(M \circ M)^r [(M \circ M \circ M)^{k-r-1} \circ (M^{s+1})^T] \times \\ & \qquad (M \circ M)^r (M^{m-k-r+1} \circ M^{n-2k-s+1})\} \Big] \\ & \left. + \text{ the same with } m \text{ and } n \text{ interchanged.} \right\}, \end{aligned} \tag{28}$$

$$\begin{aligned} V_2 = {} & \frac{mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ & \left\{ \mathrm{tr}\,[(M^{m-k+1} \circ M^{n-k+1})(M \circ M)^{k-1}] \right. \\ & \left. \quad + 2\sum_{t=1}^{k-1} \mathrm{tr}\,[(M^{m-k-t+1} \circ M^{n-k-t+1})(M \circ M)^{k+t-1}] \right\}, \end{aligned} \tag{29}$$

$$\begin{aligned} V_3 \quad = \quad & \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \sum_{t=1}^{k-1} \sum_{s=0}^{t-1} \mathrm{tr}\,\Big[(M \circ M)^s Q (M \circ M)^s \\ & \times (M^{m-k-t+1} \circ M^{n-k-s+1} + M^{n-k-t+1} \circ M^{m-k-s+1})\Big], \end{aligned} \tag{30}$$

where

$$\nu = \left\lfloor \frac{k-s}{t-s} \right\rfloor, \qquad \rho = (k-s) \bmod (t-s), \tag{31}$$

and

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

Finally,

$$V_4 = \frac{2mn}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \sum_{r,t=1}^{k-1} \mathrm{tr}\, U, \tag{33}$$

where

$$\nu = \left\lfloor \frac{k}{r+t} \right\rfloor, \qquad \zeta = k \bmod (r+t), \tag{34}$$

and

$$U = \begin{cases} \left\{(M^{\circ(2\nu+1)})^{t-1} \circ (M^{m-k-t+1})^T\right\} M^{\circ 2\nu} & \\ \quad \times \left\{(M^{\circ(2\nu+1)})^{r-1} \circ (M^{n-k-r+1})^T\right\} M^{\circ 2\nu} & \text{if } \zeta = 0, \\ \left\{(M^{\circ(2\nu+1)})^{r-\zeta+1} \circ M^{m-k-t+1}\right\} (M^{\circ(2\nu+2)})^{\zeta-1} & \\ \quad \times \left\{(M^{\circ(2\nu+1)})^{t-\zeta+1} \circ M^{n-k-r+1}\right\} (M^{\circ(2\nu+2)})^{\zeta-1} & \text{if } 0 < \zeta \le r, t, \\ \left\{(M^{\circ(2\nu+3)})^{\zeta-r-1} \circ (M^{m-k-t+1})^T\right\} (M^{\circ(2\nu+2)})^{r} & \\ \quad \times \left\{(M^{\circ(2\nu+1)})^{t-\zeta+1} \circ M^{n-k-r+1}\right\} (M^{\circ(2\nu+2)})^{r} & \text{if } r < \zeta \le t, \\ \text{as above with } m \text{ and } n \text{ interchanged} & \\ \qquad \text{and } r \text{ and } t \text{ interchanged} & \text{if } t < \zeta \le r, \\ \left\{(M^{\circ(2\nu+3)})^{\zeta-r-1} \circ (M^{m-k-t+1})^T\right\} (M^{\circ(2\nu+2)})^{t+r-\zeta+1} & \\ \quad \times \left\{(M^{\circ(2\nu+3)})^{\zeta-t-1} \circ (M^{n-k-r+1})^T\right\} (M^{\circ(2\nu+2)})^{t+r-\zeta+1} & \text{if } r, t < \zeta. \end{cases} \tag{35}$$

A full derivation of these contributions is given in the appendix.

# 4 Numerical Results

## 4.1 Computer implementation of the mean and variance.

In the supplementary material we provide an R implementation [19] of $E(D_2(k, M))$ for arbitrary $k$ and of $\mathrm{Var}\,(D_2(k, M))$ for $k \ge \omega$ using the formulae derived above. The $k > \omega$ means and variances are calculated by reducing the problem to the equivalent $\omega = 1$ calculation with effective $d^\omega \times d^\omega$ Markov matrix $\mathbb{M}$ and effective word length $k - \omega + 1$ (see Eq. (24)).

The computationally expensive parts of the computation of $\mathrm{Var}\,(D_2)$ are the sums over $r$ and $s$ occurring in Eqs. (27) and the first line of Eq. (28). These sums are facilitated for large sequence lengths $m$ and $n$ by storing powers of $\mathbb{M}$ out to convergence and by making use of the fact that the summand is essentially constant over parts of the domain of summation for which these matrix powers have converged. Although the programs are not yet fully optimised, they calculate

Var $(D_2)$ in about 30 seconds on a standard laptop computer for an alphabet of size $d = 4$, Markovian order $\omega = 3$, word lengths up to $k = 20$ and arbitrarily large sequence lengths $m$ and $n$. The variance program slows considerably for higher order Markov models as the size of $\mathbb{M}$ grows exponentially with $\omega$. Considerable gains are possible for the case $k = \omega$, as the terms $V_2$, $V_3$ and $V_4$ in the equivalent $\omega = 1$ calculation are automatically zero, and double sum in the term $V_0$ can be computed more efficiently by using the identity

$$\sum_{r=0}^{m-2}\sum_{s=0}^{n-2} \operatorname{tr}\left[(\mathbb{M}^{r+1} \circ \mathbb{M}^{s+1})(\mathbb{M}^{m-r-1} \circ \mathbb{M}^{n-s-1})\right] = \\ \sum_{i,j=1}^{d^\omega} \left(\sum_{r=0}^{m-2} (\mathbb{M}^{r+1})_{ij} (\mathbb{M}^{m-r-1})_{ji}\right) \left(\sum_{s=0}^{n-2} (\mathbb{M}^{s+1})_{ij} (\mathbb{M}^{m-s-1})_{ji}\right). \tag{36}$$

Also included in the supplementary material is a test program which generates the complete distribution of $D_2$ for short sequences for a randomly chosen Markovian model created by by choosing each matrix element from a uniform distribution on the interval $[0, 1]$ and then normalising each row sum to 1. Using this program we have confirmed the accuracy of the above mean and variance formulae to 13 significant figures for sequences up to length $m = n = 10$ for various values of values of the alphabet size $d$, Markov order $\omega$ and word length $k$. Two examples of the exact $D_2$ distribution for short sequences are shown in Figure 3.

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 $k < 1/2 \log_b n + \text{const.}$ [4] and Pólya-Aeppli in the regime $k > 2 \log_b n + \text{const.}$ [17]. 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 [21]. In earlier work on i.i.d. sequences [3], we have found in general that, for simulations of $D_2$ for moderate to long sequences, the gamma distribution provides a good interpolation between the normal and Pólya-Aeppli regimes. 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 added the Pólya-Aeppli distribution function with the mean and variance adjusted to their theoretical values to the plots, and find it to be a surprisingly close fit. Disagreement arises

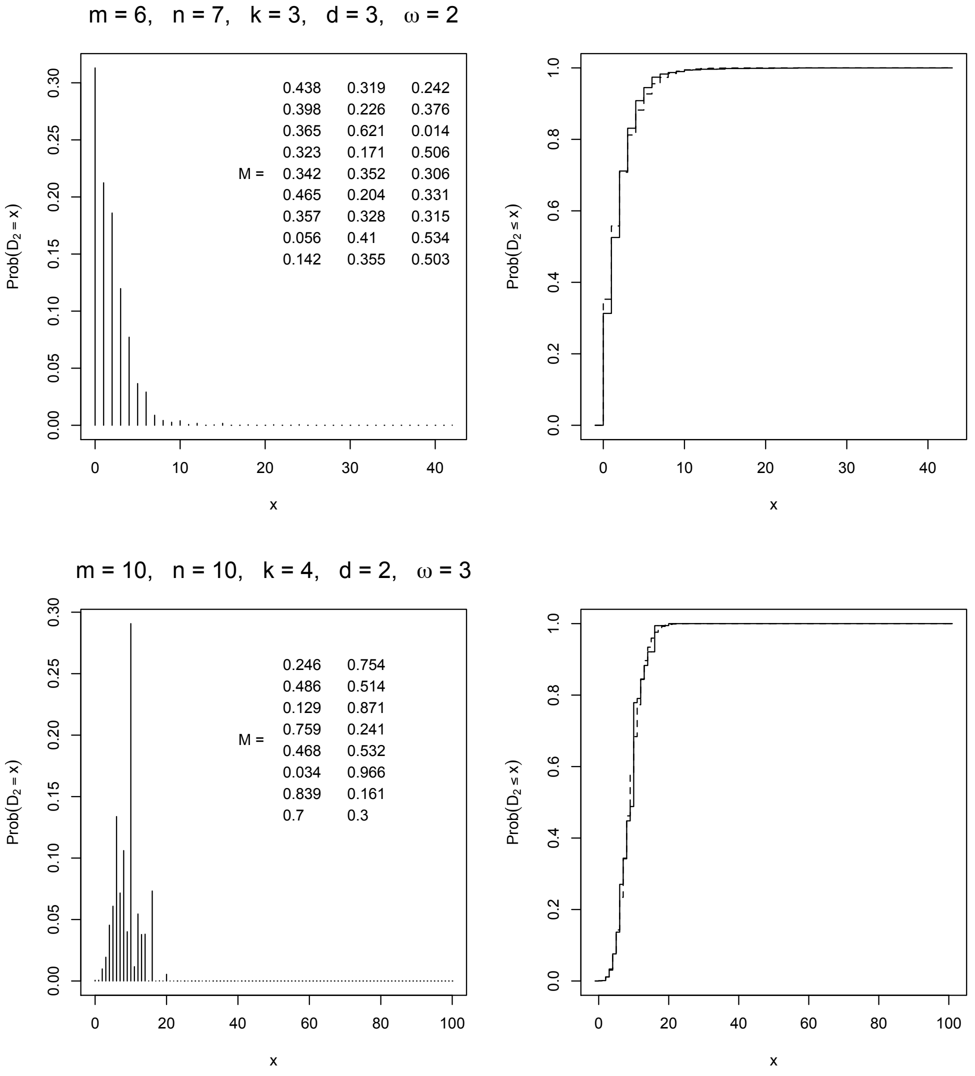

Figure 3: The exact distribution of the $D_2$ statistic for short sequences of length $m$, $n$ and words of length $k$ from a Markov models of order $\omega$ and alphabet of size $d$. The Markov matrix $M$ has been generated randomly in each case. Also shown (dashed curve) is the cumulative distribution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

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 $\infty$, albeit with very low probability).

|  |  | Order |  |  |  |
|---|---|---|---|---|---|
|  |  | 0 | 1 | 2 | 3 |
| Mean | Lower 95% | 18.84 | 24.70 | 27.66 | 28.84 |
|  | Theoretical | 18.92 | 24.73 | 27.79 | 28.97 |
|  | Empirical | 18.95 | 24.83 | 27.80 | 29.00 |
|  | Upper 95% | 19.07 | 24.96 | 27.95 | 29.15 |
| Variance | Lower 95% | 32.89 | 43.23 | 53.06 | 59.01 |
|  | Theoretical | 33.24 | 44.69 | 55.56 | 60.53 |
|  | Empirical | 33.81 | 44.44 | 54.54 | 60.65 |
|  | Upper 95% | 34.77 | 45.70 | 56.09 | 62.37 |

Table 1: Mean and variance of $D_2$ calculated from the theoretical formulae derived in Section 3, and estimated from synthetically generated data (10 000 sequence pairs) for Markov models of order $\omega = 0, 1, 2$ and 3 using Markov matrices estimated from human chromosome 1 [6]. Word length $k = 8$, alphabet size $d = 4$, sequence lengths $m = n = 1000$.

## 4.2 Comparison with simulated distributions

For sequences of realistic biological length composed of the 4-letter nucleotide alphabet it is necessary to resort to Monte Carlo simulations to investigate the $D_2$ probability distribution.

We used a combination of R scripts and the SAFT program (Sequence Alignment-Free Tool, under development) to further verify the formulae for the mean and variance, and to compare the empirical distribution of the $D_2$ statistic with the conjectured asymptotic normal, Pólya-Aeppli and gamma distributions. For this purpose, as well as using randomly generated Markov matrices, we used matrices obtained from DNA sequences occurring in nature. The supplementary material to Chor et al. [5] contains maximum likelihood estimates of Markov matrices for a number of species and for different regions within the human genome. As an example, we used the Markov matrices for human chromosome 1, with Markov orders 0, 1, 2 and 3 [6]. For each of these matrices, we used an R script that implements Algorithm 1, using the built-in random number generator of R, via the function `sample.int()`, to generate 20 000 sequences of length 1000, arranged as 10 000 pairs of cyclic sequences. The SAFT program calculated the $D_2$ statistic for each of these 10 000 pairs. We then used a second R script, based on the code in the supplementary material, to compare the mean and variance of the empirical distribution of the $D_2$ statistic with the theoretical values given by (16) and (25) to (35), to compare the empirical cumulative distribution of the $D_2$ statistic with known distributions, and to plot results.

Table 1 presents the results for the mean and variance for Markov orders 0 to 3. For the mean, the row labelled “Theoretical” is calculated from the corresponding Markov matrix using formula (16), the row labelled “Empirical” is estimated from the 10 000 values of $D_2$ obtained via SAFT, and the rows labelled “Lower 95%” and “Upper 95%” are obtained from the confidence interval returned by the R function `t.test()` that implements Student’s $t$-test. For the variance $\sigma^2$ , the row labelled “Theoretical” is calculated from the corresponding Markov matrix using formulae (25) to (35), the row labelled “Empirical” is estimated from the 10 000 values of $D_2$ obtained via SAFT, and the rows labelled “Lower 95%” and “Upper 95%” are obtained via the $\chi^2$ distribution, using the R quantile function `qchisq` and the inequality [22, (5.10.2), p. 74]

$$(N-1)s^2/\chi^2_{0.025} \leqslant \sigma^2 \leqslant (N-1)s^2/\chi^2_{0.975},$$

where $N = 10\,000$ in this case, and $s^2$ is the sample variance. In these and in a number of other simulations we have performed (data not shown) we 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.

As a general rule, and as can be seen from Table 1, we observe that both the mean and variance of $D_2$ increase markedly as the Markov order increases for fixed word length $k$ and sequence lengths $m$ and $n$. The difference between the empirical cumulative distribution functions for the different Markov orders for the parameters of Table 1 is further illustrated in Figure 4.

We compared the empirical distribution of $D_2$ for each Markov order with conjectured asymptotic distributions based on the theoretical mean and variance calculated via (16) and (25) to (35). For Markov order 3, this is illustrated by Figure 5. Here the cumulative Gamma and Normal distributions are plotted using the built-in R functions `pgamma()` and `pnorm()`, respectively, and the cumulative Pólya-Aeppli distribution is plotted using the function `pPolyaAeppli()` included in the Supplementary Materials. We observe that for these parameter values the three conjectured distributions do not differ greatly from one another, though the Pólya-Aeppli clearly gives the best fit, particularly in the important tail of the distribution relevant to estimating p-values. This trend is also observed for other parameter values. For parameters leading to large values of $E(D_2)$, the continuous normal and gamma distributions are more readily computable than the Pólya-Aeppli, and of these two the gamma is invariably observed to give a better fit.

## 4.3 Comparison with chromosomal DNA

Ultimately one hopes to use $D_2$ or similarly defined statistics as an alignment-free tool to assess the relatedness of sections of genomic sequences. To this end, it

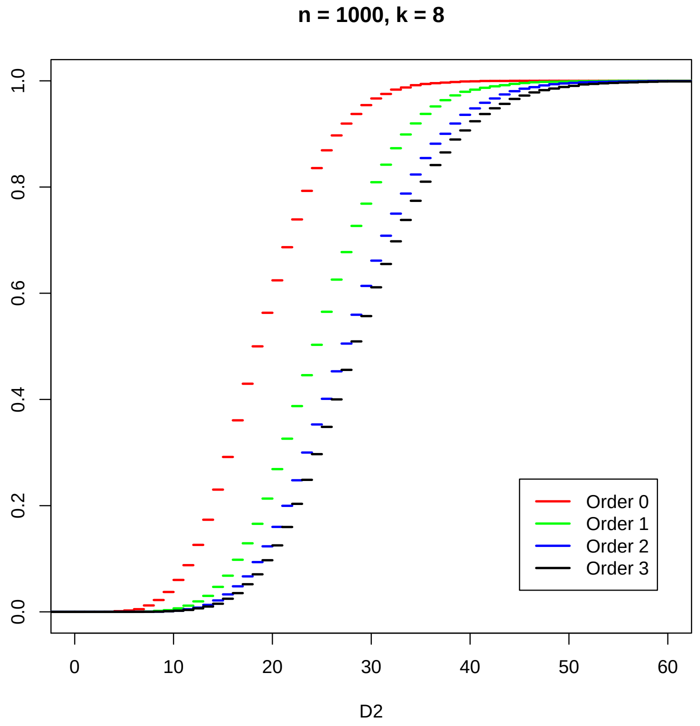

Figure 4: Comparison of empirical cumulative distribution function for simulated $D_2$ using Markov matrices for human chromosome 1 from Chor et al. [6], for orders $\omega = 0, 1, 2$ and 3. 10 000 pairs per order, word length $k = 8$, alphabet size $d = 4$, sequence lengths $m = n = 1000$.

is helpful to know to what extent genomic sequences can be modelled as Markovian sequences for the purpose of defining a null-hypothesis distribution for the $D_2$ statistic. With this in mind, we have performed some exploratory comparisons between the $D_2$ distributions obtained via simulating the Markov processes using maximum likelihood estimates of Markov matrices and the $D_2$ distribution obtained by sampling original DNA data, for example the DNA sequence from human chromosome 1 [29].

Figure 6 illustrates the comparison between the density of Gamma distributions adjusted to the theoretical means and variances from Section 3 with the empirical density of the $D_2$ distribution obtained from chromosome 1. To obtain the empirical density we took the soft masked DNA sequence for human chromosome 1 from Ensembl [29], and took uniform random samples of subsequences of length 1000, according to Knuth's Algorithm S [16, Section 3.4.2], but avoiding

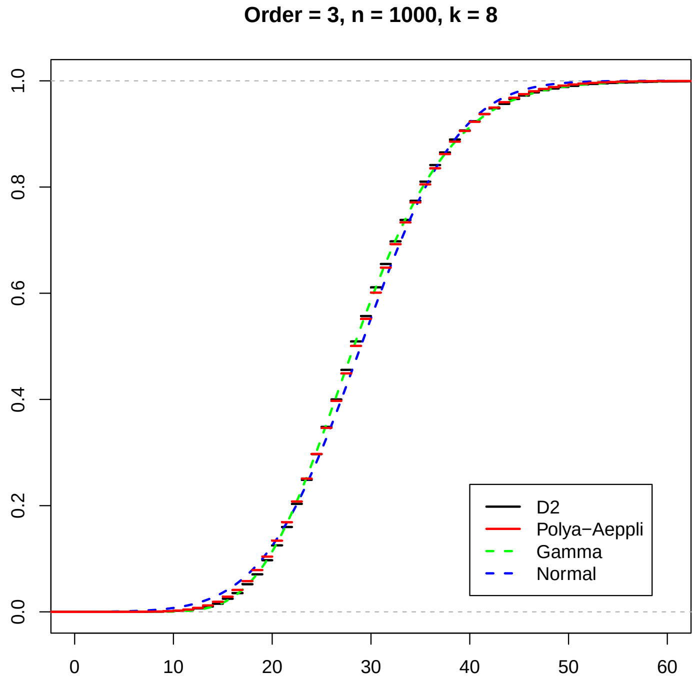

Figure 5: Comparison of Pólya-Aeppli, Gamma and Normal cumulative distributions with empirical cumulative distribution function for simulated $D_2$ using Markov order 3 matrix for human chromosome 1 from Chor et al. [6]. 10 000 pairs, word length $k = 8$, sequence lengths $m = n = 1000$.

all ambiguous and masked regions. Ensembl's masking removes repetitive regions including tandem repeats. This data source and procedure correspond to those described by Chor et al. [5] except that the Markov matrices have been estimated from Ensembl's 'soft-masked' sequences with the repeat regions (i.e. the lower case letters) ignored, whereas Chor et al. include the repeat regions. We find that, as expected, including the repeat regions leads to a skewed $D_2$ distribution with an extremelty heavy right-hand tail.

The resulting sample mean and variance, together with the theoretical values, are shown in Table 2. In general, agreement between the Markovian model and the empirical distribution improve as the Markovian order increases. For higher orders the Markovian mean overshoots slightly. The Markovian variance, on the other hand, severely underestimates the empirical variance at any order. This is consistent with earlier observations [7] that genomic word count distributions

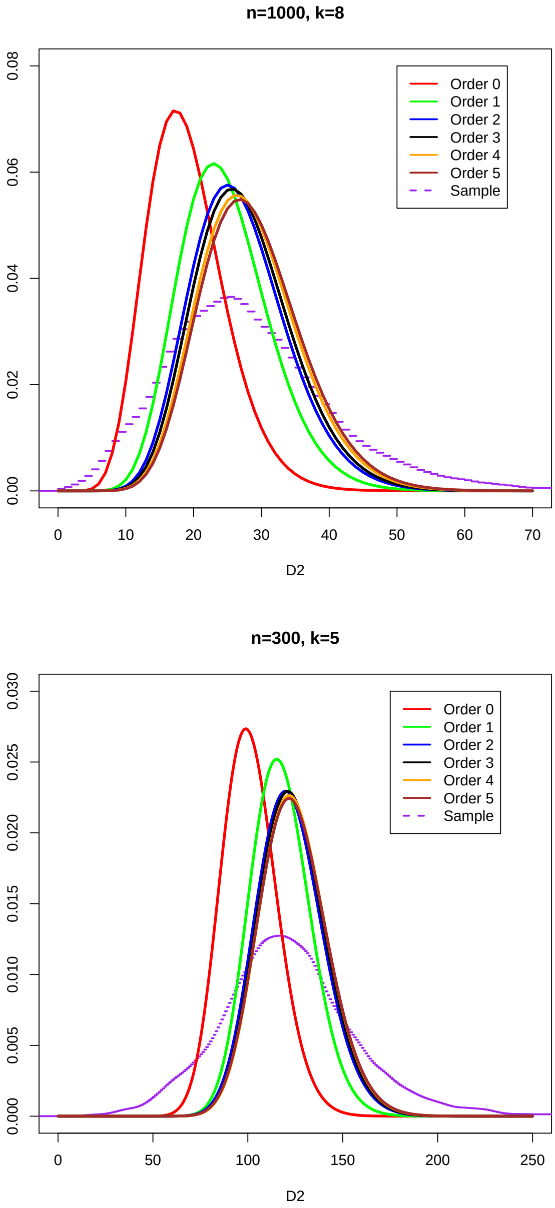

Figure 6: Density of the empirical distribution of $D_2$ from human chromosome 1 sample data [29] compared to Gamma distributions with calculated mean and variance, based on Markov models of various orders $\omega$ [6]. 10 000 sample pairs, word length $k = 8$ and sequences lengths $m = n = 1000$ (upper plot), word length $k = 5$ and sequences lengths $m = n = 300$ (lower plot).

| | Theoretical values | | | | | | Sample |
|---|---|---|---|---|---|---|---|
| | $\omega = 0$ | 1 | 2 | 3 | 4 | 5 | estimate |
| | $m = n = 1000, \quad k = 8$ | | | | | | |
| Mean | 19.08 | 24.74 | 26.90 | 27.55 | 28.30 | 28.74 | 27.66 |
| Variance | 33.58 | 44.62 | 50.99 | 52.19 | 54.37 | 56.01 | 181.1 |
| Std. Dev. | 5.795 | 6.680 | 7.141 | 7.224 | 7.373 | 7.484 | 13.46 |
| | $m = n = 300, \quad k = 5$ | | | | | | |
| Mean | 101.1 | 117.4 | 122.4 | 123.5 | 124.3 | 124.3 | 120.7 |
| Variance | 216.6 | 254.4 | 307.5 | 307.8 | 315.6 | 321.9 | 1,258. |
| Std. Dev. | 14.71 | 15.95 | 17.53 | 17.55 | 17.77 | 17.94 | 35.47 |

Table 2: Empirical estimates of the mean and variance of $D_2$ from human chromosome 1 sample data [29] (right hand column), compared to the theoretical mean and variance based on Markov models of various orders using estimated Markov matrices for human chromosome 1.

tend to have heavier tails than that predicted by Markovian models, or, to put it another way, certain $k$-mers are 'under-' or 'over-represented' within genomes.

Note also that the Markovian plots in Figure 6 suggest that $\omega = k$ may be in some sense a limiting case. Recall that the formula for the mean takes a different form for $\omega > k$ (see Eq. (16)) and that the formula derived for the variance is only valid for $\omega \leq k$ and remains intractable for $\omega > k$. We suspect that this is related to the fact that, for sufficiently long sequences, $\omega$-mer frequencies are determined by the stationary eigenvector of the Markov matrix, and that the statistics of $k$-mers for $k < \omega$ is implicit with the statistics of $\omega$-mers.

# 5 Discussion

The primary purpose of this paper is to demonstrate that it is possible to construct accurate representations of the distribution of the $D_2$ statistic under the null hypothesis of Markovian sequences without the need to resort to computationally expensive Monte Carlo simulations or to asymptotic approximations valid only when $\log n >> k$. We have demonstrated that, for sequences of moderate length of up to only a few hundred letters, and for which $\log n \approx k$, the Pólya-Aeppli distribution with parameters determined by the exact formulae for the mean and variance developed herein accurately represents the true Markovian $D_2$ distribution of any order (see Fig. 5). For comparatively longer sequences with higher $E(D_2)$, for which evaluating the Pólya-Aeppli distribution may be slow, the gamma distribution provides an acceptable approximation which is more accurate than the normal distribution.

It is known that the $D_2$ statistic itself, if used directly as a measure of sequence similarity, may perform poorly as the signal of over-representation of the same words in the query and target sequences is masked by the natural variability of word counts in each of the two sequences [17]. Variations on the theme of the $D_2$ statistic, such as the weighted, centred statistic $D_2^*$ studied in ref. [20] have been developed to circumvent this problem. In earlier work we have extended calculations of the exact mean and variance for i.i.d. sequences to weighted and centred versions of $D_2$ [2, 3], and expect that the analogous calculation for Markovian sequences will be entirely feasible.

The secondary purpose of this paper is a preliminary comparison of the Markovian $D_2$ distribution with an empirical genomic $D_2$ distribution. As a test example we have considered the empirical distribution of the $D_2$ statistic between randomly chosen segments of a single human chromosome, avoiding highly repetitive parts of the chromosome such as stretches of tandem repeats. In general, we find that the empirical distribution has much heavier tails than the Markovian distribution of any order up to $\omega = 5$ (see Fig. 6). We interpret this as a signal that the chromosome, taken as a whole, contains a number of strongly over- and under-represented $k$-mers, relative to a Markovian sequence. Thus one is tempted to conclude that a Markov model will tend to overestimate significance and give an inflated false positive rate when attempting to detect relatedness of genomic sequences.

However, this test is preliminary, and takes no account of the structure of the genome. In particular, we have not restricted ourselves to non-protein-coding segments. As current opinion is that even the non-coding part of the genome may be up to 80% functional, the possibility exists that the over- and under-represented words are restricted to segments of genome with specific, possibly yet unknown, functions. Thus the potential exists, for instance, to use $D_2$ as an exploratory probe to detect structure within the non-coding part of the genome: Using a randomly generated Markovian probe sequence (a random probe of length $m = 10,000$, say, would contain almost all 6-mers), one could calculate $D_2$ between the probe and a moving window running along the genome. This exercise would expose whether, for instance, the genome consists of a sea of 'null hypothesis' Markovian sequence containing islands of repeated motifs, or whether the genome is uniformly peppered with a particular set over-expressed words. The ability to easily calculate the null $D_2$ distribution as a function of sequence and word lengths enables the experiment to be performed readily at different resolutions. Furthermore, the property of $D_2$ that it is dominated by the natural variability in either of the two sequences being compared becomes an advantage. If a subset of words is over-represented within the moving window at a specific location in the genome, provided that subset contains some words also present in the probe sequence, its over-representation within the window will manifest as an extreme $D_2$.

# Appendix: Contributions to $\mathrm{Var}\,(D_2)$.

We derive the contributions $V_0$ to $V_4$ to $\mathrm{Var}\,(D_2)$ when $\omega = 1$ given in Section 3.3. These contributions are the partial sums $\sum_{i,i',j,j'} \mathrm{Prob}\,(I_{ij} = 1, I_{i'j'} = 1)$ contributing to Eq. (26) where, for given $(i, j)$, the indices $(i', j')$ range over the regions shown in Fig. 2. The event "$I_{ij} = 1, I_{i'j'} = 1$" means that the $k$-words beginning at sites $i$ and $i'$ in sequence $\mathbf{X}$ match the $k$-words beginning at sites $j$ and $j'$ in sequence $\mathbf{Y}$ respectively.

### Non-overlapping words in both sequences: $V_0$

Taking into account the PBCs, these are the contributions from the cases for which both $k \leq |i' - i| \leq m - k$ and $k \leq |j' - j| \leq n - k$ occur simultaneously. Consider the situation

$$\begin{aligned} i' &= (i + k + r) \bmod m, \qquad && r = 0, \ldots, m - 2k, \\ j' &= (j + k + s) \bmod n, \qquad && s = 0, \ldots, n - 2k, \end{aligned}$$

shown in Fig. 7(a). Since the two sequences are independent, applying Eq. (11) gives

$$\begin{aligned} &\mathrm{Prob}\,(I_{ij} = 1, I_{i'j'} = 1) \\ &\quad = \sum_{w,v \in \mathcal{A}^k} \mathrm{Prob}\,(X_i \ldots X_{i+k-1} = w)\mathrm{Prob}\,(X'_i \ldots X_{i'+k-1} = v) \\ &\qquad\qquad \times \mathrm{Prob}\,(Y_j \ldots Y_{j+k-1} = w)\mathrm{Prob}\,(Y'_j \ldots Y_{j'+k-1} = v) \\ &\quad = \frac{1}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\ &\qquad \sum_{w,v \in \mathcal{A}^k} (M^{m-2k-r+1})_{v_k w_1} M_{w_1 w_2} \ldots M_{w_{k-1} w_k} (M^{r+1})_{w_k v_1} M_{v_1 v_2} \ldots M_{v_{k-1} v_k} \\ &\qquad\qquad \times (M^{n-2k-s+1})_{v_k w_1} M_{w_1 w_2} \ldots M_{w_{k-1} w_k} (M^{s+1})_{w_k v_1} M_{v_1 v_2} \ldots M_{v_{k-1} v_k} \\ &\quad = \frac{\mathrm{tr}\,\left[(M^{r+1} \circ M^{s+1})(M \circ M)^{k-1}(M^{m-2k-r+1} \circ M^{n-2k-s+1})(M \circ M)^{k-1}\right]}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)}. \end{aligned} \tag{37}$$

Summing over $r$ and $s$, and including a factor of $mn$ to account for the sum over $i$ and $j$ then gives Eq. (27).

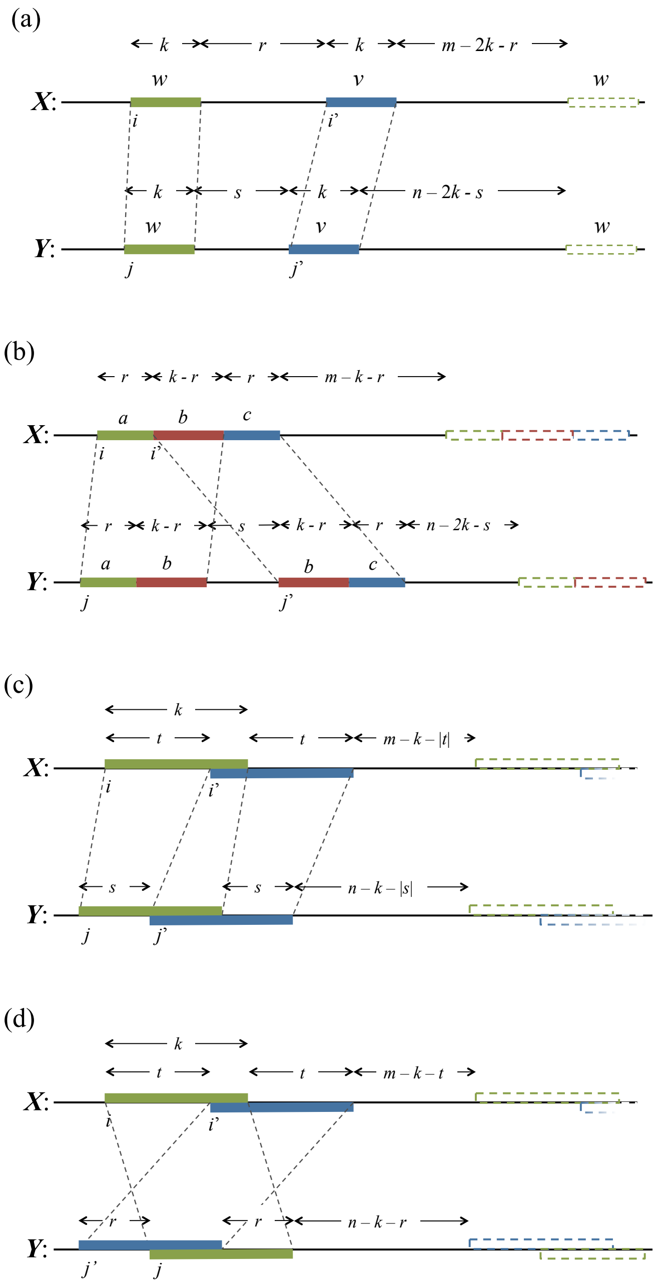

Figure 7: Arrangements of word matches contributing to (a) non-overlapping words, $V_0$, (b) crabgrass $V_1$ and (c) accordions $V_2$, $V_3$ and (d) the accordion $V_4$. Images of words due to periodic boundary conditions are shown as a dashed outline.

**Overlaps in one sequence only:** $V_1$

These are cases for which either $k \leq |j' - j| \leq n - k$ and $|i' - i| < k$ or $> m - k$ (overlaps in $\mathbf{X}$ but not in $\mathbf{Y}$), or $k \leq |i' - i| \leq m - k$ and $|j' - j| < k$ or $> n - k$ (overlaps in $\mathbf{Y}$ but not in $\mathbf{X}$). This region is referred to as the 'crabgrass' in ref. [28]. Fig. 7(b) shows the case of overlaps in $\mathbf{X}$ but not $\mathbf{Y}$, where we have set

$$\begin{aligned} r &= \begin{cases} i' - i & \text{if } |i' - i| < k, \\ i' - i - m & \text{if } |i' - i| > m - k, \end{cases} \\ j' &= (j + k + s) \bmod n, \end{aligned}$$

for $r = -k+1, \ldots, 0, \ldots, k-1$ and $s = 0, \ldots, n - 2k$. We split the common word beginning at $i$ and $j$ into a piece $a$ of length $r$ and a piece $b$ of length $k - r$, and split the common word beginning at $i'$ and $j'$ into the piece $b$ and a piece $c$ of length $r$.

Then

$$\begin{aligned} V_1 &= mn \left\{ \sum_{s=0}^{n-2k} \sum_{r=-k+1}^{k-1} \sum_{a \in \mathcal{A}^r} \sum_{b \in \mathcal{A}^{k-r}} \sum_{c \in \mathcal{A}^r} \text{Prob}\,(\text{configuration in Fig. 7(b)}) \right. \\ &\qquad \left. + \text{ a similar sum with the roles of } \mathbf{X} \text{ and } \mathbf{Y} \text{ interchanged} \right\}, \qquad (38) \end{aligned}$$

where the sums over $r$ and $s$ arise from sums over $i'$ and $j'$ for fixed $i$ and $j$, and the factor of $mn$ arises from the outer sum over $i$ and $j$. Using Eq. (11),

$$\begin{aligned}
&\sum_{a\in\mathcal{A}^r}\ \sum_{b\in\mathcal{A}^{k-r}}\ \sum_{c\in\mathcal{A}^r} \mathrm{Prob}\,(\text{configuration in Fig. 7(b)}) \\
&= \frac{1}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)} \times \\
&\quad \sum_{a,b,c} M_{a_1a_2}\dots M_{a_{r-1}a_r}M_{a_rb_1}M_{b_1b_2}\dots M_{b_{k-r-1}b_{k-r}}M_{b_{k-r}c_1} \times \\
&\qquad\qquad M_{c_1c_2}\dots M_{c_{r-1}c_r}(M^{m-k-r+1})_{c_ra_1} \times \\
&\quad M_{a_1a_2}\dots M_{a_{r-1}a_r}M_{a_rb_1}M_{b_1b_2}\dots M_{b_{k-r-1}b_{k-r}}(M^{s+1})_{b_{k-r}b_1} \times \\
&\quad M_{b_1b_2}\dots M_{b_{k-r-1}b_{k-r}}M_{b_{k-r}c_1}M_{c_1c_2}\dots M_{c_{r-1}c_r}(M^{n-2k-s+1})_{c_ra_1} \\
&= \sum_{a_1,b_1,b_{k-r},c_r} [(M\circ M)^r]_{a_1b_1}[(M\circ M\circ M)^{k-r-1}]_{b_1b_{k-r}}[(M\circ M)^r]_{b_{k-r}c_r} \times \\
&\qquad\qquad (M^{s+1})_{b_{k-r}b_1}(M^{m-k-r+1})_{c_ra_1}(M^{n-2k-s+1})_{c_ra_1} \\
&= \mathrm{tr}\,\{(M\circ M)^r[(M\circ M\circ M)^{k-r-1}\circ(M^{s+1})^T]\times \\
&\qquad\qquad (M\circ M)^r(M^{m-k-r+1}\circ M^{n-2k-s+1})\}\,, \qquad (39)
\end{aligned}$$

where the superscript $T$ indicates the matrix transpose. Eqs. (38) and (39) combine to give the crabgrass contribution Eq. (28).

### Overlaps in both sequences

The set of configurations for which the words at positions $i$, $i'$, $j$ and $j'$ overlap in both sequences simultaneously are referred to as the 'accordion' in ref. [28]. For convenience we define the folowing overlap distances (illustrated in Fig. 7(c)):

$$t = \begin{cases} i'-i & \text{if } |i'-i| < k, \\ i'-i-m & \text{if } |i'-i| > m-k, \end{cases} \qquad (40)$$

in sequence $\mathbf{X}$ and

$$s = \begin{cases} j'-j & \text{if } |j'-j| < k, \\ j'-j-n & \text{if } |j'-j| > n-k, \end{cases} \qquad (41)$$

in sequence $\mathbf{Y}$. These definitions ensure that $-k+1 \leq t, s \leq k-1$. The remaining three contributions are from the accordion.

### Diagonal part of the accordion: $V_2$

This is the contribution from those cases with $s = t$, in which case Fig. 7(c) becomes a match between the $(k + |t|)$-letter word at position $i$ in $\mathbf{X}$ and the $(k + |t|)$-letter word at position $j$ in $\mathbf{Y}$. Noting that the probability of this match is independent of $i$ and $j$, we have

$$V_2 = mn \sum_{t=-k+1}^{k-1} \mathrm{Prob}\,((X_i \dots X_{i+k+|t|-1}) = (Y_j \dots Y_{j+k+|t|-1})), \tag{42}$$

where, by analogy with Eq. (21),

$$\begin{aligned} &\mathrm{Prob}\,((X_i \dots X_{i+k+|t|-1}) = (Y_j \dots Y_{j+k+|t|-1})) \\ &\qquad = \frac{\mathrm{tr}\,[(M^{m-k-t+1} \circ M^{n-k-t+1})(M \circ M)^{k+t-1}]}{\mathrm{tr}\,(M^m)\mathrm{tr}\,(M^n)}. \end{aligned} \tag{43}$$

Combining Eqs. (42) and (43) gives Eq. (29).

### Off-diagonal part of the accordion: subcases contributing to $V_3$

The off-diagonal part of the accordion is divided into a number of subcases. Consider first the contribution from the four subcases making up the region $V_3$ in Fig. 2:

3(i): $0 \le s < t \le k - 1$;

3(ii): $-k + 1 \le s < t \le 0$;

3(iii): $-k + 1 \le t < s \le 0$ and

3(iv): $0 \le t < s \le k - 1$.

By symmetry, each subcase makes an equivalent contribution to the variance. Subcase 3(i) is shown in Fig. 8, and the required contribution takes the form

$$\begin{aligned} V_3 \;=\; 2mn \sum_{t=1}^{k-1}\sum_{s=0}^{t-1} \sum_{a,b\in\mathcal{A}^s} \sum_{c\in\mathcal{A}^\rho} \sum_{d\in\mathcal{A}^\sigma} \big[&\mathrm{Prob}\,(\text{configuration shown in Fig. 8}) + \\ &\mathrm{Prob}\,(\text{same configuration with } m \text{ and } n \text{ interchanged})\big]. \end{aligned} \tag{44}$$

To calculate the probability of the configuration, the overlapping words have been divided into repeating independent elements. Elements $a$ and $b$ are the non-overlapping parts of length $s$ at either end of the words at $j$ and $j'$ in $\mathbf{Y}$. The non-overlapping part of the words at $i$ and $i'$ in $\mathbf{X}$ are segmented into elements

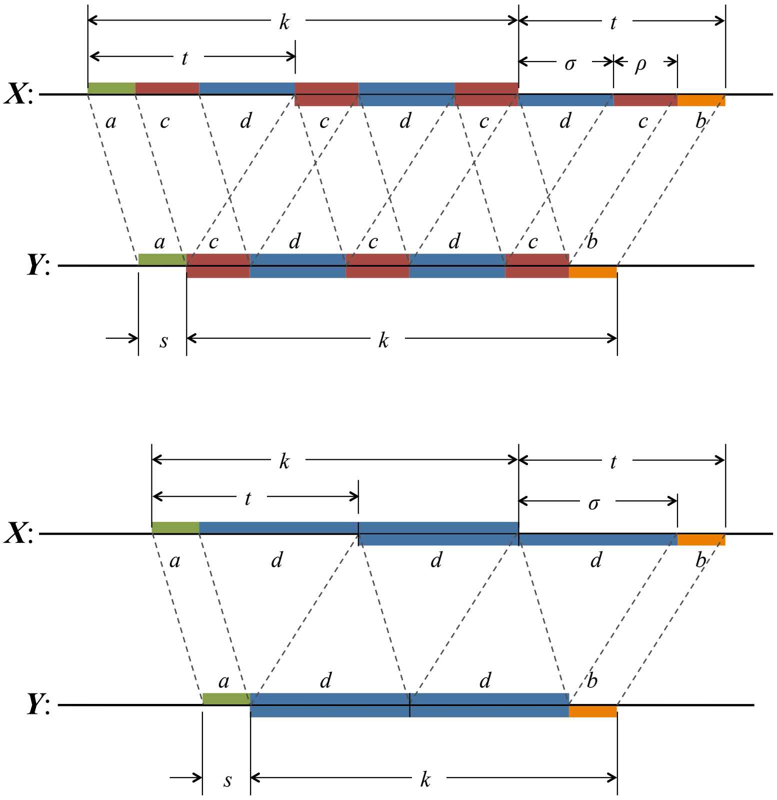

Figure 8: Arrangements of word matches contributing to subcase 3(i) when $\rho = (k - s) \bmod (t - s) > 0$ (upper figure) and $\rho = 0$ (lower figure).

$(acd)$ and $(dcb)$ shown in the upper part of Fig. (8). The segment $(cd)$ repeats an integer number $\nu$ times within the overlapping part in sequence $\mathbf{Y}$, with a segment $c$ of length $\rho$ left over. We set the length of element $d$ equal to $\sigma$. Thus

$$\nu = \left\lfloor \frac{k - s}{t - s} \right\rfloor, \quad \rho = (k - s) \bmod (t - s), \quad \text{and} \quad \sigma = t - s - \rho. \tag{45}$$

When $\rho = 0$ the element $c$ does not occur (lower part of Fig. (8)).

Using arguments similar to those for the crabgrass contribution, we have, for $\rho > 0$,

$$\begin{aligned}
&\sum_{a,b\in\mathcal{A}^s}\sum_{c\in\mathcal{A}^\rho}\sum_{d\in\mathcal{A}^\sigma} \text{Prob}\,(\text{configuration shown in Fig. 8}) = \\
&\qquad \frac{1}{\text{tr}\,(M^m)\text{tr}\,(M^n)} \sum_{a_1,a_s,b_1,b_s,c_1,c_\rho,d_1,d_\sigma\in\mathcal{A}} \\
&\qquad\qquad [(M\circ M)^{s-1}]_{a_1a_s}(M\circ M)_{a_sc_1}[(M^{\circ(2\nu+3)})^{\rho-1}]_{c_1c_\rho}\times \\
&\qquad\qquad\quad (M^{\circ(2\nu+1)})_{c_\rho d_1}[(M^{\circ(2\nu+1)})^{\sigma-1}]_{d_1d_\sigma}(M^{\circ(2\nu+1)})_{d_\sigma c_1}\times \\
&\qquad\qquad\qquad (M\circ M)_{c_\rho b_1}[(M\circ M)^{s-1}]_{b_1b_s}(M^{m-k-t+1}\circ M^{n-k-s+1})_{b_sa_1} \\
&= \frac{1}{\text{tr}\,(M^m)\text{tr}\,(M^n)}\text{tr}\,\Big[(M\circ M)^s\big\{(M^{\circ(2\nu+3)})^{\rho-1}\circ[(M^{\circ(2\nu+1)})^{\sigma+1}]^T\big\}\times \\
&\qquad\qquad (M\circ M)^s(M^{m-k-t+1}\circ M^{n-k-s+1})\Big], \qquad (46)
\end{aligned}$$

while for $\rho = 0$ we have

$$\begin{aligned}
&\sum_{a,b\in\mathcal{A}^s}\sum_{c\in\mathcal{A}^\rho}\sum_{d\in\mathcal{A}^\sigma} \text{Prob}\,(\text{configuration shown in Fig. 8}) = \\
&\qquad \frac{1}{\text{tr}\,(M^m)\text{tr}\,(M^n)} \sum_{a_1,a_s,b_1,b_s,d_1,d_\sigma\in\mathcal{A}} \\
&\qquad\qquad [(M\circ M)^{s-1}]_{a_1a_s}(M\circ M)_{a_sd_1}[(M^{\circ(2\nu+1)})^{t-s-1}]_{d_1d_\sigma}(M^{\circ(2\nu-1)})_{d_\sigma d_1}\times \\
&\qquad\qquad\quad (M\circ M)_{d_\sigma b_1}[(M\circ M)^{s-1}]_{b_1b_s}(M^{m-k-t+1}\circ M^{n-k-s+1})_{b_sa_1} \\
&= \frac{1}{\text{tr}\,(M^m)\text{tr}\,(M^n)}\text{tr}\,\Big[(M\circ M)^s\big\{(M^{\circ(2\nu+1)})^{t-s-1}\circ(M^{\circ(2\nu-1)})^T\big\}\times \\
&\qquad\qquad (M\circ M)^s(M^{m-k-t+1}\circ M^{n-k-s+1})\Big], \qquad (47)
\end{aligned}$$

Combining Eqs. (44), (45), (46) and (47) gives Eqs. (30) to (32).

**Off-diagonal part of the accordion: subcases contributing to $V_4$**

These are contributions from the subcases

4(i): $1 \le t \le k-1$, $-k+1 \le s \le -1$; and

4(ii): $1 \le s \le k-1$, $-k+1 \le t \le -1$

labelled $V_4$ in Fig. 2. In these cases either $t$ or $s$ is negative. By symmetry, each of these two subcases makes an equivalent contribution to $V_4$, so we consider subcase 4(i) and for convenience set $r = -s$ (see Fig. 7(d)). Then

$$\begin{aligned} V_4 &= 2mn \sum_{r,t=1}^{k-1} \mathrm{Prob}\left[(X_i \ldots X_{i+k-1}) = (Y_j \ldots Y_{j+k-1}),\right. \\ &\qquad\qquad \left.(X_{i+t} \ldots X_{i+t+k-1}) = (Y_{j-r} \ldots Y_{j-r+k-1})\right], \end{aligned} \tag{48}$$

where the factor $mn$ arises from a sum over $i$ and $j$, and we make use of the the fact that for periodic Markovian sequences the summand is independent of $i$ and $j$.

It is convenient to define

$$\nu = \left\lfloor \frac{k}{r+t} \right\rfloor, \qquad \zeta = k \bmod (r+t). \tag{49}$$

Here $\nu$ is the integer number of times the complete repeat unit $(Y_j \ldots Y_{j+r+t-1})$ fits inside the $k$-word $(Y_j \ldots Y_{j+k-1})$, and $\zeta$ is the number of letters remaining (see Figs. 9 and 10). Calculation of the probability occurring in Eq. (48) then proceeds in a similar fashion to that for $V_3$ by dividing the overlapping words into independent non-overlapping elements. It turns out that the configuration of elements depends on the relationship between $\zeta$, $r$ and $t$. The complete set of configurations is enumerated in Figs. 9 and 10, with the repeated elements labelled $a, b$, etcetera. The calculation is lengthy and repetitive but straightforward, and yields Eqs. (33) and (35) after recombining cases which give the same algebraic formula.

# Acknowledgements

This work was supported in part by ARC Discovery grant DP120101422.

# References

[1] B.E. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National USA*, 83:5155–5159, 1986.

[2] C. J. Burden, J. Jing, S. Forêt, and S. R Wilson. Application of $k$-word match statistics to the clustering of proteins with repeated domains. In A Colubi, K. Fokianos, E.J. Kontoghiorghes, and González-Rodríguez, editors, *Proceedings of COMPSTAT 2012, 20th International Conference on Computational Statistics*, pages 131–142, 2012.

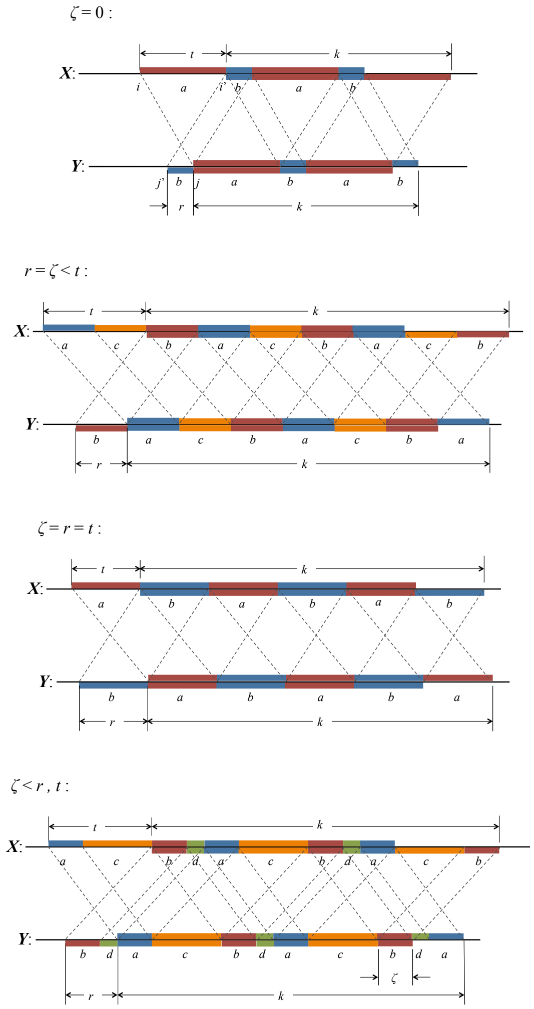

Figure 9: Arrangements of word matches contributing to V4 (first 4 cases).

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

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

[5] B. Chor, D. Horn, N Goldman, Y. Levy, and T. Massingham. Genomic DNA

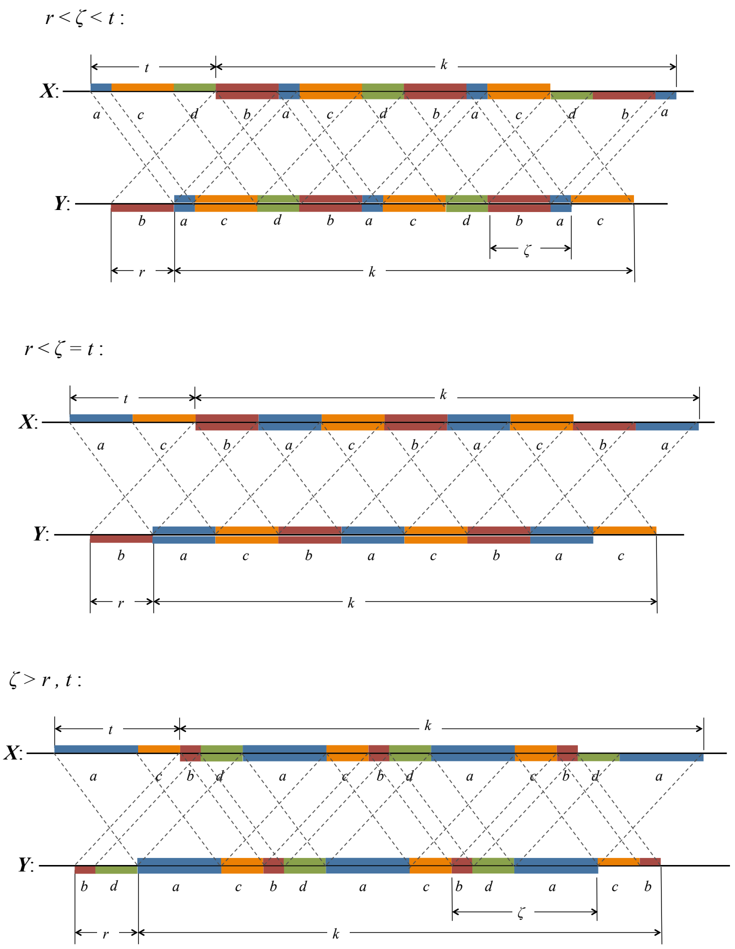

Figure 10: Arrangements of word matches contributing to V4 (final 3 cases).

$k$-mer spectra: models and modalities. *Genome Biology*, 10:R108, 2009.

[6] B. Chor, D. Horn, N Goldman, Y. Levy, and T. Massingham. $k$-mer analysis of multiple genomes, 2009. `http://www.ebi.ac.uk/goldman-srv/ChorEtAlSpectra/Spectra/HumanChromosomes/chr1/`.

[7] M. Csűrös, L. Noé, and G. Kucherov. Reconsidering the significance of genomic word frequencies. *Trends in Genetics*, 23:543–546, 2007.

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

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

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

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

[12] W. Hide, J. Burke, and D. B. Davison. Biological evaluation of d2, an algorithm for high-performance sequence comparison. *J Comput Biol*, 1(3):199–215, 1994.

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

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

[15] M. R. Kantorovitz, G. E. Robinson, and S. Sinha. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, 23(13):i249–55, 2007.

[16] Donald E. Knuth. *The Art of Computer Programming*, volume 2, Seminumerical Algorithms. Addison-Wesley, 2nd edition, 1981.

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

[18] J.K. Percus and O.E. Percus. The statistics of words on rings. *Communications on pure and applied mathematics*, 59:145–160, 2006.

[19] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012.

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

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

[22] George W. Snedecor and William G. Cochran. *Statistical Methods.* Iowa State University Press, 7th edition, 1980.

[23] G.W. Stuart, K. Moffett, and S. Baker. Integrated gene and species phylogenies from unaligned whole genome protein sequences. *Bioinformatics*, 18:100–108, 2002.

[24] G.W. Stuart, K. Moffett, and J.J. Leader. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol. Biol. Evol.*, 19:554–562, 2002.

[25] D.C. Torney, C. Burks, D. Davison, and K.M. Sirotkin. Computation of $d^2$. A measure of sequence dissimilarity. In *Computers and DNA*, Santa Fe Institute Studies in the Sciences of Complexity, pages 109–125. Addison-Wesley, New York, 1990.

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

[27] L. Wan, G. Reinert, F. Sun, and M. S. Waterman. Alignment-free sequence comparison (II): theoretical power of comparison statistics. *J. Comp. Biol.*, 17(11):1467–90, 2010.

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

[29] Wellcome Trust Sanger Institute and European Bioinformatics Institute. Ensembl Genome Browser, July 2012. Homo Sapiens DNA, `ftp://ftp.ensembl.org/pub/release-68/fasta/homo_sapiens/dna/`, file Homo_sapiens.GRCh37.68.dna_sm.chromosome.1.fa.gz.