Statistical Applications in Genetics and Molecular Biology

Volume 11, *Issue* 1 2012 *Article* 3

Alignment-free Sequence Comparison for Biologically Realistic Sequences of Moderate Length

Conrad J. Burden, *Australian National University*
Junmei Jing, *Australian National University*
Susan R. Wilson, *Australian National University*

Recommended Citation:
Burden, Conrad J.; Jing, Junmei; and Wilson, Susan R. (2012) "Alignment-free Sequence Comparison for Biologically Realistic Sequences of Moderate Length," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 1, Article 3.
DOI: 10.2202/1544-6115.1724

Alignment-free Sequence Comparison for Biologically Realistic Sequences of Moderate Length

Conrad J. Burden, Junmei Jing, and Susan R. Wilson

Abstract

The D_2 statistic, defined as the number of matches of words of some pre-specified length k, is a computationally fast alignment-free measure of biological sequence similarity. However there is some debate about its suitability for this purpose as the variability in D_2 may be dominated by the terms that reflect the noise in each of the single sequences only. We examine the extent of the problem and the effectiveness of overcoming it by using two mean-centred variants of this statistic, D_2* and D_2c. We conclude that all three statistics are potentially useful measures of sequence similarity, for which reasonably accurate p-values can be estimated under a null hypothesis of sequences composed of identically and independently distributed letters. We show that D_2 and D_2c, and to a somewhat lesser extent D_2*, perform well in tests to classify moderate length query sequences as putative cis-regulatory modules.

Author Notes: This work was funded in part by ARC discovery grant DP0987298.

1 Introduction and Background

The D_2 statistic is defined as the number of word matches of some pre-specified word length k between the two sequences of letters from a given alphabet $\mathscr{A}$. Given two sequences $\mathbf{A} = A_1, \ldots, A_{n_A}$ and $\mathbf{B} = B_1, \ldots, B_{n_B}$ of length n_A and n_B respectively, let X_w and Y_w be the number of occurrences of the k-word $w \in \mathscr{A}^k$ in $\mathbf{A}$ and $\mathbf{B}$ respectively. Then

$$D_2(k) = \sum_{w \in \mathscr{A}^k} X_w Y_w. \tag{1}$$

In a series of papers (Forêt, Kantorovitz, and Burden (2006), Kantorovitz, Booth, Burden, and Wilson (2006), Burden, Kantorovitz, and Wilson (2008), Forêt, Wilson, and Burden (2009a,b)) the D_2 statistic has been promoted as a potential tool for alignment-free comparison of biological sequences (Waterman (1995), Lippert, Huang, and Waterman (2002)). In these applications the alphabet $\mathscr{A}$ consists of 4 nucleic acids in the case of DNA sequences or 20 amino acids in the case of protein sequences. Compared with alignment-based sequence comparison methods such as BLAST (Altschul, Madden, Schaffer, Zhang, Zhang, Miller, and Lipman (1997)), alignment-free sequence comparison measures do not assume conservation of long range contiguity between sequences, and may be useful when genome shuffling, reversal or long insertions occur or for the identification of potential gene-regulatory regions from training data.

The reasoning behind the D_2 statistic is that it should detect simultaneous over-representation in both sequences of a particular subset of all possible words. However, it has been argued by Lippert et al. (2002) and Reinert, Chew, Sun, and Waterman (2009) that a potential serious shortcoming of the D_2 statistic is that the signal one is trying to detect may be hidden by the natural variability of D_2 due to noise in each of the single sequences only, as measured by $\mathrm{Var}(D_2)$ under a suitable null hypothesis. We refer to this effect as single sequence noise. More specifically, under an assumption that $\mathbf{A}$ and $\mathbf{B}$ are random sequences consisting of identically and independently distributed (iid) letters, the variance of D_2 for a pair of sequences of length n is composed of an order $O(n^3)$ part arising from independent random variations of word frequencies in each of the two sequences (i.e. single sequence noise), and an order $O(n^2)$ arising from correlated variations of word frequencies in both sequences. Therefore single sequence noise will swamp any signal of simultaneous over-representation of particular words if the sequences are sufficiently long. Consequently Reinert et al. (2009) argue that it may be more appropriate to define alignment-free word match statistics in terms of the mean-centred word counts

$$\tilde{X}_w = X_w - E[X_w] = X_w - \bar{n}_A p_w,$$
$$\tilde{Y}_w = Y_w - E[Y_w] = Y_w - \bar{n}_B p_w, \tag{2}$$

where p_w is the probability of the k-word w occurring at any given location in **A** or **B** and $\bar{n}_A, \bar{n}_B$ are the number of possible k-word locations in sequences **A** and **B** respectively.[1] Accordingly they propose two new statistics, a centred, weighted word count defined by

$$D_2^*(k) = \sum_{w \in \mathscr{A}^k} \frac{\tilde{X}_w \tilde{Y}_w}{\sqrt{n_A n_B} p_w}, \tag{3}$$

and a "self-standardized" or "Schepp" word match count,

$$D_2^S(k) = \sum_{w \in \mathscr{A}^k} \frac{\tilde{X}_w \tilde{Y}_w}{\sqrt{\tilde{X}_w^2 + \tilde{Y}_w^2}}. \tag{4}$$

The main benefit of these two new statistics reported by Reinert et al. (2009) and Wan, Reinert, Sun, and Waterman (2010) is that they have higher power to detect sequence similarity when tested on synthetic data. Their power calculations are based on alternative hypotheses simulated either from a common motif model, in which a short motif is inserted randomly into two iid sequences, or a pattern transfer model in which short motifs are randomly copied from one sequence to the other. The power under a pattern transfer model can be further improved for extremely long sequences by using localised versions of D_2^* or D_2^S which are defined as sums of pairwise comparisons of subsequences of the sequences under consideration (Liu, Wan, Li, Reinert, Waterman, and Sun (2011)).

A second reported benefit of the Schepp statistic is that it is well represented by a normal distribution. Unfortunately the Schepp statistic has a number of disadvantages for applications which involve large database searches: (i) it is computationally more intensive than the D_2 statistic, requiring a sum over all $|\mathscr{A}|^k$ possible k-words and hence a run time of order $O(|\mathscr{A}|^k(n_A+n_B))$, ($D_2$ has an extremely fast run time of order $O(k(n_A+n_B))$), (ii) unlike D_2, there is no known exact formula for its variance, which must be estimated from a simulation for each set of parameters k, n_A, n_B and letter frequencies f_a, and (iii) the generalisation to a larger alphabet, as required for protein amino-acid sequences, becomes impractical given the large number of possible k-words and high dimensionality of the parameter space.

[1] In general $\bar{n}_A = n_A - k + 1$ and $\bar{n}_B = n_B - k + 1$. Herein, however, we find it algebraically convenient to impose periodic boundary conditions on each of the the sequences and henceforth set $\bar{n}_A = n_A$ and $\bar{n}_B = n_B$. The periodic boundary conditions are a minor technicality which are easily implemented in practical applications (see Forêt et al. (2009a)).

DOI: 10.2202/1544-6115.1724

In this paper we argue that the extent of the single sequence noise problem is not serious for moderate sequence lengths encountered in many biological applications, such as searches for regulatory motifs or protein phylogeny. We also demonstrate that, provided the word probabilities p_w are specified externally and not estimated from the sequences, the D_2^* statistic can be conveniently written as an uncentred weighted word count whose mean and variance can be calculated analytically under an iid null hypothesis and whose null distribution is very well approximated by a Gamma distribution. The benefit of these results is that D_2^* shares with D_2 the properties that it is very fast to compute (with run time linear in the sequence lengths), and that accurate p-values can easily be obtained under the iid null hypothesis for biologically relevant parameter regimes. We also test and compare the performance of the D_2, D_2^* and a centred version of D_2, which we call D_2^C, with and without mismatches against a dataset of known cis-regulatory modules constructed by Kantorovitz, Robinson, and Sinha (2007).

2 The magnitude of single sequence variations

As mentioned in the introduction, the extent of the single sequence noise problem is indicated by the dominance of $\mathrm{Var}\,(D_2)$ by a contribution from noise in each of the single sequences. In the appendix it is shown for sequences of lengths n_A and n_B and the more general case of a weighted word match statistic that the variance is composed of an $O(n_A n_B)$ part consisting of correlated variations of word-match counts from their mean (i.e. the true signal) and an $O(n_A n_B(n_A+n_B))$ part composed entirely of unwanted single sequence noise. Clearly the second contribution will dominate in the limit $n_A, n_B \to \infty$.

In Fig. 1 is plotted the contribution to $\mathrm{Var}\,(D_2)$ from its $O(n_A n_B(n_A+n_B))$ part arising from single sequence noise as a fraction of the total variance under the iid null-hypothesis assumption. The calculation uses the exact formula in Forêt et al. (2009b) for $\mathrm{Var}\,(D_2(k))$, which easily splits into $O(n_A n_B)$ and $O(n_A n_B(n_A+n_B))$ parts. The range of word lengths k and sequence lengths n_A, n_B covers many cases arising in previous studies of the D_2 statistic by Forêt et al. (2006, 2009a) and Kantorovitz et al. (2007). For alphabet $\mathscr{A} = \{A, C, G, T\}$, the parameter η is an asymmetry parameter introduced by Melko and Mushegian (2004) describing the departure of a strand-symmetric nucleic acid distribution from uniformity:

$$f_C = f_G = \tfrac{1}{4}(1-\eta), \quad f_A = f_T = \tfrac{1}{4}(1+\eta), \tag{5}$$

where f_a is the frequency of occurrence of letter $a \in \mathscr{A}$ in either sequence and $-1 \leq \eta \leq 1$. Examples of more extreme letter asymmetry in genomes are the roundworm *Caenorhabditis elegans* and zebra fish *Danio rerio*, which exhibit compositional

biases in the region of $\eta = \frac{1}{3}$, whereas for most mammals the asymmetry parameter lies in the range $\eta \lesssim 0.1$ (Khuu, Sandor, DeYoung, and Ho (2007)).

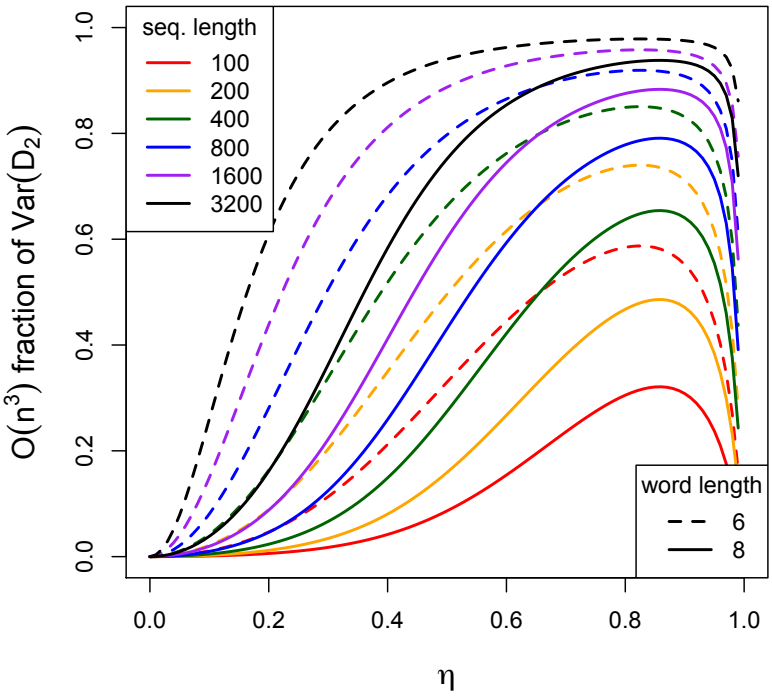

Figure 1: The fraction of $\mathrm{Var}\,(D_2(k))$ accounted for by its $O(n^3)$ single sequence noise contribution for a range of sequence lengths $n = n_A = n_B$ and word lengths k. The letter frequency asymmetry parameter η is defined in Eq. (5). Only $\eta > 0$ is shown as the curves are symmetric about $\eta = 0$.

One sees in Fig. 1 that $\mathrm{Var}\,(D_2)$ is not strongly dominated by single sequence noise for the moderate values of η occurring in nature, particularly for longer word lengths and shorter sequences. This brings into question the assertion in Reinert et al. (2009) that $D_2(k)$ is not a suitable statistic for sequence comparison, and demands further investigation. In Section 5 below we repeat an earlier analysis by Forêt et al. (2009a) of a collection of cis-regulatory model data sets constructed by Kantorovitz et al. (2007) to gauge the advantage to be gained by using the centred versions of word count statistics. These data sets are from the parts of the human and fly genomes with asymmetry parameters in the range $0.1 \lesssim \eta \lesssim 0.2$.

DOI: 10.2202/1544-6115.1724

3 The properties of D_2^*

The statistic D_2^* (Eq. (3)), which was introduced by Reinert et al. (2009), was motivated by a desire to scale the word count vectors X_w and Y_w by estimates of their standard deviations. In its original version, the quantity p_w occurring in the denominator of Eq. (3) was an estimate of the word frequency w obtained from observed letter frequencies in the sequences **A** and **B**, and was therefore a random variable. In the subsequent calculations of the properties of D_2^* by Wan et al. (2010), these word frequencies are defined as $p_w = \Pi_{a\in w} f_a$, where f_a is the prespecified probability of occurrence of letter $a \in \mathscr{A}$ under the null hypothesis in the randomly generated iid sequences **A** and **B**. In this second interpretation, p_w is an externally specified parameter and not a random variable. The choice of interpretation will affect the null distribution of the statistic D_2^*, and in this paper we chose the second interpretation, principally because it renders the statistic amenable to analytic investigation.

We now show that D_2^* can be written, up to an additive constant, as a weighted uncentred word match count. For any k-word $w \in \mathscr{A}^k$ and position $i = 1,\ldots,n_A$ in sequence **A**, define the indicator random variable

$$I_i^A(w) = \begin{cases} 1 & \text{if } (A_i \ldots A_{i+k-1}) = (w_1 \ldots w_k), \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

and similarly for $I_j^B(w)$, $j = 1,\ldots,n_B$. We also impose periodic boundary conditions on both sequences, i.e. $A_{n_A+i} = A_i$ and $B_{n_B+i} = B_i$ for $i = 1,\ldots,k-1$. The word count vectors can then be written

$$X_w = \sum_{i=1}^{n_A} I_i^A(w), \qquad Y_w = \sum_{i=1}^{n_B} I_i^B(w). \tag{7}$$

Consider the expansion

$$\sqrt{n_A n_B} D_2^*(k) = \sum_{w\in\mathscr{A}^k} \frac{X_w Y_w}{p_w} - n_B \sum_{w\in\mathscr{A}^k} X_w - n_A \sum_{w\in\mathscr{A}^k} Y_w + n_A n_B \sum_{w\in\mathscr{A}^k} p_w. \tag{8}$$

The second term of this expansion is

$$-n_B \sum_{w\in\mathscr{A}^k} X_w = -n_B \sum_{i=1}^{n_A} \sum_{w\in\mathscr{A}^k} I_i^A(w) = -n_B \sum_{i=1}^{n_A} 1 = -n_A n_B, \tag{9}$$

and similarly the third term is also equal to $-n_A n_B$. Since $\sum_{w\in\mathscr{A}} p_w = 1$, the final term is equal to $n_A n_B$. Thus D_2^* can equivalently be written as

$$\sqrt{n_A n_B} D_2^*(k) = D_2^\dagger(k) - n_A n_B, \tag{10}$$

where $D_2^\dagger$ is the weighted, uncentred word match count

$$D_2^\dagger(k) = \sum_{w \in \mathscr{A}^k} \frac{X_w Y_w}{p_w}. \tag{11}$$

3.1 Mean and variance of D_2^*

The means of $D_2^*(k)$ and $D_2^\dagger(k)$ are easily seen to be

$$E[D_2^*(k)] = 0, \quad E[D_2^\dagger(k)] = n_A n_B. \tag{12}$$

An exact analytic formula for the variance of any weighted word count statistic of the form $D_2^W = \sum_{w,v \in \mathscr{A}^k} X_w \beta_{w,v} Y_v$, where $\beta_{w,v}$ is a fixed $d^k \times d^k$ symmetric matrix defined between any two k-words w and v and $d = |\mathscr{A}|$, has been given by Burden, Jing, and Wilson (2011) (see appendix). Applying Eqs. (23) to (28) to the current case gives

$$\begin{aligned} n_A n_B \mathrm{Var}\,(D_2^*(k)) &= \mathrm{Var}\,(D_2^\dagger(k)) \\ &= n_A n_B \left[d^k + 1 - 2k + 2d \frac{d^{k-1} - 1}{d - 1} \right]. \end{aligned} \tag{13}$$

Two things are of note. Firstly the variance is independent of the letter distribution f_a. Secondly the variance is proportional to $n_A n_B$, with no third order part. Given the relationship between D_2^* and $D_2^\dagger$, this is consistent with the result proved in the appendix that the variance of the centred version of weighted word count, defined as

$$D_2^{WC}(k) = \sum_{w,v \in \mathscr{A}^k} \tilde{X}_w \beta_{w,v} \tilde{Y}_v, \tag{14}$$

is precisely the $O(n_A n_B)$ part of $\mathrm{Var}\,(D_2^W(k))$, while the third order part is entirely composed of single sequence noise. Thus we expect $D_2^\dagger$ to be free of single sequence noise as n_A, n_B become large.

3.2 Empirical distribution of D_2^*

Numerical simulations carried out by Forêt et al. (2009a,b) indicate that the Gamma distribution gives a more accurate approximation to the distribution of $D_2(k)$ than does the Normal distribution for parameter ranges typically encountered in biological applications. We have carried out analogous simulations of the $D_2^\dagger$ distributions generated from ensembles of 100,000 pairs of random iid sequences. We compare

DOI: 10.2202/1544-6115.1724

these with Normal and Gamma distributions. For comparison with the Gamma distribution in particular, it is more convenient to work with $D_2^\dagger$, whose range is the interval $[0,\infty)$, rather than D_2^*, to which it is related by a simple shift and rescaling, namely Eq. (10).

The motivation for the use of a Gamma distribution is as follows. The Normal and Pólya-Aeppli (or compound Poisson) distributions are known asymptotic distributions of the D_2 statistic in cases where the word length k is small or large respectively with respect to the logarithm of the sequence lengths (Lippert, Huang, and Waterman (2002), Burden, Kantorovitz, and Wilson (2008)). A Pólya-Aeppli random variable is the sum of a Poisson number of Geometric random variables, and 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 (Lothaire (2005)). A Gamma random variable is a fixed number of iid exponential random variables, and therefore should be an approximation to a Pólya-Aeppli random variable as the Poisson parameter increases (so that the number of clumps is narrowly distributed about its mean) and the expected number of word matches increases (so that the geometric number of words in a clump can be approximated by a continuous exponential random variable). By the central limit theorem, a Gamma distribution is also asymptotically Normal in the limit that the number of iid exponential random variables becomes large. It is therefore a logical choice for an empirical fit straddling both asymptotic regimes.

Figure 2 shows empirically generated cumulative distribution functions of the $D_2^\dagger$ distribution for typical values of sequence length, word length k and asymmetry parameter $\eta = 0$, 0.1 and 0.2 defined by Eq. (5) for a $d = 4$ letter alphabet, together with Normal, Gamma and Pólya-Aeppli distribution functions with means and variances matching the theoretical values given in Eqs. (12) and (13).

By definition, $D_2^\dagger(k)$ is a discrete random variable. For the case of a uniform letter distribution, $\eta = 0$, we have $D_2^\dagger(k) = d^k D_2(k)$, where $D_2(k)$ takes only integer values. As shown in Figure 2, the discrete nature of $D_2^\dagger(k)$ is evident and the Pólya-Aeppli approximation is very close in this case for the parameter values shown. In the biologically realistic range $\eta = 0.1$ to 0.2 however, the allowed values of $D_2^\dagger(k)$ are considerably more densely packed on the real line. For the parameter values shown, the Gamma approximation is clearly superior to the Normal approximation, though the Normal approximation improves as n increases as expected.

A more quantitative indication of the closeness of fit of empirically determined samples of $D_2^\dagger$ to model distributions can be got from the Kolmogorov-Smirnov test. Table 1 gives Kolmogorov-Smirnov p-values obtained by comparing $D_2^\dagger$ samples against Normal and Gamma distributions with means and variances matching the theoretical values given in Eqs. (12) and (13). We observe that the

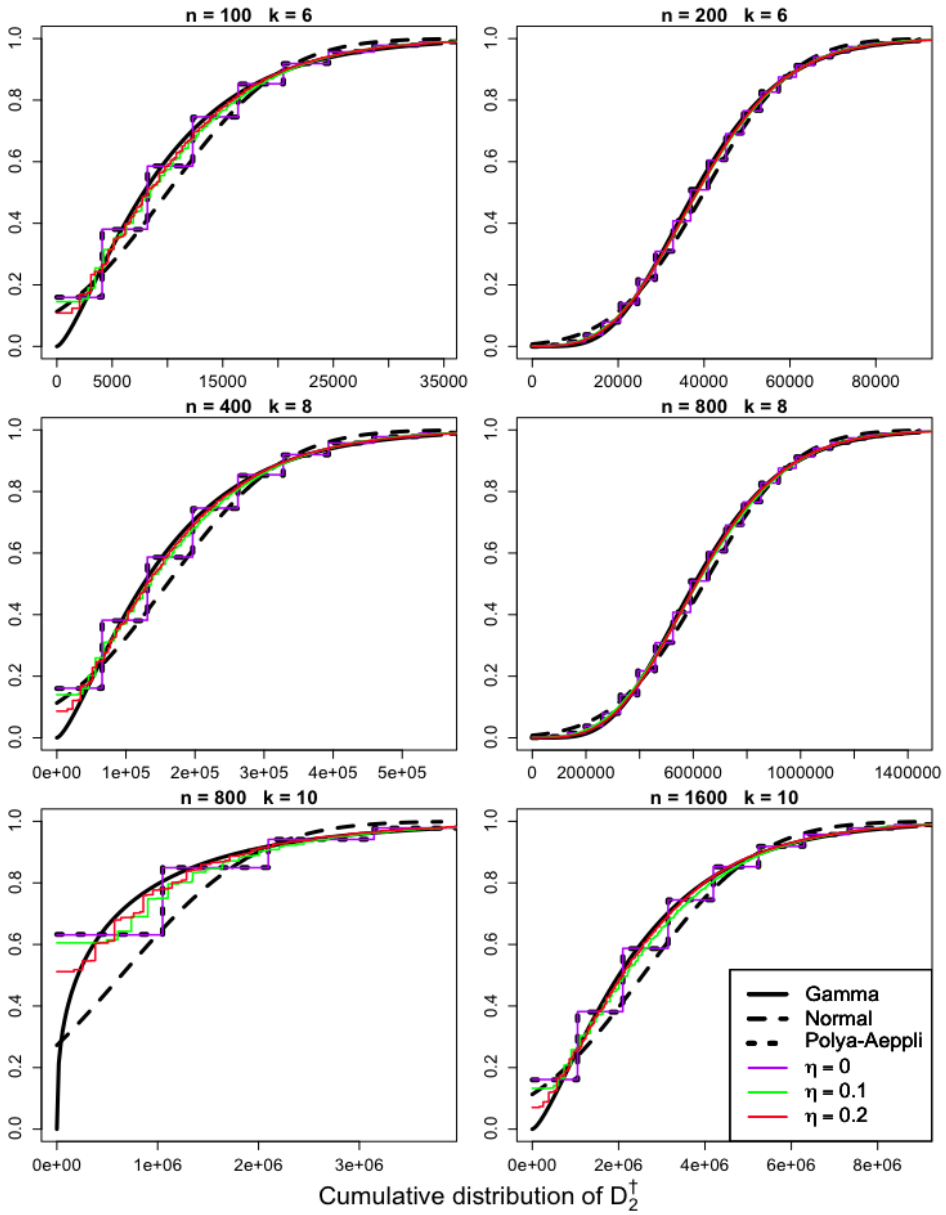

Figure 2: Empirical distribution functions of $D_2^\dagger(k)$ for a range of sequence lengths $n = n_A = n_B$, word lengths k and letter frequency asymmetry parameter η for a 4 letter alphabet, obtained from samples of 100,000 pairs of iid sequences. Also shown are Normal, Gamma and Pólya-Aeppli distribution functions with means and variances matching the theoretical values

DOI: 10.2202/1544-6115.1724

Table 1: P-values from the Kolmogorov-Smirnov test applied to samples of the $D_2^\dagger$ statistic from samples of pairs of random i.i.d. sequences of length n from a 4 letter alphabet with asymmetry parameter $\eta = 0.1$ for words of length k. The tests were carried out against Normal and Gamma distributions with means and variances calculated independently of the samples. P-values greater than 0.05 are shown in bold face.

$k \backslash n$	100	200	400	800	1600	3200
$D_2^\dagger$ vs. Normal (Sample size $= 10,000$)						
4	0.000	0.042	**0.740**	**0.758**	**0.745**	**0.876**
6	0.000	0.000	0.003	**0.340**	**0.349**	**0.142**
8	0.000	0.000	0.000	0.000	0.000	**0.251**
10	0.000	0.000	0.000	0.000	0.000	0.000
$D_2^\dagger$ vs. Gamma (Sample size $= 10,000$)						
4	**0.131**	**0.959**	**0.704**	**0.378**	**0.510**	**0.895**
6	0.000	0.000	**0.057**	**0.120**	**0.724**	**0.475**
8	0.000	0.000	0.000	0.004	**0.327**	0.016
10	0.000	0.000	0.000	0.000	0.000	0.007
$D_2^\dagger$ vs. Normal (Sample size $= 1,000$)						
4	**0.063**	**0.256**	**0.614**	**0.171**	**0.708**	**0.516**
6	0.000	0.001	0.013	**0.080**	**0.309**	**0.726**
8	0.000	0.000	0.000	0.002	**0.333**	**0.465**
10	0.000	0.000	0.000	0.000	0.000	**0.060**
$D_2^\dagger$ vs. Gamma (Sample size $= 1,000$)						
4	**0.439**	**0.662**	**0.324**	**0.263**	**0.728**	**0.492**
6	0.000	**0.807**	**0.483**	**0.372**	**0.236**	**0.602**
8	0.000	0.000	0.000	**0.423**	**0.362**	**0.951**
10	0.000	0.000	0.000	0.000	0.000	**0.285**

deviation from the respective model distributions, as indicated by small P-values, sets in rapidly as the Pólya-Aeppli regime in the bottom left hand corner of the table is approached, though in general the Gamma distribution maintains an acceptible fit further into this regime. Note also that, because the model distributions are never an exact description of the true $D_2^\dagger$ distribution, the p-values decrease as the sample sizes increase. That is, the location of a boundary beyond which the Gamma approximation becomes invalid cannot be unambiguously specified.

To explore the accuracy of the above approximations for estimating p-values in the tail of the distribution, in Figs. 3 and 4 we show Q-Q plots of the empirically

determined $D_2^\dagger$ quantiles against quantiles of the Normal and Gamma distributions. Again the Gamma distribution outperforms the Normal distribution, accurately estimating p-values out to the 99.0% percentile in most cases of biological interest. However we would caution against trusting p-values obtained from the Gamma approximation as anything other than a qualitative guide of significance beyond this point.

4 Centred exact and approximate word match statistics

In order to overcome the problem of single sequence noise, one might also define a centred version of the original D_2 statistic in terms of the centred word counts defined in Eq. (2), namely

$$D_2^C(k) = \sum_{w \in \mathscr{A}^k} \tilde{X}_w \tilde{Y}_w. \tag{15}$$

As shown in the appendix for the more general case of a weighted word match statistic, the variance of $D_2^C(k)$ is precisely the $O(n_A n_B)$ part of that of $D_2(k)$, and the remaining, potentially troublesome third order part of $\mathrm{Var}\,(D_2(k))$ is entirely composed of single sequence noise.

Another advantage of D_2^C is that, like D_2, it can be computed in order $O(k(n_A+n_B))$ time. To see this, consider the expansion

$$\begin{aligned} D_2^C(k) &= D_2(k) - n_A \sum_{w \in \mathscr{A}^k} p_w Y_w - n_B \sum_{w \in \mathscr{A}^k} p_w X_w + n_A n_B \sum_{w \in \mathscr{A}^k} p_w^2 \\ &= D_2(k) - n_A \sum_{j=1}^{n_B} p_{(B_i \ldots B_{i+k-1})} \\ &\qquad - n_B \sum_{i=1}^{n_A} p_{(A_i \ldots A_{i+k-1})} + n_A n_B \sum_{w \in \mathscr{A}^k} p_w^2, \end{aligned} \tag{16}$$

where we have used the result

$$\sum_w p_w X_w = \sum_{i=1}^{n_A} \sum_w p_w I_i^A(w) = \sum_{i=1}^{n_A} p_{(A_i \ldots A_{i+k-1})}, \tag{17}$$

which uses the indicator variables defined by Eq. (6). For iid sequences, the probability of the observed word at position i in sequence $\mathbf{A}$ is simply $p_{(A_i \ldots A_{i+k-1})} = f_{A_i} \ldots f_{A_{i+k-1}}$. Each of the second and third terms of Eq. (16) requires calculating

DOI: 10.2202/1544-6115.1724

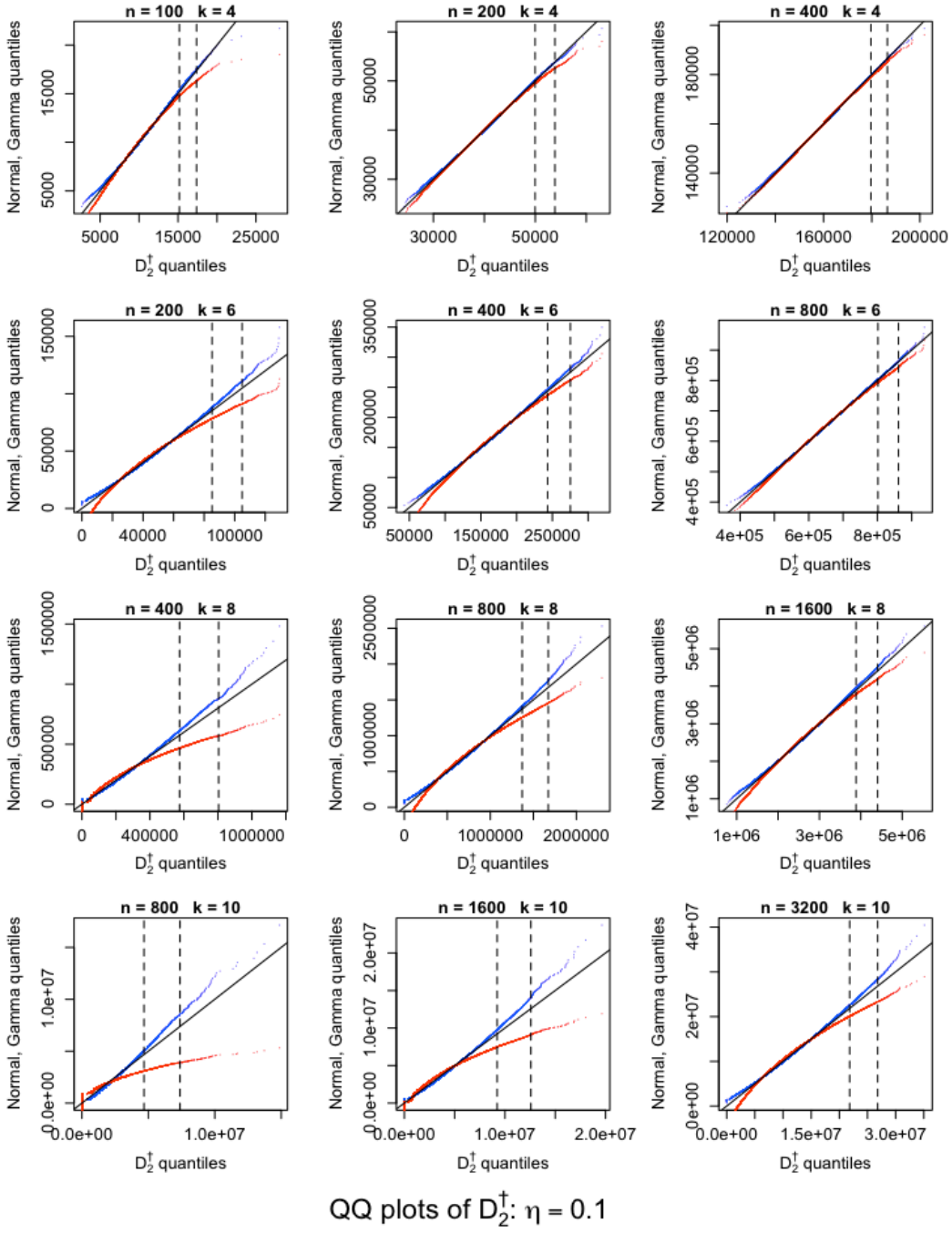

Figure 3: QQ-plots comparing the $D_2^\dagger$ statistic generated from 100,000 pairs of random iid sequences from a 4-letter alphabet with asymmetry parameter $\eta = 0.1$ with Normal (red) and Gamma (blue) distributions. The vertical dotted lines indicate the 99% and 99.9% sample percentiles.

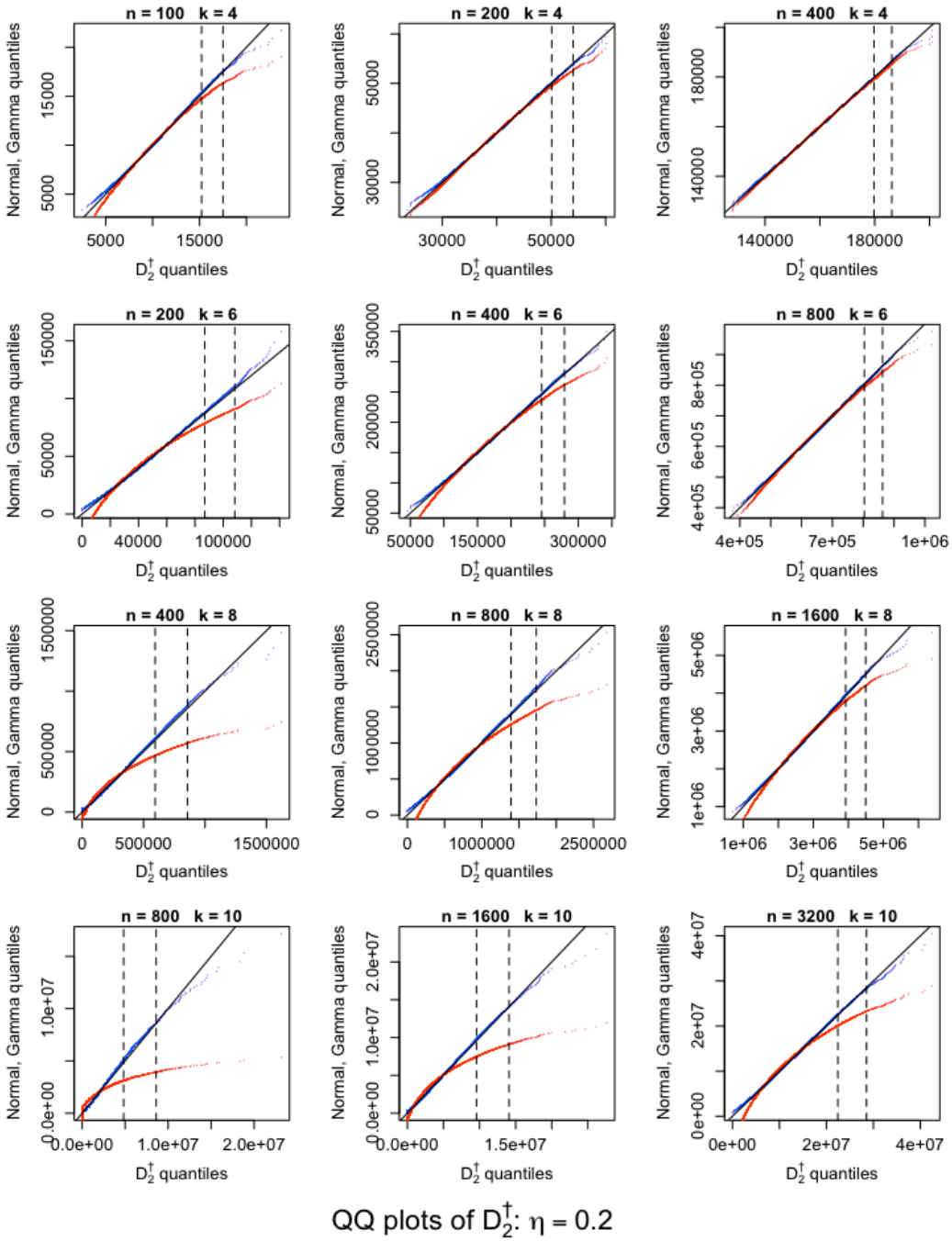

Figure 4: The same as Fig. 3, except with $\eta = 0.2$

DOI: 10.2202/1544-6115.1724

Table 2: The same as for Table 1, except for the D_2^C statistic at $\eta = 0.1$. The shifted Gamma random variable is explained in the text.

$k \setminus n$	100	200	400	800	1600	3200
D_2^C vs. Normal (Sample size $= 10,000$)						
4	0.000	0.032	**0.182**	**0.867**	**0.822**	**0.077**
6	0.000	0.000	0.001	0.026	0.011	**0.103**
8	0.000	0.000	0.000	0.000	0.000	**0.085**
10	0.000	0.000	0.000	0.000	0.000	0.000
D_2^C vs. shifted Gamma (Sample size $= 10,000$)						
4	**0.236**	**0.309**	**0.103**	**0.540**	**0.672**	**0.066**
6	0.000	0.000	**0.089**	**0.528**	**0.212**	0.023
8	0.000	0.000	0.000	0.018	**0.614**	**0.368**
10	0.000	0.000	0.000	0.000	0.000	0.000
D_2^C vs. Normal (Sample size $= 1,000$)						
4	0.004	**0.517**	**0.682**	**0.085**	**0.003**	**0.355**
6	0.000	0.004	0.017	**0.725**	**0.248**	**0.771**
8	0.000	0.000	0.000	0.000	0.001	**0.155**
10	0.000	0.000	0.000	0.000	0.000	0.000
D_2^C vs. shifted Gamma (Sample size $= 1,000$)						
4	**0.345**	**0.302**	**0.692**	0.049	0.004	**0.360**
6	0.000	**0.494**	**0.249**	**0.940**	**0.292**	**0.715**
8	0.000	0.000	0.000	**0.450**	**0.181**	**0.138**
10	0.000	0.000	0.000	0.000	0.000	**0.567**

the probabilities of each of the words occurring in either of the sequences, which requires $O(k(n_A + n_B))$ time, and the fourth term is a simple constant.

As for the previous word count statistics, we find that the empirical distribution of D_2^C is better approximated by a Gamma distribution than a Normal distribution. Since its mean is zero, we compare D_2^C to a shifted random variable $X_\Gamma - E[D_2]$, where X_Γ is a Gamma random variable whose mean is equal to the theoretical $E[D_2]$ and variance is equal to the theoretical $\mathrm{Var}\,(D_2^C)$. QQ-plots of D_2^C samples obtained from 100,000 randomly generated pairs of iid sequences are shown in Figs. 5 and 6, and Kolmogorov-Smirnov p-values for D_2^C samples from 10,000 and 1,000 pairs of i.i.d. sequences compared against Normal and shifted Gamma distributions are given in Table 2.

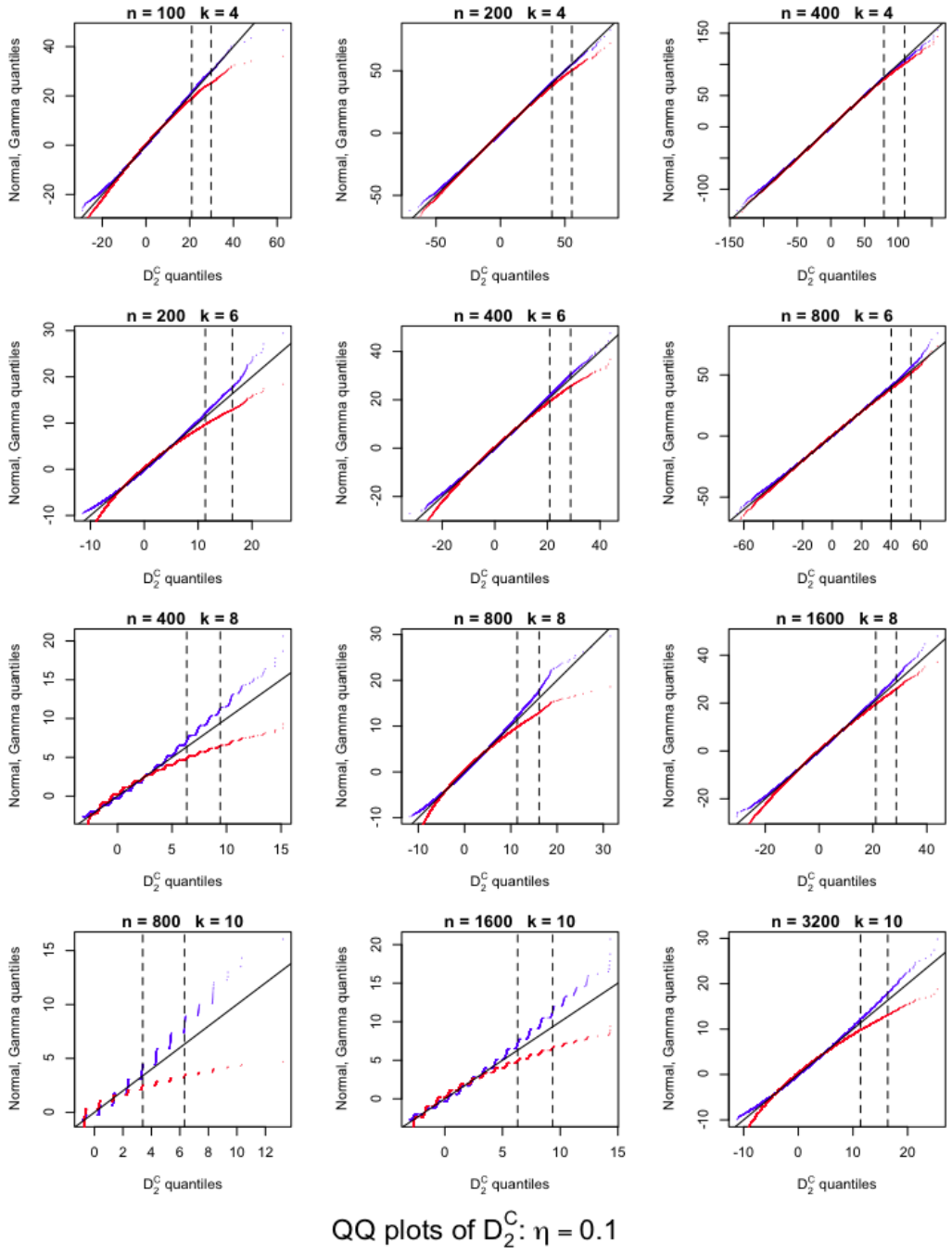

Figure 5: QQ-plots comparing the D_2^C statistic generated from 100,000 pairs of random iid sequences from a 4-letter alphabet with asymmetry parameter $\eta = 0.1$ with Normal (red) and shifted Gamma (blue) distributions. The vertical dotted lines indicate the 99% and 99.9% sample percentiles.

DOI: 10.2202/1544-6115.1724

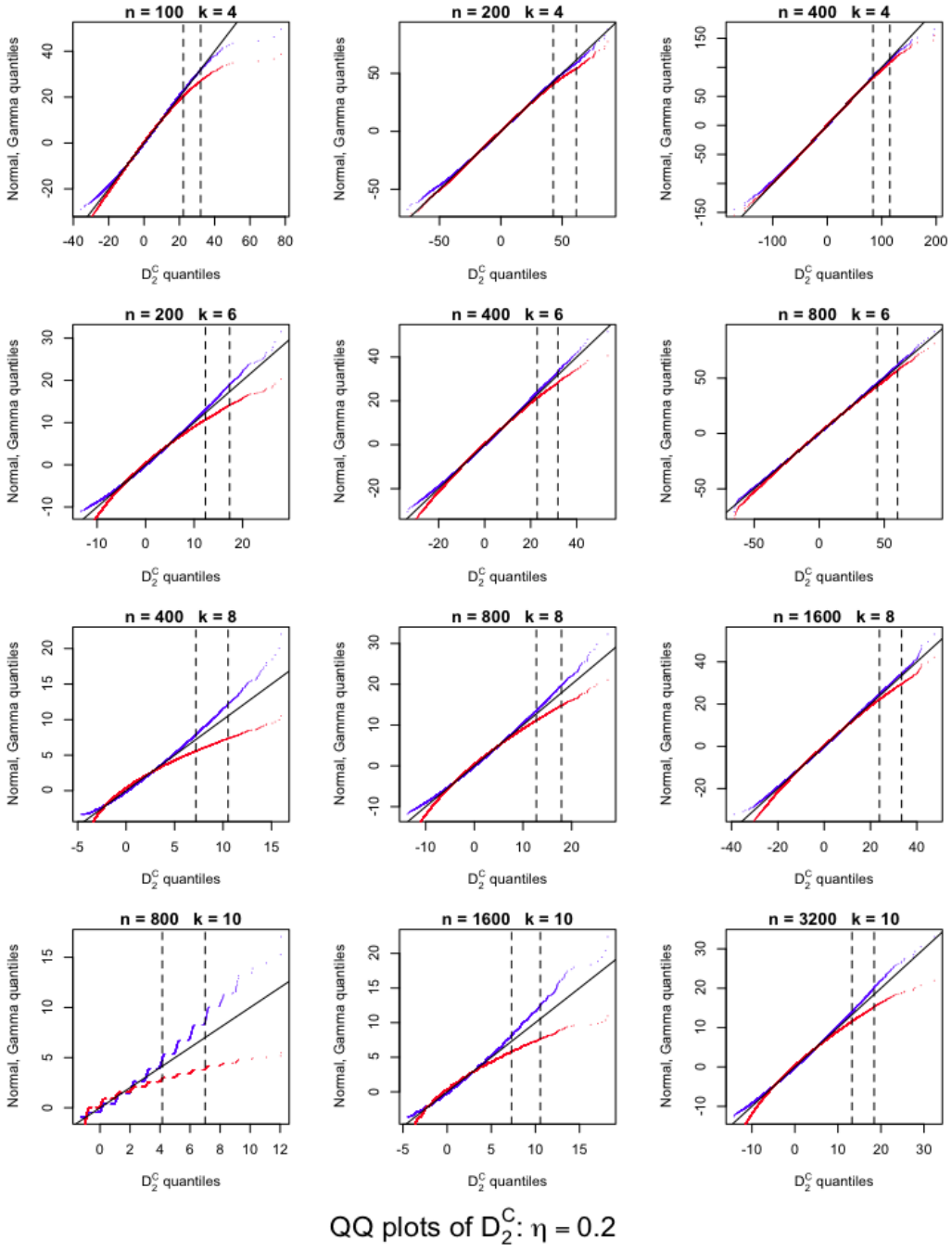

Figure 6: The same as Fig. 5, except with $\eta = 0.2$

With applications to regulatory motifs in mind, Burden et al. (2008) and Forêt et al. (2009a) considered an extension of the usual $D_2(k)$ statistic to an approximate word match statistic

$$D_2(k,t) = \sum_{w \in \mathscr{A}^k} \sum_{\{v:\Delta(w,v)\leq t\}} X_w Y_v, \quad t = 0,1,\ldots,k, \tag{18}$$

where $\Delta(w,v)$ is the number of mismatches between two words w and v. Like the exact word match statistic, this statistic was also shown in general to be well approximated by a Gamma random variable with a theoretically calculated mean and variance. Bearing in mind once again the avoidance of single sequence noise, we consider in the next section a centred version

$$D_2^C(k,t) = \sum_{w \in \mathscr{A}^k} \sum_{\{v:\Delta(w,v)\leq t\}} \tilde{X}_w \tilde{Y}_v, \quad t = 0,1,\ldots,k. \tag{19}$$

which, like $D_2^C(k)$, is well represented by a shifted Gamma random variable (data not shown). We note however that the approximate centred word match statistic $D_2^C(k,t)$ is somewhat slower to calculate than its exact match counterpart $D_2^C(k)$ as the summands of the second and third terms of Eq. (16) must be replaced by a sum of the probabilities of all words with up to t mismatches relative to the words at positions j in **B** and i in **A** respectively.

5 Application to the discovery of cis-regulatory modules

The effectiveness of the D_2 statistic as a tool for the discovery of *cis-regulatory modules* (CRMs) was recently explored by Forêt et al. (2009a). A dataset constructed by Kantorovitz et al. (2007) was used, which consisted of two parts: a 'positive' data set consisting of seven sets of sequences from *Drosophila* and human known to contain CRMs, and a 'negative' data set constructed from randomly chosen non-coding sequences from the same species. The following problem was addressed: given a set of sequences known to contain CRMs, and a query sequence, can the query sequence be classified as containing similar CRMs or not?

The following experiment was set up: each sequence in each positive set was selected in turn as the query sequence and compared to both the remaining positive sequences of this set and to the corresponding negative sequences using the D_2 statistic. The query sequences were then screened to accept only those for which the smallest p-value of all comparisons was less than 0.01. A stringent criterion was used, namely, a positive query was considered to be correctly classified if the smallest p-value was obtained with another sequence of the positive set. Here this

DOI: 10.2202/1544-6115.1724

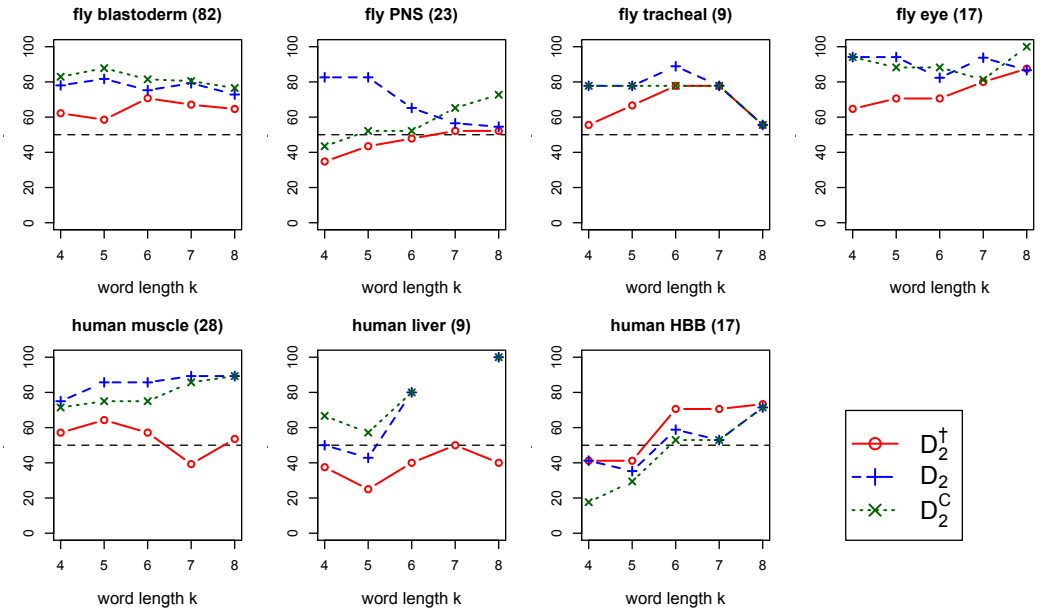

Figure 7: The percentage of times a query sequence was correctly classified as containing CRMs by testing using D_2 and $D_2^\dagger$ and D_2^C statistics. The numbers in parentheses are the number of positive control sequences in each set. Sequence lengths are typically in the range $n = 100$ to 2000, with a median close to 600. Percentages are only plotted if at least 4 query sequences survived the screening requirement that the minimum p-value should be less than 0.01.

experiment is repeated using the $D_2^\dagger$ (or equivalently, D_2^*), D_2^C and D_2 statistics. The p-values were calculated using the Gamma approximations to each statistic. The asymmetry parameters η used in the Gamma approximations were estimated from relative letter frequencies within each entire data set, and were in the range $0.1 \lesssim \eta \lesssim 0.2$.

The results are illustrated in Fig. 7. For both D_2 and D_2^C, a good sensitivity is achieved for most data sets, with typically 80% or more of the sequences correctly classified for at least one choice of word length using the above stringent criterion. In most cases the performance of $D_2^\dagger$ was noticeably poorer.

We have also carried out the above tests using the approximate word match statistics $D_2(k,t)$ and $D_2^C(k,t)$ for t up to three mismatches as test statistics. The results for the percentage of times a query sequence was correctly classified as containing CRMs are shown in Fig. 8. In each data set we observe that the optimal choice of parameters includes a combination with $t = 0$ mismatches, suggesting that

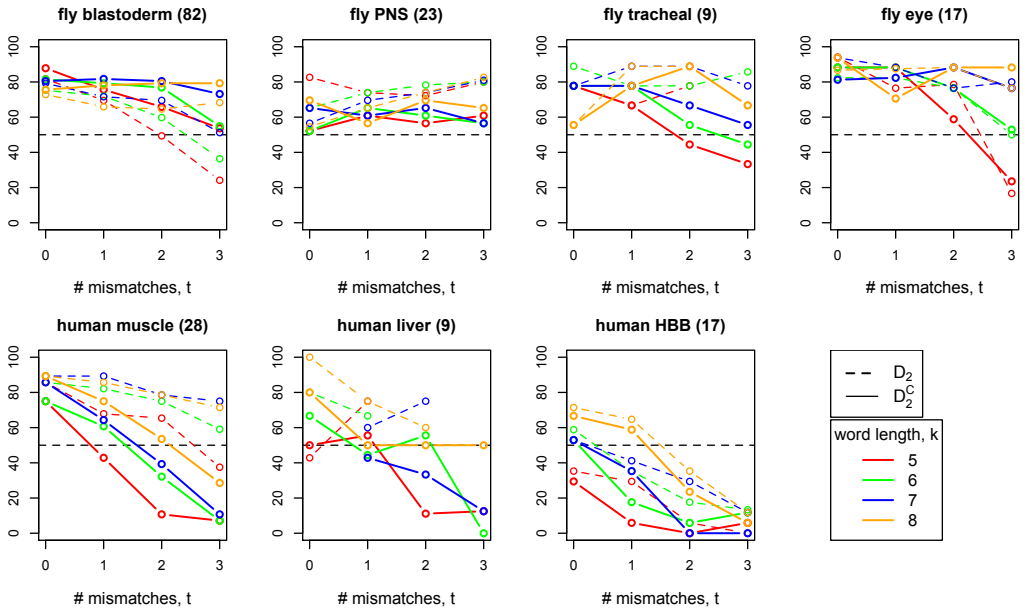

Figure 8: The percentage of times a query sequence was correctly classified as containing CRMs by testing using $D_2(k,t)$ and $D_2^C(k,t)$ statistics for up to $t = 3$ mismatches for the same data set as Fig. 7.

the extra considerable computational complexity involved in including mismatches confers no benefit.

As a general rule, in applications to the detection of CRMs, the problem of choosing between the word match statistic D_2, $D_2^\dagger$ and D_2^C, and choosing the word length k for a given data set could easily be solved by the above approach, namely by determining a set of positive sequences and using these to estimate appropriate parameters before comparing the query sequence(s) to them.

6 Discussion

Our main purpose is to examine the relative merits of variants of the D_2 word match statistic as an alignment-free method of biological sequence comparison. It has been argued by Reinert et al. (2009) and Wan et al. (2010) that, in its original form, the D_2 statistic is not suitable for biological sequence comparison, firstly because it is dominated by single sequence noise, and secondly because it performs badly in tests of its power to detect sequence relatedness in synthetically generated data. These claims are based on analyses which have concentrated on the asymptotic behaviour of D_2 and related statistics as the sequence lengths become large. Here

DOI: 10.2202/1544-6115.1724

we have concentrated our efforts on more moderate sequence lengths relevant to biological applications, such as CRMs, of up to about 3,000 letters.

The claim by Reinert et al. (2009) that the D_2 statistic is dominated by single sequence noise is based on an observation that for sequences of length n, $\mathrm{Var}\,(D_2)$ under a null iid hypothesis is composed of an $O(n^3)$ part due to fluctuations in word counts in each of the two sequences, and an $O(n^2)$ part due to correlated fluctuations in word counts in both sequences. In the Appendix we have verified the form of these contributions for the more general case of a weighted word match statistic, D_2^W defined by Eq. (20), using exact analytic formulae for the variance. Reinert et al.'s argument is that any genuine signal of simultaneous over-representation of words above the $O(n^2)$ part is masked by the spurious $O(n^3)$ part. While the claim certainly has merit in the asymptotic limit of very long sequences, we find that for parameter values relevant to many biological applications that single sequence noise is unlikely to be a serious problem. For instance, the $O(n^3)$ part contributes less than half of $\mathrm{Var}\,(D_2)$ for word lengths greater than $k = 6$ letters, sequence lengths less than about $n = 1000$ letters, and the moderate values of letter distribution asymmetry ($\eta \lesssim 0.2$) observed in most genomes (see Figure 1).

Nevertheless, it is straightforward to remove the $O(n^3)$ part of the variance by defining word match statistics in terms of the mean-centred word counts defined by Eq. (2). We have considered two such statistics, D_2^* originally proposed by Reinert et al. (2009), and a simple centred version of D_2, which we call D_2^C, defined by Eq. (15). We obtain exact analytic calculations of the variance of these statistics, and demonstrate empirically that for biologically relevant letter distributions and a broad range of word and sequence lengths, they can be extremely well approximated up to the 99th percentile by a shifted Gamma distribution by making use of the known mean and variance. In fact the statistic D_2^* is shown to be equivalent up to a simple scaling and additive constant to a weighted, non-centred statistic, $D_2^\dagger$, defined by Eq. (11).

Thus either of these statistics potentially provide a measure of sequence similarity which can be evaluated rapidly in time linear in the sequence lengths, and for which accurate p-values up to the 99th percentile under the null hypothesis of iid letters can be readily obtained.

These statistics therefore have computational advantages over the Schepp word match count D_2^S (Eq. (4)), also introduced by Reinert et al. (2009), whose reported benefit is that it is asymptotically Normal for moderate sequence lengths. However, there is no known exact formula for $\mathrm{Var}\,(D_2^S)$, which must be estimated numerically. Furthermore evaluation of D_2^S requires a sum over all possible k-words, which has a high computational cost, particularly for a large alphabet such as the set of 20 amino acids needed for protein sequences.

Rather than using simulated data, we have reanalysed a data set constructed by Kantorovitz et al. (2007) for the purpose of testing the effectiveness of similarity measures in discovering cis-reulatory modules. Our analysis indicates that D_2^C performs roughly as well as or slightly better than the previously tested exact word-match D_2 statistic, whereas $D_2^\dagger$ (that is equivalent to D_2^*) performs noticeably worse. We have also considered approximate-match versions of D_2 and D_2^C which allow a certain prespecificed number of mismatches, but we find that in general that this leads to worse performance. For detecting CRMs against training data we would therefore recommend checking both the exact $D_2(k)$ and $D_2^C(k)$ statistics using the method outlined in Section 5 to determine the most appropriate statistic and optimum word length.

Finally, we note that as with any test of sequence similarity, results must be interpreted with caution. Firstly it is difficult to judge the sensitivity of any particular statistic for a given biological problem given that it is difficult to know the true nature of the alternate hypothesis one is testing for. Choosing an alternate hypothesis to measure the power of a sequence similarity test remains as much an art form as choosing the statistic itself. Secondly, the specificity of a test depends on having accurately judged the underlying null hypothesis: obtaining a small p-value may indicate a common origin for the sequences in question, or it may simply indicate that the underlying assumptions on which the test is based do not hold. Recent studies of the k-word spectra of several entire genomes in Chor, Horn, Goldman, Levy, and Massingham (2009) suggest a second order Markovian dependency may in general be more appropriate than the iid null hypothesis used herein. In future work we intend to extend our analysis of the distributional properties of D_2, D_2^C and D_2^* to include Markovian dependencies.

Appendix: Determination of $\mathrm{Var}\left(D_2^{WC}\right)$

We give a derivation of the result that variance of the weighted, centred word match statistic, under the iid null hypothesis, is precisely the $O(n_A n_B)$ part of the variance of the uncentred statistic.

The weighted word match statistic is a generalisation of the $D_2(k)$ statistic defined by

$$D_2^W(k) = \sum_{w,v \in \mathscr{A}^k} X_w \beta_{w,v} Y_v, \tag{20}$$

where $\beta_{w,v}$ is a fixed $d^k \times d^k$ symmetric matrix defined between any two k-words w and v and $d = |\mathscr{A}|$ is the alphabet size. The β-matrix is assumed to take the form of a product: $\beta_{wv} = \beta_{(w_1,\ldots,w_k),(v_1,\ldots,v_k)} = \beta_{w_1 v_1} \ldots \beta_{w_k v_k}$. The original D_2 statistic

DOI: 10.2202/1544-6115.1724

and the $D_2^\dagger$ statistic of Section 3 are particular cases of D_2^W. Burden et al. (2011) give the formulae set out below for the mean and variance of D_2^W. As in previous work (Forêt, Wilson, and Burden (2009a,b)), periodic boundary conditions are imposed on both sequences, that is, we define $A_i = A_{i-n_A}$, $i = n_A+1,\ldots n_A+k-1$, and similarly for sequence **B**. The periodic boundary conditions are a minor technicality, easily implemented in practical applications. We further assume that the probability of the letter $a \in \mathscr{A}$ occurring at any given site in either sequence is f_a, where $\sum_{a\in\mathscr{A}} f_a = 1$.

We begin with the following definitions. For $a,b \in \mathscr{A}$, set

$$\eta_a = \sqrt{f_a}, \qquad M_{ab} = \eta_a \beta_{ab} \eta_b, \qquad \pi_t = \eta' M^{t-1} \eta, \quad t = 1,2,\ldots,$$

where $\eta' = (\eta_1,\ldots,\eta_L)$, η is the corresponding column vector and M is the $L\times L$ matrix with elements M_{ab}. We also define

$$\phi = \sum_{a,b\in\mathscr{A}} f_a f_b \beta_{ab}^2. \tag{21}$$

For the mean, one obtains by analogy with (Forêt et al., 2009b, Eq. (4)) the result

$$E[D_2^W] = n_A n_B \pi_2^k. \tag{22}$$

Writing the variance of Var (D_2^W) as a sum of cross-covariances, gives a sum of five contributions:

$$\text{Var}\,(D_2^W) = V_1 + V_2 + V_3 + V_4 + V_5. \tag{23}$$

Analogous to (Forêt et al., 2009b, Eqs. (10), (14), (17), (20) and (26)) we find

$$V_1 = n_A n_B \left(\phi^k - \pi_2^{2k}\right), \tag{24}$$

$$V_2 = n_A n_B (n_A + n_B - 4k + 2)\left[\pi_3^k + 2\sum_{s=1}^{k-1} \pi_2^{2s}\pi_3^{k-s} - (2k-1)\pi_2^{2k}\right], \tag{25}$$

$$V_3 = 2n_A n_B \left[\phi\pi_2^2 \frac{\phi^{k-1} - \pi_2^{2k-2}}{\phi - \pi_2^2} - (k-1)\pi_2^{2k}\right], \tag{26}$$

$$V_4 = 4n_A n_B \sum_{t=1}^{k-1}\sum_{s=0}^{t-1} \left({\pi_2}^{2s} \pi_{2\nu+3}{}^{\rho} \pi_{2\nu+1}{}^{t-s-\rho} - {\pi_2}^{2k}\right), \tag{27}$$

and

$$V_5 = 2n_A n_B \sum_{r,t=1}^{k-1} \left[\left(\prod_{i=1}^{t} \pi_{l_i}\right)\left(\prod_{j=1}^{r} \pi_{m_j}\right) - {\pi_2}^{2k}\right]. \tag{28}$$

In the contributions V_4 and V_5, the following definitions have been used:

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

$$\begin{aligned} l_i &= 1+2\eta+\left\{ \begin{array}{ll} 1 & \text{if } i \le \zeta \\ 0 & \text{otherwise} \end{array} \right\} + \left\{ \begin{array}{ll} 1 & \text{if } i \le \zeta - r \\ 0 & \text{otherwise} \end{array} \right\} \\ m_j &= 1+2\eta+\left\{ \begin{array}{ll} 1 & \text{if } j \le \zeta \\ 0 & \text{otherwise} \end{array} \right\} + \left\{ \begin{array}{ll} 1 & \text{if } j \le \zeta - t \\ 0 & \text{otherwise} \end{array} \right\}, \end{aligned}$$

where

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

and $\lfloor\ \rfloor$ indicates the integer part.

The centred version of the weighted word match statistic is defined as

$$D_2^{WC}(k) = \sum_{w,v \in \mathscr{A}^k} \tilde{X}_w \beta_{w,v} \tilde{Y}_v, \tag{29}$$

where $\tilde{X}$ and $\tilde{Y}$ are defined by Eq. (2). The statistics D_2^C of Section 4 and D_2^* of Section 3 are particular cases of D_2^{WC}. The mean $E[D_2^{WC}]$ is clearly zero. Here we show that Var (D_2^{WC}) is precisely the $O(n_A n_B)$ part of Var (D_2^W), and that the remaining $O(n_A n_B(n_A+n_B))$ part of Var (D_2^W) is principally composed of single sequence noise in the form of contributions from Var $(\sum_w p_w \beta_{w,v} X_v)$ and Var $(\sum_w p_w \beta_{w,v} Y_v)$.

From the above definitions,

$$\begin{aligned} \text{Var}\,(D_2^{WC}(k)) = \text{Var}\,&\left(D_2^W(k)\right) \\ &+n_B^2 \text{Var}\quad \sum_{w,v} p_w \beta_{w,v} X_v \Bigg) + n_A^2 \text{Var}\quad \sum_{w,v} p_w \beta_{w,v} Y_v \Bigg) \\ &-2n_B \text{Cov}\quad D_2^W, \sum_{w,v} p_w \beta_{w,v} X_v \Bigg) \\ &-2n_A \text{Cov}\quad D_2^W, \sum_{w,v} p_w \beta_{w,v} Y_v \Bigg). \end{aligned} \tag{30}$$

DOI: 10.2202/1544-6115.1724

First term

The first term, the variance of $D_2^W(k)$ given above, is of the form

$$\text{Var}\left(D_2^W(k)\right) = n_A n_B(n_A+n_B)U + n_A n_B V, \tag{31}$$

where U and V depend on the word length k and letter distribution f_a but not the sequence lengths n_A and n_B. More specifically,

$$U = \pi_3^k + 2\sum_{s=1}^{k-1} \pi_2^{2s}\pi_3^{k-s} - (2k-1)\pi_2^{2k}. \tag{32}$$

Second and third terms

To calculate $\text{Var}\left(\sum_w p_w \beta_{w,v} X_v\right)$ we make use of the indicator random variable $I_i^A(v)$ for the occurrence of word v at position i in sequence **A** to obtain

$$\begin{aligned}\sum_{w,v} p_w \beta_{w,v} X_v &= \sum_{i=1}^{n_A}\sum_{w,v}(f_{w_1}\beta_{w_1 v_1}\ldots f_{w_k}\beta_{w_k v_k})I_i^A(v) \\ &= \sum_{i=1}^{n_A} F_i, \end{aligned}\tag{33}$$

where

$$F_i = \left(\sum_{a\in\mathscr{A}} f_a \beta_{aA_i}\right)\ldots\left(\sum_{a\in\mathscr{A}} f_a \beta_{aA_{i+k-1}}\right). \tag{34}$$

Then

$$\text{Var}\left(\sum_w p_w \beta_{w,v} X_v\right) = \sum_{i=1}^{n_A}\text{Var}\,(F_i) + 2\sum_{1\le i<j\le n_A}\text{Cov}\,(F_i,F_j). \tag{35}$$

The first of these terms is

$$\begin{aligned}\sum_{i=1}^{n_A}\text{Var}\,(F_i) &= \sum_{i=1}^{n_A}\left\{E[F_i^2] - E[F_i]^2\right\} \\ &= \sum_{i=1}^{n_A}\left\{E\left[\left(\sum_a f_a\beta_{aA_i}\right)^2\right]^k - E\left[\sum_a f_a\beta_{aA_i}\right]^{2k}\right\} \\ &= \sum_{i=1}^{n_A}\left\{\left(\sum_{a,b,c} f_a\beta_{ab}f_b\beta_{bc}f_c\right)^k - \left(\sum_{a,b} f_a\beta_{ab}f_b\right)^{2k}\right\} \\ &= n_A\left(\pi_3^k - \pi_2^{2k}\right), \end{aligned}\tag{36}$$

where we have made use of the iid property of the sequence.

The second of these terms is a sum of covariances which are zero unless the k-words beginning at locations i and j overlap. Therefore we can set $j = i+s$ and write the second term in Eq. (35) as

$$\begin{aligned}
2\sum_{1\le i<j\le n_A}\mathrm{Cov}\,(F_i,F_j) &= 2\sum_{i=1}^{n_A}\sum_{s=1}^{k-1}\mathrm{Cov}\,(F_i,F_{i+s}) \\
&= 2\sum_{i=1}^{n_A}\sum_{s=1}^{k-1}\left\{E\left[\left(\sum_a f_a\beta_{aA_i}\right)\cdots\left(\sum_a f_a\beta_{aA_{i+k-1}}\right)\times\right.\right. \\
&\qquad\left.\left.\left(\sum_b f_b\beta_{bA_{i+s}}\right)\cdots\left(\sum_b f_b\beta_{bA_{i+s+k-1}}\right)\right]- \right. \\
&\qquad E\left[\left(\sum_a f_a\beta_{aA_i}\right)\cdots\left(\sum_a f_a\beta_{aA_{i+k-1}}\right)\right]\times \\
&\qquad\left.E\left[\left(\sum_b f_b\beta_{bA_{i+s}}\right)\cdots\left(\sum_b f_b\beta_{bA_{i+s+k-1}}\right)\right]\right\} \\
&= 2\sum_{i=1}^{n_A}\sum_{s=1}^{k-1}\left\{E\left[\sum_a f_a\beta_{aA_i}\right]^s E\left[\sum_{a,b} f_a\beta_{aA_i}f_b\beta_{bA_i}\right]^{k-s} E\left[\sum_b f_b\beta_{bA_i}\right]^s - \right. \\
&\qquad\left.E\left[\sum_a f_a\beta_{aA_i}\right]^{2k}\right\} \\
&= 2\sum_{i=1}^{n_A}\sum_{s=1}^{k-1}\left(\pi_2^s\pi_3^{k-s}\pi_2^s-\pi_2^{2k}\right) \\
&= 2n_A\left\{\sum_{s=1}^{k-1}\pi_2^{2s}\pi_3^{k-s}-(k-1)\pi_2^{2k}\right\}. \qquad (37)
\end{aligned}$$

Adding Eqs. (36) and (37) gives

$$\mathrm{Var}\left(\sum_{w,v}p_w\beta_{w,v}X_v\right) = n_A U, \qquad (38)$$

with U defined in Eq. (32). An analogous result holds for $\mathrm{Var}\,(\sum_w p_w Y_w)$. The contribution from the second and third terms of Eq. (30) is then

$$n_A n_B(n_A+n_B)U. \qquad (39)$$

DOI: 10.2202/1544-6115.1724

Fourth and fifth terms

To calculate $\mathrm{Cov}\left(D_2^W, \sum_{w,v} p_w \beta_{w,v} X_v\right)$, first note from Eqs. (20) and (7) that

$$D_2^W(k) = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \beta_{(A_i \ldots A_{i+k-1}),(B_j \ldots B_{j+k-1})}. \tag{40}$$

Then, using Eq. (33) and the fact that the covariance in the second line below is zero unless words at positions i and ℓ in sequence **A** overlap,

$$\begin{aligned} &\mathrm{Cov}\left(D_2^W, \sum_{w,v} p_w \beta_{w,v} X_v\right) \\ &= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sum_{\{\ell:|\ell-i|<k\}} \mathrm{Cov}\left(\beta_{(A_i \ldots A_{i+k-1}),(B_j \ldots B_{j+k-1})}, F_\ell\right). \\ &= \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \sum_{\{\ell:|\ell-i|<k\}} E\left[\beta_{(A_i \ldots A_{i+k-1}),(B_j \ldots B_{j+k-1})} F_\ell\right] - n_A n_B (2k-1)\pi_2^{2k}. \end{aligned} \tag{41}$$

In the last line the definition Eq. (34) and the iid assumption have been used to evaluate $E[\beta_{(A_i \ldots A_{i+k-1}),(B_j \ldots B_{j+k-1})}] \times E[F_\ell] = \pi_2^{2k}$.

For fixed positions i and $\ell = i+s$ in sequence A, where $s = 0, \ldots, k-1$, and j in sequence B, the term $E\left[\beta_{(A_i \ldots A_{i+k-1}),(B_j \ldots B_{j+k-1})} F_\ell\right]$ can be written as a sum over all combinations of letters $(a_1, \ldots, a_{k+s}, b_1, \ldots, b_k) \in \mathscr{A}^{2k+s}$ as illustrated in Fig. 9

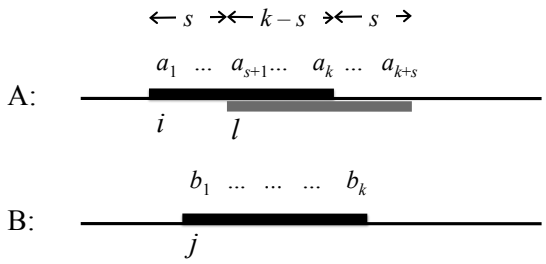

Figure 9: Arrangement of letters contributing to $E\left[\beta_{(A_i \ldots A_{i+k-1}),(B_j \ldots B_{j+k-1})} F_\ell\right]$. The diagram is shown for the case $i \le \ell < i+k$. It is straightforward to show that Eq. (42) with $s = |\ell - i|$ holds for all $i-k \le \ell \le i+k$.

as

$$
\begin{aligned}
& E\left[\beta_{(A_i\ldots A_{i+k-1}),(B_j\ldots B_{j+k-1})}F_\ell\right] \\
&= \sum_{a_1,\ldots,a_{k+s},b_1,\ldots,b_k\in\mathscr{A}} \Pr(\text{config. in Fig. 9}) \times \beta_{(a_1\ldots a_k),(b_1\ldots b_k)} \times \\
&\qquad \left(\sum_{c\in\mathscr{A}} f_c\beta_{ca_{s+1}}\right)\cdots\left(\sum_{c\in\mathscr{A}} f_c\beta_{ca_{s+k}}\right) \\
&= \sum_{a_1,\ldots,a_{k+s},b_1,\ldots,b_k\in\mathscr{A}} f_{a_1}\ldots f_{a_{k+s}}f_{b_1}\ldots f_{b_k}\times\beta_{a_1b_1}\ldots\beta_{a_kb_k}\times \\
&\qquad \sum_{c_1,\ldots,c_k\in\mathscr{A}} f_{c_1}\ldots f_{c_k}\beta_{c_1a_{s+1}}\ldots\beta_{c_ka_{s+k}} \\
&= \sum_{a_1,\ldots,a_{k+s},b_1,\ldots,b_k,c_1,\ldots,c_k\in\mathscr{A}} f_{a_1}\beta_{a_1b_1}f_{b_1}\ldots f_{a_s}\beta_{a_sb_s}f_{b_s}\times \\
&\qquad f_{c_1}\beta_{c_1a_{s+1}}f_{a_{s+1}}\beta_{a_{s+1}b_{s+1}}f_{b_{s+1}}\ldots f_{c_{k-s}}\beta_{c_{k-s}a_k}f_{a_k}\beta_{a_kb_k}f_{b_k}\times \\
&\qquad f_{c_{k-s+1}}\beta_{c_{k-s+1}a_{k+1}f_{k+1}}\ldots f_{c_k}\beta_{c_ka_{k+s}}f_{k+s} \\
&= \left(\sum_{a,b\in\mathscr{A}} f_a\beta_{ab}f_b\right)^s\left(\sum_{c,a,b\in\mathscr{A}} f_c\beta_{ca}f_a\beta_{ab}f_b\right)^{k-s}\left(\sum_{c,a\in\mathscr{A}} f_c\beta_{ca}f_a\right)^s \\
&= \pi_2^{2s}\pi_3^{k-s}. \qquad (42)
\end{aligned}
$$

A similar calculation leads to the same result for $\ell = i-s$, where $s = 1,\ldots,k-1$.

Substituting Eq. (42) into Eqs. (41), and using (32), we obtain

$$
\begin{aligned}
\text{Cov}\quad D_2^W, \sum_{w,v} p_w\beta_{w,v}X_v\Big) &= n_An_B\quad \pi_3^k + 2\sum_{s=1}^{k-1}\pi_2^{2s}\pi_3^{k-s} - (2k-1)\pi_2^{2k}\Big) \\
&= = n_An_BU, \qquad (43)
\end{aligned}
$$

with a similar result holding for $\text{Cov}\,(D_2^W, \sum_{w,v} p_w\beta_{w,v}Y_v)$. Thus the contribution to the fourth and fifth terms of Eq. (30) is

$$-2n_An_B(n_A+n_B)U. \qquad (44)$$

Finally, adding Eqs. (31), (39) and (44),

$$\text{Var}\,(D_2^{WC}(k)) = n_An_BV. \qquad (45)$$

That is, the variance of $D_2^{WC}(k)$ is precisely the $O(n_An_B)$ part of the variance of $D_2^W(k)$, as stated. Furthermore, the result (39) shows that the $O(n_An_B(n_A+n_B))$ part can be written entirely in terms of either of the single sequence variances $\text{Var}\,(p_wX_w)$ or $\text{Var}\,(p_wY_w)$.

DOI: 10.2202/1544-6115.1724

References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman (1997): "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.*, 25, 3389–402.

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

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

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

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

Forêt, S., S. R. Wilson, and C. J. Burden (2009a): "Characterizing the *D*2 statistic: Word matches in biological sequences," *Stat. Appl. Genet. Mo. B.*, 8, Article 43.

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

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

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

Khuu, P., M. Sandor, J. DeYoung, and P. S. Ho (2007): "Phylogenomic analysis of the emergence of GC-rich transcription elements," *Proceedings of the National Academy of Science*, 104, 16528–16533.

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

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

Lothaire, M. (2005): *Applied Combinatorics on Words*, Cambridge University Press.

Melko, O. M. and A. R. Mushegian (2004): "Distribution of words with a predefined range of mismatches to a DNA probe in bacterial genomes." *Bioinformatics*, 20, 67–74.

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

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

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

DOI: 10.2202/1544-6115.1724