JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 11, 2010
© Mary Ann Liebert, Inc.
Pp. 1467–1490
DOI: 10.1089/cmb.2010.0056

Alignment-Free Sequence Comparison (II): Theoretical Power of Comparison Statistics

LIN WAN,[1] GESINE REINERT,[2] FENGZHU SUN,[1,3] and MICHAEL S. WATERMAN[1,3]

ABSTRACT

Rapid methods for alignment-free sequence comparison make large-scale comparisons between sequences increasingly feasible. Here we study the power of the statistic D_2, which counts the number of matching k-tuples between two sequences, as well as D_2^*, which uses centralized counts, and D_2^S, which is a self-standardized version, both from a theoretical viewpoint and numerically, providing an easy to use program. The power is assessed under two alternative hidden Markov models; the first one assumes that the two sequences share a common motif, whereas the second model is a pattern transfer model; the null model is that the two sequences are composed of independent and identically distributed letters and they are independent. Under the first alternative model, the means of the tuple counts in the individual sequences change, whereas under the second alternative model, the marginal means are the same as under the null model. Using the limit distributions of the count statistics under the null and the alternative models, we find that generally, asymptotically D_2^S has the largest power, followed by D_2^*, whereas the power of D_2 can even be zero in some cases. In contrast, even for sequences of length 140,000 bp, in simulations D_2^* generally has the largest power. Under the first alternative model of a shared motif, the power of D_2^* approaches 100% when sufficiently many motifs are shared, and we recommend the use of D_2^* for such practical applications. Under the second alternative model of pattern transfer, the power for all three count statistics does not increase with sequence length when the sequence is sufficiently long, and hence none of the three statistics under consideration can be recommended in such a situation. We illustrate the approach on 323 transcription factor binding motifs with length at most 10 from JASPAR CORE (October 12, 2009 version), verifying that D_2^* is generally more powerful than D_2. The program to calculate the power of D_2, D_2^* and D_2^S can be downloaded from http://meta.cmb.usc.edu/d2. Supplementary Material is available at www.liebertonline.com/cmb.

Key words: alignment-free, hidden Markov model, motifs, normal approximation, power, sequence alignment, word count statistics.

[1] Molecular and Computational Biology, University of Southern California, Los Angeles, California.
[2] Department of Statistics, University of Oxford, Oxford, United Kingdom.
[3] TNLIST/Department of Automation, Tsinghua University, Beijing, P.R. China.

1. INTRODUCTION

Alignment-free sequence comparisons have received extensive attention recently (Burden et al., 2006; Forêt et al., 2006, 2009a,b; Ivan et al., 2008; Kantorovitz et al. 2007a,b). One widely used statistic for alignment free sequence comparison is the D_2 statistic that counts the number of matching k-tuples (also referred as k-words or k-grams) between the two sequences. Throughout this paper, we use *tuples* and *words* interchangeably. It was pointed out in Lippert et al. (2002) that D_2 is not appropriate for the comparison of two sequences because it is dominated by the deviation of the word counts from the corresponding expectations in each sequence. In Reinert et al. (2009), two new variants of the D_2 word count statistics, referred to as D_2^* and D_2^S, were proposed. The statistic D_2^* is based on centered counts, divided by the square root of their means, whereas D_2^S is a self-standardized statistic. More specifically, let $X_{\mathbf{w}}$ and $Y_{\mathbf{w}}$ be the numbers of occurrences of word $\mathbf{w}$ in the first and the second sequences, respectively. The D_2 statistic is defined as

$$D_2 \equiv \sum_{\mathbf{w}\in\mathcal{A}^k} X_{\mathbf{w}} Y_{\mathbf{w}}.$$

To define D_2^* and D_2^S as in [9], we first introduce the centralized count variables by

$$\tilde{X}_{\mathbf{w}} = X_{\mathbf{w}} - np_{\mathbf{w}} \text{ and } \tilde{Y}_{\mathbf{w}} = Y_{\mathbf{w}} - np_{\mathbf{w}},$$

where $p_{\mathbf{w}}$ is the probability of word $\mathbf{w}$ under the null model. Then we put

$$D_2^S = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{\sqrt{\tilde{X}_{\mathbf{w}}^2 + \tilde{Y}_{\mathbf{w}}^2}}, \text{ and } D_2^* = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{np_{\mathbf{w}}}.$$

Here we set $\frac{0}{0} = 0$.

The power of those statistics under two alternative models were explored via simulation approaches. The first alternative model is that the two sequences contain random instances of a common motif, whereas the second alternative model is a pattern transfer model, where randomly chosen DNA segments in the first sequence are used to replace corresponding segments in the second sequence.

It has been shown that, under the first alternative model, the power of both D_2^* and D_2^S is an increasing function of the sequence length for any tuple size $k \geq 2$, while the power of D_2 does not necessarily increase with sequence length and sometimes can even be smaller than the pre-specified type I error. In almost all the simulations considered, the power of D_2^* is higher than that of D_2^S. Under the second alternative model, the power of both D_2^* and D_2^S quickly reaches their plateau and does not seem to change with sequence length. The power of D_2 can decrease with sequence length in some examples.

Simulation studies can only explore very limited ranges of parameter values to compare the power of detecting the relationship between two sequences or genomes. To compare the performance of the different statistics under a broad range of evolutionary scenarios, theoretical studies of the power of these statistics are needed. In addition, it should be very useful to have an easy to use program for calculating the power of sequence comparisons using the various statistics without resorting to time consuming simulations. In this article, we achieve the following objectives: (1) to study the limiting distributions of D_2, D_2^*, and D_2^S under the two alternative models; (2) to compare the theoretical approximate mean, variance, and power of D_2, D_2^*, and D_2^S with the corresponding simulated values (we show that the approximations are reliable for D_2 and D_2^*. However, for the approximations of D_2^S to be reasonable, very long sequences are usually needed); (3) and to develop a program to calculate the power of detecting the relationship between two sequences using D_2, D_2^*, as well as D_2^S. As our calculations are based on approximations, we note that the power in this article is approximate. For easier exposition we omit the word "approximate"; any power is understood to be approximate.

The organization of the article is as follows. In Section 2, we give details of the alternative model I, and show that the distributions of D_2/n^2, D_2^*/n and D_2^S/n converge to normal distributions as the sequence length tends to infinity. Formulas for the approximate mean and variance of D_2/n^2, D_2^*/n and D_2^S/n are presented, and they are put to use to calculate the power of D_2, D_2^* and D_2^S. In Section 3, we give details of alternative model II and develop a new hidden Markov model (HMM) for generating pairs of sequences related through alternative model II. The approximate distributions of D_2, D_2^*, and D_2^S under alternative model II are then derived. These approximate distributions are not normal and are complicated. We show that the power of D_2, D_2^*, and D_2^S converges rapidly and does not change much as sequence length n increases, a phenomenon observed in the simulation studies of Reinert et al. (2009). Under the second model, we do not have an efficient method for calculating the mean and variance of D_2^S, but we are able to present methods for calculating the approximate mean of D_2 and D_2^*. In Section 4, we first describe a web-based and a R program package for calculating the power of D_2, D_2^* and D_2^S to detect the relationships between two sequences under alternative model I. We then evaluate the program by comparing the theoretical mean, variance, and power derived in this study with the corresponding simulated quantities presented in Reinert et al. (2009) and show that the approximate mean and variance are generally close to their corresponding true values when the sequence length is very large. We find that convergence for D_2^S is considerably slower than for D_2^* and for D_2. This also affects the power of the statistic—the power approximation for D_2^S is poor in the parameter regimes we considered. Hence, we concentrate on D_2 and D_2^* for the remainder of the article. Moreover, D_2 has zero power under some models, and hence cannot be used to infer the relationship between sequences under such models. For D_2 and D_2^*, the program developed in this study can be readily used to study the power of comparing sequences using k-tuples. We then extend our study to 323 transcription factor (TF) binding motifs and show the superiority of D_2^* compared to D_2 for sequence comparison for general motif patterns although there are a few exceptions where D_2 is more powerful than D_2^*. For alternative model II, we study how the means of D_2 and D_2^* change with the word length k in order to explain the observation that the power of D_2^* using $k=10$ is much higher than the power using $k=5$ in the simulation studies reported in Reinert et al. (2009). The article concludes with some discussion and potential extensions to more general background sequence models.

The results regarding the approximate distributions of D_2, D_2^*, and D_2^S and the power of detecting the relationships between the sequences using these statistics can be easily extended to sequence pairs with different background letter frequencies, sequence lengths, and motif densities. However, the notation and presentation will be more complicated. For notational simplicity and clarity of presentation, we present the results for two sequences having the same background probability distribution, sequence length, and motif density. The results for the general situations are given in the Appendix. As the proofs are very similar to the ones presented in the article, they are omitted.

2. ALTERNATIVE MODEL I

2.1. The model and the count statistics

The alternative model I renders the two sequences dependent through a common motif which is randomly distributed across the two sequences. As in Reinert et al. (2009), we model the background sequence as independent identically distributed (IID) random variables taking different letters from finite alphabet $\mathcal{A}$ with probability $p_a (a \in \mathcal{A})$. For notational convenience, we also denote $p_a^{(0)} = p_a$. For nucleotide sequences, $\mathcal{A} = \{A, C, G, T\}$ and for amino acid sequences, the $\mathcal{A}$ is the set of 20 amino acids. In general, we assume that $\mathcal{A}$ contains L letters and write $\mathcal{A} = \{0, 1, 2, \cdots, L-1\}$. For the motif instances, we use the model in Zhai et al. (2010), which is more general than the model used in Reinert et al. (2009), where fixed motifs were used. In this article and in Zhai et al. (2010), a position weight matrix (PWM) is used to describe the distribution of the nucleotides at the different positions of a motif (Stormo, 2000). For a given motif of length M, and at the m-th position of the motif, the probability that the base takes value a from $\mathcal{A}$ is $p_a^{(m)}, m = 1, 2, \cdots, M$. The motif instances are randomly distributed across the sequence with density $1-\lambda$ $(0<\lambda<1)$. That is, at each position in the sequence which is not already covered by an instance of a motif, with probability λ, a base with the background distribution is generated, and with probability $1-\lambda$, an instance of the motif of length M is generated based on the PWM for the motif. Once an instance of a motif is generated, we move to the end of the instance of the motif to repeat this process.

For the model in more detail, see Zhai et al. (2010). The sequences with random motif instances were modeled by an HMM (Rabiner, 1989). The underlying Markov chain (MC) of each sequence is denoted as $Q_1 Q_2 \cdots Q_i \cdots Q_{n+k-1}$ (i is the position index of the sequence with length $n+k-1$) which take values in $\{0, 1, 2, \cdots, M\}$. The 0 indicates that the state of the sequence is the background sequence while m $(1 \le m \le M)$ indicates the state at the m-th position of the motif. Under each state, the emission probability of each letter from $\mathcal{A}$ is denoted as $p_a^{(m)}$ ($a \in \mathcal{A}$ and $m = 0, 1, 2, \cdots, M$). The transition matrix for the underlying MC $Q_1 Q_2 \cdots Q_i \cdots Q_{n+k-1}$ is given by $T = (t_{mm'})_{(M+1)\times(M+1)}$, where $t_{00} = t_{M0} = \lambda$, $t_{01} = t_{M1} = 1-\lambda$, $t_{m,m+1} = 1$, $m = 1, 2, \cdots, M-1$, and all the other t's are 0. The MC has as stationary distribution $\pi = \frac{1}{\lambda + M(1-\lambda)}(\lambda, 1-\lambda, 1-\lambda, \cdots, 1-\lambda)$ (Zhai et al., 2010). Therefore, in stationarity, the expected fraction of the sequence that is covered by the motif instances is $M(1-\lambda)/(\lambda + M(1-\lambda))$. Unless λ is close to 1, the expected fraction of the sequence covered by inserted motif instances can be unrealistically large (Table S1; for Supplementary Material, see www.liebertonline.com/cmb). Hence we only study values of λ which are no smaller than 0.9.

Now we consider two sequences of length $n+k-1$ generated by the above HMM, $\mathbf{A} = A_1A_2\cdots A_{n+k-1}$ and $\mathbf{B} = B_1B_2\cdots B_{n+k-1}$. We let the sequence length be $n+k-1$ for notational simplicity in the remainder of the paper. Given a k-tuple $\mathbf{w} = (w_1, w_2, \ldots, w_k) \in \mathcal{A}^k$, let $X_\mathbf{w}$ and $Y_\mathbf{w}$ be the numbers of occurrences of $\mathbf{w}$ within $\mathbf{A}$ and $\mathbf{B}$, respectively; within each sequence, the occurrences could overlap. Assume that the Markov process starts in the stationary distribution. Based on Proposition 2.2 in Zhai et al. (2010), the means of $X_\mathbf{w}(n)$ and $Y_\mathbf{w}(n)$ can be calculated as

$$\mathbb{E}_\lambda X_\mathbf{w} = \mathbb{E}_\lambda Y_\mathbf{w} = nP_\lambda(\mathbf{w}),$$

where $P_\lambda(\mathbf{w}) = \sum_{m=0}^{M} \alpha_k^{(\mathbf{w})}(m)$ is the probabiltiy of the word $\mathbf{w}$ under the alternative model I. The $\alpha_i^{(\mathbf{w})}(m) = P(A_j = w_j, j = 1, 2, \cdots, i; Q_i = m), i = 1, 2, \cdots, k,$ are calculated recursively using the standard forward procedure for calculating the probability of an observation sequence based on HMM (Zhai et al., 2010; Rabiner, 1989) for $i = 1, 2\cdots$:

$$\begin{aligned}
\alpha_{i+1}^{(\mathbf{w})}(0) &= (\alpha_i^{(\mathbf{w})}(0) + \alpha_i^{(\mathbf{w})}(M))\lambda p_{w_{i+1}}^{(0)},\\
\alpha_{i+1}^{(\mathbf{w})}(1) &= (\alpha_i^{(\mathbf{w})}(0) + \alpha_i^{(\mathbf{w})}(M))(1-\lambda) p_{w_{i+1}}^{(1)},\\
\alpha_{i+1}^{(\mathbf{w})}(m) &= \alpha_i^{(\mathbf{w})}(m-1) p_{w_{i+1}}^{(m)}, \qquad (m = 2, 3, \ldots, M),
\end{aligned}$$

and

$$\begin{aligned}
\alpha_1^{(\mathbf{w})}(0) &= \frac{\lambda p_{w_1}^{(0)}}{\lambda + M(1-\lambda)},\\
\alpha_1^{(\mathbf{w})}(m) &= \frac{(1-\lambda) p_{w_1}^{(m)}}{\lambda + M(1-\lambda)}, \qquad (1 \le m \le M).
\end{aligned}$$

In particular, $P_1(\mathbf{w}) = p_\mathbf{w} = p_{w_1}p_{w_2}\cdots p_{w_k}$.

2.2. *The expectations of* D_2, D_2^* *and* D_2^S *under alternative model I*

It is easy to see that $E_\lambda(\widetilde{X}_\mathbf{w}) = n(P_\lambda(\mathbf{w}) - p_\mathbf{w})$, where $P_\lambda(\mathbf{w})$ is the probability of word $\mathbf{w}$ under the alternative model I. However, for the mean of $\frac{\tilde{X}_\mathbf{w}\tilde{Y}_\mathbf{w}}{\sqrt{\tilde{X}_\mathbf{w}^2 + \tilde{Y}_\mathbf{w}^2}}$, it is in general only known that it is non-negative, and when $\tilde{X}_\mathbf{w}$ and $\tilde{Y}_\mathbf{w}$ are IID, the mean is zero if and only if the distribution of $\tilde{X}_\mathbf{w}$ is symmetric (Novak, 2007). Note that the two sequences $\mathbf{A}$ and $\mathbf{B}$ are independent under the alternative model I. Then, we have the following theorem.

Theorem 2.1. *Assume alternative model I for the two sequences* $\mathbf{A}$ *and* $\mathbf{B}$, *and let* $P_\lambda(\mathbf{w}) = P(A_1A_2\cdots A_k = w_1w_2\cdots w_k)$ *be as calculated in Subsection 2.1. Then for the expectations of* D_2, D_2^* *and* D_2^S, *we have*

$$\begin{aligned}
\mathbb{E}(D_2) &= n^2 \sum_{\mathbf{w}\in\mathcal{A}^k} (P_\lambda(\mathbf{w}))^2,\\
\mathbb{E}(D_2^*) &= n \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{(P_\lambda(\mathbf{w}) - p_\mathbf{w})^2}{p_\mathbf{w}},\\
\textit{and } \lim_{n\to\infty} \frac{\mathbb{E}(D_2^S)}{n} &= \frac{1}{\sqrt{2}} \sum_{\mathbf{w}\in\mathcal{A}^k} |P_\lambda(\mathbf{w}) - p_\mathbf{w}|.
\end{aligned}$$

Further,

$$\lim_{n\to\infty} n\left(\frac{\mathbb{E}(D_2^S)}{n} - \frac{1}{\sqrt{2}}\sum_{\mathbf{w}\in\mathcal{A}^k}|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|\right) = -\frac{3\sqrt{2}}{8}\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\sigma_\lambda^2(\mathbf{w})}{|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|},$$

where $\sigma_\lambda^2(\mathbf{w}) = \lim_{n\to\infty}\frac{Var(X_{\mathbf{w}})}{n}$*; see also* (1) *below.*

The first two equations can be easily proven by the independence of the two sequences. The last two limit expressions can be proven by Taylor expansion (the delta method); see the proof of Theorem 2.4 for details.

2.3. *The approximate distributions of* D_2, D_2^*, *and* $\mathrm{D}_2^{\mathrm{S}}$ *under alternative model I*

The variances of D_2 and its variants are complicated. Under the null model of IID sequences, upper and lower bounds for the variance of D_2 were first explored in Lippert et al. (2002). In Kantorovitz et al. (2007b), an explicit formula for the variance of D_2 is given in the IID case. To study the power of D_2, D_2^*, and D_2^S in detecting the relationship between two sequences, we explore the approximate distributions of these statistics as the sequence length goes to infinity. Note that the distributions of D_2 and D_2^* under the null model when $(p_A^{(0)}, p_C^{(0)}, p_G^{(0)}, p_T^{(0)}) = (1/4, 1/4, 1/4, 1/4)$ have been carefully studied in Reinert et al. (2009). Therefore, in the rest of the article, we assume that $(p_A^{(0)}, p_C^{(0)}, p_G^{(0)}, p_T^{(0)}) \neq (1/4, 1/4, 1/4, 1/4)$.

For $0 < \lambda < 1$, the values of

$$\sigma_\lambda^2(\mathbf{w}) = \lim_{n\to\infty}\frac{\mathrm{Var}(X_{\mathbf{w}})}{n} \text{ and } \sigma_\lambda(\mathbf{w},\mathbf{w}') = \lim_{n\to\infty}\frac{Cov(X_{\mathbf{w}}, X_{\mathbf{w}'})}{n} \tag{1}$$

can be calculated using the method in Zhai et al. (2010), Proposition 2.3; for $\lambda = 1$, the corresponding values can be found, for example, in Reinert et al. (2009), Corollary 6.1. We denote the asymptotic variance of $\sum_{\mathbf{w}\in\mathcal{A}^k} P_\lambda(\mathbf{w})X_{\mathbf{w}}/\sqrt{n}$ in one sequence by

$$(\Sigma_\lambda)^2 = \sum_{\mathbf{w}\in\mathcal{A}^k} P_\lambda^2(\mathbf{w})\sigma_\lambda^2(\mathbf{w}) + \sum_{\mathbf{w}\neq\mathbf{w}'} P_\lambda(\mathbf{w})P_\lambda(\mathbf{w}')\sigma_\lambda(\mathbf{w},\mathbf{w}'). \tag{2}$$

The following theorem gives the approximate distributions of D_2 under the null and the alternative model I.

Theorem 2.2. *Assume that in the background model not all letters are equally likely.*
a. [Lippert et al. (2002), Theorem 4.2.] Suppose $\lambda = 1$ *(the null model that the sequences are IID). Then*

$$\lim_{n\to\infty}\sqrt{n}\left(\frac{D_2}{n^2} - \sum_{\mathbf{w}\in\mathcal{A}^k} p_{\mathbf{w}}^2\right) = Z_1,$$

where Z_1 *has normal distribution* $\mathcal{N}(0, 2(\Sigma_1)^2)$*. Here the asymptotic is valid when the sequence length tends to infinity with alphabet size, motif length, and word length kept fixed.*
b. Suppose $0 < \lambda < 1$ *(the alternative model I). Then*

$$\lim_{n\to\infty}\sqrt{n}\Big(\frac{D_2}{n^2} - \sum_{\mathbf{w}\in\mathcal{A}^k}(P_\lambda(\mathbf{w}))^2\Big) = Z_\lambda,$$

where Z_λ *has normal distribution* $\mathcal{N}(0, 2(\Sigma_\lambda)^2)$*. Here the asymptotic is valid when the sequence length tends to infinity with alphabet size, motif length, and word length kept fixed.*

On the other hand, under the null model that no motif instances are inserted, D_2^* is approximately the sum of products of dependent mean 0 normal random variables (and thus not normal). However, it is approximately normally distributed when the sequence length is large under the alternative model I, as long as $\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})}{p_{\mathbf{w}}}$ is not constant in $\mathbf{w}$, as the following theorem shows. We put

$$(\Sigma_\lambda^*)^2 = \sum_{\mathbf{w}\in\mathcal{A}^k}\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})^2}{p_{\mathbf{w}}^2}\sigma_\lambda^2(\mathbf{w}) + \sum_{\mathbf{w}\neq\mathbf{w}'}\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})(P_\lambda(\mathbf{w}')-p_{\mathbf{w}'})}{p_{\mathbf{w}}p_{\mathbf{w}'}}\sigma_\lambda(\mathbf{w},\mathbf{w}'), \tag{3}$$

with $\sigma_\lambda^2(\mathbf{w})$ and $\sigma_\lambda(\mathbf{w}, \mathbf{w}')$ given in (1).

Theorem 2.3. *a. Suppose $\lambda = 1$ (the null model that the sequences are IID). Then, in distribution,*

$$\lim_{n\to\infty} D_2^* = Z_1^* = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{Z_{\mathbf{w}}^{(1)} Z_{\mathbf{w}}^{(2)}}{p_{\mathbf{w}}},$$

where $\{Z_{\mathbf{w}}^{(1)}, \mathbf{w} \in \mathcal{A}^k\}$ and $\{Z_{\mathbf{w}}^{(2)}, \mathbf{w} \in \mathcal{A}^k\}$ are independent and have the same mean 0 normal distribution (with non-trivial covariance matrix).

b. Suppose $0 < \lambda < 1$ (the alternative model I), and that $\frac{(P_\lambda(\mathbf{w}) - p_{\mathbf{w}})}{p_{\mathbf{w}}}$ is not constant in $\mathbf{w}$. Then, in distribution,

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{D_2^*}{n} - \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{(P_\lambda(\mathbf{w}) - p_{\mathbf{w}})^2}{p_{\mathbf{w}}}\right) = Z_\lambda^*,$$

where Z_λ^* has normal distribution $\mathcal{N}(0, 2(\Sigma_\lambda^*)^2)$.

We let

$$(\Sigma_\lambda^S)^2 = \frac{1}{8}\left\{\sum_{\mathbf{w}\in\mathcal{A}^k} \sigma_\lambda^2(\mathbf{w}) + \sum_{\mathbf{w}\neq\mathbf{w}'} \mathrm{sign}(P_\lambda(\mathbf{w}) - p_{\mathbf{w}})\mathrm{sign}(P_\lambda(\mathbf{w}') - p_{\mathbf{w}'})\sigma_\lambda(\mathbf{w}, \mathbf{w}')\right\}, \quad (4)$$

where $\mathrm{sign}(x) = 1$ if $x > 0$, $\mathrm{sign}(x) = -1$ if $x < 0$, and $\mathrm{sign}(0) = 0$; again $\sigma_\lambda^2(\mathbf{w})$ and $\sigma_\lambda(\mathbf{w}, \mathbf{w}')$ are given in (1). The following theorem gives the approximate distribution of D_2^S under the null and the alternative models.

Theorem 2.4. *a. Suppose $\lambda = 1$ (the null model that the sequences are IID). Then, in distribution,*

$$\lim_{n\to\infty} \frac{D_2^S}{\sqrt{n}} = Z_1^S = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{Z_{\mathbf{w}}^{(1)} Z_{\mathbf{w}}^{(2)}}{\sqrt{(Z_{\mathbf{w}}^{(1)})^2 + (Z_{\mathbf{w}}^{(2)})^2}} \quad (5)$$

where $\{Z_{\mathbf{w}}^{(1)}, \mathbf{w} \in \mathcal{A}^k\}$ and $\{Z_{\mathbf{w}}^{(2)}, \mathbf{w} \in \mathcal{A}^k\}$ are independent and have the same mean 0 normal distribution.

b. Suppose $0 < \lambda < 1$ (the alternative model I), and assume that $P_\lambda(\mathbf{w}) - p(\mathbf{w})$ have different sign in $\mathbf{w}$. Then, in distribution,

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{D_2^S}{n} - \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}}\right) = Z_\lambda^S,$$

where Z_λ^S has normal distribution $\mathcal{N}(0, 2(\Sigma_\lambda^S)^2)$.

c. Suppose $0 < \lambda < 1$ (the alternative model I), and assume that $P_\lambda(\mathbf{w}) - p_{\mathbf{w}}$ have different sign in $\mathbf{w}$. Then, in distribution,

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{D_2^S}{n} - \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}} + \frac{3\sqrt{2}}{8n}\sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\sigma_\lambda^2(\mathbf{w})}{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}\right) = Z_\lambda^S,$$

where Z_λ^S has normal distribution $\mathcal{N}(0, 2(\Sigma_\lambda^S)^2)$.

Remark 2.1. *Since each term on the right hand side of* (5) *has a normal distribution under the null model by Reinert et al. (2009), and the terms are jointly normal, the limit of $\frac{D_2^S}{\sqrt{n}}$ is mean zero normally distributed. The variance can be estimated from the empirical distribution, as illustrated in Reinert et al. (2009).*

Replacing $\sum_{\mathbf{w}\in\mathcal{A}^k} \frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}}$ by $\sum_{\mathbf{w}\in\mathcal{A}^k} \frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}} - \frac{3\sqrt{2}}{8n}\sum_{\mathbf{w}\in\mathcal{A}^k} \frac{\sigma_\lambda^2(\mathbf{w})}{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}$ can be significant when we study the power of detecting the relationships between two sequences using D_2^S, as we shall see in Section 4.2.

The proofs of these theorems are presented in the Appendix.

2.4. The power of detecting the relationship between two sequences under alternative model I using D_2, D_2^*, and D_2^S

Knowing the asymptotic distributions of D_2, D_2^*, and D_2^S under the null and the alternative models, we are able to approximate the power of detecting the relationships between two sequences using any of the three statistics. For notational simplicity, let

$$A(\lambda)=\sum_{\mathbf{w}\in\mathcal{A}^k}P_\lambda^2(\mathbf{w}),\quad A^*(\lambda)=\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})^2}{p_{\mathbf{w}}},$$

$$A^S(\lambda)=\frac{1}{\sqrt{2}}\sum_{\mathbf{w}\in\mathcal{A}^k}|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|$$

denote the (asymptotic) means of D_2, D_2^*, and D_2^S under alternative model I. Let $\Phi(\cdot)$ be the cumulative distribution for the standard normal distribution. From Theorems 2.2, 2.3, and 2.4, we can show the following theorem to hold.

Theorem 2.5. *Assume that* $\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})^2}{p_{\mathbf{w}}}$ *and* $P_\lambda(\mathbf{w})-p_{\mathbf{w}}$ *are not constant in* $\mathbf{w}$. *Then, for any given type I error* α, *the power of detecting the relationship between two sequences against the null model that* $\lambda=1$ *using* D_2, D_2^* *and* D_2^S *can be approximated by* $1-\Phi(C(\lambda))$, $1-\Phi(C^*(\lambda))$, *and* $1-\Phi(C^S(\lambda))$, *respectively, where*

$$\begin{aligned}C(\lambda)&=-\sqrt{n}B(\lambda)+z_\alpha/(\sqrt{2}\Sigma_\lambda),\\ C^*(\lambda)&=-\sqrt{n}B^*(\lambda)+z_\alpha^*/(\sqrt{2n}\Sigma_\lambda^*),\\ \textit{and}\ C^S(\lambda)&=-\sqrt{n}B^S(\lambda)+z_\alpha^S/(\sqrt{2}\Sigma_\lambda^S)\end{aligned}$$

and

$$B(\lambda)=\frac{A(\lambda)-A(1)}{\sqrt{2}\Sigma_\lambda},\quad B^*(\lambda)=\frac{A^*(\lambda)}{\sqrt{2}\Sigma_\lambda^*},\ \textit{and}\ B^S(\lambda)=\frac{A^S(\lambda)}{\sqrt{2}\Sigma_\lambda^S}.$$

Here, z_α, z_α^*, *and* z_α^S *are the upper* α *quantile of* Z_1, Z_1^*, Z_1^S *from Theorems 2.2, 2.3, and 2.4, respectively.*

Note that we can again replace $A^S(\lambda)$ by $A_m^S(\lambda)=A^S(\lambda)-\frac{3\sqrt{2}}{8n}\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\sigma_\lambda^2(\mathbf{w})}{|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|}$ when we calculate the power of D_2^S for relative small values of sequence length n. Here the subscript m stands for *modified*.

Theorem 2.5 indicates that when sequence length n is large, the dominant terms in $C(\lambda)$, $C^*(\lambda)$, and $C^S(\lambda)$ are the first term and the second term becomes negligible when n is large. Therefore, the higher the values of the B's, the more powerful the corresponding statistic is when n is sufficiently large. In Section 4, we present some examples for values of the B's and the C's.

The tests under alternative model I make extensive use of the fact that the means of our statistics are different under the alternative model versus the null model. Under alternative model II, this will turn out not to be the case.

3. ALTERNATIVE MODEL II

In this section, we consider the second alternative model which is inspired by horizontal gene transfer. We randomly choose a certain number of segments in the first sequence and then replace the corresponding segments (position-wise) in the second sequence by the letters in the first sequence.

3.1. A second HMM model for the sequence pair **A** *and* **B**

Alternative model II is again a HMM model for the sequence pair $\mathbf{A}=A_1A_2\cdots A_{n+k-1}$ and $\mathbf{B}=B_1B_2\cdots B_{n+k-1}$. First, two IID sequences $\mathbf{A}$ and $\mathbf{B}'$ are generated. From these two sequences we construct $\mathbf{B}$ as follows. We assume that at each position which is not already covered by a chosen segment, with probability λ, the original bases of the two sequences at the position are kept. With probability $1-\lambda$, a segment of length M from the first sequence is chosen, and the same segment in the second sequence is replaced by it. Then we move to the end of the segment to start this process again. Consider an underlying Markov chain $Q_1Q_2\cdots Q_i\cdots$ defined as follows. Each Q_i takes values in $\{0, 1, 2, \cdots, M\}$, where $Q_i = 0$ indicates that, at position i, A_i and B_i are the originally generated bases, whereas $Q_\mathrm{i} = m\ (1, 2, \cdots, M)$ indicates that position i is at the m-th position of a segment which was copied from the first sequence to the second sequence. The transition matrix of $Q_1Q_2\cdots Q_i\cdots$ is given by $T = (t_{mm'})_{(M+1)\times(M+1)}$, where $t_{00} = t_{M0} = \lambda$, $t_{01} = t_{M1} = 1-\lambda$, $t_{m,m+1} = 1$, $m = 1, 2, \cdots, M-1$, and all the other t's are 0. It is easy to see that the stationary distribution of this Markov chain is $\pi = \frac{1}{\lambda + M(1-\lambda)}(\lambda, 1-\lambda, 1-\lambda, \cdots, 1-\lambda)$ (see Proposition 2.1 in Zhai et al. [2010]).

Let $C_i = (A_i, B_i)^t$. With p_a denoting the probability of letter a in the IID model, the emission probabilities are given by

$$P((A_i, B_i) = (a, b)|Q_i = 0) = p_a p_b, \quad P((A_i, B_i) = (a, b)|Q_i = m) = p_a I(a = b).$$

Then $C_1C_2\cdots C_i\cdots$ form a HMM.

3.2. The asymptotic distributions and power of D_2, D_2^*, and D_2^S *for detecting relationships between sequences under alternative model II*

Under alternative model II, the marginal distributions of the individual sequences are IID sequences and hence the means of X_w and Y_w are unchanged compared to the IID model. However, the two sequences depend on each other because they share some common segments. The following theorem shows an efficient way to calculate the covariance of the number of occurrences of word $\mathbf{w}$ in sequence $\mathbf{A}$ and the number of occurrences of word $\mathbf{w}'$ in sequence $\mathbf{B}$. These covariances are used to derive the limiting distributions of D_2, D_2^*, and D_2^S when the sequence length tends to infinity.

Theorem 3.1. *Let $X_\mathbf{w}$ and $Y_\mathbf{w}$ be the number of occurrences of word $\mathbf{w}$ in sequence $\mathbf{A}$ and $\mathbf{B}$, respectively. Assume that the MC starts in the stationary distribution. For any pair of words $(\mathbf{w}, \mathbf{w}')$, we have under alternative model II,*

a. The expectation of $X_\mathbf{w}Y_{\mathbf{w}'}$ is

$$\mathbb{E}(X_\mathbf{w}Y_{\mathbf{w}'}) = n\gamma_0(\mathbf{w}, \mathbf{w}') + \sum_{j=1}^{k-1}(n-j)(\gamma_j(\mathbf{w}, \mathbf{w}') + \gamma_j(\mathbf{w}', \mathbf{w}))$$
$$+ [n^2 - (2k-1)n + k(k-1)]p_\mathbf{w}p_{\mathbf{w}'}.$$

b. The covariance of $X_\mathbf{w}$ and $Y_{\mathbf{w}'}$ changes linearly with sequence length n, and

$$\begin{aligned}\delta_\lambda(\mathbf{w}, \mathbf{w}') &= \lim_{n\to\infty}\frac{Cov(X_\mathbf{w}, Y_{\mathbf{w}'})}{n}\\ &= \gamma_0(\mathbf{w}, \mathbf{w}') + \sum_{j=1}^{k-1}(\gamma_j(\mathbf{w}, \mathbf{w}') + \gamma_j(\mathbf{w}', \mathbf{w})) - (2k-1)p_\mathbf{w}p_{\mathbf{w}'}.\end{aligned} \tag{6}$$

c. The difference $ED_2 - n^2\sum_{\mathbf{w}\in\mathcal{A}^k}p_\mathbf{w}^2$ changes linearly with respect to sequence length n, and

$$\lim_{n\to\infty}\frac{ED_2 - n^2\sum_{\mathbf{w}\in\mathcal{A}^k}p_\mathbf{w}^2}{n} = \sum_{\mathbf{w}\in\mathcal{A}^k}\left[\gamma_0(\mathbf{w}, \mathbf{w}) + 2\sum_{j=1}^{k-1}\gamma_j(\mathbf{w}, \mathbf{w}) - (2k-1)p_\mathbf{w}^2\right]. \tag{7}$$

d. The expectation of D_2^* converges as the sequence length n tends to infinity, and

$$\lim_{n\to\infty}ED_2^* = \sum_{\mathbf{w}\in\mathcal{A}^k}\left[\frac{\gamma_0(\mathbf{w}, \mathbf{w}) + 2\sum_{j=1}^{k-1}\gamma_j(\mathbf{w}, \mathbf{w})}{p_\mathbf{w}}\right] - (2k-1). \tag{8}$$

In all the above equations, $\gamma_0(\mathbf{w}, \mathbf{w}') = P(A_l = w_l, B_l = w'_l, l = 1, 2, \cdots, k)$ can be calculated as $\gamma_0(\mathbf{w}, \mathbf{w}') = \sum_{m=0}^{M}\theta_k^{(\mathbf{w}, \mathbf{w}')}(m)$, and $\theta_i^{(\mathbf{w}, \mathbf{w}')}(m) = P(A_l = w_l, B_l = w'_l,\ l = 1, 2, \cdots, i;\ Q_i = m)$ can be calculated recursively using the following equations for $i = 1, 2, \cdots$

$$\theta_{i+1}^{(\mathbf{w},\mathbf{w}')}(0)=(\theta_i^{(\mathbf{w},\mathbf{w}')}(0)+\theta_i^{(\mathbf{w},\mathbf{w}')}(M))\lambda p_{w_{i+1}}p_{w'_{i+1}},$$

$$\theta_{i+1}^{(\mathbf{w},\mathbf{w}')}(1)=(\theta_i^{(\mathbf{w},\mathbf{w}')}(0)+\theta_i^{(\mathbf{w},\mathbf{w}')}(M))(1-\lambda)p_{w_{i+1}}I(w_{i+1}=w'_{i+1}),$$

$$\theta_{i+1}^{(\mathbf{w},\mathbf{w}')}(m)=\theta_i^{(\mathbf{w},\mathbf{w}')}(m-1)p_{w_{i+1}}I(w_{i+1}=w'_{i+1}),\qquad (m=2,3,\ \ldots,M)$$

with initial values

$$\theta_1^{(\mathbf{w},\mathbf{w}')}(0)=\frac{\lambda p_{w_1}p_{w'_1}}{\lambda+M(1-\lambda)},$$

$$\theta_1^{(\mathbf{w},\mathbf{w}')}(m)=\frac{(1-\lambda)p_{w_1}I(w_{i+1}=w'_{i+1})}{\lambda+M(1-\lambda)},\qquad (1\le m\le M).$$

Moreover $\gamma_j(\mathbf{w},\mathbf{w}')=P(A_l=w_l,B_{l+j}=w_l,l=1,\ \ldots,k)$ *can be calculated as*

$$\gamma_j(\mathbf{w},\mathbf{w}')=\gamma_0(w_{j+1}w_{j+2}\cdots w_k,w'_1w'_2\cdots w'_{k-j})\prod_{s=1}^{j}p_{w_s}\prod_{s=k-j+1}^{k}p_{w'_s},\quad j=1,2,\ \cdots k-1.$$

Similarly to the proofs of Theorems 2.2, 2.3, 2.4, we can prove the following theorem regarding the limiting distributions of D_2, D_2^*, and D_2^S. Let $\sigma_1(\mathbf{w},\mathbf{w}')=\lim_{n\to\infty}\frac{Cov(X_{\mathbf{w}},X_{\mathbf{w}'})}{n}=\lim_{n\to\infty}\frac{Cov(Y_{\mathbf{w}},Y_{\mathbf{w}'})}{n}$ and $\sigma_1^2(\mathbf{w})=\sigma_1(\mathbf{w},\mathbf{w})$, which can be calculated as in Zhai et al. (2010), and recall $\delta_\lambda(\mathbf{w},\ \mathbf{w}')$ from (6).

Theorem 3.2. *Suppose* $0<\lambda\le 1$ *and the alternative model II.*
a. Then, in distribution,

$$\lim_{n\to\infty}\sqrt{n}\left(\frac{D_2}{n^2}-\sum_{\mathbf{w}\in\mathcal{A}^k}p_{\mathbf{w}}^2\right)=\widetilde{Z}_\lambda,$$

where $\widetilde{Z}_\lambda$ *has normal distribution* $\mathcal{N}(0,2(\Lambda_\lambda)^2)$*, and*

$$(\Lambda_\lambda)^2=\sum_{\mathbf{w}\in\mathcal{A}^k}p_{\mathbf{w}}^2\sigma_1^2(\mathbf{w})+\sum_{\mathbf{w}\neq\mathbf{w}'}p_{\mathbf{w}}p_{\mathbf{w}'}\sigma_1(\mathbf{w},\mathbf{w}')+\sum_{\mathbf{w},\,\mathbf{w}'}p_{\mathbf{w}}p_{\mathbf{w}'}\delta_\lambda(\mathbf{w},\mathbf{w}').$$

b. In distribution,

$$\lim_{n\to\infty}D_2^*=\widetilde{Z}_\lambda^*=\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\widetilde{Z}_{\mathbf{w}}^{(1)}\widetilde{Z}_{\mathbf{w}}^{(2)}}{p_{\mathbf{w}}},$$

where $\{\widetilde{Z}_{\mathbf{w}}^{(1)},\mathbf{w}\in\mathcal{A}^k\}$ *and* $\{\widetilde{Z}_{\mathbf{w}}^{(2)},\mathbf{w}\in\mathcal{A}^k\}$ *have the same marginal normal distribution* $N(0,\ (\sigma_1(\mathbf{w},\ \mathbf{w}'))_{\mathbf{w},\mathbf{w}'})$ *and the covariance between* $\widetilde{Z}_{\mathbf{w}}^{(1)}$ *and* $\widetilde{Z}_{\mathbf{w}'}^{(2)}$ *is* $\delta_\lambda(\mathbf{w},\ \mathbf{w}')$.
c. In distribution,

$$\lim_{n\to\infty}\frac{D_2^S}{\sqrt{n}}=\widetilde{Z}_\lambda^S=\sum_{\mathbf{w}\in\mathcal{A}^k}\frac{\widetilde{Z}_{\mathbf{w}}^{(1)}\widetilde{Z}_{\mathbf{w}}^{(2)}}{\sqrt{(\widetilde{Z}_{\mathbf{w}}^{(1)})^2+(\widetilde{Z}_{\mathbf{w}}^{(2)})^2}}$$

where $\{\widetilde{Z}_{\mathbf{w}}^{(1)},\mathbf{w}\in\mathcal{A}^k\}$ *and* $\{\widetilde{Z}_{\mathbf{w}}^{(2)},\mathbf{w}\in\mathcal{A}^k\}$ *are the same as in part (b).*

Based on the above theorem, we can obtain the approximate power of D_2, D_2^*, and D_2^S for detecting the relationships between two sequences under the alternative model II.

Theorem 3.3. *Suppose* $0<\lambda<1$ *and the alternative model II. For a given type I error* α*, let* $\widetilde{z}_\alpha$, $\widetilde{z}_\alpha^*$*, and* $\widetilde{z}_\alpha^S$ *be the upper* α *quantile for* $\widetilde{Z}_1$, $\widetilde{Z}_1^*$*, and* $\widetilde{Z}_1^S$*, respectively. Then the corresponding power of* $\widetilde{Z}_\lambda$, $\widetilde{Z}_\lambda^*$*, and* $\widetilde{Z}_\lambda^S$ *under the alternative model II when* $\lambda<1$ *is asymptotically* $P\{\widetilde{Z}_\lambda\ge\widetilde{z}_\alpha\}$, $P\{\widetilde{Z}_\lambda^*\ge\widetilde{z}_\alpha^*\}$*, and* $P\{\widetilde{Z}_\lambda^S\ge\widetilde{z}_\alpha^S\}$*, respectively.*

Since $\widetilde{Z}_1$ is normally distributed with mean 0, the threshold value $\widetilde{z}_\alpha>0$ if $\alpha<0.5$. From this theorem, it is clear that the power of the three statistics for detecting the relationships between the two sequences does not increase with sequence length n when n is sufficiently large, which is consistent with the simulation results in Reinert et al. (2009). The theoretical results presented here explain that none of the three statistic is what would be most desirable for detecting the relationships between sequences under the alternative model II. One unsolved problem is what statistics we should use under alternative model II.

4. RESULTS

In this section, we describe an online implementation and a stand-alone R program for calculating the power of detecting the relationships between two sequences under the alternative model I using any of the statistics studied in this article. Then we compare the mean, variance, and power of the statistics D_2, D_2^* and D_2^S derived using our formula with the simulated quantities for the situations in Reinert et al. (2009). As an illustration of the difficulties involved, we present the results for the relatively simple two letter sequences under alternative model I in the supplementary material. In particular, this simple case shows that in some cases D_2 will have zero power for detecting the relationship between two sequences when they share a common motif. In some scenarios, however, we see that D_2 can be more powerful than D_2^* and D_2^S. It also shows that the convergence of the mean and variance of D_2^S to their theoretical limit is very slow, which affects the approximate power calculation; in the parameter region which we considered, the theoretical approximate power of D_2^S differs so considerably from the power under simulation that we do not recommend using D_2^S for moderate sequence lengths. Finally, the power of detecting the relationships between two sequences when any of the 323 motifs with motif length at most 10 in JASPAR (Sandelin et al., 2004) (October 12, 2009 version) are present in the two sequences are given. For alternative model II, we give an explanation why the power of D_2^* using $k=10$ is much higher than using $k=2$, 3, 4, 5 for the parameters in simulation studies (Reinert et al., 2009).

4.1. *A program for calculating the power of detecting the relationships between two sequences under alternative model I*

To facilitate the use of D_2^* or D_2^S for sequence or genome comparison and for evaluation of statistical power for detecting the relationship between the sequences, a web-based online program (http://meta.cmb.usc.edu/d2) and a stand-alone R program were developed to calculate the power of sequence comparison using these statistics. We describe the program for the above model. However, the program can be easily extended to the general scenario of different background letter frequencies, sequence lengths, and motif densities as in the supplementary materials. The inputs of the program are:

1. The background nucleotide or amino acid frequencies $p_l^{(0)}, l=0,1, \cdots L-1$ of the two sequences **A** and **B** under study;
2. the nucleotide or amino acid frequencies $p_l^{(m)},\ l=0,1, \cdots L-1,\ m=1,2,\cdots,M$ at each position of the motif (PWM);
3. the lengths n of the sequences **A** and **B**;
4. the motif density, $1-\lambda$, for the sequences **A** and **B**;
5. the type I error, α.

For each set of parameters, the program first calculates the mean $P_\lambda(\mathbf{w}) = \mathbb{E}_\lambda(X_{\mathbf{w}})$ for any word $\mathbf{w}$ and the covariance $\sigma_\lambda(\mathbf{w},\ \mathbf{w}') = \mathrm{Cov}(X_{\mathbf{w}},\ X_{\mathbf{w}'})$ for two words $\mathbf{w}$ and $\mathbf{w}'$, related to sequence **A**. The corresponding quantities related to sequence **B** are also calculated. Secondly, the program calculates the approximate variance, $2(\Sigma_\lambda)^2$, $2(\Sigma_\lambda^*)^2$, and $2(\Sigma_\lambda^S)^2$ of D_2, D_2^*, and D_2^S using formulas derived in Theorems 2.2, 2.3, and 2.4, respectively. Thirdly, for the given type I error α, the threshold values z_α, z_α^*, and z_α^S for the corresponding statistics D_2, D_2^* or D_2^S in Theorem 2.5 are calculated. Since the cumulative distribution functions of Z_1^* and Z_1^S are not readily available, a simulation based method is used to obtain the threshold values. A large number of independent sequence pairs are simulated according to the specified letter frequencies and the sequence lengths, and the empirical distributions of D_2^* and D_2^S are estimated. The threshold values are estimated by the upper α% quantile of the simulated values of each statistic. Finally, the values of $C(\lambda)$, $C^*(\lambda)$, and $C^S(\lambda)$, and thus the power using the corresponding statistics in Theorem 2.5 is calculated.

We use the program to study the power of detecting the relationship between related sequences under alternative model I using the different statistics. In Subsection 4.2, we present the results for the parameter sets used in Reinert et al. (2009) and compare the results derived using our program with the simulated quantities in previous studies. In Subsection 4.3, we present the power of the various statistics for comparing the relationships between sequences when any of the motifs with motif length at most 10 in JASPAR (Sandelin et al., 2004) are present in both sequences.

4.2. Comparison of theoretical mean, standard deviation, and power of D_2, D_2^*, *and* D_2^S *with their corresponding simulated values from Reinert et al. (2009)*

In this subsection, we present some numerical results on the mean, standard deviation, and power of detecting the relationships between two sequences for the three statistics D_2, D_2^*, and D_2^S under the alternative model I using the same set of parameters as in Reinert et al. (2009). The objective is to see how close the corresponding quantities calculated using our formulas approximate the true values. We let the background letter frequencies for the two sequences be $p_A = p_{T=\frac{1}{6}}$, $p_C = p_G = \frac{1}{3}$. The inserted motif is "AGCCA" and the motif density $1 - \lambda = 0.01$. The size of the k-tuple is $k = 5$. We used 10,000 simulations to find the threshold values $z_{0.05}$, $z^*_{0.05}$, and $z^S_{0.05}$. The type I error α was set at 0.05 and 0.01.

For scaled D_2, D_2^*, and D_2^S defined respectively by

$$ND_2 = \sqrt{n}\left(\frac{D_2}{n^2} - \sum_{\mathbf{w}\in\mathcal{A}^k} p_{\mathbf{w}}^2\right), \quad ND_2^* = D_2^*/\sqrt{n}, \quad ND_2^S = D_2^S/\sqrt{n},$$

from Theorems 2.2, 2.3, and 2.4, it can be seen that the (approximate) means are $\sqrt{n}(A(\lambda) - A(1))$, $\sqrt{n}A^*(\lambda)$ and $\sqrt{n}A^S(\lambda)$, respectively. Similarly, the approximate variance of ND_2, ND_2^* and ND_2^S are $2(\Sigma_\lambda)^2$, $2(\Sigma^*_\lambda)^2$, and $2(\Sigma^S_\lambda)^2$, respectively.

Table 1 shows the simulated mean and standard deviation of ND_2, ND_2^*, and ND_2^S, respectively, and their corresponding limits. Surprisingly, the approximate mean and standard deviation of ND_2 are within 15% of their limit even when the sequence length is just 1Kbp. For D_2^*, the simulated mean is roughly the same as the theoretical limit and the simulated standard deviation is within 21% of its theoretical limit when the sequence length is at least 1Kbp. However, the simulated mean of D_2^S is much smaller than its limit. The corrected mean for D_2^S is very different from the

Table 1. Comparison of Simulated Mean and Variance of ND_2, ND_2^*, and ND_2^S for Different Sequence Length n with the Corresponding Theoretical Limits (the last row), with $(p_A, p_C, p_G, p_T) = (1/6, 1/3, 1/3, 1/6)$, $\lambda = 0.99$, Motif = "AGCCA", and Word Length $k = 5$

	D_2		D_2^*		D_2^S		
$n * 10^{-4}$	$\frac{END_2 * 10^4}{\sqrt{n}}$	$\sigma(ND_2) * 10^3$	$\frac{END_2^* * 10}{\sqrt{n}}$	$\sigma(ND_2^*)$	$\frac{END_2^S}{\sqrt{n}} * 10^2$	$A_m^S * 10^2$	$\sigma(ND_2^S) * 10$
0.1	0.92	4.09	1.34	2.41	4.35	−1032	6.80
0.12	0.92	4.12	1.33	2.34	4.03	−859	7.01
0.14	0.94	4.08	1.34	2.34	3.73	−735	7.15
0.16	0.97	4.03	1.34	2.27	3.60	−642	6.99
0.18	0.98	4.01	1.35	2.23	3.51	−570	7.07
0.2	0.98	3.95	1.35	2.24	3.38	−512	7.05
0.3	0.99	3.90	1.33	2.14	3.35	−339	7.32
0.4	1.01	3.86	1.34	2.11	3.48	−252	7.39
0.5	1.02	3.85	1.34	2.09	3.57	−200	7.47
0.6	1.03	3.84	1.34	2.08	3.66	−165	7.61
1	1.03	3.82	1.34	2.05	3.98	−96	7.90
2	1.04	3.80	1.34	2.03	4.48	−44	8.38
20	1.04	3.71	1.34	2.00	6.46	2.8	9.32
1000	1.05	3.76	1.34	1.95	7.90	7.89	7.83
Theory	1.05	3.76	1.34	1.99	7.99	7.99	7.72

As before, σ denotes standard deviation.

simulated mean, too, probably because the difference between $P_\lambda(\mathbf{w}) - p_\mathbf{w}$ for most 5-tuples are very small; both approximations do not work well in this parameter regime. Therefore, while we expect that the power formulas we derived should approximate the true power of D_2 and D_2^* well even for sequences of over 1Kbp long the power formula for D_2^S can significantly over-estimate the true power.

Table 2 shows the theoretical approximate power of D_2, D_2^*, and D_2^S calculated using our formulas and the simulated power with the same setting as in Table 1. The results show that the approximations are very close for D_2 and D_2^*. However, the theoretical approximate power based on the first approximation significantly over-estimates, while the approximate power based on the second approximation significantly under-estimates the simulated power for D_2^S, in the parameter regime we consider.

As the approximate power for D_2^S is not accurate in the parameter regimes we have considered, in the following, we only show the results related to D_2 and D_2^* using the theoretical approximate power. Figure 1 shows the values of $C(\lambda)$ and $C^*(\lambda)$ (upper panel) and the power of D_2 and D_2^* for detecting the relationships between pairs of sequences (lower panel) as a function of sequence length and the word length k when $\lambda = 0.99$. It should be noted that the power is a decreasing function of the values of C's and the smaller the values of C, the higher the power of the corresponding statistic is. From the left panel related to D_2, it can be seen that, when $k = 2$ or 3, the value of C actually increases and that the power $1 - \Phi(C(\lambda))$ decreases

Table 2. Comparison of the Theoretical and the Simulated Power Under Alternative Model I for Different Values of Sequence Length, with $(p_A, p_C, p_G, p_T) = (1/6, 1/3, 1/3, 1/6)$, $\lambda = 0.99$, Motif = "AGCCA", and Word Length $k = 5$

	D_2		D_2^*		D_2^S		
$n * 10^{-4}$	*Theory*	*Simulated*	*Theory*	*Simulated*	*Theory1*	*Theory2*	*Simulated*
Type I error $\alpha = 5\%$							
0.1	21	20	85	81	87	0	33
0.12	25	23	91	88	94	0	39
0.14	29	26	95	93	98	0	45
0.16	32	29	97	97	99	0	52
0.18	32	29	98	98	100	0	57
0.2	38	35	99	99	100	0	62
0.3	49	45	100	100	100	0	81
0.4	59	55	100	100	100	0	93
0.5	66	63	100	100	100	0	97
0.6	73	71	100	100	100	0	99
1	90	89	100	100	100	0	100
2	99	99	100	100	100	0	100
20	100	100	100	100	100	100	100
1000	100	100	100	100	100	100	100
Type I error $\alpha = 1\%$							
0.1	4	5	71	66	72	0	16
0.12	7	8	82	77	83	0	18
0.14	8	9	88	84	91	0	21
0.16	11	11	93	92	96	0	29
0.18	11	11	96	96	98	0	33
0.2	14	14	97	97	99	0	36
0.3	22	20	100	100	100	0	60
0.4	31	28	100	100	100	0	81
0.5	38	36	100	100	100	0	90
0.6	45	43	100	100	100	0	96
1	71	70	100	100	100	0	100
2	96	96	100	100	100	0	100
20	100	100	100	100	100	100	100
1000	100	100	100	100	100	100	100

As before, σ indicates standard deviation.

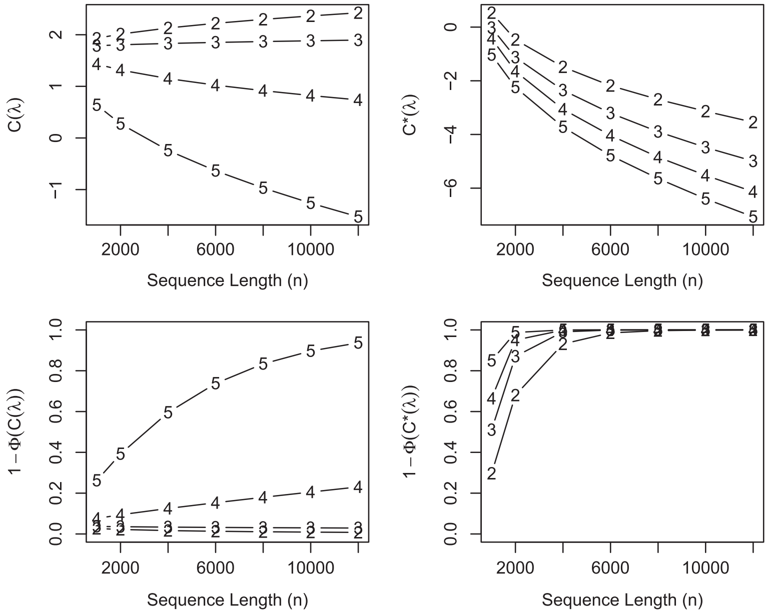

FIG. 1. The values of $C(\lambda)$ and $C^{*}(\lambda)$ (upper panels) and the power of $D2$ and D_2^* (lower panels) for detecting the relationships between sequence pairs related through alternative model I for different values of word size $k=2$, 3, 4, 5 and sequence length n. The parameters were set at $p_A=p_T=1/6$, $p_C=p_G=1/3$, $\lambda=0.99$, and type I error $\alpha=0.05$.

with the sequence length. For given sequence length and word size k, the power of D_2^* is generally higher than the power of D_2. All these conclusions are consistent with the simulation studies in Reinert et al. (2009). Comparing the two figures in the lower panel of Figure 1 here with Figures 1 and 2 in Reinert et al. (2009), respectively, we can see that the the theoretical power is slightly higher than the simulated power, but the difference is generally small, less than 10% in all the situations considered.

Simulation studies can only explore the influence of a relatively small range of parameter sets on the power of the different tests. With the theoretical results presented in this paper, we are able to explore a much larger parameter space. Theorem 2.5 shows that the power of D_2 and D_2^* is mainly determined by $B(\lambda)$ and $B^{*}(\lambda)$, respectively. The higher the values of B's, the more powerful the test is. Therefore, we also plot the values of $B(\lambda)$ and $B^{*}(\lambda)$ for $k=2$, 3, 4, 5 and $\lambda=0.93$ or 0.99 (Fig. 2). Again it is shown that $B^{*}(\lambda)$ is generally larger than $B(\lambda)$ indicating that D_2^* is generally more powerful than D_2. We note that both B and B^{*} decrease when λ increases. The smaller λ is, the larger is the probability of inserting a motif, and the eaiser it is to detect a difference from the null model.

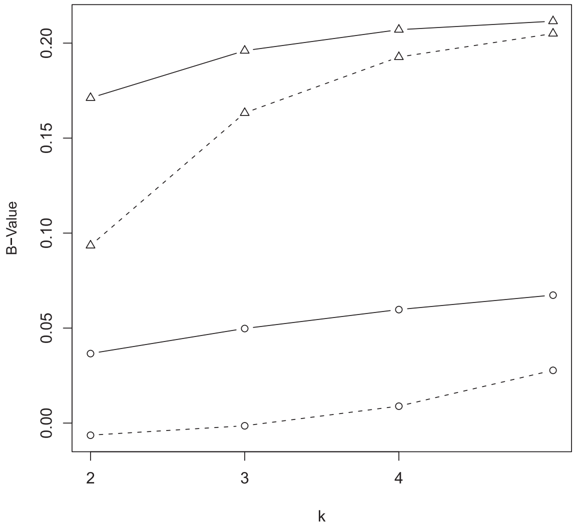

FIG. 2. The values of $B(\lambda)$ and $B^{*}(\lambda)$ for $\lambda=0.93$, 0.99 and $k=2$, 3, 4, 5. Dashed lines refer to B and solid lines to B^{*}; triangles refer to $\lambda=0.93$ and circles to $\lambda=0.99$. $B(0.99)$, dash line with circle points; $B(0.99)$, dash line with triangle points; $B^{*}(0.99)$, solid line with circle points; $B^{*}(0.99)$, solid line with triangle points.

4.3. The power of D_2 *and* D_2^* *for comparing two sequences when motifs in JASPAR are present*

Since the approximate distribution of D_2^S in Theorem 2.4 requires very long sequences and the resulting formula for calculating the power of D_2^S significantly over-estimates the true power, we will not consider D_2^S in the following. We next investigate whether the relative performance of D_2 and D_2^* for comparing the relationships between two sequences holds for a large class of motifs. To achieve this objective, we downloaded the transcription factor (TF) binding sites in the database JASPAR CORE (Sandelin et al., 2004) as motifs and studied the power of detecting the relationship between two sequences if such motifs are present in the sequences. The same letter frequencies for the background as in Reinert et al. (2009) are used. The theoretical formulas obtained in this paper make such large scale comparisons possible. Due to the long computational time required when the motif length is large, we only consider motifs with length at most 10.

A total of 323 transcription factor binding profiles with length at most 10 from JASPAR CORE (Sandelin et al., 2004) (October 12, 2009 version) are currently available. These motifs represent the most abundant publicly available knowledge regarding nucleotide sequence motifs. The corresponding PWMs are used to insert motifs as in alternative model I. Based on these assumptions, we can calculate the values of $B(\lambda)$, $B^*(\lambda)$, $C(\lambda)$, $C^*(\lambda)$, and the corresponding power for different values of word length k and motif density $1-\lambda$. The resulting figures and the corresponding letter frequencies in each position for all the motifs are presented in the supplementary material. From this large-scale study, we can conclude that D_2^* is more powerful than D_2 in more than 90% of the motifs. An example motif profile "MA0003" for which D_2 is more powerful than D_2^* is given in Figure 3. Note that in this motif, the overall frequencies of (A, C, G, T) in the motif are (0.11, 0.40, 0.40 0.09).

We then calculate the mean overall letter frequencies of (A, C, G, T) in those motifs for which D_2 is more powerful than D_2^* for at least three of the $k=2$, 3, 4, 5 ($\lambda=0.93$) and the corresponding frequencies are (0.08, 0.22, 0.57, 0.13). On the other hand, the mean overall letter frequencies of (A, C, G, T) in the other motifs are (0.30, 0.22, 0.25, 0.23). Under the background sequence model with (A, C, G, T) frequencies equal to (1/6, 1/3, 1/3, 1/6), in general, the GC content of the motifs for which D_2 is more powerful than D_2^* is higher than that of the other motifs under the background model considered in this article. If the background sequence model is changed, the PWM of motifs for which D_2 outperforms D_2^* should also change. As a general rule, D_2 outperforms D_2^* if the letter frequencies in a motif are close to the background letter frequencies.

Since we found that, in most situations, the power of D_2 can be even smaller than the type I error, whereas the power of D_2^* always approaches 1 for sequence length tending to infinity, we do not suggest using D_2 in general situations even if it can potentially perform well in some special cases.

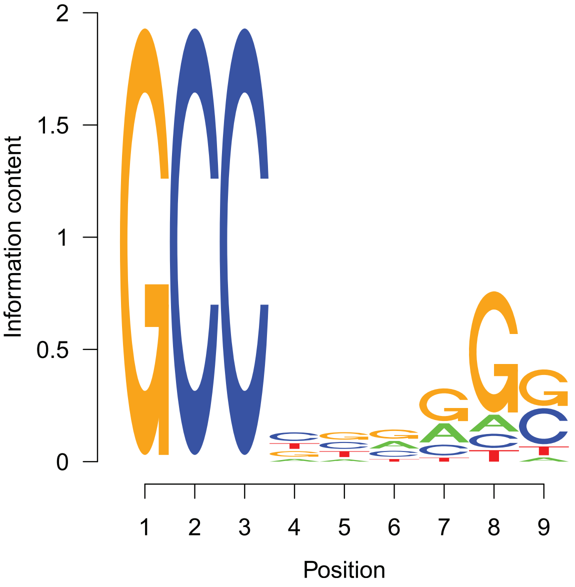

FIG. 3. The sequence LOGO of motif "MA0003".

4.4. The power of D_2, D_2^*, *and* D_2^S *for detecting the relationships between two sequences under alternative model II*

Previous simulation studies have shown that, under alternative model II, the power of D_2 is less than 0.4 and decreases with sequence length n when the word size k is 2 to 6. Actually, we can show that when n is large, the power of D_2 is always less than 0.5 for any parameter set. Note that Theorem 3.3 shows that the power of D_2 is approximately $P(\widetilde{Z}_\lambda \geq \widetilde{z}_\alpha)$. Since $\widetilde{z}_\alpha$ is positive and $\widetilde{Z}_\lambda$ is approximately normally distributed, the power is less than 0.5 when the sequence length is large for any set of parameters. However, similar arguments will not work for D_2^* and D_2^S since the distributions of $\widetilde{Z}_\lambda^*$ and $\widetilde{Z}_\lambda^S$ are not normal when $\lambda < 1$. This shows that D_2 does not have enough power to detect the relationship between sequences under alternative model II. So we will not study D_2 further under alternative model II. Previous simulation studies also showed that D_2^S is less powerful than D_2^*. So we now concentrate on further understanding D_2^* and D_2^S.

Theorems 3.2 and 3.3 show that the power of detecting the relationships between two sequences related through the alternative model II using any of D_2, D_2^*, and D_2^S reaches its plateau quickly as the sequence length increases, and the limit is generally much smaller than 1. Theorems 3.2 and 3.3 justify the simulation results that the simulated power by any of the statistics tends to a limit which is typically less than 1 when sequence length goes to infinity (Reinert et al., 2009), which was quite intriguing at the time of the simulation studies. Let T be any one of statistics D_2, D_2^*, and D_2^S. It is theoretically shown here that the primary reason for the power of T to be stable with respect to sequence length n is that there exist constants a_n and b_n such that $U_{\lambda,n} = a_n(T - b_n)$ approximates non-degenerate random variables U_λ under both the null model ($\lambda = 1$) and the alternative model $\lambda < 1$. Although U_λ is stochastically decreasing with respect to λ, the power of the test approaches a constant $P(U_\lambda \geq u_\alpha)$, where $P(U_1 \geq u_\alpha) = \alpha$. In order for the power of T to increase with respect to sequence length n and to finally reach 1, we need that (1) $U_{1,n}$ approximates a non-degenerate random variable U_1 under the null model ($\lambda = 1$), and (2) $U_{\lambda,n}$ tends to infinity as n tends to infinity.

Another interesting observation from previous simulation studies is that the power of D_2^* seems to increase with the length, k, of word pattern used (see Figure 8 in Reinert et al. (2009)). In order to explain this phenomenon, we study the mean D_2^* as a function of word length k. We are aware that in general the power of a test depends on the distributions of the test statistics under the null and the alternative hypothesis, not just the mean and/or the variance. However, as an explanation to the intriguing observation, we try to see if $\mathbb{E}(D_2^*)$ increases with k when other parameters are fixed. Theorem 3.1 (d) shows that

$$\lim_{n\to\infty}(ED_2^*) = \sum_{\mathbf{w}\in\mathcal{A}^k}\left[\frac{\gamma_0(\mathbf{w}) + 2\sum_{j=1}^{k-1}\gamma_j(\mathbf{w},\mathbf{w})}{p_{\mathbf{w}}}\right] - (2k-1) = S(\lambda, k).$$

Figure 4 shows the relationship between $S(\lambda, k)$ and $\lambda \in (0.9, 1)$ for $k = 2, 4, 6, 8, 10$. It can be seen that $S(\lambda, k)$ increases with k for any $\lambda \in (0.9, 1.0)$, as does the discrepancy between $S(\lambda, k)$ and $S(1, k)$ for $\lambda < 1$. As our statistic is based on comparing the means of the counts under the two models, this partially explains that the power of D_2^* increases with word length k.

5. DISCUSSION

Alignment-free sequence comparison has become increasingly important as new sequencing technologies can generate enormous amount of sequence data in a relative short time and at low cost. However, the statistics used for alignment-free sequence comparison are usually ad-hoc, and it is not clear whether such ad-hoc statistics can actually find the relationships between sequences. It is also important to know under which evolutionary models the statistics are meaningful. One of the widely discussed and studied statistics for alignment free sequence comparison is the D_2 statistic. Previously simulation studies have shown the limitations of D_2 in detecting the relationships between sequences under a common motif model (alternative model I) and a pattern transfer model (alternative model II). It was shown that the power of D_2 can even be smaller than the pre-specified type I error under some situations. Two new statistics, D_2^* and D_2^S, were developed to overcome the inherent problems of D_2 and simulation studies showed their superior performance compared to D_2 (Reinert et al., 2009).

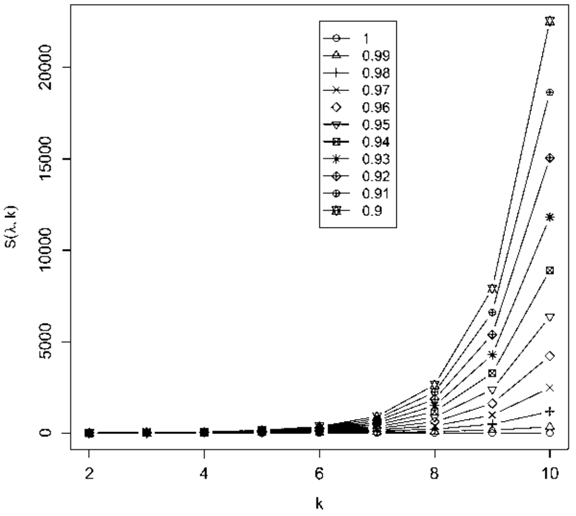

FIG. 4. The values of $S(\lambda, k) = \lim_{n\to\infty} \mathbb{E}(D_2^*)$ as a function of motif density λ and word length k, $\lambda = 0.9$ to 1.0 by step 0.01, and $k = 2, \cdots, 10$

However, the approximate distributions of these statistics were not known at the time of the study (Reinert et al., 2009), and thus, it was not possible to give a theoretical formula to calculate the power of the different tests. Having the limiting distribution of the test statistics can help us design algorithms to calculate the power. With the power calculator, we are able to explore a large range of the parameter space and study how the parameters individually and collectively contribute to the power of the tests. The theoretical studies also give insights into when and how the test statistics can be applied to compare sequences. In this paper, we carried out a systematic theoretical study of the power of D_2, D_2^* and D_2^S for detecting the relationships between sequences under alternative models I and II. Under alternative model I, we provided an easy to use program to calculate the power of the test statistics D_2 and D_2^* for different combinations of parameters. Using the program, we then obtained the theoretical power and compared with the simulated power using the same parameters as in Reinert et al. (2009) and showed that they are generally close, thus validating the usefulness of our program. However, the convergence of D_2^S to our theoretical limit is very slow and the approximation is only reasonable for very long sequences. We then carried out a large-scale comparison of D_2 and D_2^* statistics for sequence comparison under alternative model I when the motif is any one of the 323 motifs with length at most 10 in JASPAR CORE. Our program made such a large-scale comparison possible. We verified the relative performance of D_2 and D_2^* observed in previous studies, i.e. D_2^* is generally more powerful than D_2. Under alternative model II, we theoretically showed that the power of the three statistics tends to a constant, usually less than 1. We also gave some reasons why the power of D_2^* increases with the word size k.

This study has several limitations regarding the models of the background sequences and the foreground motif models. The IID model was used to model the background sequence. It is known that the genomes of organisms are hierarchically organized (Mantegna et al., 1994) and simple IID models cannot fully describe the background sequences; instead high-order Markovian models could be more appropriate. Similarly, the positions of the motifs are assumed independent and again this assumption can be violated in many motifs. To incorporate such complexity into our model, high-order HMMs can potentially be used; the calculations would then become much more involved. Although the extensions to higher order HMM are conceptually simple, heavy computational issues need to be solved.

We made several simple assumptions regarding the distribution of the motifs along the sequences as in Reinert et al. (2009). First it was assumed that the motifs are uniformly distributed along the sequences. Motifs can cluster together in some regions and may be sparse in other regions of the sequences. If such inhomogeneity is known to be present, an inhomogeneous HMM can be used to model the distribution of motifs by assuming large motif density λ in motif-clustered regions and low motif density λ in sparse motif regions. If such motif-clustered and motif-sparse regions are unknown, but suspected, we can assume that λ is a random variable following certain distributions. Second, we considered the presence of just one motif along the sequences. In many situations, several motif patterns work together to form modules. How to model such sequences is a problem for future studies. Third, we emphasize that the three statistics we consider here are most likely not optimal and other more powerful statistics may possibly be constructed. Fourth, applying these statistics to practical examples is another topic for future research.

In this article, we theoretically showed that, under alternative model II, the power of D_2, D_2^*, and D_2^S converges to a value that is generally much less than 1 when the sequence length tends to infinity. Therefore, they are not appropriate to test for relationships between sequences under this model. The obvious important question is which statistics based on word counts should be used for testing against this model instead.

6. APPENDIX A: PROOFS OF THE THEOREMS

In this Appendix, we prove the theorems in the main text.

A.1. Proofs of Theorems 2.2–2.5 under alternative model I

Proof of Theorem 2.2. From the definition of D_2, we have

$$\begin{aligned}\frac{D_2}{n^2} &= \sum \frac{X_{\mathbf{w}}}{n}\frac{Y_{\mathbf{w}}}{n}\\ &= \sum \left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w}) + P_\lambda(\mathbf{w})\right)\left(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w}) + P_\lambda(\mathbf{w})\right)\\ &= \sum \left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right)\left(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right) + \sum P_\lambda(\mathbf{w})\left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right)\\ &\quad + \sum P_\lambda(\mathbf{w})\left(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right) + \sum (P_\lambda(\mathbf{w}))^2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\sqrt{n}\Big(\frac{D_2}{n^2} - \sum_{\mathbf{w}} (P_\lambda(\mathbf{w}))^2\Big)\\ &= \quad \frac{1}{\sqrt{n}}\sum_{\mathbf{w}} \sqrt{n}\Big(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\Big)\sqrt{n}\Big(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\Big)\\ &\quad + \sum_{\mathbf{w}} P_\lambda(\mathbf{w})\Big(\sqrt{n}\Big(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\Big) + \sqrt{n}\Big(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\Big)\Big).\end{aligned} \tag{9}$$

It has been shown in Zhai et al. (2010), Proposition 2.4, for $0 < \lambda < 1$, and in Reinert et al. (2009), Proposition 6.1, for $\lambda = 1$, that, in distribution,

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right) = \lim_{n\to\infty} \sqrt{n}\left(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right) = N(0, \sigma_\lambda^2(\mathbf{w})), \tag{10}$$

where $\sigma_\lambda^2(\mathbf{w}) = \lim_{n\to\infty} \frac{\mathrm{Var}(X_{\mathbf{w}})}{n}$. Therefore, the first term in equation (9) tends to 0 when $n \to \infty$, with alphabet size fixed, and

$$\sqrt{n}\left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right) + \sqrt{n}\left(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right) \to \mathcal{N}(0, 2\sigma_\lambda^2(\mathbf{w})).$$

Let $\sigma_\lambda(\mathbf{w}, \mathbf{w}') = \lim_{n\to\infty} \frac{\mathrm{Cov}(X_{\mathbf{w}}, X_{\mathbf{w}'})}{n}$ which can be calculated as in Zhai et al. (2010) for $0 < \lambda < 1$, and as in Reinert et al. (2009) for $\lambda = 1$. Since $\{X_{\mathbf{w}}, \mathbf{w} \in \mathcal{A}^k\}$ and $\{Y_{\mathbf{w}}, \mathbf{w} \in \mathcal{A}^k\}$ are independent, the second term in (9) is asymptotically normal with mean 0 and variance $2(\Sigma_\lambda)^2$. Theorem 2.2 is proved.

We note that the proof of Theorem 2.2 breaks down when all letters are equally likely, as then with $p = p_{\mathbf{w}}$,

$$\sum_{\mathbf{w}} p\sqrt{n}\left(\frac{X_{\mathbf{w}}}{n}-p\right)=0$$

and thus the second term in (9) vanishes.

Proof of Theorem 2.3. The proof of Theorem 2.3 is similar to the proof of Theorem 2.2. The first part can be easily proved using the normal approximation Corollary 6.1 in Reinert et al. (2009) for the individual centered word counts, which also holds when all letters are equally likely. To prove the second part, note that

$$\begin{aligned}\frac{D_2^*}{n} &= \sum\frac{1}{p_{\mathbf{w}}}\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)\left(\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)+\sum\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})}{p_{\mathbf{w}}}\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)\\ &\quad+\sum\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})}{p_{\mathbf{w}}}\left(\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)+\sum\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})^2}{p_{\mathbf{w}}}.\end{aligned}$$

It follows from the normal approximation for individual word counts that, in distribution,

$$\sqrt{n}\sum\frac{1}{p_{\mathbf{w}}}\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)\left(\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)\to 0 \text{ as } n\to\infty.$$

Therefore, in distribution,

$$\begin{aligned}&\lim_{n\to\infty}\sqrt{n}\left(\frac{D_2^*}{n}-\sum\frac{(P_\lambda(W)-p_{\mathbf{w}})^2}{p_{\mathbf{w}}}\right)\\ &\quad=\lim_{n\to\infty}\sum\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})}{p_{\mathbf{w}}}\sqrt{n}\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})+\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right).\end{aligned}$$

For $0<\lambda<1$, under the assumption that $\frac{(P_\lambda(\mathbf{w})-p_{\mathbf{w}})}{p_{\mathbf{w}}}$ is not constant in w, this expression has a normal distribution with mean 0 and variance $2(\Sigma_\lambda^*)^2$, where (Σ_λ^*) is given in (3). Theorem 2.3 is proved.

Proof of Theorem 2.4. The first part of Theorem 2.4 has been proved in Theorem 2.1 in Reinert et al. (2009). We only present the outline for the proof of the second part. Using Taylor expansion, it is straightforward to show that for any $a\neq 0$ and (x, y) in the neighborhood of (0,0),

$$\frac{(x+a)(y+a)}{\sqrt{(x+a)^2+(y+a)^2}}=\frac{|a|}{\sqrt{2}}+\frac{\mathrm{sign}(a)}{2\sqrt{2}}(x+y)+O(x^2+y^2),$$

where $O(x^2+y^2)$ indicates a term such that there exists a constant C with

$$|O(x^2+y^2)|\leq C(x^2+y^2).$$

For each word $\mathbf{w}$, let $a=P_\lambda(\mathbf{w})-p_{\mathbf{w}}$, $x=\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})$, and $y=\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})$. Then, with this Taylor expansion,

$$\begin{aligned}\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{n\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}} &= \frac{|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|}{\sqrt{2}}+\frac{\mathrm{sign}(P_\lambda(\mathbf{w})-p_{\mathbf{w}})}{2\sqrt{2}}\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})+\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)\\ &\quad+O\left(\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)^2+\left(\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)^2\right).\end{aligned} \tag{11}$$

Taking expectations in (11) we obtain that

$$E_\lambda\left(\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{n\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}}\right)=\frac{|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|}{\sqrt{2}}+O\left(E_\lambda\left(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)^2+E_\lambda\left(\frac{Y_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w})\right)^2\right).$$

As $E_\lambda(\frac{X_{\mathbf{w}}}{n}-P_\lambda(\mathbf{w}))^2=\frac{1}{n}\mathrm{Var}_\lambda(\frac{X_{\mathbf{w}}}{\sqrt{n}})=O(n^{-1})$, we obtain that the asymptotic mean of $\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{n\sqrt{\tilde{X}_{\mathbf{w}}^2+\tilde{Y}_{\mathbf{w}}^2}}$ equals $\frac{|P_\lambda(\mathbf{w})-p_{\mathbf{w}}|}{\sqrt{2}}$.

Moreover, summing Equation (11) over all the word patterns $\mathbf{w} \in \mathcal{A}^k$, we have

$$\begin{aligned}
&\sqrt{n}\left(\frac{D_2^S}{n} - \sum_{\mathbf{w}\in\mathcal{A}^k}\frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}}\right)\\
&= \sqrt{n}\sum\left(\frac{\tilde{X}_{\mathbf{w}}\tilde{Y}_{\mathbf{w}}}{n\sqrt{\tilde{X}_{\mathbf{w}}^2 + \tilde{Y}_{\mathbf{w}}^2}} - \frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}}\right)\\
&= \sum\frac{\text{sign}(P_\lambda(\mathbf{w}) - p_{\mathbf{w}})}{2\sqrt{2}}\sqrt{n}\left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w}) + \frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right)\\
&\quad + \frac{1}{\sqrt{n}}\sum O\left(n\left(\frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right)^2 + n\left(\frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})\right)^2\right).
\end{aligned}$$

Similar as in the proof of Theorem 2.2, under the assumption that $P_\lambda(\mathbf{w}) - p(\mathbf{w})$ is not constant in w, we see that $\sqrt{n}\left(\frac{D_2^S}{n} - \sum_{\mathbf{w}\in\mathcal{A}^k}\frac{|P_\lambda(\mathbf{w}) - p_{\mathbf{w}}|}{\sqrt{2}}\right)$ is asymptotically normal with mean 0 and variance $2(\Sigma_\lambda^S)^2$.

For the last assertion, we refine the Taylor expansion to

$$\frac{(x+a)(y+a)}{\sqrt{(x+a)^2 + (y+a)^2}} \approx \frac{|a|}{\sqrt{2}} + \frac{sign(a)}{2\sqrt{2}}(x+y) - \frac{3}{8\sqrt{2}|a|}\{x^2 - 2xy + y^2\},$$

and using $a = P_\lambda(\mathbf{w}) - p_{\mathbf{w}}$, if $P_\lambda(\mathbf{w}) - p_{\mathbf{w}} \neq 0$, $x = \frac{X_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})$, and $y = \frac{Y_{\mathbf{w}}}{n} - P_\lambda(\mathbf{w})$, taking expectations completes the proof of Theorem 2.4.

Proof of Theorem 2.5. The proof of the three equations are roughly the same, and thus we only give the proof for the first equation.

Note that under the alternative model I, we expect that the k-tuple counts for the two sequences are more correlated than that for two random sequences. Therefore we use one-sided test. For fixed type I error α, based on Theorem 2.2 (a), we find z_α such that $P\{Z_1 \geq z_\alpha\} = \alpha$. Under the null hypothesis that $\lambda = 1$, $n^{-2}D_2 - \sum p_{\mathbf{w}}^2$ has approximate mean zero, whereas under the alternative $\lambda < 1$, the approximate mean of $n^{-2}D_2 - \sum p_{\mathbf{w}}^2$ will not be zero. We reject the null hypothesis if $Z_1 > z_\alpha$, which is approximately equivalent to $D_2 > n^2A(1) + z_\alpha\sqrt{n^3}$. The power for D_2 is

$$\begin{aligned}
1 - \beta &= P_\lambda(D_2 > n^2A_1(1) + z_\alpha\sqrt{n^3})\\
&= P_\lambda\left(\frac{D_2 - n^2A(\lambda)}{\sqrt{2n^3(\Sigma_\lambda)^2}} \geq \frac{n^2A(1) + z_\alpha\sqrt{n}^3 - n^2A(\lambda)}{\sqrt{2}n^3(\Sigma_\lambda)^2}\right)\\
&\approx 1 - \Phi(C(\lambda)).
\end{aligned}$$

The last approximation holds because of Theorem 2.2 (b).

A.2. Proofs of Theorems 3.1, 3.2, and 3.3

Proof of Theorem 3.1. We calculate $\mathbb{E}(X_{\mathbf{w}}Y_{\mathbf{w}'})$ for any two words $\mathbf{w}$ and $\mathbf{w}'$ of length k. Let

$$\begin{cases} I_{\mathbf{w}}^A(i) = I(A_iA_{i+1}\cdots A_{i+k-1} = w_1w_2\cdots w_k) \\ I_{\mathbf{w}'}^B(i) = I(B_iB_{i+1}\cdots B_{i+k-1} = w_1'w_2'\cdots w_k') \end{cases},$$

then

$$X_{\mathbf{w}} = \sum_{i=1}^{n} I_{\mathbf{w}}^A(i), \qquad \text{and } Y_{\mathbf{w}'} = \sum_{i=1}^{n} I_{\mathbf{w}'}^B(i).$$

Thus

$$
\begin{aligned}
\mathbb{E}X_{\mathbf{w}}Y_{\mathbf{w}'} &= \mathbb{E}\sum_{i=1}^{n}\sum_{j=1}^{n} I_{\mathbf{w}}^{A}(i)I_{\mathbf{w}'}^{B}(j) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\, I_{\mathbf{w}}^{A}(i)I_{\mathbf{w}'}^{B}(j)\\
&= \sum_{i=1}^{n}\mathbb{E}\, I_{\mathbf{w}}^{A}(i)I_{\mathbf{w}'}^{B}(i) + \sum_{j=1}^{k-1}\sum_{i=1}^{n-j}\mathbb{E}\, I_{\mathbf{w}}^{A}(i)I_{\mathbf{w}'}^{B}(i+j) + \sum_{j=1}^{k-1}\sum_{i=1}^{n-j}\mathbb{E}\, I_{\mathbf{w}'}^{B}(i)I_{\mathbf{w}}^{A}(i+j)\\
&\quad + \sum_{j=k}^{n-1}\sum_{i=1}^{n-j}\mathbb{E}\, I_{\mathbf{w}}^{A}(i)I_{\mathbf{w}'}^{B}(i+j) + \sum_{j=k}^{n-1}\sum_{i=1}^{n-j}\mathbb{E}\, I_{\mathbf{w}'}^{B}(i)I_{\mathbf{w}}^{A}(i+j)\\
&= n\gamma_0(\mathbf{w},\mathbf{w}') + \sum_{j=1}^{k-1}(n-j)(\gamma_j(\mathbf{w},\mathbf{w}') + \gamma_j(\mathbf{w}',\mathbf{w})) + 2\sum_{j=k}^{n-1}\sum_{i=1}^{n-j} p_{\mathbf{w}}p_{\mathbf{w}'}\\
&= n\gamma_0(\mathbf{w},\mathbf{w}') + \sum_{j=1}^{k-1}(n-j)(\gamma_j(\mathbf{w},\mathbf{w}') + \gamma_j(\mathbf{w}',\mathbf{w}))\\
&\quad + [n^2-(2k-1)n+k(k-1)]p_{\mathbf{w}}p_{\mathbf{w}'},
\end{aligned}
$$

where $\gamma_0(\mathbf{w},\mathbf{w}') = P(A_l = w_l, B_l = w'_l,\ l = 1, 2, \cdots, k)$ and $\gamma_j(\mathbf{w},\mathbf{w}) = P(A_l = w_l, B_{j+l} = w'_l,\ l = 1, 2, \cdots, k)$. Part (a) of the theorem is proved.

Note that

$$\mathrm{Cov}(X_{\mathbf{w}}, Y_{\mathbf{w}'}) = \mathbb{E}(X_{\mathbf{w}}, Y_{\mathbf{w}'}) - n^2 p_{\mathbf{w}}p_{\mathbf{w}'}.$$

Then part (b) can be easily deduced from part (a).

Part (c) and (d) can be proved by the definition of D_2 and D_2^*, respectively, and by part (b) above by letting $\mathbf{w} = \mathbf{w}'$. The recursion follows as in Reinert et al. (2009).

Proof of Theorem 3.2. The proofs of parts (a), (b), and (c) of the theorem are similar to that of Theorems 2.2–2.4, respectively.

(a) As in the proof of Theorem 2.2, we have

$$
\begin{aligned}
\sqrt{n}\left(\frac{D_2}{n^2} - \sum p_{\mathbf{w}}^2\right) &= \frac{1}{\sqrt{n}}\sum \sqrt{n}\left(\frac{X_{\mathbf{w}}}{n} - p_{\mathbf{w}}\right)\sqrt{n}\left(\frac{Y_{\mathbf{w}}}{n} - p_{\mathbf{w}}\right)\\
&\quad + \sum p_{\mathbf{w}}\left(\sqrt{n}\left(\frac{X_{\mathbf{w}}}{n} - p_{\mathbf{w}}\right) + \sqrt{n}\left(\frac{Y_{\mathbf{w}}}{n} - p_{\mathbf{w}}\right)\right). \qquad (12)
\end{aligned}
$$

Under alternative model II, the marginal sequences are IID, and hence $\sqrt{n}\left(\frac{X_{\mathbf{w}}}{n} - p_{\mathbf{w}}\right)$ converges to a mean zero normal variable, call the asymptotic variance M_1; and $\sqrt{n}\left(\frac{Y_{\mathbf{w}}}{n} - p_{\mathbf{w}}\right)$ converges to the same limit. As the two count vectors are asymptotically jointly normal, we obtain that, in distribution,

$$
\sqrt{n}\begin{pmatrix} \frac{X_{\mathbf{w}}}{n} & - & p_{\mathbf{w}}, & \mathbf{w}\in\mathcal{A}^k\\ \frac{Y_{\mathbf{w}}}{n} & - & p_{\mathbf{w}}, & \mathbf{w}\in\mathcal{A}^k \end{pmatrix} \to N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} M_1 & \Delta_\lambda\\ \Delta_\lambda & M_1\end{pmatrix}\right), \qquad (13)
$$

where $M_1 = (\sigma_1(\mathbf{w},\mathbf{w}'))_{\mathbf{w},\mathbf{w}'}$ and $\Delta_\lambda = (\delta_\lambda(\mathbf{w},\mathbf{w}'))_{\mathbf{w},\mathbf{w}'}$.

Therefore, the first term in Equation 12 tends to 0 as n tends to infinity. The second term tends to a normal distribution with mean 0 and variance $2(\Lambda_\lambda)^2$. Part (a) is proved. Parts (b) and (c) follow directly from the normal approximation (13).

Proof of Theorem 3.3. The proof of this theorem is similar to the proof of Theorem 2.5. For illustration only, we prove the claim for the power of D_2^*. From Theorem 3.2 (b) with $\lambda = 1$, we can choose $\widetilde{z}_\alpha^*$ such that

$$P(\widetilde{Z}_1^* \geq \widetilde{z}_\alpha^*) = \alpha.$$

We reject the null hypothesis that the two sequences are not related if $D_2^* \geq \widetilde{z}_\alpha^*$. We use one sided test since the mean of D_2^* is expected to be greater than 0 under the alternative model. From Theorem 3.2 (b), the test has an approximate type I error α under the null hypothesis $\lambda = 1$.

The power is the probability that the null model is rejected under the alternative model II $\lambda < 1$. Thus, the power is

$$1 - \beta = P_\lambda(D_2^* \geq \widetilde{z}_\alpha^*) \approx P\{\widetilde{Z}_\lambda^* \geq \widetilde{z}_\alpha^*\}.$$

APPENDIX B: LIMIT DISTRIBUTIONS OF D_2, $\mathrm{D_2}^*$, AND $\mathrm{D_2}^{\mathrm{S}}$ WHEN THE TWO SEQUENCES HAVE DIFFERENT LETTER FREQUENCIES, MOTIF DENSITIES, AND SEQUENCE LENGTHS

For simplicity of presentation, we have so far assumed that the two sequences have the same letter frequency, motif density, and sequence length. The theorems in the main text can be easily extended to the general situations. Let n_X be the length and $1 - \lambda_X$ be the motif density for sequence **A**. Let $p_{\mathbf{w}}^X$ be the probability of pattern **w** under the null model and $P_{\lambda_X}^X(\mathbf{w})$ be the probability of word pattern w as calculated in subsection 2.1 for sequence **A**. Let $(\Sigma_{\lambda_X}^X)^2$ and $(\Sigma_{\lambda_X}^{X*})^2$ be similarly defined as in equations 2 and 3, respectively, by replacing λ with λ_X. Similar notation can be defined for sequence **B**; here we use the superscript or subscript Y. We define D_2 and D_2^S similarly as above by replacing $p_{\mathbf{w}}$ by $p_{\mathbf{w}}^X$ or $p_{\mathbf{w}}^Y$ appropriately. Let $C_{XY} = n_X/n_Y$. For simplicity of presentation, we also define $C_{YX} = n_Y/n_Y = 1/C_{XY}$. Under the general model, we redefine D_2^* as

$$D_2^* = \sum_{\mathbf{w} \in \mathcal{A}^k} \frac{\widetilde{X}_{\mathbf{w}} \widetilde{Y}_{\mathbf{w}}}{\sqrt{n_X p_{\mathbf{w}}^X}\sqrt{n_Y p_{\mathbf{w}}^Y}}.$$

In this general setting,

$$\frac{D_2^S}{\sqrt{n_x n_y}} = \sum_{\mathbf{w} \in \mathcal{A}^k} \frac{(\widetilde{X}_{\mathbf{w}}/n_X)(\widetilde{Y}_{\mathbf{w}}/n_Y)}{\sqrt{C_{XY}\widetilde{X}_{\mathbf{w}}^2/n_X^2 + C_{YX}\widetilde{Y}_{\mathbf{w}}^2/n_Y^2}}. \tag{14}$$

From the law of large numbers we deduce that, in distribution and almost surely, $\frac{\widetilde{X}_{\mathbf{w}}}{n_X} \rightarrow P_{\lambda_X}^X(\mathbf{w}) - p_{\mathbf{w}}^X$, and a similar statement holds for $\frac{Y_w}{n_Y}$. Hence, we abbreviate in connection with the asymptotic means, see Theorem 5.1

$$A^g(\lambda_X, \lambda_Y) = \sum_{\mathbf{w} \in \mathcal{A}^k} P_{\lambda_X}^X(\mathbf{w}) P_{\lambda_Y}^Y(\mathbf{w}),$$

$$A^{g*}(\lambda_X, \lambda_Y) = \sum_{\mathbf{w} \in \mathcal{A}^k} \frac{(P_{\lambda_X}^X(\mathbf{w}) - p_{\mathbf{w}}^X)(P_{\lambda_Y}^Y(\mathbf{w}) - p_{\mathbf{w}}^Y)}{\sqrt{p_{\mathbf{w}}^X p_{\mathbf{w}}^Y}},$$

$$A^{gS}(\lambda_X, \lambda_Y) = \sum_{\mathbf{w} \in \mathcal{A}^k} \frac{(P_{\lambda_X}^X(\mathbf{w}) - p_{\mathbf{w}}^X)(P_{\lambda_Y}^Y(\mathbf{w}) - p_{\mathbf{w}}^Y)}{\sqrt{C_{XY}(P_{\lambda_X}^X(\mathbf{w}) - p_{\mathbf{w}}^X)^2 + C_{YX}(P_{\lambda_Y}^Y(\mathbf{w}) - p_{\mathbf{w}}^Y)^2}},$$

where and in the following, the superscript "g" indicates the general model. In analogy to Theorems 2.1, 2.2, 2.3, 2.4, and 2.5, we have the following theorems. As the proofs are very similar to the ones presented in the article, they are omitted.

Theorem 5.1. *Under alternative model I for the two sequences as described above, the expectations of* D_2, D_2^* *and* D_2^S *can be calculated as follows.*

$$\mathbb{E}(D_2) = n_X n_Y A^g(\lambda_X, \lambda_Y),$$

$$\mathbb{E}(D_2^*) = \sqrt{n_X n_Y} A^{g*}(\lambda_X, \lambda_Y),$$

$$\textit{and} \lim_{n \to \infty} \frac{\mathbb{E}(D_2^S)}{\sqrt{n_X n_Y}} = A^{gS}(\lambda_X, \lambda_Y).$$

The limiting distributions of D_2, D_2^*, and D_2^S under the general model are given as follows.

Theorem 5.2. *Assume that in the background model not all letters are equally likely.*

a. Suppose $\lambda_X = \lambda_Y = 1$ *(the null model that the sequences are independent). Then*

$$\lim_{n\to\infty}(n_X n_Y)^{\frac{1}{4}}\left(\frac{D_2}{n_X n_Y} - \sum_{\mathbf{w}\in\mathcal{A}^k} p^X_{\mathbf{w}} p^Y_{\mathbf{w}}\right) = Z^g_1,$$

where Z^g_1 *has normal distribution* $\mathcal{N}(0, \sqrt{C_{YX}}(\Sigma^X_1)^2 + \sqrt{C_{XY}}(\Sigma^Y_1)^2)$*. Here the asymptotics is valid when the sequence length tends to infinity with alphabet size, motif length, and word length kept fixed.*

b. Suppose $0 < \lambda_X,\ \lambda_Y < 1$ *(the alternative model I). Then*

$$\lim_{n\to\infty}(n_X n_Y)^{\frac{1}{4}}\left(\frac{D_2}{n_X n_Y} - A^g(\lambda_X, \lambda_Y)\right) = Z^g_{\lambda_X,\,\lambda_Y},$$

where $Z^g_{\lambda_X,\,\lambda_Y}$ *has normal distribution* $\mathcal{N}(0, \sqrt{C_{YX}}(\Sigma^X_{\lambda_X})^2 + \sqrt{C_{XY}}(\Sigma^Y_{\lambda_Y})^2)$*. Here the asymptotics is valid when the sequence length tends to infinity with alphabet size, motif length, and word length kept fixed.*

For D^*_2, we have:

Theorem 5.3. *a. Suppose* $\lambda_X = \lambda_Y = 1$ *(the null model that the sequences are independent). Then, in distribution,*

$$\lim_{n\to\infty} D^*_2 = Z^{g*}_1 = \sum_{\mathbf{w}\in\mathcal{A}^k} \frac{Z^{(g1)}_{\mathbf{w}} Z^{(g2)}_{\mathbf{w}}}{\sqrt{p^X_{\mathbf{w}}}\sqrt{p^Y_{\mathbf{w}}}},$$

where $\{Z^{(g1)}_{\mathbf{w}}, \mathbf{w}\in\mathcal{A}^k\}$ *and* $\{Z^{(g2)}_{\mathbf{w}}, \mathbf{w}\in\mathcal{A}^k\}$ *are independent and have mean 0 normal distributions (with non-trivial covariance matrix).*

b. Suppose $0 < \lambda < 1$ *(the alternative model I), and that* $\frac{(P_\lambda(\mathbf{w}) - p_{\mathbf{w}})}{p_{\mathbf{w}}}$ *is not constant in* **w**. *Then, in distribution,*

$$\lim_{n\to\infty}(n_X n_Y)^{\frac{1}{4}}\left(\frac{D^*_2}{\sqrt{n_X n_Y}} - A^{g*}(\lambda_X, \lambda_Y)\right) = Z^{g*}_{\lambda_X,\,\lambda_Y},$$

where $Z^{g*}_{\lambda_X,\,\lambda_Y}$ *has normal distribution* $\mathcal{N}\left(0, \sqrt{C_{YX}}(\Sigma^{X*}_{\lambda_X})^2 + \sqrt{C_{XY}}(\Sigma^{Y*}_{\lambda_Y})^2\right)$.

In order to state the limit distribution for D^S_2, we let

$$u_{\mathbf{w}} = \frac{C_{YX}(P^Y_{\lambda_Y}(\mathbf{w}) - p^Y_{\mathbf{w}})^3}{\left\{C_{XY}(P^X_{\lambda_X}(\mathbf{w}) - p^X_{\mathbf{w}})^2 + C_{YX}(P^Y_{\lambda_Y}(\mathbf{w}) - p^Y_{\mathbf{w}})^2\right\}^{\frac{3}{2}}},$$

$$v_{\mathbf{w}} = \frac{C_{XY}(P^X_{\lambda_X}(\mathbf{w}) - p^X_{\mathbf{w}})^3}{\left\{C_{XY}(P^X_{\lambda_X}(\mathbf{w}) - p^X_{\mathbf{w}})^2 + C_{YX}(P^Y_{\lambda_Y}(\mathbf{w}) - p^Y_{\mathbf{w}})^2\right\}^{\frac{3}{2}}},$$

and

$$\left(\Sigma^{XS}_{\lambda_X}\right)^2 = \sum_{\mathbf{w}\in\mathcal{A}^k}(u_{\mathbf{w}})^2(\sigma^X_{\lambda_X})^2(\mathbf{w}) + \sum_{\mathbf{w}\neq\mathbf{w}'} u_{\mathbf{w}} u_{\mathbf{w}'} \sigma^X_{\lambda_X}(\mathbf{w}, \mathbf{w}'),$$

$$\left(\Sigma^{YS}_{\lambda_Y}\right)^2 = \sum_{\mathbf{w}\in\mathcal{A}^k}(v_{\mathbf{w}})^2(\sigma^Y_{\lambda_Y})^2(\mathbf{w}) + \sum_{\mathbf{w}\neq\mathbf{w}'} v_{\mathbf{w}} v_{\mathbf{w}'} \sigma^Y_{\lambda_Y}(\mathbf{w}, \mathbf{w}').$$

The following theorem gives the approximate distribution of D^S_2 under the null and the alternative models for the general situation.

Theorem 5.4. *a. Suppose* $\lambda_X = \lambda_Y = 1$ *(the null model that the sequences are independent). Then, in distribution,*

$$\lim_{n\to\infty}\frac{D_2^S}{(n_X n_Y)^{\frac{1}{4}}} = Z_1^{gS} = \sum_{\mathbf{w}\in\mathcal{A}^k}\frac{Z_{\mathbf{w}}^{(g1)}Z_{\mathbf{w}}^{(g2)}}{\sqrt{\sqrt{C_{XY}}(Z_{\mathbf{w}}^{(g1)})^2+\sqrt{C_{YX}}(Z_{\mathbf{w}}^{(g2)})^2}}, \tag{15}$$

where $\{Z_{\mathbf{w}}^{(g1)}, \mathbf{w}\in\mathcal{A}^k\}$ and $\{Z_{\mathbf{w}}^{(g2)}, \mathbf{w}\in\mathcal{A}^k\}$ are independent and have mean 0 normal distribution.

b. Suppose $0<\lambda_X$, $\lambda_Y<1$ (the alternative model I), and assume that both $P_{\lambda_X}^X(\mathbf{w})-p_{\mathbf{w}}^X$and $P_{\lambda_Y}^Y(\mathbf{w})-p_{\mathbf{w}}^Y$ are not constant in **w**. *Then, in distribution,*

$$\lim_{n\to\infty}(n_X n_Y)^{\frac{1}{4}}\left(\frac{D_2^S}{\sqrt{n_X n_Y}} - A^{gS}(\lambda_X, \lambda_Y)\right) = Z_{\lambda_X, \lambda_Y}^{gS},$$

where $Z_{\lambda_X, \lambda_Y}^{gS}$ has normal distribution $\mathcal{N}(0, \sqrt{C_{YX}}(\Sigma_{\lambda_X}^{XS})^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^{YS})^2)$.

The proof of Theorem 5.4 is sketched as follows. Similarly as for (14),

$$\frac{D_2^S}{(n_X n_Y)^{\frac{1}{4}}} = \sum_{\mathbf{w}\in\mathcal{A}^k}\frac{(\widetilde{X}_{\mathbf{w}}/\sqrt{n_X})(\widetilde{Y}_{\mathbf{w}}/\sqrt{n_Y})}{\sqrt{\sqrt{C_{XY}}\widetilde{X}_{\mathbf{w}}^2/n_X+\sqrt{C_{YX}}\widetilde{Y}_{\mathbf{w}}^2/n_Y}}.$$

For part (a), under the null hypothesis, we have that, in distribution,

$$\widetilde{X}_{\mathbf{w}}/\sqrt{n_X}\to Z_{\mathbf{w}}^{(g1)}, \quad \widetilde{Y}_{\mathbf{w}}/\sqrt{n_Y}\to Z_{\mathbf{w}}^{(g2)}.$$

For part (b), we can write

$$\frac{\widetilde{X}_{\mathbf{w}}}{n_X} = \frac{X_{\mathbf{w}}}{n_X} - P_{\lambda_X}^X(\mathbf{w}) + P_{\lambda_X}^X(\mathbf{w}) - p_{\mathbf{w}}^X,$$

Then we use Taylor expansion for the function $g_{\mathbf{w}}(x, y)$ given by

$$g_{\mathbf{w}}(x, y) = \frac{(x+P_{\lambda_X}^X(\mathbf{w})-p_{\mathbf{w}}^X)(y+P_{\lambda_Y}^Y(\mathbf{w})-p_{\mathbf{w}}^Y)}{\sqrt{C_{XY}(x+P_{\lambda_X}^X(\mathbf{w})-p_{\mathbf{w}}^X)^2+C_{YX}(y+P_{\lambda_Y}^Y(\mathbf{w})-p_{\mathbf{w}}^Y)^2}},$$

at $(x, y)=(0, 0)$, as well as (14).

From Theorems 5.2, 5.3, and 5.4, we are able to calculate the power of detecting the relationships between sequences **A** and **B** under the general model.

Theorem 5.5. *Assume that $(P_{\lambda_X}^X(\mathbf{w})-p_{\mathbf{w}}^X)^2/p_{\mathbf{w}}^X$, $(P_{\lambda_Y}^Y(\mathbf{w})-p_{\mathbf{w}}^Y)^2/p_{\mathbf{w}}^Y$ and $P_{\lambda_X}^X(\mathbf{w})-p_{\mathbf{w}}^X$ are not constant in* **w**. *Then, for any given type I error α, the power of detecting the relationship between two sequences* **A** *and* **B** *against the null model that $\lambda_X=\lambda_Y=1$ using D_2, D_2^* and D_2^S can be approximated by $1-\Phi(C^g(\lambda_X, \lambda_Y))$, $1-\Phi(C^g(\lambda_X, \lambda_Y))$, $1-\Phi(C^{g*}(\lambda_X, \lambda_Y))$, and $1-\Phi(C^{gS}(\lambda_X, \lambda_Y))$, respectively, where*

$$\begin{aligned}
C^g(\lambda_X, \lambda_Y) &= -(n_X n_Y)^{\frac{1}{4}}B^g(\lambda_X, \lambda_Y)+z_\alpha^g/\sqrt{\sqrt{C_{YX}}(\Sigma_{\lambda_X}^X)^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^Y)^2},\\
C^{g*}(\lambda_X, \lambda_Y) &= -(n_X n_Y)^{\frac{1}{4}}B^{g*}(\lambda_X, \lambda_Y)+z_\alpha^{g*}/((n_X n_Y)^{\frac{1}{4}}\sqrt{\sqrt{C_{YX}}(\Sigma_{\lambda_X}^{X*})^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^{Y*})^2}),\\
C^{gS}(\lambda_X, \lambda_Y) &= -(n_X n_Y)^{\frac{1}{4}}B^{gS}(\lambda_X, \lambda_Y)+z_\alpha^{gS}/\sqrt{\sqrt{C_{YX}}(\Sigma_{\lambda_X}^{XS})^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^{YS})^2}
\end{aligned}$$

and

$$\begin{aligned}
B^g(\lambda_X, \lambda_Y) &= \frac{A^g(\lambda_X, \lambda_Y)-A^g(1, 1)}{\sqrt{\sqrt{C_{YX}}(\Sigma_{\lambda_X}^X)^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^Y)^2}},\\
B^{g*}(\lambda_X, \lambda_Y) &= \frac{A^{g*}(\lambda_X, \lambda_Y)}{\sqrt{\sqrt{C_{YX}}(\Sigma_{\lambda_X}^{X*})^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^{Y*})^2}},\\
B^{gS}(\lambda_X, \lambda_Y) &= \frac{A^{gS}(\lambda_X, \lambda_Y)}{\sqrt{\sqrt{C_{YX}}(\Sigma_{\lambda_X}^{XS})^2+\sqrt{C_{XY}}(\Sigma_{\lambda_Y}^{YS})^2}}.
\end{aligned}$$

Here, $z_\alpha^g, z_\alpha^{g*}$, and z_α^{gS} are the upper α quantile of Z_1^g, Z_1^{g*}, Z_1^{gS} from Theorems 5.2, 5.3, and 5.4, respectively.

The alternative model II can equally be extended to the situation of different letter frequencies in the two sequences; we omit the details here.

ACKNOWLEDGMENTS

L.W. was supported by NIH grant no. P50 HG 002790 and by NIH grant no. R21AG032743. G.R. was supported in part by EPSRC grant no. GR/R52183/01, and by BBSRC and EPSRC through OCISB. F.S. was supported by NIH grants no. P50 HG 002790 and R21AG032743 and NSFC grants 60928007 and 60805010. M.S.W. was supported by NIH grant no. P50 HG 002790 and by NIH grant no. R21AG032743.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Burden, C.J., Kantorovitz, M.R., and Wilson, S.R. 2006. Approximate word matches between two random sequences. *Ann. Appl. Probab.* 18:1–21.

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

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

Forêt, S., Wilson, S.R., and Burden, C.J. 2009b. Characterizing the D2 statistic: word matches in biological sequences. *Stat. Appl. Genet. Mol. Biol.* 8:43.

Ivan, A., Halfon, M.S., and Sinha, S. 2008. Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. *Genome Biol.* 9:R22.

Kantorovitz, M.R., Booth, H.S., Burden, C.J., and Wilson, S.R. 2007a. Asymptotic behavior of *k*-word matches between two uniformly distributed sequences. *J. Appl. Probab.* 44:788–805.

Kantorovitz, M.R., Robinson, G.E., and Sinha, S. 2007b. A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics* 23:i249–i255.

Lippert, R.A., Huang, H.Y., and Waterman, M.S. 2002. Distributional regimes for the number of *k*-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA*, 100:13980–13989.

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., et al. 1994. Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* 73:3169–3172.

Novak, S.Y. 2007. A new characterization of the normal law. *Stat. Probabil. Lett.* 77:95–98.

Reinert, G., Chew, D., Sun, F.Z., et al. 2009. Alignment-free sequence comparison (I): Statistics and power. *J. Comput. Biol.* 16:1615–1634.

Rabiner, L.R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* 77:257–286.

Stormo, G.D. 2000. DNA binding sites: representation and discovery. *Bioinformatics* 16:16–23.

Sandelin, A., Alkema, W., Engström, P., et al. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32:D91–D94.

Zhai, Z.Y., Ku, S.Y., Luan, Y.H., et al. 2010. The power of detecting enriched patterns: an HMM approach. *J. Comput. Biol.* 17:581–592.

Address correspondence to:
Dr. Michael S. Waterman
Molecular and Computational Biology
University of Southern California
1050 Childs Way, RRI 201
Los Angeles, CA 90089-2910

E-mail: msw@usc.edu