J. Appl. Prob. **44,** 788–805 (2007)
Printed in England
© *Applied Probability Trust* 2007

ASYMPTOTIC BEHAVIOR OF k-WORD MATCHES BETWEEN TWO UNIFORMLY DISTRIBUTED SEQUENCES

M. R. KANTOROVITZ,* *Australian National University and University of Illinois*
H. S. BOOTH,**
C. J. BURDEN,*** **** *Australian National University*
S. R. WILSON,*** ***** *Australian National University*

Abstract

Given two sequences of length n over a finite alphabet $\mathcal{A}$ of size $|\mathcal{A}| = d$, the D_2 statistic is the number of k-letter word matches between the two sequences. This statistic is used in bioinformatics for EST sequence database searches. Under the assumption of independent and identically distributed letters in the sequences, Lippert, Huang and Waterman (2002) raised questions about the asymptotic behavior of D_2 when the alphabet is uniformly distributed. They expressed a concern that the commonly assumed normality may create errors in estimating significance. In this paper we answer those questions. Using Stein's method, we show that, for large enough k, the D_2 statistic is approximately normal as n gets large. When $k = 1$, we prove that, for large enough d, the D_2 statistic is approximately normal as n gets large. We also give a formula for the variance of D_2 in the uniform case.

Keywords: Stein's method; count vector; k-word matches; sequence comparison

2000 Mathematics Subject Classification: Primary 62E20; 92D20; 60F99

1. Introduction

Methods for alignment-free sequence comparison are among the more recent tools being developed for sequence analysis in biology [14]. A disadvantage in the classical Smith–Waterman local alignment algorithm [11], which is implemented in search algorithms such as FASTA and BLAST, is that it assumes conservation of contiguity between homologous segments. In particular, it overlooks the occurrence of genetic shuffling [16]. Alignment-free sequence comparison methods are used to compensate for this problem.

A natural alignment-free comparison of two sequences is the number of k-letter word matches between the sequences. This statistic is referred to as D_2 in [9]. It can be computed in linear time in the length of the sequences, which is also an advantage over the nonlinear local alignment algorithms. The D_2 statistic is used extensively for EST sequence database searches; see, e.g. [3], [4], [10], and in the software package STACK [6].

Received 28 September 2005; revision received 20 July 2006.

* Postal address: Department of Mathematics, University of Illinois, Urbana, IL 61801, USA.
Email address: ruth@math.uiuc.edu
** H. S. Booth died 26 May 2005.
*** Postal address: Mathematical Sciences Institute, Australian National University, Canberra, ACT 0200, Australia.
**** Email address: conrad.burden@anu.edu.au
***** Email address: sue.wilson@anu.edu.au

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

In [9], Lippert *et al.* started a rigorous study of D_2 using the model of independent letters in DNA sequences. A formula for the expectation was computed as well as upper and lower bounds for the variance. Limiting distributions, as the length of the sequences, n, and the size of the word, k, get large, were derived in some cases. Lippert *et al.* used Stein–Chen methods (see [5] and [12]) to obtain the following result. When $k/\log_c n > 2$, D_2 has a compound Poisson asymptotic behavior. The logarithmic base c is defined by $c = (\sum_{a\in\mathcal{A}} f_a^2)^{-1}$, where f_a is the probability of a letter taking the value a. As pointed out in [1] and [15], the compound Poisson approximation is meaningful in this region only when $\mathrm{E}(D_2)$ is not too small. To control this degenerate case, the linear restriction $k = 2\log_c n + C$ was added.

Another asymptotic regime was identified in [9] under the assumption that the underlying distribution of the alphabet is *nonuniform*. In this case, Lippert *et al.* proved that D_2 has a normal asymptotic behavior when $k/\log_c n < \frac{1}{6}$. However, their method of proof breaks down in the *uniform case*. Lippert *et al.* [9] gave an example showing that in the degenerate uniform case, when $k = 1$ and the size of the alphabet is $d = 2$, D_2 is not asymptoticly normal as $n \to \infty$. They suggested that a limiting normal distribution may not always occur when k is small and the letters are uniformly distributed. They also raised the concern that commonly assumed normality may create errors in estimating significance.

Following results from simulations, the following two conjectures were made in [9] regarding the uniform case.

Conjecture 1. *When $k = 1$, D_2 is approximately normal for appropriately large enough d and n.*

Conjecture 2. *For large enough k, D_2 should be approximately normal as $n \to \infty$. Simulations in [9] for $d = 4$ suggested that, when $k \geq 2$ and $n > 2^{k-3} \times 100$, a good normal approximation already occurs.*

In this paper we address Conjectures 1 and 2. When $k = 1$, the following theorem says that, for large enough d, the standardized statistic $(D_2 - \mathrm{E}(D_2))/\sigma(D_2)$ is approximately normal as n gets large.

Theorem 1. *For $k = 1$,*

$$\lim_{d\to\infty}\lim_{n\to\infty}\left|\Pr\left(\frac{D_2 - \mathrm{E}(D_2)}{\sigma(D_2)} \leq x\right) - \Phi(x)\right| = 0,$$

where Φ is the standard normal distribution function.

Theorem 2 states that, for large enough k, the standardized statistic $(D_2 - \mathrm{E}(D_2))/\sigma(D_2)$ is approximately normal as n gets large. Simulations in [9] (see Section 4, below) show that, for $d = 4$, normal behavior already occurs when $k = 2$ and n is in the hundreds. The proof of the following theorem uses Stein's method.

Theorem 2. *We have*

$$\lim_{k\to\infty}\lim_{n\to\infty}\left|\Pr\left(\frac{D_2 - \mathrm{E}(D_2)}{\sigma(D_2)} \leq x\right) - \Phi(x)\right| = 0.$$

We give a formula for the variance of D_2 in the uniform case in the following result.

Theorem 3. *We have*

$$\operatorname{var}(D_2(n)) = \bar{n}^2\left[\left(\frac{1}{d}\right)^k - \left(\frac{1}{d}\right)^{2k}\right] + 2\bar{n}^2\left[\frac{(1/d)^{k+1}(1-(1/d)^{k-1})}{1-(1/d)} - \frac{k-1}{d^{2k}}\right], \quad (1)$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

where $\bar{n} = n - k + 1$*. In particular, when* $k = 1$*,*

$$\operatorname{var}(D_2(n)) = \frac{n^2(d-1)}{d^2}.$$

The organization of this paper is as follows. Section 2 is devoted to preliminaries. In Section 3 we prove Theorem 1 and Theorem 2. In Section 4 we briefly discuss simulations. In Section 5 we derive a formula for the variance of D_2 in the uniform case (Theorem 3).

2. Preliminaries

We follow the notation and terminology in [9]. Let $\boldsymbol{A} = A_1A_2\cdots A_n$ and $\boldsymbol{B} = B_1B_2\cdots B_n$ be two sequences with independent and identically distributed (i.i.d.) letters. The letters are taken from a finite set of alphabet $\mathcal{A}$ of size $d = |\mathcal{A}|$.

The $D_2 = D_2(n,k)$ statistic is defined to be the number of k-letter word (abbreviated as 'k-word') matches (including overlaps) between the two sequences $\boldsymbol{A}$ and $\boldsymbol{B}$. One way to compute this statistic is

$$D_2 = \sum_{(i,j)\in I} Y_{(i,j)},$$

where $Y_{(i,j)}$ is the k-word match indicator (starting) at position $(i,\ j)$ (position i in sequence $\boldsymbol{A}$ and j in $\boldsymbol{B}$). The index set I is

$$I = \{(i,\ j) \in \mathbb{N}\times\mathbb{N} \colon 1 \le i \le n-k+1,\ 1 \le j \le n-k+1\}.$$

For convenience, we write $\bar{n}$ for $n-k+1$.

The mean of $D_2(n)$ is easily computed from the above expression as follows. For $a \in \mathcal{A}$, write f_a for the probability of a letter in the sequence taking the value a. Then

$$\mathrm{E}(Y_{(i,j)}) = \Pr(Y_{(i,j)} = 1) = \biggl(\sum_{a\in\mathcal{A}} f_a^2\biggr)^k \tag{2}$$

and

$$\mathrm{E}(D_2(n)) = \sum_{(i,j)\in I} \mathrm{E}(Y_{(i,j)}) = \bar{n}^2\biggl(\sum_{a\in\mathcal{A}} f_a^2\biggr)^k.$$

When the alphabet is uniformly distributed, i.e. $f_a = 1/d$ for all $a \in \mathcal{A}$, we have

$$\mathrm{E}(D_2(n)) = \frac{\bar{n}^2}{d^k}. \tag{3}$$

For the variance, upper and lower bounds were given in [9].

Another way to think of D_2 is as the inner product of the vectors of word counts. More explicitly, let $\mathcal{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{d^k}\}$ be the set of all k-words on the alphabet $\mathcal{A}$. For $\boldsymbol{w} \in \mathcal{W}$, let $N_{\boldsymbol{w}}^{\boldsymbol{A}} = N_{\boldsymbol{w}}^{\boldsymbol{A}}(n)$ be the number of times the word $\boldsymbol{w}$ appears in the sequence $\boldsymbol{A}$ (overlaps allowed). Then $\boldsymbol{N}^{\boldsymbol{A}}(n) = (N_{\boldsymbol{w}_1}^{\boldsymbol{A}}(n), \ldots, N_{\boldsymbol{w}_{d^k}}^{\boldsymbol{A}}(n))$ is the count vector for the sequence $\boldsymbol{A}$. Similarly, define the count vector for the sequence $\boldsymbol{B}$ as $\boldsymbol{N}^{\boldsymbol{B}}(n) = (N_{\boldsymbol{w}_1}^{\boldsymbol{B}}(n), \ldots, N_{\boldsymbol{w}_{d^k}}^{\boldsymbol{B}}(n))$. Then we obtain

$$D_2(n) = \langle \boldsymbol{N}^{\boldsymbol{A}}(n), \boldsymbol{N}^{\boldsymbol{B}}(n)\rangle = \sum_{\boldsymbol{w}\in\mathcal{W}} N_{\boldsymbol{w}}^{\boldsymbol{A}}(n) N_{\boldsymbol{w}}^{\boldsymbol{B}}(n).$$

The following central limit theorem is known for the count vector.

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

Theorem 4. ([15, Theorem 12.5].) *Let $\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m\}$ be a set of words on a given alphabet $\mathcal{A}$. Let $\boldsymbol{N}(n) = (N_{\boldsymbol{w}_1}(n), \ldots, N_{\boldsymbol{w}_m}(n))$ be the count vector for $\boldsymbol{W}$ in a random sequence of length n. Then, $n^{-1/2}\boldsymbol{N}(n)$ is asymptotically normal with mean $n^{1/2}\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{\mu}$ is the limiting mean vector*

$$\boldsymbol{\mu} = \lim_{n\to\infty} n^{-1}(\mathrm{E}(N_{\boldsymbol{w}_1}(n)), \ldots, \mathrm{E}(N_{\boldsymbol{w}_m}(n)))$$

and $\boldsymbol{\Sigma}$ is the limiting covariance matrix with elements

$$\sigma_{i,j} = \sigma_{\boldsymbol{w}_i,\boldsymbol{w}_j} = \lim_{n\to\infty} n^{-1}\operatorname{cov}(N_{\boldsymbol{w}_i}(n), N_{\boldsymbol{w}_j}(n)) \quad \textit{for } 1 \le i,\ j \le m.$$

A formula for the covariance matrix is given in [15, Chapter 12]. Here we summarize the results when applied to our model of i.i.d. letters and for words of the same length. First we need the following notation. Let $\boldsymbol{A} = A_1A_2\cdots A_n$ be a sequence of i.i.d. letters. Let $\boldsymbol{u} = (u_1, \ldots, u_k)$ and $\boldsymbol{v} = (v_1, \ldots, v_k)$ be two words of length k. We write $\pi_{\boldsymbol{u}}$ for the probability of seeing $\boldsymbol{u}$. In the notation of (2), we have

$$\pi_{\boldsymbol{u}} = \prod_{i=1}^{k} f_{u_i}.$$

Note that when the alphabet is uniformly distributed,

$$\pi_{\boldsymbol{u}} = \frac{1}{d^k}. \tag{4}$$

Next, we define the overlap indicator

$$\beta_{\boldsymbol{u},\boldsymbol{v}}(j) = \begin{cases} 1 & \text{if } u_{j+1} = v_1, \ldots, u_k = v_{k-j}, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

That is, $\beta_{\boldsymbol{u},\boldsymbol{v}}(j) = 1$ if the last $k - j$ letters of $\boldsymbol{u}$ match the first $k - j$ letters of $\boldsymbol{v}$.

We define an indicator that a word $\boldsymbol{u}$ occurs starting at position i in the sequence $\boldsymbol{A}$ by

$$\mathfrak{I}_{\boldsymbol{u}}(i) = \begin{cases} 1 & \text{if } A_i = u_1, \ldots, A_{i+k-1} = u_k, \\ 0 & \text{otherwise.} \end{cases}$$

Note that $N_{\boldsymbol{u}}(n) = \sum_{j=1}^{\bar{n}} \mathfrak{I}_{\boldsymbol{u}}(j)$. Hence,

$$\mathrm{E}(N_{\boldsymbol{u}}(n)) = \bar{n}\pi_{\boldsymbol{u}}. \tag{6}$$

Finally, denote by $\mathrm{P}_{\boldsymbol{u}}(j)$ the probability of seeing the subword made out of the last j letters of $\boldsymbol{u}$, i.e.

$$P_{\boldsymbol{u}}(j) = \begin{cases} \prod_{i=k-j+1}^{k} f_{u_i} & \text{for } 0 < j < k, \\ 1 & \text{otherwise.} \end{cases}$$

We are now ready to state the formula for the covariance matrix.

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

Proposition 1. ([15, Corollary 12.1].) *Let $\boldsymbol{A} = A_1A_2\cdots A_n$ be a sequence of i.i.d. letters. Let $\boldsymbol{u} = (u_1, \dots, u_k)$ and $\boldsymbol{v} = (v_1, \dots, v_k)$ be two words of length k. Write $N_{\boldsymbol{u}}(n)$ and $N_{\boldsymbol{v}}(n)$ for the count of $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively, in the sequence $\boldsymbol{A}$. Then the elements of the limiting covariance matrix $\boldsymbol{\Sigma} = (\sigma_{\boldsymbol{u},\boldsymbol{v}})$ are given by*

$$\begin{aligned}\sigma_{\boldsymbol{u},\boldsymbol{v}} &= \lim_{n\to\infty} n^{-1}\operatorname{cov}(N_{\boldsymbol{u}}(n), N_{\boldsymbol{v}}(n)) \\ &= \pi_{\boldsymbol{u}}\sum_{j=0}^{k-1}\beta_{\boldsymbol{u},\boldsymbol{v}}(j)\,\mathrm{P}_{\boldsymbol{v}}(j) + \pi_{\boldsymbol{v}}\sum_{j=0}^{k-1}\beta_{\boldsymbol{v},\boldsymbol{u}}(j)\,\mathrm{P}_{\boldsymbol{u}}(j) - \pi_{\boldsymbol{u}}\pi_{\boldsymbol{v}}(2k-1) - \pi_{\boldsymbol{u}}\beta_{\boldsymbol{v},\boldsymbol{u}}(0). \qquad (7)\end{aligned}$$

Remark 1. From (6), the limiting mean is

$$\boldsymbol{\mu} = \lim_{n\to\infty}\frac{\bar{n}(\pi_{\boldsymbol{w}_1}, \dots, \pi_{\boldsymbol{w}_{d^k}})}{n} = (\pi_{\boldsymbol{w}_1}, \dots, \pi_{\boldsymbol{w}_{d^k}}).$$

In the uniform case, by (4), we have

$$\boldsymbol{\mu} = \frac{1}{d^k}(1, 1, \dots, 1). \qquad (8)$$

3. Asymptotic behavior

For the rest of this paper we assume that the underlying distribution of the alphabet is uniform.

In this section we prove our main results. In Section 3.1 we prove Theorem 1 and in Section 3.2 we prove Theorem 2. We start with a few observations.

From the discussion in Section 2, we have

$$D_2(n) = \langle \boldsymbol{N}^{\boldsymbol{A}}(n), \boldsymbol{N}^{\boldsymbol{B}}(n)\rangle = \sum_{i=1}^{d^k} N_i^{\boldsymbol{A}}(n)N_i^{\boldsymbol{B}}(n).$$

By Theorem 4, we have convergence in distribution, i.e.

$$n^{-1/2}\boldsymbol{N}^{\boldsymbol{A}}(n) - n^{1/2}\boldsymbol{\mu} \xrightarrow{\mathrm{D}} \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}^{\boldsymbol{A}}$$

as $n \to \infty$, and the same for B, where $\boldsymbol{Z}^{\boldsymbol{A}}$ and $\boldsymbol{Z}^{\boldsymbol{B}}$ are independent multivariate standard normal random vectors. Hence,

$$\langle n^{-1/2}\boldsymbol{N}^{\boldsymbol{A}}(n) - n^{1/2}\boldsymbol{\mu}, n^{-1/2}\boldsymbol{N}^{\boldsymbol{B}}(n) - n^{1/2}\boldsymbol{\mu}\rangle \xrightarrow{\mathrm{D}} \langle \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}^{\boldsymbol{B}}\rangle.$$

Expanding the left-hand side, we obtain

$$n^{-1}D_2(n) - \langle \boldsymbol{\mu}, \boldsymbol{N}^{\boldsymbol{B}}(n)\rangle - \langle \boldsymbol{N}^{\boldsymbol{A}}(n), \boldsymbol{\mu}\rangle + n||\boldsymbol{\mu}||^2 \xrightarrow{\mathrm{D}} \langle \boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle. \qquad (9)$$

Remark 2. For each row of the limiting covariance matrix $\boldsymbol{\Sigma}$, the sum of the entries equals 0. To see this, note that in the count vector $\boldsymbol{N}(n)$,

$$\sum_{j=1}^{d^k} N_j(n) = \bar{n}. \qquad (10)$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

Hence, for each fixed i,

$$\begin{aligned}\sum_j \sigma_{ij} &= \lim_{n\to\infty} n^{-1} \sum_{j=1}^{d^k} \operatorname{cov}(N_i(n), N_j(n)) \\ &= \lim_{n\to\infty} n^{-1} \operatorname{cov}\biggl(N_i(n), \sum_{j=1}^{d^k} N_j(n)\biggr) \\ &= \lim_{n\to\infty} n^{-1} \operatorname{cov}(N_i(n), \bar{n}) \\ &= 0.\end{aligned}$$

Proposition 2. *For fixed k and d,*

$$\frac{D_2(n) - \mathrm{E}(D_2(n))}{\sigma(D_2(n))} \xrightarrow{\mathrm{D}} \frac{\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle}{\sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)}$$

as $n \to \infty$.

Proof. Since

$$\frac{D_2(n) - \mathrm{E}(D_2(n))}{\sigma(D_2(n))} = \frac{n^{-1}D_2(n) - \mathrm{E}(n^{-1}D_2(n))}{\sigma(n^{-1}D_2(n))},$$

it is enough to show that

$$n^{-1}D_2(n) - \mathrm{E}(n^{-1}D_2(n)) \xrightarrow{\mathrm{D}} \langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle. \tag{11}$$

Let $e(n,k) = -\langle \boldsymbol{\mu}, \boldsymbol{N}^{\boldsymbol{B}}(n)\rangle - \langle \boldsymbol{N}^{\boldsymbol{A}}(n), \boldsymbol{\mu}\rangle + n||\boldsymbol{\mu}||^2$ be the 'correcting term' on the left-hand side of (9). By (8) and (10), we have

$$\langle \boldsymbol{N}^{\boldsymbol{A}}(n), \boldsymbol{\mu}\rangle = d^{-k} \sum_{j=1}^{d^k} N_j^{\boldsymbol{A}}(n) = \bar{n} d^{-k},$$

and the same for B. By (8), we obtain

$$||\boldsymbol{\mu}||^2 = \frac{d^k}{d^{2k}} = d^{-k}.$$

Hence,

$$e(n,k) = -2\bar{n}d^{-k} + nd^{-k} = -(n-2k+2)d^{-k}. \tag{12}$$

By (3), applied to the uniform case, we have

$$-\mathrm{E}(n^{-1}D_2(n)) = -\frac{\bar{n}^2}{nd^k}. \tag{13}$$

The case in which $\boldsymbol{k} = \boldsymbol{1}$ is straightforward. Here, the correcting term in (12) is $e(n,1) = -n/d$. Conversely, when $k=1$, (13) becomes

$$-\mathrm{E}(n^{-1}D_2(n)) = -\frac{n^2}{nd} = -\frac{n}{d},$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

which is precisely $e(n, 1)$. Therefore, by (9),

$$n^{-1}D_2(n) - \mathrm{E}(n^{-1}D_2(n)) = n^{-1}D_2(n) + e(n, 1) \xrightarrow{\mathrm{D}} \langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B} \rangle$$

as $n \to \infty$.

For the general case, from (12) and (13) we obtain

$$-\mathrm{E}(n^{-1}D_2(n)) = -\frac{\overline{n}^2}{nd^k} = \frac{-n + 2k - 2}{d^k} - \frac{(k-1)^2}{nd^k} = e(n, k) - \frac{(k-1)^2}{nd^k}.$$

Hence,

$$n^{-1}D_2(n) - \mathrm{E}(n^{-1}D_2(n)) = n^{-1}D_2(n) + e(n, k) - \frac{(k-1)^2}{nd^k}.$$

Since $(k-1)^2/nd^k \to 0$ as $n \to \infty$, (11) holds.

We now look at the variance of $\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B} \rangle$.

Lemma 1. *We have*

$$\operatorname{var}(\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B} \rangle) = \sum_{i,j} \sigma_{ij}^2.$$

Proof. First note that $\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B} \rangle = \sum_{i,j} \sigma_{ij} Z_j^{\boldsymbol{A}} Z_i^{\boldsymbol{B}}$. A direct computation gives

$$\operatorname{cov}(Z_j^{\boldsymbol{A}} Z_i^{\boldsymbol{B}}, Z_t^{\boldsymbol{A}} Z_s^{\boldsymbol{B}}) = \begin{cases} 0 & \text{if } (i, j) \neq (s, t), \\ 1 & \text{if } (i, j) = (s, t). \end{cases} \tag{14}$$

Therefore,

$$\operatorname{var}(\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B} \rangle) = \sum_{i,j} \operatorname{var}(\sigma_{ij} Z_j^{\boldsymbol{A}} Z_i^{\boldsymbol{B}}) = \sum_{i,j} \sigma_{ij}^2.$$

3.1. The case in which $k = 1$

We begin by computing the limiting covariance matrix $\boldsymbol{\Sigma}$ using Proposition 1. Since $\pi_{\boldsymbol{u}} = \pi_{\boldsymbol{v}} = 1/d$ by (4), the limiting covariance matrix is the $d \times d$ matrix

$$\boldsymbol{\Sigma} = \frac{1}{d} \begin{pmatrix} 1 - 1/d & -1/d & \dots & -1/d \\ -1/d & 1 - 1/d & \dots & -1/d \\ \vdots & \vdots & \ddots & \vdots \\ -1/d & -1/d & \dots & 1 - 1/d \end{pmatrix} = \frac{1}{d}\boldsymbol{I} - \frac{1}{d^2} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}, \tag{15}$$

where $\boldsymbol{I}$ is the $d \times d$ identity matrix. By (8), the limiting mean is

$$\boldsymbol{\mu} = \frac{1}{d}(1, \dots, 1).$$

Lemma 2. *For* $k = 1$,

$$\operatorname{var}(\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B} \rangle) = \frac{d-1}{d^2}.$$

Proof. By Lemma 1,

$$
\begin{aligned}
\operatorname{var}(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle) &= \sum_{i,j} \sigma_{ij}^2 \\
&= \frac{1}{d^2}\Bigg(\sum_{i=1}^{d}\Bigg[\bigg(1-\frac{1}{d}\bigg)^2 + (d-1)\bigg(\frac{1}{d}\bigg)^2\Bigg]\Bigg) \qquad \text{(by (15))} \\
&= \frac{d-1}{d^2}.
\end{aligned}
$$

Lemma 3. *For $k = 1$, as $d \to \infty$, the asymptotic distributions of*

$$
\frac{\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle)} \quad \textit{and} \quad \frac{\langle (1/d) \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle)}
$$

are the same, provided that they exist.

Proof. It is enough to show that the variance of the difference of the two statistics approaches 0 as $d \to \infty$. By (15), $(\boldsymbol{\Sigma} - (1/d)\boldsymbol{I})_{ij} = -1/d^2$. Hence,

$$
\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle - \left\langle \frac{1}{d} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \right\rangle = -\frac{1}{d^2} \sum_{i,j} Z_j^A Z_i^B
$$

and

$$
\begin{aligned}
\operatorname{var}\left(\frac{\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle - \langle (1/d) \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle)}\right) &= \frac{\operatorname{var}(-(1/d^2) \sum_{i,j} Z_j^A Z_i^B)}{\operatorname{var}(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle)} \\
&= \frac{(1/d^4) \sum_{i,j} 1}{(d-1)/d^2} \qquad \text{(by (14) and Lemma 2)} \\
&= \frac{d^2/d^4}{(d-1)/d^2} \\
&= \frac{1}{d-1} \\
&\to 0 \quad \text{as } d \to \infty.
\end{aligned}
$$

Proof of Theorem 2. By Proposition 2, it is enough to show that, as $d \to \infty$,

$$
\frac{\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle)} \xrightarrow{\mathrm{D}} \mathcal{N}(0, 1).
$$

By Lemmas 2 and 3,

$$
\frac{\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle)} \quad \text{and} \quad \frac{\langle (1/d) \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sqrt{(d-1)}/d}
$$

have the same asymptotic behavior as $d \to \infty$. Now,

$$
\frac{\langle (1/d) \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}} \rangle}{\sqrt{(d-1)}/d} = \frac{(1/d) \sum_{i=1}^{d} Z_i^A Z_i^B}{\sqrt{(d-1)}/d} = \frac{\sum_{i=1}^{d} Z_i^A Z_i^B}{\sqrt{d-1}},
$$

where the summands, $\{Z_i^{\boldsymbol{A}} Z_i^{\boldsymbol{B}}\}_{i=1,\dots,d}$, are i.i.d. with mean 0 and variance 1. Hence, by the central limit theorem we obtain

$$\frac{\sum_{i=1}^{d} Z_i^{\boldsymbol{A}} Z_i^{\boldsymbol{B}}}{\sqrt{d-1}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

3.2. The general case

In this section we show that, for large enough k, the standardized statistic D_2 is approximately normal as n gets large (Theorem 2). Simulations in [9] (see Section 4, below) show that, for $d = 4$, normal behavior already occurs when $k = 2$ and n is in the hundreds.

To understand the limiting covariance matrix we need the following lemma.

Lemma 4. *Let* $\boldsymbol{u}$ *be a k-word. Then the following results hold.*

(i) *We have*

$$\sigma_{\boldsymbol{u},\boldsymbol{u}} = \frac{1}{d^k} - \frac{2k-1}{d^{2k}} + O\left(\frac{1}{d^{k+j}}\right)$$

for some $1 \le j \le k-1$.

(ii) *For a k-word* $\boldsymbol{v} \ne \boldsymbol{u}$, $\sigma_{\boldsymbol{u},\boldsymbol{v}} = -(2k-1)/d^{2k} + O(1/d^{k+j})$ *for some* $1 \le j \le k-1$. *Moreover, for each* $j = 1, 2, \dots, k-1$ *and for a given* $\boldsymbol{u}$, *there are at most* $2d^j$ *k-words,* $\boldsymbol{v}$, *with*

$$\sigma_{\boldsymbol{u},\boldsymbol{v}} = \frac{1}{d^{k+j}} - \frac{2k-1}{d^{2k}} + O\left(\frac{1}{d^{k+j}}\right).$$

When $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *have no overlaps (i.e. when the overlap indicator* $\beta_{\boldsymbol{u},\boldsymbol{v}}(j)$, *defined in (5), equals* 0 *for all* j*), then* $\sigma_{\boldsymbol{u},\boldsymbol{v}} = -(2k-1)/d^{2k}$.

Proof. We examine the terms in (7). Since the alphabet is uniformly distributed and the letters in the sequences are assumed to be i.i.d., we have

$$\pi_{\boldsymbol{u}} = \pi_{\boldsymbol{v}} = \frac{1}{d^k},$$
$$\mathrm{P}_{\boldsymbol{u}}(j) = \mathrm{P}_{\boldsymbol{v}}(j) = \frac{1}{d^j}.$$

When $\boldsymbol{u} = \boldsymbol{v}$, $\beta_{\boldsymbol{v},\boldsymbol{u}}(0) = 1$; hence,

$$\begin{aligned}\sigma_{\boldsymbol{u},\boldsymbol{u}} &= \frac{1}{d^k}\sum_{j=0}^{k-1}\beta_{\boldsymbol{u},\boldsymbol{v}}(j)\frac{1}{d^j} + \frac{1}{d^k}\sum_{j=0}^{k-1}\beta_{\boldsymbol{v},\boldsymbol{u}}(j)\frac{1}{d^j} - \frac{2k-1}{d^{2k}} - \frac{1}{d^k}\\ &= \frac{1}{d^k} + \frac{1}{d^k}\sum_{j=1}^{k-1}\beta_{\boldsymbol{u},\boldsymbol{v}}(j)\frac{1}{d^j} + \frac{1}{d^k}\sum_{j=1}^{k-1}\beta_{\boldsymbol{v},\boldsymbol{u}}(j)\frac{1}{d^j} - \frac{2k-1}{d^{2k}}\\ &= \frac{1}{d^k} - \frac{2k-1}{d^{2k}} + O\left(\frac{1}{d^{k+j}}\right), \end{aligned} \tag{16}$$

where $j = \min\{j\colon 1 \le j \le k-1, \text{ and } \beta_{\boldsymbol{u},\boldsymbol{v}}(j) = 1 \text{ or } \beta_{\boldsymbol{v},\boldsymbol{u}}(j) = 1\}$.

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

When $\boldsymbol{u} \neq \boldsymbol{v}$, $\beta_{\boldsymbol{v},\boldsymbol{u}}(0) = \beta_{\boldsymbol{u},\boldsymbol{v}}(0) = 0$; hence,

$$\begin{aligned}\sigma_{\boldsymbol{u},\boldsymbol{v}} &= \frac{1}{d^k}\sum_{j=1}^{k-1}\beta_{\boldsymbol{u},\boldsymbol{v}}(j)\frac{1}{d^j} + \frac{1}{d^k}\sum_{j=1}^{k-1}\beta_{\boldsymbol{v},\boldsymbol{u}}(j)\frac{1}{d^j} - \frac{2k-1}{d^{2k}} \\ &= \frac{1}{d^{k+j}} - \frac{2k-1}{d^{2k}} + O\biggl(\frac{1}{d^{k+j}}\biggr),\end{aligned}$$

where $j = \min\{j : 1 \le j \le k-1, \text{ and } \beta_{\boldsymbol{u},\boldsymbol{v}}(j) = 1 \text{ or } \beta_{\boldsymbol{v},\boldsymbol{u}}(j) = 1\}$.

Note that, for each j and a given $\boldsymbol{u}$, there are d^j possible k-words $\boldsymbol{v}$ for which $\beta_{\boldsymbol{u},\boldsymbol{v}}(j) = 1$, since $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{k-j}$ are determined by the overlap with the last $k-j$ letters of $\boldsymbol{u}$, and there are d^j choices for the last j letters of $\boldsymbol{v}$. Repeating the argument for $\beta_{\boldsymbol{v},\boldsymbol{u}}(j)$, we find that the number of k-words $\boldsymbol{v}$ for which $j = \min\{j : 1 \le j \le k-1, \text{ and } \beta_{\boldsymbol{u},\boldsymbol{v}}(j) = 1 \text{ or } \beta_{\boldsymbol{v},\boldsymbol{u}}(j) = 1\}$ is at most $2d^j$. This completes the proof.

We want to show that $\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle / \sigma(\langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle) \xrightarrow{\text{D}} \mathcal{N}(0,1)$ as $k \to \infty$. It is convenient to rescale by a factor of d^k. That is, our aim is to show that

$$\frac{\langle d^k \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle}{\sigma(d^k \langle \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)} \xrightarrow{\text{D}} \mathcal{N}(0,1) \quad \text{as } k \to \infty.$$

Lemma 5. *We have*

$$\operatorname{var}(\langle d^k \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle) \ge d^k - 4k.$$

Proof. We obtain

$$\begin{aligned}\operatorname{var}(\langle d^k \boldsymbol{\Sigma} \boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle) &= d^{2k}\sum_{i,j}\sigma_{ij}^2 \qquad \text{(by Lemma 1)} \\ &\ge d^{2k}\sum_i \sigma_{ii}^2 \\ &\ge d^{2k}\sum_i\biggl(\frac{1}{d^k} - \frac{2k-1}{d^{2k}}\biggr)^2 \qquad \text{(by (16))} \\ &\ge d^{2k}\sum_i\biggl(\frac{1}{d^{2k}} - \frac{4k}{d^{3k}}\biggr) \\ &= d^{3k}\biggl(\frac{1}{d^{2k}} - \frac{4k}{d^{3k}}\biggr) \\ &= d^k - 4k.\end{aligned}$$

Construction 1. For $m < k$, we decompose the limiting covariance matrix as follows. Let

$$d^k \boldsymbol{\Sigma} = \boldsymbol{T}(m) + \boldsymbol{R}(m),$$

where, using Lemma 4,

$$\boldsymbol{T}(m)_{\boldsymbol{u},\boldsymbol{v}} = \begin{cases} 0 & \text{if } \sigma_{\boldsymbol{u},\boldsymbol{v}} = -(2k-1)/d^{2k}, \text{ i.e. no overlaps,} \\ 0 & \text{if } \sigma_{\boldsymbol{u},\boldsymbol{v}} = 1/d^{j+k} - (2k-1)/d^{2k} + O(1/d^{j+k}) \text{ with } m \le j \le k-1, \\ d^k \sigma_{\boldsymbol{u},\boldsymbol{v}} & \text{otherwise,} \end{cases}$$

and $R(m) = d^k \boldsymbol{\Sigma} - T(m)$.

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

This means that the diagonal terms of $T(m)$ are $T(m)_{\boldsymbol{u},\boldsymbol{u}} = d^k\sigma_{\boldsymbol{u},\boldsymbol{u}}$ and all the nonzero off-diagonal terms are of the form $1/d^j - (2k-1)/d^k + O(1/d^j)$, with $j < m$. All the terms of the remainder, $R(m)$, are $O(1/d^m)$.

We now have $\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle = \langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle + \langle R(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle$. Next we show that the contribution of $\langle R(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle$ to $\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle$ can be made as small as we want, as $k \to \infty$, by picking a large enough m.

Lemma 6. *We have*

$$\frac{\operatorname{var}(\langle R(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)}{\operatorname{var}(\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)} = O\left(\frac{1}{d^m}\right).$$

Proof. As in Lemma 1, we obtain

$$\operatorname{var}(\langle R(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle) = \operatorname{var}\left(\sum_{\boldsymbol{u},\boldsymbol{v}} R(m)_{\boldsymbol{u}\boldsymbol{v}} Z^{A}_{\boldsymbol{v}} Z^{\boldsymbol{B}}_{\boldsymbol{u}}\right) = \sum_{\boldsymbol{u},\boldsymbol{v}} (R(m)_{\boldsymbol{u}\boldsymbol{v}})^2.$$

By Construction 1 and Lemma 4, we have

$$\begin{aligned}\sum_{\boldsymbol{u},\boldsymbol{v}} (R(m)_{\boldsymbol{u}\boldsymbol{v}})^2 &= \sum_{\boldsymbol{u}}\left(\sum_{\boldsymbol{v}} (R(m)_{\boldsymbol{u}\boldsymbol{v}})^2\right)\\ &= \sum_{\boldsymbol{u}}\left(\sum_{t=m}^{k-1} d^t O\left(\frac{1}{d^{2t}}\right)\right)\\ &= \sum_{\boldsymbol{u}} O\left(\frac{1}{d^m}\right)\\ &= d^k O\left(\frac{1}{d^m}\right).\end{aligned}$$

Hence, By Lemma 5, we have

$$\frac{\operatorname{var}(\langle R(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)}{\operatorname{var}(\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)} \le \frac{\operatorname{var}(\langle R(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)}{d^k - 4k} = \frac{d^k O(1/d^m)}{d^k - 4k} = O\left(\frac{1}{d^m}\right).$$

The next lemma states that the variance of the difference between

$$\frac{\langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle}{\sigma(\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)} \quad\text{and}\quad \frac{\langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle}{\sigma(\langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)}$$

can be made as small as we want, as $k \to \infty$.

Lemma 7. *We have*

$$\operatorname{var}\left(\frac{\langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle}{\sigma(\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)} - \frac{\langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle}{\sigma(\langle T(m)\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle)}\right) = O\left(\frac{1}{d^m}\right).$$

Proof. Again, as in Lemma 1, we have

$$\operatorname{var}(\langle d^k\boldsymbol{\Sigma}\boldsymbol{Z}^{\boldsymbol{A}}, \boldsymbol{Z}^{\boldsymbol{B}}\rangle) = d^{2k}\sum_{i,j}\sigma_{ij}^2,$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

$$\operatorname{var}(\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle) = \sum_{i,j} T(m)_{ij}^2,$$

$$\operatorname{var}(\langle R(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle) = \sum_{i,j} R(m)_{ij}^2.$$

By Construction 1 we have $d^{2k}\sum_{i,j}\sigma_{ij}^2 = \sum_{i,j} T(m)_{ij}^2 + \sum_{i,j} R(m)_{ij}^2$ (since we are simply rearranging the terms of the sum on the left-hand side). Hence,

$$\operatorname{var}(\langle d^k \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle) = \operatorname{var}(\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle) + \operatorname{var}(\langle R(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle). \tag{17}$$

Now,

$$\begin{aligned}
&\operatorname{var}\biggl(\frac{\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle d^k \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} - \frac{\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle)}\biggr)\\
&\quad= \frac{(\sigma(\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle) - \sigma(\langle d^k \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle))^2}{\operatorname{var}(\langle d^k \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)}\\
&\quad\le \frac{\operatorname{var}(\langle R(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle)}{\operatorname{var}(\langle d^k \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} \qquad \text{(by (17))}\\
&\quad= O\biggl(\frac{1}{d^m}\biggr) \qquad \text{(by Lemma 6).}
\end{aligned}$$

We now concentrate on the asymptotic behavior (as $k \to \infty$) of the term

$$\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle = \sum_{\boldsymbol{u},\boldsymbol{v}} T(m)_{\boldsymbol{uv}} Z_{\boldsymbol{v}}^{\boldsymbol{A}} Z_{\boldsymbol{u}}^{\boldsymbol{B}}. \tag{18}$$

We will need the following central limit theorem for the sum of the dependent random variables. It is a variation on Stein's result [13].

Theorem 5. ([7, Theorem 4.2].) *Let $X_1, \dots, X_N$ be random variables satisfying $|X_i - \mathrm{E}(X_i)| \le M$ almost surely, for $i = 1, \dots, N$, $\mathrm{E}(\sum_{i=1}^N X_i) = \lambda$, $\operatorname{var}(\sum_{i=1}^N X_i) = \sigma^2$, and $(1/N)\,\mathrm{E}(\sum_{i=1}^N |X_i - \mathrm{E}(X_i)|) = \mu$. Let $S_i \subset \{1, \dots, N\}$ be such that $j \in S_i$ if and only if $i \in S_j$, and (X_i, X_j) is independent of $\{X_k\}_{k \notin S_i \cup S_j}$ for $i, j = 1, \dots, N$. Then, for $D = \max_{1 \le i \le N} |S_i|$,*

$$\biggl|\Pr\biggl(\frac{\sum_{i=1}^N X_i - \lambda}{\sigma} \le w\biggr) - \Phi(w)\biggr| \le 7\frac{N\mu}{\sigma^3}(DM)^2.$$

We want to apply Theorem 5 to the sum in (18), with $X_{\boldsymbol{u},\boldsymbol{v}} = T(m)_{\boldsymbol{uv}} Z_{\boldsymbol{v}}^{\boldsymbol{A}} Z_{\boldsymbol{u}}^{\boldsymbol{B}}$, but first we need to approximate the summands by bounded random variables.

For $Z \sim \mathcal{N}(0, 1)$, let $\widetilde{Z}$ be the truncation of Z at $b > 0$. That is, $\widetilde{Z}$ has probability density function

$$f_{\widetilde{Z}}(z) = \frac{\phi(z)}{\Phi(b) - \Phi(-b)} \quad \text{for } |z| < b,$$

where ϕ is the standard normal probability density function. Then, $\mathrm{E}(\widetilde{Z}) = 0$ and

$$\operatorname{var}(\widetilde{Z}) = \biggl(1 - \frac{2b\phi(b)}{2\Phi(b) - 1}\biggr); \tag{19}$$

see, for example, [8]. In what follows, we take $b = b(k) = d^{k/a}$, where $a > 0$ is a constant.

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

For $\mathbf{Z} = (Z_1, \ldots, Z_N) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, write $\widetilde{\mathbf{Z}}$ for $(\widetilde{Z}_1, \ldots, \widetilde{Z}_N)$. Let

$$W(m) = \sum_{\boldsymbol{u},\boldsymbol{v}} T(m)_{\boldsymbol{u}\boldsymbol{v}} Z_{\boldsymbol{v}}^{\boldsymbol{A}} Z_{\boldsymbol{u}}^{\boldsymbol{B}} = \langle T(m)\mathbf{Z}^{\boldsymbol{A}}, \mathbf{Z}^{\boldsymbol{B}} \rangle$$

and

$$\widetilde{W}(m) = \sum_{\boldsymbol{u},\boldsymbol{v}} T(m)_{\boldsymbol{u}\boldsymbol{v}} \widetilde{Z}_{\boldsymbol{v}}^{\boldsymbol{A}} \widetilde{Z}_{\boldsymbol{u}}^{\boldsymbol{B}} = \langle T(m)\widetilde{\mathbf{Z}}^{\boldsymbol{A}}, \widetilde{\mathbf{Z}}^{\boldsymbol{B}} \rangle.$$

Since $\widetilde{\mathbf{Z}} \xrightarrow{\mathrm{D}} \mathbf{Z}$ as $k \to \infty$, it follows, from the mapping theorem (see, for example, [2, Theorem 29.2]), that

$$\widetilde{W}(m) \xrightarrow{\mathrm{D}} W(m) \quad \text{as } k \to \infty.$$

Corollary 1. *As $k \to \infty$, the asymptotic distributions of*

$$\frac{W(m)}{\sigma(W(m))} \quad \text{and} \quad \frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))}$$

are the same, provided that they exist.

Lemma 8. *We have* $\operatorname{var} \widetilde{W}(m) \geq (d^k - 4k)(1 + o(k))$.

Proof. As in the proof of Lemma 1 and using (19), we obtain

$$\operatorname{cov}(\widetilde{Z}_{\boldsymbol{v}}^{\boldsymbol{A}} \widetilde{Z}_{\boldsymbol{u}}^{\boldsymbol{B}}, \widetilde{Z}_{\boldsymbol{v}'}^{\boldsymbol{A}} \widetilde{Z}_{\boldsymbol{u}'}^{\boldsymbol{B}}) = \begin{cases} 0 & \text{if } (\boldsymbol{u}, \boldsymbol{v}) \neq (\boldsymbol{u}', \boldsymbol{v}'), \\ \left(1 - \dfrac{2b\phi(b)}{2\Phi(b) - 1}\right)^2 & \text{if } (\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u}', \boldsymbol{v}'). \end{cases}$$

Hence,

$$\operatorname{var} \widetilde{W}(m) = \left(1 - \frac{2b\phi(b)}{2\Phi(b) - 1}\right)^2 \sum_{\boldsymbol{u},\boldsymbol{v}} T(m)_{\boldsymbol{u},\boldsymbol{v}}^2.$$

The rest of the proof follows the proof of Lemma 5, i.e.

$$\begin{aligned} \operatorname{var} \widetilde{W}(m) &\geq \left(1 - \frac{2b\phi(b)}{2\Phi(b) - 1}\right)^2 \sum_{\boldsymbol{u}} T(m)_{\boldsymbol{u}\boldsymbol{u}}^2 \\ &= \left(1 - \frac{2b\phi(b)}{2\Phi(b) - 1}\right)^2 d^{2k} \sum_{\boldsymbol{u}} \sigma_{\boldsymbol{u}\boldsymbol{u}}^2 \\ &\geq \left(1 - \frac{2b\phi(b)}{2\Phi(b) - 1}\right)^2 (d^k - 4k) \qquad \text{(by Lemma 5)} \\ &= (1 + o(k))(d^k - 4k) \qquad (\text{since } 2b\phi(b)/(2\Phi(b) - 1) \to 0 \text{ as } k \to \infty). \end{aligned}$$

Lemma 9. *Fix m and let $b = b(k) = d^{k/a}$ with $a > 12$. Then we have*

$$\frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))} \xrightarrow{\mathrm{D}} \mathcal{N}(0, 1) \quad \textit{as } k \to \infty.$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

Proof. Let $X_{\boldsymbol{u},\boldsymbol{v}} = T(m)_{\boldsymbol{uv}}\widetilde{Z}^{\boldsymbol{A}}_{\boldsymbol{v}}\widetilde{Z}^{\boldsymbol{B}}_{\boldsymbol{u}}$. We want to apply Theorem 5 to

$$\widetilde{W}(m) = \sum_{\boldsymbol{u},\boldsymbol{v}} X_{\boldsymbol{u},\boldsymbol{v}}. \tag{20}$$

With the notation of Theorem 5, we have the following. By Lemma 4, the number N of nonzero terms in the sum in (20) satisfies

$$N \le 2d^k(1 + d + \cdots + d^m) = O(d^{k+m}).$$

Similarly, since $X_{\boldsymbol{u},\boldsymbol{v}}$ and $X_{\boldsymbol{u}',\boldsymbol{v}'}$ are independent if $\boldsymbol{u} \neq \boldsymbol{u}'$ and $\boldsymbol{v} \neq \boldsymbol{v}'$, we have that the dependency neighborhood $S_{\boldsymbol{u},\boldsymbol{v}}$ satisfies

$$|S_{\boldsymbol{u},\boldsymbol{v}}| \le 4(1 + d + \cdots + d^m).$$

Hence,

$$D = \max_{\boldsymbol{u},\boldsymbol{v}} |S_{\boldsymbol{u},\boldsymbol{v}}| = O(d^m).$$

Also note that $|T(m)_{\boldsymbol{uv}}| = O(1)$ (see Lemma 4), so $|X_{\boldsymbol{u},\boldsymbol{v}}| = O(b^2)$ and $\mu = (1/N) \times \sum_{\boldsymbol{u},\boldsymbol{v}} |X_{\boldsymbol{u},\boldsymbol{v}}| = O(b^2)$. Hence, from Theorem 5, with $M = O(b^2)$ we have

$$\begin{aligned}
\left|\Pr\left(\frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))} \le x\right) - \Phi(x)\right| &\le 7\frac{N\mu}{\sigma^3}(DM)^2 \\
&\le \frac{Cd^{k+m}b^2d^{2m}b^4}{(\sigma(\widetilde{W}(m)))^3} \qquad \text{(where } C \text{ is a constant)} \\
&= \frac{Cd^kd^{3m}(d^{k/a})^6}{(\sigma(\widetilde{W}(m)))^3} \\
&\le \frac{Cd^kd^{3m}d^{6k/a}}{((d^k - 4k)(1 + o(k)))^{3/2}} \qquad \text{(by Lemma 8)} \\
&\to 0 \quad \text{as } k \to \infty \text{ for } a > 12.
\end{aligned}$$

Proof of Theorem 4. By Proposition 2, it is enough to show that

$$\frac{\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} \xrightarrow{\text{D}} \mathcal{N}(0, 1).$$

Let $\varepsilon > 0$. We have,

$$\begin{aligned}
&\left|\Pr\left(\frac{\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} \le x\right) - \Phi(x)\right| \\
&\quad \le \left|\Pr\left(\frac{\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle \boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} \le x\right) - \Pr\left(\frac{\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle d^k\boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} \le x\right)\right| \\
&\quad\quad + \left|\Pr\left(\frac{\langle T(m)\boldsymbol{Z^A}, \boldsymbol{Z^B}\rangle}{\sigma(\langle d^k\boldsymbol{\Sigma Z^A}, \boldsymbol{Z^B}\rangle)} \le x\right) - \Pr\left(\frac{W(m)}{\sigma(W(m))} \le x\right)\right| \\
&\quad\quad + \left|\Pr\left(\frac{W(m)}{\sigma(W(m))} \le x\right) - \Pr\left(\frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))} \le x\right)\right| \\
&\quad\quad + \left|\Pr\left(\frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))} \le x\right) - \Phi(x)\right|.
\end{aligned} \tag{21}$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

By Lemma 6, for large enough k, we can find $m = m(k)$ such that

$$\left|\Pr\left(\frac{\langle \mathbf{\Sigma Z^A}, \mathbf{Z^B}\rangle}{\sigma(\langle \mathbf{\Sigma Z^A}, \mathbf{Z^B}\rangle)} \leq x\right) - \Pr\left(\frac{\langle T(m)\mathbf{Z^A}, \mathbf{Z^B}\rangle}{\sigma(\langle d^k \mathbf{\Sigma Z^A}, \mathbf{Z^B}\rangle)} \leq x\right)\right| < \frac{\varepsilon}{4}.$$

By Lemma 7, for large enough k and m, we have

$$\left|\Pr\left(\frac{\langle T(m)\mathbf{Z^A}, \mathbf{Z^B}\rangle}{\sigma(\langle d^k \mathbf{\Sigma Z^A}, \mathbf{Z^B}\rangle)} \leq x\right) - \Pr\left(\frac{W(m)}{\sigma(W(m))} \leq x\right)\right| < \frac{\varepsilon}{4}.$$

By Corollary 1, for large enough k, we obtain

$$\left|\Pr\left(\frac{W(m)}{\sigma(W(m))} \leq x\right) - \Pr\left(\frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))} \leq x\right)\right| < \frac{\varepsilon}{4}.$$

By Lemma 9, for large enough k and $b = d^{k/13}$, we have

$$\left|\Pr\left(\frac{\widetilde{W}(m)}{\sigma(\widetilde{W}(m))} \leq x\right) - \Phi(x)\right| < \frac{\varepsilon}{4}.$$

Hence, in (21), for large enough k, we obtain

$$\left|\Pr\left(\frac{\langle \mathbf{\Sigma Z^A}, \mathbf{Z^B}\rangle}{\sigma(\langle \mathbf{\Sigma Z^A}, \mathbf{Z^B}\rangle)} \leq x\right) - \Phi(x)\right| < \varepsilon.$$

4. Simulations

Simulations by Lippert *et al.* [9, Table 2] for $d = 4$ produced our Table 1. It shows that, for $k \geq 2$, normal behavior occurs for

$$2^{k-3} \times 100 < n.$$

For $k = 1$, Table 1 shows that, with $d = 4$, there is no apparent asymptotic normal behavior as n gets large.

However, in support of Theorem 1, the simulation in Figure 1 shows that, for $k = 1$, normal behavior already occurs when $d = 16$ and $n = 400$. The Kolmogorov–Smirnov p-value for this simulation was 0.8093. Simulations with $d = 10$ and $n = 800$ resulted in a similar good fit. We used the statistical language R® and used 2 500 sample points (to be consistent with the simulations in [9]).

TABLE 1: Kolmogorov–Smirnov p-values for uniform D_2 compared with normal ($d = 4$); see [9, Table 2].

n	$2^0 \times 10^2$	$2^1 \times 10^2$	$2^2 \times 10^2$	$2^3 \times 10^2$	$2^4 \times 10^2$	$2^5 \times 10^2$	$2^6 \times 10^2$	$2^7 \times 10^2$
$k = 1$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$k = 2$	0.038	0.158	0.287	0.475	0.078	0.191	0.258	0.009
$k = 3$	0.048	0.154	0.121	0.552	0.708	0.226	0.811	0.311
$k = 4$	0.000	0.057	0.048	0.813	0.689	0.658	0.982	0.692
$k = 5$	0.000	0.000	0.234	0.189	0.773	0.108	0.083	0.087
$k = 6$	0.000	0.000	0.001	0.071	0.087	0.720	0.067	0.452
$k = 7$	0.000	0.000	0.000	0.000	0.028	0.696	0.269	0.068
$k = 8$	0.000	0.000	0.000	0.000	0.000	0.063	0.657	0.054
$k = 9$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.321
$k = 10$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

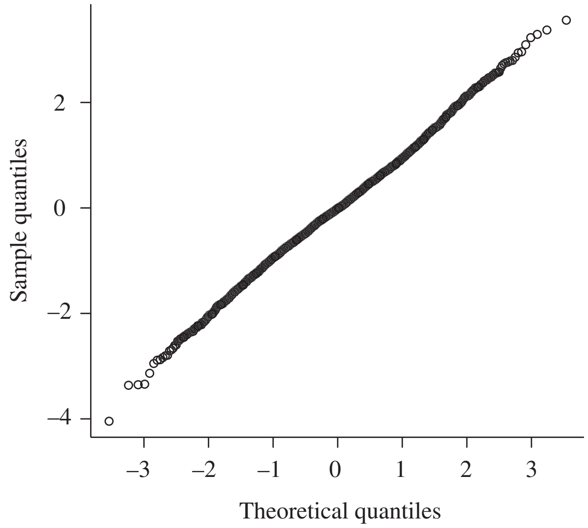

FIGURE 1: Normal Q-Q plot for uniform D_2 with $k = 1$, $d = 16$, and $n = 400$.

5. A formula for var$(D_2(n))$ in the uniform case

Proof of Theorem 3. With the notation of Section 2, we have

$$\operatorname{var}(D_2(n)) = \operatorname{var}\biggl(\sum_{(i,j)\in I} Y_{(i,j)}\biggr) = \sum_{(i,j)\in I} \operatorname{var}(Y_{(i,j)}) + \sum_{(i,j)\neq(s,t)} \operatorname{cov}(Y_{(i,j)}, Y_{(s,t)}).$$

The first term in (1) comes from the sum of the variances. To shorten notation, let $u = (i, j)$, then

$$\operatorname{var}(Y_u) = \mathrm{E}(Y_u^2) - (\mathrm{E}(Y_u))^2 = \mathrm{E}(Y_u) - (E(Y_u))^2 = \frac{1}{d^k} - \frac{1}{d^{2k}}.$$

Hence, summing up over all possible us, we obtain

$$\sum_{u\in I} \operatorname{var}(Y_u) = \bar{n}^2\biggl(\frac{1}{d^k} - \frac{1}{d^{2k}}\biggr).$$

Following the notation and terminology in [9], let $J_u = \{v = (s, t)\colon |s - i| < k \text{ or } |t - j| < k\}$ be the dependency neighborhood of Y_u. It can be decomposed into two parts, *accordion* and *crabgrass*, $J_u = J_u^{\mathrm{a}} \cup J_u^{\mathrm{c}}$, where

$$J_u^{\mathrm{a}} = \{v = (s, t) \in J_u\colon |s - i| < k \text{ and } |t - j| < k\} \quad \text{and} \quad J_u^{\mathrm{c}} = J_u \setminus J_u^{\mathrm{a}}.$$

We compute the cross covariances, $\operatorname{cov}(Y_u, Y_v)$, by looking at the following cases.

Case 1. ($v \notin J_u$.) In this case, Y_u and Y_v are independent; hence, $\operatorname{cov}(Y_u, Y_v) = 0$.

Case 2. ($v \in J_u^{\mathrm{c}}$.) We claim that in this case $\operatorname{cov}(Y_u, Y_v) = 0$. To see this let $u = (i, j)$ and $v \in J_u^{\mathrm{c}}$. By symmetry of the covariance, we may assume that $v = (i + t, j')$, where $|j - j'| \geq k$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

and $0 \le t < k$. Then, by direct computation, we have

$$\begin{aligned}\mathrm{E}(Y_u Y_v) = \Pr(Y_u = 1, Y_v = 1) &= \sum_{(a_1,\dots,a_{k+t})\in\mathcal{A}^{k+t}} \frac{1}{d^{3k+t}}\\ &= \frac{d^{k+t}}{d^{3k+t}}\\ &= \frac{1}{d^{2k}}.\end{aligned}$$

Hence,

$$\operatorname{cov}(Y_u, Y_v) = \mathrm{E}(Y_u Y_v) - \mathrm{E}(Y_u)\,\mathrm{E}(Y_v) = \frac{1}{d^{2k}} - \frac{1}{d^{2k}} = 0.$$

We note that in fact, in this case, Y_u and Y_v are independent.

Case 3. (v *is on the main diagonal of* J_u^{a}.) In this case, $v = (i+t, j+t)$, where $-k < t < k$ and $t \ne 0$. Here, we claim that $\operatorname{cov}(Y_u, Y_v) = 1/d^{k+|t|} - 1/d^{2k}$. As above, to prove this claim it is enough to show that $\Pr(Y_u = 1, Y_v = 1) = 1/d^{k+|t|}$. By symmetry, we may assume that $t > 0$, which gives

$$\begin{aligned}\Pr(Y_u = 1, Y_v = 1) &= \sum_{(a_1,\dots,a_{k+t})\in\mathcal{A}^{k+t}} \Pr(\text{a specific } (k+t)\text{-word match at the } (i,j) \text{ position})\\ &= \sum_{(a_1,\dots,a_{k+t})\in\mathcal{A}^{k+t}} \frac{1}{d^{2(k+t)}}\\ &= \frac{d^{k+t}}{d^{2(k+t)}}\\ &= \frac{1}{d^{k+t}}.\end{aligned}$$

Case 4. ($v \in J_u^{\mathrm{a}} \setminus \{$*main diagonal*$\}$.) In this case, $v = (i+s, j+t)$, where $s \ne t$, $0 < |s|$, and $|t| < k$. Here, we claim that $\operatorname{cov}(Y_u, Y_v) = 0$. The proof is again by direct computation. By symmetry, we may assume that $s, t > 0$. It is enough to show that $\Pr(Y_u = 1, Y_v = 1) = 1/d^{2k}$. Indeed, it is straightforward to check that

$$\Pr(Y_u = 1, Y_v = 1) = \sum_{(a_1,\dots,a_{s+t})\in\mathcal{A}^{s+t}} \frac{1}{d^{2k+s+t}} = \frac{d^{s+t}}{d^{2k+s+t}} = \frac{1}{d^{2k}}.$$

Finally, summing up over all the cross covariances we obtain the second term of (1), i.e.

$$\begin{aligned}\sum_u \sum_{v\ne u} \operatorname{cov}(Y_u, Y_v) &= \sum_u \sum_{\substack{-k<t<k\\ t\ne 0}} \biggl(\frac{1}{d^{k+|t|}} - \frac{1}{d^{2k}}\biggr)\\ &= \sum_u 2\biggl[\biggl(\sum_{t=1}^{k-1} \frac{1}{d^{k+t}}\biggr) - \frac{k-1}{d^{2k}}\biggr]\\ &= \sum_u 2\biggl[\frac{(1/d^{k+1})(1-1/d^{k-1})}{1-1/d} - \frac{k-1}{d^{2k}}\biggr]\\ &= 2\bar{n}^2\biggl[\frac{(1/d^{k+1})(1-1/d^{k-1})}{1-1/d} - \frac{k-1}{d^{2k}}\biggr].\end{aligned}$$

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press

Acknowledgement

This work was funded in part by ARC Discovery grant number DP0559260.

References

[1] Barbour, A. and Chryssaphinou, O. (2001). Compound Poisson approximation: a user's guide. *Ann. Appl. Prob.* **11,** 964–1002.

[2] Billingsley, P. (1995). *Probability and Measure*, 3rd edn. John Wiley, New York.

[3] Burke, J., Davison, D. and Hide, W. (1999). d2 cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.* **9,** 1135–1142.

[4] Carpenter, J. E., Christoffels, A., Weinbach, Y. and Hide, W. A. (2002). Assessment of the parallelization approach of d2 cluster for high-performance sequence clustering. *J. Comput. Chem.* **23,** 755–757.

[5] Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Prob.* **3,** 534–545.

[6] Christoffels, A. *et al.* (2001). STACK: sequence tag alignment and consensus knowledgebase. *Nucleic Acids Res.* **29,** 234–238.

[7] Dembo, A. and Rinott, Y. (1996). Some examples of normal approximations by Stein's method. In *Random Discrete Structures* (IMA Vol. Math. Appl. **76**), Springer, New York, pp. 25–44.

[8] Johnson, N. L. and Kotz, S. (1970). *Distributions in Statistics. Continuous Univariate Distributions. 1*. Houghton Mifflin Co., Boston, MA.

[9] Lippert, R. A., Huang, H and Waterman, M. S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proc. Nat. Acad. Sci. USA* **99,** 13980–13989.

[10] Miller, R. T. *et al.* (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* **9,** 1143–1155.

[11] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Molec. Biol.* **147,** 195–197.

[12] Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, Vol. II, University of California Press, Berkeley, pp. 583–602.

[13] Stein, C. (1986). *Approximate Computation of Expectations*. Institute of Mathematical Statistics, Hayward, CA.

[14] Vinga, S. and Almeida, J. S. (2003). Alignment-free sequence comparison – a review. *Bioinformatics* **19,** 513–523.

[15] Waterman, M. S. (1995). *Introduction to Computational Biology*. Chapman & Hall, New York.

[16] Zhang, Y. X. *et al.* (2002). Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* **415,** 644–646.

https://doi.org/10.1239/jap/1189717545 Published online by Cambridge University Press