Bernoulli **9**(2), 2003, 183–199

Upper bounds and importance sampling of *p*-values for DNA and protein sequence alignments

HOCK PENG CHAN

Department of Statistics and Applied Probability, Faculty of Science, National University of Singapore, Singapore 119260, Singapore. E-mail: stachp@nus.edu.sg

We show in general how the substitution matrix and gap penalty function for local sequence alignments can be chosen such that the score statistic grows at a logarithmic rate when the two sequences are unrelated. The method used is the construction of a mixture distribution in which sequences with large scores are generated with uniformly higher likelihood. This distribution is also used for the importance sampling of the p-value of the score. An upper bound of this p-value is computed and compared against the simulated value.

Keywords: exponential tilting; importance sampling; p-value; sequence alignments

1. Introduction

In the past decade, there has been tremendous progress in the understanding of the asymptotic behaviour of the local alignment score of two sequences; see, for example, Arratia and Waterman (1994), Dembo *et al*. (1994), Neuhauser (1994), Siegmund and Yakir (2000a; 2000b) and references therein. Heuristical approximations of the p-value of the scores have been obtained by Mott and Tribe (1999). Of interest are sequences of nucleotides or amino acids; a comprehensive account of the background to this topic is provided in Waterman and Vingron (1994) and Waterman (1995). The software program BLAST (see Altschul *et al*. 1990), currently in widespread use, implements an efficient search algorithm to approximate the scores, which are assigned according to the number and length of the gaps and the quality of the matches in the alignment, as measured from a substitution matrix.

When an equation (to be defined in Section 2) that is dependent on the substitution matrix and gap penalty function has a positive solution, a distribution Q can be constructed such that sequences having large scores are generated with uniformly higher likelihood compared to the null model in which the two sequences are unrelated. By using a change-of-measure argument, we obtain an upper bound for the p-value of the local alignment score that decays exponentially, thus ensuring that the score grows logarithmically with respect to the length of the sequences. The distribution Q can also be used to perform importance sampling of the p-value of the score. For efficient computation, an algorithm is presented that computes the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$ recursively, so that the computation time needed to estimate the p-value by generating the sequences from the distribution Q is comparable to that of direct Monte Carlo.

The hidden Markov model, which justifies the local alignment score as a maximum likelihood statistic (see Durbin *et al*. 1998), is introduced in Section 4. This model suggests a modification of the substitution matrix which would ensure that the equation defined in Section 2 has a positive solution. In Section 5, upper bounds of the p-value are obtained using random walk theory and exponential tilting. These upper bounds are then compared against the importance sampling estimates of the p-value in Section 6. The choice of appropriate gap penalty functions is also examined, and the paper concludes with an example that computes an upper bound for the longest common subsequence problem.

2. Choosing substitution matrices and gap penalty functions in the logarithmic region

Let $\mathbf{x} = \{x_1, \ldots, x_m\}$ and $\mathbf{y} = \{y_1, \ldots, y_n\}$ be two sequences of independent and identically distributed random variables taking values in a finite alphabet $\mathcal{A}$, with x_i having distribution μ and y_j distribution ν. Let $g : \{0, 1, \ldots\} \to [0, \infty]$ be a non-decreasing gap penalty function with $g(0) = 0$ and $K : \mathcal{A} \times \mathcal{A} \to [-\infty, \infty)$ a substitution matrix. We call $\mathbf{z} = \{(i_t, j_t) : 1 \leqslant t \leqslant u\}$ an alignment of u $(= u_{\mathbf{z}})$ matches if $1 \leqslant i_1 < \ldots < i_u \leqslant m$, $1 \leqslant j_1 < \ldots < j_u \leqslant n$ and, for all $1 \leqslant t \leqslant u-1$, either $i_{t+1} = i_t + 1$ or $j_{t+1} = j_t + 1$ (or both). For each alignment $\mathbf{z}$, define

$$S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{u} K(x_{i_t}, y_{j_t}) - \sum_{t=1}^{u-1} g(i_{t+1} - i_t - 1 + j_{t+1} - j_t - 1). \tag{2.1}$$

Let $\mathcal{Z}$ denote the class of all alignments and let the local alignment score

$$H_{m,n} = H_{m,n}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{z} \in \mathcal{Z}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}). \tag{2.2}$$

$K(x, y)$ is large when $x = y$ and also if x can be substituted easily by y or vice versa in the evolutionary process. Gaps are allowed in the alignment $\mathbf{z}$ to model for the insertion and deletion of segments in the sequences but are penalized through the gap penalty function. Hence a large value of $H_{m,n}$ indicates a strong possibility that a segment each of $\mathbf{x}$ and $\mathbf{y}$ is descended from a recent common ancestor.

Let $|\cdot|$ denote the number of elements in a finite set. $|\mathcal{Z}|$ increases exponentially with m, n and hence an affine penalty function of the form $g(k) = \Delta + \delta k$ for $k \geqslant 1$ is used, as the value of $H_{m,n}$ can then be computed in $O(mn)$ time and memory (Gotoh 1982). $\Delta + \delta$ is known as the gap opening penalty, and δ the gap extension penalty.

Consider the null hypothesis in which $\mathbf{x}$ and $\mathbf{y}$ are unrelated. If $\mathrm{E}[K(x_1, y_1)] > 0$, then $H_{m,n}$ increases linearly with $\min(m, n)$ and we lie in the so-called linear region. In this situation, $H_{m,n}$ is not useful for determining if $\mathbf{x}$ and $\mathbf{y}$ are related. It has been shown that under the conditions

$$\mathrm{E}[K(x_1, y_1)] < 0 \quad \text{and} \quad P\{K(x_1, y_1) > 0\} > 0, \tag{2.3}$$

if $g(1) = \infty$ (Dembo *et al*. 1994) or if Δ increases at a logarithmic rate (Siegmund and Yakir 2000b), then $H_{m,n}$ is of the order of $\log(mn)$ and we are said to lie in the logarithmic region. No conclusions have been drawn, however, for the case of g finite and fixed. In practice, for such g, the transition between the linear and logarithmic region is determined empirically. In the next theorem we shall show, using the underlying model, how (K, g) can be chosen appropriately to lie in the logarithmic region.

Theorem 1. *Let* (K, g) *be chosen such that*

(I) *the convex function*

$$h(\theta) = \left(1 + 2\sum_{k\geqslant 1} \mathrm{e}^{-\theta g(k)}\right) \sum_{x,y\in A} \mathrm{e}^{\theta K(x,y)}\mu(x)\nu(y)$$

has a positive solution, with the larger solution denoted by $\tilde{\theta}$.

Then:

(i) $P\{H_{m,n} \geqslant b\} \leqslant nm\mathrm{e}^{-\tilde{\theta}b}$;
(ii) *for any* $\epsilon > 0$, $P\{H_{m,n} \geqslant (1+\epsilon)\log(nm)/\tilde{\theta}\} \to 0$ *as* $nm \to \infty$;
(iii) $\lim_{n\to\infty} n^{-1}\mathrm{E}H_{n,n} = 0$.

Remark. Let the moment generating function

$$\Lambda(\theta) = \sum_{x,y\in\mathcal{A}} \mathrm{e}^{\theta K(x,y)}\mu(x)\nu(y). \tag{2.4}$$

As $\Lambda'(0) = \mathrm{E}[K(x_1, y_1)]$, under condition (2.3), $\Lambda(\theta) < 1$ for some $\theta > 0$ and hence if g chosen large enough, condition (I) is satisfied. Furthurmore, as $g(1) \to \infty$, $\tilde{\theta} \to \theta^*$, the positive root of the equation $\Lambda(\theta) = 1$. It follows from Dembo *et al*. (1994) that $P\{H_{m,n} \leqslant (1-\epsilon)\log(mn)/\theta^*\} \to 0$ as $nm \to \infty$ for all $\epsilon > 0$. Thus Theorem 1 is consistent with the work of Siegmund and Yakir (2000b) in the sense that, for large g, m and n, $H_{m,n}/\log(mn)$ is approximately $1/\theta^*$ under the null hypothesis.

Proof. Let $s > 0$ be such that

$$1 + 2\sum_{k\geqslant 1} \mathrm{e}^{-\tilde{\theta} g(k)} = \mathrm{e}^s. \tag{2.5}$$

Then by condition (I),

$$f_{\tilde{\theta}}(x, y) = \mathrm{e}^{\tilde{\theta}K(x,y)+s}\mu(x)\nu(y), \qquad \text{for } x, y \in \mathcal{A}, \tag{2.6}$$

is a probability mass function.

Let $H_{i,j} = H_{i,j}(x_1, \ldots, x_i, y_1, \ldots, y_j)$. We shall construct a mixture distribution Q with state space $\mathcal{A}^m \times \mathcal{A}^n$ in the following manner:

1. Pick (i_1, j_1) uniformly from $\{1, \ldots, m\} \times \{1, \ldots, n\}$ and let $x_i \sim \mu$ for $i < i_1$, $y_j \sim \nu$ for $j < j_1$ and $(x_{i_1}, y_{j_1}) \sim f_{\tilde{\theta}}$.
2. Define recursively, for $t \geqslant 1$, $i_{t+1} = i_t + 1 + \tau_t$ and $j_{t+1} = j_t + 1 + \sigma_t$, where
$$P\{(\tau_t, \sigma_t) = (k, 0)\} = P\{(\tau_t, \sigma_t) = (0, k)\} = \mathrm{e}^{-\tilde{\theta} g(k) - s}, \qquad \text{for } k = 0, 1, \ldots . \tag{2.7}$$
If $i_{t+1} \leqslant m$ and $j_{t+1} \leqslant n$, let $x_i \sim \mu$ for $i_t < i < i_{t+1}$, $y_j \sim \nu$ for $j_t < j < j_{t+1}$ and $(x_{i_{t+1}}, y_{j_{t+1}}) \sim f_{\tilde{\theta}}$.
3. Repeat step 2 until $u = \min\{t : H_{i_t, j_t} \geqslant b, i_{t+1} > m$ or $j_{t+1} > n\}$, let $x_i \sim \mu$ for $i > i_u$ and $y_j \sim \nu$ for $j > j_u$. Let $\mathbf{z} = \{(i_t, j_t) : 1 \leqslant t \leqslant u\}$, $Q_{\mathbf{z}}$ be the measure of $(\mathbf{x}, \mathbf{y})$ generated together with alignment $\mathbf{z}$ and $Q = \sum_{z \in \mathcal{Z}} Q_{\mathbf{z}}$.

If $(\mathbf{x}, \mathbf{y})$ belongs to the set $A = \{(\mathbf{x}, \mathbf{y}) : H_{m,n}(\mathbf{x}, \mathbf{y}) \geqslant b\}$, then there exists an alignment $\mathbf{z}$ such that $S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b$ and $H_{i_u - 1, j_u - 1} < b$. Then as $s > 0$ (see (2.5)),

$$\begin{aligned}
\frac{\mathrm{d}Q}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) &\geqslant \frac{\mathrm{d}Q_{\mathbf{z}}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) \\
&= (nm)^{-1} \exp\left\{\sum_{t=1}^{u} [\tilde{\theta} K(x_{i_t}, y_{j_t}) + s]\right\} \exp\left\{-\sum_{t=1}^{u-1} [\tilde{\theta} g(i_{t+1} - i_t - 1 + j_{t+1} - j_t - 1) + s]\right\} \\
&= (nm)^{-1} \exp[\tilde{\theta} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) + s] \\
&\geqslant (nm)^{-1} \mathrm{e}^{\tilde{\theta} b}.
\end{aligned} \tag{2.8}$$

(i) then follows from (2.8) since
$$P(A) = \mathrm{E}_Q\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \mathbf{1}_A\right). \tag{2.9}$$
(ii) follows directly from (i) by letting $b = (1 + \epsilon)\tilde{\theta}^{-1} \log(nm)$, while (iii) follows from (ii) and because $n^{-1} H_{n,n}$ is bounded above by $\max_{x, y \in \mathcal{A}} K(x, y)$. □

3. Importance sampling

Let $G(\mathbf{x}_r, \mathbf{y}_r) = \sup_{\mathbf{z} \in \mathcal{Z}} [S_{\mathbf{z}}(\mathbf{x}_r, \mathbf{y}_r) - g(r - i_u) - g(r - j_u)]$ and $f_r(\theta) = \log(\mathrm{E} \exp[\theta G(\mathbf{x}_r, \mathbf{y}_r)])$ be its log moment generating function. In Bundschuh (2002), r is chosen to be moderately large and $f_r(\theta)$ is approximated by a Monte Carlo estimate $\hat{f}_r(\theta)$ using an importance sampling algorithm in which a modification of step 2 (see (2.7)) is executed recursively with θ^*, the positive root of $\Lambda(\theta) = 1$, replacing $\tilde{\theta}$. Then p-values of the local alignment score can be estimated by fitting a conjectured asymptotic Gumbel-type distribution with the root of $\hat{f}_r(\theta) = 0$ as one of its parameters.

In this paper, we propose using the mixture distribution Q or a modified version of it for importance sampling of $p = P\{H_{m,n} \geqslant b\}$ directly. One advantage of our estimator is that it does not rely on any asymptotic theory of $H_{m,n}$ and hence can also be used for an independent verification of numerical approximations based on such asymptotics. We consider the following unbiased estimators of p: the importance sampling estimator,

$$\hat{p}_{\mathrm{I}} = B^{-1} \sum_{\ell=1}^{B} \frac{\mathrm{d}P}{\mathrm{d}Q}(\mathbf{x}^{(\ell)}, \mathbf{y}^{(\ell)}) \mathbf{1}_{\{H_{m,n}(\mathbf{x}^{(\ell)}, \mathbf{y}^{(\ell)}) \geqslant b\}}, \tag{3.1}$$

where $(\mathbf{x}^{(\ell)}, \mathbf{y}^{(\ell)})$, $1 \leqslant \ell \leqslant B$, are generated independently from Q; and the direct Monte Carlo estimator,

$$\hat{p}_{\mathrm{D}} = B^{-1} \sum_{\ell=1}^{B} \mathbf{1}_{\{H_{m,n}(\mathbf{x}^{(\ell)}, \mathbf{y}^{(\ell)}) \geqslant b\}}, \tag{3.2}$$

where $(\mathbf{x}^{(\ell)}, \mathbf{y}^{(\ell)})$, $1 \leqslant \ell \leqslant B$, are generated independently from $\mu^m \times \nu^n$. By (2.8),

$$\begin{aligned} \mathrm{var}(\hat{p}_{\mathrm{I}}) &= B^{-1}\left(\mathrm{E}_Q \left\{ \left[\frac{\mathrm{d}P}{\mathrm{d}Q}(\mathbf{x}, \mathbf{y}) \mathbf{1}_{\{H_{m,n} \geqslant b\}} \right]^2 \right\} - p^2 \right) \qquad (3.3) \\ &= B^{-1}\left(\mathrm{E}_P \left[\frac{\mathrm{d}P}{\mathrm{d}Q}(\mathbf{x}, \mathbf{y}) \mathbf{1}_{\{H_{m,n} \geqslant b\}} \right] - p^2 \right) \\ &\leqslant B^{-1} p(nm\mathrm{e}^{-\tilde{\theta} b} - p), \end{aligned}$$

where E_P refers to $(\mathbf{x}, \mathbf{y}) \sim \mu^m \times \nu^n$. Hence $\mathrm{var}(\hat{p}_{\mathrm{I}})/\mathrm{var}(\hat{p}_{\mathrm{D}}) \leqslant (nm\mathrm{e}^{-\tilde{\theta} b} - p)/(1-p) \to 0$ as $b \to \infty$. The use of exponential tilting for the importance sampling of large-deviation probabilities as in (3.1) has a long history and have been used successfully in many sequential analysis and change-point detection problems (Siegmund 1976; Lai and Shan 1999; Chan and Lai 1999; 2000).

The finite-state automaton which has been used to describe the computation of the local alignment score for affine penalty functions $g(k) = \Delta + \delta k$, for $k \geqslant 1$ (Durbin *et al*. 1998), can also be used to understand the recursive computation of the likelihood ratio $(\mathrm{d}P/\mathrm{d}Q)(\mathbf{x}, \mathbf{y})$. Let there be three states : M signifing a match; I_x signifying an unaligned letter in the $\mathbf{x}$ sequence; and I_y signifying an unaligned letter in the $\mathbf{y}$ sequence. A starting point is picked from $\{1, \ldots, m\} \times \{1, \ldots, n\}$ with starting state M, and the sequence of states generated from the transitions $M \leftrightarrow I_x$ and $M \leftrightarrow I_y$ will determine the alignment $\mathbf{z}$. For example, if the starting point is $(2, 3)$ and the sequence of states is $M \to I_x \to M \to M \to I_y \to I_y \to M$, then the sequence of pairs 'emitted' by these states is (x_2, y_3), $(x_3, -)$, (x_4, y_4), (x_5, y_5), $(-, y_6)$, $(-, y_7)$, (x_6, y_8) and the alignment $\mathbf{z} = \{(2, 3), (4, 4), (5, 5), (6, 8)\}$, with $-$ denoting a gap space.

Let $W_1, \ldots, W_r$ denote the sequence of states corresponding to alignment $\mathbf{z}$. Then by (2.8), we can write

$$\frac{\mathrm{d}Q_{\mathbf{z}}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) = (nm)^{-1} \mathrm{e}^{s} \prod_{k=1}^{r} L_k,$$

where

$$L_k = \begin{cases} e^{\tilde{\theta} K(x_{i_t}, y_{j_t})}, & \text{if } W_k \text{is the } t\text{th } M \text{ state },\\ e^{-\tilde{\theta}(\Delta+\delta)}, & \text{if } W_k = I_x \text{ or } I_y \text{ and } W_{k-1} = M,\\ e^{-\tilde{\theta}\delta}, & \text{if } W_k = W_{k-1} = I_x \text{ or } I_y. \end{cases} \tag{3.4}$$

For known sequences $\mathbf{x}$ and $\mathbf{y}$, we shall construct counters $V_M(i, j)$, $V_X(i, j)$ and $V_Y(i, j)$ for $0 \leqslant i \leqslant m$, $0 \leqslant j \leqslant n$, to update the sum of likelihood over $\mathcal{Z}$ when the finite-state automaton visits the states M, I_x and I_y respectively at (i, j). For example, the alignment $\mathbf{z} = \{(2, 3), (4, 4), (5, 5), (6, 8)\}$ contributes a likelihood of

$$\begin{aligned} (nm)^{-1}\, e^{s+\tilde{\theta}K(x_2,y_3)} &\text{ to } V_M(2, 3),\\ (nm)^{-1}\, e^{s+\tilde{\theta}[K(x_2,y_3)-(\Delta+\delta)]} &\text{ to } V_X(3, 3),\\ (nm)^{-1}\, e^{s+\tilde{\theta}[K(x_2,y_3)-(\Delta+\delta)+K(x_4,y_4)]} &\text{ to } V_M(4, 4),\\ \vdots \qquad &\qquad \vdots\\ (nm)^{-1}\, e^{s+\tilde{\theta}[K(x_2,y_3)-(\Delta+\delta)+K(x_4,y_4)+K(x_5,y_5)-(\Delta+\delta)-\delta]} &\text{ to } V_Y(6, 7). \end{aligned}$$

In the simulation of Q, we consider the finite-state automaton to have stopped whenever a match state M at (i, j) satisfies $H_{i,j} \geqslant b$ (see step 3 in the construction of Q in the proof of Theorem 1). We shall define counter $V_{\mathrm{E}}(i, j)$ to record the sum of likelihood when this occurs. Thus if $H_{6,8} < b$, $\mathbf{z}$ contributes a likelihood of $(nm)^{-1}e^{s+\tilde{\theta}S_{\mathbf{z}}(\mathbf{x},\mathbf{y})}$ to $V_M(6, 8)$ while if $H_{6,8} \geqslant b$, the likelihood is contributed instead to $V_{\mathrm{E}}(6, 8)$. The values of these counters can be obtained from the initialization $V_M(i, j) = V_X(i, j) = V_Y(i, j) = 0$ when $i = 0$ or $j = 0$ and the recurrence relations

$$\begin{aligned} V_{\mathrm{E}}(i, j) &= e^{\tilde{\theta}K(x_i,y_j)}[V_M(i-1, j-1) + V_X(i-1, j-1)\\ &\quad + V_Y(i-1, j-1) + (nm)^{-1}e^s]\mathbf{1}_{\{H_{i,j} \geqslant b\}},\\ V_M(i, j) &= e^{\tilde{\theta}K(x_i,y_j)}[V_M(i-1, j-1) + V_X(i-1, j-1)\\ &\quad + V_Y(i-1, j-1) + (nm)^{-1}e^s]\mathbf{1}_{\{H_{i,j} < b\}}\\ V_X(i, j) &= e^{-\tilde{\theta}(\Delta+\delta)}V_M(i-1, j) + e^{-\tilde{\theta}\delta}V_X(i-1, j),\\ V_Y(i, j) &= e^{-\tilde{\theta}(\Delta+\delta)}V_M(i, j-1) + e^{-\tilde{\theta}\delta}V_Y(i, j-1), \end{aligned} \tag{3.5}$$

for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$, where the term $(nm)^{-1}e^s$ accounts for a new match starting at (i, j). The likelihood ratio is

$$\frac{\mathrm{d}Q}{\mathrm{d}P}(\mathbf{x},\mathbf{y}) = \sum_{z\in\mathcal{Z}} \frac{\mathrm{d}Q_{\mathbf{z}}}{\mathrm{d}P}(\mathbf{x},\mathbf{y}) \tag{3.6}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} V_{\mathrm{E}}(i,j) + \sum_{i=m \text{ or } j=n} V_M(i,j)$$

$$+ \sum_{i=1}^{m-1}\sum_{j=1}^{n-1}[(1-\mathrm{e}^{-s})(\mathrm{e}^{-(m-i-1)\delta} + \mathrm{e}^{-(n-j-1)\delta})/2]V_M(i,j),$$

where the coefficient of $V_M(i, j)$ in the last line of (3.6) is the probability under Q that the match at (i, j) is the last one given that $H_{i,j} < b$. Numerical examples involving the estimator (3.1) using the algorithm (3.5)–(3.6) will be presented in Section 6 and compared with the direct Monte Carlo estimator (3.2). The p-values computed are then used to examine the sharpness of an upper bound of $P\{H_{m,n} \geqslant b\}$ derived in Section 5.

4. On substitution matrices and the hidden Markov model

Let q be a probability mass function on $\mathcal{A}\times\mathcal{A}$ and let H_1 be the hypothesis that there exists an alignment $\mathbf{z}$ between two sequences $\mathbf{x}$ and $\mathbf{y}$ such that

$$P\{(x_i, y_j) = (x, y)\} = \begin{cases} q(x, y), & \text{if } (i, j) = (i_t, j_t) \text{ for some } 1 \leqslant t \leqslant u, \\ \mu(x)\nu(y), & \text{otherwise.} \end{cases} \tag{4.1}$$

Let H_0 be the hypothesis that

$$P\{(x_i, y_j) = (x, y)\} = \mu(x)\nu(y), \qquad \text{for all } 1 \leqslant i \leqslant m \text{ and } 1 \leqslant j \leqslant n. \tag{4.2}$$

If

$$K(x, y) = \log\left[\frac{q(x, y)}{\mu(x)\nu(y)}\right] \tag{4.3}$$

and $g(1) = \infty$ (no gaps allowed), then $H_{m,n}$ (see (2.1) and (2.2)) is a maximum likelihood ratio statistic for testing H_0 against H_1. PAM and BLOSUM matrices are substitution matrices of the form (4.3) and differ only in the derivation of q.

For the affine penalty function $g(k) = \Delta + \delta k$ with Δ and δ finite, the score $H_{m,n}$ can also be expressed as a maximum likelihood ratio statistic by considering the hidden Markov model as discussed in Durbin *et al.* (1998, Chapter 4). In the hidden Markov model, the states M, I_x and I_y follow a Markov chain with transition matrix

$$\begin{pmatrix} -2\alpha & \alpha & \alpha \\ 1-\epsilon & \epsilon & 0 \\ 1-\epsilon & 0 & \epsilon \end{pmatrix}$$

with $0 < \alpha < 1/2$ and $0 < \epsilon < 1$. Let

$$K(x,\ y) = \log\left[\frac{q(x,\ y)}{\mu(x)\nu(y)}\right] + \log(1-2\alpha), \qquad \text{for } x,\ y \in \mathcal{A}, \tag{4.4a}$$

$$\Delta = -\log\left[\frac{\alpha(1-\epsilon)}{1-2\alpha}\right] + \log\epsilon \tag{4.4b}$$

$$\delta = -\log\epsilon. \tag{4.4c}$$

The alignment $\mathbf{z}$ with u matches, υ gap spaces $(-)$ and w gaps has $u-1-w$ transitions from M to M, w each from M to I and from I to M, and $\upsilon - w$ from I to I and hence, taking the likelihood of the alignment $\mathbf{z}$ into account, the likelihood ratio between H_1 and H_0 for an alignment $\mathbf{z}$ is

$$\begin{aligned}
&\left(\prod_{t=1}^{u}\frac{q(x_{i_t},\ y_{j_t})}{\mu(x_{i_t})\nu(y_{j_t})}\right)(1-2\alpha)^{u-1-w}\alpha^{w}(1-\epsilon)^{w}\epsilon^{\upsilon-w}\\
&= \exp\left(\sum_{t=1}^{u} K(x_{i_t},\ y_{j_t}) - w\Delta - \upsilon\delta\right)/(1-2\alpha)\\
&= \mathrm{e}^{S_{\mathbf{z}}(\mathbf{x},\mathbf{y})}/(1-2\alpha).
\end{aligned}$$

Thus $S_{\mathbf{z}}(\mathbf{x},\ \mathbf{y}) - \log(1-2\alpha)$ is the log-likelihood ratio statistic for the alignment $\mathbf{z}$ and $H_{m,n}$ is a maximum likelihood ratio statistic.

Lemma 1. *If* $(K,\ g)$ *are defined as in* (4.4)*, then condition* (I) *is satisfied.*

Proof .By (2.4) and (4.4),

$$\begin{aligned}
h(1) &= \left(1 + 2\sum_{k\geqslant 1}\mathrm{e}^{-(\Delta + k\delta)}\right)\sum_{x,y\in\mathcal{A}}\mathrm{e}^{K(x,y)}\mu(x)\nu(y). \qquad (4.5)\\
&= \left(1 + 2\frac{\alpha(1-\epsilon)}{(1-2\alpha)(1-\epsilon)}\right)\sum_{x,y\in\mathcal{A}} q(x,\ y)(1-2\alpha)\\
&= \sum_{x,y\in\mathcal{A}} q(x,\ y) = 1.
\end{aligned}$$

□

By (4.4a), we can carry out a correction of $\log(1-2\alpha)$ in the PAM and BLOSUM matrices. Then by Lemma 1 and Theorem 1, $(K,\ g)$ would automatically lie in the logarithmic region. In practice, however, such corrections are not performed as α is often considered to be small and the correction of $\log(1-2\alpha)$ would then not be significant.

5. A sharper upper bound

While Theorem 1(i) is useful for showing that (K, g) lies in the logarithmic region, it can be too crude for estimating p-values. The next theorem provides a sharper upper bound for $P\{H_{m,n} \geqslant b\}$. A numerical implementation of this theorem will be illustrated in an example in Section 6.

Theorem 2. *Let* (K, g) *be chosen to satisfy condition* (I) *and define* $\psi(\theta) = \log h(\theta)$. *Let* $b > 0$ *and* $u_0 = \lceil b/\max_{x,y\in\mathcal{A}} K(x, y)\rceil$ *be the minimal number of matches needed for an alignment score to exceed b, where* $\lceil x\rceil$ *is the smallest integer greater than or equal to x. Let* χ *be a measure with support on* $[\tilde{\theta}, \infty)$ *and define*

$$\eta = \min_{u\in\mathbf{Z},u_0\leqslant u\leqslant \min(m,n)} u\int e^{-u\psi(\theta)}d\chi(\theta).$$

Then

$$P\{H_{m,n} \geqslant b\} \leqslant nm\int(b\theta + 1)e^{-\theta b}d\chi(\theta)/\eta. \tag{5.1}$$

If $\{K(x, y) : x, y \in A\} \cup \{g(k) : k \geqslant 1\}$ *is lattice with span* κ *and b is a multiple of* κ, *we can replace the right-hand side of* (5.1) *by* $nm \int[b\kappa^{-1}(1 - e^{-\kappa\theta}) + 1]e^{-\theta b}d\chi(\theta)/\eta$.

Proof. Let $\mathcal{Z}_{i^*,j^*} = \{\mathbf{z} \in \mathcal{Z} : (i^*, j^*) \in \mathbf{z}\}$ where $(i^*, j^*) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$, and let the random variable

$$U_{i^*,j^*} = \max\{|\mathbf{z}| : \mathbf{z} \in \mathcal{Z}_{i^*,j^*}, S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b\}, \tag{5.2}$$

where, by convention, $\max\varnothing = 0$. We shall show later in the proof that

$$\ell_\theta(= \ell_{\theta,i^*,j^*}) = \sum_{u=1}^{\infty} e^{-u\psi(\theta)}P\left\{\max_{\mathbf{z}\in\mathcal{Z}_{i^*,j^*}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b;\ U_{i^*,j^*} = u\right\} \leqslant (b\theta + 1)e^{-\theta b} \tag{5.3}$$

for all $\theta \geqslant \tilde{\theta}$. Assuming first that (5.3) is true,

$$\sum_{i^*=1}^{m}\sum_{j^*=1}^{n}\sum_{u=1}^{\infty} e^{-u\psi(\theta)}P\left\{\max_{\mathbf{z}\in\mathcal{Z}_{i^*,j^*}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b;\ U_{i^*,j^*} = u\right\} \leqslant nm(b\theta + 1)e^{-\theta b}. \tag{5.4}$$

Let the random variable $U = \max\{|\mathbf{z}| : \mathbf{z} \in \mathcal{Z},\ S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b\}$. Now if $H_{m,n} \geqslant b$ and $U = u$, we can find some $\mathbf{z}_0$ such that $S_{\mathbf{z}_0}(\mathbf{x}, \mathbf{y}) \geqslant b$ with $|\mathbf{z}_0| = u$. Let $\mathbf{z}_0 = \{(i_t, j_t) : 1 \leqslant t \leqslant u\}$. Then, for each (i_t, j_t), $1 \leqslant t \leqslant u$, $\mathbf{z}_0 \in \mathcal{Z}_{i_t,j_t}$ and $U_{i_t,j_t} = u$. Hence it follows that

$$\sum_{i^*=1}^{m}\sum_{j^*=1}^{n}\sum_{u=1}^{\infty} e^{-u\psi(\theta)}P\left\{\max_{\mathbf{z}\in\mathcal{Z}_{i^*,j^*}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b,\ U_{i^*,j^*} = u\right\} \geqslant \sum_{u=1}^{\infty} ue^{-u\psi(\theta)}P\{H_{m,n} \geqslant b,\ U = u\}. \tag{5.5}$$

From (5.4)–(5.5) and integrating θ over the measure χ,

$$\eta P\{H_{m,n} \geqslant b\} \leqslant \sum_{u=u_0}^{\min(m,n)} u \int e^{-u\psi(\theta)} d\chi(\theta) P\{H_{m,n} \geqslant b,\, U = u\} \tag{5.6}$$

$$\leqslant nm \int (b\theta + 1) e^{-\theta b} d\chi,\, (\theta),$$

and Theorem 2 is proved.

To show (5.3), we observe that

$$\max_{\mathbf{z}\in\mathcal{Z}_{i^*,j^*}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{z}\in\mathcal{Z}^{(1)}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) + \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) - K(x_{i^*}, y_{j^*}), \tag{5.7}$$

where $\mathcal{Z}^{(1)}(= \mathcal{Z}^{(1)}_{i^*,j^*}) = \{\mathbf{z} \in \mathcal{Z} : (i_1, j_1) = (i^*, j^*)\}$ and $\mathcal{Z}^{(2)}(= \mathcal{Z}^{(2)}_{i^*,j^*}) = \{\mathbf{z} \in \mathcal{Z} : (i_u, j_u) = (i^*, j^*)\}$. In other words, $\mathcal{Z}^{(1)}$ consists of all alignments with (i^*, j^*) as the first match and $\mathcal{Z}^{(2)}$ consists of all alignments with (i^*, j^*) as the last match. For $\mathbf{z} \in \mathcal{Z}^{(2)}$, let $S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) - K(x_{i^*}, y_{j^*})$.

If $S'_{\mathbf{z}^{(2)}}(\mathbf{x}, \mathbf{y}) = w$ for some $\mathbf{z}^{(2)} \in \mathcal{Z}^{(2)}$ and $S_{\mathbf{z}^{(1)}}(\mathbf{x}, \mathbf{y}) \geqslant b - w$ for some $\mathbf{z}^{(1)} \in \mathcal{Z}^{(1)}$, then by (5.7), $S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \mathbf{b}$, where $\mathbf{z} = \mathbf{z}^{(1)} \cup \mathbf{z}^{(2)}$, and hence by (5.2), $U_{i^*,j^*} \geqslant U^{(2)} + U^{(1)}_{b-w} - 1$, where

$$U^{(2)} = \min\left\{ |\mathbf{z}| : \mathbf{z} \in \mathcal{Z}^{(2)},\, S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{z}'\in\mathcal{Z}^{(2)}} S_{\mathbf{z}'}(\mathbf{x}, \mathbf{y}) \right\}, \tag{5.8a}$$

$$U^{(1)}_c = \min\{|\mathbf{z}| : \mathbf{z} \in \mathcal{Z}^{(1)},\, S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant c\} \tag{5.8b}$$

Then as $\psi(\theta) \geqslant 0$ for $\theta \geqslant \tilde{\theta}$, it follows that

$$\ell_\theta = \sum_{u=1}^{\infty} e^{-u\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}_{i^*,j^*}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b,\, U_{i^*,j^*} = u \right\} \tag{5.9}$$

$$\leqslant \int_{-\infty}^{\infty} \sum_{v=1}^{\infty} e^{-(v-1)\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \in dw,\, U^{(2)} = v \right\}$$

$$\times \sum_{r=1}^{\infty} e^{-r\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(1)}} S_z(\mathbf{x}, \mathbf{y}) \geqslant b - w,\, U^{(1)}_{b-w} = r \right\}.$$

We shall show in the Appendix, using a technique similar to the proof of Theorem 1, that

$$\sum_{r=1}^{\infty} e^{-r\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(1)}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant c,\, U^{(1)}_c = r \right\} \leqslant \min(1, e^{-\theta c}). \tag{5.10}$$

Let $U^{(2)}_c = \min\{|\mathbf{z}| : \mathbf{z} \in \mathcal{Z}^{(2)},\, S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant c\}$. Then $U^{(2)} \geqslant U^{(2)}_c$ for all $c \leqslant \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y})$. By (5.9), (5.10) and considering the two separate cases $c > 0$ and $c \leqslant 0$, it follows that

$$\ell_\theta \leqslant \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} \left(\mathrm{e}^{-\theta b} \int_{-\infty}^{b} \mathrm{e}^{\theta w} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \in \mathrm{d}w,\ U^{(2)} = v \right\} \right. \tag{5.11}$$

$$\left. + P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b;\ U^{(2)} = v \right\} \right)$$

$$= \theta \mathrm{e}^{-\theta b} \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} \int_{-\infty}^{b} \mathrm{e}^{\theta\gamma} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma,\ U^{(2)} = v \right\} \mathrm{d}\gamma$$

$$\leqslant \theta \mathrm{e}^{-\theta b} \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} \int_{-\infty}^{b} \mathrm{e}^{\theta\gamma} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma,\ U^{(2)}_{\gamma} = v \right\} \mathrm{d}\gamma$$

since $\theta \int_{-\infty}^{w} \mathrm{e}^{\theta\gamma} \mathrm{d}\gamma = \mathrm{e}^{\theta w}$ for $w \leqslant b$. We shall also show in the Appendix, as in (5.10), that

$$\sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma,\ U^{(2)}_{\gamma} = v \right\} \leqslant \min(1, \mathrm{e}^{-\theta\gamma}). \tag{5.12}$$

Equation (5.3) then follows from (5.9) and (5.11)–(5.12), bringing the summation in the last line of (5.11) inside the integral and considering the two cases $\gamma \leqslant 0$ and $\gamma > 0$. For the lattice case, it follows by (5.10), (5.12), the arguments of (5.9), (5.11) and $(1 - \mathrm{e}^{-\kappa\theta}) \sum_{\gamma \leqslant w, \gamma\in\kappa\mathbf{Z}} \mathrm{e}^{\theta\gamma} = \mathrm{e}^{\theta w}$ for $w \in \kappa\mathbf{Z}$, that

$$\ell(\theta) \leqslant \sum_{w\in\kappa\mathbf{Z}} \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = w;\ U^{(2)} = v \right\}$$

$$\times \sum_{r=1}^{\infty} \mathrm{e}^{-r\psi(\theta)} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(1)}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b - w,\ U^{(1)}_{b-w} = r \right\}$$

$$\leqslant \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} \left(\mathrm{e}^{-\theta b} \sum_{w \leqslant b, w\in\kappa\mathbf{Z}} \mathrm{e}^{\theta w} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = w,\ U^{(2)} = v \right\} \right.$$

$$\left. + P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b + \kappa,\ U^{(2)} = v \right\} \right)$$

$$= (1 - \mathrm{e}^{-\kappa\theta}) \mathrm{e}^{-\theta b} \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} \sum_{\gamma \leqslant b, \gamma\in\kappa\mathbf{Z}} \mathrm{e}^{\theta\gamma} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma,\ U^{(2)} = v \right\}$$

$$\leqslant (1 - \mathrm{e}^{-\kappa\theta}) \mathrm{e}^{-\theta b} \sum_{v=1}^{\infty} \mathrm{e}^{-(v-1)\psi(\theta)} \sum_{\gamma \leqslant b, \gamma\in\kappa\mathbf{Z}} \mathrm{e}^{\theta\gamma} P\left\{ \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma,\ U^{(2)}_{\gamma} = v \right\}$$

$$\leqslant [b\kappa^{-1}(1 - \mathrm{e}^{-\kappa\theta}) + 1] \mathrm{e}^{-\theta b},$$

and hence the result for the lattice case follows from the arguments of (5.4)–(5.6). □

6. Examples

Example 1. Consider $|\mathcal{A}| = 4$ and $\mu = \nu$ uniform distributions on $\mathcal{A}$. Let

$$K(x, y) = \begin{cases} 1, & \text{if } x = y \\ -1 & \text{if } x \neq y, \end{cases}$$

and gap penalty $g(k) = \Delta + k$. Simulations are performed using estimators (3.1) and (3.2), while for an upper bound using (5.1) we consider χ to be a discrete measure with

$$\chi((1 + r/100)\tilde{\theta}) = \mathrm{e}^{r/q}, \qquad \text{for } r = 0, \ldots, 99. \tag{6.1}$$

Various values of q were tried, and it was found that $q = 40$ gave the sharpest upper bound. These values are compared against the simulated values in Table 1.

It would seem from (5.6) that a good choice of χ would be such that $\eta_u = u \int \mathrm{e}^{-u\psi(\theta)} \mathrm{d}\chi(\theta)$ is close to η for a wide range of values of $u \geqslant u_0$. This is true for (6.1) with $q = 40$ as can be seen in Table 2, thus suggesting that χ can be chosen essentially dependent only on (K, g).

Table 1. Estimates of $P\{H_{500,500} \geqslant b\}$. For simulations, $B = 10\,000$ repetitions.

b	Δ	$\tilde{\theta}$	Upper bound (5.1)	Direct MC (3.2)	Importance sampling (3.1)
15	4	1.069	9.764×10^{-3}	$(5.4 \pm 1.0) \times 10^{-3}$	$(6.036 \pm 0.051) \times 10^{-3}$
17	4		11.208×10^{-4}	$(2.0 \pm 2.0) \times 10^{-4}$	$(6.962 \pm 0.062) \times 10^{-4}$
19	4		12.967×10^{-5}	0	$(7.802 \pm 0.073) \times 10^{-5}$
15	5	1.090	7.909×10^{-3}	$(5.4 \pm 1.0) \times 10^{-3}$	$(5.457 \pm 0.040) \times 10^{-3}$
17	5		8.705×10^{-4}	0	$(6.086 \pm 0.046) \times 10^{-4}$
19	5		9.656×10^{-5}	0	$(6.714 \pm 0.054) \times 10^{-5}$

Table 2. List of values of η_u for $q = 40$.

$\Delta = 4$										
u	15	16	17	18	19	20	21	22	23	24
η_u	224.9	223.3	221.9	220.8	219.9	219.1	218.4	217.9	217.5	217.1
u	25	26	27	28	29	30	31	32	33	34
η_u	216.8	216.6	216.5	216.3	216.3	216.2	216.2	216.2	216.3	216.4
$\Delta = 5$										
u	15	16	17	18	19	20	21	22	23	24
η_u	252.7	251.1	249.7	248.7	247.8	247.1	246.5	246.1	245.7	245.4
u	25	26	27	28	29	30	31	32	33	34
η_u	245.2	245.1	245.0	244.9	244.9	244.9	245.0	245.1	245.2	245.3

Example 2. It follows by applying Theorem 1 on the BLOSUM62 matrix that the gap penalty functions $g(k) = 18 + k$ and $g(k) = 13 + 3k$ lie in the logarithmic region. The upper bound obtained using Theorem 2 with the discrete measure (6.1) for optimal q is approximately 10 times the asymptotic upper bound computed in Storey and Siegmund (2001), as can be seen in Table 3.

For smaller gap penalty functions, for example, $g(k) = 11 + k$ or $g(k) = 9 + 2k$, condition (I) fails to hold. Rather than computing the upper bound, we compute a Monte Carlo estimate based on a modification of the mixture distribution Q. Let $\Theta_1 = \{\theta : \Lambda(\theta) < 1\}$, where $\Lambda(\theta)$ is defined in (2.4). For $\theta \in \Theta_1$, let $c(\theta) \in (0, 1)$ satisfy the equation

$$\Lambda(\theta)(1 + 2c(\theta)\sum_{k\geqslant 1} \mathrm{e}^{-\theta g(k)}) = 1.$$

Pick $\hat{\theta}$ to maximize $c(\theta)$ over Θ_1 and define $\mathrm{e}^s = 1 + 2c(\hat{\theta})\sum_{k\geqslant 1}\mathrm{e}^{-\hat{\theta} g(k)}$ instead of (2.5). Simulate $(\mathbf{x}, \mathbf{y})$ in steps 1–3 of the proof of Theorem 1 with $\hat{\theta}$ replacing $\tilde{\theta}$ and (2.7) replaced by

$$P\{(\tau_t, \sigma_t) = (0, 0)\} = \Lambda(\hat{\theta}),$$

$$P\{(\tau_t, \sigma_t) = (k, 0)\} = P\{\tau_t, \sigma_t) = (0, k)\} = \Lambda(\hat{\theta})c(\hat{\theta})\mathrm{e}^{-\hat{\theta} g(k)}, \qquad \text{for } k = 1, 2, \ldots .$$

The recursive computation of V_X and V_Y in (3.5) is replaced by

$$V_X(i, j) = \mathrm{e}^{-\hat{\theta}(\Delta+\delta)} c(\hat{\theta}) V_M(i-1, j) + \mathrm{e}^{-\hat{\theta}\delta} V_X(i-1, j),$$

$$V_Y(i, j) = \mathrm{e}^{-\hat{\theta}(\Delta+\delta)} c(\hat{\theta}) V_M(i, j-1) + \mathrm{e}^{-\hat{\theta}\delta} V_Y(i, j-1).$$

The simulation results in Table 4 show that the importance sampling estimator is effective even in the estimation of a probability of the order of 10^{-6}, whereas the direct Monte Carlo estimator breaks down completely. The numerical estimates from Altschul and Gish (1996) are determined empirically by fitting a Gumbel-type distribution, while the estimates from

Table 3. Estimates of $P\{H_{500,500} \geqslant 81\}$.

Δ	δ	$\tilde{\theta}$	q	Upper bound (5.1)	Storey–Siegmund
18	1	0.2882	12	2.87×10^{-6}	0.21×10^{-6}
13	3	0.2857	13	3.57×10^{-6}	0.22×10^{-6}

Table 4. Estimates of $P\{H_{500,500} \geqslant 81\}$. For simulations, $B = 5000$ repetitions.

Δ	δ	Importance sampling (3.1)	Direct MC (3.2)	Storey–Siegmund	Altschul–Gish
11	1	$(2.74 \pm 0.15) \times 10^{-6}$	0	4.6×10^{-6}	2.3×10^{-6}
9	2	$(1.484 \pm 0.072) \times 10^{-6}$	0	2.2×10^{-6}	1.3×10^{-6}

Storey and Siegmund (2001) are based on the theoretical results of Siegmund and Yakir (2000b).

By Lemma 1 and (4.4) (with a $\sqrt{2}$ log base as used in BLOSUM matrices) it also follows that a correction of 0.5 in every entry of the substitution matrix would ensure that the local alignment score increases at a logarithmic rate when $g(k) = 7 + 2k$. A refinement of condition (I) is possible which we believe will show that $g(k) = 11 + k$ lies in the logarithmic region. However, the verification of this is computationally very intensive.

Example 3. Let $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\mathbf{y} = \{y_1, \ldots, y_n\}$ and let $\mu = \nu$ be uniform distributions on $\mathcal{A} = \{0, 1\}$. Consider the problem of finding the expected length of the longest common subsequence between $\mathbf{x}$ and $\mathbf{y}$. This is equivalent to letting

$$K(x, y) = \begin{cases} 1, & \text{if } x = y, \\ -\infty, & \text{if } x \neq y, \end{cases}$$

$g(k) = 0$ for all k, and finding $n^{-1}\mathrm{E}H_{n,n}$. For any consecutive string of 1s or consecutive string of 0s, there is no loss of generality in trying to align the beginning of the string first. Hence $H_{n,n}(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{z}\in\mathcal{Z}_1} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y})$, where $\mathcal{Z}_1$ is the class of all alignments in $\mathcal{Z}$ such that, for $1 \leqslant t \leqslant u-1$, $x_i \neq x_{i_{t+1}}$ for all $i_t < i < i_{t+1}$ and $y_j \neq y_{j_{t+1}}$ for all $j_t < j < j_{t+1}$. Let $K_a(0, 0) = K_a(1, 1) = 1 - a$, $K_a(0, 1) = K_a(1, 0) = -\infty$, $g_a(k) = ka/2$ for $k \geqslant 1$, and $H^{(a)}_{m,n} = \max_{\mathbf{z}\in\mathcal{Z}_1} S^{(a)}_{\mathbf{z}}(\mathbf{x}, \mathbf{y})$, where $S^{(a)}\mathrm{z}(\mathrm{x}, \mathrm{y})$ is the score for alignment $\mathbf{z}$ using the pair (K_a, g_a). Then

$$S^{(a)}_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) - (i_u - i_1 + j_u - j_1 + 2)a/2 \geqslant S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) - na,$$

implying that

$$n^{-1} H^{(a)}_{n,n}(\mathbf{x}, \mathbf{y}) \geqslant n^{-1} H_{n,n}(\mathbf{x}, \mathbf{y}) - a, \qquad \text{for all } \mathbf{x}, \mathbf{y}. \tag{6.2}$$

Let us consider the following condition:

(II) There exists a positive solution θ_a to the equation

$$\left(1 + 2\sum_{k\geqslant 1} \mathrm{e}^{-(\theta g_a(k) + k\log 2)}\right) \sum_{x,y\in\mathcal{A}} \mathrm{e}^{\theta K_a}(x, y)\mu(x)\nu(y) = 1.$$

If (II) is satisfied, we let $s_a \geqslant 0$ be such that

$$\mathrm{e}^{-s_a} = \sum_{x,y\in\mathcal{A}} \mathrm{e}^{\theta_a K_a(x,y)}\mu(x)\nu(y) = \mathrm{e}^{\theta_a(1-a)}/2. \tag{6.3}$$

We can generate distribution $Q^{(a)}$ as in the proof of Theorem 1 with

$$P\{(\tau_t, \sigma_t) = (k, 0)\} = P\{(\tau_t, \sigma_t) = (0, k)\} = \mathrm{e}^{-(\theta_a g_a(k) + k\log 2) - s_a}, \qquad \text{for } k = 0, 1, \ldots,$$

in place of (2.7), and whenever $i_{t+1} \leqslant m$ and $j_{t+1} \leqslant n$ in step 2 we let $x_i = 1 - x_{i_{t+1}}$ for $i_t < i < i_{t+1}$ and $y_j = 1 - y_{j_{t+1}}$ for $j_t < j < j_{t+1}$ instead of uniformly distributed on $\mathcal{A}$. Note that, by (6.3),

$$P\{(x_{i_t}, y_{j_t}) = (x, y)\} = e^{\theta_a} K_{a(x,y)+s_a} = \begin{cases} 1/2, & \text{if } (x, y) = (0, 0), \\ 1/2, & \text{if } (x, y) = (1, 1), \\ 0, & \text{otherwise.} \end{cases}$$

If $H^{(a)}_{n,n} \geqslant b$, then there exists $\mathbf{z} \in \mathcal{Z}_1$ with u matches and v gap spaces $(-)$ such that $S^{(a)}_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant b$ and $H_{i_u-1, j_u-1} < b$, so that

$$\begin{aligned} \frac{\mathrm{d}Q^{(a)}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) &\geqslant \frac{\mathrm{d}Q^{(a)}_{\mathbf{z}}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) \\ &= n^{-2} \exp[\theta_{\mathrm{a}} S^{(a)}_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) + s_a - v \log 2]/(1/2)^v \geqslant n^{-2} e^{\theta_a b}, \end{aligned} \tag{6.4}$$

since $s_a \geqslant 0$. The factor of $(1/2)^v$ in the second line of (6.4) is due to the fact that x_i is fixed for $i_t < i < i_{t+1}$ and y_j fixed for $j_t < j < j_{t+1}$ when generated under $Q^{(a)}$, whereas it has probability $1/2$ of taking either 0 or 1 under P. By (6.4), $\lim_{n\to\infty} n^{-1} \mathrm{E} H^{(a)}_{n,n} = 0$ as in Theorem 1. Hence $\lim_{n\to\infty} n^{-1} \mathrm{E} H_{n,n} \leqslant a$ by (6.2). It can be shown numerically that condition (II) is satisfied for $0 < a < 0.858\,68$ so that $\lim_{n\to\infty} n^{-1} \mathrm{E} H_{n,n} \leqslant 0.858\,68$. This is a modest improvement on the upper bound of $0.866\,66$ obtained by Chvátal and Sankoff (1975).

Appendix

Proof of (5.10). We need only consider $c > 0$, since the case $c < 0$ follows from $\psi(\theta) \geqslant 0$. Let $s_1(\theta)$ be such that $1 + 2\sum_{k\geqslant 1} e^{-\theta g(k)} = e^{s_1}(\theta)$ and $s_2(\theta)$ such that

$$f_\theta(x, y) = e^{\theta K(x,y) + s_2(\theta)} \mu(x) \nu(y), \qquad \text{for } x, y \in \mathcal{A},$$

is a probability mass function. Then by condition (I),

$$\psi(\theta) = \log h(\theta) = s_1(\theta) - s_2(\theta). \tag{A.1}$$

Construct a mixture distribution $Q^{(1)}$ as follows:

1. Let $(i_1, j_1) = (i^*, j^*)$, $x_i \sim \mu$ for $i < i_1$, $y_j \sim \nu$ for $j < j_1$ and $(x_{i_1}, y_{j_1}) \sim f_\theta$.
2. Define recursively, for $t \geqslant 1$, $i_{t+1} = i_t + 1 + \tau_t$ and $j_{t+1} = j_t + 1 + \sigma_t$, where

$$P\{(\tau_t, \sigma_t) = (k, 0)\} = P\{(\tau_t, \sigma_t) = (0, k)\} = e^{-\theta g(k) - s_1(\theta)}, \qquad \text{for } k = 0, 1, \ldots. \tag{A.2}$$

   If $i_{t+1} \leqslant m$ and $j_{t+1} \leqslant n$, let $x_i \sim \mu$ for $i_t < i < i_{t+1}$, $y_j \sim \nu$ for $j_t < j < j_{t+1}$ and $(x_{i_{t+1}}, y_{j_{t+1}}) \sim f_\theta$.
3. Let $\mathbf{z}^{(t)} = \{(i_k, j_k) : 1 \leqslant k \leqslant t\}$ and repeat step 2 until $U = \min\{t : S_{\mathbf{z}^{(t)}} \geqslant c$ or $i_{t+1} > m$ or $j_{t+1} > n\}$. Let $\mathbf{z} = \mathbf{z}^{(U)}$, $x_i \sim \mu$ for $i > i_U$ and $y_j \sim \nu$ for $j > j_U$. Let $Q^{(1)}_{\mathbf{z}}$ be the measure of $(\mathbf{x}, \mathbf{y})$ generated together with alignment $\mathbf{z}$ and $Q^{(1)} = \sum_{\mathbf{z}\in\mathcal{Z}^{(1)}} Q^{(1)}_{\mathbf{z}}$.

If $(\mathbf{x}, \mathbf{y}) \in A_u = \{(\mathbf{x}, \mathbf{y}) : \max_{\mathbf{z}\in\mathcal{Z}^{(1)}} S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant c,\ U_c^{(1)} = u\}$ (see (5.8b)), then there exists some $\mathbf{z} \in \mathcal{Z}^{(1)}$ with u matches such that $S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant c$. By the arguments of (2.8) and (A.1),

$$\frac{\mathrm{d}Q^{(1)}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) \geqslant \frac{\mathrm{d}Q^{(1)}_{\mathbf{z}}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) = \exp[\theta S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) + us_2(\theta) - (u-1)s_1(\theta)] \geqslant \exp[\theta c - u\psi(\theta)] \tag{A.3}$$

since $s_1(\theta) \geqslant 0$. By (A.3),

$$\sum_{u=1}^{\infty} \mathrm{e}^{-u\psi(\theta)} P(A_u) = \sum_{u=1}^{\infty} \mathrm{e}^{-u\psi(\theta)} \mathrm{E}_{Q^{(1)}}\left(\frac{\mathrm{d}P}{\mathrm{d}Q^{(1)}}(\mathbf{x}, \mathbf{y})\mathbf{1}_{A_u}\right) \leqslant \mathrm{e}^{-\theta c} \sum_{u=1}^{\infty} (\mathrm{E}_{Q^{(1)}} \mathbf{1}_{A_u}) \leqslant \mathrm{e}^{-\theta c}. \tag{A.4}$$

□

Proof of (5.12). As $\psi(\theta) \geqslant 0$, we need only consider $\gamma > 0$. In this proof, we shall label the matches of $\mathbf{z} = \{(i_t, j_t) : 1 \leqslant t \leqslant v\}$ in decreasing order instead of the conventional increasing order. Thus $i_1 > \ldots > i_v$. Construct a mixture distribution $Q^{(2)}$ as follows:

1. Let $(i_1, j_1) = (i^*, j^*)$, $x_i \sim \mu$ for $i \geqslant i_1$ and $y_j \sim \nu$ for $j \geqslant j_1$.
2. Define recursively, for $t \geqslant 1$, $i_{t+1} = i_t - 1 - \tau_t$ and $j_{t+1} = j_t - 1 - \sigma_t$, where (τ_t, σ_t) are distributed as in (A.2). If $i_{t+1} \geqslant 1$ and $j_{t+1} \geqslant 1$, let $x_i \sim \mu$ for $i_t > i > i_{t+1}$, $y_j \sim \nu$ for $j_t > j > j_{t+1}$ and $(x_{i_{t+1}}, y_{j_{t+1}}) \sim f_\theta$.
3. Let $\mathbf{z}^{(t)} = \{(i_k, j_k) : 1 \leqslant k \leqslant t\}$ and repeat step 2 until $U = \min\{t : S'_{\mathbf{z}^{(t)}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma$ or $i_{t+1} < 1$ or $j_{t+1} < 1\}$. Let $\mathbf{z} = \mathbf{z}^{(U)}$, $x_i \sim \mu$ for $i < i_U$ and $y_j \sim \nu$ for $j < j_U$. Let $Q^{(2)}_{\mathbf{z}}$ be the measure of $(\mathbf{x}, \mathbf{y})$ generated together with alignment $\mathbf{z}$ and $Q^{(2)} = \sum_{\mathbf{z}\in\mathcal{Z}^{(2)}} Q^{(2)}_{\mathbf{z}}$.

Now if $(\mathbf{x}, \mathbf{y}) \in A^{(2)}_v = \{(\mathbf{x}, \mathbf{y}) : \max_{\mathbf{z}\in\mathcal{Z}^{(2)}} S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma,\ U^{(2)}_\gamma = v\}$ (see the line after (5.10)), then there exists some $\mathbf{z} \in \mathcal{Z}^{(2)}$ with v matches such that $S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) \geqslant \gamma$. By the arguments of (2.8) and (A.1),

$$\frac{\mathrm{d}Q^{(2)}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) \geqslant \frac{\mathrm{d}Q^{(2)}_{\mathbf{z}}}{\mathrm{d}P}(\mathbf{x}, \mathbf{y}) = \exp\{\theta S'_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) + (v-1)[s_2(\theta) - s_1(\theta)]\} \geqslant \exp[\theta\gamma - (v-1)\psi(\theta)],$$

and (5.12) can be shown by the arguments of (A.4). □

Acknowledgements

I would like to thank T.L. Lai for introducing me to various importance sampling techniques and David Siegmund for helpful discussions that led to improvements in this manuscript. I would also like to thank a referee for his comments and also for pointing out an important reference.

References

Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Methods in Enzymology*, **266**, 460–480.

Altschul, S.F., Gish, W., Miller, W., Meyers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arratia, R. and Waterman, M.S. (1994) A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, **4**, 200–225.

Bundschuh, R. (2002) Rapid significance alignment with gaps, *J. Comput. Biol.*, **9**, 243–260.

Chan, H.P. and Lai, T.L. (1999) Importance sampling for Monte Carlo evaluation of boundary crossing probabilities in hypothesis testing and changepoint detection. Technical Report, Department of Statistics, Stanford University.

Chan, H.P. and Lai, T.L. (2000) Asymptotic approximations for error probabilities of sequential or fixed sample size tests in exponential families. *Ann. Statist.*, **28**, 1638–1669.

Chvátal, V. and Sankoff, D. (1975) Longest common subsequences of two random sequences. *J. Appl. Probab.*, **12**, 306–315.

Dembo, A., Karlin, S. and Zeitouni, O. (1994) Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Probab.*, **22**, 2022–2039.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis*. Cambridge: Cambridge University Press.

Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705.

Lai, T.L. and Shan, J. (1999) Efficient recursive algorithms for detection of abrupt changes in signal and control systems. *IEEE Trans. Automat. Control*, **44**, 952–966.

Mott, R. and Tribe, R. (1999) Approximate statistics of gapped alignments. *J. Comput. Biol.*, **6**, 91–112.

Neuhauser, C. (1994) A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Statist.*, **22**, 1603–1629.

Siegmund, D. (1976) Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, **4**, 673–684.

Siegmund, D. and Yakir, B. (2000a) Tail probabilities for the null distribution of scanning statistics, *Bernoulli*, **6**, 191–213.

Siegmund, D. and Yakir, B. (2000b) Approximate p-values for local sequence alignments. *Ann. Statist.*, **28**, 657–680.

Storey, J.D. and Siegmund, D. (2001) Approximate p-values for local sequence alignments: numerical studies. *J. Comput. Biol.*, **8**, 549–556.

Waterman, M.S. (1995) *Introduction to Computational Biology: Maps, Sequences and Genomes*. London: Chapman & Hall.

Waterman, M.S. and Vingron, M. (1994) Sequence comparison and Poisson approximation. *Statist. Sci*., **9**, 367–381.

Received August 2001 and revised May 2002