Accelerated convergence and robust asymptotic regression of the Gumbel scale parameter for gapped sequence alignment

2005 J. Phys. A: Math. Gen. 38 97

(http://iopscience.iop.org/0305-4470/38/1/006)

J. Phys. A: Math. Gen. **38** (2005) 97–108 doi:10.1088/0305-4470/38/1/006

Accelerated convergence and robust asymptotic regression of the Gumbel scale parameter for gapped sequence alignment

Yonil Park, Sergey Sheetlin and John L Spouge

National Center for Biotechnology Information, National Library of Medicine,
Bethesda MD 20894, USA

E-mail: spouge@ncbi.nlm.nih.gov

Received 28 September 2004, in final form 2 November 2004
Published 8 December 2004
Online at stacks.iop.org/JPhysA/38/97

Abstract
Searches through biological databases provide the primary motivation for studying sequence alignment statistics. Other motivations include physical models of annealing processes or mathematical similarities to, e.g., first-passage percolation and interacting particle systems. Here, we investigate sequence alignment statistics, partly to explore two general mathematical methods. First, we model the global alignment of random sequences heuristically with Markov additive processes. In sequence alignment, the heuristic suggests a numerical acceleration scheme for simulating an important asymptotic parameter (the Gumbel scale parameter λ). The heuristic might apply to similar mathematical theories. Second, we extract the asymptotic parameter λ from simulation data with the statistical technique of robust regression. Robust regression is admirably suited to 'asymptotic regression' and deserves to be better known for it.

PACS numbers: 02.50.Ga, 87.10.+e, 87.15.Cc

1. Introduction

Computational tools for sequence alignment are indispensable to modern molecular biology. Nowadays, the functional, structural and evolutionary relationships of a novel protein or nucleic acid sequence are often inferred by finding similar sequences of known function in a database. Because subsequences of a biological sequence (e.g., a protein or nucleic acid) often contribute to its functionality, a 'local alignment', which compares subsequences (Smith and Waterman 1981), is often more sensitive in determining relationships than a 'global alignment', which compares entire sequences (Needleman and Wunsch 1970).

0305-4470/05/010097+12$30.00

Local alignment is therefore more important in database applications than global alignment (Altschul *et al* 1990, 1997, Schaffer *et al* 2001). From a physical perspective, however, both global and local alignments are interesting, because they can provide approximate models for different types of annealing, e.g., between either entire molecules of DNA or between DNA subsequences (Waterman *et al* 1987). The mathematics of sequence alignment draws heavily from path optimization, as follows (Needleman and Wunsch 1970).

Let $\mathbf{A} = A_0A_1A_2\ldots$ and $\mathbf{B} = B_0B_1B_2\ldots$ be two infinite sequences drawn from a finite alphabet L, e.g., the amino acid alphabet {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} or the nucleotide alphabet {A, C, G, T}. Let $S : L \times L \mapsto \mathbb{R}$ denote a 'scoring matrix'. In a physical application, $S(a, b)$ represents free energy dissipated when, e.g., nucleic acids a and b from two different DNA molecules form hydrogen bonds. In database applications, $S(a, b)$ quantifies some type of similarity between a and b, e.g., the PAM and BLOSUM scoring matrices quantify the evolutionary similarity between two amino acids (Dayhoff *et al* 1978, Henikoff and Henikoff 1992).

The alignment graph Γ of the sequence-pair $(\mathbf{A}, \mathbf{B})$ is a directed, weighted lattice graph in two dimensions, as follows. The vertices v of Γ are non-negative integer points (i, j). (Below, ':=' denotes a definition, e.g., the natural numbers are $\mathbb{N} := \{1, 2, 3, \ldots\}$; i, j, k, m, n and g are integers throughout the paper.) Three sets of directed edges e come out of each vertex $v = (i, j)$: northward, northeastward and eastward. One northeastward edge goes into $(i + 1, j + 1)$ with weight $S(A_i, B_j)$. For each $g > 0$, one eastward edge goes into $(i + g, j)$ and one northward edge goes into $(i, j + g)$; both are assigned the same weight $-\Delta(g) < 0$. The deterministic function $\Delta : \mathbb{N} \mapsto (0, \infty)$ is called the 'gap penalty'. Affine gap penalties $\Delta(g) = a + bg$ are typical in database searches.

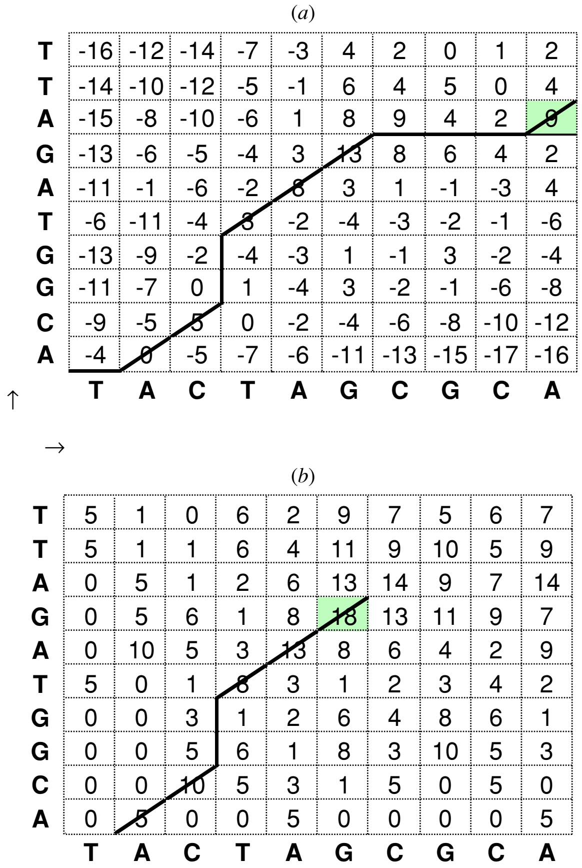
A directed path $\pi = (v_0, e_1, v_1, e_2, \ldots, e_k, v_k)$ in Γ is a finite, alternating sequence of vertices and edges that starts and ends with a vertex. For each $i = 1, 2, \ldots, k$, the directed edge e_i comes out of vertex v_{i-1} and goes into vertex v_i. We say that the path π starts at v_0 and ends at v_k.

Denote subsequences of a sequence $\mathbf{A}$ by $\mathbf{A}[i, m] = A_iA_{i+1}\ldots A_m$. Every gapped alignment of the subsequences $\mathbf{A}[i, m]$ and $\mathbf{B}[j, n]$ corresponds to exactly one directed path that starts at $v_0 = (i, j)$ and ends at $v_k = (m, n)$ (see figure 1). The alignment's score is the 'path weight' $W_\pi := \sum_{i=1}^{k} W(e_i)$.

Define the 'global score' $S_{ij} := \max_\pi W_\pi$, where the maximum is taken over all paths π starting at $v_0 = (0, 0)$ and ending at $v_k = (i, j)$. The paths π starting at v_0 and ending at v_k with weight $W_\pi = S_{ij}$ are 'optimal global paths' and correspond to 'optimal global alignments' between $\mathbf{A}[0, i]$ and $\mathbf{B}[0, j]$. Define the 'corner maximum' $S_n := S_{nn}$, the 'edge maximum' $E_n := \max\{\max_{0\leqslant i\leqslant n} S_{in}, \max_{0\leqslant j\leqslant n} S_{nj}\}$, and the 'global maximum' $M := \sup_{n\geqslant 0} E_n$. (The single subscript in S_n and E_n indicates that the variables correspond to a square $[0, n] \times [0, n]$, and not a general rectangle $[0, m] \times [0, n]$.)

Define also the 'local score' $\hat{S}_{ij} := \max_\pi W_\pi$, where the maximum is taken over all paths π ending at $v_k = (i, j)$, regardless of their starting point. Define the 'local maximum' $\hat{M}_{mn} := \max_{0\leqslant i\leqslant m, 0\leqslant j\leqslant n} \hat{S}_{ij}$. The paths π ending at $v_k = (i, j)$ with local score $W_\pi = \hat{S}_{ij} = \hat{M}_{mn}$ are 'optimal local paths' corresponding to the 'optimal local alignments' between subsequences of $\mathbf{A}[0, m]$ and $\mathbf{B}[0, n]$. 'Ungapped local alignment' is the case where $\Delta(g) \equiv \infty$ identically, because then no optimal local path includes a northward or eastward edge, i.e., as the terminology suggests, gaps are absent from optimal ungapped local alignments.

Now, we introduce randomness. Choose each letter in the sequences $\mathbf{A}$ and $\mathbf{B}$ randomly from a fixed distribution on the alphabet L. Under this 'independent letters' model, each random optimal global alignment score can be viewed as a variant type of first-passage



(*a*)

T	-16	-12	-14	-7	-3	4	2	0	1	2
T	-14	-10	-12	-5	-1	6	4	5	0	4
A	-15	-8	-10	-6	1	8	9	4	2	9
G	-13	-6	-5	-4	3	13	8	6	4	2
A	-11	-1	-6	-2	8	3	1	-1	-3	4
T	-6	-11	-4	3	-2	-4	-3	-2	-1	-6
G	-13	-9	-2	-4	-3	1	-1	3	-2	-4
G	-11	-7	0	1	-4	3	-2	-1	-6	-8
C	-9	-5	5	0	-2	-4	-6	-8	-10	-12
A	-4	0	-5	-7	-6	-11	-13	-15	-17	-16
	T	**A**	**C**	**T**	**A**	**G**	**C**	**G**	**C**	**A**

(*b*)

T	5	1	0	6	2	9	7	5	6	7
T	5	1	1	6	4	11	9	10	5	9
A	0	5	1	2	6	13	14	9	7	14
G	0	5	6	1	8	18	13	11	9	7
A	0	10	5	3	13	8	6	4	2	9
T	5	0	1	8	3	1	2	3	4	2
G	0	0	3	1	2	6	4	8	6	1
G	0	0	5	6	1	8	3	10	5	3
C	0	0	10	5	3	1	5	0	5	0
A	0	5	0	0	5	0	0	0	0	5
	T	**A**	**C**	**T**	**A**	**G**	**C**	**G**	**C**	**A**

Figure 1. Gapped alignment scores and the corresponding directed paths for two subsequences $\mathbf{A}[0, 10]$ = TACTAGCGCA and $\mathbf{B}[0, 10]$ = ACGGTAGATT of sequences drawn from the nucleotide alphabet {A, C, G, T}. Both figures use the standard nucleotide scoring matrix, with $S(a, b) = 5$ if $a = b$ and -4 otherwise, and the affine gap penalty $\Delta(g) = 3 + 2g$. The vertex (i, j) is in the northeast corner of the cell (i, j), with the origin $(0, 0)$ at the southwest corner of each figure. In (*a*), the cell (i, j) displays the global score S_{ij}, calculated from equation (3). The optimal global path ending at the point $(10, 8)$, e.g., consists of 12 edges, in the following order: 1 east of length 1, 2 northeast, 1 north of length 2, 3 northeast, 1 east of length 3, and 1 northeast. The optimal global path corresponds to the global sequence alignment of $\mathbf{A}[0, 10]$ and $\mathbf{B}[0, 8]$, TAC− −TAGCGCA and −ACGGTAG− − −A. The global score $S_{10,8} = -5 + 5 + 5 - 7 + 5 + 5 + 5 - 9 + 5 = 9$ is the sum of the corresponding edges and represents the path of greatest weight starting at $(0, 0)$ and ending at $(10, 8)$. The six most northeasterly corner maxima are $S_5 = -2$, $S_6 = 3$, $S_7 = 8$, $S_8 = 4$, $S_9 = 0$, $S_{10} = 2$, the corresponding edge maxima are $E_5 = 3$, $E_6 = 8$, $E_7 = 13$, $E_8 = 9$, $E_9 = 6$, $E_{10} = 9$, and the global maximum M for $\mathbf{A}$ and $\mathbf{B}$ is no less than 13, the largest global score shown. In (*b*), each cell (i, j) displays the corresponding local score $\hat{S}_{ij}$, which can be calculated from a recursion similar to equation (3) (Smith and Waterman 1981). A local score of 0 indicates that no path of positive weight ends at the corresponding point. The optimal local path ending at the point $(6, 7)$, e.g., consists of 7 edges, in the following order: 2 northeast, 1 north of length 2, and 3 northeast. The optimal local path corresponds to the local sequence alignment of, e.g., $\mathbf{A}[0, 10]$, and $\mathbf{B}[0, 10]$, AC− −TAG and ACGGTAG. The local score $\hat{S}_{6,7} = 5 + 5 - 7 + 5 + 5 + 5 = 18$ is the sum of the corresponding edges and represents the path of greatest weight ending at $(6, 7)$. The score is also the greatest weight of any path within the rectangle displayed, so it also corresponds to the local maximum score $\hat{M}_{10,10} = 18$.

percolation time. Under certain conditions, the distribution of the random local maximum $\hat{M}_{mn}$ approximates the following Gumbel extreme value distribution (Galombos 1978, Aldous 1989):

$$\mathbb{P}(\hat{M}_{mn} > y) \approx 1 - \exp[-Kmn \exp(-\lambda y)]. \tag{1}$$

The Gumbel distribution in equation (1) has 'location parameter' K and 'scale parameter' λ.

By default, BLASTP, the popular BLAST computer program for searching protein databases, uses the BLOSUM62 scoring matrix, the affine gap penalty $\Delta(g) = 11 + g$, and the empirically determined Robinson and Robinson amino acid frequencies for its random letter frequencies. Our simulations generally adhered to the BLASTP defaults, for which the Gumbel parameters are known to extraordinary accuracy: $\lambda \approx 0.267$ and $K \approx 0.041$ (Altschul *et al* 2001).

The Gumbel distribution in equation (1) has been the subject of intense research effort for about 15 years. For ungapped local alignment (i.e., for $\Delta(g) \equiv \infty$), a rigorous proof of equation (1) yields formulae for the Gumbel parameters λ and K (Dembo *et al* 1994). For gapped local alignment, few rigorous results are available, although some approximate analytical studies are extant (Mott 1999, Mott 2000, Siegmund and Yakir 2000, Storey and Siegmund 2001). In the absence of a rigorous theory for gapped local alignment, computer simulations confirm the validity of equation (1) (Mott 1992, Waterman and Vingron 1994, Altschul and Gish 1996, Olsen *et al* 1999); in the absence of formulae, they also provide estimates of λ and K (Smith *et al* 1985, Collins *et al* 1988, Mott 1992, Mott and Tribe 1999).

Presently, the BLAST program offers its users a narrow choice indeed in their alignment parameters, because it needs to pre-compute the Gumbel parameters λ and K offline. If λ and K could be computed online (in, say, less than 1 s) before searching a database, a user's alignment parameters could be arbitrary. Recently, much research has been directed accordingly, towards speeding the estimation of the Gumbel parameters.

Several methods of estimating λ and K improve on crude simulation of local alignments, e.g., the '*declumping method*' (Waterman and Vingron 1994) and the '*island method*' (Olsen *et al* 1999). Both methods are based on '*islands*', a concept explained elsewhere with mathematical rigor (Spouge 2004). Unfortunately, the methods still require minutes to estimate the Gumbel parameters, too slow for online computation.

BLAST uses only large values of y for the tail in equation (1), so errors in λ have a much greater practical impact than errors in K. The relative errors in λ generally must be less than 1% to 4%; the relative errors in K, less than about 10%. The rigorous theory of global alignment (Arratia and Waterman 1994) shows, however, that

$$\lambda = \lim_{y \to \infty} \frac{-\ln \mathbb{P}(M > y)}{y}. \tag{2}$$

Thus, if K is considered inessential, simulations of global alignment can determine λ alone (Bundschuh 2002b). At comparable accuracies for λ, global alignments use sequence lengths $m = n \approx 100$; local alignments, $m = n \approx 600$ (Altschul *et al* 2001). Consequently, Bundschuh's idea of using global alignment to estimate λ speeds computation by a factor of at least 5. Bundschuh also makes in passing several interesting conjectures about gapped global alignment.

This paper modifies Bundschuh's estimate for λ by modelling gapped global alignment heuristically as a Markov additive process (MAP) (Cinlar 1975, Asmussen 2003). Although it is really just a loose analogy, the heuristic MAP model generalizes Bundschuh's estimate for λ, reducing noticeably the sequence length required for a given accuracy in λ. It also gives some quantitative insight into Bundschuh's conjectures.

Sequence alignment is a physical model for the annealing of nucleic acids and has mathematical analogies to familiar models of condensed matter like first-passage percolation or interacting particle systems (Bundschuh 2002a, Uchiyama *et al* 2004). MAP analogies might be extended to these models, as well. In addition, our statistical techniques have some general interest. We need to extract an asymptotic parameter, the Gumbel scale parameter λ, with an *asymptotic regression*. Usually, asymptotic parameters are extracted from simulation results only after establishing arbitrary cut-offs, to ensure that a regression considers only data from the asymptotic regime. In extracting λ here, however, we use the statistical technique of robust regression, which removes any need for arbitrary cut-offs. Our results suggest that robust regression might be generally useful for asymptotic regression.

The layout of this paper is as follows. Section 2 presents our methods. Section 2.1 gives our algorithm for computing the global score S_{ij}. Section 2.2 heuristically models global gapped alignment as a Markov additive process (MAP). The MAP model suggests several new equations for the Gumbel scale parameter λ. Section 2.3 solves the new equations and estimates the simulation error in their roots. Section 2.4 indicates that robust regression can extract the asymptotic parameter λ from a series of finite estimates for λ. Section 3 presents numerical results; and section 4, our discussion.

Generally, we present the three components of our paper (sequence alignment, Markov additive processes and robust asymptotic regression) as independently as possible.

2. Methods

2.1. The standard global sequence alignment algorithm for affine gaps

For affine gaps $\Delta(g) = a + bg$, the global score can be calculated with the recursion

$$S_{ij} = \max\{S_{i-1,j-1} + S(A_i, B_j), C_{ij}, D_{ij}\}, \tag{3}$$

where

$$C_{ij} = \max\{S_{i,j-1} - a - b, C_{i,j-1} - b\}, \qquad D_{ij} = \max\{S_{i-1,j} - a - b, D_{i-1,j} - b\}$$

and boundary conditions

$$S_{00} = 0, S_{i0} = S_{0i} = -\Delta(i) \quad \text{for} \quad i > 0, \qquad C_{i0} = D_{0i} = -\infty \quad \text{for} \quad i, j \geqslant 0$$

(Waterman 1995).

2.2. Markov additive process

Here, we present Markov additive processes (MAP) as a heuristic model for global sequence alignment. The MAP model suggests several generalizations of Bundschuh's estimate for λ (Bundschuh 2002b). Because rigorous, general definitions of a MAP can be found elsewhere (Asmussen 2003), the following gives a relatively informal description of the essentials for our application. We use the standard asymptotic notation $\sim$, O and o (e.g., Erdélyi (1956) p 5).

Imagine a Markov chain $\{J_n\}$ on a finite state space $\mathbb{J}$ containing $|\mathbb{J}|$ elements. Let $\mathbf{P} = \|p_{ij}\|$ denote the $|\mathbb{J}| \times |\mathbb{J}|$ transition matrix of $\{J_n\}$, so $p_{ij} = \mathbb{P}(J_{n+1} = j | J_n = i)$ for $n = 0, 1, 2, \ldots$. For simplicity, assume that $\mathbf{P}$ is strictly positive, so $p_{ij} > 0$ for all $i, j = 1, 2, \ldots, |\mathbb{J}|$. The Markov chain $\{J_n\}$ then has a stationary distribution given by a strictly positive $1 \times |\mathbb{J}|$ row vector $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi}\mathbf{1} = 1$, where $\mathbf{1}$ is the $|\mathbb{J}| \times 1$ column vector whose elements are all 1.

Run the Markov chain $\{J_n\}$, and take its states as given. Now, consider another sequence $\{Y_n\}$ of random variables, where the distribution of Y_n is determined by the transition $J_{n-1} \to J_n$ of the Markov chain, i.e., Y_m and Y_n carry the same distribution if $J_{m-1} = J_{n-1} = i$ and $J_m = J_n = j$. In that case, let the generic (doubly subscripted) random variable Y_{ij} carry the distribution that Y_m and Y_n share. (Note that although Y_m and Y_n carry the same distribution, randomness can still give them different values.) The main random variables of interest in a MAP are the sums $T_n = \sum_{m=0}^{n} Y_m$ for $n = 0, 1, \ldots$, where $Y_0 = 0$.

MAPs have the following heuristic analogy to global alignment. Identify the Markov chain states J_n in the MAP with the vertices of Γ within the rectangle $[0, n] \times [0, n]$. The sum T_n in the MAP can then be identified with either S_n or E_n, whichever happens to be more convenient. We make no pretence that the MAP analogy with global alignment is in any way precise, but it leads to some interesting conjectures.

In the MAP, define the matrix $\mathbf{P}_\theta = \|p_{ij}\mathbb{E}[\exp(\theta Y_{ij})]\|$, and let $r(\theta)$ be its spectral radius (i.e., the maximum absolute value of any eigenvalue of $\mathbf{P}_\theta$). The theory of MAPs states that the equation $r(\lambda) = 1$ has a unique positive root $\lambda > 0$, and that the maximum $M = \max_{n \geqslant 0} T_n$ satisfies the asymptotic equality $\mathbb{P}(M > y) \sim c\,\mathrm{e}^{-\lambda y}$ for some $c > 0$ as $y \to \infty$. If T_n is identified with E_n, M becomes identified with the global maximum in sequence alignment, and the Gumbel scale parameter in equation (2) becomes identified with the root λ of the equation $r(\lambda) = 1$.

We need just a little more MAP theory. Let the indicator function $\mathbb{I}[J_n = j]$ equal 1 if $J_n = j$ and 0 otherwise, so $\exp(\theta T_n)\mathbb{I}[J_n = j]$ equals $\exp(\theta T_n)$ if $J_n = j$ and 0 otherwise. Then, $\mathbb{E}\{\exp(\theta T_n)\mathbb{I}[J_n = j] | J_0 = i\}$ is the expectation of $\exp(\theta T_n)\mathbb{I}[J_n = j]$ when the starting state is $J_0 = i$. Induction shows that the matrix moment generating function $\|\mathbb{E}\{\exp(\theta T_n)\mathbb{I}[J_n = j] | J_0 = i\}\|$ equals $\mathbf{P}_\theta^n$, the nth power of the matrix $\mathbf{P}_\theta$. Thus, if the Markov chain $\{J_n\}$ starts in a state J_0 with distribution $\boldsymbol{\gamma}$, matrix algebra yields

$$\mathbb{E}[\exp(\theta T_n)] = \boldsymbol{\gamma}\mathbf{P}_\theta^n \mathbf{1}. \tag{4}$$

Now, define $K_n(\theta) := \ln(\mathbb{E}[\exp(\theta T_n)])$, the cumulant generating function of T_n. A spectral (eigenvalue) decomposition of the matrix $\mathbf{P}_\theta$ in equation (4) (Seneta 1981) suggests that

$$K_n(\theta) = n \ln\{r(\theta)\} + C_0 + O(\varepsilon^n), \tag{5}$$

where $0 \leqslant \varepsilon < 1$ is determined by the magnitude of the subdominant eigenvalue of $\mathbf{P}_\theta$, and C_0 is a constant independent of θ and n. For $0 \leqslant m < n$, we can accelerate the convergence in equation (5) as $n \to \infty$ by writing

$$\ln\left(\frac{\mathbb{E}[\exp(\theta T_n)]}{\mathbb{E}[\exp(\theta T_m)]}\right) = K_n(\theta) - K_m(\theta) = (n - m)\ln\{r(\theta)\} + O(\varepsilon^m). \tag{6}$$

For $m < n$, let λ_{mn} denote the root of the equation

$$\mathbb{E}[\exp(\lambda_{mn} T_n)] = \mathbb{E}[\exp(\lambda_{mn} T_m)]. \tag{7}$$

Because $r(\lambda) = 1$, a linear Taylor approximation around λ yields $\ln\{r(\lambda_{mn})\} \approx r'(\lambda)(\lambda_{mn} - \lambda)$, so equation (6) becomes

$$(n - m)r'(\lambda)(\lambda_{mn} - \lambda) = O(\varepsilon^m), \tag{8}$$

i.e., with $n - m$ fixed, λ_{mn} converges geometrically to λ as $m \to \infty$.

2.3. Numerical schemes for estimating λ

Denote λ_{mn} by $\lambda_{mn}^{[S]}$ (or $\lambda_{mn}^{[E]}$), if T_n and T_m in equation (7) are replaced by S_n and S_m (or E_n and E_m). Bundschuh examined $\lambda_{0n}^{[S]}$ (Bundschuh 2002b), i.e., the numerical scheme corresponding to

$$\mathbb{E}[\exp(\lambda_{0n} S_n)] = 1. \tag{9}$$

He conjectured that $\lambda_{0n} = \lambda + Cn^{-1} + O(n^{-2})$, which is consistent with equation (8), where $\lambda_{0n} - \lambda = \{(n-0)r'(\lambda)\}^{-1}O(\varepsilon^0) = O(n^{-1})$ for $m = 0$.

To estimate $\mathbb{E}[\exp(\theta S_n)]$, Bundschuh used importance sampling, basing it on the known distribution of optimal subsequence pairs in so-called 'hybrid sequence alignment' (Yu and Hwa 2001). We estimate $\mathbb{E}[\exp(\theta S_n)]$ and $\mathbb{E}[\exp(\theta E_n)]$ for equation (7) with Bundschuh's importance sampling method. Details of the method can be found elsewhere (Bundschuh 2002b).

All simulations used affine gap penalties $\Delta(g) = a + bg$. For S_n, we followed Bundschuh, who recommended zero boundary conditions ($S_{ij} = 0$ for $i = 0$ or $j = 0$) to promote rapid convergence. For E_n, however, we required that $\mathbb{E}[E_n] < 0$, so we used the standard boundary conditions ($S_{i0} = S_{0i} = -a - bi$ for $i \geqslant 1$).

To estimate λ_{mn} from simulations, define $f(T_m, T_n; \theta) := \exp(\theta T_n) - \exp(\theta T_m)$, with $N^{-1}\sum_1^N f(T_m, T_n; \theta)$ being the average of $f(T_m, T_n; \theta)$ over N realizations. The random root $\theta = \Lambda_{mn}$ of $\sum_1^N f(T_m, T_n; \theta) = 0$ is our estimate for λ_{mn} from the N realizations.

Let s_{mn} be the standard error of Λ_{mn}. We calculated s_{mn} as follows. Expand $N^{-1}\sum_1^N f(T_m, T_n; \theta)$ in a Taylor series as far as the linear term and substitute $\theta = \Lambda_{mn}$:

$$0 = N^{-1}\sum_1^N f(T_m, T_n; \Lambda_{mn}) \approx N^{-1}\sum_1^N f(T_m, T_n; \lambda_{mn}) + (\Lambda_{mn} - \lambda_{mn})N^{-1}\sum_1^N f'(T_m, T_n; \lambda_{mn}). \tag{10}$$

In the final expression, approximate $N^{-1}\sum_1^N f'(T_m, T_n; \lambda_{mn}) \approx \mathbb{E}[f'(T_m, T_n; \lambda_{mn})]$, introducing a relative error that vanishes in probability as $N \to \infty$. Because

$$\Lambda_{mn} - \lambda_{mn} \approx -N^{-1}\frac{\sum_1^N f(T_m, T_n; \lambda_{mn})}{\mathbb{E}f'(T_m, T_n; \lambda_{mn})}, \tag{11}$$

square equation (11) and take expectations to derive the approximation in equation (12) between the second and third expressions:

$$s_{mn}^2 := \mathbb{E}(\Lambda_{mn} - \lambda_{mn})^2 \approx N^{-1}\frac{\mathbb{E}\{[f(T_m, T_n; \lambda_{mn})]^2\}}{[\mathbb{E}f'(T_m, T_n; \lambda_{mn})]^2} \approx N^{-1}\frac{\mathbb{E}\{[f(T_m, T_n; \Lambda_{mn})]^2\}}{[\mathbb{E}f'(T_m, T_n; \Lambda_{mn})]^2}. \tag{12}$$

The final approximation introduces another relative error that vanishes in probability. The final expression is estimated directly from the simulation.

2.4. Robust regression method for λ

Let $\bar{n}$ be the maximum sequence length in the simulations. We can produce estimates $\Lambda_{mn} \pm s_{mn}$ of λ_{mn} for $0 \leqslant m < n \leqslant \bar{n}$. From the data $\Lambda_{mn} \pm s_{mn}$, we want to extract an overall estimate Λ_∞ for $\lambda := \lambda_\infty := \lim_{m\to\infty}\lambda_{mn}$. We could extract the estimate by establishing (e.g., by eye) some *ad hoc* cut-off $\underline{m}$, then defining Λ_∞ as some weighted average of $\{\Lambda_{mn} : \underline{m} \leqslant m < n \leqslant \bar{n}\}$. Instead of introducing *ad hoc* cut-offs, however, we adapted the statistical technique of robust regression to regress the asymptotic constant λ_∞ systematically.

With an eye to generalizations in section 4, we now describe multivariate linear robust regression (Ryan 1996). Consider a regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is a matrix of independent ('regressor') variables; $\mathbf{y}$, $\boldsymbol{\beta}$ and $\mathbf{e}$ are column vectors of appropriate dimension containing the dependent ('response') variables $\mathbf{y}$, the fitted parameters $\boldsymbol{\beta}$ and the random errors ('residuals') $\mathbf{e}$. Denote the elements of, e.g., $\mathbf{y}$ by y_i. Let $s_i^2 := \operatorname{var} y_i = \operatorname{var} e_i$ be the measure of dispersion; let $e_i^* := e_i/s_i$ be the 'normalized residuals'.

Each of the three classes of robust regression estimators (*M-estimators*, *bounded influence estimators* and *high breakdown point estimators*) has its own extensive literature. Here, we consider *M-estimators* (Huber 1964, 1973). Given some 'criterion function' ρ (described further below), the *M-estimator* $\hat{\boldsymbol{\beta}}$ solves the minimization problem $\min \sum_i \rho(e_i^*)$, i.e.,

$$\hat{\boldsymbol{\beta}} = \arg\min \sum_i \rho(e_i^*). \tag{13}$$

The global minimum in equation (13) can be found by general global minimization methods or by iteratively reweighted least-squares methods (Ryan 1996). Statistics favours the iterative methods, because as a by-product, they estimate the covariance matrix $\operatorname{cov} \hat{\boldsymbol{\beta}} = [\operatorname{cov}(\hat{\beta}_i, \hat{\beta}_j)]$, thereby assigning an error to every estimate. Because they can be trapped in local minima, however, they require a judicious initial guess for $\hat{\boldsymbol{\beta}}$. We do not need iterative methods here, because equation (12) already estimates our errors.

The derivative $\psi = \rho'$ is called the 'influence function'. Ordinary weighted least squares, e.g., correspond to the linear influence function $\psi(e^*) = e^*$. Many influence functions have been proposed (Montgomery *et al* (2001) p 388). The Andrews function is particularly well suited to asymptotic regression: it is $\psi(e^*) = \sin(e^*/a)$ for $e^* \in [-a\pi, a\pi]$ and 0 otherwise, typically with $a = 1.5$ or 2 (Andrews 1974). For asymptotic regression, the Andrews function with $a = 2$ was superior to the other established influence functions we tried (data not shown).

The superiority was predictable. Given a close approximation to the true regression line, if a normalized residual is large, the corresponding point is probably not in the asymptotic regime. Because the point's apparent bias greatly exceeds its estimated error, it should not be permitted to influence the asymptotic regression. In other words, if $e_i^* > 2\pi$, e.g., the point's influence should be $\psi(e_i^*) = 0$, as in the Andrews function.

The criterion $e_i^* > 2\pi$ for $a = 2$ gives large biases some influence on the regression. Smaller values of a considerably roughen the graph of $\sum_i \rho(e_i^*)$, however, generating many local minima. For our purposes, the value $a = 2$ seemed best (data not shown).

Since we wished to extract an estimate Λ_∞ for $\lambda := \lambda_\infty := \lim_{m\to\infty} \lambda_{mn}$ from the data $\Lambda_{mn} \pm s_{mn}$, we considered the (trivial) regression model $\mathbf{y} = \mathbf{1}\beta + \mathbf{e}$, where $\mathbf{y}$ is a column vector consisting of the Λ_{mn} in any order, $\mathbf{1}$ is a column vector whose elements are all 1, and $\beta = \Lambda_\infty$ is the estimated Gumbel scale parameter λ_∞. To avoid local minima, we derived the robust estimate Λ_∞ from equation (13) by direct global minimization (i.e., we formed a fine mesh and tested the function values at each mesh point).

3. Numerical results

Figures 2 and 3 display some estimates of the Gumbel scale parameter λ in gapped local alignment. The figures show results for the BLOSUM62 scoring matrix with the usual affine gap cost $\Delta(g) = 11 + g$ for a gap of length g. Other common scoring matrices gave similar results (data not shown).

Figure 2 compares the estimates for $\lambda_{(n-5)n}^{[S]}$ and $\lambda_{(n-5)n}^{[E]}$ from equation (7) to the estimate for $\lambda_{0n}^{[S]}$ from Bundschuh's equation (9). The lag of 5 in $\lambda_{(n-5)n}$ was the best compromise between rapid, geometric convergence and small standard errors (data not shown). Figure 2 displays the standard errors $s_{(n-5)n}$ in equation (12) as error bars. It plots the estimates for $\lambda_{(n-5)n}^{[S]}$ and $\lambda_{(n-5)n}^{[E]}$ against the sequence length n of the global alignments, up to $n = 40$, every point representing 1000 000 realizations. The horizontal line represents the previous

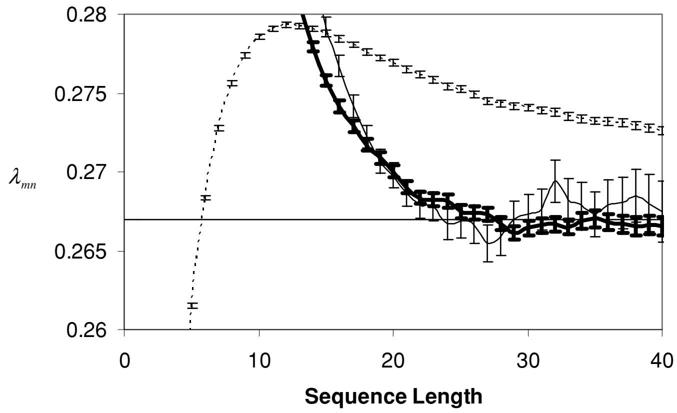


Figure 2. Plot of estimates for $\lambda^{[S]}_{0n}$, $\lambda^{[S]}_{(n-5)n}$, $\lambda^{[E]}_{(n-5)n}$ against sequence length n for the BLOSUM62 scoring matrix with an affine gap cost of 11 + g for a gap of length g, with random sequences whose letters are chosen according to the empirical Robinson amino acid frequency (Robinson and Robinson 1991). Each point represents 1000 000 random sequence pairs generated by importance sampling (Bundschuh 2002b). The error bars indicate standard errors. The dotted line indicates estimates of $\lambda^{[S]}_{0n}$; a solid line, $\lambda^{[S]}_{(n-5)n}$; and a thick solid line, $\lambda^{[E]}_{(n-5)n}$. The horizontal line $\lambda = 0.267$ represents the previous best estimate of the asymptotic constant λ (Altschul *et al* 2001).

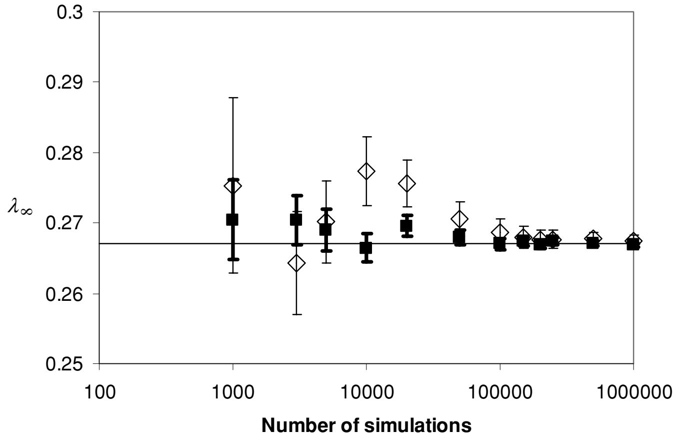


Figure 3. Plot of estimates λ_∞ obtained via the robust regression method using all $\Lambda_{mn} \pm s_{mn}$ for $0 \leqslant m < n \leqslant 50$ against different simulation numbers. The horizontal line $\lambda = 0.267$ again indicates the previous best estimate. The robust regression estimate λ_∞ using $\Lambda^{[S]}_{mn}$ is shown with ◇ and solid error bar; λ_∞ using $\Lambda^{[E]}_{mn}$, with ■ and thick solid line error bar.

best estimate $\lambda \approx 0.267$ (Altschul *et al* 2001). Both $\lambda^{[S]}_{(n-5)n}$ and $\lambda^{[E]}_{(n-5)n}$ appear to converge geometrically to the line, as predicted in equation (8). Specifically, $\lambda^{[E]}_{(n-5)n}$ crosses the line $\lambda = 0.267$ at $n = 28$; $\lambda^{[S]}_{(n-5)n}$, at $n = 24$. In contrast, $\lambda^{[S]}_{0n}$ is still far away from the line at $n = 40$ and only crosses it at $n = 122$ (data not shown). In fact, $\lambda^{[E]}_{(n-5)n}$ appears to converge to 0.2667, which is consistent with the previous best estimate $\lambda \approx 0.267$. Because E_n incorporates more information from the simulation and so presumably has a smaller variance than S_n, $\lambda^{[E]}_{(n-5)n}$ shows less Monte Carlo fluctuation than $\lambda^{[S]}_{(n-5)n}$. Figure 2 indicates that equation (7) estimates λ accurately from sequence lengths as short as $n = 30$.

Figure 3 plots the robust regression estimates of λ_∞ with their standard error bars against different numbers N of realizations. In each case, we estimated λ_∞ from all the estimates λ_{mn} for $0 \leqslant m < n \leqslant \bar{n} = 50$, using the (trivial) regression model in section 2.4, setting $a = 2$ in Andrews' influence function. As an estimate, $\lambda^{[E]}_\infty := \lim_{m\to\infty} \lambda^{[E]}_{mn}$ is more accurate than $\lambda^{[S]}_\infty := \lim_{m\to\infty} \lambda^{[S]}_{mn}$, but as $N \to \infty$, both converged asymptotically to the Gumbel scale parameter λ. A 2.8 GHz Pentium 4 processor with 0.5 GB RAM computed $\lambda^{[E]}_\infty$ for $N = 1000$ with 1.29% relative error in 5.7 s.

Regression of all our estimates $\lambda^{[E]}_{mn}$ yielded $\lambda \approx 0.2667 \pm 0.0004$.

4. Discussion

Our paper contains two techniques of general mathematical interest. First, we used Markov additive processes (MAPs) to provide a heuristic model for global sequence alignment, a variant of first-passage percolation. Second, we used robust regression to extract asymptotic parameters without resorting to any *ad hoc* cut-offs. Here, we examine some consequences and generalizations of the two techniques.

Our first technique, the MAP heuristic, might generalize to certain types of physical models (e.g., first-passage percolation), serving to accelerate convergence to asymptotic parameters there as well. In the present context, it accelerated the convergence of Bundschuh's estimate $\lambda^{[S]}_{0n}$ of the Gumbel scale parameter $\lambda = \lambda_\infty$. His estimate $\lambda^{[S]}_{0n}$ converges harmonically, with $\lambda^{[S]}_{0n} - \lambda_\infty = O(n^{-1})$. Our MAP heuristic provided a simple rationale for introducing equation (7), based on the Perron–Frobenius theorem about the dominant eigenvalue of a positive matrix (Seneta 1981). Equation (7) then suggested an examination of our quantities $\lambda^{[S]}_{mn}$ and $\lambda^{[E]}_{mn}$. Both quantities appeared to converge geometrically, with $\lambda_{mn} - \lambda_\infty = (n-m)^{-1} O(\varepsilon^m)$ as the MAP heuristic predicts, and with $\lambda^{[E]}_{mn}$ having smaller standard error than $\lambda^{[S]}_{mn}$ in practice. At a maximum simulated sequence length of $\bar{n} = 24$, $\lambda^{[S]}_{mn}$ achieved about the same accuracy as $\lambda^{[S]}_{0n}$ achieved at $\bar{n} = 122$. Bundschuh had heuristics for speeding the computation of $\lambda^{[S]}_{0n}$ (Bundschuh 2002b), and although we did not implement them, his heuristics apply to $\lambda^{[S]}_{mn}$ and $\lambda^{[E]}_{mn}$ as well. Our summary estimate of λ, presently the most accurate one published, is $\lambda \approx 0.2667 \pm 0.0004$.

Our second technique, robust regression, can extract asymptotic parameters from simulation data without using *ad hoc* cut-offs to define an asymptotic regime. Historically, robust regression was suggested as a alternative to ordinary least-squares regression, being preferable if the errors $\mathbf{e}$ are not Gaussian, or if occasional outliers (x_i, y_i) from an unknown distribution contaminate the data (Huber 1964, 1973). In asymptotic regression, the data collected outside the asymptotic regime can be regarded as contaminating outliers. Interestingly, Internet searches on keywords like 'robust regression' and 'asymptotic analysis' did not return any examples of robust asymptotic regression. Thus, while robust regression was originally developed with other aims in mind and appears not to have been applied to asymptotic regression, the results suggest that it is admirably suited to this application.

Consider a general asymptotic regression problem: some simulation data points $(x_i, y_i \pm s_i)$ $(i = 0, \ldots, m)$ are collected to fit an asymptotic series $y := y(x) = \sum_{j=0}^{n} \phi_j(x)\beta_j + e$, where $\phi_{i+1}(x) = o\{\phi_i(x)\}$ $(j = 0, \ldots, n)$ and $e = O\{\phi_{n+1}(x)\}$ (Erdélyi 1956). Define the $m \times n$ matrix $\mathbf{X} = [\phi_j(x_i)]$ and apply the robust regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with the Andrews influence function and $a = 2$. If a residual $e_i = O\{\phi_{n+1}(x_i)\}$ exceeds 2π times the simulation spread s_i, the point $(x_i, y_i \pm s_i)$ has no influence in robust regression, in accord with *ad hoc* intuition. Step-wise robust asymptotic regressions, where $\{\beta_1, \ldots, \beta_j\}$ is estimated before β_{j+1}, are also possible.

In a typical robust regression, the spread s_i in a data point $(x_i, y_i \pm s_i)$ is estimated with a median error rather than a standard error (Montgomery *et al* 2001), because medians are less sensitive to an occasional contaminating outlier point. In asymptotic regression, however, *every* data point (x, y) with a particular abscissa $x = x_i$ has an identical bias. Thus, the median and standard error defend equally well against these asymptotic biases, i.e., not at all. In our application, moreover, the median error was expensive to compute. It required partitioning the realizations from the simulation into, e.g., k equal subsamples, and then estimating $\lambda_{mn(i)}$ for each subsample $(i = 1, \ldots, k)$. With $\bar{\lambda}_{mn} := k^{-1} \sum_{i=1}^{k} \lambda_{mn(i)}$, median errors $s_{mn} := k^{-1/2}\, \text{median}_{i=1,\ldots,k} |\bar{\lambda}_{mn} - \lambda_{mn(i)}|$ can be estimated. Standard errors can be estimated analogously. In our application, the approximate standard error from equation (12) required less computation and performed as well as subsample estimates of the median or standard error. In robust asymptotic regression, we saw no reason to prefer any measure of spread to our approximate standard error. Thus, in our application, the only specific adaptation we made to robust asymptotic regression was measuring spreads with the approximate standard error.

The typical robust regression algorithm performs sequential weighted least-squares regressions, iteratively adjusting regression weights according to the current regression estimates. The iterations can easily contain specific adaptations to asymptotic regression. Given the current regression estimates, iterations typically adjust the weights independently for each data point. Robust asymptotic regression might therefore be improved by grouping points together to reduce the regression influence outside asymptotic regions. Section 2.4 (in the discourse about the influence $\psi(e_i^*) = 0$ under the criterion $e_i^* > 2\pi$) hints that specific adaptations might indeed lead to improvements.

Finally, this paper used *global* sequence alignments to estimate the scale parameter λ from the Gumbel distribution in equation (1) for *local* alignment. We also have methods for estimating the Gumbel location parameter K from *global* sequence alignments (manuscript in preparation). Finally, no rigorous mathematical proof for the Gumbel distribution in equation (1) is known. We speculate that MAPs will be instrumental in such a proof.

References

Aldous D 1989 *Probability Approximations via the Poisson Clumping Heuristic* (New York: Springer)

Altschul S F, Bundschuh R, Olsen R and Hwa T 2001 The estimation of statistical parameters for local alignment score distributions *Nucleic Acids Res.* **29** 351–61

Altschul S F and Gish W 1996 Local alignment statistics *Methods Enzymol.* **266** 460–80

Altschul S F, Gish W, Miller W, Myers E W and Lipman D J 1990 Basic local alignment search tool *J. Mol. Biol.* **215** 403–10

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

Andrews D F 1974 A robust method for multiple linear regression *Technometrics* **16** 523–31

Arratia R and Waterman M S 1994 A phase transition for the score in matching random sequences allowing deletions *Ann. Appl. Probab.* **4** 200–25

Asmussen S 2003 *Applied Probability and Queues* (New York: Springer)

Bundschuh R 2002a Asymmetric exclusion process and extremal statistics of random sequences *Phys. Rev.* E **65** 031911

Bundschuh R 2002b Rapid significance estimation in local sequence alignment with gaps *J. Comput. Biol.* **9** 243–60

Cinlar E 1975 *Introduction to Stochastic Processes* (Upper Saddle River, NJ: Prentice-Hall)

Collins J F, Coulson A F and Lyall A 1988 The significance of protein sequence similarities *Comput. Appl. Biosci.* **4** 67–71

Dayhoff M O, Schwartz R M and Orcutt B C (ed) 1978 A model of evolutionary change in proteins *Atlas of Protein Sequence and Structure* (Silver Spring, MD: National Biomedical Research Foundation) pp 345–52

Dembo A, Karlin S and Zeitouni O 1994 Limit distributions of maximal non-aligned two-sequence segmental score *Ann. Probab.* **22** 2022–39

Erdélyi A 1956 *Asymptotic Expansions* (New York: Dover)

Galombos J 1978 *The Asymptotic Theory of Extreme Order Statistics* (New York: Wiley)

Henikoff S and Henikoff J G 1992 Amino acid substitution matrices from protein blocks *Proc. Natl Acad. Sci. USA* **89** 10915–9

Huber P J 1964 Robust estimation of a location parameter *Ann. Math. Stat.* **35** 73–101

Huber P J 1973 Robust regression: asymptotics, conjectures and Monte Carlo *Ann. Stat.* **1** 799–821

Montgomery D C, Peck E A and Vining G G 2001 *Introduction to Linear Regerssion Analysis* (New York, NY: Wiley)

Mott R 1992 Maximum-likelihood-estimation of the statistical distribution of Smith–Waterman local sequence similarity scores *Bull. Math. Biol.* **54** 59–75

Mott R 1999 Local sequence alignments with monotonic gap penalties *Bioinformatics* **15** 455–62

Mott R 2000 Accurate formula for p-values of gapped local sequence and profile alignments *J. Mol. Biol.* **300** 649–59

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

Needleman S B and Wunsch C D 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins *J. Mol. Biol.* **48** 443–53

Olsen R, Bundschuh R and Hwa T 1999 Rapid assessment of extremal statistics for gapped local alignment *Proc. 7th Int. Conf. Intelligent Systems for Molecular Biology* pp 211–22

Robinson A B and Robinson L R 1991 Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins *Proc. Natl Acad. Sci. USA* **88** 8880–4

Ryan T P 1996 *Modern Regression Methods* (New York: Wiley)

Schaffer A A, Aravind L, Madden T L, Shavirin S, Spouge J L, Wolf Y I, Koonin E V and Altschul S F 2001 Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements *Nucleic Acids Res.* **29** 2994–3005

Seneta E 1981 *Non-Negative Matrices and Markov Chain* (New York: Springer)

Siegmund D and Yakir B 2000 Approximate p-values for local sequence alignments *Ann. Stat.* **28** 657–80

Smith T F and Waterman M S 1981 Identification of common molecular subsequences *J. Mol. Biol.* **147** 195–7

Smith T F, Waterman M S and Burks C 1985 The statistical distribution of nucleic acid similarities *Nucleic Acids Res.* **13** 645–56

Spouge J L 2004 Path reversal and islands in the gapped alignment of random sequences *J. Appl. Probab.* at press

Storey J D and Siegmund D 2001 Approximate p-values for local sequence alignments: numericalx studies *J. Comput. Biol.* **8** 549–56

Uchiyama M, Sasamoto T and Wadati M 2004 Asymmetric simple exclusion process with open boundaries and Askey–Wilson polynomials *J. Phys. A: Math. Gen.* **37** 4985–5002

Waterman M S 1995 *Introduction to Computational Biology* (London/Boca Raton, FL: Chapman and Hall/CRC)

Waterman M S, Gordon L and Arratia R 1987 Phase transitions in sequence matches and nucleic acid structure *Proc. The National Academy of Sciences of The United States of America* vol 84, pp 1239–43

Waterman M S and Vingron M 1994 Rapid and accurate estimates of statistical significance for sequence data base searches *Proc. Natl Acad. Sci. USA* **91** 4625–8

Yu Y K and Hwa T 2001 Statistical significance of probabilistic sequence alignment and related local hidden Markov models *J. Comput. Biol.* **8** 249–82