JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 5, 2001
Mary Ann Liebert, Inc.
Pp. 549–556

Approximate P-Values for Local Sequence Alignments: Numerical Studies

JOHN D. STOREY and DAVID SIEGMUND

ABSTRACT

Siegmund and Yakir (2000) have given an approximate p-value when two independent, identically distributed sequences from a finite alphabet are optimally aligned based on a scoring system that rewards similarities according to a general scoring matrix and penalizes gaps (insertions and deletions). The approximation involves an infinite sequence of difficult-to-compute parameters. In this paper, it is shown by numerical studies that these reduce to essentially two numerically distinct parameters, which can be computed as one-dimensional numerical integrals. For an arbitrary scoring matrix and affine gap penalty, this modified approximation is easily evaluated. Comparison with published numerical results show that it is reasonably accurate.

Key words: local alignment, affine gap penalty, p-value, Markov renewal theory.

1. INTRODUCTION

Comparison of a new gene (DNA sequence) or protein (amino acid sequence) with the existing sequences in a database can be an important first step in learning the structure and/or function of the new sequence. Local pairwise alignment of sequences plays an important role in such database searches.

The quality of an alignment is determined by the total similarity score of letters at aligned positions and the number of gaps (insertions/deletions) in each sequence. (These concepts will be defined precisely in what follows.) Smith and Waterman (1981) and Altschul *et al.* (1990) have developed rapid and effective computational algorithms for local pairwise comparison of DNA or amino acid sequences. In addition, one would like to evaluate the statistical significance of sequences showing a particular level of similarity (e.g., Arratia *et al.*, 1990; Dembo *et al.*, 1994; Altschul and Gish, 1996).

Approximate p-values when gaps are not allowed have been given by a number of authors, e.g., Arratia *et al.* (1990), for a special scoring function and more generally by Dembo *et al.* (1994). The form of the approximation in the ungapped case has been conjectured to be valid also in the gapped case (Waterman and Vingron, 1994) and has been empirically fit to simulated and to real data to obtain values for parameters in the approximating formula (cf. Waterman and Vingron, 1994; Altschul and Gish, 1996). While these approximations seem to be reasonably accurate, development of each approximation requires substantial statistical analysis and computation that go beyond the basic data of the problem: the scoring matrix and the gap costs. More recent results in this direction have been obtained by Altschul *et al.* (2001).

Department of Statistics, Stanford University, Stanford, CA 94305.

For affine gap costs, Mott and Tribe (1999) obtain a useful approximation by an interesting heuristic argument. Assuming the "gap open" penalty is sufficiently large, Siegmund and Yakir (2000) provide a somewhat different approximation, which depends on infinitely many parameters. They also conjecture that for numerical purposes a further approximation depending on two of these parameters would suffice. The purpose of this paper is to show that a) this conjecture is correct, b) the modified approximation is reasonably accurate, and c) it can be rapidly computed directly from basic data of the problem. In addition to better theoretical understanding of how the p-value depends on the parameters of the problem, this approximation requires effectively zero computing time, so it could be implemented on line for a scoring matrix of the user's choice.

Section 2 gives our notation and states the formal approximation of Siegmund and Yakir (2000), which depends on the parameters $\lambda_0, \lambda_1, \cdots$. Section 3 explains the probabilistic meaning of these parameters. Section 4 contains simulations based on the representations of Section 3, which show for several common scoring matrices that the λ_r for $r \geq 1$ are essentially constant. Using the two parameters λ_0 and $\Lambda = \lim_r \lambda_r$, one obtains a much simpler approximation, which is shown numerically to be in reasonable agreement with the approximation of Atlschul and Gish (1996) based on their empirically estimated parameters. Convenient integral representations for λ_0, Λ are given in an appendix.

2. NOTATION AND APPROXIMATION

Consider two finite sequences $\mathbf{x}$ and $\mathbf{y}$ from a finite alphabet. Thus, $\mathbf{x} = x_1 x_2 \cdots x_m$ and $\mathbf{y} = y_1 y_2 \cdots y_n$ with $x_i, y_j \in \mathcal{A}$. We assume throughout that $x_1, \ldots, x_m$ are independently distributed with $\mathrm{P}_0\{x_i = \alpha\} = \mu_\alpha$ for all i; and similarly $y_1, \ldots, y_n$ are independently distributed with $\mathrm{P}_0\{y_j = \beta\} = \nu_\beta$ for all j. Moreover, the x's and y's are independent.

A *candidate alignment*, $\mathbf{z} = \{(i_t, j_t) : 1 \leq t \leq k\}$, for some $1 \leq i_1 < i_2 < \cdots < i_k \leq m$ and $1 \leq j_1 < j_2 < \cdots < j_k \leq n$, specifies that x_{i_t} and y_{j_t} are *aligned* for all $t = 1, \cdots, k$. The other x's with subscripts between i_1 and i_k and the other y's with subscripts between j_1 and j_k are said to be *unaligned*. Note that there may be other letters, at the beginning or end of the two sequences, that are neither aligned nor formally designated as unaligned. We assume that either $i_{t+1} = i_t + 1$ or $j_{t+1} = j_t + 1$ for all $1 \leq t < k$; i.e., there can be unaligned letters in only one sequence at a time.

With each candidate alignment $\mathbf{z}$ we associate a score $S_\mathbf{z} = S_\mathbf{z}(\mathbf{x}, \mathbf{y})$. Aligned letters x_i and y_j are scored according to a similarity matrix $K(x_i, y_j)$. We assume a penalty of δ for each unaligned letter. The total number of unaligned letters is $l = (i_k - i_1 - k + 1 + j_k - j_1 - k + 1)$. A gap is the interval of unaligned letters that begins with a value $t + 1$ such that $i_t = j_t$ and either $i_{t+1} > i_t + 1$ or $j_{t+1} > j_t + 1$ and ends with the next aligned pair after (x_{i_t}, y_{j_t}). Each gap is assessed a cost Δ in addition to the cost δ of each unaligned letter. Frequently one refers to Δ as the "gap open" and δ as the "gap extension" cost. The score of the candidate alignment $\mathbf{z}$ is $S_\mathbf{z} = S_\mathbf{z}(\mathbf{x}, \mathbf{y}) = \Sigma_{t=1}^k K(x_{i_t}, y_{j_t}) - \Delta j - \delta l$, where j is the number of gaps and l is the total number of unaligned letters, or equivalently the total length of all gaps. Note that there are no costs assessed for letters that are neither aligned nor unaligned.

Within the collection $\mathcal{Z}$ of all candidate alignments, one can identify the best alignment—the one with the highest score. The p-value of the best score under the null assumption that the sequences $\mathbf{x}$ and $\mathbf{y}$ are independent random samples from the given alphabet is

$$\mathrm{P}_0(\max_{\mathbf{z} \in \mathcal{Z}} S_\mathbf{z} \geq b), \tag{1}$$

where b is the observed value of the score for the best alignment and P_0 is the null probability described above.

We assume that

$$\mathrm{E}_0 K(x, y) < 0 \text{ and } \mathrm{P}_0\{K(x, y) > 0\} > 0. \tag{2}$$

Let $\psi(\theta) = \log \mathrm{E}_0 \exp[\theta K(x, y)]$. Then $\exp[\theta K(\alpha, \beta) - \psi(\theta)]\mu_\alpha \nu_\beta$ defines an exponential family of probabilities indexed by θ, which for $\theta = 0$ reduces to the original probability $Q_0(\alpha, \beta) = \mu_\alpha \nu_\beta$.

It follows from (2) by convexity that there is a unique value $\theta^* > 0$ for which $\psi(\theta^*) = 0$. Let $\mu_0 = \psi'(\theta^*) > 0$.

Let $Q_1(\alpha, \beta) = \exp[\theta^* K(\alpha, \beta)] Q_0(\alpha, \beta)$ and note that $\mu_0 = \mathrm{E}_1 K(x, y)$. Also let $Q_{i,j}$ be defined to be the product probability that gives x the marginal distribution it has under Q_i and y the marginal distribution it has under Q_j. We assume that

$$\mathrm{E}_1 K(x, y) - \mathrm{E}_{i,j} K(x, y) > 0 \tag{3}$$

for all i, j. This condition is easily checked and excludes the case that $K(\alpha, \beta)$ is a sum of a function of α and a function of β.

We shall require the technical assumptions that $\min(m, n)/b \to \infty$ and that for some $0 < \epsilon < 1$, $mn\exp\{-(\theta^* b)^{1-\epsilon}\}$ is bounded. The second assumption puts the probability of interest into the domain of large deviations.

To simplify the presentation of the main approximation, we assume that $K(x, y)$ is a nonarithmetic random variable and discuss the more complicated arithmetic case below. The main result of Siegmund and Yakir (2000) is as follows.

Theorem 1. *Assume conditions (2) and (3) hold and that*

$$\theta^* \Delta = \log(\theta^* b) + C \tag{4}$$

for some constant C. There exist parameters $\lambda_0, \lambda_1, \cdots$ *in* (0,1] *such that as* $b \to \infty$,

$$\mathrm{P}_0\left(\max_{\mathbf{z} \in \mathcal{Z}} S_{\mathbf{z}} \geq b\right) \sim mne^{-\theta^* b}(\theta^* \mu_0)^{-1} \lambda_0^2 \sum_{j=0}^{\infty} [(2be^{-\theta^* \Delta}/\mu_0)^j / j!] \sum_{l=j}^{\infty} e^{-\theta^* \delta l} \sum \lambda_1^{i_1} \cdots \lambda_{l-j+1}^{i_{l-j+1}}, \tag{5}$$

where the innermost summation extends over the $\binom{l-1}{j-1}$ *terms having* $i_1 + \cdots + i_{l-j+1} = j$ *and* $i_1 + 2i_2 + \cdots + (l - j + 1)i_{l-j+1} = l$. *An upper bound for the indicated sums is* $1/[\exp(\theta^* \delta) - 1]$.

Remarks. (i) A principal consequence of the numerical studies reported below is that the constants λ_r for $r \geq 1$ are essentially all equal, say, to a single parameter Λ. Hence the right hand side of (5) can reasonably be approximated by the much simpler expression

$$mn(\theta^* \mu_0)^{-1} \lambda_0^2 \exp(-\theta^* b\{1 - 2(\theta^* \mu_0)^{-1} \Lambda e^{-\theta^* \Delta}/[e^{\theta^* \delta} - 1]\}). \tag{6}$$

(ii) In applications, the random variables $K(x_i, y_j)$ will be arithmetic of span h, typically $h = 1$. In this case it does not seem possible to give a precise asymptotic expression for the desired probability, which can be shown to lie asymptotically in an interval bounded below by $h\theta^*/(e^{h\theta^*} - 1)$ and above by $(1 - e^{-h\theta^*})/(h\theta^*)$ times the quantitity on the right hand side of (5).

(iii) As explained by Siegmund and Yakir (2000) the condition (4) can be viewed as a diagnostic for the accuracy of the approximation (5). If we suppose that b and Δ are given and (4) defines the value of C, then one should be careful in applying (5) when C is small, say, $C < -1.75$.

3. THE CONSTANTS λ_r, $r = 0, 1, 2, \cdots$

In the rest of this paper we assume that $K(x_i, y_j)$ is arithmetic with span $h = 1$.

The constant λ_0 appears in the approximation of Dembo *et al.* (1994), where gaps are not permitted. It has the same probabilistic meaning as the other λ_r, but it is defined relative to a random walk process rather than a Markov chain and consequently is easily evaluated numerically. It has been thoroughly discussed in the context of sequential analysis (e.g., Woodroofe, 1982; Siegmund, 1985).

To be more precise and prepare for the more complex case of $\lambda_r (r \geq 1)$ given below, let P_1 denote the probability under which $(x_1, y_1), (x_2, y_2), \cdots$ are independent and identically distributed with the distribution Q_1 defined following display (2). The log likelihood ratio of the first n pairs (x_i, y_i) under P_1

relative to P_0 is $\ell_n = \theta^* S_n$, where $S_n = \sum_1^n K(x_i, y_i)$. It follows from the arguments of Siegmund and Yakir (2000) that

$$\lambda_0 = \lim_N N^{-1} E_0[\exp(\max_{n \le N} \ell_n)], \tag{7}$$

which after summation by parts and a change of measure equals

$$\lim N^{-1}(1 - e^{-\theta^*}) \sum_j E_1[\exp\{-\theta^*(S_{\tau(j)} - j)\}; \tau(j) \le N], \tag{8}$$

where $\tau(j) = \min\{n : S_n \ge j\}$ for $j \ge 1$ (and $\tau(0)$ is similarly defined with a strict inequality). Since $\mu_0 = E_1 K(x_1, y_1) > 0$, by the strong law of large numbers $P_1\{\tau(j) \le N\}$ converges to 1 for $j < (1-\epsilon)N\mu_0$ and to 0 for $j > (1+\epsilon)N\mu_0$ for arbitrary $0 < \epsilon < 1$. It follows from the renewal theorem that

$$\nu_0 = \lim_j E_1[\exp\{-\theta^*(S_{\tau(j)} - j)\}] \tag{9}$$

exists. From (5)–(7) and elementary analysis one obtains

$$\lambda_0 = \mu_0(1 - e^{-\theta^*})\nu_0. \tag{10}$$

In addition, ν_0 can be expressed in terms of the distribution of the ladder variables $(\tau(0), S_{\tau(0)})$ (e.g., Siegmund, 1985, Chapter 8) and then can be evaluated numerically as a one-dimensional inverse Fourier transform (cf. Woodroofe [1982] for nonarithmetic variables and Tu and Siegmund [1999] for the arithmetic case). For completeness the result of Tu and Siegmund (1999) is stated in the appendix.

Except for the numerical evaluation of ν_0, the preceding argument applies with minor modifications to the λ_r for $r \ge 1$. Consider two sequences $\mathbf{x}$ and $\mathbf{y}$ of length N, for some integers $N \to \infty$. Given r and k, $1 \le k < N - r$, consider the alignment which matches the first k x's with the first k y's and the x's between $k+1$ and $N-r$ with the y's between $k+r+1$ and N. Let $P_k^{(r)}$ denote the probability measure that makes the aligned pairs x_i, y_j independent and identically distributed with joint distribution Q_1 and leaves the distribution of the other x_i and y_j unchanged. Let $\ell_k^{(r)}$ denote the log likelihood ratio of $P_k^{(r)}$ relative to $P_0^{(r)}$, so $\ell_k^{(r)} = \theta^*[\sum_1^k K(x_i, y_i) + \sum_{k+1}^{N-r} K(x_i, y_{i+r})]$ and $\ell_k^{(r)} - \ell_1^{(r)} = \theta^* \sum_2^k [K(x_i, y_i) - K(x_i, y_{i+r})]$. Siegmund and Yakir (2000) obtain the representations (cf. (7))

$$\lambda_r = \lim_N N^{-1} E_0^{(r)} \left[\exp(\max_{k \le N} \ell_k^{(r)})\right] = \lim_N N^{-1} E_1^{(r)} \left[\exp(\max_{k \le N} \{\ell_k^{(r)} - \ell_1^{(r)}\})\right]. \tag{11}$$

Let $P^{(r)}$ denote the extension of $P_{N-r}^{(r)}$ to the infinitely long sequence $\{x_i, y_i, 1 \le i < \infty\}$. Note that $\zeta_k^{(r)} = \ell_k^{(r)} - \ell_{k-1}^{(r)} = \theta^*[K(x_k, y_k) - K(x_k, y_{k+r})]$ is a function of x_k, y_k and y_{k+r} only. Under $P^{(r)}$ the $\zeta_k^{(r)}$ are identically distributed; and $\zeta_k^{(r)}$ is independent of any $\zeta_l^{(r)}$ for l such that $(l \bmod r) \ne (k \bmod r)$. For each $0 \le i < r$, the process $\{\zeta_k^{(r)} : 1 \le k \le N - r, (k \bmod r) = i\}$ is a first order Markov chain. Thus $\ell_k^{(r)} - \ell_1^{(r)} = \sum_2^k \zeta_i^{(r)}$ is an additive functional of an rth order Markov chain.

By (11) and the argument leading from (7) to (10), we see that

$$\lambda_r = \mu_r(1 - e^{-\theta^*})\nu_r, \tag{12}$$

where $\mu_r = E^{(r)}[K(x_2, y_2) - K(x_2, y_{2+r}]$ is constant in r, and the limit defining ν_r (exactly as in (9)) is guaranteed to exist by the renewal theorem for additive functionals of a Markov chain applied to the process of ladder variables, e.g., Athreya *et al.* (1978).

Although we obtain a representation structurally identical to (10), numerical computations are much more complicated than for a random walk. Asmussen (1989) has used a representation of the limit defining ν_r in terms of ladder variables to evaluate such a parameter for a similar process. The problem is also discussed by Karlin and Dembo (1992), although they do not appear to have developed a numerical algorithm. In the

following section we use the representation in (11) for simulating ν_r. In this regard, note that simulation requires only the generation of independent, identically distributed, discrete, random variables, while an analytic approach must deal with the Markovian dependence described above. From the probabilistic meaning of the ν_r, it seems natural to conjecture that they are effectively constant in $r \geq 1$; and since $\zeta_2^{(r)}, \cdots, \zeta_{r+1}^{(r)}$ are independent and identically distributed, with a distribution not depending on r, the limit of ν_r as $r \to \infty$ can be evaluated in terms of a random walk (albeit a different random walk from that entering into ν_0 above).

4. NUMERICAL RESULTS

Our goal in this section is two-fold: (i) to show by simulation that it is reasonable to regard the constants λ_r for $r \geq 1$ as a single constant, so that the complicated approximations given in Theorem 1 can be replaced by the much simpler expresson (6), which is very easy to evaluate numerically via the algorithm given in the appendix; and (ii) to compare the results of our approximation to selected results in the literature reported by Altschul and Gish (1996). We use the amino acid frequencies reported by Robinson and Robinson (1991) and the scoring matrices BLOSUM62, BLOSUM50 (Henikoff and Henikoff (1992), and PAM250 (e.g., Dayhoff, 1978).

Tables 1 and 2 give estimated values of ν_r for the BLOSUM62 and PAM250 substitution matrices, respectively. The estimates are based on 50,000 repetitions of a Monte Carlo experiment and are given for various values of r. Since ν_r is represented as a limit as $j \to \infty$ and large values of j require proportionately long Monte Carlo runs, the estimates are also tabled for different values of j. (The process $\ell_k^{(r)} - \ell_1^{(r)}$ increases at rate μ_1, so a rough estimate, which ignores edge effects, of the mean value of $\tau(j)$ is j/μ_1. Table 3 gives values of μ_1.) Standard deviations of the estimates are about 0.0013–0.0014. The estimates are essentially constant in both r and j. Very similar results have been obtained for the BLOSUM50 substitution matrix and hence are omitted. *We conclude from these simulations that it is reasonable to use the simplified approximation (6) instead of the much more complicated (5).*

Table 3 gives the numerical values of different parameters that are required for implementation of the approximation (6). In particular, instead of ν_r for $r \geq 1$, we propose to use $\nu_\infty = \lim_r \nu_r$, which involves independent, identically distributed, random variables and hence can be computed by the method given in the appendix. Then $\Lambda = \mu_1(1 - e^{-\theta^*})\nu_\infty$. For ease in making numerical comparisons with published values, we continue to use the parameters in the form given above. We also give parameters relevant to the scale that Altschul and Gish (1996) call "nats," which arises from multiplying K given in an arbitrary scale by θ^*. An increase of one unit in the threshold $a = \theta^* b$ results in a change of the probability (1) by approximately the factor e^{-1}. In this scale, the mean values μ_r become the Kullback–Leibler information numbers $I_r = \theta^* \mu_r$ $(r = 0, 1)$.

In Table 3, the values given for ν_0 and ν_∞ have been obtained by the method given in the appendix, which requires a one-dimensional numerical integration. It is easily automated and very fast to evaluate. To the extent that the values of ν_r do change with r, the value given for ν_∞ in Table 3 should be a better approximation for large r, when the increments of the process $\ell_k^{(r)} - \ell_1^{(r)}$ are independent over long stretches. Although the value of ν_∞ is close to the values given in Table 1, it is outside the range of

TABLE 1. SIMULATED VALUES OF ν_r FOR BLOSUM62 MATRIX[a]

	r	*1*	*2*	*4*	*10*	*20*	*50*
j							
5		0.539	0.546	0.548	0.546	0.548	0.546
7		0.559	0.566	0.565	0.562	0.564	0.561
10		0.549	0.554	0.557	0.555	0.553	0.554
20		0.548	0.554	0.554	0.557	0.553	0.555
30		0.550	0.553	0.555	0.557	0.555	0.554

[a]Estimates are based on 50,000 repetitions of a Monte Carlo experiment, for various values of j.

Table 2. Simulated Values of ν_r for PAM250 Matrix[a]

r	1	2	4	10	20	50
j						
5	0.609	0.619	0.615	0.609	0.606	0.606
7	0.600	0.611	0.613	0.607	0.602	0.603
10	0.591	0.602	0.608	0.603	0.601	0.600
20	0.588	0.592	0.599	0.605	0.604	0.600
30	0.590	0.595	0.600	0.604	0.603	0.602

[a]Estimates are based on 50,000 repetitions of a Monte Carlo experiment, for various values of j.

Table 3. Parameters of Selected Substitution Matrices

	BLOSUM62	*BLOSUM50*	*PAM250*
θ^*	0.318	0.232	0.229
ν_0	0.624	0.607	0.662
μ_0	1.267	1.448	0.969
I_0	0.403	0.336	0.222
λ_0	0.214	0.182	0.131
ν_∞	0.564	0.556	0.597
μ_1	2.192	2.496	1.751
I_1	0.697	0.579	0.401
Λ	0.337	0.287	0.214

Table 4. p-Values[a]

K	Δ	δ	b	$p \times 10^6$
BLOSUM62	11	1	81	4.6
				2.3
	9	1	95	3.5
				3.2
	9	2	81	2.2
				1.3
	6	2	95	7.1
				5.2
BLOSUM50	10	2	130	7.1
				3.9
	11	2	125	3.0
				1.5
	10	3	120	1.5
				1.1
	9	3	120	4.9
				2.9
PAM250	12	3	110	5.8
				1.6
	11	3	115	5.1
				1.4
	10	3	122	4.3
				0.9
	11	2	130	8.9
				1.6

[a]Upper entry is our approximation; lower entry from Altschul and Gish (1996); $m = n = 500$.

reasonable sampling variability. The values of ν_∞ for the BLOSUM50 and PAM250 matrices are within the range of reasonable sampling variability of the Monte Carlo estimates for large r.

To verify that these values are at least reasonable, one can compare them to the case of a random walk having normally distributed increments, for which there is the simple approximation $\nu \approx \exp[-\rho(2I)^{1/2}]$, where $\rho = 0.583\cdots$ (cf. Siegmund, 1985). This gives $\nu_0 \approx 0.625$ and $\nu_\infty \approx 0.503$ for the BLOSUM62 matrix and similar results for the other two cases.

For implementation of the approximation (6), we also make edge corrections to m and n, by replacing m with $m' = m - b/\mu_0$ and by a similar correction to n. This correction is justified by the observation that an (ungapped) interval that achieves a score of b is roughly of length b/μ_0. Similar edge corrections are used by Altschul and Gish (1996) and by Mott and Tribe (1999).

Table 4 gives p-values for $m = n = 500$ and selected values of Δ, δ, and b. The first entry given is the approximation (6); the second uses the extreme value approximation suggested by Waterman and Vingron (1994) and by Altschul and Gish (1996) with the empirically determined parameters given in Altschul and Gish (1996). Our approximation is typically about 2–5 times as large as the Altschul–Gish approximation. Very similar results hold for other values of $m = n$ in the range [300,1000].

The practical advantage of our approximation is that it requires only the almost trivial evaluation of θ^* and related parameters and computation of two one dimensional numerical integrals. Hence it can be easily evaluated for a scoring matrix chosen by the user without restriction to the few cases for which the empirical research necessary to estimate the parameters of the Altschul–Gish approximation has been carried out.

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation Grant DMS-0072523 and by an NSF Graduate Research Fellowship.

5. APPENDIX

Theorem 2. *Let $x_1, x_2, \cdots, x_n$ be independent and identically distributed arithmetic random variables with positive mean μ, finite positive variance σ^2, and span h. Let $\phi(t)$ be the characteristic function of x_1 and $\xi(t) = -\log(1 - \phi(t)) = \sum_{n=1}^{\infty}[\phi^n(t)/n]$. Let $S_n = \sum_{i=1}^{n} x_i$. Then for arbitrary $\alpha > 0$,*

$$\sum_{n=1}^{\infty}\left[\frac{1}{n}E(e^{-\alpha S_n^+})\right] + \log(\mu/h) = \frac{1}{2\pi}\int_0^{2\pi} dt \left\{\xi\left(\frac{t}{h}\right)\cdot\frac{e^{-\alpha h - it}}{1 - e^{-\alpha h - it}} + \frac{\xi\left(\frac{t}{h}\right) + \log\left(\frac{\mu}{h}\cdot(1 - e^{it})\right)}{1 - e^{it}}\right\}. \tag{13}$$

Let $S(\alpha)$ denote the left hand side of (11). Siegmund (1985, p. 175) shows that ν_0 and ν_∞ are of the form $h\exp[-S(\theta^*)]/[1 - e^{-\theta^* h}]$. The values given in Table 2 were obtained from this expression.

REFERENCES

Altschul, S.F., Bundschuh, R., Olsen, R., and Hwa, T. 2001. The estimation of statistical parameters for local alignment score distributions. *Nucl. Acids Res.* 29, 351–361.

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

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

Altschul, S.F., Madden, T.L., Schäffer, 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. *Nucl. Acids Res.* 25, 3389–3402.

Arratia, R., Goldstein, L., and Gordon L. 1989. Two moments suffice for Poisson approximation: The Chen–Stein method. *Ann. Probab.* 17, 9–25.

Arratia, R., Gordon, L., and Waterman, M.S. 1990. The Erdös-Rényi Law in distribution for coin tossing and sequence matching. *Ann. Statist.* 18, 539–570.

Asmussen, S. 1989. Risk theory in a Markokvian environment. *Scand. Actuarial J.*, 69–100.

Athreya, K.B., McDonald, D., and Ney, P. 1978. Limit theorems for semi-Markov processes and renewal theory for Markov chains. *Ann. Probab.* 6, 788–797.

Dayhoff, M.O. 1978. *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–358. National Biomedical Research Foundation, Washington, DC.

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

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Nat. Acad. Sci. USA* 89, 10915–10919.

Karlin, S., and Dembo, A. 1992. Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Probab.* 24, 113–140.

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

Pearson, W.R. 1995. Comparison of methods for searching protein databases. *Protein Sci.* 4, 1145–1160.

Robinson, A.B., and Robinson, L.R. 1991. Distribution of glutamine and asparagine residues and their near neighbors in peptides and proteins. *Proc. Nat. Acad. Sci. USA* 88, 8880–8884.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals.* Springer-Verlag, New York.

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

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.

Tu, I.-P., and Siegmund, D. 1999. The maximum of a function of a Markov chain and application to linkage analysis. *Adv. Appl. Probab.* 31, 510–531.

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

Waterman, M., and Vingron, M. 1994. Sequence comparison and Poisson approximation. *Statistical Science* 9, 367–381.

Woodroofe, M. 1982. *Nonlinear Renewal Theory in Sequential Analysis.* SIAM, Philadelphia.

Yakir, B., and Pollak, M. 1998. A new representation for a renewal-theoretic constant appearing in asymptotic approximations of large deviations. *Ann. Appl. Probab.* 8, 749–774.

Address correspondence to:
David Siegmund
Stanford University
Department of Statistics
Sequoia Hall 140
Stanford, CA, 94305-4065

E-mail: dos@stat.stanford.edu