The Annals of Statistics
2000, Vol. 28, No. 3, 657–680

APPROXIMATE *p*-VALUES FOR LOCAL SEQUENCE ALIGNMENTS

BY DAVID SIEGMUND[1] AND BENJAMIN YAKIR[2]

Stanford University and The Hebrew University

Assume that two sequences from a finite alphabet are optimally aligned according to a scoring system that rewards similarities according to a general scoring scheme and penalizes gaps (insertions and deletions). Under the assumption that the letters in each sequence are independent and identically distributed and the two sequences are also independent, approximate *p*-values are obtained for the optimal local alignment when either (i) there are at most a fixed number of gaps, or (ii) the gap initiation cost is sufficiently large. In the latter case the approximation can be written in the same form as the well-known case of ungapped alignments.

1. Introduction. An important step in learning the function of a new gene (DNA sequence) or protein (amino acid sequence) is to compare the new sequence with existing sequences in a data base search, for example, of the DNA data base GenBank maintained by the National Institutes of Health. In addition to the sequence information, these data bases contain whatever is known about the function of a gene/protein, which may be comparatively easy to determine in experimental organisms, for example, baker's yeast or mice. Evolutionary theory holds that genes/proteins having a similar function are likely to have evolved from a common ancestor through mutation. Hence one hopes that by finding in the data base sequences similar to the new sequence one can make an educated guess about its function.

The major problem of sequence comparison is algorithmic determination of sequence similarity [cf. Smith and Waterman (1981), Altschul, Gish, Miller, Myers and Lipman (1990)]. In addition one would like to evaluate the statistical significance of sequences showing a particular level of similarity [e.g., Arratia, Gordon and Waterman (1990), Dembo, Karlin and Zeitouni (1994), Altschul and Gish (1996)]. For a more detailed discussion of the scientific background and introduction to the computational and statistical issues, see Altschul and Gish (1996), Durbin, Eddy, Krogh and Mitchison (1998), Waterman and Vingron (1994) and Waterman (1995). Pearson (1995) makes an empirical comparison of different methods. Altschul, Madden, Schäffer, Zhang, Zhang, Miller and Lipman (1997) discuss recent additions to the BLAST program in a paper that at the same time is a readable introduction to many basic issues.

---

[1]Supported in part by NSF Grant DMS-97-04324.

[2]Supported in part by the Israeli Academy of Science and by the US–Israel Binational Science Foundation.

AMS 1991 *subject classifications*. Primary 62M40; secondary 92D10.

Key words and phrases. Sequence alignment, *p*-value, gaps, large deviations.

We shall be concerned with the optimal local alignment of two sequences of letters from a common, finite alphabet. 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 one or the other sequence. These concepts will be made more precise below. Our goal in this paper is to give new p-value approximations. Approximations when gaps are not allowed have been given by a number of authors, for example, Arratia, Gordon and Waterman (1990) for a special scoring function and more generally by Dembo, Karlin and Zeitouni (1994). Neuhauser (1994) has obtained an approximation in the case that a fixed number of gaps in allowed, but aligned letters are required to match.

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)]. Recently Mott and Tribe (1999) have obtained a useful heuristic approximation for a general scoring scheme with gaps. Their work is complementary to ours. They heuristically piece together results for ungapped alignments to obtain useful numerical approximations for gapped alignments; we have attempted to obtain a mathematically precise approximation, which turns out to be structurally similar although it differs in the value of important parameters.

Except for Mott and Tribe, the authors cited above employ the Chen–Stein method to obtain a Poisson approximation. Our method is a modification of that introduced by Yakir and Pollak (1998) for one-dimensional random fields and extended to multidimensional fields by Siegmund and Yakir (2000). Our basic approximation is one of large deviations. We also show how that approximation can be converted into a Poisson approximation.

2. Notation and assumptions. 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 $x_i, y_j \in \mathscr{A}$. 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.

These sequences are to be aligned. A *candidate alignment* $\mathbf{z} = \{(i_t, j_t): 1 \le t \le k\}$, for some $1 \le i_1 < i_2 < \cdots < i_k \le m$ and $1 \le j_1 < j_2 < \cdots < j_k \le n$, specifies that x_{i_t} and y_{j_t} are *aligned* for all $t = 1, \ldots, k$. 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 $1 \le t < k$, that is, 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 $K(x_i, y_j)$. For the first of two cases that we consider, we assume a penalty of δ for each

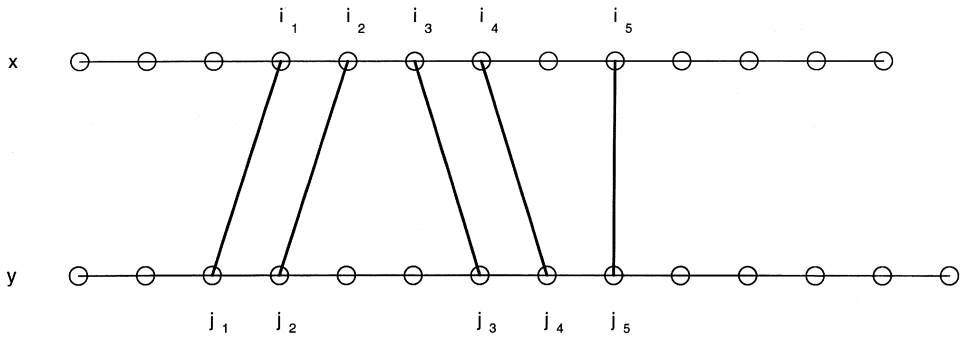

FIG. 1. *A candidate local alignment of words of length* $m = 13$ *and* $n = 14$ *having* $k = 5$ *aligned letters,* $l = 3$ *unaligned letters and* $j = 2$ *gaps.*

unaligned letter. The total number of unaligned letters is $l = (i_k - i_1 - k + 1 + j_k - j_1 - k + 1)$, so $S_{\mathbf{z}} = S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{k} K(x_{i_t}, y_{j_t}) - \delta l$. When we use this formulation, we also assume that all unaligned letters lie in at most a fixed number j of *gaps*, where 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}). We also consider the case where the maximum number of gaps is not fixed, but 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. In this case $S_{\mathbf{z}} = S_{\mathbf{z}}(\mathbf{x}, \mathbf{y}) = \sum_{t=1}^{k} K(x_{i_t}, y_{j_t}) - \Delta j - \delta l$, where j is the number of gaps and l is the total number of unaligned letters, or equivalently the total length of all gaps. Figure 1 shows a candidate alignment with $k = 5$ aligned letters and $l = 3$ unaligned letters within $j = 2$ gaps. Note that there are no costs assessed for letters that are neither aligned nor unaligned. [This is the essential difference between the local alignments discussed in this paper and global alignments. See Waterman (1995).]

Given a collection $\mathcal{Z}$ of 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. Our main focus will be to approximate this probability for large values of b, m and n and for an appropriate collection $\mathcal{Z}$ of candidate alignments.

Given a candidate alignment $\mathbf{z}$, we construct an alternative probability measure $\mathrm{P}_{\mathbf{z}}$ for the sequences $\mathbf{x}$ and $\mathbf{y}$ and consider the log-likelihood ratio $\ell_{\mathbf{z}} = \log[\mathrm{P}_{\mathbf{z}}(\mathbf{x}, \mathbf{y})/\mathrm{P}_0(\mathbf{x}, \mathbf{y})]$ of the probability measure $\mathrm{P}_{\mathbf{z}}$ over the original null measure P_0. Our construction will be such that the score function will

be proportional to a penalized log-likelihood ratio. Notice that under P_0 the random part in $S_{\mathbf{z}}$, namely $\sum_t K(x_{i_t}, y_{j_t})$, is a sum of independent, identically distributed random variables. We assume that

$$E_0 K(x, y) < 0 \quad \text{and} \quad P_0\{K(x, y) > 0\} > 0. \tag{2}$$

Let $\psi(\theta) = \log E_0 \exp[\theta K(x, y)]$. Then $f_\theta(\alpha, \beta) = \exp[\theta K(\alpha, \beta) - \psi(\theta)]\mu_\alpha \nu_\beta$ defines an exponential family of probabilities indexed by θ. Note that $f_0(\alpha, \beta)$ is the original product probability $\mu_\alpha \nu_\beta$; but except in the trivial case that $K(\alpha, \beta)$ is a sum of a function of α and a function of β, x and y are not generally independent under f_θ for $\theta \neq 0$.

From (2) it follows that there exists a unique value $\theta^* > 0$ for which $\psi(\theta^*) = 0$. We define $P_{\mathbf{z}}$ to be the probability that (i) makes the aligned pairs in $\mathbf{z}$ independent with the distribution of f_{θ^*} and (ii) makes all other x_i and y_j independent of each other and of the aligned pairs with the distributions μ and ν, respectively. The log likelihood ratio of $P_{\mathbf{z}}$ relative to P_0 is $\ell_{\mathbf{z}} = \theta^* \sum_{t=1}^k K(x_{i_t, j_t})$. The event of interest from (1) can be rewritten in terms of a penalized log likelihood ratio. Let $g(\mathbf{z})$ equal $\theta^* \delta l$ or $\theta^*(\delta l + \Delta j)$ for the two cases described above. Also let $a = \theta^* b$. Then we can rewrite (1) as

$$P_0\{\max_{\mathbf{z} \in \mathscr{Z}} [\ell_{\mathbf{z}} - g(\mathbf{z})] \geq a\}. \tag{3}$$

In the first instance we take $\mathscr{Z}$ to be the set of all matching patterns having at most j gaps; in the latter case the restriction on the number of gaps is not required.

In the case of protein alignment, the most important examples of functions $K(\alpha, \beta)$ are designed to reflect both the rarity of the amino acid and the ease with which one amino acid can change to another through mutation. See, for example, Durbin, Eddy, Krogh and Mitchison (1998) for a detailed discussion. A scientifically artificial, but mathematically illuminating example is to put $K(\alpha, \beta)$ equal to 1 or $-\xi(\xi > 1)$ according as $\alpha = \beta$ or not. Then $\sum_{t=1}^k K(x_{i_t}, y_{j_t}) = (1 + \xi)X_k - \xi k$; where under P_0, X_k is binomially distributed with parameters k, p_0 and

$$p_0 = \sum_{\alpha \in \mathscr{A}} \mu_\alpha \nu_\alpha$$

is the probability of a pair to match. In this case the probability measure $P_{\mathbf{z}}$ gives elevated probability to the event $\{x_{i_t} = y_{j_i}\}$ compared to what may be expected when the sequences are independent. Let $p_1 > p_0$ be the unique solution to $\xi = \log\{(1 - p_0)/(1 - p_1)\}/\log(p_1/p_0)$. The $P_{\mathbf{z}}$-probability of the event $\{x_{i_t} = \alpha, y_{j_t} = \beta\}$ equals $(p_1/p_0)\mu_\alpha \nu_\beta$ if $\alpha = \beta$ and equals $[(1 - p_1)/(1 - p_0)]\mu_\alpha \nu_\beta$ if $\alpha \neq \beta$. Under $P_{\mathbf{z}}$ the distribution of X_k is Binomial (k, p_1).

3. Main results. Observe that the probability in (3) has the upper bound

$$P_0\Big(\max_{\mathbf{z} \in \mathscr{Z}} (\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\Big) \leq \sum_{\mathbf{z} \in \mathscr{Z}} P_0(\ell_{\mathbf{z}} - g(\mathbf{z}) \geq a) \leq e^{-a} \sum_{\mathbf{z} \in \mathscr{Z}} \exp(-g(\mathbf{z}))\Big).$$

The collection $\mathscr{J}$ is a union of the subcollections $\mathscr{J}_{j,l,k}$ where $\mathscr{J}_{j,l,k}$ consists of all candidate alignments with k aligned letters and j gaps of total length $l \geq j$. The cardinality of the collection $\mathscr{J}_{j,l,k}$ is approximately $mn2^j\binom{k-1}{j}\binom{l-1}{j-1}$. For large values of m and n, a candidate alignment can begin at approximately any of mn positions. If there are to be k aligned letters, there are $2^j\binom{k-1}{j}$ ways of opening j gaps, and there are $\binom{l-1}{j-1}$ ways to divide l unaligned letters among these j gaps (the number of ways to put l balls into j boxes so that each box contains at least one ball). Understanding the geometry of $\mathscr{J}$, in particular that k must be restricted to be proportional to a, will lead to approximations to the probability rather than crude upper bounds.

The approximating formula involves an infinite number of constants $\lambda_0, \lambda_1, \ldots$. The constant λ_0 is defined in terms of fluctuations of a random walk. It arises in the case of ungapped alignments [Dembo, Karlin and Zeitouni (1994)] and has been thoroughly studied in the context of sequential analysis [e.g., Woodroofe (1982), Siegmund (1985)]. For $r \geq 1$, the parameter λ_r is similar, but it is defined in terms of a Markov chain induced by a gap of length r. It seems plausible from their probabilistic meaning that the λ_r for $r \geq 1$ are roughly equal. A more complete discussion is given in Section 6.

Define I and σ^2 by the relations

$$\mathrm{E}_{\mathbf{z}}\ell_{\mathbf{z}} = kI$$

and

$$\mathrm{var}_{\mathbf{z}}\ell_{\mathbf{z}} = k\sigma^2.$$

We shall require the technical assumptions that $\min(m, n)/a \to \infty$, and for Theorems 1 and 2 that $mn \exp\{-a^{1-\varepsilon}\}$ is bounded. The second assumption puts the probability of interest into the domain of large deviations.

To simplify the presentation of our main results we assume here that $K(x, y)$ is a nonlattice random variable and limit our discussion of the more complicated lattice case to a remark following Lemma 4. The notation $a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$.

THEOREM 1. *Suppose $\mathscr{J}$ consists of candidate alignments having at most j gaps and $g(\mathbf{z}) = \delta l$. There exist positive constants $\lambda_r \leq 1$, which are defined following Lemma 4 and characterized in Section 6, such that as $a \to \infty$, the probability* (3)

$$\sim mn \exp(-a)[I^{-1}\lambda_0^2(2a/I)^j/j!]\sum_{l=j}^{\infty} e^{-\theta^*\delta l}\sum \lambda_1^{i_1}\cdots\lambda_{l-j+1}^{i_{l-j+1}},$$

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 $1/[\exp(\theta^*\delta) - 1]$.

Although the cost structure of Theorem 1 is scientifically artificial, its proof contains the essential ingredients needed for Theorem 2 given below. Note

also that except for the large deviation normalization, the result of Dembo, Karlin and Zeitouni (1994) for the ungapped case is the special case $j = 0$.

Now suppose that $\mathscr{Z}$ consists of all candidate alignments without any restriction on the number of gaps, $\mathscr{Z}_j$ is the subset of $\mathscr{Z}$ consisting of candidate alignments having exactly j gaps, and $g(\mathbf{z}) = \theta^*(\delta l + \Delta j)$. The argument of Theorem 1 applied to $\mathscr{Z}_j$ with a replaced by $a + \theta^*\Delta j$ formally yields

$$\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{Z}_j}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\Big) \sim mne^{-a}I^{-1}\lambda_0^2\big[\big(2ae^{-\theta^*\Delta}/I\big)^j/j!\big] \times \sum_{l=j}^{\infty} e^{-\theta^*\delta l} \sum \lambda_1^{i_1}\cdots\lambda_{l-j+1}^{i_{l-j+1}}.$$

A natural conjecture is that we can replace $\mathscr{Z}_j$ by $\mathscr{Z}$ on the left-hand side and sum the right-hand side over j to obtain the approximation

$$\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{Z}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\Big) \sim mne^{-a}I^{-1}\lambda_0^2 \sum_{j=0}^{\infty}\big[\big(2ae^{-\theta^*\Delta}/I\big)^j/j!\big] \times \sum_{l=j}^{\infty} e^{-\theta^*\delta l} \sum \lambda_1^{i_1}\cdots\lambda_{l-j+1}^{i_{l-j+1}}. \tag{4}$$

We are unable, however, to prove (4) in general. In particular if the cost Δ of initiating a new gap is fixed while m and n become large, there may be cases when the score under the null hypothesis can be improved asymptotically by candidate alignments with a very large number of gaps. This is usually undesirable biologically, and can be controlled mathematically by assuming Δ to be sufficiently large that j cannot be too large without incurring an unacceptable penalty. [In particular, this serves to keep the problem well within the so-called "logarithmic domain"; see Waterman and Vingron (1994).]

The preceding considerations are formalized in the following theorem.

THEOREM 2. *Let $\mathscr{Z}$ be the set of all candidate alignments and $g(\mathbf{z}) = \theta^*(\delta l + \Delta j)$. Assume $\theta^*\Delta = \log(a) + C$ for some constant C. Then* (4) *holds as $a \to \infty$. The innermost summation on the right-hand side of* (4) *has the same meaning as in Theorem* 1.

If we replace the condition of subexponential growth of mn in Theorem 2 with the hypothesis that $mn\exp(-a)$ converges to a finite, positive limit, then a general result of Arratia, Goldstein and Gordon (1989) allows us to turn Theorem 2 into a Poisson approximation.

THEOREM 3. *Suppose the conditions of Theorem* 2 *hold, but that $mn\exp(-a)$ converges to a finite, positive limit. Let Q denote the right-hand side of* (4). *Then*

$$\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{Z}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\Big) - [1 - \exp(-Q)] \to 0$$

as $a \to \infty$.

REMARKS. If the λ_r for $r \geq 1$ are replaced by an upper bound Λ, the right-hand side of (4) is bounded by

$$mnI^{-1}\lambda_0^2 \exp\bigl(-a\{1 - 2I^{-1}\Lambda e^{-\theta^*\Delta}/[e^{\theta^*\delta} - 1]\}\bigr). \tag{5}$$

In Section 6 we give an upper bound that usually improves on the inequality $\lambda_r \leq 1$ stated in Theorem 1. It is also shown in Section 6 that when K takes on only two values, the λ_r are all equal, so in that case (4) simplifies to (5) with $\Lambda = \lambda_1$. Numerical evidence that the λ_r, $r \geq 1$, are effectively constant in many cases of interest and the accuracy of (5) as an approximation to (3) will be discussed elsewhere.

4. Discussion. The technical requirement of Theorem 2, that $\theta^*\Delta = \log a + C$, is awkward conceptually since the parameter Δ must be chosen before one knows the values of a of interest. However, the condition can be viewed as a diagnostic: for particular values of Δ; θ^* and a it can always be formally satisfied, but this may require a large value of $|C|$. Presumably a large positive value of C is of no concern. That would indicate that there effectively are no gaps, while formal application of (4) would lead to roughly the same numerical result as Theorem 1 with $j = 0$, that is, when gaps are forbidden. A large negative value of C serves as a warning that the approximation may be poor for that value of a. With regard to Theorem 3, the condition can also be viewed as a requirement that Δ be sufficiently large compared to $\log \log(mn)$, which is conceptually more natural since it does not explicitly involve the threshold a.

The p-value in the ungapped case ($j = 0$ in Theorem 1) is often written in the (Poisson) form

$$\mathrm{P}_0\Bigl\{\max_{\mathbf{z}} S_{\mathbf{z}} \geq b\Bigr\} \approx 1 - \exp\{-mnK \exp(-\lambda b)\}.$$

Waterman and Vingron (1994) and Althschul and Gish (1996), among others, have conjectured that an approximation of this form is valid in the gapped case as well and have put considerable ingenuity into developing such an approximation for numerical applications. The most common approach has been to use simulated and/or actual data to estimate values of K and λ (and in some cases to introduce modifications to m and n as an edge correction). We see from Theorems 2 and 3 that the conjectured form of the approximation is correct, at least when the technical hypothesis concerning Δ is satisfied. Thus Theorems 2 and 3 can be viewed as a theoretical justification of current practice. From this point of view, the complications arising from the different λ_r are unimportant.

Mott and Tribe (1999) obtain heuristically an approximation of the form of (5) with Λ equal to λ_0. They suggest a modified approximation, which they study numerically and conclude is a good approximation for small values of a parameter that in our notation equals $\{2I^{-1}\Lambda \exp(-\theta^*\Delta)/[\exp(\theta^*\delta) - 1]\}$. For fixed δ this translates to the requirement that Δ be large. It also suggests that our requirement on Δ might be replaced by the hypothesis that $\theta^*[\Delta +$

$\delta] = \log(a) + C$. While this modified hypothesis would widen slightly the scope of application of Theorems 2 and 3, and appears to require relatively inconsequential changes in the proofs given below, from an aesthetic point of view we would prefer to be able to weaken the condition substantially or even eliminate it entirely.

We conjecture that the technical requirement of Theorem 2 can be weakened to $\theta^*\Delta = \gamma \log a + C$ for some $1/2 < \gamma \le 1$. Under this hypothesis the values of j that contribute to the final expression are all $o(a^{1/2})$, hence $o(k^{1/2})$ for the important values of k (cf. Lemmas 8 and 9 below). Many of the required modifications go through easily, but others (e.g., Lemma 4) require substantially more refined analysis. However, if we would try to replace the lower bound on γ by anything smaller than 1/2, the important values of j would in general exceed $k^{1/2}$. It is easy to see from reasoning as in the birthday problem, or in calculating the distribution of the minimum distance between consecutive order statistics when sampling from a uniform distribution, that when j exceeds $k^{1/2}$, then a nonnegligible proportion of gaps are separated by only short intervals of aligned pairs, and the argument given below based on the independence that arises from having relatively long blocks of aligned pairs between gaps, breaks down.

5. Proofs. Following Yakir and Pollak (1998) and Siegmund and Yakir (1999), we use a likelihood-ratio identity in order to transform the problem from one of computing the P_0-probability of a rare event to a new measure under which the same event is much more likely to occur. For an intuitive description of the method, see Siegmund and Yakir (2000). It is convenient to use the notational convention that the symbol $E(X; A)$ denotes $E(X1_A)$ when 1_A is the indicator of the event A. The likelihood ratio transformation to finite positive measure is as follows:

$$\begin{aligned}
&\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{J}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big) \\
&\quad = \mathrm{E}_0\left[\frac{\sum_{\mathbf{z}\in\mathscr{J}} \exp\{\ell_{\mathbf{z}} - g(\mathbf{z})\}}{\sum_{\mathbf{u}\in\mathscr{J}} \exp\{\ell_{\mathbf{u}} - g(\mathbf{u})\}}; \max_{\mathbf{u}\in\mathscr{J}}\{\ell_{\mathbf{u}} - g(\mathbf{u})\} \ge a\right] \qquad (6)\\
&\quad = \sum_{\mathbf{z}\in\mathscr{J}} e^{-g(\mathbf{z})}\mathrm{E}_{\mathbf{z}}\left[\left(1\Big/\sum_{\mathbf{u}\in\mathscr{J}} \exp\{\ell_{\mathbf{u}} - g(\mathbf{u})\}\right); \max_{\mathbf{u}\in\mathscr{J}}\{\ell_{\mathbf{u}} - g(\mathbf{u})\} \ge a\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
&e^a\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{J}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big) \\
&\quad = \sum_{\mathbf{z}\in\mathscr{J}} e^{-g(\mathbf{z})}\mathrm{E}_{\mathbf{z}}\left[\frac{\max_{\mathbf{u}\in\mathscr{J}} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{J}} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))} \exp\Big(-\Big[\max_{\mathbf{u}\in\mathscr{J}}(\ell_{\mathbf{u}} - g(\mathbf{u})) - a\Big]\Big);\right. \qquad (7)\\
&\qquad\qquad \left.\max_{\mathbf{u}\in\mathscr{J}}(\ell_{\mathbf{u}} - g(\mathbf{u})) \ge a\right].
\end{aligned}$$

The analysis is carried through by approximating the terms in this sum.

We will eventually want to approximate the p-values for the collection $\mathcal{J}$ of all possible $\mathbf{z}$'s. We begin with the simpler case that the unaligned letters are concentrated in a finite number of gaps. In other words, using the usual dot matrix representation of the candidate alignment, where a dot at the position (i, j) indicates that x_i and y_j are aligned, runs of dots along diagonal lines represent intervals of consecutive aligned pairs and jumps, horizontal or vertical, represent gaps, our restriction is that the number of jumps is at most j, for some fixed j. Equivalently, the statement is that the number of runs is at most $j+1$. No restriction is put on the number of aligned pairs or on the overall number of unaligned letters.

Until further notice we assume that $\mathcal{J}$ consists of matching patterns having at most j gaps and the function g is given by $g(\mathbf{z}) = \theta^* \delta l$, where l is the total number of unaligned letters.

Consider a smaller collection $\mathcal{J}_j \subset \mathcal{J}$, defined to be the collection of all $\mathbf{z} \in \mathcal{J}$ for which the k aligned pairs satisfy $(1-\varepsilon_1)a/I < k < (1+\varepsilon_1)a/I$, the number of gaps is *exactly* j, the overall number of unaligned letters, l, is bounded by $\varepsilon_2 a^{1/2}$ for some small $\varepsilon_2 > \varepsilon_1 > 0$, to be specified more precisely later. We intend to approximate the P_0-probability that the maximum over $\mathbf{z} \in \mathcal{J}_j$ of $\ell_{\mathbf{z}} - g(\mathbf{z})$ exceeds a. This probability is smaller than the probability (4), but as we will see in Lemma 8 in the Appendix, the ratio between the probabilities converges to 1. Thus, an approximation of the probability for $\mathcal{J}_j$ will effectively give us the approximation for the probability for $\mathcal{J}$.

The considerations that led to (7), applied to $\mathcal{J}_j$, yield

$$
\begin{aligned}
&e^a \mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathcal{J}_j}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big) \\
&\quad = \sum_{\mathbf{z}\in\mathcal{J}_j} e^{-g(\mathbf{z})} \mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathcal{J}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathcal{J}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))} \exp\Big(-\Big[\max_{\mathbf{u}\in\mathcal{J}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) - a\Big]\Big); \\
&\qquad\qquad \max_{\mathbf{u}\in\mathcal{J}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) \ge a \Bigg] \\
&\quad = \sum_{l=j}^{\varepsilon_2 a^{1/2}} e^{-\theta^* \delta l} \sum_{\mathbf{z}\in\mathcal{J}_{j,l}} \mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathcal{J}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}{\sum_{\mathcal{J}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))} \\
&\qquad\qquad \times \exp\Big(-\Big[\max_{\mathcal{J}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) - a\Big]\Big); \max_{\mathbf{u}\in\mathcal{J}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) \ge a \Bigg],
\end{aligned}
$$

where

$$
\mathcal{J}_{j,l} = \{\mathbf{z} \in \mathcal{J}_j \colon \mathbf{z} \text{ has } j \text{ gaps of total length } l\}.
$$

Next we approximate each one of the expectations in the above sum. In broad outline we show that the fraction

$$
\frac{\max_{\mathbf{u}\in\mathcal{J}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathcal{J}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}
$$

is asymptotically independent of

$$\exp\Big\{-\Big[\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big\}1\Big\{\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}}-g(\mathbf{u}))\ge a\Big\},$$

and we evaluate the expectations of the two factors separately. The second factor is evaluated by a local limit theorem in Lemma 4, while the first is shown to equal approximately a product of independent random variables, the expectation of which leads to the product of the λ_r. The proof is divided into a series of lemmas, which to a certain extent parallel the argument of Siegmund and Yakir (2000). Some technical results are deferred to the Appendix.

Let $\mathscr{D}_{\mathbf{z}}$ be the collection of all $\mathbf{u}\in\mathscr{D}_j$ which agree with $\mathbf{z}$ in all but at most $(\log a)^2$ terms. Note that although both $\mathscr{D}_{\mathbf{z}}$ and $\mathscr{D}_j$ involve subscripted versions of $\mathscr{D}$, there should be no confusion between the integer j and the matching pattern $\mathbf{z}$. The usefulness of $\mathscr{D}_{\mathbf{z}}$ stems from the observation that for most $\mathbf{z}$, g is constant on $\mathscr{D}_{\mathbf{z}}$. In the first lemma it is shown that the collection $\mathscr{D}_j$ can be replaced by the much smaller collection $\mathscr{D}_{\mathbf{z}}$, without changing the expectation by much.

LEMMA 1. *Let $\varepsilon>0$ be given. Then for all $\mathbf{z}\in\mathscr{D}_j$,*

$$\begin{aligned}
&\mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_j}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_j}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Bigg(-\Big[\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Bigg);\\
&\qquad\qquad\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}}-g(\mathbf{u}))\ge a\Bigg]\\
&\le \mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Big(-\Big[\max_{\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big);\\
&\qquad\qquad\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))\ge a\Bigg]+a^{-1}
\end{aligned}$$

and

$$\begin{aligned}
&(1+\varepsilon)\mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_j}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_j}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Big(-\Big[\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big);\\
&\qquad\qquad\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}}-g(\mathbf{u}))\ge a\Bigg]\\
&\ge \mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Big(-\Big[\max_{\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big);\\
&\qquad\qquad\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))\ge a\Bigg]-a^{-1},
\end{aligned}$$

provided that a is large enough.

PROOF. Let $W = \sum_{\mathbf{u}\in\mathscr{D}_j} \exp\{(\ell_{\mathbf{u}}-\ell_{\mathbf{z}})-(g(\mathbf{u})-g(\mathbf{z}))\}$, $W_{\mathbf{z}} = \sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}} \exp\{(\ell_{\mathbf{u}}-\ell_{\mathbf{z}})-(g(\mathbf{u})-g(\mathbf{z}))\}$ and $\overline{W}_{\mathbf{z}} = W - W_{\mathbf{z}}$. Likewise, $Q_{\mathbf{z}} = \max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}} \exp\{(\ell_{\mathbf{u}}-\ell_{\mathbf{z}})-(g(\mathbf{u})-g(\mathbf{z}))\}$, $\overline{Q}_{\mathbf{z}} = \max_{\mathbf{u}\in\mathscr{D}_j\setminus\mathscr{D}_{\mathbf{z}}} \exp\{(\ell_{\mathbf{u}}-\ell_{\mathbf{z}})-(g(\mathbf{u})-g(\mathbf{z}))\}$ and $Q = \max(Q_{\mathbf{z}}, \overline{Q}_{\mathbf{z}})$. Note that $W > W_{\mathbf{z}} > 1$, $Q \geq Q_{\mathbf{z}} \geq 1$, $W \geq Q$ and $W_{\mathbf{z}} \geq Q_{\mathbf{z}}$.

On the one hand

$$\begin{aligned}
&e^a \mathrm{E}_{\mathbf{z}}\big[1/(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W); \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q \geq e^a\big] \\
&\quad = e^a \mathrm{E}_{\mathbf{z}}\big[1/(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W); \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q_{\mathbf{z}} \geq e^a, \overline{Q}_{\mathbf{z}} \leq 1\big] \\
&\qquad + \mathrm{E}_{\mathbf{z}}\big[e^a/(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W); \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q \geq e^a, \overline{Q}_{\mathbf{z}} \geq 1\big] \\
&\quad \leq e^a \mathrm{E}_{\mathbf{z}}\big[1/(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W_{\mathbf{z}}); \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q_{\mathbf{z}} \geq e^a\big] + \mathrm{P}_{\mathbf{z}}(\overline{W}_{\mathbf{z}} > \varepsilon).
\end{aligned}$$

On the other hand

$$\begin{aligned}
&e^a \mathrm{E}_{\mathbf{z}}\big[1/(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W); \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q \geq e^a\big] \\
&\quad \geq e^a \mathrm{E}_{\mathbf{z}}\big[1/[(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W_{\mathbf{z}})(1 + \overline{W}_{\mathbf{z}}/W_{\mathbf{z}})]; \\
&\qquad\qquad \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q_{\mathbf{z}} \geq e^a, \overline{W}_{\mathbf{z}} \leq \varepsilon\big] \\
&\geq (1+\varepsilon)^{-1} e^a \mathrm{E}_{\mathbf{z}}\big[1/(\exp(\ell_{\mathbf{z}} - g(\mathbf{z}))W_{\mathbf{z}}); \exp(\ell_{\mathbf{z}} - g(\mathbf{z}))Q_{\mathbf{z}} \geq e^a\big] - \mathrm{P}_{\mathbf{z}}(\overline{W}_{\mathbf{z}} > \varepsilon).
\end{aligned}$$

In order to show that $\mathrm{P}_{\mathbf{z}}(\overline{W}_{\mathbf{z}} > \varepsilon) \leq 1/a$ we divide the set $\mathscr{D}_j\setminus\mathscr{D}_z$ into two subsets. The first subset contains all $\mathbf{u}$'s which agree with $\mathbf{z}$ in all but (at most) $a^{1-\varepsilon}$ terms. Note that the number of terms in this subset is polynomial in a, and on this subset $g(\mathbf{u}) - g(\mathbf{z}) \geq 0$. [Actually $g(\mathbf{u}) - g(\mathbf{z}) = 0$, although the weaker relation is all that is required.] The rest of the matching patterns $\mathbf{u}$ differ from $\mathbf{z}$ in more that $a^{1-\varepsilon}$ terms; the number of these terms is mn times a polynomial in a. It is enough to show that for arbitrary c, for all a large enough,

$$\mathrm{P}_{\mathbf{z}}(\ell_{\mathbf{u}} - \ell_{\mathbf{z}} \geq -c \log a) \leq 1/a^c, \tag{8}$$

for all $\mathbf{u}$ in the first subset, and

$$\mathrm{P}_{\mathbf{z}}\big\{\ell_{\mathbf{u}} - \ell_{\mathbf{z}} \geq -c(a^{1/2} + \log mn)\big\} \leq 1/(mna^{c/2}), \tag{9}$$

for all $\mathbf{u}$ in the second subset.

The random variable $\ell_{\mathbf{u}} - \ell_{\mathbf{z}}$ is a sum of a sequence of random elements over the set of all indices on which $\mathbf{u}$ and $\mathbf{z}$ differ. The elements in the sequence can be separated into pairs that do not share a common index with other pairs and those that do. The former type of pairs form a sequence of independent random variables and the latter type form a finite state Markov chain. Hence by a large deviations result for finite Markov chains [e.g., Lezaud (1988)] or by the Azuma–Hoeffding inequality for sums of uniformly bounded martingale differences [cf. Williams (1991)], we have for some $\rho > 0$,

$$\mathrm{P}_{\mathbf{z}}(\ell_{\mathbf{u}} - \ell_{\mathbf{z}} \geq -\rho s) \leq ce^{-\rho s},$$

where s is the number of pairs that are not identical in both patterns. From the definition of $\mathscr{D}_z$ we obtain (8). From the definition of $\mathscr{D}_j$ and the fact that mn is assumed to be subexponential, we obtain inequality (9). $\square$

Lemmas 2 and 3 provide technical bounds to guarantee the independence required in the application of a local limit theorem in Lemma 4.

LEMMA 2. *For all sufficiently large a,*

$$\begin{aligned}
&\mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Big(-\Big[\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big);\\
&\qquad\qquad \max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))\ge a\Bigg]\\
&\le a^{-1}+\mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Big(-\Big[\max_{\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big);\\
&\qquad\qquad a\le\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))\le a+\log a\Bigg].
\end{aligned}$$

PROOF. Note that

$$\begin{aligned}
&\mathrm{E}_{\mathbf{z}}\Bigg[\frac{\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}}-g(\mathbf{u}))}\exp\Big(-\Big[\max_{\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))-a\Big]\Big);\\
&\qquad\qquad a+\log a<\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))\Bigg]\le a^{-1},
\end{aligned}$$

hence the claim. $\square$

Let $\tilde{\mathbf{z}}=\bigcap_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\mathbf{u}$. From the definition of $\mathscr{D}_{\mathbf{z}}$ it follows that $\tilde{\mathbf{z}}$ consists of those aligned pairs from $\mathbf{z}$ that are not within $(\log a)^2$ terms of a gap nor of the initial or final aligned pairs of $\mathbf{z}$. We observe that since $\tilde{\mathbf{z}}$ contains gaps in both sequences simultaneously, it is not a candidate alignment, although this fact plays no role in what follows. The term $\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u}))$ is a sum of two independent terms: $\ell_{\tilde{\mathbf{z}}}-g(\tilde{\mathbf{z}})$ and $q_{\mathbf{z}}=\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}(\ell_{\mathbf{u}}-g(\mathbf{u})-\ell_{\tilde{\mathbf{z}}}+g(\tilde{\mathbf{z}}))$. In the next lemma we show that $q_{\mathbf{z}}$ is of controllable order.

LEMMA 3. *Let $\varepsilon>0$ be given. Then $\mathrm{P}_{\mathbf{z}}(q_{\mathbf{z}}>\varepsilon a^{1/2})\le 1/a$, provided that a is sufficiently large.*

PROOF. Note that for all $\mathbf{u}\in\mathscr{D}_{\mathbf{z}}$, $|g(\tilde{\mathbf{z}})-g(\mathbf{u})|\le 2j\theta^*\delta(\log a)^2$. Thus,

$$\mathrm{P}_{\mathbf{z}}\big(q_{\mathbf{z}}>\varepsilon a^{1/2}\big)\le\mathrm{P}_{\mathbf{z}}\big(\ell_{\mathbf{z}}-\ell_{\tilde{\mathbf{z}}}>a^{1/2}\varepsilon/3\big) \tag{10}$$

$$+\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}}\mathrm{P}_{\mathbf{z}}\big(\ell_{\mathbf{u}}-\ell_{\mathbf{z}}\ge a^{1/2}\varepsilon/3\big). \tag{11}$$

However, $\ell_{\mathbf{u}} - \ell_{\mathbf{z}}$ is a log-likelihood ratio relative to $P_{\mathbf{z}}$, thus $P_{\mathbf{z}}(\ell_{\mathbf{u}} - \ell_{\mathbf{z}} \geq a^{1/2}\varepsilon/3) \leq \exp\{-a^{1/2}\varepsilon/3\}$. Since the cardinality of $\mathscr{J}_{\mathbf{z}}$ is $O([2\log a]^{2j})$, it follows that the sum in (11) is $o(1/a)$. Regarding the right-hand side of (10), notice that $\ell_{\mathbf{z}} - \ell_{\tilde{\mathbf{z}}}$ is a sum of at most a constant times $(\log a)^2$ independent, identically distributed random variables. The desired bound thus follows from simple large deviation estimates. □

In the next lemma we condition on $\{q_{\mathbf{z}} = q\}$ and integrate with respect to the distribution of $\ell_{\tilde{\mathbf{z}}}$, which is independent of $q_{\mathbf{z}}$. Notice that the aligned pairs in $\tilde{\mathbf{z}}$ are independent.

LEMMA 4. *Let ϕ denote the standard normal probability density function. Let $\varepsilon > 0$ be given. Let I and σ^2 be defined by $E_{\mathbf{z}}\ell_{\mathbf{z}} = kI$ and $\mathrm{var}_{\mathbf{z}}\ell_{\mathbf{z}} = k\sigma^2$. Let $\tilde{a} = a - q + g(\tilde{\mathbf{z}})$, for $-\varepsilon_2 a^{1/2}\delta\theta^* \leq q - g(\tilde{\mathbf{z}}) \leq \varepsilon_2 a^{1/2}/I$. Then, for large a, uniformly in $q - g(\tilde{\mathbf{z}})$,*

$$E_{\mathbf{z}}\big[\exp(-(\ell_{\tilde{\mathbf{z}}} - \tilde{a}));\tilde{a} \leq \ell_{\tilde{\mathbf{z}}} \leq \tilde{a} + \log a\big] \sim \frac{(1+O(\varepsilon))}{k^{1/2}\sigma}\phi\left(\frac{(a-Ik)(1+O(\varepsilon))}{k^{1/2}\sigma}\right). \tag{12}$$

PROOF. Lemma 5 in the Appendix applies to give (12), but with $\tilde{a}$ and $\tilde{k}$ on the right-hand side. By considering separately the two cases, $|kI - a|/k^{1/2} < c$ and $c < |kI - a|/k^{1/2} < \varepsilon_1 k^{1/2}$, we see that we can replace $\tilde{k}$ and $\tilde{a}$ by k and a, respectively. □

REMARK. The scoring matrices K of interest in applications take on sufficiently many distinct values, especially for the 20 letter amino acid alphabets of protein sequence analysis, that our assumption that $\ell_{\mathbf{z}}$ is nonlattice seems quite reasonable. If $\ell_{\mathbf{z}}$ is lattice with span h, say, it does not seem feasible to give an exact asymptotic expression for the behavior of the left-hand side of (12), although one can bound that expectation asymptotically to lie in the interval $[(1-\varepsilon)h/(e^h - 1), (1+\varepsilon)h/(1 - e^{-h})]$. This interval of uncertainty filters through to the statement of Theorems 1–3 in the lattice case.

We now turn to consideration of

$$E_{\mathbf{z}}\left[\frac{\max_{\mathbf{u}\in\mathscr{J}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})}{\sum_{\mathbf{u}\in\mathscr{J}_{\mathbf{z}}}\exp(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})}\right],$$

which leads to the constants λ_0 and λ_r for $r \geq 1$. The constant λ_0 arises from discrepancies between $\mathbf{u}$ and $\mathbf{z}$ at the two ends of the candidate alignment; it is the same constant that enters into the analysis of ungapped alignments and has been thoroughly studied in the context of sequential analysis [e.g., Woodroofe (1982), Siegmund (1985)].

Consider the special alignment $\mathbf{z} = \{(1,1)\}$ and let $\mathbf{u} = \{(1,1), \ldots, (u,u)\}$, for $u = 1, \ldots, t$. Then

$$\lambda_0 = \lim_{t\to\infty} E_{\mathbf{z}}\left[\frac{\max_{1\leq u\leq t}\exp(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})}{\sum_{1\leq u\leq t}\exp(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})}\right],$$

which by the argument of Siegmund and Yakir (2000) [see also (16) in Section 6 below] has the alternative representation

$$\lambda_0 = \lim_t t^{-1} \mathrm{E}_0 \exp\bigl[\max_{1\le u\le t}(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})\bigr]. \tag{13}$$

Since $\ell_{\mathbf{u}} - \ell_{\mathbf{z}} = \theta^* \sum_{i=2}^{u} K(x_i, y_i)$ is a sum of independent, identically distributed random variables, computable expressions for λ_0 can be obtained along the lines of Appendix A of Siegmund and Yakir (2000) and Chapter 8 of Siegmund (1985).

The constants λ_r are structurally similar, but they arise from discrepancies between $\mathbf{u}$ and $\mathbf{z}$ in the neighborhood of the gaps in $\mathbf{z}$. As a consequence they involve Markov chains induced by the measure $\mathrm{P}_{\mathbf{z}}$ and are substantially more complicated to evaluate numerically. Let $\mathbf{z}$ be a candidate alignment of length k, $k \ge a^{1/2}$, which contains one gap of length r. Assume this is a gap in the y's and that it is at least $(\log a)^2$ terms from the ends of $\mathbf{z}$. Let

$$\lambda_r^* = \lim_{a\to\infty} \mathrm{E}_{\mathbf{z}}\biggl[\frac{\max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}^*} \exp(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})}{\sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}^*} \exp(\ell_{\mathbf{u}} - \ell_{\mathbf{z}})}\biggr],$$

where $\mathscr{D}_{\mathbf{z}}^*$ is the set of all $\mathbf{u}$'s that agree with $\mathbf{z}$ in all but (at most) $t = (\log a)^2$ pairs, and agree completely at the ends of $\mathbf{z}$. Obviously, $\lambda_r^* \le 1$. For a gap in the x's the analogous constant λ_r^{**} will in general be different. The constant λ_r appearing in Theorems 1–3 is $\lambda_r = (\lambda_r^* + \lambda_r^{**})/2$. See Section 6 for alternative representations.

We are now in a position to prove Theorem 1.

PROOF OF THEOREM 1. In Lemma 8 of the Appendix we show that it suffices to prove the theorem with $\mathscr{D}$ replaced by $\mathscr{D}_j$, so there are *exactly* j gaps. To simplify the argument we also assume that $\lambda_r^* = \lambda_r^{**} = \lambda_r$, which would be the case if K is symmetric and the marginal distributions μ and ν are equal. (These are often satisfied in applications.) We remove this assumption at the end of the proof. Lemmas 1, 2 and 3 can be summarized by saying that the term

$$\mathrm{E}_{\mathbf{z}}\biggl[\frac{\max_{\mathbf{u}\in\mathscr{D}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}{\sum_{\mathbf{u}\in\mathscr{D}_j} \exp(\ell_{\mathbf{u}} - g(\mathbf{u}))} \exp\Bigl(-\Bigl[\max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) - a\Bigr]\Bigr); \max_{\mathbf{u}\in\mathscr{D}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) \ge a\biggr]$$

can be approximated, up to an additive error of size $1/a$ and an arbitrary small relative error, by

$$\mathrm{E}_{\mathbf{z}}\biggl[\frac{Q_{\mathbf{z}}}{W_{\mathbf{z}}} \exp\bigl(-[\ell_{\tilde{\mathbf{z}}} - g(\tilde{\mathbf{z}}) + q_{\mathbf{z}} - a]\bigr); \; a \le \ell_{\tilde{\mathbf{z}}} - g(\tilde{\mathbf{z}}) + q_{\mathbf{z}} \le a + \log a; \; 0 \le q_{\mathbf{z}} \le \varepsilon a^{1/2}\biggr], \tag{14}$$

where $Q_{\mathbf{z}} = \max_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}} \exp\{\ell_{\mathbf{u}} - g(\mathbf{u}) - \ell_{\mathbf{z}} + g(\mathbf{z})\}$ and $W_{\mathbf{z}} = \sum_{\mathbf{u}\in\mathscr{D}_{\mathbf{z}}} \exp\{\ell_{\mathbf{u}} - g(\mathbf{u}) - \ell_{\mathbf{z}} + g(\mathbf{z})\}$. In addition $(Q_{\mathbf{z}}, W_{\mathbf{z}}, q_{\mathbf{z}})$ are independent of $\ell_{\tilde{\mathbf{z}}} - g(\tilde{\mathbf{z}})$. Thus,

by Lemma 4, we get that the expectation in (14) divided by the product of $\mathrm{E}_{\mathbf{z}}[Q_{\mathbf{z}}/W_{\mathbf{z}}]$ and

$$\frac{1}{k^{1/2}\sigma}\phi\Big(\frac{(a-Ik)[1+O(\varepsilon)]}{k^{1/2}\sigma}\Big) = \frac{1}{k^{1/2}\sigma}\phi\Big(\frac{(k-a/I)[1+O(\varepsilon)]}{k^{1/2}\sigma/I}\Big), \tag{15}$$

eventually has values in the interval $(1-\varepsilon, 1+\varepsilon)$.

Consider the expectation $\mathrm{E}_{\mathbf{z}}[Q_{\mathbf{z}}/W_{\mathbf{z}}]$. It is obvious that it is nonnegative and bounded above by 1. We claim that for most $\mathbf{z} \in \mathscr{Q}_j$ it is approximately equal to $\lambda_0^2\lambda_1^{i_1}\cdots\lambda_{l-j+1}^{i_1-j+1}$, where i_r is the number of gaps in $\mathbf{z}$ of length r. In particular, it is independent of k. Indeed, let $\mathbf{z}$ be such that the number of aligned pairs in the candidate alignment between consecutive gaps and between the outmost gaps and the ends is more than $2(\log a)^2$. It follows that $\mathscr{Q}_{\mathbf{z}}$ contains only $\mathbf{u}$'s with the same number of gaps as $\mathbf{z}$. Hence $g(\mathbf{u})$ is a constant and cancels out. For such $\mathbf{z}$'s, the random variable $Q_{\mathbf{z}}$ is a maximum of independent random variable, each of which is a maximum of likelihood ratios over disjoint regions, and $W_{\mathbf{z}}$ can be factored as a product of sums of independent likelihood ratios, the sums taken over the same disjoint regions. Hence, the random variable $Q_{\mathbf{z}}/W_{\mathbf{z}}$ can be factored into $j+2$ independent random variables, j of which correspond to the gaps while the other two correspond to the edges of $\mathbf{z}$. The expectation for each of the j factors can be approximated by an appropriate λ_r, where r is the length of the corresponding gap, and the expectation of each edge factor can be approximated by λ_0.

Hence,

$$\begin{aligned}
&e^a \mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{Q}_j}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big) \\
&\quad = \sum_{l=j}^{\varepsilon_2 a^{1/2}} e^{-\theta^*\delta l} \sum_{\mathbf{z}\in\mathscr{Q}_{j,l}} \mathrm{E}_{\mathbf{z}}\Big[\frac{\max_{\mathscr{Q}_j}\exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}{\sum_{\mathscr{Q}_j}\exp(\ell_{\mathbf{u}} - g(\mathbf{u}))}\exp(-[\max_{\mathscr{Q}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) - a]); \\
&\qquad\qquad \max_{\mathbf{u}\in\mathscr{Q}_j}(\ell_{\mathbf{u}} - g(\mathbf{u})) \ge a\Big] \\
&\quad \sim mn\lambda_0^2 \sum_{l=j}^{\varepsilon_2 a^{1/2}} \sum_{k=(1-\varepsilon_1)a/I}^{(1+\varepsilon_1)a/I} e^{-\theta^*\delta l} 2^j \binom{k-1}{j} \frac{1}{k^{1/2}\sigma}\phi\Big(\frac{k-a/I}{k^{1/2}\sigma/I}\Big) \sum \lambda_1^{i_1}\cdots\lambda_{l-j+1}^{i_l-j+1} \\
&\quad \sim nm\lambda_0^2 I^{-1}\Big(\frac{2a}{I}\Big)^j \frac{1}{j!}\sum_{l=j}^{\infty} e^{-\theta^*\delta l} \sum \lambda_1^{i_1}\cdots\lambda_{l-j+1}^{i_{l-j+1}},
\end{aligned}$$

which is the result stated in Theorem 1.

In the preceding argument, where λ_r^* is assumed the same as λ_r^{**}, the 2^{i_r} possible choices for the i_r gaps of size r to be in either the x or the y sequence each contribute a factor λ_r. In general, for some $i = 0, 1, \ldots, i_r$, i of the gaps will be in the x's and $i_r - i$ in the y's. There are $\binom{i_r}{i}$ ways to distribute these gaps, which contribute a factor $\lambda_r^{*i}\lambda_r^{**(i_r-i)}$. When these possibilities are summed over i, we obtain $(\lambda_r^* + \lambda_r^{**})^{i_r} = 2^{i_r}\lambda_r^{i_r}$, as before. □

The proof of Theorem 2 involves similar but more complicated arguments along the lines of those used to prove Theorem 1. To obtain an upper bound, we show that for arbitrarily large j_1, (4) holds uniformly over $j < j_1$. We then use the assumption about Δ and Lemmas 10–13 of the Appendix to show that other matching patterns make a negligible contribution. To obtain a lower bound, we first replace $\mathscr{J}$ in (7) by $\cup_{j<j_1}\mathscr{J}_j$, where $\mathbf{z} \in \mathscr{J}_j$ is now restricted by the additional requirements that any pair of gaps is separated by a distance of at least $(\log a)^2$ and $l < l_1$ for a large but fixed value of $l_1(\gg j_1)$. Since different $\mathbf{z}$ in different $\mathscr{J}_j$ must differ in at least $(\log a)^2$ positions, extending the lower bound of Lemma 1 to this class of matching patterns requires only trivial changes. Since the indicated restriction leaves the cardinality of the (similarly restricted) $\mathscr{J}_{j,l,k}$ essentially unchanged when k is of order a, the rest of the proof proceeds as before.

To prove Theorem 3, suppose that $mne^{-a} \to x$. Let $\widetilde{\mathscr{J}} = \cup_{j\le j_1}\mathscr{J}_j$, where j_1 is a large but fixed integer. By essentially trivial modifications in the proofs of Lemmas 9–12 in the Appendix, we see that for all sufficiently large j_1, $\mathrm{P}(\cup_{\mathbf{z}\in\mathscr{J}\setminus\widetilde{\mathscr{J}}}\{\ell_{\mathbf{z}} - g(\mathbf{z}) \ge a\}) \le \varepsilon$, so we can confine our attention to the set $\widetilde{\mathscr{J}}$. Note that the elements of this set, represented as paths in two-dimensional grid, are of restricted dimensions: a path that begins at the point (i_1, i_2) is contained in the square

$$\big\{(i_1, i_2), (i_1 + ca, i_2), (i_1 + ca, i_2 + ca), (i_1, i_2 + ca)\big\},$$

where $c = I^{-1}(1 + \varepsilon_1) + \varepsilon_2$.

For a candidate alignment in $\widetilde{\mathscr{J}}$, the dot matrix representation between the two sequences is contained in a rectangle of size $m \times n$ in the two-dimensional lattice, which can be subdivided into squares of size $a^2 \times a^2$. Let α be the index of a typical square and denote by $\mathscr{J}_\alpha$ the set of candidate alignments which intersect with the square α. Note that though $\mathscr{J}_\alpha$'s are not disjoint, $\mathscr{J}_\alpha$ can have common elements with $\mathscr{J}_\beta$ only if $\mathscr{J}_\beta$ is a neighbor. Denote by B_α the neighborhood of α. (The neighborhood contains α.) Let X_α be the indicator of the event $\{\max_{\mathbf{z}\in\mathscr{J}_\alpha}[\ell_{\mathbf{z}} - g(\mathbf{z})] \ge a\}$. It follows that X_α and X_β are independent, provided that $\beta \notin B_\alpha$.

Consider $W = \sum_\alpha X_\alpha$ and note that $\mathrm{E}_0 W \to x\lambda$ as $a \to \infty$. According to Arratia, Goldstein and Gordon (1989), the difference between the probability of the event $\{W > 0\} = \{\max_{\mathbf{z}\in\widetilde{\mathscr{J}}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\}$ and the quantity $1 - \exp(-\mathrm{E}_0 W)$ is bounded by $2(b_1 + b_2)$, where

$$b_1 = \frac{mn}{a^4}\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{J}_\alpha}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big)^2,$$

$$b_2 = \frac{8mn}{a^4}\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{J}_\alpha}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a, \max_{\mathbf{u}\in\mathscr{J}_\beta}(\ell_{\mathbf{u}} - g(\mathbf{u})) \ge a\Big),$$

with $\alpha \neq \beta \in B_\alpha$. From Theorem 2 we see that $b_1 = O(\lambda x e^{-a})$. In order to bound the term b_2, note that

$$\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{D}_\alpha}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a, \max_{\mathbf{z}\in\mathscr{D}_\beta}(\ell_{\mathbf{u}} - g(\mathbf{u})) \geq a\Big)$$
$$\leq \mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{D}_\alpha}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\Big)^2 + 3\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{D}_\alpha\cap\mathscr{D}_\beta}(\ell_{\mathbf{z}} - g(\mathbf{z})) \geq a\Big).$$

Theorem 3 now follows since the last probability is asymptotically proportional to $a^3 e^{-a}$.

6. The constants $\lambda_r, r = 1, 2, \ldots$. In this section we discuss different representations of the parameters λ_r for $r \geq 1$ and give a useful upper bound.

Consider two sequences $\mathbf{x}$ and $\mathbf{y}$ of length t, for some integer $t \to \infty$. Given r and u, $1 \leq u < t - r$, let $\ell_{\mathbf{u}} = \ell_u$ be the log-likelihood ratio statistic calculated for the alignment $\mathbf{u}$ which matches the first u x's with the first u y's and the x's between $u+1$ and $t-r$ with the y's between $u+r+1$ and t. Then $\ell_u = \theta^*[\sum_1^u K(x_i, y_i) + \sum_{u+1}^{t-r} K(x_i, y_{i+r})]$, so $\zeta_u = \ell_u - \ell_{u-1}$ is a function of x_u, y_u and y_{u+r} only. The ζ_u are identically distributed, and ζ_u is independent of any ζ_v for v such that $(v \bmod r) \neq (u \bmod r)$. For each $0 \leq i < r$ the process $\{\zeta_u : 1 \leq u \leq t-r, (u \bmod r) = i\}$ is a first order Markov chain. The process $\{\ell_u - \ell_1, 1 \leq u \leq t-r\}$ is an additive functional of an rth order Markov chain.

For any $\varepsilon t < z < (1-\varepsilon)t$ the $\mathrm{P}_{\mathbf{z}}$-distribution of the process $\{\ell_u - \ell_z : |u - z| < \varepsilon t\}$ does not depend on z. Using exponential bounds similar to those used in Lemma 1, we see that

$$\left|\mathrm{E}_z\left[\frac{\max_u e^{\ell_u}}{\sum_u e^{\ell_u}}\right] - \mathrm{E}_{t/2}\left[\frac{\max_u e^{\ell_u}}{\sum_u e^{\ell_u}}\right]\right| \leq \varepsilon,$$

where u ranges over $|u - z| < \varepsilon t$. By summing over all z's in the range $\varepsilon t < z < (1-\varepsilon)t$ and changing back the measure from $\sum_{z=\varepsilon t}^{(1-\varepsilon)t} \mathrm{P}_{\mathbf{z}}$ to the original null measure P_0, it can be shown that

$$\left|\mathrm{E}_z\left[\frac{\max_u e^{\ell_u}}{\sum_u e^{\ell_u}}\right] - \frac{1}{t}\mathrm{E}_0\Big[\max_u e^{\ell_u}\Big]\right| \leq \varepsilon.$$

Hence we obtain the representations [cf. (13)]

$$\lambda_r^* = \lim_t \frac{1}{t}\mathrm{E}_0\Big[\max_u e^{\ell_u}\Big] = \lim_t \frac{1}{t}\mathrm{E}_1\Big[\max_u \exp(\ell_u - \ell_1)\Big]. \tag{16}$$

Let P_∞ denote the extension of P_{t-r} to infinitely long sequences $\{x_i, y_i,\ 1 \leq i < \infty\}$. For $x > 0$ let $\tau_x = \inf\{u : \ell_u - \ell_1 > x\}$. By the argument of Hogan and Siegmund [(1986), Lemma 3.4] [see also Siegmund and Yakir (2000), Appendix A], we see that when $\ell_u - \ell_1$ is nonarithmetic the right-hand term in (16) equals

$$\lim_t t^{-1}\mathrm{E}_\infty(\ell_{t-r} - \ell_1) \lim_{x\to\infty} \mathrm{E}_\infty \exp\big[-(\ell_{\tau_x} - \ell_1 - x)\big]. \tag{17}$$

The first limit equals $I_1 = \theta^* \mathrm{E}_\infty[K(x_2, y_2) - K(x_2, y_{2+r})]$. The second is guaranteed to exist by the renewal theorem for additive functionals of a Markov chain [see Athreya, McDonald and Ney (1978)], applied to the process of ladder variables.

Comparison of (13) and (16) shows the relation of λ_0 and the λ_r, but evaluation of the latter is complicated by the fact that the underlying process involves a Markov chain instead of independent identically distributed random variables. Since the second factor in (17) is bounded by 1, the rough upper bound $\lambda_r \le I_1$ would yield a general asymptotic upper bound of the form of (5) with $\Lambda = I_1$. Asmussen (1989) has used the representation of the limit in (17) in terms of ladder variables to evaluate an analogous constant in a related problem. The problem is also discussed by Karlin and Dembo (1992), although their algorithm appears not to have been implemented. The representation (17) can also be made the basis of simulation of the λ_r, which can be accomplished in parallel for different values of r by virtue of the structure described above. Since $\zeta_2, \ldots, \zeta_{r+1}$ are independent and identically distributed, the limit of λ_r as $r \to \infty$ can be evaluated in terms of a random walk. This suggests to the optimist that it may be possible to approximate the different λ_r by a single value, which would lead to a substantially simpler overall approximation of the form of (5).

In the very special case that $K(\alpha, \beta)$ equals 1 or $-\xi$ according as α equals or is different from β, it is possible to evaluate $\lambda_1^* = \lambda_1^{**}$ explicitly. Recall from the final paragraph of Section 2 that in this case p_1 is defined to satisfy $\xi = \log[(1-p_0)/(1-p_1)]/\log(p_1/p_0)$, and $\theta^* = \log(p_1/p_0)$.

By simple algebra

$$\ell_u - \ell_1 = \theta^* \sum_{i=2}^{u} \bigl[K(x_i, y_i) - K(x_i, y_{i+r})\bigr] = \theta^*(1+\xi)S_u,$$

where $S_u = \sum_{i=2}^{u}[1\{x_i = y_i\} - 1\{x_i = y_{i+r}\}]$. The terms in the summation can take on only the values $1, 0$ and -1. Because of the simple structure of the increments of S_u, $S_\tau = x + 1$ for integral x whenever $\tau < t - r$. Hence trivial modifications in the argument leading to (17) to account for the difference between the arithmetic and nonarithmetic cases leads to

$$\bigl(1 - \exp(-\theta^*(1+\xi))\bigr) \lim_t t^{-1} \mathrm{E}_\infty[S_t],$$

which is easily evaluated to give

$$\lambda_r = \bigl(1 - \exp(-\theta^*(1+\xi))\bigr) \times \Bigl(p_1 - e^{-2\theta^*\xi} \sum_\alpha \mu_\alpha \nu_\alpha \bigl[\mu_\alpha(e^{\theta^*(1+\xi)} - 1) + 1\bigr]\bigl[\nu_\alpha(e^{\theta^*(1+\xi)} - 1) + 1\bigr]\Bigr).$$

APPENDIX

Here we include some technical lemmas that are used in the proofs of Theorems 1–3. The first two are given as general results about sums of independent random variables from a exponential family of distributions. The notation

used here for Lemmas 5–6 is local; it is not consistent with the notation used elsewhere in the paper.

Let $X_1, \ldots, X_k$ be independent random variables from an exponential family of distributions with likelihood ratio of the form $\exp[\theta x - \psi(\theta)]$. Assume that $\psi(0) = 0$, $\dot{\psi}(0) = I > 0$, and $\ddot{\psi}(0) = \sigma^2$. Assume also that the distribution of X_1 is nonlattice. Let $S_k = X_1 + \cdots + X_k$.

LEMMA 5. *For any $\varepsilon > 0$, for all sufficiently large a uniformly in $|kI - a| \le \varepsilon k$*

$$\begin{aligned} &\mathrm{E}_0\{\exp(-(S_k - a)); a \le S_k \le a + \log a\} \\ &\quad = \phi\left\{\frac{(kI - a)[1 + O(\varepsilon)]}{k^{1/2}\sigma}\right\}\frac{[1 + O(\varepsilon)]}{k^{1/2}\sigma}. \end{aligned} \tag{18}$$

PROOF. Fix $\delta > 0$. Then

$$\begin{aligned} &\mathrm{E}_0(\exp(-(S_k - a)); a \le S_k \le a + \log a) \\ &\quad \le \sum_{i=0}^{\lceil \log a/\delta \rceil} e^{-i\delta}\mathrm{P}_0(S_k \in [a + i\delta, a + (i+1)\delta]). \end{aligned} \tag{19}$$

For each i, let $\theta = \theta(k, a, i)$ be the solution to the equation

$$\mathrm{E}_\theta X_1 = \dot{\psi}(\theta) = (a + i\delta)/k. \tag{20}$$

A likelihood ratio identity yields the relation

$$\begin{aligned} &\mathrm{P}_0(S_k \in [a + i\delta, a + (i+1)\delta]) \\ &\quad = \mathrm{E}_\theta(\exp(-(\theta S_k - k\psi(\theta))); S_k \in [a + i\delta, a + (i+1)\delta]) \\ &\quad = \exp(k(\psi(\theta) - \dot{\psi}(\theta)\theta))\mathrm{E}_\theta(\exp(-\theta(S_k - a - i\delta)); S_k \in [a + i\delta, a + (i+1)\delta]) \\ &\quad \le \exp(k(\psi(\theta) - \dot{\psi}(\theta)\theta))e^{|\theta|\delta}\mathrm{P}_\theta(S_k \in [a + i\delta, a + (i+1)\delta]). \end{aligned}$$

Taylor expansion of ψ around θ yields, since $\psi(0) = 0$,

$$\psi(\theta) - \dot{\psi}(\theta)\theta = -\ddot{\psi}(\theta_1)\theta^2/2,$$

for some $\theta_1 \in [0, \theta]$. Moreover, from (20) we get that

$$(a + i\delta)/k - I = \dot{\psi}(\theta) - \dot{\psi}(0) = \theta\ddot{\psi}(\theta_2),$$

where, again, $\theta_2 \in [0, \theta]$. Hence by straightforward algebra, we obtain

$$k(\psi(\theta) - \dot{\psi}(\theta)\theta) = -\frac{(kI - a - i\delta)^2}{2k\sigma^2}[1 + O(\varepsilon)]. \tag{21}$$

Note also that since $|kI - a| < \varepsilon k$,

$$(kI - a - i\delta)^2/k > (kI - a)^2/k - 3\varepsilon\delta i. \tag{22}$$

To bound the probability

$$\mathrm{P}_\theta(S_k \in [a + i\delta, a + (i+1)\delta]) = \mathrm{P}_\theta(S_k - a + i\delta \in [0, \delta]) \tag{23}$$

we consider separately the terms in (19) with $i \le c$ and with $i > c$. For the former we apply a local limit theorem [e.g., Durrett (1996), page 134] and for the latter the upper bound constant times $k^{-1/2}$ [cf. Chung (1974), page 177]. Using these bounds in conjunction with (20)–(23), letting $c \to \infty$, then $\delta \to 0$, we obtain an asymptotic upper bound of the form indicated in the statement of the lemma. A lower bound follows by a similar argument. $\square$

Now suppose $\theta > 0$ and let $\ell_k = \theta S_k - k\psi(\theta)$.

LEMMA 6. *For any positive μ for any real x,*

$$\mathrm{P}_0(\ell_k \ge x) \le \exp\big\{-(\mu/\theta)x - k[(\mu/\theta)\psi(\theta) - \psi(\mu)]\big\}.$$

PROOF. Simple algebra leads to the relation

$$\big\{\theta S_k - k\psi(\theta) \ge x\big\} = \big\{\mu S_k - k\psi(\mu) \ge (\mu/\theta)x + k[(\mu/\theta)\psi(\theta) - \psi(\mu)]\big\}.$$

The lemma thus follows from an exponential Chebyshev's inequality, since $\mathrm{E}_0 \exp\{\mu S_k - k\psi(\mu)\} = 1$. $\square$

We now return to the notation used earlier in the paper. In particular k denotes the number of aligned pairs in a candidate alignment $\mathbf{z}$. The following lemma gives a bound on the tail of the distribution of $\ell_{\mathbf{z}}$ that is useful in showing that for an optimal alignment, k must be about a/I.

LEMMA 7. *There exists a constant c such that for any η and for any real x,*

$$\mathrm{P}_0(\ell_{\mathbf{z}} \ge x) \le \exp[-x - \eta(kI - x) + kc\eta^2].$$

In particular,

$$\mathrm{P}_0\{\ell_{\mathbf{z}} \ge a + g(\mathbf{z})\} \le \exp[-a - \eta(kI - a) - (1-\eta)g(\mathbf{z}) + kc\eta^2].$$

PROOF. Assume $\eta > 0$. Choose $\mu = \theta(1-\eta)$. A Taylor expansion of the function ψ around θ gives

$$\psi(\mu) = \psi(\theta) - \eta\theta\psi'(\theta) + \theta^2\eta^2\psi''(\tilde{\theta})/2,$$

for some $\theta(1-\eta) \le \tilde{\theta} \le \theta$. The proof follows from Lemma 6 with $\mu = \theta(1-\eta)$ and from the fact that $\psi''(\tilde{\theta})$ is bounded. Essentially the same proof also applies for $\eta < 0$. $\square$

LEMMA 8. *Let $g(\mathbf{z}) = \theta^*\delta l$ and let $\mathscr{J}$ be the set of all matching patterns having at most j gaps. Then*

$$\frac{P_0(\max_{\mathbf{z}\in\mathscr{J}_j}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a)}{P_0(\max_{\mathbf{z}\in\mathscr{J}}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a)} \longrightarrow 1 \qquad \text{as } a \to \infty. \tag{24}$$

PROOF. Since the left-hand side of (24) is bounded from above by 1, it is enough to show that

$$P_0\Big(\max_{\mathbf{z}\in\mathscr{Z}\setminus\mathscr{Z}_j}(\ell_{\mathbf{z}}-g(\mathbf{z}))\ge a\Big)=o(nma^je^{-a}),$$

as $a\to\infty$.

Recall that $\mathscr{Z}$ is a union of the subcollections $\mathscr{Z}_{i,l,k}$ and that the cardinality of $\mathscr{Z}_{i,l,k}$ is approximately $mn2^i\binom{k-1}{i}\binom{l-1}{i-1}$. Also, for any $0\le i\le j$, $\mathscr{Z}_i=\cup_{k,\,i\in\mathscr{I}}\mathscr{Z}_{i,l,k}$, with $\mathscr{I}=\{k,l\colon l\le\varepsilon_2a^{1/2},\ |k-a/I|\le\varepsilon_1a/I\}$. The same considerations that led to earlier results can be used in order to show that

$$P_0\Big(\max_{\mathbf{z}\in\mathscr{Z}_i}(\ell_{\mathbf{z}}-g(\mathbf{z}))\ge a\Big)=O(mna^ie^{-a}).$$

It follows that

$$\sum_{i<j}P_0\Big(\max_{\mathbf{z}\in\mathscr{Z}_i}(\ell_{\mathbf{z}}-g(\mathbf{z}))\ge a\Big)=o(mna^je^{-a}).$$

Let $\widetilde{\mathscr{I}}_i$ denote the set of all pairs (k,l) which do not belong to $\mathscr{I}_i$. For any i, $0\le i\le j$, consider $\widetilde{\mathscr{Z}}_i=\cup_{l,\,k\in\widetilde{\mathscr{I}}_i}\mathscr{Z}_{i,k,l}$. From Lemma 7, first taking $\eta=a^{-1/2}$, then $\eta=-a^{-1/2}$, and taking ε_1 sufficiently small compared to ε_2, we obtain

$$\begin{aligned}
&(mn)^{-1}e^a\mathrm{P}_0\Big(\max_{\mathbf{z}\in\widetilde{\mathscr{Z}}_i}(\ell_{\mathbf{z}}-g(\mathbf{z}))\ge a\Big)\\
&\le\sum_{k,l\in\widetilde{\mathscr{I}}_i}2^i\binom{k-1}{i}\binom{l-1}{i-1}e^a\mathrm{P}_0(\ell_{\mathbf{z}}\ge a+\theta^*\delta l)\\
&\le\sum_{k>(1+\varepsilon_1)(a/I)}\ \sum_{l<\varepsilon_2a^{1/2}}2^i\binom{k-1}{i}\binom{l-1}{i-1}\exp(-\theta^*\delta l(1-a^{-1/2})-(kI-a)/a^{1/2}+ck/a)\\
&\quad+\sum_{|kI-a|<\varepsilon_1a}\ \sum_{l>\varepsilon_2a^{1/2}}2^i\binom{k-1}{i}\binom{l-1}{i-1}\exp(-\theta^*\delta l(1-a^{-1/2})-(kI-a)/a^{1/2}+ck/a)\\
&\quad+\sum_{k<(1-\varepsilon_1)(a/I)}\ \sum_{l<\varepsilon_2a^{1/2}}2^i\binom{k-1}{i}\binom{l-1}{i-1}\exp(-\theta^*\delta l-\varepsilon_1a^{1/2}+c/I)\\
&<[(ca)^{i+1}/i!][\exp(-\varepsilon_1a^{1/2}/2)+\exp\{-(\varepsilon_2^2-\varepsilon_1)a^{1/2}+ci\}+\exp(-\varepsilon_1a^{1/2})],
\end{aligned}$$

hence the proof. □

The following four lemmas provide appropriate upper bounds for Theorems 2 and 3. Let $g(\mathbf{z})=\theta^*(\delta l+\Delta j)$, and assume as in Theorems 2 and 3 that $\theta^*\Delta=\log(a)+C$. Let j_1 be a large integer, to be specified later, and $j_2=(\log a)^2$.

LEMMA 9. *For $\widetilde{\mathscr{D}}_j$ defined as in Lemma* 8,

$$\sum_{j<j_2} \mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{D}_j,\, k<ca}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big)$$

divided by the right-hand side of (4) *converges to* 0 *as* $a \to \infty$.

The proof of Lemma 9 follows the pattern of Lemma 8, except now we must also incorporate the cost of j gap intervals and sum over j. A review of the proof of Lemma 8 shows that the final inequality holds uniformly in $i < j_2$ provided $k < ca$. The rest is straightforward. □

LEMMA 10. *As* $a \to \infty$,

$$\mathrm{P}_0\Big(\max_{\mathbf{z}\in\mathscr{D},\, j>j_2,\, k<ca}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big) = o(mne^{-a}).$$

To prove Lemma 10, we use the inequality

$$\mathrm{P}_0\{\ell_{\mathbf{z}} - g(\mathbf{z}) \ge a\} \le \exp(-a - g(z)),$$

to see that the probability of interest is bounded by

$$mne^{-a} \sum_{j>j_2} \sum_{l\ge j} \sum_{k\le ca} 2^j k^j \binom{l-1}{j-1} \exp(-\theta^*\{\Delta j + \delta l\})/j!,$$

which simple calculations show is bounded by

$$mne^{-a} \sum_{j>j_2} (2ca)^{j+1} \exp(-\theta^* j\Delta)/[e^{\theta^*\delta} - 1]^j/j!.$$

Since $e^{-\theta^*\Delta} = e^C/a$, the series is easily seen to converge to 0 (e.g., by a tail estimate for the Poisson distribution). □

LEMMA 11. *For sufficiently large* c,

$$\mathrm{P}_0\Big\{\max_{\mathbf{z}\in\mathscr{D},\, k>ca}(\ell_{\mathbf{z}} - g(\mathbf{z})) \ge a\Big\} = o(mne^{-a}).$$

To prove Lemma 11 we use again the bound of Lemma 7 with $\eta = a^{-1/2}$ to get the bound

$$mne^{-a}e^c \sum_{k>ca} \exp(-(kI - a)/a^{1/2}) \sum_{j\ge 0} \sum_{l\ge j} \frac{2^j k^j}{j!} \binom{l-1}{j-1} \times \exp(-(1 - a^{-1/2})\theta^*(\Delta j + \delta l)).$$

This time we sum first over j to obtain an exponential factor bounded by $\exp(k/a^{1-\varepsilon})$, then sum over the values $k > ca$. □

LEMMA 12. *For arbitrary* $\varepsilon > 0$, *for all sufficiently large* j_1,

$$\sum_{j=j_1}^{j_2} \sum_{\mathbf{z}\in\mathscr{D}_j} \mathrm{P}_0\{\ell_z \geq a + g(\mathbf{z})\} < \varepsilon mne^{-a}.$$

To prove Lemma 12 we write

$$\mathrm{P}_0\{\ell_z \geq a + g(\mathbf{z})\} = \exp[-a - g(\mathbf{z})]\mathrm{E}_{\mathbf{z}}\{\exp\{-[\ell_{\mathbf{z}} - a - g(\mathbf{z})]\}; [\cdots] \geq 0\}$$

and reason as in the proof of Lemma 4 and the now established pattern of Lemmas 8–11 to obtain the stated result. □

Acknowledgments. The first author thanks the Australian National University and the University of Cambridge for their hospitality and Richard Mott for helpful conversations about the subject of the paper. The authors thank two referees for their careful reading and helpful suggestions concerning the manuscript.

REFERENCES

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. and LIPMAN, D. J. (1990). Basic local alignment search tool. *J. Molecular Biol.* **215** 403–410.

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

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. *Nucleic Acids Research* **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 Markovian 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.

CHUNG, K. L. (1974). *A Course in Probability Theory.* Academic Press, New York.

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 Univ. Press.

DURRETT, R. (1990). *Probability: Theory and Examples*. Duxbury Press, Belmont, CA.

HOGAN, M. and SIEGMUND, D. (1986). Large deviations for the maximum of some random fields, *Adv. in Appl. Math.* **7** 2–22.

KARLIN, S. and DEMBO, A. (1992). Limit distributions of maximal segmental score among Markov-dependent parital sums. *Adv. in Appl. Probab.* **24** 113–140.

LEZAUD, P. (1998). Chernoff-type bound for finite Markov chains. *Ann. Appl. Probab.* **8** 849–867.

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.

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

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer, New York.

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

SMITH, T. F. and WATERMAN, M. S. (1981). Identification of common molecular subsequences. *J. Molecular Biol.* **147** 195–197.

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. *Statist. Sci.* **9** 367–381.

WILLIAMS, D. (1991). *Probability and Martingales*. Cambridge Univ. Press.

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.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
E-MAIL: dos@stat.stanford.edu

DEPARTMENT OF STATISTICS
THE HEBREW UNIVERSITY
JERUSALEM
ISRAEL
E-MAIL: msby@mscc.huji.ac.il