Approximate Statistics of Gapped Alignments

RICHARD MOTT¹ and ROGER TRIBE²

ABSTRACT

A heuristic approximation to the score distribution of gapped alignments in the logarithmic domain is presented. The method applies to comparisons between random, unrelated protein sequences, using standard score matrices and arbitrary gap penalties. It is shown that gapped alignment behavior is essentially governed by a single parameter, \( \alpha \), depending on the penalty scheme and sequence composition. This treatment also predicts the position of the transition point between logarithmic and linear behavior. The approximation is tested by simulation and shown to be accurate over a range of commonly used substitution matrices and gap-penalties.

Key words: alignment, significance, probability, sequences

1. INTRODUCTION

This paper addresses the problem of how to determine the statistical significance of gapped alignments between protein sequences. It is well known that, in general, correct alignments between real, related sequences may contain gaps, which could reasonably have arisen by insertions or deletions in ancestral sequences. Consequently, routine use is made of sequence comparison algorithms, which can insert gaps as appropriate in order to maximize an alignment score. We are interested in the distribution of these alignment scores between random unrelated sequences, so that we may judge the significance of alignments between real sequences objectively. For surveys of this subject, see Waterman (1995); Durbin et al. (1998); and Apostolico and Giancarlo (1998).

The problem can be summarized as follows: The inputs are two finite sequences of symbols, a symbol scoring matrix and a gap-penalty function. Any local alignment of the sequences can be scored and ranked according to this matrix and penalty. One is typically interested either in the maximum-scoring local alignment (Smith and Waterman, 1981; Gotoh, 1982), or in all nonintersecting local alignments scoring greater than some threshold value (Altschul et al., 1990, 1997). We want to know whether an observed score could have arisen by chance under an appropriate model of sequence randomness, which for protein sequences is usually that each amino acid residue occurs with a certain frequency independently of its neighbors.

It is this classical, frequentist view of the problem that we will consider here. Radically different approaches to assessing statistical significance are described elsewhere (Zhu et al., 1998, Bayesian; and Milosavljevic, 1994, algorithmic significance).

Currently, there only exists a usable rigorous theory of classical alignment statistics when gaps are not allowed (Arratia et al., 1988; Karlin and Altschul, 1990; Karlin and Dembo, 1992; Dembo et al., 1994). The basic result, a Poisson approximation, is that if the expected score for aligning two symbols at random is negative, then there exist positive constants \( K, \lambda \), depending on the score matrix and the sequences

¹Wellcome Trust Centre for Human Genetics, Oxford, U.K.
²Mathematics Institute, University of Warwick, Coventry, U.K.
compositions, such that the distribution function of an ungapped score \( Z \) between long random sequences of lengths \( m, n \) follows an extreme-value distribution as \( m, n, t \to \infty, \log(m/n) \to 1 \):

\[
\Pr(Z > t) \sim 1 - \exp(-Kmne^{-\lambda t})
\]

(1)

This approximation can be understood by thinking of the score \( Z \) as the maximum of \( mn \) weakly dependent variables, each with an exponential tail probability, with \( K \) measuring the degree of dependence. For large \( t \), the formula can be simplified to \(-\log \Pr(Z > t) \sim -\log(Kmn) + \lambda t\), i.e., \( \log p \)-values are linearly related to scores.

Two complicating factors must be taken into account when applying this formula: (i) alignment scoring schemes take discrete values, so no fixed value of \( K \) gives a good approximation for all \( t \), and it must be replaced by calculable values \((K^-, K^+)\) that give upper and lower bounds. (However, provided \( t \) is restricted to achievable score values only, then \( K^- \) gives the correct approximation.) (ii) \( m, n \) should be replaced by effective lengths \( m', n' \) which allow for the fact that high-scoring alignments cannot start too near the end of either sequence.

There are theoretical results for special cases of the gapped alignment problem (Karlin and Altschul, 1993; Neuhauser, 1994), which do not use gap-penalties explicitly, but little has been proved in the general case. One cannot for example write down a formula for the distribution of optimal Smith-Waterman alignment scores even for a simple affine gap penalty. However, there is a great deal of empirical evidence (Mott, 1992; Waterman and Vingron, 1994b; Altschul and Gish, 1996; Spang and Vingron, 1998) indicating that the extreme-value theory underlying the ungapped case carries over, provided the gap-penalties are severe enough. The unsolved problem is to compute new values of \( K, \lambda \) and the effective lengths \( m', n' \), all of which will now depend on the gap penalty as well.

In current practice, the statistical significance of gapped alignments is estimated by a variety of techniques which broadly fall into two categories: (i) fitting a parametric exponential or extreme-value distribution to observed scores (either from a databank search or from simulations; Pearson and Lipman, 1988; Collins et al., 1988; Mott, 1992; Waterman and Vingron, 1994b; Altschul and Gish, 1996; Pearson, 1998; Spang and Vingron, 1998); (ii) sum statistics that combine the \( p \)-values of nearby ungapped similarities in a consistent way, but which do not use gap penalties (Altschul and Gish, 1996; Karlin and Altschul, 1993; Altschul et al., 1997).

The asymptotic formula (1) for ungapped alignments only applies when the substitution matrix has negative expected score. It does not hold when this value is positive. When the expectation passes through zero a phase transition occurs: instead of defining short regions of high similarity, alignments frequently span the entire lengths of the sequences, even when no genuine similarity is present. A qualitatively similar phenomenon occurs with gapped alignment (Arratia and Waterman, 1994; Waterman et al., 1987; Drasdo et al., 1998). For strong or infinite gap-penalties, the expected score grows logarithmically with sequence length, but as the gap penalties are decreased a threshold is reached, beyond which scores grow linearly with length. Consequently, a useful theory of gapped alignments should not only compute the parameters but also predict the transition point.

In this paper, an approximate statistical model for gapped alignments in the logarithmic domain is presented. It predicts gapped behaviour, including transition points, and provides heuristic formulas for assessing significance without the need for simulation or model fitting. The predictions are shown by numerical simulation to be quite accurate over a range of commonly used scoring matrices and gap penalties.

### 2. STATISTICS OF UNGAPPED ALIGNMENTS

We first recapitulate the statistical theory of matching random sequences when gaps are not allowed. We shall give a heuristic treatment of these results that introduces concepts, quantities, and arguments that are needed in the case of gapped alignments. A rigorous treatment is found in (Arratia et al., 1988; Dembo and Karlin, 1991; Karlin and Dembo, 1992; Dembo et al., 1994).

Consider the comparison of two random, but independent infinite sequences \( U_i, V_j \), of lengths \( m, n \). In the first sequence, the letter \( a \) occurs with probability \( p_a \), and in the second with probability \( q_a \). In both sequences, all the letters are independently distributed. The score for matching letters \( a, b \) is \( S_{ab} \). The mass function \( h(x) \) of the score, \( M \), obtained by drawing a letter at random from each sequence is

\[
\Pr(M = x) = h(x) = \sum_{ij:S_{ij}=x} p_i q_j
\]

(2)
which is assumed to have negative expectation, i.e., \( \sum_{a,b} p_{a,b} S_{a,b} < 0 \). Each element of the \((m \times n)\) "dot matrix" \( M_{ij} \equiv S_{i,j} \) is a random variable with mass function \( h(x) \). The diagonal \( d \) consists of all pairs \((i, j)\) such that \( i - j = d \). Along a diagonal \( d \) the \( M_{ij} \) are jointly statistically independent, although there are small (and ignorable) dependencies across diagonals.

The statistical theory of ungapped alignments is based on the behavior of sums of consecutive match scores along diagonals. For fixed coordinate \((i, j)\), define the partial sums

\[
Y_{ijk} = \begin{cases} 
0 & k = 0 \\
\sum_{n=0}^{n<k} M_{i+n,j+n} & k > 0
\end{cases}
\]

i.e., \( Y_{ijk} \) is the accumulated score of matching a string of \( k \) consecutive letter pairs, starting with the pair \((i, j)\), the \( i \)th letter from the first sequence and the \( j \)th from the second, and continuing along the diagonal \( d \). As \( k \) varies, this generates a random walk with negative drift. Because of this drift, \( Y_{ijk} \rightarrow -\infty \) as \( k \rightarrow \infty \) with probability 1, and indeed the walk frequently never attains a positive score. Of primary interest is the maximum attained by this walk. For reasons which will become clear later, one must distinguish between two types of maximum:

(i) The unconstrained maximum sum \( Z'_{ij} = \max_{k \geq 0} Y_{ijk} \). That is, \( Z'_{ij} \) is the score of the maximum partial sum starting at \((i, j)\), or zero. Note that, as \( k \) increases, the partial sums may take negative values before achieving their maximum score. Define \( C'_{ij} \) to be the coordinate where the walk first attains its maximum, and set \( C'_{ij} = (i, j) \) if the walk never attains a positive score. The nontruncated maximum \( Z''_{ij} = \max_{k > 0} Y_{ijk} \) will also be needed.

(ii) Let \( T_{ij} \) be the first value of \( k \) for which \( Y_{ijk} \) is negative. Then the constrained sum \( Z_{ij} \) is the maximum attained by the \( Y_{ijk} \) such that \( k \leq T_{ij} \), i.e., \( Z_{ij} = \max_{k \leq T_{ij}} Y_{ijk} \). \( C_{ij} \) is defined similarly. Clearly, \( Z_{ij} \leq Z'_{ij} \) and \( C_{ij} \) occurs before \( C'_{ij} \) (Figure 1).

The random variables \( Z', Z \) have related tail behavior for their distributions. First consider the unconstrained walk: fix \((i, j)\) and set \( Z' \equiv Z'_{ij}, M_k \equiv M_{i+k,j+k} \). Let \( G(t) = \Pr(Z' \leq t) \). Then, for \( t > 0 \),

\[
G(t) = \Pr(\max(0, M_0, M_0 + M_1, M_0 + M_1 + M_2, \ldots) \leq t) \\
= \Pr(M_0 + \max(0, M_1, M_1 + M_2, \ldots) \leq t) \\
= \sum_{m=-\infty}^{m \leq t} \Pr(M_0 = m) \Pr(\max(0, M_1, M_1 + M_2, \ldots) \leq t - m) \\
= \sum_{m=-\infty}^{m \leq t} h(m) G(t - m)
\]  

(3)

FIG. 1. A realization of a random walk with negative drift, illustrating the constrained and unconstrained maxima, at coordinates \((C,Z)\) and \((C',Z')\), respectively. Descending ladder points are marked with circles.
This is Lindley’s dam equation, also known as the Weiner-Hopf equation. The distributions of \( Z' \), \( Z'' \) are the same for \( t > 0 \). [Note that the probability that the walk never attains a positive or zero score is \( \Pr(Z'' < 0) = G(0) - \Pr(Z'' = 0) \).

The dam equation has an asymptotic solution
\[
(1 - G(t))e^{\lambda t} \to s \text{ as } t \to \infty
\]  
for some strictly positive constant \( s \), and where \( \lambda \) is the positive root of the equation
\[
1 = E(e^{\lambda M}) = \sum_m h(m)e^{\lambda m}
\]  
Feller (1972) and Iglehart (1972) provide a formula for \( s \) in terms of fluctuation sum identities. (Appendix 7.1). Alternatively, \( s \) may be obtained via a recurrence (Appendix 7.2).

Iglehart (1972) proved that the distribution of the constrained maximum, \( F(t) = \Pr(Z \leq t) \), has the similar asymptotic form
\[
(1 - F(t))e^{\lambda t} \to r \text{ as } t \to \infty
\]  
where \( \lambda \) is the same as above but with a different formula for \( r \) (Appendix 7.1).

Karlin and Dembo (1992) extended Iglehart’s results to derive statistics for ungapped similarity scores. First, we define a local ungapped similarity to be a section of a diagonal, starting at \((i, j)\) and ending at \((i + k - 1, j + k - 1)\) say, such that no other section of the diagonal intersecting this region has a higher score. The maximal local similarity over all the diagonals, with score \( Z_{\text{max}} \), is that found by the Smith-Waterman algorithm when gaps are forbidden.

Choose one fixed diagonal and consider the random walk generated by the partial sums \( Y_1, Y_2, \ldots \) starting from the left-hand end of the diagonal (using an obvious notation). \( Y_n \) is a descending strict ladder point if \( Y_n < Y_{n+1} \) for all \( r < n \). Define a clump to be an excursion in the random walk above its previous ladder point \( n \). The score of the clump is the maximum height attained above \( Y_n \) during the excursion. These clumps form identically distributed, independent events.

The key observation is that the maximum clump score agrees with the maximum of the scores of the local ungapped similarities. Since successive clump scores are independent, they are easier to study analytically. The term “high scoring pair” (HSP) is normally defined as a local ungapped similarity, but in the light of the above observation we will use it to refer to that segment of a clump that achieves the clump score.

Under this definition, the start of each HSP must be immediately preceded by a ladder point. Let \( l \) be the expected length of interval between successive ladder points (including immediately adjacent ladder points of length 1, corresponding to HSPs with score 0). Then the probability that a randomly-selected coordinate marks the start of an HSP is \( \rho = 1/l \), which may be calculated by Equation (25) (Asmussen, 1987). It then follows that there are on average \( pmn \) HSPs on the dot-matrix, and \( Z_{\text{max}} = \max_{ij} Z_{ij} \) will be distributed like the maximum of \( pmn \) independent copies of the random variable \( Z \), provided one ignores the weak dependence between diagonals. Equation (1) then follows.

Here we introduce another line of argument, related to the Poisson Clumping Heuristic (Aldous, 1989; Waterman and Vingron, 1994b; Waterman, 1995), which also computes the probability that a given coordinate is at the start of a HSP, and show the two methods are equivalent. This will be needed later for gapped alignments.

Suppose \( n + 1 \) marks the start of a clump. Imagine constructing the backwards partial sums \( M_n, M_n + M_{n-1}, M_n + M_{n-1} + M_{n-2}, \ldots \). This generates a random walk with the same distribution as the forward walk considered above. Take the nontruncated unconstrained maximum \( Z'' \) over this walk. Then \( n + 1 \) marks the start of a clump if \( Z'' < 0 \). For if \( Z'' \geq 0 \), then the local alignment containing \( n + 1 \), but starting where the backwards walk attained its maximum, has at least as high a score. Conversely, if \( Z'' < 0 \) then the score of an alignment containing \( n + 1 \) cannot be improved by starting earlier. Therefore, we claim \( \rho = \Pr(Z'' < 0) = 1/l \). This equality is in fact a standard result in random walk theory (Asmussen, 1987); \( \Pr(Z'' < 0) \) is the probability that a walk starting from zero never attains a positive score, and \( l \) is the expected time for the walk to first take a negative value. For a numerical example, see Appendix 7.2.

Now the overall maximum \( Z_{\text{max}} \) is if there are no clumps with scores \( > t \). The expected number of such clumps is
\[
\mu(t) = mnp(1 - F(t)) \sim mnp e^{-\lambda t},
\]
ignore length correction effects (Altschul and Gish, 1996). Under the assumption that the diagonals may be treated as independent, the Poisson Clumping Heuristic says that when \( t \) is large the number of such clumps will have a Poisson distribution approximately, so

\[
\Pr(Z_{\text{max}} < t) = e^{-\mu(t)} \sim \exp(-mn\rho \rho)
\]

relations is \( K^* \) in (Karlin and Altschul, 1990; Karlin and Dembo, 1992).

As noted above, the discrete nature of the scoring scheme means that \( r \) should in fact be replaced by upper and lower bounds

\[
(K^-, K^+) = (\rho, e^{\delta} \rho)
\]

where \( \delta \) is the smallest span of score values (usually 1), because

\[
\Pr(Z_{\text{max}} \in (t, t + \delta)) \sim \exp(-mn\rho \rho) - \exp(-mn\rho \rho)
\]

where \( \delta \) is the smallest span of score values (usually 1), because

\[
\Pr(Z_{\text{max}} \in (t, t + \delta)) = \exp(-mnK^+e^{-\lambda}) - \exp(-mnK^-e^{-\lambda})
\]

Length correction effects are discussed later.

3. THE GREEDY EXTENSION MODEL (GEM)

Now consider the distribution of the maximal local similarity when gaps are allowed. In outline our approach is: (a) to approximate the optimal Smith-Waterman local alignment score \( W_{SW} \) by the score \( W_{GEM} \), of a slightly suboptimal alignment generated by a certain greedy algorithm; (b) find an approximation to the greedy algorithm’s distribution function \( P_{GEM}(t) \approx \Pr(W_{GEM} \leq t) \); (c) it is plausible that \( 1 - P_{GEM}(W_{SW}) \) is a useful estimate of the statistical significance of \( W_{SW} \), provided both of these approximations are close when the sequences compared are random. This is the key point. Of course, any algorithm that scores alignments the same way as the Smith-Waterman must return an alignment with score not exceeding the optimal alignment’s. Consequently \( P_{GEM}(t) \geq \Pr(W_{SW} \leq t) \) and the method may overestimate statistical significance.

It may seem inappropriate at first sight to consider greedy algorithms because they often give poor results when applied to related sequences, and certainly their use is not recommended in practice. The greedy algorithm is introduced solely as a model to help us understand alignments between random, independent sequences, where we claim its behaviour is close to the optimal alignment’s, but with the advantage that its statistical properties are more tractable.

Neuhauser (1994) described a greedy algorithm for aligning DNA sequences that does not allow mismatches and limits gaps to have a maximum length. Using the Chen-Stein method, a rigorous Poisson approximation of the distribution of the scores generated by the algorithm was obtained. The algorithm described here is similar in spirit to Neuhauser’s, but deals with the more realistic situation of general substitution matrices (with negative expectation) and arbitrary gap penalties, in particular the popular affine penalty \( g(n) = A + Bn \), where \( n \) is the length of the gap and \( A, B \) are gap-open and gap-extension parameters. Our mathematical treatment is based on the random walk theory outlined above.

GEM is defined as follows (Figure 2):

(i) Imagine the dot-matrix surface of comparison is very large, and for the moment just consider local alignments well away from the edges. The algorithm begins by considering as potential starting positions only those coordinates \( (i, j) \) that are starts for ungapped HSPs. Using the substitution scoring matrix \( S \), construct the optimal ungapped local alignment starting at \( (i, j) \), which ends at \( C_{ij} = (k - 1, l - 1) \) say. This will have score \( Z_{ij} = Z \), say, which was analysed earlier. By definition, no improvement to \( Z_{ij} \) is possible by extending further down this diagonal.

(ii) From \( (k, l) \), scan the neighboring diagonals on either side to find the best place to extend the alignment. To reach a diagonal requiring the insertion of \( n \) gaps in the either sequence, i.e., starting at coordinates \( (k, l + n) \) or \( (k + n, l) \), one must pay a penalty \( g(n) \). To get the value of this particular extension, subtract \( g(n) \) from the maximal ungapped diagonal score \( Z'_{k,l+n} \) (say) that would be obtained by running the ungapped unconstrained algorithm from this location. The greedy extension from \( (k, l) \) is defined as the extension with maximum score \( U_1 \) over all possible insertions in either sequence, i.e.,

\[
U_1 = \max_{n>0}(Z'_{k,l+n} - g(n), Z_{k+n,l} - g(n))
\]
One must use the unconstrained maxima $Z'$, because in general any path connecting the one HSP to another may need to pass through a negative-scoring diagonal segment. After this step, the end of the alignment will have moved to $C'_{k,l+n}$, say, and the score will be $Z + U_1$.

It is possible $U_1 < 0$—in fact, it will be a requirement in what follows that the average greedy extension score is negative—but the algorithm still takes the step, even if it reduces the score of the alignment. A special case occurs when the greedy extension step does not move diagonally, i.e., comprises just a gap of length 1 with cost $-g(1)$. When this happens, the algorithm must stop because it is unable to extend the alignment. We call this condition **sticking**.

(iii) Assuming sticking has not occurred, the algorithm proceeds by repeating the greedy extension step, producing a series of moves with scores $U_1, U_2, \ldots$, generating the $U$-walk $U_1, U_1 + U_2, \ldots$. Thus a GEM local alignment starts at an HSP and comprises a series of (gap, diagonal) pairs such that each pair is locally the best move possible. Once the algorithm sticks, then all successive steps are null.

(iv) It remains to define the termination point in the $U$-walk. Let $T(Z)$ be the first time the $U$-walk drops below $-Z$, or the time till sticking. In a manner analogous to the ungapped algorithm, the score $W$ for the greedy alignment starting at $(i, j)$ is defined as the maximum over all the constrained partial sums

$$W(Z) = \max(0, Z, Z + U_1, Z + U_1 + U_2, \ldots, Z + U_1 + \cdots + U_{T(Z)}),$$

$$= Z + \max(0, U_1, U_1 + U_2, \ldots, U_1 + \cdots + U_{T(Z)}) = Z + V(Z),$$

say.

This construction ensures that the alignment is locally optimal; if the $U$-walk were allowed to dip below $-Z$ then a higher-scoring local alignment could be formed by removing that part of the walk prior to this point. This is the same as saying there is no ladder point in the walk 0, $Z, Z + U_1, Z + U_1 + U_2, \cdots$ until some time after $T(Z)$.

[If $V$, $V'$ are the constrained and unconstrained maxima attained by the $U$-walk, then $V \leq V(Z) \leq V'$. The asymptotic behavior of all three random variables is similar (Iglehart, 1972), a fact which will be exploited later.]

(v) The maximal GEM alignment is the maximum-scoring greedy alignment, with score $W_{GEM}$ say, found by running the algorithm from all starting positions.
There are two immediate reasons why the GEM algorithm may under-perform relative to the optimal Smith-Waterman:

(a) the algorithm keeps on a diagonal \( d \) for as long as possible, even if jumping sooner onto another diagonal \( d' \) would improve the score. The gain in jumping earlier is equal to the difference in scores of two parallel diagonal sections of the same length; if this is very large then it is likely the previous greedy extension step would have preferred \( d' \) over \( d \) in the first place.

(b) If there is a choice of more than one diagonal, then taking the local greedy step may trap the algorithm into a local minimum. This is a more serious objection. Nevertheless, one is primarily interested in the case of strong gap penalties, when it is very rare for more than one possible extension to have a positive score.

Informally, one may argue that only high-scoring alignments are of interest. These are dominated by very highly scoring collections of nearby HSPs, which are scarcer than individual HSPs and hence unlikely to overlap. The optimal Smith-Waterman alignment is obtained by linking these up and the method of linking is not crucially important to the score.

Now the successive extension scores \( U \) are identically distributed, but they are not independent, because (a) the weak dependence between diagonals, and (b) each greedy extension from \((i, j)\) searches through the entire quadrant beyond \((i, j)\). This imposes a constraint on the subsequent extensions, “sticking” being the extreme case. However, if an extension is long then the constraint is weak, and the maximum-scoring alignment is likely to involve longer runs along diagonals. Therefore, the higher-scoring alignments will resemble realizations from a random walk with negative drift.

Consequently, by using a GEM and making the approximation that successive extensions are independent, the mathematics involved in analysing the distribution of the score \( W_{GEM} \) becomes more tractable, indeed it is the same mathematics as for the ungapped case, once we know how the score of a single greedy extension is distributed.

### 3.1. Distribution of the greedy extension score

Let \( \phi(t) \) be the probability that a single greedy extension has score \( \leq t \). Treating the diagonals as independent, and approximating \( G(t) \) to be exponential \( \forall t \geq g(1) \), Equation (9) implies

\[
\phi(t) = \prod_{n>0} G(t + g(n))^2 = \prod_{n>0} (1 - se^{-\lambda(t+g(n))})^2
\]

In Appendix 7.2, a recursion, Equation (26), is derived which computes \( G(t) \) exactly for small \( t \), so we are able to check the assumption that \( G(t) \) is exponential for small \( t \). We found that the rate at which \((1 - G(t))e^{\lambda t}\) approaches \( s \) depends on the “smoothness” of \( h(\cdot) \), i.e., of the substitution matrix \( S \). The approximation holds even for \( t \approx 1 \) for the PAM250 matrix; in contrast for PAM120 there is quite violent oscillation before convergence at \( t \approx 10 \).

For the moment restrict attention to the affine gap penalty \( g(n) = A + Bn \). Then

\[
\phi(t) = \prod_{n>0} (1 - se^{-\lambda(t+1) + \lambda Bn})^2
\]

Set \( \beta(t) = se^{-\lambda(t+1)} \). Then, taking logs,

\[
\log \phi(t) = 2 \sum_{n>0} \log(1 - \beta(t)e^{-\lambda Bn})
\]

\[
= -2 \sum_{n>0} \sum_{m>0} \frac{\beta(t)^m e^{-\lambda Bmn}}{m}
\]

\[
= -2 \sum_{m>0} \frac{\beta(t)^m}{m} \sum_{n>0} e^{-\lambda Bmn}
\]

\[
= -2 \sum_{m>0} \frac{\beta(t)^m}{m} \frac{e^{-\lambda Bm}}{(1 - e^{-\lambda Bm})}
\]

Equation (11) cannot be summed explicitly, but we can show that \( \phi_1(t) \leq \phi(t) \leq \phi_2(t) \forall t \), where

\[
\phi_1(t) \equiv \max 0, (1 - a e^{-\lambda t})^2
\]

\[
\phi_2(t) \equiv \exp(-ae^{-\lambda t})
\]
FIG. 3. The distribution function of a single greedy extension, for the BLOSUM62 substitution matrix and \( A = 10, B = 2 \). \( \phi \) is the product formula from Equation (10). \( \phi_1, \phi_2 \) are the lower and upper bounds from Equations (12) and (13).

and

\[
\alpha = 2se^{-\lambda(A+B)}/(1 - e^{-\lambda B})
\]  

The upper bound \( \phi_2(t) \) is obtained simply by truncating Equation (11) to its the first term. The lower bound \( \phi_1(t) \) is derived by using the inequality \( 1 - x^m \geq (1 - x)^m \) for \( x \in [0, 1], m \geq 1 \), so Equation (10) implies

\[
\log \phi(t) \geq -2 \sum_{m \geq 1} \frac{(se^{-\lambda t(A+B)})^m}{m(1 - e^{-\lambda B})^m} = -2 \sum_{m \geq 1} \frac{\alpha^m e^{-\lambda mt}}{m2^m} = \log((1 - (\alpha/2)e^{-\lambda t})^2)
\]

whenever \( \alpha e^{-\lambda t}/2 \leq 1 \). Combining with the trivial inequality \( \phi(t) \geq 0 \) we obtain Equation (12).

Both bounds have the same tail behavior \( 1 - \alpha e^{-\lambda t} \) as \( t \to \infty \), but differ for small \( t \). The bounds are very close when \( \alpha \) is small, i.e., when either \( A \) or \( B \) is large. Figure 3 plots both the product function, Equation (10), and the bounds, Equations (12) and (13), for the BLOSUM62 substitution matrix and \( A = 10, B = 2 \). The two bounds are virtually identical for positive \( t \).

Our analysis will be entirely based around these bounds. Comparison with simulations will let us distinguish between discrepancies that are due to the purely technical approximations made above, and the more profound assumption that the GEM and SW algorithms are essentially equivalent.

Note that all the information about the gap penalties is encoded in \( \alpha \): different penalty schemes that produce the same \( \alpha \) are statistically indistinguishable according to this analysis. Note also that if \( B \) is not too small then \( \log \alpha \) is a linear function of \( A + B \), so it is more natural to work in terms of \( \log \alpha \).

One can write down some results for other gap penalties. It has been suggested that nonaffine gap penalties, such as the logarithmic \( A + B \log n \) and power law \( A + Bn^C \), may be more suitable for certain applications, for example when the existence of very long gaps is suspected, (Benner et al., 1993, Gu and Li, 1995). An efficient algorithm that finds local alignments for monotonically increasing gap penalties is described elsewhere (Mott, 1999).
Approximate Statistics of Gapped Alignments

For an arbitrary gap penalty \( g(n) \), define the quantity

\[
\mathcal{G}(\lambda, g) = \sum_{n=0}^{\infty} e^{-\lambda g(n)}.
\]

Equation (15) converges for \( \lambda \) large enough, and the same holds for any gap penalty that grows sub-logarithmically at least as fast as \( n \) or \( n \log n \). \( \mathcal{G}(\lambda, g) \) is called the gap generating function in the case of logarithmic growth. A similar analysis to the affine case shows that \( \phi(t) \) still has the same upper bound, Equation (13), with \( \alpha = 2sG \). The weakest gap penalty for which Equation (15) converges is \( g(n) = A + B \log n, \lambda B > 1 \), in which case 

\[
\mathcal{G} = e^{-\lambda A} \sum_{n=0}^{\infty} n^{-\lambda B}.
\]

These results suggest that any gap penalty at least as strong as this will have the same upper bound, Equation (13), and that provided \( \alpha \) can be calculated for the gap penalty of interest, the statistical development will be similar.

Weaker gap penalties, such as the length-independent gap penalty \( g(n) = A \) (Mott, 1997), produce a degenerate distribution \( \phi(t) = 0 \) \( \forall t \). However, by adding the constraint \( g(n) = \infty \) \( \forall n > B \), i.e., there is a maximum gap length \( B \), \( \mathcal{G} \) converges.

3.2 Calculation of the rate constant and the transition point

From here on, the discrete-valued random variable \( U \) will be approximated by a continuous one with distribution function \( \phi(t) \) bounded by Equations (12) and (13). The random walk theory says that, provided \( E(U) < 0 \), the score of the three maxima \( V', V(Z), V \) have distributions with exponential tails of rate \( \lambda(\alpha) \equiv \lambda \theta(\alpha), \) say, where \( \lambda \) is given by Equation (5) and

\[
1 = E(e^{\lambda \theta U}) = \int e^{\lambda \theta u} d\phi(u)
\]

Equation (16)

\( \theta(\alpha) \) measures how much the rate constant has been deflated relative to ungapped alignments using the same substitution matrix. Let \( \theta_1(\alpha), \theta_2(\alpha) \) be the corresponding exponential moments for \( \phi_1(u), \phi_2(u) \), i.e.,

\[
1 = \int e^{\lambda \theta_0} d\phi_i(u), \quad i = 1, 2
\]

Then the rate constant for \( \phi(t) \) satisfies \( \theta_1(\alpha) \leq \theta(\alpha) \leq \theta_2(\alpha) \). The integral equations for \( \theta_1(\alpha) \) can be simplified by using the substitution \( w = \alpha e^{-\lambda u} \): For the lower bound \( \theta_1 \), let \( u_0 = \log(\alpha/2)/\lambda \), i.e., the solution to \( \alpha e^{-\lambda u_0}/2 = 1 \). Then

\[
1 = \lambda \alpha \int_{u_0}^{\infty} e^{\lambda \theta_0} (1 - (\alpha/2)e^{-\lambda u}) e^{-\lambda u} \, du
= \int_0^{\frac{\alpha}{2}} \alpha^\theta w^{-\theta} (1 - w/2) \, dw
= \frac{2(\alpha/2)^{\theta_1}}{(1 - \theta_1)(2 - \theta_1)}.
\]

Equation (17)

Similarly for the upper bound \( \theta_2 \), (and omitting the intermediate integral substitution),

\[
1 = \int_{-\infty}^{-\infty} e^{\lambda \theta_0} \alpha \lambda e^{-\lambda u} \exp(-\alpha e^{-\lambda u}) \, du
= \alpha^{\theta_2} \Gamma(1 - \theta_2)
\]

Equation (18)

where \( \Gamma(\cdot) \) is the gamma function.

Equations (17) and (18) are the main results of the paper. They are easy to solve numerically for \( \theta_1, \theta_2 \) and the solutions are guaranteed to be in the interval \( (0, 1) \). \( \theta_1(\alpha), \theta_2(\alpha) \) are graphed against \( \alpha \) (log scale) in Figure 4. For \( \alpha < 0.1 \), the bounds are almost the same.

For an affine penalty structure, where \( \alpha \) is related to the gap penalties by Equation (14), as \( \alpha \to 0 \) the gap penalties rise so that gaps become very rare and \( \theta(\alpha) \to 1 \). The ungapped case corresponds to \( \alpha = 0 \). Conversely, as \( \alpha \) increases, the gap penalties decrease, \( \theta(\alpha) \to 0 \) and \( E(U) \to 0 \), at which point the phase transition from logarithmic to linear behavior occurs.
Hence the transition point $\alpha_{\text{crit}}$ will be bounded between the solutions $\alpha_1, \alpha_2$ of $0 = \int u d\phi_i(u)$ $i = 1, 2$. In Figure 4, these are also the points where $\theta_i(\alpha) = 0$. The lower bound transition point is found by solving

$$0 = \int t d\phi_1(u) = \alpha_1 \int_0^\infty u(1 - (\alpha_1/2)e^{-\lambda u})e^{-\lambda u} du$$

$$= (\alpha_1/2\lambda) \int_0^{2/\alpha_1} \log(w)(\alpha_1 w - 2) dw, \quad \text{substituting } w = e^{-\lambda u}$$

$$= (3/2 + \log(\alpha_1/2))/\lambda.$$  

This is solved when $\alpha_1 = 2e^{-3/2} = 0.44626 \cdots$.

Using the fact that the mean of the Gumbel distribution $\phi_2(u)$ is $(\gamma + \log \alpha)/\lambda$, the upper bound transition point occurs when $\alpha_2 = e^{-\gamma} = 0.56145 \cdots$.

For either bound, the values of the gap penalty parameters $A, B$ at which the transition behavior occurs lie on the contour $\alpha_i = \alpha(A, B)$ where $\alpha(A, B)$ is shorthand for Equation (14). Many alignment programs reparameterize the gap penalty to $g(n) = C + B(n - 1)$, where $C \equiv g(1) = A + B$, so that the cost of the initial gap of length 1 implicitly includes the gap-open charge. The mathematics is essentially unchanged, but the transition behavior is clearer; the contour becomes

$$\alpha_1 e^{-\lambda B} + 2s e^{-\lambda C} = \alpha_i$$  \hspace{1cm} (19)\]

Figure 5 shows transition contours for the commonly used substitution matrices PAM120, PAM250 (Dayhoff et al., 1983), BLOSUM50, and BLOSUM62 (Henikoff and Henikoff, 1992) for random protein sequences with compositions typical for human. The lower-bound contours (not shown) were very similar. Each contour partitions the $(C, B)$ space into two regions—that part containing the origin, which corresponds to linear behavior, and the other where logarithmic alignments occur. The line $C = B$ corresponds to $A = 0$; it is unlikely in practice that penalties to the left of this line would be used, as they correspond to the case $A < 0$. Note that for $C < \log(2s/\alpha_1)/\lambda$ alignments are in the linear domain, independent of $B$. This means that if $g(1)$ is sufficiently weak, alignments will always be linear, even if $B$ is infinite.
Thus, the GEM predicts bounds for the exponential rate for alignments with gaps, and the penalty threshold at which the alignments explode. These predictions were tested by by simulation. Sets of 10,000 pairs of random protein sequences were generated with amino acid composition typical for human proteins, and each pair compared using the Smith-Waterman algorithm. The PAM120, PAM250, BLOSUM50, and BLOSUM62 substitution matrices were used with gap-open penalties \( A = 0, 1 \ldots 10 \) and with the gap extension penalty \( B \) allowed to vary over the range 0–10 in steps of 0.25. Sequence lengths were held constant at \( m = n = 300 \). For each simulation set, an extreme-value distribution \( \exp(-Kmne^{-\lambda}) \) was fitted using maximum-likelihood to estimate \( \hat{K}, \hat{\lambda} \) (Mott, 1992; Waterman and Vingron, 1994b). One expects \( \hat{\lambda} \approx \lambda(\alpha) \) when the gap penalties are in the logarithmic domain.

The results for all the simulations are shown on Figure 4. The estimate \( \hat{\theta}(\alpha) = \hat{\lambda}/\lambda \) is plotted against \( \alpha \) (log scale). Comparison with the graphs for \( \theta_1(\alpha), \theta_2(\alpha) \) show the model agrees very closely with the simulations when \( \alpha < 0.06 \) approximately. For larger values of \( \alpha \), the model slightly overestimates \( \theta \) until saturation effects (i.e., when the lengths of the alignments and the sequences become comparable) dominate, and the fitted estimates stop decreasing and the model breaks down. In general, the lower bound \( \theta_1(\alpha) \) is closer to the fitted values. For numerical examples, see Tables 1 and 2.

Recalling that the expected score of a random variable with exponential tail behavior varies inversely with the rate constant \( \lambda \), the fact that the simulations imply the predicted rate constant is too large is consistent with the GEM algorithm finding slightly poorer (i.e., lower scoring) alignments compared with SW. Other, less likely, reasons for the discrepancy are dependence between diagonals, dependence between successive steps in the GEM alignment, and the behavior of \( G(t) \) for small \( t \) being nonexponential in some cases.

The true transition point, \( \alpha_{\text{crit}} \), can be estimated approximately from Figure 4, by extrapolating by eye a curve fitted through the data points, from the point before saturation effects take hold (around \( \theta \approx 0.6 \)) onto the abscissa. This rough method implies \( 0.3 < \alpha_{\text{crit}} < 0.5 \). However, \( \alpha_{\text{crit}} \) is only applicable to very long sequences. Shorter sequences, which are typically encountered in practice, will be subject to the saturation effects illustrated here, with gap penalties inside the logarithmic region, and any theory based on asymptotic behavior will break down before the transition point is reached.

A different perspective on the data is illustrated in Figure 6, which compares the predicted and simulated exponential rates \( \lambda(B) \) for the PAM250 matrix as a function of the gap extension penalty \( B \), for constant gap open penalty \( A = 0, 3, 6, 9 \). The plots show that provided \( A \) is reasonably strong (i.e., \( A \geq 6 \) in these examples) and the gap extension penalty \( B \) not too close to the transition point there is very good agreement. In other words, amongst the set of gap penalties that have the same \( \alpha \), those with large \( A \) will give GEM
alignments closest to SW. Also, GEM remains accurate near to the transition point for large $A$. Results for the other matrices were similar.

Studies on the choice of gap penalty in databank searches (Pearson, 1998) generally recommend a high gap-open and low gap extension penalty because this maximizes the discrimination between random alignments (which contain few gaps) and genuine similarities, which can contain long gaps. For example, BLAST uses $g(n) = 11 + n$ with the BLOSUM62 matrix, for which $\alpha \approx 0.09$. Thus, GEM is accurate for penalties which are of practical use, although, of course, in these ranges, the statistical behavior need not be very different from the ungapped case.

**FIG. 6.** Comparison of the GEM bounds (solid lines) and maximum-likelihood estimates (points) for the exponential rate $\lambda$ for the PAM250 matrix as a function of the gap extension penalty $B$, for constant gap open penalty: $A = 0$ (A), $A = 3$ (B), $A = 6$ (C), $A = 9$ (D).
3.3. Distribution of $W_{\text{GEM}}$

It remains to estimate $K$ to complete the distribution of $W_{\text{GEM}}$. This section uses additional approximations; mathematicians may wish to hold their noses.

Gapped local alignments will occur in clumps with a much more complicated geometry than in the ungapped case. There can be more than one chain of greedy extensions ending at the same HSP, rather like tributaries joining a main river. To obtain a Poisson approximation, it is necessary to estimate the number of local alignment clumps in which at least one alignment score is $>t$. Waterman and Vingron (1994a,b) fitted a Poisson model by directly counting the number of clumps with scores of $>t$ from a number of simulated comparisons. Here we will attempt to derive analytic estimates.
At first sight, this is a difficult problem because of the branching, interrelated structure of a clump. To circumvent these obstacles, restrict attention to the case of stringent gap penalties, i.e., small \( \alpha \), where the expected number of positive-scoring extensions from the end of an HSP is much less than unity. Then it is plausible to assume the geometry of a clump is in fact linear—i.e., that it has no branches.

Define the coordinate \((i, j)\) to be a clump start if these conditions apply: (i) \((i, j)\) is the start of an ungapped HSP of positive score, with probability in the interval \((K^{-}, K^{+})\); (ii) the unconstrained version of the GEM algorithm run backwards from \((i, j)\) never attains positive score (equivalent to \((i, j)\) being a ladder point). By the argument used earlier for ungapped alignments, this event has probability \(\Pr(V' < 0)\).

Assuming that the average greedy extension score is negative, i.e., \(E(U) < 0\), so that the alignments will terminate, the tail behaviors of the constrained and unconstrained scores \(V', V\) of a GEM alignment are

\[
(1 - \Pr(V \leq t))e^{\lambda t} \rightarrow R \text{ as } t \rightarrow \infty.
\]

\[
(1 - \Pr(V' \leq t))e^{\lambda t} \rightarrow S \text{ as } t \rightarrow \infty.
\]

The positive constants \(R, S\) could be calculated using fluctuation sum formulae applied to a discrete version of \(\phi(t)\). However, one may exploit the fact that when \(\alpha\) is small \(\phi\) has an exponential tail, i.e., for \(t > 0\),

\[
\phi(t) \approx 1 - \alpha e^{-\lambda t}.
\]

In this special case, the solution to the dam Equation (3) is exponential \(\forall t > 0\), and \(S = 1 - \theta\) (Feller, 1972; Appendix 7.3). Consequently, \(\Pr(V' < t) \approx \theta\).

Furthermore, if it is assumed that \(V\) also has an exactly exponential distribution, then \(R = 1 - \Pr(V = 0) = 1 - \phi(0) = 1 - e^{-\alpha}\). When \(\alpha\) is small, then \(R = S + O(\alpha^2)\).
Provided attention is restricted to small $\alpha$, say $< 0.1$, where the bounds $\theta_1(\alpha) \approx \theta_2(\alpha)$, in the following $\theta$ may be replaced by either value without affecting the answers overmuch.

Now a local greedy alignment starts with a diagonal HSP of strictly positive score $Z$, followed by the constrained maximum $V(Z)$ attained by a GEM walk. Using the bounds $V \leq V(Z) \leq V'$, it follows that $\Pr(V > t) \leq \Pr(V(Z) > t) \leq \Pr(V' > t)$, which are tight since $R \approx S$.

Let $W = Z + V$, $W' = Z + V'$. Then, making the approximation $\Pr(Z > t \mid Z > 0) \approx e^{-\lambda t}$, we have

\[
\Pr(W > t) = \Pr(Z + V > t \mid V > 0)\Pr(V > 0) + \Pr(Z > t)\Pr(V = 0) \\
= R\Pr(Z + V > t \mid V > 0) + (1 - R)\Pr(Z > t) \\
= \frac{\theta R}{1 - \theta} e^{-\lambda t} + \left[ 1 - R - \frac{\theta R}{1 - \theta} \right] e^{-\lambda t} \\
\sim \frac{\theta(1 - e^{-\alpha})}{1 - \theta} e^{-\lambda t} \text{ as } t \to \infty
\]

Similarly,

\[
\Pr(W' > t) = \theta e^{-\lambda t}
\]

Therefore, modulo the various assumptions and approximations made so far, $\Pr$ (the coordinate marks the start of a gapped alignment of score $> t$) lies in the interval

\[
(K^-(-\alpha)e^{-\lambda t}, K^+(\alpha)e^{-\lambda t} \equiv \left( K^-\theta(1 - e^{-\alpha}) e^{-\lambda t}, K^+\theta e^{-\lambda t} \right)
\]

and the distribution of $W_{GEM}$ satisfies

\[
\Pr(W_{GEM} \leq t) \in (\exp(-mnK^+(\alpha)e^{-\lambda t}), \exp(-mnK^-(-\alpha)e^{-\lambda t}))
\]

Note that as $\alpha \to 0$ Equation (21) reduces to the bounds, Equation (8), for ungapped alignments.

It could be argued that it is inconsistent to treat the leading HSP as discrete-valued but the GEM walk as continuous. A more accurate derivation would expand the error bounds to account for this. However, we judged it was not worthwhile in view of the large number of approximations already made and the edge corrections added in the next section.

4. EDGE EFFECTS

Before testing the accuracy of Equation (21), one must take into account edge effects (Waterman and Vingron, 1994b; Altschul and Gish, 1996; Spang and Vingron, 1998). The formulas for the maximum-scoring ungapped similarity need to be corrected before they will give accurate $p$ values. The sequence lengths $m, n$ must be adjusted to allow for the fact that high-scoring alignments physically cannot start close to the ends of either sequence. It has been found that $p$ values calculated using Equation (1) agree much more closely with simulations if $m, n$ are replaced by effective lengths $m' = m - E(l_{HSP})$, $n' = n - E(l_{HSP})$, where $E(l_{HSP})$ is the expected length of an HSP.

Altschul and Gish (1996) calculated estimates for $E(l_{HSP})$ as follows: From the properties of the extreme-value distribution, the expected score of the largest HSP is approximately $\log(Kmn)/\lambda$. It is also known that the expected score of a pair of aligned symbols within an HSP converges to

\[
E_{HSP} = \sum xh(x)e^{\lambda x}
\]

(Arratia et al., 1988; Karlin and Altschul, 1990). Consequently,

\[
E(l_{HSP}) \approx \log(Kmn)/(\lambda E_{HSP})
\]

Fortunately, it is not necessary to know the alignment length very accurately for the purpose of length correction.
FIG. 7. Plot of the average adjusted alignment length $I_{SW} - E(I_{HSP})$ against $\alpha$ for all simulations, where $m = n = 300$ is held constant. The straight lines $260\alpha - 10$ (solid line A) and $-3.599 - 529.519\alpha + 55.820\alpha \log mn$ (dashed line B) are also shown.

The edge effect can be stronger with gapped alignments, when $\alpha$ is not too small, because they will be longer on average. In the simulations performed above, the mean length $I_{SW}$ of each SW alignment (i.e., number of aligned residues + total length of gaps) was also computed for every combination of gap penalty and matrix. One expects that the GEM alignment length $I_{GEM} \approx I_{SW}$ for strong gap penalties.

Now it is plausible that $E(I_{GEM})$ may be partitioned into $E(I_{GEM}) = E(I_{HSP}) + E(I_{gap})$, where $I_{HSP}$ is the length of the leading HSP, which may be got from Equation (22), and $I_{gap}$ that of the remaining gapped alignment. Figure 7 plots $I_{SW} - E(I_{HSP})$ against $\alpha$ for all the simulations, where $m = n = 300$ is held constant. Figure 7 shows that $E(I_{gap})$ is a function of $\alpha$ alone for $\alpha < 0.2$, and furthermore has a roughly linear relationship provided $\alpha < 0.1$. To a good approximation, the effects of the matrix and gap penalty are factored into $\alpha$. Detailed examination of Figure 7 close to $\alpha = 0$ does indicate some separation according to the matrix used, but this is insignificant.

Over the range $\alpha < 0.1$, one finds

$$\ell(\alpha) \approx 260\alpha - 10$$

Consequently, the expected length of a gapped alignment can be estimated as

$$E(I_{GEM}) \approx E(I_{HSP}) + \ell(\alpha)$$

and this approximation should hold for any reasonable choice of matrix, gap penalty, and sequence composition.

More generally, allowing the sequence lengths to vary, one may assume that for small $\alpha$

$$E(I_{gap}) = \ell(\alpha, m, n) \approx \ell_0 + \ell_1\alpha + \ell_2\alpha \log mn$$

for unknown constants $\ell_0, \ell_1, \ell_2$. To test this prediction, 474 pairs of sequences were drawn at random from the SWISSPROT protein sequence databank. For each pair, 10,000 comparisons were performed on randomly generated sequences using the lengths and compositions of the originals. The BLOSUM62 matrix with gap penalty $9 + 2n$ was used throughout; $\alpha$ varied depending on the sequences’ compositions, from 0.0104 to 0.154, with median 0.0575. For each data set, $E(I_{gap})$ was calculated, and then Equation (24) was fit to these values. The least squares estimates were $\ell_0 = -3.599, \ell_1 = -529.519, \ell_2 = 55.820, R^2 = 0.901$. All estimated coefficients were highly significant. Other, more complex models were also investigated; only
modest gains in goodness-of-fit were observed. Figure 8 plots the observed and predicted values of \( E(l_{gap}) \) for the data. In Figure 7, the predicted straight line \(-3.599 - 529.519\alpha + 55.820\alpha \log(300 \times 300)\) is superimposed. Over the range \( \alpha < 0.1 \), the predicted fit is reasonably accurate, so Equation (24) may be used to predict the length of any alignment provided \( \alpha < 0.1 \) and the sequence lengths are not too small.

5. VALIDATION

One is now in a position to test the accuracy of the GEM predictions for \( K^-(\alpha) \), \( K^+(\alpha) \), and for scores thresholds at various levels of statistical significance. Figure 9 plots the logarithm of the maximum-likelihood estimate \( k_{mn} \) against \( \alpha \) (log scale) for all simulations using the BLOSUM62 matrix. Length correction was performed using Equation (23). The bounds \( K^-(\alpha) \), \( K^+(\alpha) \) are shown as “error bars.”

Now the equations satisfied by the maximum likelihood estimators of \( \lambda, K \) are very complicated if one considers, as one should, the data as following a discrete distribution whose distribution function agrees at score values with the extreme value distribution. A more tractable set of equations can be derived for approximate ML estimators by supposing that the data comes from a continuous extreme value distribution, which was how our estimates were calculated. The error in this approximation can be estimated, and to first order the true estimator for \( \lambda \) agrees with the simpler approximate estimator. However, the true estimator for \( K \) is a multiplicative factor \( \exp(\delta \lambda / 2) \) larger than the approximate estimator, where \( \delta \) is the span of the score values. For this reason, we plot the approximate ML estimators for \( K \) against an error bar showing the interval \( K^-(\alpha), K^+(\alpha) \).

Figure 9 shows that for \( \alpha < 0.1 \) approximately, the fitted estimates are within this interval. The results for the other matrices (not shown) were similar except for the PAM250 matrix, which slightly exceeded the bounds \( \alpha \rightarrow 0 \). Interestingly, this discrepancy is not due to the theory developed here, but perhaps to the ungapped length correction, Equation (22), being too strong for this matrix. This point emphasizes the fact that although edge corrections have proved very useful, their theoretical basis is yet to be established.

To get a realistic perspective on the magnitude of the uncertainty in \( K \), we note that for typical proteins, matrices and gap-penalties, the effects of length correction and the ML approximation changes \( \log(K_{mn}) \) by about unity. This corresponds to changing the \( p \)-value \( \approx \exp(\log(K_{mn}) - \lambda t) \) by about a factor 3.

Table 1 compares the GEM and ML estimates of \( K, \lambda \) for the commonly used default scoring schemes for BLAST and FASTA, for a variety of sequence compositions. FASTA uses a default scoring scheme
(BLOSUM50, $10 + 2n$) for which $\alpha \approx 0.16 - 0.18$, $\theta \approx 0.7 - 0.74$. BLAST’s default scoring scheme (BLOSUM62, $11 + 1n$) is more stringent, with $\alpha \approx 0.075 - 0.085$, $\theta \approx 0.9$.

Table 1 indicates that the GEM estimates consistently overestimate $\lambda(\alpha)$, compared with the MLE, but that the discrepancy is small—less than 5%. Also, although the default FASTA parameters are outside the region $0 < \alpha < 0.1$, the GEM estimates are still usable.

The GEM bounds for $K(\alpha)$ are less precise, but they too are sufficiently accurate. This is illustrated in Table 2, which gives the threshold scores $t$ for achieving $p$-values $10^{-6}$ and $10^{-8}$, for the six cases in Table 1. When $p$ is small, the threshold $t(p) \approx (\log(Kmn) - \log(p))/\lambda$. The interval $[10^{-8}, 10^{-6}]$ corresponds to the very important region of borderline statistical significance in a databank search; sequence comparisons with $p$ values $<10^{-8}$ probably reflect genuine similarities (or are due to nonrandom artefacts such as a common repeated elements), whilst those with $p$-values $>10^{-6}$ are more likely to be random. It is therefore important that significance thresholds are accurate in this region. The GEM thresholds are given as intervals corresponding to using $K^- (\alpha)$ or $K^+ (\alpha)$. The Table shows that the MLE estimates are usually inside the GEM intervals, or are just outside.

6. DISCUSSION

This paper provides strong evidence that the behavior of the optimal gapped alignment score is similar to that of the GEM, which provides simple formulae, depending on a single additional parameter $\alpha$, for assessing statistical significance. We conjecture that similar statistics apply to any gap penalty for which $\alpha$ may be calculated. The model also gives useful predictions on the position of the transition point from logarithmic to linear behavior. GEM uses existing ideas from ungapped alignment statistics and greedy alignment approximations. The key new idea, that an alignment can be broken into a series of greedy extension steps, each of which resembles a step in a random walk, owes something to the renormalization ideas in statistical physics.

What are the advantages and disadvantages of using the theoretical GEM predictions for assessing statistical significance, over parametric function fitting? The use of formulae is more satisfying mathematically, can save computer time by avoiding simulation, and is free from problems of outlier detection in databank searches (Mott, 1992; Pearson, 1998). In addition, even where the formulae are not too accurate, they enable one to
choose suitable values for gap penalties with arbitrary substitution matrices and predict which side of the transition point a given set of penalties will lie.

Against this, function fitting methods are accurate closer to the transition point because parameter estimates are adjusted implicitly to take into account edge and saturation effects. The GEM formulae are unsafe when the expected alignment length is a significant fraction of the sequences' lengths, and when these lengths are short or very different. Unusual sequence features can also be taken into account by curve fitting; e.g., if a sequence contains a repeated subunit, then its effective length is less (Mott, 1992; Spang and Vingron, 1998).

We are currently investigating a semiparametric approach to improve the accuracy of the GEM estimates; if we assume that \( \lambda(\alpha) \) and \( K(\alpha) \) are unknown functions of \( \alpha \) (our GEM formulae being mere approximations), then these functions may be estimated once from a set of simulations and then applied to all future cases. For example, the irregular line in Figure 4 plots the moving average of \( \hat{\theta}(\alpha) \)—a smoothed version of which could be used as a less biased estimator.

Beyond its obvious use, likely practical applications of the GEM formulae include:

(a) Design of scoring schemes. It is generally accepted that the most sensitive gap-penalty for databank searching is in the logarithmic domain, with a high gap-open \( A \) with a small gap-extension \( B \), and fairly close to the transition point. Since the GEM estimates are very accurate in this region, one could plot graphs of \( \lambda(\alpha) \) against \( B \), as in Figure 6, to choose reasonable values. For example, if a particular gap-penalty, and hence \( \alpha \), has been shown to work well for an existing matrix, then a good initial choice of gap-penalties for a new matrix might be to choose penalties which give a similar \( \alpha \). Further investigation of the link, if any, between sensitivity and \( \alpha \) is required.

(b) Postprocessing the results of databank searches to remove false positives. Popular databank search programs assume that sequence composition is uniform (and hence variations in in it ignorable) across the databank. BLAST (Altschul et al., 1990; Altschul and Gish, 1996; Altschul et al., 1997) assesses gapped alignment significance against a lookup-table for a standard composition typical for many protein sequences. FASTA (Pearson and Lipman, 1988; Pearson, 1998) fits an extreme-value distribution to the scores, taking account of variation in sequence length, but ignoring composition. A consequence of the assumption of uniformity is that some databank sequences with abnormal compositions may actually have a value \( \alpha \) much larger than the average, and hence their score \( p \) values will appear abnormally significant. Now it is possible to fit a model that includes composition effects (Mott, 1992), but since this requires the solution of Equation (5) for every comparison, the search is slower. One could calculate GEM estimates \( (K(\alpha), \lambda(\alpha)) \) from first principles for every comparison, but this would be equally tardy because the ungapped quantities \( K^-, \lambda \) are still needed. Instead, one could reassess just those comparisons judged potentially significant, e.g., with \( p \) values of \( <10^{-5} \), by the initial search.

Clearly, the results presented in this paper will be useful, but they are not a complete solution. An ideal theory of gapped alignment behavior should not only provide accurate asymptotic estimates for \( (K, \lambda) \) for all scoring schemes, but also deal with nonasymptotic behavior, i.e., edge and saturation effects. This is a challenging problem.

7. APPENDIX

7.1. Fluctuation sum formulae

For completeness, we give formulae from Iglehart (1972) for \( s, r \) in Equations (4) and (6), using the modifications in Karlin and Altschul (1990) and Karlin and Dembo (1992) to cope with lattice (discrete-valued) variables. Let \( Y_n \) be the sum of \( n \) iid random variables with mass function given by Equation (2). Let \( E(X; X < 0) \) be the expectation of the random variable \( X \) restricted to negative values, i.e., \( \sum_{x<0} x \Pr(X = x) \). Let \( \lambda \) be defined as in Equation (5), and \( \delta \) be the smallest span of score values. Let

\[
P = \exp \left\{ -\frac{1}{k} \Pr(Y_k \geq 0) \right\}
\]

\[
Q = \exp \left\{ -\sum_{k>0} \frac{1}{k} E(e^{\lambda Y_k}; Y_k < 0) \right\}
\]


These quantities may be calculated numerically by successive autoconvolution of the mass function $h(\cdot)$. Then

$$s = \frac{\delta}{e^{\lambda \delta} - 1} \frac{P Q}{E(Y_1 e^{\lambda Y_1})}$$

$$r = s Q$$

Karlin and Dembo (1992) showed that the expected distance $l$ between successive ladder points is

$$l = 1/\rho = 1/P$$

and that

$$(K^-, K^+) = (r \rho, r \rho e^{\lambda \delta})$$

7.2. Recurrence formulae

In order to demonstrate the equivalence of the Poisson Clumping Heuristic (Aldous, 1989; Waterman and Vingron, 1994b) and the ladder point arguments of Karlin and Dembo (1992), define $G_n(t) = \Pr(\max(0, M_0, M_0 + M_1, \ldots M_0 + M_1 + \cdots + M_n) \leq t)$. By a similar argument to that used to derive Equation (3), one can show that

$$G_n(t) = \sum_{i \leq t} G_{n-1}(t-i) h(i)$$

This is unsuitable for numerical calculation, so put $H_n(t) = e^{\lambda t} (1 - G_n(t))$. Then

$$H_n(t) = H_0(t) + \sum_{i \leq t} H_{n-1}(t-i) e^{\lambda i} h(i)$$

This is a type of renewal equation and has asymptotic solution $H_n(t) \to H(t)$ as $n \to \infty$. Also $H(t) \to s$ as $t \to \infty$. The equation can be solved iteratively for $n = 2, 3 \ldots$ to obtain estimates for $G(0) = H(0)$ and $s$.

Write $g(t) = \Pr(Z^r = t) = G(t) - G(t-1)$. Then the probability $\rho = 1/1$ that the random walk never exits from negative values is

$$\rho = G(0) - \Pr(Z^r = 0) = G(0) - h(0)G(0) - \sum_{m < 0} h(m)g(-m)$$

Implementations of the two methods for computing $s$, $\rho$ yield the same result, allowing for rounding error. For example, with the PAM250 matrix, gap penalty $9 + 3n$, the fluctuation formulae give $(s, \rho) = (0.5293, 0.3347)$, while the recurrence gives $(0.5289, 0.3343)$. Note however that since the recurrence method does not give a formula for $r$, it cannot be used to get all the quantities of interest. However, the recurrence does provide accurate estimates of $G(t)$ for small $t$, which we used to check that $H(t)$ is effectively constant for $t > A$.

7.3. Exact solution of the Dam equation

We have to show that if $\phi(t) = 1 - \alpha e^{-\lambda t}$ when $t > 0$ then $S = 1 - \theta$, where $E(e^{\lambda \theta x}) = 1$. This is readily shown by substitution. Using an obvious notation, we have:

$$\int_{-\infty}^t (1 - Se^{-\lambda \theta(t-x)}) d\phi(x) = \int_0^t + \int_{-\infty}^0$$

$$\int_0^t = \int_0^t (1 - Se^{-\lambda \theta(t-x)}) \alpha e^{-\lambda x} dx$$

$$= \frac{\alpha}{\lambda} + e^{-\lambda t} \alpha \frac{S}{\lambda (1 - \theta)} - e^{-\lambda \theta t} \frac{S \alpha}{\lambda (1 - \theta)}$$

$$\int_{-\infty}^0 = \int_{-\infty}^0 (1 - Se^{-\lambda \theta(t-x)}) d\phi(x)$$

$$= \phi(0) + Se^{-\lambda \theta t} \int_{-\infty}^0 e^{\lambda \theta x} d\phi(x)$$
but by the definition of $\theta$

$$
\int_{-\infty}^{\infty} e^{\lambda x} d\phi(x) = \int_{-\infty}^{0} e^{\lambda x} d\phi(x) + \frac{\alpha}{\lambda(1-\theta)} = 1
$$

and by the law of total probability

$$
\int_{-\infty}^{\infty} d\phi(x) = \phi(0) + \frac{\alpha}{\lambda} = 1
$$

so

$$
\int_{0}^{T} + \int_{-\infty}^{0} = 1 - Se^{-\lambda T}
$$

provided $S = 1 - \theta$.

ACKNOWLEDGMENT

We thank Dr. Stephen Altschul for many valuable comments and suggestions.

Note Added in Proof

D. Siegmund and B. Yakir (unpublished work) have recently found a theoretical asymptotic solution for certain storing regimes.

REFERENCES


Mott, R.F. 1999. Local sequence alignments with monotonic gap penalties. *Bioinformatics.* (Submitted.)

Address reprint requests to:
Richard Mott
Wellcome Trust Centre for Human Genetics
University of Oxford
Windmill Road
Oxford OX3 7BN
U.K.

Richard.Mott@well.ox.ac.uk

Received for publication August 16, 1998; accepted as revised October 25, 1998.
This article has been cited by:


4. Richard Mott. Alignment: Statistical Significance. [CrossRef]


6. S. Sheetlin. 2005. The Gumbel pre-factor k for gapped local alignment can be estimated from simulations of global alignment. *Nucleic Acids Research* 33:15, 4987-4994. [CrossRef]


11. References 337-348. [CrossRef]


