# The Estimation of Statistical Parameters for Local Alignment Score Distributions

Stephen F. Altschul[1], Ralf Bundschuh[2], Rolf Olsen[2] and Terence Hwa[2]
[1]National Center for Biotechnology Information, National Library of Medicine,
National Institutes of Health, Bethesda, MD 20894
[2]Department of Physics, University of California at San Diego,
9500 Gilman Drive, La Jolla, CA 92093-0319

August 6, 2001

**Abstract**

The distribution of optimal local alignment scores of random sequences plays a vital role in evaluating the statistical significance of sequence alignments. These scores can be well described by an extreme-value distribution. The distribution's parameters depend upon the scoring system employed and the random letter frequencies; in general they cannot be derived analytically, but must be estimated by curve fitting. For obtaining accurate parameter estimates, a form of the recently described "island" method has several advantages. We describe this method in detail, and use it to investigate the functional dependence of these parameters on finite-length edge effects.

## 1 Introduction

Local sequence alignment is perhaps the most widely used tool of computational molecular biology, with most protein and DNA database search programs (1-4) implementing heuristic versions of local alignment algorithms (5,6). These algorithms seek the highest-scoring alignment of segments from the two sequences being compared. An alignment's score is calculated by adding *substitution scores*, defined for each aligned pair of letters, and *gap scores* for each run of letters in one segment aligned with null characters inserted into the other.

A key question is what alignment scores may be expected to occur purely by chance. This question is generally addressed by analyzing the distribution of optimal alignment scores from random or real but unrelated sequences. We confine attention to random sequences, defined as strings of independent letters chosen with fixed *background* probabilities, because they are easier to control and study. Depending upon the details of the alignment scoring system and the background letter probabilities, the optimal score for the alignment of two random sequences of length $n$ tends to grow proportionally either to $n$ or to $\log(n)$ (7-10). The linear scoring regime corresponds to optimal alignments that tend to involve virtually the entire sequences; the logarithmic regime, with substitution and gap scores that are on average more negative, corresponds to optimal alignments that are relatively short. Many alignments representing true biological relationships involve only segments of the sequences compared, but these will tend to be outscored by long "random alignments" when a scoring system in the linear regime is employed. Therefore, attention has focused primarily on scoring systems in the logarithmic regime, and we deal here exclusively with such scores.

In the asymptotic limit of long sequences, optimal local alignment scores follow an *extreme value distribution* (11), described by two parameters $\lambda$ and $K$. For the type of scoring system in most general use, these parameters cannot be calculated but must instead be estimated by random simulation. Most directly, one may generate optimal alignment scores for a large number of random sequence pairs, and fit an extreme value distribution to these scores. Recently, an alternative approach has been described; it uses scores for local alignment "islands" generated by a slight modification of the Smith-Waterman algorithm (12). We will discuss in detail the implementation and application of the island parameter estimation method, and compare it to the direct method in several ways. The island method has a number of useful features: 1) It renders explicit a tradeoff between parameter estimate bias and stochastic error, and allows this tradeoff to be easily controlled; 2) It estimates accurately the tail behavior of score distributions for small-length comparisons; 3) It allows parameter estimates to be obtained for arbitrary length sequence comparisons, including the infinite-length limit. In some circumstances, the first two of these features can be transferred advantageously to the direct method, appropriately modified. For asymptotic parameter estimation, however, the island method has a clear speed advantage.

## 2 The direct estimation of statistical parameters

An asymptotic theory for local alignment scores has been developed for the case in which no gaps are permitted. In brief, for the comparison of random sequences of sufficient lengths $m$ and $n$, the *number* of distinct local alignments with score at least $x$ is approximately Poisson distributed, with mean

$$E(x) \approx Kmne^{-\lambda x}, \qquad [1]$$

where $\lambda$ and $K$ are easily calculated parameters (13,14). This implies that the optimal alignment score $S'$ approximately follows an extreme value distribution (11), with

$$\text{Prob}\left(S' \geq x\right) \approx 1 - \exp\left(-Kmne^{-\lambda x}\right). \qquad [2]$$

For local alignments that allow gaps, no asymptotic score distribution has been established analytically. However, computational experiments strongly suggest that equations [1] and [2] apply to this type of alignment as well (12,15-22). The key to using equations [1] and [2] is the accurate estimation of the statistical parameters $\lambda$ and $K$. Perhaps the most direct approach to estimating these parameters for a fixed scoring system and set of background letter frequencies is to generate a large number of pairs of random sequences of equal length $n$, and find the optimal local alignment score for each pair. From these scores one may calculate maximum-likelihood estimates $\widehat{K}$ and $\widehat{\lambda}$ for the statistical parameters in equation [2] (23). If $R$ scores are generated, the ratio $\widehat{\lambda}/\lambda$ is approximately normally distributed, with mean 1 and standard error $0.78/\sqrt{R}$ (23). Note that the estimates developed in (23) assume continuous data, whereas alignment scores are almost always discrete. If the scale parameter $\lambda$ times the lattice spacing of possible scores is small, the error introduced by assuming continuous scores is minor. One may, however, derive maximum-likelihood estimates $\widehat{\lambda}$ and $\widehat{K}$ that explicitly assume discrete scores (Appendix).

Because $\lambda$ enters equations [1] and [2] exponentially, accurate estimates of $\lambda$ are particularly important. Marginally significant alignments from current database searches typically have a scaled score $\lambda x > 25$, for which even a 4% error in $\lambda$ leads to an estimated $E$-value in error by greater than a factor of 2.7. Thus, standard errors of less than 2%, or even 1% in $\widehat{\lambda}$ may be desirable.

## 3 The island method

Recently, Olsen et al. (12) proposed the island method for estimating $\lambda$ and $K$; it is a variant of ideas introduced by Waterman and Vingron (18,19) that translates into a very efficient algorithm. Rather than finding optimal alignment scores for pairs of random sequences, they propose generating scores for each island (as defined below) in a path graph. To generate sufficiently many scores for accurate parameter estimation, a single large, or multiple smaller pairwise comparisons may be used.

Briefly, the Smith-Waterman algorithm generates a score for each cell $C$ in a path graph, corresponding to the highest-scoring local alignment ending at $C$ (5). This local alignment starts at a specific anchoring cell, and an *island* consists of all cells with identical anchor (Figure 1). The score assigned to an island is the maximum score of the cells it contains. A simple modification of the Smith-Waterman algorithm, involving only a fixed amount of extra computation per cell, allows one to record which island each cell belongs to, and to keep track of each island's score. Note also that as one moves row by row through a path graph with $n$ columns, there can be at most $O(n)$ islands represented on any given row. This allows one to tabulate all island scores generated by an $m \times n$ path graph in $O(mn)$ time, and using only $O(n)$ space.

Island scores correspond to *distinct* locally optimal alignments, and thus the number of islands with score at least $x$ should be well described by equation [1] when $x$ is sufficiently large. The island method generates maximum-likelihood estimates of $\lambda$ and $K$ from equation [1], while the direct method generates these estimates from equation [2].

The concept of two or more local alignments being distinct is a subtle one, and a variety of definitions have been proposed (6,12,24,25). The differences among these definitions are relevant more for the comparison of real than random sequences. Because using any reasonable definition of distinct alignments should yield equivalent statistical results, the advantage of the "island" (12) over the "declumping" definition (18,19,24,25) for parameter estimation is its algorithmic efficiency.

In general, equation [1] becomes increasingly accurate for larger values of $x$, so to obtain a good estimate for $\lambda$ one should confine attention to islands whose score attains at least some threshold value $c$. Assume the set $I_c$ of such islands has cardinality $R_c$, and let $\overline{S}_c$ be the mean score in excess of $c$ of these islands:

$$\overline{S}_c = \frac{1}{R_c} \sum_{i \in I_c} [S(i) - c] \qquad [3]$$

where $S(i)$ is the score of island $i$. Then, assuming island scores are integral, with unit lattice spacing, the maximum-likelihood estimate (Appendix) for $\lambda$ is:

$$\widehat{\lambda}_c = \ln\left(1 + \frac{1}{\overline{S}_c}\right). \qquad [4]$$

The standard error of $\widehat{\lambda}_c/\lambda$ is:

$$\sigma = \frac{e^{\lambda} - 1}{\lambda\sqrt{e^{\lambda}}} \frac{1}{\sqrt{R_c}} \approx \frac{1 + \lambda^2/24}{\sqrt{R_c}}, \qquad [5]$$

where the approximation holds to better than 0.05% for $\lambda < 1$. If the island scores were continuous, the maximum-likelihood estimate $\widehat{\lambda}_c$ would instead be simply $1/\overline{S}_c$, and the standard error of $\widehat{\lambda}_c/\lambda$ would be $1/\sqrt{R_c}$.

In conjunction with $\widehat{\lambda}_c$, the maximum-likelihood estimate for $K$ is:

$$\widehat{K}_c = \frac{R_c e^{\widehat{\lambda}_c c}}{A}, \qquad [6]$$

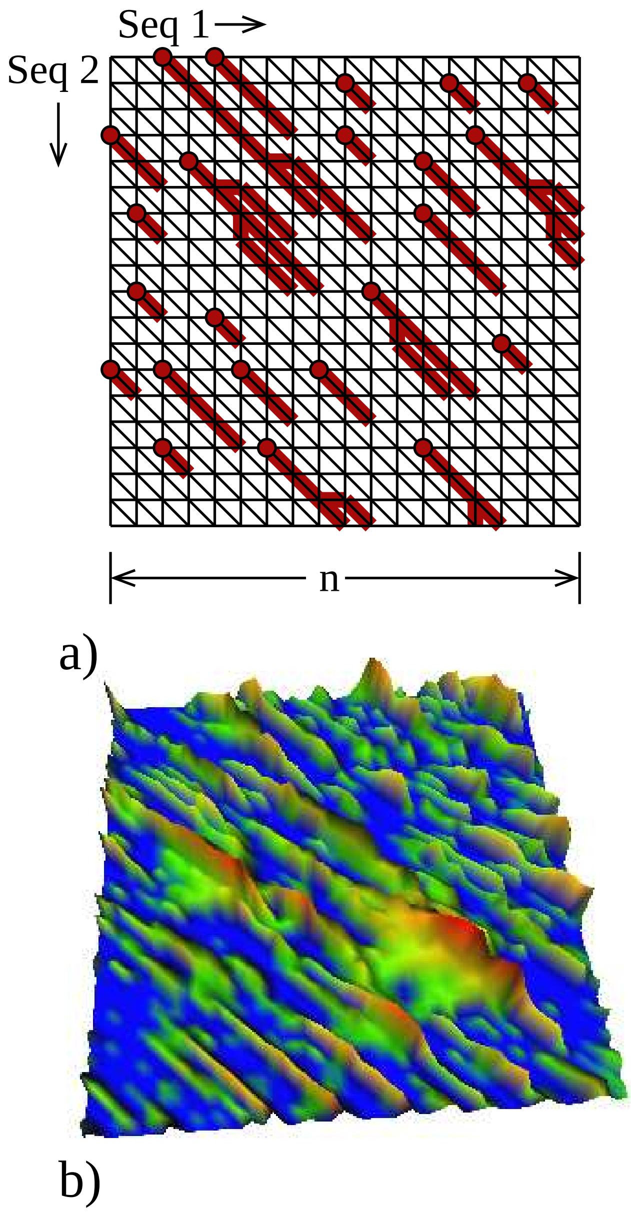


Figure 1: Islands in a local alignment path graph. (a) is a schematic representation of the path graph. In every cell $C$ the red line recalls the choice made by the optimization procedure of the Smith-Waterman algorithm. By these lines, all the cells with non-zero scores are partitioned into islands according to which anchoring points (circles) they are connected to. (b) shows the score landscape on a $50\times50$ path graph. The score at every cell of the path graph is represented by its height above the surface and color-coded with zero scores corresponding to blue areas and increasingly red colors for higher scores. The example shown is generated with a BLOSUM-62 scoring matrix, and a score $-(11+k)$ for each gap of length $k$. The islands are easily seen.

where $A$ is the aggregate "area" of the search space from which the collection of islands were drawn. If a single pair of sequences, of lengths $m$ and $n$, were compared to generate the islands, then $A = mn$; if $B$ such comparisons were performed, then $A = Bmn$.

The parameters $\lambda$ and $K$ of equations [1] and [2] properly apply only in the limit of infinite length sequences. If one uses either the island or direct method to estimate $\lambda$ for sequences of finite length, one obtains estimates with an observable finite-length bias. As will be discussed below, this bias can be explained in terms of "edge effects", for which a simple correction can be applied to the lengths $m$ and $n$ in equations [1] and [2]. The resulting formulas retain the asymptotic values of $\lambda$ and $K$, so it is desirable to avoid any finite-length bias in the estimation of these parameters. We note here that, by eliminating edge effects, the island method can estimate asymptotic values of $\lambda$ and $K$ directly. This is done by embedding a length $n \times n$ sequence comparison within a larger $(n + 2b) \times (n + 2b)$ comparison, with a border of length $b$ on each side (Figure 2). Only islands anchored within the central $n \times n$ region are recorded. When $b$ is sufficiently large, edge effects are essentially abolished.

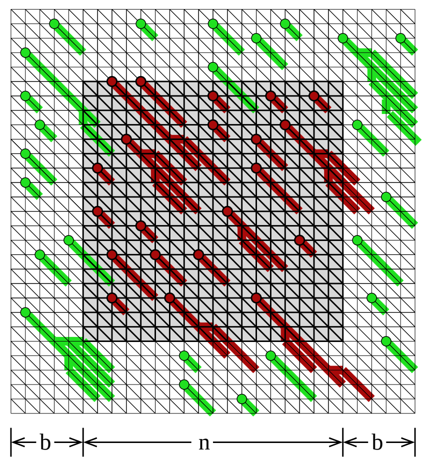


Figure 2: Schematic representation of a path graph used to avoid edge effects in the estimation of $\lambda$ and $K$ via the island method. The $n \times n$ scoring lattice (gray square in the middle) is surrounded by a border of width $b$. Only islands that are anchored within the central $n \times n$ area (shown in dark red) are counted. Islands anchored outside this area (green) are ignored. Note that some of the ignored islands reach into the inner area and some of the accepted islands reach into the border region since the classification of an island depends only on the position of its anchor (circles); borders thus are required on all sides to suppress edge effects properly.

## 4 The tradeoff of speed, bias, and precision

Because of $\lambda$'s exponential role in equations [1] and [2], accurate estimates for $\lambda$ are far more important than those for $K$, and we shall therefore focus on the estimation of $\lambda$. A key question for applying the island method effectively is how to choose an appropriate threshold parameter $c$ for use in equation [4].

While we believe that the qualitative features presented here are true independent of the scoring system used, below we will illustrate the issues involved in choosing $c$ using a specific example. To obtain extremely accurate parameter estimates for this case study, we performed a massive random simulation for a particular local alignment scoring system. Specifically, we used a set of standard amino acid frequencies for proteins (26) to generate over 92,000 pairs of length-7000 "random amino acid sequences". We compared each pair using the BLOSUM-62 amino acid substitution matrix (27), in conjunction with *affine* gap scores (28-31) of $-(11+k)$ for gaps of length $k$. To suppress edge effects, scores were tabulated only for islands anchored within the central $5000\times5000$ square of each pairwise comparison; approximately $10^{12}$ total island scores were recorded. Using equations [3]-[6], estimates of $\lambda$ and $K$ were obtained from these data for a range of the cutoff scores $c$; the results are summarized in Table 1 and the values of $\widehat{\lambda}$ are plotted in Figure 3.

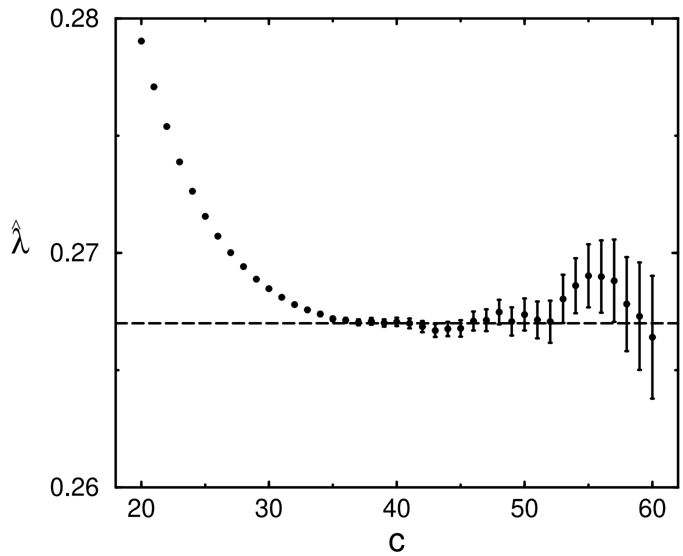


Figure 3: Estimates $\widehat{\lambda}_c$ obtained via the island method with different cutoffs $c$. Standard errors for the estimates are shown with error bars. The plotted horizontal line indicates the best estimate of the asymptotic $\lambda$. Details of the simulation are given in the legend to Table 1.

While the estimates $\widehat{\lambda}_c$ of Table 1 should be essentially free of edge effect bias, there is another systematic and easily understood bias (12) evident for small values of $c$. Optimal local alignments with low score are unlikely to contain a gap, as will be discussed further below, and for low thresholds $\widehat{\lambda}_c$ is therefore biased towards the higher $\lambda$ applicable to local alignments that exclude gaps. In this example, $\widehat{\lambda}_c$ falls monotonically for $c \geq 20$, until it reaches the value 0.2670 at $c = 37$; thereafter, $\widehat{\lambda}_c$ appears to fluctuate randomly about this value. Of course a yet larger simulation, yielding smaller stochastic errors, might detect systematic bias even beyond $c = 37$.

There is a tension between the bias of $\widehat{\lambda}_c$ and its precision, for the larger the value of $c$ chosen, the fewer the islands that attain score $c$, and the larger the standard error of $\widehat{\lambda}_c$. To illustrate the point, consider a realistically sized random simulation, 10,000 times smaller than that shown in Table 1, which would require about 2 minutes on a modern workstation. The systematic bias in the $\widehat{\lambda}_c$ from such a simulation should be the same as seen in Table 1, but the standard errors will be 100 times larger. Table 2 shows the resulting tradeoff between bias and precision. The best tradeoff probably occurs near $c = 28$, where the sum (about 2.2%) of the bias and the standard error is minimized. As the size of the random simulation grows, the bias at a given cutoff remains fixed, whereas the standard error decreases. Thus in general the optimal tradeoff for larger simulations will tend to occur at higher values of $c$.

| $c$ | $R_c$ | $\widehat{\lambda}_c$ | $\widehat{K}_c$ |
|---|---|---|---|
| 20 | 508087143 | 0.2790 ± 0.0000 ( 0.00%) | 0.058 |
| 21 | 382046389 | 0.2771 ± 0.0000 ( 0.01%) | 0.056 |
| 22 | 288047946 | 0.2754 ± 0.0000 ( 0.01%) | 0.053 |
| 23 | 217666586 | 0.2739 ± 0.0000 ( 0.01%) | 0.051 |
| 24 | 164854001 | 0.2726 ± 0.0000 ( 0.01%) | 0.050 |
| 25 | 125090432 | 0.2716 ± 0.0000 ( 0.01%) | 0.048 |
| 26 | 95080777 | 0.2707 ± 0.0000 ( 0.01%) | 0.047 |
| 27 | 72367615 | 0.2700 ± 0.0000 ( 0.01%) | 0.046 |
| 28 | 55135823 | 0.2694 ± 0.0000 ( 0.01%) | 0.045 |
| 29 | 42040928 | 0.2689 ± 0.0000 ( 0.02%) | 0.044 |
| 30 | 32087753 | 0.2685 ± 0.0000 ( 0.02%) | 0.044 |
| 31 | 24502349 | 0.2681 ± 0.0001 ( 0.02%) | 0.043 |
| 32 | 18721366 | 0.2678 ± 0.0001 ( 0.02%) | 0.043 |
| 33 | 14312497 | 0.2676 ± 0.0001 ( 0.03%) | 0.042 |
| 34 | 10945852 | 0.2674 ± 0.0001 ( 0.03%) | 0.042 |
| 35 | 8372081 | 0.2672 ± 0.0001 ( 0.03%) | 0.042 |
| 36 | 6407611 | 0.2671 ± 0.0001 ( 0.04%) | 0.042 |
| 37 | 4904102 | 0.2670 ± 0.0001 ( 0.05%) | 0.041 |
| 38 | 3755281 | 0.2671 ± 0.0001 ( 0.05%) | 0.042 |
| 39 | 2874422 | 0.2670 ± 0.0002 ( 0.06%) | 0.041 |
| 40 | 2201167 | 0.2671 ± 0.0002 ( 0.07%) | 0.042 |
| 41 | 1684893 | 0.2670 ± 0.0002 ( 0.08%) | 0.041 |
| 42 | 1289490 | 0.2669 ± 0.0002 ( 0.09%) | 0.041 |
| 43 | 986932 | 0.2667 ± 0.0003 ( 0.10%) | 0.041 |
| 44 | 756060 | 0.2668 ± 0.0003 ( 0.12%) | 0.041 |
| 45 | 579087 | 0.2668 ± 0.0004 ( 0.13%) | 0.041 |
| 46 | 443934 | 0.2671 ± 0.0004 ( 0.15%) | 0.042 |
| 47 | 339913 | 0.2671 ± 0.0005 ( 0.17%) | 0.042 |
| 48 | 260519 | 0.2675 ± 0.0005 ( 0.20%) | 0.042 |
| 49 | 199117 | 0.2671 ± 0.0006 ( 0.22%) | 0.042 |
| 50 | 152595 | 0.2674 ± 0.0007 ( 0.26%) | 0.042 |
| 51 | 116705 | 0.2671 ± 0.0008 ( 0.29%) | 0.042 |
| 52 | 89323 | 0.2671 ± 0.0009 ( 0.34%) | 0.042 |
| 53 | 68605 | 0.2680 ± 0.0010 ( 0.38%) | 0.044 |
| 54 | 52570 | 0.2686 ± 0.0012 ( 0.44%) | 0.045 |
| 55 | 40242 | 0.2690 ± 0.0013 ( 0.50%) | 0.046 |
| 56 | 30746 | 0.2690 ± 0.0015 ( 0.57%) | 0.046 |
| 57 | 23481 | 0.2688 ± 0.0018 ( 0.65%) | 0.046 |
| 58 | 17888 | 0.2678 ± 0.0020 ( 0.75%) | 0.043 |
| 59 | 13662 | 0.2673 ± 0.0023 ( 0.86%) | 0.042 |
| 60 | 10427 | 0.2664 ± 0.0026 ( 0.98%) | 0.039 |

Table 1: Island method estimates for $\lambda$ and $K$. A total of 92,441 pairs of length 7000 random sequences were generated using a set of standard amino acid frequencies (26). Island scores were generated for each pair using an extension of the Smith-Waterman algorithm (5), modified for affine gap scores (28). Substitutions were scored using the BLOSUM-62 matrix (27), and gaps of length $k$ were assessed the score $-(11 + k)$. Scores were recorded only for islands anchored within the central 5000×5000 square of each pairwise comparison. Maximum-likelihood estimates for $\lambda$ and $K$ were obtained using equations [4]-[6].

| $c$ | Bias (%) | Standard error (%) |
|---|---|---|
| 22 | 3.1 | 0.6 |
| 23 | 2.6 | 0.7 |
| 24 | 2.1 | 0.8 |
| 25 | 1.7 | 0.9 |
| 26 | 1.4 | 1.0 |
| 27 | 1.1 | 1.2 |
| 28 | 0.9 | 1.3 |
| 29 | 0.7 | 1.5 |
| 30 | 0.6 | 1.8 |
| 31 | 0.4 | 2.0 |
| 32 | 0.3 | 2.3 |
| 33 | 0.2 | 2.6 |
| 34 | 0.1 | 3.0 |
| 35 | 0.1 | 3.5 |
| 36 | 0.0 | 4.0 |
| 37 | 0.0 | 4.5 |

Table 2: Tradeoff between bias and precision in the estimation of $\lambda$. The bias in the estimation of $\lambda$ is calculated from Table 1, assuming $\lambda$'s true value is 0.2670. The standard error assumes an experiment generating 1/10,000 the number of island scores shown in Table 1.

For a given simulation one may estimate well the standard error at any given $c$, but not the bias; if one could estimate bias, one could correct for it. The analysis of a relatively small simulation given in Table 2 is possible only because a much larger simulation has in fact been performed. In practice, one must choose the $c$ at which to estimate $\lambda$ without knowing to any certainty how much bias it entails. We have investigated automatic procedures for choosing $c$, and found several reasonable methods, but none for which an argument of optimality can be advanced. In outline, $\widehat{\lambda}$ decreases systematically for increasing $c$, until its increasing standard error obscures any further change. It is at this point that the cutoff $c$ should be chosen.

## 5 Edge effects and their correction

Independently of the type of bias in estimating $\lambda$ described above, $\widehat{\lambda}$ varies substantially as a function of $m$ and $n$ when $\lambda$ is estimated from traditional borderless (i.e., $b = 0$) $m \times n$ sequence comparisons (20). One may therefore argue that one's estimate of $\lambda$ and $K$ should depend upon the lengths of the real sequences to which they will be applied (22). We here take the alternative view that the length-dependence of $\widehat{\lambda}$ is merely an artifact of finite-length sequence comparison edge effects, and that a correction for these effects is best applied to $m$ and $n$ in equations [1] and [2] rather than to $\lambda$ and $K$.

The central idea of the "edge effect" correction is that high-scoring local alignments from the comparison of two random sequences have an expected length $l(x)$, dependent upon their score $x$, and therefore cannot begin arbitrarily close to the end of either sequence. Accordingly, in place of $m$ and $n$ in equations [1] and [2], the "effective" lengths of the sequences should be taken to be $m' = m - l(x)$ and $n' = n - l(x)$ (20).

Empirically, the mean length $l(x)$ of high-scoring random alignments with sufficiently large score $x$ depends linearly on $x$:

$$l(x) = \alpha x + \beta. \quad [7]$$

We will discuss in a later section the interpretation of $\alpha$ and $\beta$, but note here that these parameters may be estimated by recording the lengths as well as the scores of optimal island alignments. The length of a gapped alignment is interpreted as the average length of the two segments it involves.

For the island method, the way that edge effects bias $\widehat{\lambda}$ is easy to understand. The decay in the observed number of alignments with score at least $x$ is steeper than would be estimated from equation [1] because the effective lengths $m'$ and $n'$ shrink with increasing $x$. Some simple calculus suggests the apparent $\lambda$ from the comparison of sequences of sufficient lengths $m$ and $n$ should be given approximately by

$$\widetilde{\lambda}(m,n) = \lambda + \alpha\left(\frac{1}{m} + \frac{1}{n}\right). \quad [8]$$

For the specific scoring system studied in the massive random simulation above, we estimate $\alpha = 1.90 \pm 0.02$ (see discussion below). Therefore, we expect the apparent $\lambda$ for $n \times n$ comparisons to follow the equation

$$\widetilde{\lambda}(n,n) = 0.2670 + \frac{3.80}{n}. \quad [9]$$

To test this theory, we used the island method to estimate $\lambda$ for the same scoring system studied in the simulation above. We generated islands from many $n \times n$ random sequence comparisons, but with no border for suppressing edge effects. Sufficient comparisons were performed to yield over $10^6$ islands with score at least 37 for each of the 12 lengths $n$ studied; as described above, using this threshold eliminates almost all cutoff-based bias. The resulting maximum-likelihood estimates $\widehat{\lambda}(n,n)$ have a standard error of 0.1%, and are shown as open circles in Figure 4. Given our small uncertainty in $\lambda$ and $\alpha$, for $n > 400$ ($1/n < 0.0025$ in the figure) the data fit the theory of equation [9] to within stochastic error (i.e. two standard deviations). Furthermore, $\widehat{\lambda}(n,n)$ deviates from equation [9] by less than 0.5% for $n > 218$ ($1/n < 0.0045$), and by less than 1% throughout the range studied. For each $n$, we calculated a $\chi^2$ goodness-of-fit test to the geometric distribution; in all 12 cases, the data fit the model with a $p$-value above 0.09.

We emphasize that we do not argue that the line plotted in Figure 4 is more accurate in describing score distribution tail behavior than the experimental $\widehat{\lambda}(n,n)$ produced by the island method. Rather, the good agreement implies the correction we recommend for finite lengths $m$ and $n$ should be sufficiently accurate for comparing proteins of typical size. In evaluating the statistical significance of actual sequence comparisons, one may apply edge-effect corrections either to the sequence lengths, as we suggest, or to $\lambda$, but one should not combine the two corrections. We emphasize further that equation [8] does not permit one to estimate $\lambda$ accurately from a "finite-size" simulation that estimates $\widetilde{\lambda}(m,n)$, because such a simulation will not yield an estimate of the asymptotic value of $\alpha$.

The island method with borders allows one to estimate the "infinite-length" or asymptotic parameters $\lambda$ and $K$ directly, and simultaneously to estimate, as described below, the edge-effect correction parameters $\alpha$ and $\beta$. A single simulation that estimates these four parameters thus permits the statistical evaluation of comparisons of sequences of arbitrary length.

## 6 Comparison of the direct and island methods

For estimating the asymptotic parameters $\lambda$, $K$, $\alpha$ and $\beta$, the island method has a distinct speed advantage over the direct method, as we will discuss below. However, it is easiest first to compare the two methods on the problem of estimating $\widetilde{\lambda}(n,n)$ studied in the previous section. In this

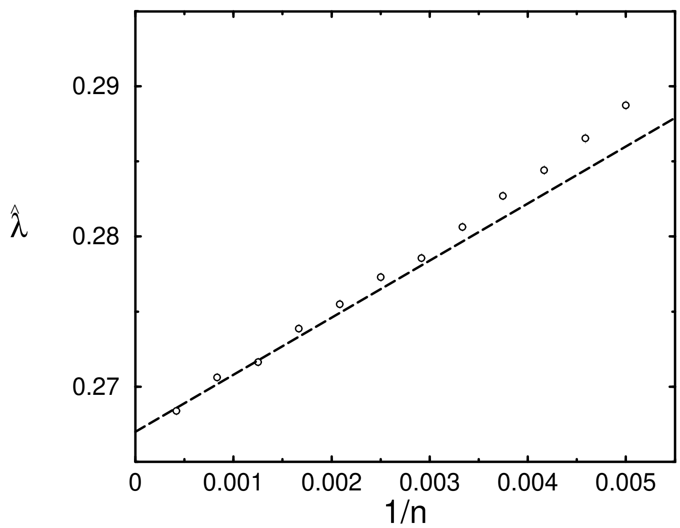


Figure 4: Estimates $\widehat{\lambda}$ derived from borderless $n \times n$ sequence comparisons by the island method as a function of $1/n$. Approximately 1,000,000 islands with score at least 37 were generated to produce the estimates, which thus have a standard error of 0.1%; the size of the symbols represents one standard error. The plotted line represents the theory of equation [9] for the apparent $\widetilde{\lambda}(n,n)$. The scoring system and random sequence model are the same as those described in the legend to Table 1.

"finite size" case, the methods have contrasting advantages. To achieve a standard error $\sigma$ in $\widehat{\lambda}(n,n)/\widetilde{\lambda}(n,n)$, the island method must generate approximately $1/\sigma^2$ data points (see equation [5]), while the direct method need generate only about $0.61/\sigma^2$ points (23). Furthermore, the algorithm for generating island scores requires more computation than that for generating maximal local alignment scores, because it must keep track to which island the score of each path graph cell belongs. Our implementation and timing experiments show the direct method uses only about 70% the time per cell that the island method does. These two factors combined yield a speed advantage of about 240% for the direct method. On the other hand, the island method may generate multiple data points from each $n \times n$ sequence comparison. The expected number of such points depends both upon the length of the sequences being compared and upon the threshold score $c$ as given by equation [1]. The total speed advantage of the island over the direct method is then $Kn^2e^{-\lambda c}/2.4$. In our case study we have been employing $c = 37$ and very many data points to obtain extremely accurate parameter estimates, but as stated above $c = 28$ would be appropriate for a comparison of more typical accuracy. At this threshold, the comparison of two sequences of length 340 yields about 2.4 islands on average, counterbalancing the direct method's speed advantages. For comparisons larger than this, the island method will be faster than the direct method, and slower for smaller comparisons.

This analysis, however, tells only part of the story, because the biases of the direct and island methods in estimating $\widetilde{\lambda}(n,n)$ vary with $n$. To study the extent of this bias, for each length $n$ considered in the previous section we generated sufficient data points for both the direct method and the island method with $c = 28$ to produce estimates $\widehat{\lambda}(n,n)$ with standard error 0.1%. We then compared these estimates to the independent and effectively unbiased estimates (also with standard error 0.1%) shown by the points plotted in Figure 4; the resulting estimates of bias are given in Table 3. For sequence lengths $n \leq 343$ the direct method tends to overestimate $\widetilde{\lambda}(n,n)$ by greater than 1%. Some reflection reveals why this should be the case. For the scoring system under study, approximately 81% of all optimal alignments from 343×343 comparisons have score less than 37, and over 7% have score less than 28. As we learned from our analysis of the island method, including low scoring, largely ungapped, alignments introduces noticeable bias into estimates of $\lambda$. The problem is amplified for the direct method because, due to the extremely fast decay of the left-hand tail of the extreme value distribution, the data points upon which the maximum-likelihood estimate most strongly depends are those with lowest score.

| $n$ | Percent bias | | | Speed ratio | |
|---|---|---|---|---|---|
| | Island | Direct | Censored | Island:Direct | Island:Censored |
| 2400 | +0.9 | -0.2 | -0.2 | 60 | 60 |
| 1200 | +0.7 | 0.0 | 0.0 | 14 | 14 |
| 800 | +0.8 | +0.4 | +0.4 | 6 | 6 |
| 600 | +0.6 | +0.5 | +0.5 | 3 | 3 |
| 480 | +0.5 | +0.8 | +0.6 | 2 | 2 |
| 400 | +0.4 | +0.8 | +0.3 | 1.4 | 1.5 |
| 343 | +0.4 | +1.2 | +0.4 | 1.0 | 1.2 |
| 300 | +0.3 | +1.4 | +0.5 | 0.8 | 1.1 |
| 267 | +0.1 | +1.4 | +0.3 | 0.6 | 1.0 |
| 240 | 0.0 | +1.6 | +0.1 | 0.5 | 0.9 |
| 218 | -0.1 | +1.4 | -0.1 | 0.4 | 0.8 |
| 200 | 0.0 | +1.6 | -0.1 | 0.3 | 0.8 |

Table 3: Bias in the estimation of $\widetilde{\lambda}(n,n)$ of the island, direct, and censored direct methods, and their relative speeds. Maximum likelihood estimates $\widehat{\lambda}$ were derived using the island, direct and censored methods from $n \times n$ sequence comparisons. For the island method with cutoff score $c = 28$, approximately $10^6$ scores were generated, yielding a standard error of 0.1%. For the direct method, approximately $6.1 \times 10^5$ scores were generated, yielding a standard error of 0.1%. For the censored method, sufficient scores above the threshold $c = 28$ were generated to yield a standard error of 0.1% (23); this number ranged from $6.1 \times 10^5$ for $n = 2400$ to $8.4 \times 10^5$ for $n = 200$. Biases are calculated assuming the correct values for $\widetilde{\lambda}(n,n)$ are those of the points plotted in Figure 4. Speed ratios are the computation times required, respectively, by the direct and censored methods divided by the time required by the island method; high speed ratios favor the island method.

Borrowing from our analysis of the island method, it is possible to greatly reduce the bias of the direct method by basing its maximum-likelihood estimate only on those scores that reach a minimum threshold $c$ (23), see also the Appendix. This refinement is achieved at a cost in speed, however, because not every $n \times n$ comparison will yield a data point, and because such "censoring" increases the number of data points required to achieve a given standard error (23). For example, only 56% of 200×200 comparisons have a maximal alignment score at least 28, and with this degree of censoring the number of data points required for a given error increases by 40% (23). For this size comparison, the censored direct method is thus 2.5 times slower than the unmodified method, while still 20% faster than the island method. A fuller analysis gives the speed advantage to the island over the censored method only for comparisons larger than about 280×280. Of course as the size of the comparisons grows, so does the island method's relative speed advantage (Table 3), reaching a factor greater than 3 for comparisons of size 600×600.

The island method has a major speed advantage for comparisons of size $\geq$ 800×800, but it appears to have a corresponding disadvantage with respect to bias (Table 3). This arises because over 99.8% of the optimal alignment scores from 800×800 comparisons are at least 31, and at least 34 from 1200×1200 comparisons. Effectively, the score "threshold" for the direct method increases with comparison size, while we have used a fixed score threshold of 28 for all the island comparisons in Table 3. Were this threshold raised for large comparisons, it would be possible to achieve equivalent bias to the direct method, while retaining a greater than threefold speed advantage. However, for large comparisons, one has the option with the island method to choose greater speed over smaller bias.

If all the island method had to offer were a two to three fold speed advantage for comparisons in the size range 500×500 to 800×800, it would hardly constitute a significant advance. However, our main point is that in lieu of estimating statistical parameters for various finite-size comparisons, a single estimate of the asymptotic $\lambda$ and $K$, along with the edge-effect correction parameters $\alpha$ and $\beta$ will suffice. In this context, the island method has major advantages to the direct method. Most simply, the island method can accurately estimate the asymptotic $\lambda$ by increasing the dimensions of its comparisons to such an extent that finite-size effects become negligible. Because the number of data points the island method generates grows in proportion to the area of its comparisons, there is no loss in speed. In contrast, as we have seen, the direct method pays a heavy penalty in speed as the size of its comparisons grow.

To avoid unduly increasing the comparison size, one might consider adding borders to direct method comparisons, as described above for the island method (Figure 2). This, however, imposes substantial computational overhead. First, one must record where local alignments are "rooted", to avoid counting local alignments rooted outside the central square. The extra computation per cell is similar to keeping track of which island a cell belongs to and increases run time by a factor over 1.4. Second, borders can greatly increase the computational area of medium-sized comparisons. For example, a border of moderate length 200 (see the next section) increases the area of a 600×600 comparison by a factor of 2.8. The two effects combined would slow such a comparison down by a factor close to 4. In contrast, for asymptotic parameter estimation, borders may be added to the island method comparisons at essentially no computational cost — first because the island method must record the roots of local alignments in any case; second because the comparisons' underlying dimensions may be enlarged arbitrarily, rendering inconsequential the additional area entailed by the inclusion of borders.

In conclusion, for finite-size parameter estimation, the island method begins to have a speed advantage only for the comparison of sequences of moderate length. However, for the asymptotic parameter estimation we recommend, the island method has a speed advantage to the direct method approaching an order of magnitude.

## 7 The estimation of $\alpha$ and $\beta$

For optimal local alignments of a given score $x$, the standard deviation in the distribution of alignment lengths is large: about the same as the mean length. Nevertheless, the mean length can be seen to grow approximately linearly with $x$, as illustrated by data from the massive simulation above, plotted in Figure 5. The slope of this dependence does not approach its asymptotic value until $x$ is sufficiently large. Therefore, as with estimates of $\lambda$, estimates of the parameters $\alpha$ and $\beta$ in equation [7] are best calculated by confining attention to alignments with score greater than or equal to a threshold value $c$. In Table 4 we give, for various thresholds, estimates of $\alpha$ and $\beta$ obtained by linear regression on the lengths of the optimal island alignments. Once again, choosing a threshold that balances bias and stochastic error is to some degree arbitrary. We show in Figure

| $c$ | $\hat{\alpha}_c$ | $\hat{\beta}_c$ |
|---|---|---|
| 33 | 1.840 $\pm$ 0.002 | -26.9 $\pm$ 0.1 |
| 34 | 1.847 $\pm$ 0.002 | -27.2 $\pm$ 0.1 |
| 35 | 1.852 $\pm$ 0.002 | -27.4 $\pm$ 0.1 |
| 36 | 1.858 $\pm$ 0.003 | -27.7 $\pm$ 0.1 |
| 37 | 1.864 $\pm$ 0.003 | -27.9 $\pm$ 0.1 |
| 38 | 1.869 $\pm$ 0.004 | -28.2 $\pm$ 0.2 |
| 39 | 1.873 $\pm$ 0.005 | -28.4 $\pm$ 0.2 |
| 40 | 1.877 $\pm$ 0.005 | -28.5 $\pm$ 0.2 |
| 41 | 1.874 $\pm$ 0.006 | -28.4 $\pm$ 0.3 |
| 42 | 1.877 $\pm$ 0.007 | -28.5 $\pm$ 0.3 |
| 43 | 1.88 $\pm$ 0.01 | -28.8 $\pm$ 0.4 |
| 44 | 1.89 $\pm$ 0.01 | -29.0 $\pm$ 0.5 |
| 45 | 1.89 $\pm$ 0.01 | -29.3 $\pm$ 0.6 |
| 46 | 1.91 $\pm$ 0.01 | -30.2 $\pm$ 0.7 |
| 47 | 1.90 $\pm$ 0.02 | -30 $\pm$ 1 |
| 48 | 1.89 $\pm$ 0.02 | -29 $\pm$ 1 |
| 49 | 1.88 $\pm$ 0.02 | -29 $\pm$ 1 |
| 50 | 1.91 $\pm$ 0.03 | -30 $\pm$ 1 |
| 51 | 1.89 $\pm$ 0.03 | -29 $\pm$ 2 |
| 52 | 1.90 $\pm$ 0.03 | -30 $\pm$ 2 |
| 53 | 1.94 $\pm$ 0.04 | -32 $\pm$ 2 |
| 54 | 1.96 $\pm$ 0.05 | -33 $\pm$ 3 |

Table 4: The estimation of $\alpha$ and $\beta$. Estimates for $\alpha$ and $\beta$ were obtained by linear regression of alignment length vrs. score for islands with score at least $c$. Details of the simulation are given in the legend to Table 1.

5 the line implied by the estimates $\hat{\alpha} = 1.90$ and $\hat{\beta} = -30$, yielded by the threshold $c = 47$. These estimates agree within stochastic error to those for all $c \geq 44$.

While the standard error for $\hat{\alpha}$ is 1% at $c = 47$, one is forced to settle for much larger errors in simulations of more realistic size. However, $\alpha$ and $\beta$ are used only to correct the lengths of the sequences being compared, and the significance of alignment scores depends only linearly upon these lengths. Therefore it is generally quite acceptable to estimate $\alpha$ to within 10 or even 20%. The data generated to provide reasonably accurate estimates of the far more important parameter $\lambda$ easily suffice for this purpose.

At a score of 95, the highest score achieved in this simulation, the predicted mean length is less than 150. Therefore, even though the standard deviation of the alignment length is approximately equal to the mean length, the border of length 1000 used in our simulation should be much more than sufficient for estimating the asymptotic values of the parameters $\lambda$, $K$, $\alpha$ and $\beta$, corresponding to "infinite length" comparisons. For comparisons performed without borders, or with borders of insufficient length, estimates of $\alpha$ and $\beta$ deviate from the asymptotic values, just as estimates of $\lambda$ were shown to deviate above.

The expected length of gapped alignments with high score clearly places limits on the applicability of formulas [1] and [2] to the comparison of short sequences, even after edge effects have been corrected for. Specifically, if the expected length of an optimal alignment is longer than the shorter

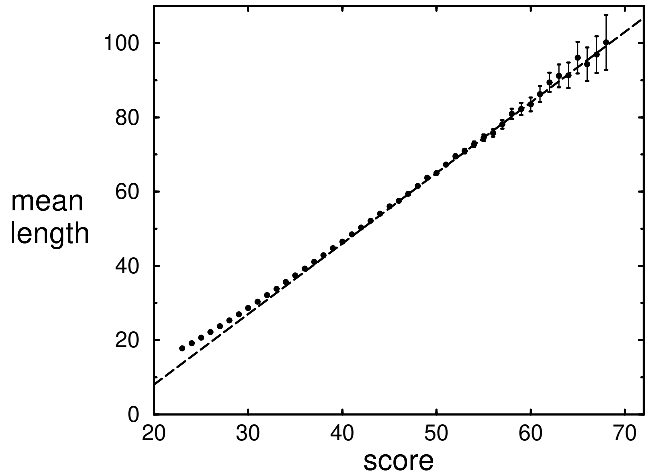


Figure 5: The mean length $l(x)$ of optimal island alignments, as a function of the alignment score $x$. Error bars, representing one standard error, grow with score primarily because the number of alignments on which the mean length estimates are based decreases. The plotted line represents a linear regression on the data for scores $\geq 47$. Details of the simulation are given in the legend to Table 1.

of the two sequences being compared, then one has effectively entered the realm of global sequence comparison, to which our theory no longer applies. This is perhaps best seen as an indication that the combination of substitution and gap costs being employed are tailored for too "distant" similarities, and that a scoring system with a greater *relative entropy* should be used instead (32).

## 8 Relative entropy and the relation of $\alpha$ to $\beta$

It has recently been established under certain simplifying assumptions that in the no-gap case, the edge-effect correction outlined above is the proper first-order correction to equations [1] and [2] for finite-length sequences (J.L. Spouge, personal communication). For high-scoring local alignments without gaps, it can be shown (33) that the average length of alignments with score $x$ is well approximated by

$$l\left(x\right) \approx \alpha_u x = \frac{\lambda_u}{H_u} x, \qquad [10]$$

where $H_u$ is the relative entropy of the scoring system in nats (32), and the subscripts $u$ indicate we are speaking of ungapped alignments. It is therefore reasonable to define, and estimate, the relative entropy per amino acid pair for gapped alignments by the formula

$$H_g = \frac{\lambda_g}{\alpha_g}, \qquad [11]$$

where the subscript $g$ indicates the gapped case.

Given this definition, we estimate $H_g$ for the scoring system studied above to be 0.141 nats $\pm$ 2%. Note that for the identical scoring system, Altschul and Gish (20) obtained the much greater estimate of 0.25 nats for $H_g$, due primarily to their assumption that $\beta$ is 0 in equation [7]. This assumption yields a good estimate of $H_g$ only in the limit of very large scores $x$, a limit not nearly approached in simulations of practical size.

| Matrix | BLOSUM-45 | BLOSUM-62 | BLOSUM-80 | PAM-70 | PAM-30 |
|---|---|---|---|---|---|
| $\alpha_u$ | 0.9113 | 0.7916 | 0.5222 | 0.3250 | 0.1938 |
| $\beta_u$ | -5.7 | -3.2 | -1.6 | -0.7 | -0.3 |
| Gap existence | 14 | 11 | 10 | 10 | 9 |
| Gap extension | 2 | 1 | 1 | 1 | 1 |
| $\alpha_g$ | $1.92 \pm 0.03$ | $1.90 \pm 0.02$ | $1.07 \pm 0.02$ | $0.70 \pm 0.01$ | $0.48 \pm 0.01$ |
| $\beta_g$ | $-37.2 \pm 1.6$ | $-29.7 \pm 1.0$ | $-12.5 \pm 0.8$ | $-8.1 \pm 0.5$ | $-5.9 \pm 0.3$ |
| $2G(\alpha_u - \alpha_g) + \beta_u$ | $-38.0 \pm 1.0$ | $-29.8 \pm 0.5$ | $-13.7 \pm 0.4$ | $-9.0 \pm 0.3$ | $-6.0 \pm 0.2$ |

Table 5: The estimation of $\beta$ using $\alpha$. Estimates for $\alpha$ and $\beta$ were obtained by linear regression of alignment length vrs. score, for scores attaining at least a cutoff value. Sequences were generated using a set of standard amino acid frequencies (26). Substitution scores were from either the BLOSUM (27) or PAM (34,35) series. Affine gap scores charged an existence penalty for each gap, and an extension penalty for each residue within a gap. Cutoffs scores were chosen sufficiently high to avoid detectable bias in estimating $\alpha$. Sufficient data points were generated to estimate $\alpha$ with a standard error of less than 2%. For each scoring system studied, this required over 500 pairs of random sequences, recording islands anchored within the central $5000 \times 5000$ square of each pairwise comparison. Borders of length 1000 were used for the BLOSUM-45 and BLOSUM-62 scoring systems, and of length 500 for all others.

Given that for ungapped alignments $\beta_u$ is near zero, as seen by experiment (see Table 5 for some examples), one may ask why $\beta_g$ should be distinctly negative. An understanding is to realize that for a scoring system in which a gap of length one has score $-G$, at each end of an optimal alignment there must be a section with score $+G$ that does not include gaps. The average lengths of these sections will be described better by the ungapped than by the gapped $\alpha$. This is a much stronger effect than the fact that an optimal alignment may not begin or end with a negatively scoring aligned pair of letters, which causes $\beta_u$ to be slightly negative. Together, these two effects lead to the prediction that the parameter $\beta_g$ can be approximated by the formula

$$\beta_g \approx 2G\,(\alpha_u - \alpha_g) + \beta_u. \qquad [12]$$

For the particular scoring system and random letter frequencies we have been studying, $G = 12$, $\alpha_u = 0.79$, and $\beta_u = -3.2$. In conjunction with our estimate of $1.90 \pm 0.02$ for $\alpha_g$, this yields an estimate of $-29.8 \pm 0.5$ for $\beta_g$, which coincides with the experimental value of $-30 \pm 1$ within the precision of measurement. Similar agreement is found for other gap costs and scoring systems that are not too close to the log-linear transition (see Table 5).

Equation [12] suggests that, with a knowledge of the easily accessible $\alpha_u$ and $\beta_u$, the estimation of $\alpha_g$ alone is sufficient for the edge-effect correction. In practice, however, estimating $\beta_g$ requires no more work than estimating $\alpha_g$, so one might as well use the experimental value.

## 9 Discussion and conclusion

It was originally claimed that the primary advantage of the island over the direct method for estimating statistical parameters lay in speed (12). We have shown here that this is substantially true only for asymptotic parameter estimation. However, we also have argued that edge-effect parameters allow a single estimation of asymptotic parameters to replace all finite-size parameter estimates. Furthermore, the island method permits simple maximum-likelihood estimation of $\lambda$

that accounts for discrete score data, and it allows for simultaneous parameter estimation using various score thresholds $c$, and thus the controlled trade-off of systematic bias and stochastic error.

The parameter $\lambda$ depends not only upon the scoring system employed, but also upon the letter frequencies of the sequences being compared. In practice, $\lambda$ may sometimes vary by over 10% from one pair of sequences to another, due merely to variations in sequence composition. Yet, in the context of a database search, it is simply too time consuming to reestimate $\lambda$ for each pairwise comparison of potential interest: one moderately accurate estimate of $\lambda$ requires as much time as searching a typical current database using standard heuristic methods (4). Thus, while one may precompute highly accurate estimates of $\lambda$ for a fixed "standard" composition, isn't this accuracy vitiated by varying compositions?

Two solutions to the problem of varying background frequencies have been proposed, both of which can make use of accurate parameter estimation procedures. Altschul et al. (4) have suggested that for non-standard letter frequencies, the substitution scores be rescaled so as to set the calculable (13) parameter $\lambda_u$ equal to that for the original substitution scores used with standard frequencies. The conjecture is that the precalculated $\lambda_g$ will then apply to gapped alignments using the rescaled substitution scores in the context of the non-standard frequencies. This procedure has been implemented with good results (36). Alternatively, Mott (22) has used random simulations for a very large number of different scoring systems, gap costs, sequence compositions and sequence lengths to derive an empirical formula for $\lambda$, dependent upon variables calculable from the scoring system, letter frequencies, and sequence lengths. Because the values of $\widehat{\lambda}$ used in deriving this formula were calculated by the direct method, frequently with short sequences, some improvement in Mott's formula may be obtainable using the methods described here. To be more conservative, statistical parameters may be based upon residue compositions within sequence regions containing the aligned segments of interest. In general, by improving the precision with which statistical parameters are estimated for local sequence alignment, more accurate judgments can be rendered concerning the biological relevance of protein and DNA sequence similarities.

## 10 Acknowledgments

We thank Dr. John Spouge for helpful conversations. This research is supported by the National Science Foundation through grants no DMR-9971456 and DBI-9970199. R.B. and T.H. are grateful to the hospitality of the N.C.B.I. through its Scientific Visitors Program. In addition, R.B. acknowledges a Hochschulsonderprogramm III fellowship of the DAAD, R.O. acknowledges an LJIS fellowship by the Wellcome-Burroughs Fund, and T.H. a Beckman Young Investigator Award.

## 11 Appendix: Maximum likelihood fitting

In this appendix we explain the maximum likelihood fitting technique in the presence of discrete scores. In the case of the extreme value distribution this extends the more commonly used maximum likelihood fitting for continuous scores as it is, e.g., presented by Lawless (23). By analogy to some analytical results on discrete extreme value distributions (37), for small lattice spacing we expect only small deviations in the estimated parameters due to the discreteness of the scores as long as we perform *uncensored* fits. However, for alignment scores it is often necessary to estimate the parameters only for a subset of the observed scores. For such a *censored* fit, the discreteness of the scores must be taken into account, as discussed here, to obtain correct maximum likelihood estimates of the parameters of the underlying distribution.

Throughout the appendix we will assume that sufficiently large island scores $S$ follow a geometric distribution

$$\text{Prob}\,(S = x) = Dp^x, \tag{13}$$

where we write $p = \exp(-\lambda)$ in order to emphasize the discrete character of the scores $S$. In the simplest case, the distribution is of this geometric form for all $x \geq 0$ which fixes the prefactor $D$ through normalization to $D = 1 - p$. Let us assume that we observed $N$ islands with the scores $x_1$, ..., $x_N$ and we wish to find the value of $p$ (i.e., $\lambda$) which best describes these observed scores. Since the probability of observing these scores is just the product of their individual probabilities the logarithm of the total probability (i.e., the log-likelihood) is

$$\ln \text{Prob}\,(x_1, ..., x_N) = \ln p \sum_{i=1}^{N} x_i + N \ln (1 - p)\,. \tag{14}$$

The best value $\widehat{p}$ of $p$ is the one which maximizes this expression. We can obtain it by equating the first derivative of this expression to zero. This yields after some simple algebra

$$\widehat{\lambda} = -\ln \widehat{p} = \ln \left( 1 + \frac{1}{\frac{1}{N} \sum_{i=1}^{N} x_i} \right). \tag{15}$$

In our application, the distribution of island scores is not of the geometric form [13] for all scores $x$. It follows this form only asymptotically for large scores $x$. In this case, the prefactor $D$ is no longer fixed by normalization. Rather, it depends on the shape of the distribution for small $x$. In order to get a good estimate for $p$ we have to perform a censored fit, i.e., we keep only those island scores $x$ with value at least $c$. The integer cutoff $c$ is chosen so that the geometric form [13] is a reasonable description of the data. This is commonly called Type I censoring (23). We expect the censored scores to be distributed according to the restricted probabilities

$$\text{Prob}\,(S = x | S \geq c) = \frac{Dp^x}{\sum_{k=c}^{\infty} Dp^k} = (1 - p)\, p^{x-c} \tag{16}$$

which is independent of the unknown normalization factor $D$. If, out of $N$ total scores, the scores $x_1$, ..., $x_M$ attain a value of $c$ or larger, the logarithm of the probability is

$$\ln \text{Prob}\,(x_1, ..., x_M | x_1 \geq c, ..., x_M \geq c) = \ln p \sum_{i=1}^{M} (x_i - c) + M \ln (1 - p)\,. \tag{17}$$

This log-likelihood function is identical to the one without censoring presented in equation [14] except for the shift of all scores by the cutoff $c$. Thus, the optimal value $\widehat{\lambda}$ is given by

$$\widehat{\lambda} = -\ln \widehat{p} = \ln \left[ 1 + \frac{1}{\frac{1}{M} \sum_{i=1}^{M} (x_i - c)} \right]. \tag{18}$$

From this expression it becomes obvious why it is important to take the discreteness of the scores into account for censored fits. The maximum likelihood estimate $\widehat{\lambda}$ depends explicitly on the cutoff, as $c$ appears in equation [18], but the set of scores $x_1$, ..., $x_M$ remains unchanged as the cutoff $c$ is varied between two adjacent integers. Therefore, it is important to demand that $c$ be integral, taking the discreteness of the scores into account.

In order to get an estimate $\widehat{K}$ for the other distribution parameter $K$, we have to employ the expected number $E(x)$ of islands with score at least $x$ given by equation [1]. If we observe $R_c$

islands with score at least $c$ in $B$ pairwise comparisons we get $R_c \approx BE(c) \approx B\widehat{K}mn\widehat{p}^c$ which can be rearranged into equation [6] for the maximum likelihood estimate $\widehat{K}$.

If we choose the direct rather than the island method to estimate $\lambda$, we are interested in the distribution

$$\text{Prob}\,(S' = x') = \text{Prob}\,(S' \leq x') - \text{Prob}\,(S' \leq x' - 1) \qquad [19]$$

of the optimal local alignment scores. The optimal local alignment score $S'$ is the maximum of all the approximately $\rho mn$ island scores of the two sequences compared, where $\rho$ describes the typical island density. Thus, the two probabilities on the right hand side of equation [19] can be expressed by the distribution of island scores $S$, and $S'$ is distributed according to

$$\text{Prob}\,(S' = x') = \text{Prob}\,(S \leq x')^{\rho mn} - \text{Prob}\,(S \leq x' - 1)^{\rho mn}\,. \qquad [20]$$

Since each island score follows the geometric distribution [13] we get

$$\text{Prob}\,(S' = x') = \left(1 - D\frac{p^{x'+1}}{1-p}\right)^{\rho mn} - \left(1 - D\frac{p^{x'}}{1-p}\right)^{\rho mn} \approx \exp\left(-Kmnp^{x'+1}\right) - \exp\left(-Kmnp^{x'}\right) \qquad [21]$$

with $K = \rho D/(1-p)$. The last approximation is justified since $Dp^{x'}/(1-p)$ is a small number for all the scores $x'$ that we are interested in. Equation [21] suggests that the optimal alignment scores are indeed extreme value distributed. The only influence of the discreteness of the scores is that the probability $\text{Prob}(S' = x')$ is given by finite differences of the extreme value distribution instead of being a proper probability density, i.e., the derivative of the extreme value distribution.

Let us now assume that we performed $B$ comparisons. We again choose a cutoff $c$, and keep only the $M$ optimal local alignment scores $x'_1$, ..., $x'_M$ that are greater than or equal to $c$. (If we are interested in an uncensored fit, we can always choose $c = 0$ since all local alignment scores are positive.) These scores are expected to follow the distribution

$$\text{Prob}\,(S' = x'|S' \geq c) = \frac{\text{Prob}\,(S' = x')}{\sum_{k=c}^{\infty} \text{Prob}\,(S' = k)} = \frac{\exp\left(-Kmnp^{x'+1}\right) - \exp\left(-Kmnp^{x'}\right)}{1 - \exp\left(-Kmnp^{c}\right)}. \qquad [22]$$

The logarithm $L(p, K) = \ln \text{Prob}(x'_1, ..., x'_M | x'_1 \geq c, ..., x'_M \geq c)$ of the probability of observing the censored scores $x'_1$, ..., $x'_M$ then becomes

$$L\,(p, K) = -Kmn \sum_{k=1}^{M} p^{x'_k} + \sum_{k=1}^{M} \ln\left[\exp\left(Kmn\,(1-p)\,p^{x'_k}\right) - 1\right] - M \ln\left[1 - \exp\left(-Kmnp^c\right)\right]. \qquad [23]$$

As before, the best estimates $\widehat{p} = \exp(-\widehat{\lambda})$ and $\widehat{K}$ of the two parameters $p$ and $K$, given the observed data $x'_1$, ..., $x'_M$, are the ones which maximize the function $L(p, K)$. We could try to find this maximum by taking the derivatives of $L(p, K)$ with respect to $p$ and $K$ and equating them to zero. However, this leads to a pair of equations that can only be solved numerically. Therefore, it is better to directly use a numerical minimization algorithm applied to the function $-L(p, K)$. We used the downhill simplex method in two dimensions (38). In order to improve convergence, it can be conveniently started at the values of $\widehat{p}$ and $\widehat{K}$ which are obtained by the relatively simple uncensored, continuous extreme value fit to the data. Then, it converges rapidly towards the global minimum $(\widehat{p}, \widehat{K})$ of the function $L(p, K)$.

## 12 Literature


1. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.

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

3. Gish, W. and States, D.J. (1993) Identification of protein coding regions by database similarity search. *Nature Genet.* **3**, 266–272.

4. 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 Res.* **25**, 3389–3402.

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

6. Sellers, P.H. (1984) Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol.* **46**, 501–514.

7. Waterman, M.S., Gordon, L. and Arratia, R. (1987) Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl. Acad. Sci. USA* **84**, 1239–1243.

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

9. Dembo, A., Karlin, S., Zeitouni, O. (1994) Critical phenomena for sequence matching with scoring. *Ann. Prob.* **22**, 1993–2021.

10. Vingron, M. and Waterman, M.S. (1994) Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.* **235**, 1–12.

11. Gumbel, E.J. (1958) *Statistics of extremes.* Columbia University Press, New York, NY.

12. Olsen, R., Bundschuh, R. and Hwa, T. (1999) Rapid assessment of extremal statistics for gapped local alignment. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D., Glasgow, J., Mewes, H.-W. and Zimmer, R. (eds.), *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park, CA, pp. 211–222.

13. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.

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

15. Smith, T.F., Waterman, M.S. and Burks, C. (1985) The statistical distribution of nucleic acid similarities. *Nucl. Acids Res.* **13**, 645–656.

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

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

18. Waterman, M.S. and Vingron, M. (1994) Rapid and accurate estimates of statistical significance for sequence database searches. *Proc. Natl. Acad. Sci. USA* **91**, 4625–4628.

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

20. Altschul, S.F. and Gish, W. (1996) Local alignment statistics. *Meth. Enzymol.* **266**, 460–480.

21. Pearson, W.R. (1998) Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.

22. Mott, R. (2000) Accurate formula for P-values of gapped local sequence and profile alignments. *J. Mol. Biol.* **300**, 649–659.

23. Lawless, J.F. (1982) *Statistical models and methods for lifetime data.* Wiley, New York, NY, pp. 141–202.

24. Altschul, S.F. and Erickson, B.W. (1986) Locally optimal subalignments using nonlinear similarity functions. *Bull. Math. Biol.* **48**, 633–660.

25. Waterman, M.S. and Eggert, M. (1987) A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons. *J. Mol. Biol.* **197**, 723–728.

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

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

28. Gotoh, O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705–708.

29. Fitch, W.M. and Smith, T.F. (1983) Optimal sequence alignments. *Proc. Natl. Acad. Sci. USA* **80**, 1382–1386.

30. Altschul, S.F. and Erickson, B.W. (1986) Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.* **48**, 603–616.

31. Myers, E.W. and Miller, W. (1988) Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11–17.

32. Altschul, S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* **219**, 555–565.

33. Dembo, A. and Karlin, S. (1991) Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Prob.* **19**, 1737–1755.

34. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) A model of evolutionary change in proteins. In Dayhoff, M.O. (ed.), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3, pp. 345–352.

35. Schwartz, R.M. and Dayhoff, M.O. (1978) Matrices for detecting distant relationships. In Dayhoff, M.O. (ed.), *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Washington, DC, Vol. 5, Suppl. 3, pp. 353–358.

36. Schäffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**, 1000–1011.

37. Arratia, R., Gordon, L. and Waterman, M.S. (1986) An extreme value theory for sequence matching. *Ann. Stat.* **14**, 971–993.

38. Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical recipes in C. The art of scientific computing.* Second Edition. Cambridge University Press, New York, NY, pp. 408–412.