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

July 11, 2000

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 focussed 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 λ 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 advantages:

1. It renders explicit a tradeoff between parameter estimate bias and stochastic error, and allows this tradeoff to be easily controlled;

2. It allows parameter estimates to be obtained for arbitrary length sequence comparisons, including the infinite-length limit;

3. It estimates accurately the tail behavior of score distributions for small-length comparisons.

For parameter estimation at fixed levels of bias and precision, the island method appears to have a speed advantage, but this advantage is not as great as previously thought.

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

$$E = Kmne^{-\lambda x}, \qquad [1]$$

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

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

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 λ 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 $\hat{K}$ and $\hat{\lambda}$ for the statistical parameters in equation [2] (23). If R scores are generated, the ratio $\hat{\lambda}/\lambda$ is approximately normally distributed, with mean 1 and standard deviation $0.78/\sqrt{R}$ (23).

Because λ enters equations [1] and [2] exponentially, accurate estimates of λ are particularly important. Marginally significant alignments from current database searches typically have a scaled score $\lambda x > 25$, for which even a 2% error in λ leads to a greater than 65% error in estimated E-value. Thus, standard errors of less than 1%, or even 0.5% in $\hat{\lambda}$ are desirable.

Note that the development in (23) assumes continuous data, whereas alignment scores are almost always discrete. If the scale parameter λ times the lattice spacing δ of possible scores is small, the error introduced by assuming continuous scores is minor. By analogy to analytic results for discrete extreme value distributions (24), $\hat{\lambda}/\lambda$ should be slightly biased, with mean $1-(\lambda\delta/2\pi)^2$. If $\lambda\delta < 0.28$, for example, the correction needed to unbias $\hat{\lambda}$ is smaller than 0.2%. The less important $\hat{K}$ requires a corresponding correction.

3 The island method

Recently, Olsen et al. (12) proposed the island method for estimating λ 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 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 λ 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 (6,12,25,26). 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 (18,19,25,26) 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 λ 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 of these islands in excess of c:

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

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

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

The standard deviation of $\hat{\lambda}_c/\lambda$ 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 $\lambda < 1$. Were the island scores not discrete, the maximum-likelihood estimate $\hat{\lambda}_c$ would instead be simply $1/\overline{S}_c$, and the standard deviation of $\hat{\lambda}_c/\lambda$ would be $1/\sqrt{R_c}$.

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

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

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 λ 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 λ 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 limiting values of λ and K, so it is desirable to correct for any finite-length bias in the estimation of these parameters. We will describe below a method for making such a correction, but here note that, by eliminating edge effects, the island method can estimate limiting values of λ 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.

4 The tradeoff of speed, bias, and precision

Because of λ's exponential role in equations [1] and [2], accurate estimates for λ are far more important than those for K, and we shall therefore focus on the estimation of λ. 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 (27) to generate 92,441 pairs of length-7000 "random amino acid sequences". We compared each pair using the BLOSUM-62 amino acid substitution matrix (28), in conjunction with *affine* gap scores (29-32) of $-(11 + k)$ for gaps of length k. To suppress edge effects, scores were tabulated only for islands anchored within the central 5000×5000 square of each pairwise comparison; approximately 10^{12} total island scores were recorded. Using equations [3]-[6], estimates of λ and K were obtained from these data for a range of the cutoff scores c; the results are summarized in Table 1 and the results for $\hat{\lambda}$ are plotted in Figure 3.

While the estimates $\hat{\lambda}_c$ 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, and for low thresholds $\hat{\lambda}_c$ is therefore biased towards the higher λ applicable to local alignments that exclude gaps. In this example, $\hat{\lambda}_c$ falls monotonically for $c \geq 20$, until it reaches the value 0.2670 at $c = 37$; thereafter, $\hat{\lambda}_c$ 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 $\hat{\lambda}_c$ 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 $\hat{\lambda}_c$. To illustrate the point, consider a realistically sized random simulation, 2500 times smaller than that shown in Table 1, which would require about 10 minutes on a modern workstation. The systematic bias in the $\hat{\lambda}_c$ from such a simulation should be the same as seen in Table 1, but the standard errors will be 50 times larger. Table 2 shows the resulting tradeoff between bias and precision. The best tradeoff probably occurs near $c = 31$, where the sum 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.

For a given simulation one has direct knowledge of the standard error at any given c, but not of 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 λ 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, $\hat{\lambda}$ decreases systematically for increasing c, until its increasing standard error obscures any further change. It is at this point — for instance near the first c for which $\hat{\lambda}_c < \hat{\lambda}_{c+1}$ — that the cutoff c should be chosen.

5 Edge effects and their correction

Independently of the type of bias in estimating λ described above, $\hat{\lambda}$ varies substantially as a function of m and n when λ is estimated from traditional borderless (i.e., $b = 0$) $m \times n$ sequence comparisons (20). One may therefore argue that one's estimate of λ 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 $\hat{\lambda}$ 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 λ 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).

As we will discuss below, the mean length $l(x)$ of high-scoring random alignments with sufficiently large score x empirically depends linearly on x:

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

By recording the lengths as well as the scores of optimal island alignments, one may estimate the parameters α and β. 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 $\hat{\lambda}$ 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 λ from the comparison of sequences of lengths m and n should be given approximately by

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

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 λ for $n \times n$ comparisons to follow the equation

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

To test this theory, we used the island method to estimate λ 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 $\hat{\lambda}(n,n)$ have a standard error of 0.1%, and are shown as open circles in Figure 4. Given our small uncertainty in λ and α, for $n > 400$ ($1/n < 0.0025$ in the figure) the data fit the theory of equation [9] within stochastic error (i.e. two standard deviations). Furthermore, $\hat{\lambda}(n,n)$ deviates from theory 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 χ^2 goodness-of-fit test to the geometric distribution [1]; the results are shown in Table 3. In all 12 cases, the data fit the distribution with a p-value above 0.09.

While the predictable behavior of $\widetilde{\lambda}(m,n)$ allows one either to estimate λ from finite-sized borderless comparisons, or to estimate $\widetilde{\lambda}(m,n)$ from $\hat{\lambda}$ for "infinite" comparisons, we recommend neither procedure in practice. Our point in this section is merely to explain the observed behavior of $\hat{\lambda}(m,n)$ as a consequence of edge effects. In evaluating the statistical significance of actual sequence comparisons, one may correct, as described above, either sequence lengths or λ for edge effects, but one should not combine the two corrections.

We recommend estimating the "infinite-length" parameters directly using the island method with borders. Estimates are thereby obtained not only for λ and K as discussed above, but also for the "infinite-length" α and β as described in section 7 below. Once these four "infinite-length" parameters are known, only the lengths m and n should be corrected for edge effects.

6 Critique of the direct method of estimating λ

In order to compare the island and direct methods, for each length n studied in the previous section we generated over 10^6 pairs of random length-n sequences, and for each pair found the optimal local alignment score using the same scoring system as above. We fit an extreme value distribution to these scores by maximum-likelihood (23), correcting for the discreteness of the data, as described in section 2 above. The resulting $\hat{\lambda}(n,n)$ are shown as filled circles in Figure 4, and corresponding χ^2 goodness-of-fit tests in Table 3.

It is clear from Table 3 that for $n <= 600$ the data begin to fit the extreme value distribution rather poorly, and from Figure 4 it is evident that for the same range ($1/n > 0.0015$) the direct and island method $\hat{\lambda}(n,n)$ begin to diverge substantially from one another. Some reflection reveals why this should be the case. For the scoring system under study, approximately half of all optimal alignments from 600×600 comparisons have score less than 37. As we learned from our analysis of the island method, including alignments below this score begins to introduce noticeable bias into estimates of λ, because equations [1] and [2] are valid only the limit of sufficiently high scores. 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. In contrast, the island method estimates λ using only the highest-scoring alignments, whose distribution is of course that of primary interest.

For $n \geq 800$ ($1/n < 0.0015$), the island and direct method $\hat{\lambda}$'s differ by less than 0.5%, and the χ^2 goodness-of-fit test indicates these comparisons to be of sufficient size that the extreme value

distribution [2] applies. However, for random sequences of this length, each pairwise comparison yields a single data point for estimating λ by the direct method, but on average more than one point for the island method. Thus, except at $n = 800$, the direct takes substantially longer than the island method to generate $\hat{\lambda}$'s with the same standard error.

For the three cases with $n \geq 800$ ($1/n < 0.0015$), Figure 4 shows that the direct method does not conform to the theory of equation [9] as well as the island method. This is most likely due to the non-uniform reliance the direct method places upon data points, with those of lowest score receiving greatest emphasis. This renders the edge effect of smaller consequence to the direct than to the island method. However, because we currently have no theory for the magnitude of the edge effect on the direct method, we cannot estimate accurately the "infinite" λ from $\hat{\lambda}(n, n)$ of the direct method.

The defects of the direct method can be mitigated, but the remedies only move it closer to the island method. For example, maximum-likelihood estimation may be applied to only the highest-scoring fraction of the optimal alignment scores generated (23). As with raising the island method cutoff score, this decreases systematic bias, but at the cost of greater stochastic error. Also, to suppress edge-effect bias, borders may be added to sequence comparisons, as shown in Figure 2. This, however, imposes substantial computational overhead. One must record where local alignment scores are "rooted", which the island method does in any case. More importantly, because the direct method produces only one data point per comparison, the length of the sequences compared must be kept small for efficiency. Borders of sufficient length to be effective will thus add greatly to the overall running time. In contrast, because it may extract many data points from a single comparison, the island method may compare arbitrarily long sequences, rendering inconsequential the extra computational cost imposed by borders.

7 The estimation of α and β

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 λ, estimates of the parameters α and β 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 α and β 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 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, α and β 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 α within 10 or even 20%. The data generated to provide reasonably accurate estimates of the far more important parameter λ 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 limiting values of parameters λ, K, α and β, corresponding to "infinite length" comparisons. For comparisons performed without borders, or with borders of insufficient length, estimates of α and β deviate from the limiting values, just as estimates of λ 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 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 (33).

8 Relative entropy and the relation of α to β

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 (34). For high-scoring local alignments without gaps, it can be shown (35) that the average length of alignments with score x is well approximated by

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

where H_u is the relative entropy of the scoring system in nats (33), 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 β 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.

Given that for ungapped alignments β_u is near zero, as seen by experiment (see Table 5 for some examples), one may ask why β_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 α. 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 β_u to be slightly negative. Together, these two effects lead to the prediction that the parameter β_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 $\beta_u = -3.7 \pm 0.2$. In conjunction with our estimate of 1.90 ± 0.02 for α_g, this yields an estimate of -30.3 ± 0.5 for β_g, which coincides with the experimental value of -30 ± 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 α_u and β_u, the estimation of α_g alone is sufficient for the edge-effect correction. In practice, however, estimating β_g requires no more work than estimating α_g, so one might as well use the experimental value.

9 Discussion and conclusion

There are multiple advantages of the island over the direct method for estimating statistical parameters for gapped local sequence alignments. It was originally claimed that the primary advantage lay in speed (12). An improvement in speed is apparent, but for estimates with equivalent bias and stochastic error, this advantage is not as great as at first claimed. The main advantages lie rather in the ease with which systematic errors can be controlled, and with which limiting values of λ, K,

α and β can be estimated. Specifically, (i) the island method easily permits maximum-likelihood estimation of λ to account for discrete score data; (ii) it allows for simple, simultaneous parameter estimation using various score thresholds c, and thus the controlled trade-off of systematic bias and stochastic error; (iii) it permits the use of borders to estimate directly the limiting values for λ and K, and the edge-effect parameters α and β; because the number of data points grows with the size of the comparisons performed, very large comparisons are feasible, rendering inconsequential the added computational cost of a sufficiently large border.

The parameter λ depends not only upon the scoring system employed, but also upon the letter frequencies of the sequences being compared. In practice, λ 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 λ for each pairwise comparison of potential interest. While one may precompute highly accurate estimates of λ 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 λ_u equal to that for the original substitution scores used with standard frequencies. The conjecture is that the precalculated λ_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 λ, dependent upon variables calculable from the scoring system, letter frequencies, and sequence lengths. Because the $\hat{\lambda}$'s 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. In general, by improving the precision with which statistical parameters are estimated for local sequence alignment, more accurate judgements 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 grant no DMR-9971456. 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 Hoch-

schulsonderprogramm 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 Literature

1. Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448.

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

3. Gish, W. and States, D.J. (1993) *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) *Nucleic Acids Res.* **25**, 3389–3402.

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

6. Sellers, P.H. (1984) *Bull. Math. Biol.* **46**, 501–514.

7. Waterman, M.S., Gordon, L. and Arratia, R. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1239–1243.

8. Arratia, R. and Waterman, M.S. (1994) *Ann. Appl. Prob.* **4**, 200–225.

9. Dembo, A., Karlin, S., Zeitouni, O. (1994) *Ann. Prob.* **22**, 1993–2021.

10. Vingron, M. and Waterman, M.S. (1994) *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) 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) *Proc. Natl. Acad. Sci. USA* **87**, 2264–2268.

14. Dembo, A., Karlin, S., Zeitouni, O. (1994) *Ann. Prob.* **22**, 2022–2039.

15. Smith, T.F., Waterman, M.S. and Burks, C. (1985) *Nucl. Acids Res.* **13**, 645–656.

16. Collins, J.F., Coulson, A.F.W. and Lyall, A. (1988) *Comput. Appl. Biosci.* **4**, 67–71.

17. Mott, R. (1992) *Bull. Math. Biol.* **54**, 59–75.

18. Waterman, M.S. and Vingron, M. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 4625–4628.

19. Waterman, M.S. and Vingron, M. (1994) *Stat. Sci.* **9**, 367–381.

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

21. Pearson, W.R. (1998) *J. Mol. Biol.* **276**, 71–84.

22. Mott, R. (2000) *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. Arratia, R., Gordon, L. and Waterman, M.S. (1986) *Ann. Stat.* **14**, 971–993.

25. Altschul, S.F. and Erickson, B.W. (1986) *Bull. Math. Biol.* **48**, 633–660.

26. Waterman, M.S. and Eggert, M. (1987) *J. Mol. Biol.* **197**, 723–728.

27. Robinson, A.B. and Robinson, L.R. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 8880–8884.

28. Henikoff, S. and Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919.

29. Gotoh, O. (1982) *J. Mol. Biol.* **162**, 705–708.

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

31. Altschul, S.F. and Erickson, B.W. (1986) *Bull. Math. Biol.* **48**, 603–616.

32. Myers, E.W. and Miller, W. (1988) *Comput. Appl. Biosci.* **4**, 11–17.

33. Altschul, S.F. (1991) *J. Mol. Biol.* **219**, 555–565.

34. Spouge, J.L. *J. Appl. Prob.*, submitted.

35. Dembo, A. and Karlin, S. (1991) *Ann. Prob.* **19**, 1737–1755.

36. Schäffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) *Bioinformatics* **15**, 1000–1011.

12 Tables

c	R_c	$\hat{\lambda}_c$	$\hat{K}_c$
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 λ and K

c	Bias (%)	Standard error (%)
22	3.1	0.3
23	2.6	0.3
24	2.1	0.4
25	1.7	0.4
26	1.4	0.5
27	1.1	0.6
28	0.9	0.7
29	0.7	0.8
30	0.6	0.9
31	0.4	1.0
32	0.3	1.2
33	0.2	1.3
34	0.1	1.5
35	0.1	1.7
36	0.0	2.0
37	0.0	2.3

Table 2: Tradeoff between bias and precision in the estimation of λ

n	Geometric			Extreme value		
	χ^2	d.f.	p-value	χ^2	d.f.	p-value
2400	45.6	41	0.29	55.7	51	0.30
1200	26.6	41	0.96	58.2	50	0.20
800	25.6	41	0.97	71.8	50	0.023
600	45.5	40	0.25	120.7	49	6×10^{-8}
480	30.9	40	0.85	127.0	49	8×10^{-9}
400	34.4	40	0.72	139.7	48	7×10^{-11}
343	25.2	40	0.97	141.7	48	4×10^{-11}
300	51.0	39	0.094	166.9	47	2×10^{-15}
267	44.0	39	0.27	175.2	47	1×10^{-16}
240	43.5	39	0.29	191.9	47	2×10^{-19}
218	43.6	39	0.28	234.1	46	5×10^{-27}
200	39.2	38	0.42	272.0	46	8×10^{-34}

Table 3: χ^2 goodness-of-fit test for the geometric (island method) and extreme value (direct method) distributions (d.f. denotes degrees of freedom)

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

Table 4: The estimation of α and β

Matrix	BLOSUM-45	BLOSUM-62	BLOSUM-80	PAM-70	PAM-30
α_u	0.9114	0.7916	0.5209	0.3570	0.1939
β_u	-6.22 ± 0.11	-3.7 ± 0.2	-2.0 ± 0.2	-0.91 ± 0.13	-0.55 ± 0.03
Gap existence	14	11	10	10	9
Gap extension	2	1	1	1	1
α_g	1.87 ± 0.03	1.90 ± 0.02	1.07 ± 0.02	0.71 ± 0.02	0.46 ± 0.01
β_g	-36 ± 1	-30 ± 1	-13.5 ± 0.4	-8.9 ± 0.4	-5.3 ± 0.2
$2G(\alpha_u - \alpha_g) + \beta_u$	-37 ± 1	-30.3 ± 0.5	-14.1 ± 0.4	-8.7 ± 0.4	-5.8 ± 0.2

Table 5: The estimation of β using α

13 Figure legends

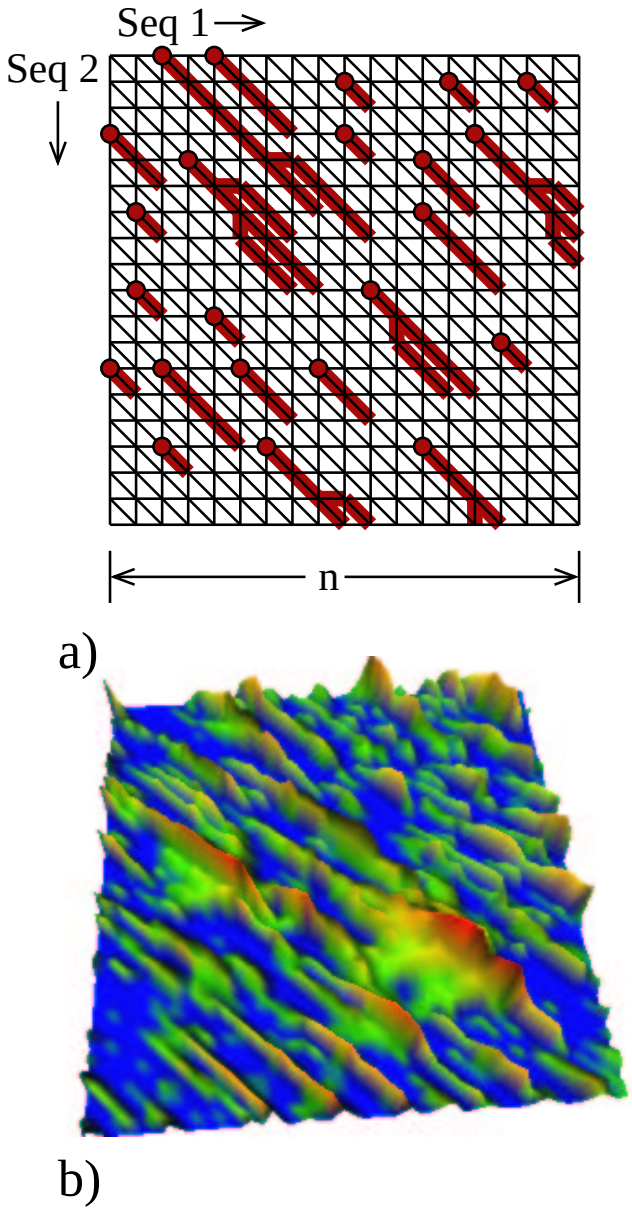

FIGURE 1. Islands on the path graph of Smith-Waterman alignment. (a) is a schematic representation of the path graph. In every cell C the red line recalls the choice made by the optimization procedure of the Smith-Waterman algorithm. By these lines, all the cells with non-zero scores are partitioned into islands according to which anchoring points (circles) they are connected to. (b) shows the score landscape on a 50×50 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.

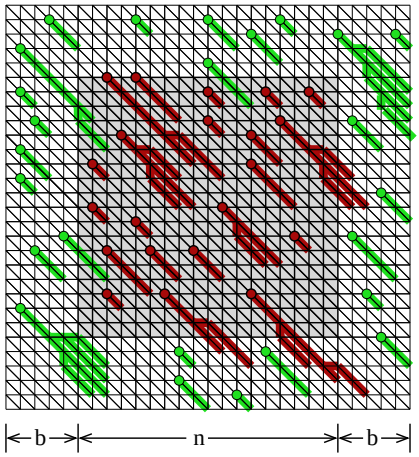

FIGURE 2. Schematic representation of a path graph used to avoid edge effects in the estimation of λ and K via the island method. The $n \times n$ scoring lattice (gray square in the middle) is surrounded by a border of width b. Only islands which are anchored within the central $n \times n$ area (shown 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).

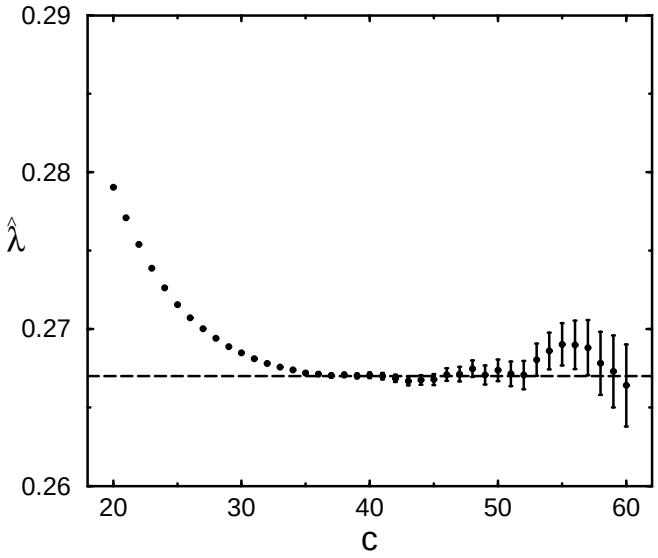

FIGURE 3. Estimates $\hat{\lambda}_c$ obtained via the island method with different cutoffs c. Standard deviations for the estimates are shown with error bars. The plotted horizontal line indicates the best estimate of the asymptotic λ.

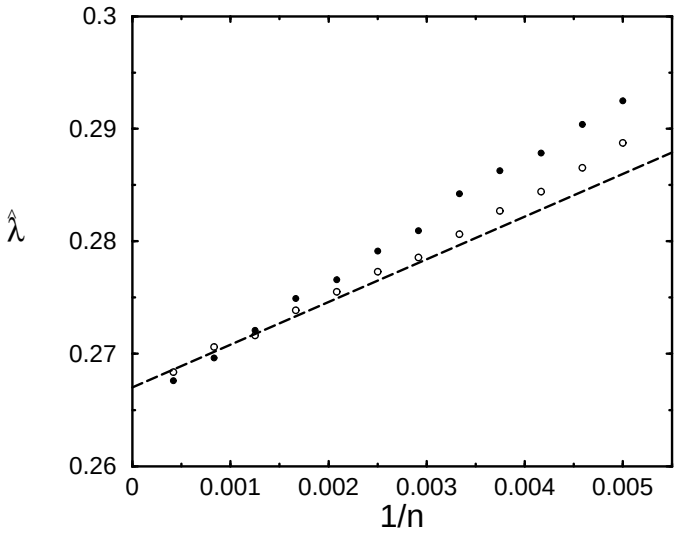

FIGURE 4. Estimates $\hat{\lambda}$ derived from borderless $n \times n$ sequence comparisons by the island method (open circles) and direct method (filled circles), as a function of $1/n$. The size of the symbols equals one standard deviation for the estimates. The plotted line represents the theory of equation [9] for $\tilde{\lambda}(\alpha, \mu)$.

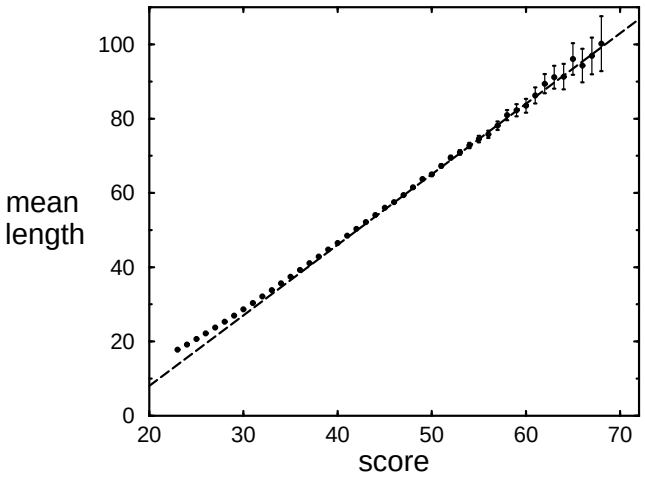

FIGURE 5. The mean length $l(x)$ of optimal island alignments, as a function of the alignment score x. Error bars, representing one standard deviation, grow with score primarily because the number of alignments on which the mean length estimates are based decreases. The plotted line represents a linear regression on the data for scores ≥ 47.