Rapid Assessment of Extremal Statistics for Gapped Local Alignment

Rolf Olsen, Ralf Bundschuh, and Terence Hwa
Department of Physics
University of California at San Diego
La Jolla, CA 92093-0319, U.S.A.
Phone: +1 (619) 534-7256 Fax: +1 (619) 534-7697
e-mail: rolf@cezanne.ucsd.edu, rbund@ucsd.edu, hwa@ucsd.edu

Abstract

The statistical significance of gapped local alignments is characterized by analyzing the extremal statistics of the scores obtained from the alignment of random amino acid sequences. By identifying a complete set of linked clusters, "islands," we devise a method which accurately *predicts* the extremal score statistics by using only one to a few pairwise alignments. The success of our method relies crucially on the link between the statistics of island scores and extremal score statistics. This link is motivated by heuristic arguments, and firmly established by extensive numerical simulations for a variety of scoring parameter settings and sequence lengths. Our approach is several orders of magnitude faster than the widely used shuffling method, since island counting is trivially incorporated into the basic Smith-Waterman alignment algorithm with minimal computational cost, and all islands are counted in a single alignment. The availability of a rapid and accurate significance estimation method gives one the flexibility to fine tune scoring parameters to detect weakly homologous sequences and obtain optimal alignment fidelity.

Keywords: sequence alignment, homology search, statistical significance, extremal statistics

Introduction

Modern molecular biology needs accurate determinations of sequence homology for the identification and classification of proteins, and the reconstruction of phylogenic trees (Waterman 1994; Doolittle 1996). Computationally efficient sequence alignment algorithms have been developed to accomplish this task. These algorithms come in two classes. For database searches, the most commonly used are gapless alignments such as the original BLAST (Altschul *et al.* 1990). More sophisticated is the Smith-Waterman algorithm (Smith and Waterman 1981) which allows for the insertion of gaps. The latter is needed to detect *weakly* homologous sequences (Pearson 1991).

Both alignments with and without gaps are designed to work in the "local alignment" regime, where the alignment scores of unrelated sequences are *typically* very small, so that the occurrence of "unusually" large scores in this regime can be attributed to sequence homology. However, even unrelated sequences can occasionally give large scores in the local alignment regime. Although these events are rare, they become important when one attempts a search over the ever-expanding sequence databases. It is therefore imperative to understand *quantitatively* the statistics of these rare, high-scoring events, in order to estimate the statistical *significance* of a high-scoring alignment.

In the case of gapless alignment, it is known rigorously (Karlin and Altschul 1990, 1993; Karlin and Dembo 1992) that the distribution of alignment scores of random sequences is the Gumbel or extreme value distribution (Gumbel 1958), which has a much broader (i.e., exponential) tail than that of the Gaussian distribution. The Gumbel distribution is specified completely by two constants, whose values can be computed exactly by solving some algebraic equations involving the scoring matrix used. Assuming that the alignments of unrelated biological sequences can be modeled by that of random sequences, one can then specify the probability of observing high-scoring alignments by chance alone, thereby quantifying the *statistical significance* of an alignment.

For the case of gapped alignment, there is no theory available to predict the distribution of alignment scores for random sequences. It has been conjectured (based on ample numerical evidence) that the score distribution is still of the Gumbel form (Smith *et al.* 1985; Collins *et al.* 1988; Mott 1992; Waterman and Vingron 1994a, 1994b; Altschul and Gish 1996). However, estimating the two Gumbel parameters for arbitrary scoring systems has turned out to be a very challenging task. The straightforward method is to generate a background population of alignment scores by repeated alignments of shuffled copies of the two sequences in question. This is enormously time-consuming, as thousands of such shuffles are needed for each set of scoring parameters (i.e., point-substitution matrices and gap costs). Consequently, current generations of gapped

alignments have been restricted to only a few scoring parameter settings for which the background distribution has been precomputed (Altschul and Gish 1996; Altschul *et al.* 1997).

The availability of a significance measure at the preset scoring parameter settings alleviates the problem of false-positives in database searches. However, it does not address the issue of *false-negatives*. More specifically, a statistically insignificant result obtained at the preset scoring parameters does not necessarily mean that the two sequences being aligned are unrelated. Also, a statistically significant "optimal alignment" obtained at the preset parameters does not necessarily mean that the alignment is truly the best possible one. Indeed, the detection of *weak* sequence homology requires careful choice of scoring parameters. This has been investigated empirically by Vingron and Waterman (1994) for sequences whose "true homology" is known. Systematic studies of alignment "fidelity" (i.e., the extent to which sequence homology is retrieved) for alignment of correlated synthetic sequences have also been reported by Drasdo *et al.* (1998a, 1998b) and Olsen *et al.* (1999). For the purpose of detecting weak sequence homology, it is necessary to *scan* in the space of scoring parameters to look for the alignment with the highest statistical significance. This demands a method for *rapid* and *accurate* assessment of gapped alignment statistics, which is the subject of this study.

In an attempt to circumvent the time-consuming shuffling method, Waterman and Vingron (1994a) proposed a "declumping" method, which extracts a list of "clumps" for each pair of random sequences. By fitting the scores of the top several hundred large clumps to Poisson statistics, they were able to estimate the two Gumbel parameters using only 10 shuffled sequences. Unfortunately, the declumping procedure has to proceed clump by clump in order to extract the top scoring clumps, and finding these clumps is rather time consuming, especially if the aligned segment becomes long. For practical purposes, the shuffling method is in fact recommended over the declumping method since it turns out that the declumping procedure takes about the same time per clump as another alignment of the whole lattice (Hardy and Waterman 1997).

In this manuscript, we present an efficient alternative to the shuffling or declumping method: For each pairwise alignment of random sequences, we identify a population of "islands" whose peak scores determine the statistics of the alignment score. These islands are conceptually similar to the clumps of Waterman and Vingron. However, unlike the clumps, the islands can be easily found by adding a few lines to the Smith-Waterman algorithm. Consequently, hundreds and thousands of islands can be scored and extracted in a single alignment with minimal computational effort. This leads to a gain of > 100 times in speed over the shuffling method of significance estimation. Our main assertion, that the statistics of such island scores can be used to *predict* the extremal statistics of the alignment score of unrelated random sequences, is supported by a heuristic argument along with extensive numerical simulations for a variety of score parameter settings and realistic sequence lengths of several hundred amino acids.

The paper is outlined as follows: First we review the theory of extremal statistics, the main analytic results on gapless alignments of random sequences, and the basics of gapped local alignment. We then focus on the alignment of random amino acid sequences. The notion of "island" is introduced; we describe how they can be identified and discuss some of their properties, in particular, their (lack of) mutual correlations. Next, the statistics of the islands are investigated in detail. We demonstrate through extensive numerical simulation that the island score statistics are directly related to the extremal statistics of alignment scores. This empirical fact is then used to predict extremal score statistics based on the island statistics generated from one or a few pairwise alignments. A number of details concerning the theory of extremal statistics, the construction of point-substitution scoring matrices, and a simple algorithm for island counting are relegated to the appendices.

Alignment Statistics

Rare Event Statistics

The general theory of the statistics of rare events is well established (Gumbel 1958; Galambos 1978). Since this theory will be crucial to the method of significance assessment that we will present in this study, we will first review briefly its basic aspects.

Given a set of independent and identically distributed (iid) random variables $x_1, x_2, \ldots$, with a distribution which decays reasonably fast for large values of the x_i's, e.g., $\Pr\{x_i > x\} \propto \exp(-\alpha x^\gamma)$ for $x \to \infty$ and $\alpha, \gamma > 0$, the distribution of the random variables

$$X_n \equiv \max\{x_1, \ldots, x_n\} \tag{1}$$

for large n is known to obey the *Gumbel distribution*

$$\Pr\{X_n > x\} = 1 - \exp(-\kappa e^{-\lambda x}). \tag{2}$$

The distribution (2) is *universal*, in that its *form* does not depend on the specifics of the distribution of the x_i's: It is, first of all, completely independent of the details of the distribution of the x_i's at small values. Moreover, the form of (2) does not depend on the values of the parameters such as α, γ and n. The latter only enter (2) through the values of λ and κ, the only parameters of the Gumbel distribution. Given the asymptotic distribution of x_i, the corresponding Gumbel parameters can be straightforwardly derived, as given explicitly in Appendix A. Especially simple is the case of *asymptotically* Poisson-distributed x_i's, where the expected number of these variables exceeding a certain value x is given by

$$\mathcal{N}(x) = \mathcal{N}_0\, e^{-\alpha x} + \mathcal{N}_1(x), \tag{3}$$

with $\mathcal{N}_1(x) \ll \mathcal{N}(x)$ for $x \to \infty$. In this case, the Gumbel parameters become simply

$$\lambda = \alpha \qquad \text{and} \qquad \kappa = \mathcal{N}_0. \tag{4}$$

Gapless Alignment

A well understood application of the statistics of extreme values to sequence alignment is *gapless alignment* as implemented, e.g., in BLAST (Altschul *et al.* 1990). In this case, it has been rigorously shown (Karlin and Dembo 1992) that the distribution of the maximal score of the alignment of two random sequences is of the Gumbel form (2); furthermore, explicit formulae are given for the two Gumbel parameters. In order to set up the framework we will use to discuss Smith-Waterman alignment with gaps, we start with a non-rigorous review of the treatment of gapless alignment by Karlin and Dembo.

Consider two query amino acid sequences $\mathcal{A} = \{a_1 a_2 \ldots a_N\}$, and $\mathcal{A}' = \{a'_1 a'_2 \ldots a'_{N'}\}$, where a and a' each denotes one of the twenty amino acids, and N, $N' \sim N$ denote the lengths of the sequences. Gapless alignment compares all *consecutive* amino acids $a_i a_{i-1} \ldots a_{i-\ell}$ in a segment of the sequence $\mathcal{A}$ with a segment $a'_j a'_{j-1} \ldots a'_{j-\ell}$ of the sequence $\mathcal{A}'$. The computational task is to find the i, j, and ℓ which give the *highest* total score Σ for a given "scoring matrix" $s_{a,a'}$. The scoring matrix reflects one's prior knowledge of the likelihood of a mutation between the amino acids a and a'; examples are the PAM or BLOSSUM matrices (Dayhoff *et al.* 1978; Henikoff and Henikoff 1992).

The optimization task called for in gapless alignment can be easily accomplished by introducing an auxiliary quantity, $S_{i,j}$, and using the algorithm

$$S_{i,j} = \max\{S_{i-1,j-1} + s_{a_i,a'_j}, 0\}, \tag{5}$$

with the "initial condition" $S_{0,k} = 0 = S_{k,0}$. The quantity $S_{i,j}$ is the optimal score of the above consecutive subsequences (optimized over ℓ); and the global optimal score is obtained as

$$\Sigma = \max_{1 \le i \le N, 1 \le j \le N'} S_{i,j}. \tag{6}$$

In order to evaluate the statistical significance of the resulting Σ, it is necessary to know the distribution of Σ for the gapless alignment of two *random* amino acid sequences, whose elements a_k's are generated independently from the same frequencies p_a as the query sequences, and scored with the same matrix $s_{a,a'}$. For random sequences, one can take $j = i$ in (5) without loss of generality. Eq. (5) then becomes a discrete Langevin equation, with

$$S_{i,i} \equiv S(i) = \max\{S(i-1) + s(i), 0\}, \tag{7}$$

where the "noise" $s(i) \equiv s_{a_i,a'_i}$ is uncorrelated and given by the distribution

$$\Pr\{s_i > s\} = \sum_{\{a,a' | s_{a,a'} > s\}} p_a p_{a'}. \tag{8}$$

From the construction of the scoring matrices (see Appendix B), the noise has the property

$$\sum_{a,a'=1}^{20} p_a p_{a'} s_{a,a'} < 0. \tag{9}$$

The "dynamics" of the evolution equation (7) is qualitatively as follows: The score $S(i)$ starts at zero. If the next local score $s(i+1)$ is negative — which is the more typical case due to the condition (9) — then S remains zero. But if the next local score is positive, then S will increase by that amount. Once it is positive, $S(i)$ performs a "random walk" with independent increments $s(i)$. Due to the condition (9), there is a *negative drift* which forces $S(i)$ to eventually return to zero. After it is reset to zero, the whole process starts over again. The qualitative "temporal" behavior of the score $S(i)$ is depicted in Fig. 1.

Figure 1: Sketch of the total score as a function of sequence position in gapless local alignment.

From the figure, it is clear that the "score landscape" can be divided into a series of *islands* of positive scores, separated by "oceans" where $S = 0$. Each such island originates from a single jump out of the zero-score state and terminates when the zero-score state is reached again. Since each of these islands depends on a different subset of independent random numbers $s(i)$, the islands are *statistically independent* of each other. The same statistical independence applies to the maxima of different islands. Let the maximal score of the k^{th} island be σ_k. Since the global optimal score in (6) can be alternatively written as $\Sigma = \max\{\sigma_1, \sigma_2, \ldots\}$, the distribution of Σ is given by the distribution of the σ_k's through the theory of extremal statistics described above.

Karlin and Dembo (1992) have shown that this island peak score distribution is given by the asymptotic Poisson form (3), with $\alpha = \lambda$ (see (4)) given implicitly by the the unique positive solution of the equation

$$\sum_{a,a'=1}^{20} p_a p_{a'} \exp(\lambda s_{a,a'}) = 1, \tag{10}$$

and $\mathcal{N}_0 = KNN'$, K given by a more complicated function of the scoring matrix. Thus the distribution of Σ can be calculated exactly for gapless alignment for any scoring matrices satisfying (9), making the statistical analysis of gapless alignment results straightforward. For the set of PAM score matrices (Dayhoff *et al.* 1978) used throughout this study, the formula (10)

gives $\lambda = \ln 2$ independent of the PAM distance d. This simple result originates from the fact that PAM scores are log-odd scores; see Appendix B.

Smith-Waterman Alignment

We now turn to the Smith-Waterman alignment algorithm which allows for insertions and deletions (indels). In addition to the scoring matrix $s_{a,a'}$ which we will take to be the PAM matrices parameterized by the PAM distance d, gap penalties need to be provided. For the sake of clarity, we concentrate on the simplest *linear* gap function, which increments the gap cost by δ per length of the gap. Our method easily generalizes to affine gap functions, which include an additional gap initiation cost, as we will demonstrate towards the end.

Alignment Paths and the Dynamic Programming Algorithm. It will be convenient to adopt the directed path representation for sequence alignment (Needleman and Wunsch 1970); an example is shown in Fig. 2 for a specific pair of sequences. In this figure, all the diagonal bonds correspond to gaps. So the score of an alignment path (high-lighted in the figure) gets a contribution $-\delta$ for *each* diagonal bond along that path. The horizontal bonds of the lattice correspond

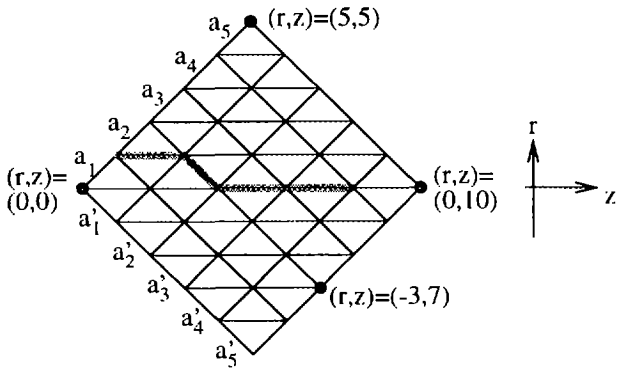

Figure 2: Local alignment of two sequences $a_1a_2a_3a_4a_5$ and $a'_1a'_2a'_3a'_4a'_5$ represented as a directed path on the alignment lattice: the diagonal bonds correspond to gaps in the alignment. The horizontal bonds are amino acid comparisons. The highlighted alignment path $r(\tau)$ therefore corresponds to one possible alignment of subsequences, namely $a_2{-}a_3a_4$ to $a'_1a'_2a'_3a'_4$. This path contains one gap. It is also shown how the coordinates r and z are used to identify the nodes of the lattice.

to pairings of amino acids from the two sequences. As in gapless alignment, such a bond contributes a score s_{a_i,a'_j}. To simplify the notation, we will refer to the nodes and bonds of the lattice via the coordinates r and z as shown in Fig. 2, and use

$$s(r,z) \equiv s_{a_{\frac{r+z+1}{2}},a'_{\frac{z-r+1}{2}}}. \tag{11}$$

With this representation, the task of local gapped alignment is to find the highest scoring path in the lattice for a given set of $\{s(r,z)\}$. The Smith-Waterman algorithm (1981) does this by computing the maximal score $S(r,z)$ of an alignment path ending at the lattice point (r,z) using the dynamic programming scheme

$$S(r,z+1) = \max\left\{\begin{array}{l} S(r+1,z)-\delta \\ S(r-1,z)-\delta \\ S(r,z-1)+s(r,z) \\ 0 \end{array}\right\}, \tag{12}$$

supplemented by the global conditions that $S(0,0) = 0$ and $S(r,z) = -\infty$ beyond the boundaries of the diamond-shaped lattice. The global optimal score is again $\Sigma = \max_{r,z} S(r,z)$.

Gapped Local Alignment. As in the dynamics of gapless alignment (7), the possibility to choose zero in Eq. (12) prevents the score $S(r,z)$ from becoming negative, thereby "filtering out" very unrelated sequences. In order to detect weak similarities, it is also undesirable to have this score grow arbitrarily for unrelated sequences. While the latter is ensured by the condition (9) for gapless alignment, the same condition is not sufficient for gapped alignment. This is because the alignment path has the possibility of gaining large positive scores in $s_{a,a'}$ by using lots of gaps, if the gap cost is sufficiently small. In this case, it is known (Waterman *et al.* 1987; Arratia and Waterman 1994) that the score S (and hence Σ) for the alignment of two random sequences will grow linearly with the sequence length N, unless the gap cost δ exceeds a finite critical value $\delta_c > 0$. The value of δ_c depends strongly on the scoring matrix used and the amino acid frequencies. Approximate analytical expressions for δ_c have been obtained for simple scoring matrices (Bundschuh and Hwa 1999). However, for the PAM matrices used here, the threshold $\delta_c(d)$ is known only numerically. In this study, we will be concerned exclusively with the regime $\delta > \delta_c(d)$ where the score S does not run away for unrelated sequences; this is called the regime of local alignment.

In this regime, the statistics of the optimal score Σ for random sequences is not known theoretically, other than the scaling of its mean,

$$\langle\Sigma\rangle \sim \log N, \tag{13}$$

where $\langle\ldots\rangle$ denotes average over the ensemble of random sequences. Numerically, there is ample evidence (Smith *et al.* 1985; Collins *et al.* 1988; Mott 1992; Waterman and Vingron 1994a, 1994b; Altschul and Gish 1996) that the distribution of Σ is again of the Gumbel form (2). [Note that (13) is compatible with the Gumbel distribution (2), if κ scales as a power of N as what appears to be the case based on the numerics.] These findings are not so surprising given the universality of the Gumbel distribution mentioned in the beginning of this section, even though no proof of this result exists so far. From a practical stand point, a more relevant issue is to devise an efficient way of estimating the parameters λ and κ of the conjectured Gumbel distribution for gapped local alignment. This is what we will address from here on.

Islands in Gapped Local Alignment

Defining the Islands

From the general discussion on extremal statistics given above, it is clear that if the notion of "islands" used in gapless alignment can be properly generalized, then the statistics of Σ can be straightforwardly *derived* from the distribution of the island peaks. More precisely, the notion of an island refers to a specific way of grouping the scores $S(r,z)$ into *non-overlapping* sets, with the requirement that

(a) the maximal scores $\{\sigma_1, \sigma_2, \ldots\}$ of the islands are independent of each other;

(b) $\Sigma = \max\{\sigma_k\}$.

The requirement (b) is easily satisfied if the islands are defined such that each lattice point (r,z) with $S(r,z) > 0$ belongs to an island. The condition (a) is more subtle and lead Waterman and Vingron (1994a, 1994b) to the introduction of their clumps. Below, we suggest a candidate which is conceptually similar but can be obtained much more efficiently. We will give a simple algorithm to find such islands and show empirically that they yield the correct extremal statistics of Σ.

In the local alignment regime, the dynamics of Smith-Waterman alignment is conceptually very similar to the dynamics of gapless alignment: There is a general trend to drive down the score $S(r,z)$, although the fourth alternative in Eq. (12) prevents it from ever becoming negative. By random occurrence of a positive pairing score $s(r_0,z_0)$ at a point (r_0,z_0), $S(r_0,z_0+1)$ can remain positive by the third alternative in Eq. (12), thereby leading to a series of positive scores at $S(r_0,z_0+3), S(r_0,z_0+5)$, etc. It can also lead to positive scores at $S(r_0\pm1,z_0+2), S(r_0\pm2,z_0+4), \ldots$ due to the first and second alternative in Eq. (12). Because of the general tendency towards decreasing scores eventually all positive scores originating from the point (r_0,z_0) will be driven back to zero. We propose each *collection of positive scores originating from the same starting event* to be an "island" for gapped alignment. These islands can be identified by slightly modifying the dynamic programming algorithm (12) and come at virtually no cost in computation time and memory[1]; an example of the necessary algorithm is given in Appendix C.

By our construction, every lattice point of the alignment lattice which has a positive score belongs to exactly one island. In the rare cases where degeneracy arises, i.e., when more than one of the four alternatives in Eq. (12) leads to the same maximum, the implementation in Appendix C provides an explicit hierarchy to island assignment[2]. Thus the islands we defined satisfy the requirement (b) stated above.

Correlation Between Islands

We next examine the mutual correlation of the islands: Recall that each score $S(r,z)$ represents the total score of the optimal path $r^*(\tau)$ ending at the lattice point (r,z). From our definition of the islands, the other end of this optimal path is at the point (r_0,z_0), the initiation point of the island to which the point (r,z) belongs. Thus, an island is a collection of lattice points linked together to their common initiation point respectively by their associated optimal paths. Fig. 3 shows a typical collection of the islands along with the optimal paths linking each point of the island to its initiation point (the solid circles). Note that there is no overlap of the linked clusters, although the clusters can be situated next to each other. Since each score $S(r,z)$ represents the sum of the random pairing scores $s(r^*(\tau),\tau)$ and gap costs *along* its associated optimal path $r^*(\tau)$, it follows that $S(r,z)$ from different islands do not share any common pairing scores in their compositions. Thus, if all the horizontal bonds $s(r,z)$'s are uncorrelated, then we expect the scores S's from different islands (and hence the peak score σ_k of the different islands) to be uncorrelated as well.

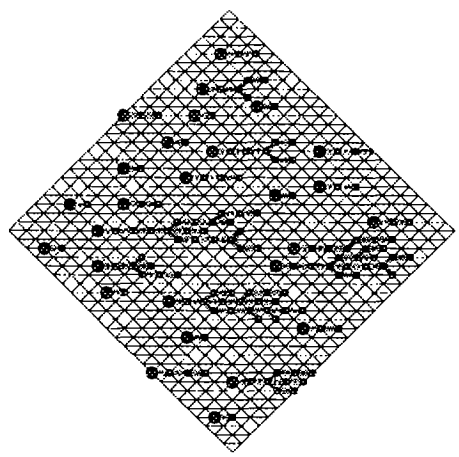

Figure 3: Sketch of some islands on the alignment lattice. The lattice sites with a positive score are marked with small dots. The bonds which have been chosen in the maximization process Eq. (12) are highlighted and together show the optimal path which ends at each point with a positive score. Each of these paths goes back on an island initiation event which is marked by a large dot.

In fact, subtle correlations in $s(r,z)$'s do exist even for random sequences, due to the fact that there are

[1]Note that the minimum memory requirement for the islands scales *linearly* with the length of the sequences being aligned, while the memory requirement for the declumping method scales quadratically with the lengths; see Appendix C.

[2]In our numerics to be reported below, the degeneracies are actually resolved by the use of random numbers. Since these cases are very rare given the variety of different entries in the scoring matrices, the precise way they are resolved should make no difference.

$N \times N'$ different entries of the pairing scores on the alignment lattice, while the two random sequences contain only $N + N'$ elements. In several recent studies (Hwa and Lässig 1996; Drasdo *et al.* 1998a; Bundschuh and Hwa 1999), we have found that this subtle form of correlations was very weak and practically made no difference to a number of statistical properties which one can compute for the case of completely independent s. We do not expect this correlation effect to be important for island statistics either.

To test our assertion of the statistical independence of the islands, we numerically computed the correlation of nearby island peak scores. More precisely, we computed the ratio

$$R = \frac{\langle \sigma\sigma' \rangle - \langle \sigma \rangle^2}{\langle \sigma^2 \rangle - \langle \sigma \rangle^2} \tag{14}$$

where $\langle \sigma \rangle$ and $\langle \sigma^2 \rangle$ are the first and second moment of the distribution of the island peak score σ's, as obtained from the array $\sigma(i)$ defined in the algorithm of Appendix C. $\langle \sigma\sigma' \rangle$ denotes the moment of the *joint* distribution of the scores of one island peak σ with the island peak σ' of its *nearest neighbor* on the alignment lattice. If the scores of nearby island peaks are strongly correlated, then the ratio R should approach 1, while if they are completely uncorrelated, $R = 0$. This ratio was computed for gapped local alignment with $\delta = 2.9$ and PAM-250 matrices; we found $R \approx 0.1$ upon averaging over 300 pairwise alignments. If we consider larger islands, e.g., if we compute R using only those σ's above a threshold value σ_0, then the correlation effects decreases significantly. For example, at $\sigma_0 = 7.5$, the correlation ratio is reduced to $R \approx -0.001$ for the same parameter setting.

Accepting the statistical independence of the islands, we can predict the statistics of Σ from the empirically measured island-peak distribution. In this way, extremal alignment statistics can be obtained from one or a few pairwise alignment of random sequences, while a direct estimate will require an enormous number of pairwise alignments. In the following section, we will compare our predictions with the direct estimates for a range of parameter settings.

Extremal Statistics Prediction

As mentioned in the previous sections, the global optimal score Σ for gapped local alignment of random sequences is expected to be Gumbel distributed. We have also verified this empirically in the numerics below. Moreover, we extract the Gumbel parameters for various scoring parameters, and compare them to the prediction of extremal value theory using the (independently measured) statistics of peak island scores.

Direct Empirical Estimate

For each of the various scoring parameter settings we studied, we ran many (i.e., over one million) pairwise alignments of random amino acid sequences generated according to the known background frequencies p_a. The global optimal score Σ from each alignment was recorded, and a histogram of Σ was constructed for each scoring parameter setting; see Fig. 4. The histogram was then fitted to the Gumbel probability density function, obtained by taking the derivative of the distribution (2). The Gumbel parameters λ and κ were the only free parameters of the fit. The quality of the fit can be seen from an inspection of Fig. 4. Note in particular the good agreement between the data and the fit deep in the exponential tails of the distributions.

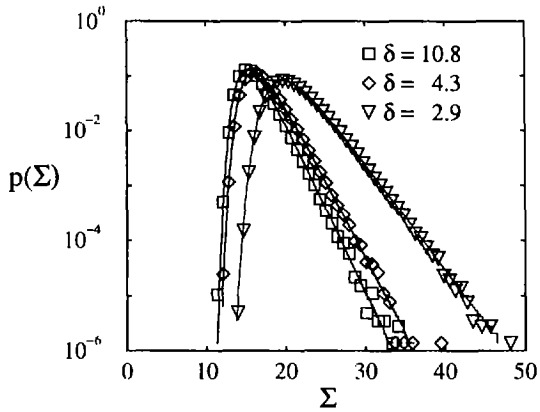

Figure 4: Semi-log plots of the measured probability densities $p(\Sigma)$ for the global optimal score Σ of an alignment with a PAM-250 scoring matrix and a sequence length of $N = 700$. The symbols denote measured values at different gap costs. The solid lines are best fits to the Gumbel probability density $p(\Sigma) = \lambda\kappa \exp(-\lambda\Sigma - \kappa e^{-\lambda\Sigma})$. The values of λ and κ derived from these fits are summarized in Table 1.

These empirical estimates of the Gumbel parameters were carried out using Smith-Waterman alignments with PAM-120 and PAM-250 scoring matrices, at the gap cost[3] of $\delta = \infty$, 10.8, 4.3, and 2.9. The alignment was performed for sequences of lengths $N = N' = 350$ and $N = N' = 700$. Table 1 gives the extracted values of the Gumbel parameters λ and κ for the different scoring parameter settings. The dependence of κ on the sequence length N is expected; however, a weak but noticeable dependence of λ on N is also observed, even for $\delta = \infty$ where the exact result $\lambda = \ln 2$ is expected (see Appendix B). This is an example of the "finite-size effect" resulting from the fact that the finite alignment grid N limits the size of the longest possible island and hence its associated scores. By examining the empirical values of λ for increasing N's, we verified that $\lambda(N)$ indeed converged towards the expected asymptotic value for the case $\delta = \infty$. Similar finite-size effects exist for alignment at finite gap cost. We did not pursue the

[3]This work was done using symmetrized natural log versions of the PAM matrices. All results have been re-scaled to the log base 2 scoring system to be consistent with previous works in the literature. The gap penalties used corresponded to $\delta = \infty$, 7.5, 3, and 2 on the natural log scale.

PAM 120				
	λ		κ	
δ	$N=350$	$N=700$	$N=350$	$N=700$
∞	0.6991	0.6938	20800	78400
10.8	0.6983	0.6943	20500	79200
4.3	0.6458	0.6401	14000	53000
2.9	0.5123	0.5035	5600	20300

PAM 250				
	λ		κ	
δ	$N=350$	$N=700$	$N=350$	$N=700$
∞	0.7075	0.7016	11000	43400
10.8	0.7043	0.6977	10600	40900
4.3	0.6395	0.6286	6500	24200
2.9	0.4758	0.4391	2190	5700

Table 1: Values of the Gumbel distribution parameters extracted by collecting global optimal scores from more than a million alignments. The errors are estimated[4] to be 0.25% for λ and 1.5% for κ.

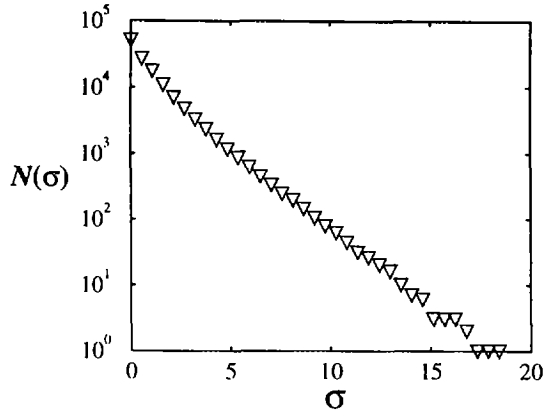

Figure 5: Semi-log plot of the island distribution of a single alignment of two random sequences of length $N = 700$ with a PAM 250 scoring matrix and a gap cost of $\delta = 2.9$.

asymptotic values of λ for these cases, as amino acid sequences are rarely any longer than the length scale we probe here.

Island Peak Distribution

We next examined the statistics of the islands. From the algorithm presented in Appendix C, we obtained in the array $\sigma(i)$ the peak score of each island i after every pairwise alignment. From this array, we recorded the number $\mathcal{N}(\sigma)$ of the score $\sigma(i)$'s exceeding a value σ. The result is shown in Fig. 5 for a representative parameter setting. One sees that $\mathcal{N}(\sigma)$ has an exponential-like tail, although the finite number of islands in a single alignment limits a clear resolution of the tail. To obtain better statistics, we repeated this process for 1000 pairs of random amino acid sequences, and computed the ensemble averaged function $\langle\mathcal{N}(\sigma)\rangle$. The results are shown in Fig. 6 for the different parameter settings.

From Fig. 6, we see that the island peak scores are evidently well described by the asymptotic Poisson statistics (3) for sufficiently large scores (e.g, $\sigma > 8$). The statistics is definitely not Poisson for the smaller scores and can be absorbed into the $\mathcal{N}_1$ term of (3). There, the island peak distributions are dominated by the large number of one-site or few-site islands initiated by the large positive pairing scores appearing in the entries of the scoring matrix $s_{a,a'}$. Thus, the distributions at small σ's largely reflect the distribution of scores as specified by the scoring matrices themselves and are not related to the introduction of gaps[5]. We will refer to the statistics associated with the small islands as "microscopic" statistics.

Given the empirical form of $\mathcal{N}(\sigma)$, we can compute the extremal statistics of the global optimal score Σ, assuming the (numerically verified) independence of the peak island scores. The asymptotic Poisson form shown in Fig. 6 then immediately leads to the Gumbel distribution of Σ, with the two Gumbel parameters λ and κ given by the exponential tails of $\mathcal{N}(\sigma)$ via Eq. (4). Since we already have the values of the Gumbel parameters from the direct empirical estimate of the extremal statistics, we directly plotted $\kappa e^{-\lambda\sigma}$ onto each of the island distributions shown in Fig. 6 for the corresponding parameter settings. We find a remarkable agreement of the slope and amplitude of $\langle\mathcal{N}(\sigma)\rangle$ with these exponentials beyond the small-σ regime for a variety of scoring parameters. Note that this agreement is like glue: it holds even in the presence of non-negligible finite size effects (e.g. the raised values of the λ's of the length 350 sequences relative to those of the length 700 sequences). In this sense, the Gumbel parameters as obtained from the exponential tail of $\mathcal{N}(\sigma)$ are even *more* reliable than the exact asymptotic result of Karlin and Altschul (1990) for gapless alignment ! The latter of course does not provide a finite-size correction, which, however, is relevant if one is comparing short sequences.

The congruence between the asymptotic statistics of the island peaks σ and the global optimal score Σ strongly indicates that it is the high scoring population of the peak island scores that generates the Gumbel distribution of the global optimal scores Σ. This becomes even more remarkable when one notes the score scales

[4]For seven of the parameter settings the 1.1 million high-scores were generated in subsets of 100,000 alignments. λ's and κ's were computed by fits to these subset distributions. We then had seven distributions of λ's and κ's. From each of these we calculated the error in the mean. The error values quoted above are the largest of the seven errors in the mean of λ and κ.

[5]In fact, a moment of reflection reveals that for $\sigma < \delta$, $\mathcal{N}(\sigma)$ must be independent of δ. Therefore, the distribution $\mathcal{N}(\sigma)$ converges to that of the gapless distribution at the small σ end for all δ's. Clearly, no information about gapped alignment can be extracted from the small islands.

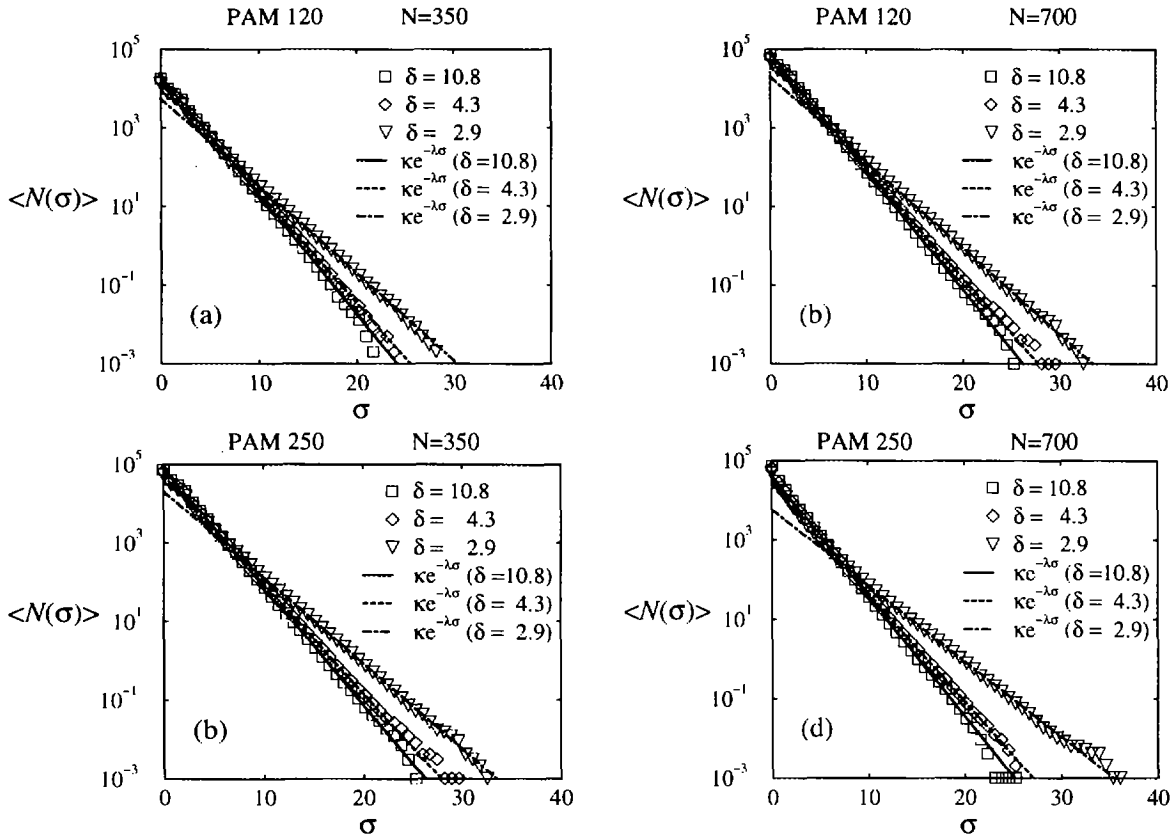

Figure 6: Island distributions for different scoring parameters. The symbols are the measured distributions (squares: $\delta = 10.8$; diamonds: $\delta = 4.3$; triangles: $\delta = 2.9$.) The lines are *not* fits to the data but they denote the Poisson distribution $\kappa e^{-\lambda\sigma}$, using the values of λ and κ from Table 1. The latter are extracted from the distribution of the alignment score Σ's, as obtained directly from over a million pairwise alignments of random sequences. The different plots are for different PAM distances and system sizes: (a) PAM 120, $N = 350$; (b) PAM 120, $N = 700$; (c) PAM 250, $N = 350$; (d) PAM 250, $N = 700$.

for the two asymptotic regimes: Take for example the alignments at $\delta = 2.9$ and PAM-250. The asymptotic statistics of σ for this parameter setting is obtained in the range $10 < \sigma < 30$ (see Fig. 6(d)), while the asymptotics in Σ (Fig. 4) extended to the score value of $40 \sim 50$, which was not yet reached by the island scores collected in Fig. 6(d). Thus, we infer that the Gumbel distribution of Σ is in fact generated by the very large (and rare) islands at $\sigma = 40 \sim 50$, and these large islands are described by the *same* statistical law (namely, the asymptotic Poisson statistics) as the intermediate-sized islands found in Fig. 6(d). The correspondence between the intermediate and large score statistics is certainly not a coincidence. It is a manifestation of the fact that there exists only one *single scale*, the typical score scale λ^{-1}, which governs the score statistics. This single-scale property is apparently a robust statistical property of local gapped alignment; it has been verified in a different context by Hwa and Lässig (1998) and Drasdo *et al.* (1998b), who used scaling theory to relate scores at different parameter settings[6]. It is a very useful property which we will exploit in the next section, in order to *extrapolate* the large-score statistics from the behavior at much smaller scores.

It should be noted here that the correspondence between the peak island score statistics $\mathcal{N}(\sigma)$ and the extremal statistics of Σ does *not* necessarily rely on the asymptotic Poisson form of $\mathcal{N}(\sigma)$. As indicated in Appendix A, the Gumbel parameters can be computed from any reasonably fast decaying function[7] $\mathcal{N}$. How-

[6] In the work of Hwa and Lässig (1998), the quantity which played the role of the typical score λ^{-1} was the saturation score $S_{\rm sat}$.

[7] Since the scores $S(r, z)$ are themselves long-range correlated in z (Hwa and Lässig 1996), the asymptotic distribution of the island peaks may not be Poisson according to recent works of Duffield and O'Connel (1995) and Narayan (1999). This however, would not be observable unless one is aligning very long sequences very close to the boundary of the local alignment regime, i.e., for $\delta \to \delta_c^+$, not relevant for the practical purpose of aligning biological sequences of finite lengths.

ever, since the results we obtained in Fig. 6 are so well described by the asymptotic Poisson form, we will assume this form in the following section.

Rapid Assessment of Extremal Statistics

From the results presented in the previous sections, we have established two empirical findings for gapped local alignment: the statistical independence of the peak island scores and their asymptotic Poisson statistics extending from intermediate to the very large score scales. We have provided heuristic arguments and independent numerical tests supporting both findings, although no mathematical proof is available. In what follows, we shall take these findings as assumptions, and exploit them for rapid assessment of extremal statistics.

We wish to extract the asymptotic island-peak distribution (and hence the Gumbel parameters) using very few pairwise alignments. The difficulty is that $\mathcal{N}(\sigma)$ is limited on the small-σ side by the microscopic statistics and on the large-σ side by the limited number of counts (see Fig. 5). To illustrate what can be achieved, we took the island-peak distribution from a *single* alignment (such as the one shown in Fig. 5), and fitted the exponential form $\kappa e^{-\lambda\sigma}$ to an *intermediate* regime of peak scores, $\sigma_{\text{lower}} < \sigma < \sigma_{\text{uppper}}$. The upper cutoff σ_{upper} was chosen such that $\mathcal{N}(\sigma)$ is above 30 counts, and the lower cutoff σ_{lower} was chosen according to the largest positive entry of the scoring matrix used. Specifically, we chose $\sigma_{\text{lower}} = 1.3 \cdot \max\{s_{a,a'}\}$, which takes on the value of 7 for the PAM-250 matrix and 8 for the PAM-120 matrix.

The results of the extracted values of λ and $\ln\kappa$ from 1 and 10 pairwise alignments are listed in Table 2. Compared to the empirical values (the second column of Table 2) taken from the direct estimate using over one million alignments, we see that a single alignment gives a reasonable prediction of the Gumbel parameters. For example, the statistical uncertainty on the (more important) λ-parameter is $\pm 8\%$, with practically no systematic bias. It can be further reduced if several alignments (of shuffled sequences) are used: for 10 alignments the statistical uncertainty is reduced to approximately $\pm 4\%$ (see column 4 of Table 2). A similar accuracy range was obtained for $\ln\kappa$, which determines the mean of the Gumbel distribution.

We also implemented our island counting method for an affine gap function on the regular diamond shaped scoring lattice. We studied the value of λ for a PAM-250 scoring matrix and a gap cost of $12 + 3k$ for a gap of length k which corresponds to the parameter settings[8] of Waterman and Vingron (1994a, 1994b). The direct empirical estimate using more than a million alignments gives $\lambda = 0.2128$ for sequences pairs of length $N = 300$ and $\lambda = 0.2024$ for $N = 900$. Obtaining the histogram of the island score distribution from a *single* alignment and fitting a Poisson distribution in the region $\sigma_{\text{lower}} < \sigma < \sigma_{\text{upper}}$ as defined previously, we find $\lambda = 0.211 \pm 0.049$ for the length 300 sequences and $\lambda = 0.202 \pm 0.024$ for the length 900 sequences. These result indicate that our method can be reliably extended also to the affine gap function using a single alignment. Computationally, island counting costed approximately 30% extra in time over the bare minimum affine-gap algorithm. Our results agree well with the values obtained by Waterman and Vingron (1994a) using the more time consuming declumping method, on 10 sequence pairs of 300 clumps each. However, we were not able to explicitly compare the time factors of the two methods, since no clump search algorithm was provided there.

PAM 120 λ-values			
δ	direct	1 Al.	10 Al.
∞	0.6938	0.697 ± 0.055	0.693 ± 0.026
10.8	0.6943	0.697 ± 0.055	0.693 ± 0.026
4.3	0.6401	0.645 ± 0.050	0.640 ± 0.025
2.9	0.5035	0.517 ± 0.037	0.509 ± 0.019

PAM 120 $\ln\kappa$-values			
δ	direct	1 Al.	10 Al.
∞	11.27	11.3 ± 0.5	11.3 ± 0.3
10.8	11.28	11.3 ± 0.5	11.3 ± 0.3
4.3	10.88	10.9 ± 0.4	10.9 ± 0.2
2.9	9.92	10.1 ± 0.3	10.0 ± 0.2

PAM 250 λ-values			
δ	direct	1 Al.	10 Al.
∞	0.7016	0.691 ± 0.076	0.697 ± 0.031
10.8	0.6977	0.691 ± 0.076	0.697 ± 0.031
4.3	0.6286	0.628 ± 0.069	0.627 ± 0.027
2.9	0.4391	0.463 ± 0.049	0.448 ± 0.021

PAM 250 $\ln\kappa$-values			
δ	direct	1 Al.	10 Al.
∞	10.68	10.6 ± 0.6	10.6 ± 0.3
10.8	10.62	10.6 ± 0.6	10.6 ± 0.3
4.3	10.09	10.1 ± 0.6	10.1 ± 0.2
2.9	8.65	8.8 ± 0.4	8.8 ± 0.2

Table 2: Average values of the Gumbel distribution parameters extracted by fitting an exponential law to the tail of the island distribution collected from 1 and 10 alignments of sequences of a length $N = 700$.

Summary and Outlook

In this study, we investigated the extremal statistics of local gapped alignments of random amino acid sequences. We identified a complete set of linked clusters — the islands — which can be very efficiently counted with minimal addition to the Smith-Waterman alignment algorithm. We established a firm empirical link

[8] In order to be able to compare directly to Waterman and Vingron's results we here use their scoring system in which the scoring matrix has been rescaled by an extra factor of $10 \log_2 10$.

between the statistics of the peak island scores and the extremal statistics of the alignment score for a variety of scoring parameter settings. The validity of this link hinged upon the statistical independence of the islands, which was supported by heuristic arguments along with direct numerical computation of nearby island correlations. Accepting their independence, the island statistics can be used to *predict* the parameters of the extremal Gumbel distribution. By further extrapolating the intermediate score statistics to the very large score values, one can accomplish the Gumbel parameter prediction by using only a single to a few pairwise alignments. The success for such extrapolation is grounded upon the *scaling properties* of score statistics as discussed by Hwa and Lässig (1996, 1998) and Drasdo *et al.* (1998a, 1998b).

Our method is conceptually similar to the declumping method of Waterman and Vingron (1994a, 1994b), but is faster: In our approach, the collection of island statistics can be directly *incorporated* into the alignment algorithm, resulting in only $\sim$ 30% increase in computational time over the most stripped down version of the Smith-Waterman algorithm. This is an improvement over the declumping method which removes the clumps one by one after the main scoring process is completed, and thus has to find the highest scoring clump on the lattice after each removal of the previous clump. We have not been able to make a direct time comparison with the declumping method. However, it is by orders of magnitude faster than the recommended shuffling method. Additionally, our method can be implemented with a memory requirement which is linear in the length of the sequences while the declumping method inherently requires $O(N^2)$ memory.

We have thus demonstrated the feasibility of a method for the *rapid* and *accurate* assessment of gapped alignment statistics. The availability of such a method will make it possible to find optimal scoring parameters quickly, i.e., on the fly. This capability is central to detecting and enhancing the fidelity of alignments of weakly homologous sequences (Vingron and Waterman 1994; Drasdo *et al.* 1998a, 1998b; Olsen *et al.* 1999). High-fidelity alignment has not been practical so far due to the extraordinary amount of computing resources needed to evaluate the statistical significance. It is hoped that the results reported here will promote the implementation of this general approach to yield alignments of much higher quality than those attainable using the available search tools.

Acknowledgments. The authors have benefited from discussions with S.F. Altschul, M. Gribskov, M. Lässig, and J.D. Moroz. The participation of P. Bermel during the initial stage of this work is much appreciated. RO gratefully acknowledges the support of an LJIS fellowship through the Wellcome-Burroughs Foundation, and RB a Hochschulsonderprogramm III fellowship of the DAAD. TH is supported by a Beckman Young Investigator Award and a Sloan research fellowship.

Appendices

Appendix A: Parameters of the Gumbel Distribution. We wish to compute the parameters of the Gumbel distribution (2) satisfied by the random variable $X_n = \max\{x_1, \ldots, x_n\}$, for independent random variable x_i's. The statistics of the x_i's is given via

$$\mathcal{N}(x) = \mathcal{N}_0 x^\beta \exp[-\alpha x^\gamma] + \mathcal{N}_1(x), \qquad (15)$$

which is the expected number of the x_i's exceeding some value x. Here $\alpha, \gamma > 0$ and β are arbitrary constants, and $\mathcal{N}_1(x)$ is assumed negligible compared to the first term for large enough x. The parameters of the Gumbel distribution of the X_i's can be calculated using the scheme presented by Galambos (1978). Applied to the distribution (15) these parameters depend on the solution y of the equation

$$\mathcal{N}_0 y^\beta \exp[-\alpha y^\gamma] = 1, \qquad (16)$$

which is easy to find numerically once the parameters α, β, γ, and $\mathcal{N}_0$ are given. The parameters of the Gumbel distribution can then be calculated as $\lambda = \alpha\gamma y^{\gamma-1}$ and $\kappa = \exp[\alpha\gamma y^\gamma]$.

In the special case without a power law prefactor, i.e., $\beta = 0$, Eq. (16) can be solved for y and we get

$$\lambda = \gamma\alpha^{1/\gamma}(\ln \mathcal{N}_0)^{1-\frac{1}{\gamma}} \quad \text{and} \quad \kappa = \mathcal{N}_0^\gamma, \qquad (17)$$

which, for $\gamma = 1$, reproduces the result (4) quoted for the asymptotic Poisson distribution.

Appendix B: PAM Matrices and Gapless Alignment. In this appendix, we review the properties of PAM matrices which we used as the scoring matrices for amino acid comparison. PAM (percent accepted mutations) matrices are constructed assuming that all the amino acids in a protein sequence mutate independently from their neighbors. The mutation is described by a 20×20 transition matrix $t_{a,a'}$ which gives the probability to find a mutation to amino acid a given that the original amino acid was a'. This PAM transition matrix is normalized such that each application of it leads to a change in an amino acid in 1% of the cases. Substitution rates exceeding one percent can be generated by repeated application of the transition matrix. These powers t^d of the basic transition matrix t are transition matrices $q_{a,a'}$ themselves, i.e., they have the interpretation of probabilities to find amino acid a given that the original amino acid was a'.

By definition of relative probabilities any such transition matrix fulfills the conditions

$$\sum_{a=1}^{20} q_{a,a'} = 1 \quad \text{and} \quad q_{a,a'} \cdot p_{a'} = q_{a',a} \cdot p_a \qquad (18)$$

where p_a are the amino acid frequencies.

The log-odds *scoring matrix* connected to a given transition matrix $q_{a,a'}$ is defined by

$$s_{a,a'} = \log_2\left[\frac{q_{a,a'}}{p_a}\right]. \qquad (19)$$

Note that the scoring matrix s is symmetric due to the second relation in (18). The symmetry in s is necessary since in scoring, one doesn't usually know which of the amino acids a and a' is the ancestor. We get the scoring matrix at a PAM distance d by applying this general scheme to the special transition matrix $p = t^d$.

If one applies the condition (10) for the solution of the Gumbel parameter λ to the above-defined scoring matrix, one finds the identity

$$Q(\lambda) \equiv \sum_{a,a'=1}^{20} p_a p_{a'} \exp[\lambda s_{a,a'}] = \sum_{a,a'=1}^{20} p_a p_{a'} \left(\frac{q_{a,a'}}{p_a}\right)^{\frac{\lambda}{\ln 2}}$$

due to the second condition in (18). Further applying the first property in (18), one finds that $Q(\lambda) = 1$ for $\lambda = \ln 2$, which is the unique positive solution of Eq. (10). Note that the solution is *independent* of the choice of the transition matrix and thus, especially of the PAM distance d. This result underscores the "naturalness" of log-odd scores in the context of gapless alignment, the meaning of which has been explored in detail by Altschul (1991). Note also, that the condition (9) is automatically fulfilled for an arbitrary log-odds scoring matrix, since the relative entropies $-\sum_{a=1}^{20} p_a \log_2[q_{a,a'}/p_a]$ of the distributions p_a and $q_{a,a'}$ for fixed a' on the right hand side of the relation

$$\sum_{a,a'=1}^{20} p_a p_{a'} s_{a,a'} = \sum_{a'=1}^{20} p_{a'} \sum_{a=1}^{20} p_a \log_2 [q_{a,a'}/p_a]$$

cannot be negative.

Appendix C: Algorithm for Island Assignment. The assignment of the different points of the alignment lattice to the different islands can be included very straightforwardly into the dynamic programming algorithm (12) with minimal computational effort. We introduce an additional array $I(r,z)$ which, upon completion of the algorithm, will contain the "island number" the lattice point (r,z) belongs to. We also need an array $\sigma(i)$ which holds the maximum score of the i^{th} island. The basic dynamic programming algorithm is expanded as follows:

```
number_of_islands:=0
loop over all z
  loop over all r
    maximum:=0.0
    island:=0
    if S(r+1,z) − δ>maximum then
      maximum:=S(r+1,z) − δ
      island:=I(r+1,z)
    end if
    if S(r−1,z) − δ>maximum then
      maximum:=S(r−1,z) − δ
      island:=I(r−1,z)
    end if
    if S(r,z−1) + s(r,z)>maximum then
      maximum:=S(r,z−1) + s(r,z)
      if S(r,z−1) = 0 then
        number_of_islands:=number_of_islands+1
        island:=number_of_islands
        σ(island):=maximum
      else
        island:=I(r,z−1)
        if maximum>σ(island) then
          σ(island):=maximum
        end if
      end if
    end if
    S(r,z+1):=maximum
    I(r,z+1):=island
  end loop
end loop
```

The lines not set in bold face constitute the usual dynamic programming algorithm (12) for linear gap cost. In order to calculate the island peak distribution, we only need to insert those lines in bold face. Their function is to assign the island number to the point $(r, z+1)$. This island number depends on which alternative of Eq. (12) gives the maximum. If the last alternative of (12) is chosen and the score is set to zero, then there is no island at $(r, z+1)$ and the island number is set to zero. Otherwise, the path from one of the points $(r+1, z)$, $(r-1, z)$, or $(r, z-1)$ is responsible for the maximum in Eq. (12). In this case, the island number of the point the path comes from is assigned to $(r, z+1)$, or if it is an island initiation event, a new island number is generated and assigned. If the maximum score of the current island might have changed, it has to be updated, too.

Note that the arrays $S(r,z)$ and $I(r,z)$ as presented here for notational simplicity are of the size $N \cdot N'$. For the purpose of constructing a histogram of the island score distribution, it is sufficient to keep track of only the "current" configuration $S(r)$, island number $I(r)$, and the maximum score σ of the "currently" active islands at each "time" step t. This renders the memory requirement *linear* in the length of the sequences. Note also that this algorithm can be easily extended to the case of affine gap cost. We found that the inclusion of island counting and island statistics resulted in a net increase of approximately 30% compared to the bare minimum Smith-Waterman algorithm with affine gap function.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403–410.

Altschul, S.F. 1991. Substitution Matrices from an Information Theoretic Perspective. *J. Mol. Biol.* 119:555–565.

Altschul, S.F. and Gish, W. 1996. Local Alignment Statistics. *Methods in Enzymology* 266:460–480.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997.

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389–3402.

Arratia, R. and Waterman, M.S. 1994. A Phase Transition for the Score in Matching Random Sequences Allowing Deletions. *Ann. Appl. Prob.* 4:200-225.

Bundschuh, R. and Hwa, T. 1999. An Analytic Study of the Phase Transition Line in Local Sequence Alignment with Gaps. To appear in *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, Israil S., *et al.*, eds, New York, NY: ACM press.

Collins, J.F., Coulson, A.F.W., and Lyall, A. 1988. The significance of protein sequence similarities. it CABIOS 4:67–71.

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. 1978. A Model of Evolutionary Change in Proteins. In *Atlas of Protein Sequence and Structure*, Dayhoff M.O. and Eck, R.V., eds., 5 supp. 3:345 358, Natl. Biomed. Res. Found.

Doolittle, R.F. 1996. *Methods in Enzymology* 266. San Diego, Calif.: Academic Press.

Drasdo, D., Hwa, T., and Lässig, M. 1998a. Scaling laws and similarity detection in sequence alignment with gaps. submitted to J. Comp. Biol. (E-print: physics/9802023).

Drasdo, D., Hwa, T., and Lässig, M. 1998b. A statistical theory of sequence alignment with gaps. In *Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology*, Glasgow, J., *et al.* eds, 52-58. Menlo Park, Calif.: AAAI Press.

Duffield, N.G. and O'Connel, N. 1995. Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Camb. Phil. Soc.* 118:363–374.

Galambos, J. 1978. *The Asymptotic Theory of Extreme Order Statistics.* New York, NY: John Wiley & Sons.

Gumbel, E.J. 1958. *Statistics of Extremes.* New York, NY: Columbia University Press.

Hardy, P. and Waterman, M.S. 1997. *The Sequence Alignment Software Library at USC.* From `http://www-hto.usc.edu/software/`.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 89:10915–10919.

Hwa, T. and Lässig, M. 1996. Similarity detection and localization. *Phys. Rev. Lett.* 76:2591–2594.

Hwa, T., and Lässig, M. 1998. Optimal detection of sequence similarity by local alignment. In *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, Israil S., *et al.*, eds, 109–116. New York, NY: ACM Press.

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.

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

Karlin, S., and Altschul, S.F. 1993. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA* 90:5873–5877.

McClure, M.A., Vasi, T.K., and Fitch, W.M. 1994. Comparative analysis of multiple protein-sequence alignment methods. *Molecular Biology and Evolution* 11:571–592.

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

Narayan, O. 1999. *Exact asymptotic queue length distribution for fractional Brownian traffic.* Forthcoming.

Needleman, S.B., and Wunsch, C.D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443–453.

Olsen, R., Hwa, T., and Lässig, M. 1999. Optimizing Smith-Waterman Alignments. *Pacific Symposium on Biocomputing* 4:302-313.

Pearson, W.R. 1991. Searching Protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11:635-650.

Smith, T.F., and Waterman, M.S. 1981. Comparison of biosequences. *Adv. Appl. Math.* 2:482–489.

Smith, T.F., Waterman, M.S., and Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* 13:645-656.

Vingron, M. and Waterman, M.S. 1994. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Bio.* 235:1 12.

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

Waterman, M.S. 1994. *Introduction to Computational Biology.* London, UK: Chapman & Hall.

Waterman, M.S. and Vingron, M. 1994a. Sequence Comparison Significance and Poisson Approximation. *Stat. Sci.* 9:367-381.

Waterman, M.S. and Vingron, M. 1994b. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. U.S.A.* 91:4625–4628.