Statistical Significance and Extremal Ensemble of Gapped Local Hybrid Alignment

Yi-Kuo Yu*, **Ralf Bundschuh**†, **and Terence Hwa**†

* Department of Physics, Florida Atlantic University, Boca Raton, FL 33431-0991

† Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319

Abstract

A "semi-probabilistic" alignment algorithm which combines ideas from Smith-Waterman and probabilistic alignment is proposed and studied in detail. It is predicted that the score statistics of this "hybrid" algorithm is of the universal Gumbel form, with the key Gumbel parameter λ taking on a *fixed* asymptotic value for a wide variety of scoring parameters. We have also characterized the "extremal ensemble", i.e., the collection of sequence pairs exhibiting similarities that a given scoring system is most sensitive to. Based on this extremal ensemble, a simple recipe for the computation of the "relative entropy", and from it the correction to λ due to finite sequence length is also given. This allows us to assign p-values to the alignment results for arbitrary scoring parameters and gap costs. The predictions compare well with direct numerical simulations for a broad range of sequence lengths with various choices of the substitution scores and affine gap parameters.

Key words: sequence alignment; statistical significance; maximum likelihood; hidden Markov model

Introduction

Computer-assisted sequence comparison tools such as BLAST (Altschul *et al.*, 1990) and FASTA (Pearson, 1988) have become an integral part of modern molecular biology. They reveal evolutionary relationships between protein sequences and therefore provide a basis for the functional identification of new genes and for the construction of phylogenic trees. Two types of algorithms have been used: those which search for the *optimal* alignment (as exemplified by the algorithm of Smith and Waterman (1981)), and those which identify *likely* alignments (as exemplified by the hidden-Markov model (HMM) based "Sequence Alignment Modules" (Hughey & Krogh, 1996)). In each case, the quality of the alignment is summarized by an alignment score S; the latter is typically taken to be the logarithm of the total likelihood in the probabilistic approaches. However, such an alignment score is assigned to *any* pair of sequences, also to biologically completely unrelated ones (e.g., to pairs of random sequences.) In order to be able to distinguish *true evolutionary relationships* from *random similarities* it is an important goal common to all algorithms to understand the probability distribution function $\mathrm{pdf}(\mathsf{S})$ of the score S for the appropriate null models. The knowledge of this distribution gives the possibility of assigning p-values, i.e., the probabilities that a high score could have arisen by chance, to alignment results. These p-values quantify the amount of surprise behind a given alignment score. Only if a score is sufficiently surprising, the underlying alignment is considered to be of evolutionary origin.

Rigorous results on such background statistics are known only for the gapless alignment, whose score distribution follows the so-called Gumbel form (Gumbel, 1958),

$$\mathrm{pdf}(\mathsf{S}) = KMN\lambda \exp\left[-\lambda\mathsf{S} - KMNe^{-\lambda\mathsf{S}}\right], \quad (1)$$

for long sequence lengths M and N (Arratia *et al.*, 1988; Karlin & Altschul, 1990, 1993; Karlin & Dembo, 1992). Explicit formulae relating the hundreds of alignment parameters to the two Gumbel parameters λ and K are available (Karlin & Altschul, 1990). For gapped sequence alignment with large enough gap cost, the score distribution is also empirically known to obey Gumbel statistics (Smith *et al.*, 1985; Collins *et al.*, 1988; Mott, 1992; Waterman & Vingron, 1994a, 1994b; Altschul & Gish, 1996, Olsen *et al.*, 1999). However, the dependence of the two Gumbel parameters on the hundreds of scoring parameters is generally so complicated that it is very difficult to determine the Gumbel parameters in an efficient enough manner to render them useful.

This problem is partially overcome in (gapped) BLAST by pre-computing the null statistics for a fixed set of scoring parameters. However, this is a severe restriction on the flexibility of the method and leads, e.g., to wrong predictions of p-values for query sequences with unusual amino acid compositions. Even more importantly, the restriction to a small set of scoring parameters for which the null statistics is pre-computed becomes prohibitive for

[1] related (p)re-prints available at `http://matisse.ucsd.edu/~hwa/pub.html`.

the use of *position-specific* scoring functions (Henikoff & Henikoff, 1994) as they are needed for detailed modeling of protein families, folds, etc. Because of this problem the iterative similarity search algorithm PSI-BLAST (Altschul *et al.*, 1997) is currently limited to *uniform* gap costs which is an unfortunate drawback.

Position-specific scoring systems are naturally incorporated in probabilistic (e.g., the HMM-based) alignment algorithms. However, only very little is known about the statistics of the log-likelihood score, even at the empirical level.

In this paper, we describe a "semi-probabilistic" alignment algorithm which is a *hybrid* of the Smith-Waterman and the probabilistic alignment algorithms. Our hybrid algorithm has the same computational complexity as the Smith-Waterman and the probabilistic algorithms, with computation time scaling as $O(M \cdot N)$; also, its sensitivity in detecting sequence homology is comparable to or better than the existing algorithms (Yu & Hwa, 1999). The key advantage of the hybrid algorithm is that its score statistics can be characterized theoretically. Moreover, the ensemble of rare sequence pairs responsible for the high-scoring events can be characterized. This "extremal ensemble" consists of sequence pairs exhibiting similarities that a given scoring system is most sensitive to. The knowledge of the connection between scoring systems and their extremal ensembles is very useful in *constructing* the optimal scoring parameters for a given model of sequence evolution. This is analogous to how the Karlin-Altschul theory of gapless alignment can be used to guide the selection of the appropriate amino-acid substitution score (Karlin & Altschul, 1990). In the following sections, we will first describe the algorithm followed by characterization of score distribution and extremal ensemble. A numerical test of the prediction of the theory will be presented at the end. More technical issues are relegated to the appendices.

Algorithms

Consider two sequences $\mathbf{a} = [a_1, a_2, ..., a_M]$ and $\mathbf{b} = [b_1, b_2, ..., b_N]$ of lengths M and N, with elements a_i and b_j taken from a finite character set χ. We will employ a frequently used null model with independently identically distributed letters where the probability of having two sequences is given by the distribution function

$$P_0[\mathbf{a}, \mathbf{b}] = \prod_{\substack{1 \le m \le M \\ 1 \le n \le N}} p(a_m) \cdot p(b_n), \qquad (2)$$

with $p(a)$ being the background frequency of the element a, and $\sum_{a \in \chi} p(a) = 1$.

Probabilistic Global Alignment

We first review probabilistic global alignment of the sequences **a** and **b**. We adopt the approach of Bishop and Thompson (1986), evaluating probabilistic global alignment as the likelihood of observing the sequences **a** and **b** given a fictitious evolution model producing pairs of *related* sequences **a** and **b**. For the purpose of illustration, let us consider the following simple version of this evolution model: Start with empty sequences **a** and **b** and go through the hidden Markov model illustrated in Fig. 1.

- Until one of the sequences reaches the desired length N, there is a probability ν for a "deletion step" and the same probability ν for an "insertion step".
  - if the "insertion mode" is selected, generate a new element a according to the background frequencies $p(a)$ and append it to sequence **a**.
  - if the "deletion mode" is selected, generate a new element b according to the background frequencies $p(b)$ and append it to sequence **b**.
  - if neither deletion nor insertion is selected, generate a *pair of elements* (a, b) according to some *joint probability distribution* $\mathcal{P}(a, b)$[1] and append a to sequence **a** and b to sequence **b**.
- If one of the sequences reaches the desired length N generate random elements according to the background frequencies $p(a)$ and append them to the shorter of the two sequences until they both have the length N.

The "weight" $W[\mathbf{a}, \mathbf{b}]$ for a specific random sequence **a** mutating into a sequence **b** can be computed iteratively (Bishop & Thompson, 1986) by introducing an auxiliary variable $\mathcal{W}_{i,j}$:

$$\begin{aligned} \mathcal{W}_{i,j} &= w(a_i, b_j) \cdot \mathcal{W}_{i-1,j-1} \\ &\quad + \nu \cdot [\mathcal{W}_{i-1,j} + \mathcal{W}_{i,j-1}], \end{aligned} \qquad (3)$$

with

$$w(a, b) = (1 - 2\nu) \frac{\mathcal{P}(a, b)}{p(a)p(b)} \qquad (4)$$

being the net substitution probability. W is then obtained as

$$\begin{aligned} W[\mathbf{a}, \mathbf{b}] &= [2\nu + w(a_M, b_N)] \cdot \mathcal{W}_{M-1,N-1} \\ &\quad + \sum_{i=1}^{M-1} [\nu + w(a_i, b_N)] \cdot \mathcal{W}_{i-1,N-1} \\ &\quad + \sum_{j=1}^{N-1} [\nu + w(a_M, b_j)] \cdot \mathcal{W}_{M-1,j-1}. \end{aligned} \qquad (5)$$

[1] The joint probability distribution $\mathcal{P}(a, b)$ is often chosen as $\mathcal{T}(b|a)p(a)$ where $\mathcal{T}(b|a)$ is the *transition probability* for a mutation from element a into element b.

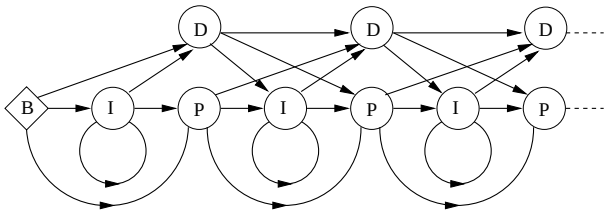

Figure 1: Schematics of the hidden Markov model $\mathcal{M}$ for sequence evolution. The different states are B for the "begin" state, I for the "insertion" states, D for the "deletion" states, and P for the "pair emission" state. The arrows indicate the allowed transitions between the states, with transition probabilities as given in the text. Sequence elements are "emitted" according to the following rules: An element a is emitted into sequence $\mathbf{a}$ with probability $p(a)$ every time the state I is visited, and an element b is emitted into sequence $\mathbf{b}$ with probability $p(b)$ every time the state D is visited. In state P a pair (a, b) is emitted according to some joint distribution $\mathcal{P}(a, b)$ and elements a and b are appended to sequences $\mathbf{a}$ and $\mathbf{b}$ respectively.

where $\mathcal{W}_{1\le m\le M,1\le n\le N}$ is obtained by iterating the recursion relation (3) from the initial conditions

$$\mathcal{W}_{i\ge 0,j=0} = \nu^i \qquad 1 \le i \le M \text{ and} \tag{6}$$

$$\mathcal{W}_{i=0,j\ge 0} = \nu^j \qquad 1 \le j \le N. \tag{7}$$

The recursion relation (3) is nothing but the probabilistic version of the Needleman-Wunsch global alignment algorithm (Needleman & Wunsch, 1970) with linear gap cost. Alternatively, the Needleman-Wunsch algorithm is just the Viterbi version of Eq. (3). In the context of alignment, ν controls the gap penalty and $w(a, b)$ the substitution cost. Note that due to the condition (4), we have

$$\sum_{[\mathbf{a},\mathbf{b}]} W[\mathbf{a},\mathbf{b}] \cdot P_0[\mathbf{a},\mathbf{b}] = 1, \tag{8}$$

which is nothing but the statement of probability conservation for the different ways sequences can be mutated into each other. Note that $W[\mathbf{a},\mathbf{b}] \cdot P_0[\mathbf{a},\mathbf{b}]$ is the likelihood of generating the sequence pair by the mutation model (Thorne *et al.*, 1991, 1992).

Probabilistic Local Alignment

Local alignment identifies subsequences, e.g., $\hat{\mathbf{a}} = [a_{m'}, a_{m'+1}, ..., a_m]$ and $\hat{\mathbf{b}} = [b_{n'}, b_{n'+1}, ..., b_n]$ with $1 \le m' \le m \le M$ and $1 \le n' \le n \le N$, whose mutual global alignment score $S(m', n'; m, n)$ is the highest, especially in cases where the global alignment of the complete sequences yields negative total scores, i.e., $S(1, 1; M, N) < 0$. Instead of directly optimizing over the four variables m', n', m, and n, the Smith-Waterman algorithm (Smith & Waterman, 1981) proceeds by first computing an auxiliary score $H(m, n) = \max_{m',n'} S(m', n'; m, n)$ by slightly modifying the Needleman-Wunsch algorithm. Then it maximizes over the H's. This procedure maintains the computational complexity at $O(M \cdot N)$ as in global alignment.

The same strategy can be adopted in probabilistic local alignment if in order to be really looking for local similarities the scoring parameters are chosen such that the total weight of global alignment is small, i.e., such that $\ln \mathcal{W}_{M,N} < 0$.

Let $\mathcal{W}_{m',n';m,n}$ be the likelihood of the global probabilistic alignment of the subsequences $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. It is computed by applying the recursion Eq. (5) with the intial conditions Eqs. (6) and (7) to the sequences $[a_{m'}, a_{m'+1}, ..., a_m]$ and $[b_{n'}, b_{n'+1}, ..., b_n]$. The likelihood is then given by $\mathcal{W}_{m',n';m,n} \equiv \mathcal{W}_{m-m'+1,n-n'+1}$.

Then, we introduce an auxiliary variable

$$Z_{m,n} \equiv 1 + \sum_{m'=1}^{m} \nu^{m'} + \sum_{n'=1}^{n} \nu^{n'} + \sum_{\substack{1\le m'\le m\\ 1\le n'\le n}} \mathcal{W}_{m',n';m,n}, \tag{9}$$

which is the total weight of all subsequence pairs including (a_m, b_n), plus the null alignments. $Z_{m,n}$ can be computed according to the probabilistic version of the Smith-Waterman algorithm,

$$Z_{i,j} = 1 + w(a_i, b_j) \cdot Z_{i-1,j-1} + \nu \cdot [Z_{i-1,j} + Z_{i,j-1}], \tag{10}$$

with the boundary conditions $Z_{0,j} = Z_{i,0} = 1$, for $1 \le i \le m$ and $1 \le j \le n$. Alternatively, the Smith-Waterman algorithm can be viewed as the Viterbi version of (10).

The total weight $\mathcal{Z}$ characterizing the fully probabilistic alignment is obtained as the sum of the Z's, i.e., as

$$\mathcal{Z} = \sum_{\substack{1\le m\le M\\ 1\le n\le N}} Z_{m,n}. \tag{11}$$

However, even the shape of the distribution of the fully probabilistic local alignment score $\ln \mathcal{Z}$ is not very well understood rendering the calculation of p-values for this score very hard. To overcome this difficulty, we introduce a maximum-log-likelihood (MLL) score

$$\mathsf{S} = \max_{\substack{1\le m\le M\\ 1\le n\le N}} \ln Z_{m,n} \tag{12}$$

to characterize the quality of the alignment. Eqs. (10) and (12) define the hybrid algorithm which we will focus on from here on.

Statistics of Hybrid Alignment

In this section we will show that the score distribution of hybrid alignment is a Gumbel distribution with $\lambda = 1$ independent of the scoring system. We will moreover characterize the extremal ensemble of the algorithm, i.e., the sequence pairs exhibiting similarities that a given scoring system most sensitive to. This knowledge about the extremal ensemble will help us to characterize the deviation of the Gumbel parameters due to the finite sequence length.

Score Landscape and Islands

For the Smith-Waterman algorithm, Olsen *et al.* (1999) utilized the "score landscape" $H(m,n)$ to characterize the tail of the Gumbel distribution. The landscape consists of a collection of *essentially uncorrelated* positive-scoring "islands", separated by a "sea" at $H = 0$. The peak scores of the islands are found to follow Poisson statistics, from which the Gumbel parameters λ and K can be directly derived. Olsen *et al.*'s study indicates clearly that the key to understanding the Gumbel distribution is to characterize the probability tail of obtaining a *single* large island.

Due to our definition (12) of the MLL score as a maximum, the distribution of S is again expected to be of the Gumbel form. (This is not true for the score $\ln \mathcal{Z}$ of the fully probabilistic local alignment, since the score $\mathcal{Z}$ is defined in Eq. (11) via the *sum* rather than the *max* operation.) Similar to what was found by Olsen *et al.* (1999), the pertinent score landscape $\ln Z_{m,n}$ for hybrid alignment consists of islands which are *essentially uncorrelated*. Instead of being separated by the "sea" of zero scores, our MLL islands do have some minor positive score background in between. Since we are interested in the high scoring islands, the minor positive-score background does not affect the identification of the high-scoring islands, see Fig. 2 for example. Since the MLL score is the maximum of many of these uncorrelated island peak scores, the statistics of the MLL score (i.e., the Gumbel parameters) can be deduced if the statistics

$$G(h) \equiv \Pr\{\text{peak island score} > h\}$$

of the individual island peak scores is known.

Thus, it is our goal to calculate this peak score distribution. We will do so in several steps. First, we compute the auxiliary quantity

$$D(h|L) \equiv \sum_{\{\mathbf{a},\mathbf{b}\}} \delta(h - \ln W_L[\mathbf{a},\mathbf{b}]) \cdot P_0[\mathbf{a},\mathbf{b}], \quad (13)$$

that a *global* probabilistic alignment of two sequences of length L will have the score about h. Then, we will relate this distribution to the island peak score distribution. Basically, we will establish that high-scoring global alignments which contribute to $D(h|L)$

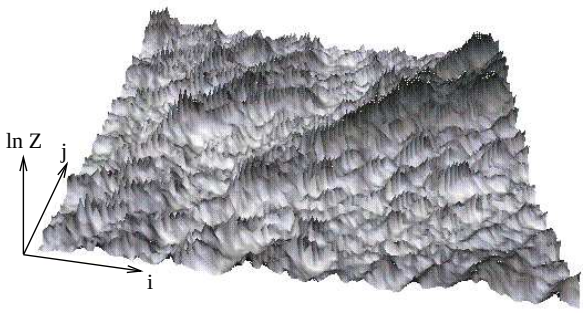

Figure 2: The $\ln Z$ landscape from aligning two random sequences. This figure is a projection of a three dimensional plot. One sequence is laid along the i direction while the other is laid along the j direction. The MLL score $\ln Z_{i,j}$ is then plotted along the third direction labeled by $\ln Z$. The gray scale is used in such a way that the larger the MLL score, the darker the point $(i, j, \ln Z)$. As shown in this figures, a sea of small ripples separate one medium-sized island from a less significant one.

correspond to high-scoring islands which contribute to $G(h)$. We find that

$$G(h) \sim e^{-\lambda h} \quad (14)$$

with $\lambda = 1$. Since S is the maximum of a large number of independent random island peak scores obeying the distribution (14), it has Gumbel statistics (Gumbel, 1958) with the same λ. This strategy of computing local alignment score statistics using the statistics of *global* alignment has been examined in detail in the context of Smith-Waterman alignments and applied to the special scoring system corresponding to the Longest Common Subsequence problem (Bundschuh, 2000).

Large-score Statistics

The first step in the derivation outlined above is very important since it does not only give us an expreession for the auxiliary quantity $D(h|L)$. It also gives us information on the *extremal ensemble*, i.e., on the *typical* sequence pairs that lead to high-scoring global alignments and thus high-scoring islands. Therefore, we devote this whole section to this computation.

We start by characterizing the distribution $D(h|L)$ for $h \gg 1$ by a simple maximization principle. Instead of computing $D(h|L)$ directly, let us first consider a different (but related) quantity, the probability $\mathcal{D}$ that the sum of the score $\ln W_\ell$ from $\mathcal{N} \gg 1$ independent global alignments of random sequences of length ℓ is $\mathcal{H}$, i.e.,

$$\mathcal{D}_\ell(\mathcal{H}|\mathcal{N}) = \sum_{\{h_j\}} \delta\left(\mathcal{H} - \sum_{j=1}^{\mathcal{N}} h_j\right) \prod_{j=1}^{\mathcal{N}} D(h_j|\ell), \quad (15)$$

where h_j is the score of the j^{th} draw. Using the definition (13) for $D(h|\ell)$ above, we find (with the help of Stirling's formula)

$$\mathcal{D}_\ell(\mathcal{H}|\mathcal{N}) \approx \sum_{\{Q_\ell\}} \exp\left[\mathcal{N}\sum_{\{\mathbf{a},\mathbf{b}\}} Q_\ell[\mathbf{a},\mathbf{b}] \ln\left(\frac{P_0[\mathbf{a},\mathbf{b}]}{Q_\ell[\mathbf{a},\mathbf{b}]}\right)\right] \cdot\delta\left(\mathcal{H} - \mathcal{N}\sum_{\{\mathbf{a},\mathbf{b}\}} Q_\ell[\mathbf{a},\mathbf{b}] \ln W_\ell[\mathbf{a},\mathbf{b}]\right), \quad (16)$$

where $Q_\ell[\mathbf{a},\mathbf{b}]$ is the fraction among the $\mathcal{N}$ draws that contains a particular sequence pair $[\mathbf{a},\mathbf{b}]$ of lengths ℓ.

For $\mathcal{H} \gg 1$, the right-hand side of Eq. (16) can be evaluated in saddle point approximation with the result

$$\mathcal{D}_\ell(\mathcal{H}|\mathcal{N}) \approx e^{-\lambda\mathcal{H}} \cdot \delta(\mathcal{H} - \mathcal{N}\ell\alpha) \quad (17)$$

where

$$\alpha \equiv \ell^{-1} \sum_{\{\mathbf{a},\mathbf{b}\}} Q^*_\ell[\mathbf{a},\mathbf{b}] \ln W_\ell[\mathbf{a},\mathbf{b}] \quad (18)$$

and the $Q^*_\ell[\mathbf{a},\mathbf{b}]$ are given by the saddle point condition

$$Q^*_\ell[\mathbf{a},\mathbf{b}] = W^\lambda_\ell[\mathbf{a},\mathbf{b}] \cdot P_0[\mathbf{a},\mathbf{b}]. \quad (19)$$

The value of λ is fixed by the normalization condition for Q^*:

$$1 = \sum_{[\mathbf{a},\mathbf{b}]} Q^*_\ell[\mathbf{a},\mathbf{b}] = \sum_{[\mathbf{a},\mathbf{b}]} W^\lambda_\ell[\mathbf{a},\mathbf{b}] P_0[\mathbf{a},\mathbf{b}]. \quad (20)$$

Comparing (20) to the normalization condition (8) for the correlated sequence pairs generated from the evolution model, we see that the solution to Eq. (20) is

$$\lambda = 1. \quad (21)$$

This implies that

$$Q^*_\ell[\mathbf{a},\mathbf{b}] = W_\ell[\mathbf{a},\mathbf{b}] \cdot P_0[\mathbf{a},\mathbf{b}], \quad (22)$$

which describes the sequence configurations contributing significantly to the large-$\mathcal{H}$ events. We shall refer to this ensemble of sequence pairs as the *extremal ensemble*. Inserting $W_\ell = Q^*_\ell/P_0$ into Eq. (18), we see that $\alpha > 0$ is nothing but the relative entropy (per length) between the extremal and the random ensemble. Since $W_\ell[\mathbf{a},\mathbf{b}] \cdot P_0[\mathbf{a},\mathbf{b}]$ is also the likelihood of obtaining the *correlated* sequence pair $[\mathbf{a},\mathbf{b}]$ from the evolution model, Eq. (18) shows that α is also the average score (per length) of the global probabilistic alignment of the correlated sequence pair $[\mathbf{a},\mathbf{b}]$. Eq. (17) states that there exists a *preferred* number of draws, $\mathcal{N}^* = \mathcal{H}/(\ell \cdot \alpha)$, which maximizes the probability of observing high scores $\mathcal{H}$.

On first sight, the distribution Eq. (17) seems not to be normalized. However, Eq. (17) describes only the high-$\mathcal{H}$ component of the full distribution $\mathcal{D}_\ell(\mathcal{H}|\mathcal{N})$. Most of the weight of this distribution is by our assumptions in the region of $\mathcal{H} < 0$. Eq. (17) only applies to the region $\mathcal{H} \gg 1$ and therefore does not have to be normalized by itself.

What does $\mathcal{D}_\ell(\mathcal{H}|\mathcal{N})$ have to do with the quantity of interest $D(h|L)$, which describes the probability of obtaining a score h from the global alignment of a *single* sequence pair of length L ? The statistics of the global alignment of correlated sequences has been studied (see, e.g., Hwa & Nattermann, 1995) in terms of the related problem of directed polymers in a random medium (Hwa & Lässig, 1996), and the results have been elaborated in the context of sequence alignment by Drasdo *et al.* (1998). These studies found that $\ln W_L$ can be decomposed into a sum of essentially independent pieces of some length ξ. Thus, the score $\ln W_L$ of the high-scoring sequence pairs can be broken into a sum of *statistically-independent* pieces, each corresponding to the score $\ln W_l$ of a pair of subsequences of length $l \ll L$ as long as $l > \xi$. Then

$$\begin{aligned} D(h|L) &\approx \mathcal{D}_l(h|L/l) \quad (23)\\ &= \sum_{\{h_j\}} \delta\left(h - \sum_{j=1}^{L/l} h_j\right) \prod_{j=1}^{L/l} D(h_j|l)\\ &\approx e^{-h} \cdot \delta(h - L\alpha), \quad (24) \end{aligned}$$

where the last line follows from Eq. (17). The approximation (23) can be further justified as explained in Appendix A. Here, we advertise that we indeed get the announced Poisson statistics with $\lambda = 1$ for the probabilistic global alignment scores which generate the islands together with the information that these high scores are created by pairs of correlated subsequences $(\mathbf{a},\mathbf{b})$ as given by the extremal ensemble $Q^*_\ell[\mathbf{a},\mathbf{b}]$.

Island peak scores

We now turn to derive the island peak score distribution $G(h)$ from the result Eq. (24). This will still require several steps. First, we will calculate the statistics of $Z_{m,n}$ for a fixed choice of (m,n). Afterwards, we will use this statistics to obtain $G(h)$. The connection between the distribution $D(h|L)$ and the statistics of $Z_{m,n}$ for a given (m,n) is made by the observation that $Z_{m,n}$ for a given (m,n) is for large values statistically equivalent to the quantity

$$\overline{Z} \equiv 1 + \sum_{L=1}^{\infty} \overline{W}_L \quad (25)$$

where

$$\overline{W}_L \equiv W[a_1 \ldots a_L, b_1 \ldots b_L]. \quad (26)$$

The derivation of this equivalence is quite technical and therefore relegated to appendix B. It also shows, how large values of Zm,n are generated by sequence configurations which can be described in terms of the

extremal ensemble $Q^*[\mathbf{a},\mathbf{b}]$. Here, we will exploit it by noting that $\overline{W}_L$ is distributed according to the distribution $D(h|L)$. For each L Eq. (24) tells us that roughly $\overline{W}_L$ takes the value $\exp(\alpha L)$ with probability $\exp(-\alpha L)$ and the value 0 with the remaining probability $1-\exp(-\alpha L)$. Of course, the $\overline{W}_L$ are not statistically independent from each other. However, due to the exponential separation of the possible values of the $\overline{W}_L$ for different L, the probability $\Pr\{\overline{Z}=z\}$ for some large enough z is very well described by the probability that $\overline{W}_{L_0}$ or one of the $\overline{W}_L$ with L very close to $L_0 \equiv \frac{1}{\alpha}\ln z$ takes its non-zero value independently of the other $\overline{W}_L$. The values of $\overline{W}_L$ with $L < L_0$ do not matter since they contribute only little to the sum Eq. (25) and the probability for $\overline{W}_L$ with $L > L_0$ being different from zero is exponentially smaller. Thus, $\Pr\{\overline{Z}=z\} \sim \frac{1}{z}$ or equivalently

$$\Pr\{\ln Z_{m,n} = \overline{h}\} \approx \Pr\{\ln \overline{Z} = \overline{h}\} \sim e^{-\overline{h}}. \qquad (27)$$

Finally, we want to relate this distribution of $Z_{m,n}$ at fixed (m,n) to the island peak score distribution $G(h)$. Again, the exponential dependence of the score distribution on the score $\overline{h}$ is essential. The fact, that the probability to find a score $\overline{h}$ at a given point (m,n) on the scoring lattice is an exponential in the score implies that increments in score from one lattice point to the next are essentially independent of the actual score $\overline{h}$ at this lattice point. Specifically, the probability of finding an even higher score at some other neighboring point (m',n') is essentially independent of the score $\overline{h}$ itself either. Thus, for any (m,n) with $Z_{m,n}$ sufficiently large, the probability of (m,n) being an island peak point is some number which depends on the value of $Z_{m,n}$ at most very weakly. Therefore, the probability for an island peak score being h is approximately proportional to the probability of the score $\ln Z_{m,n}$ being h for a fixed (m,n). We calculated the latter with the result given in Eq. (27). Therefore we have the exponential statistics $G(h) \sim e^{-h}$ for the island peak score distribution which implies the Gumbel statistics of S with $\lambda = 1$ as discussed above.

Sequence Length Correction

In order to verify the arguments and derivations leading to this result and the extremal ensemble $Q^*_\ell[\mathbf{a},\mathbf{b}]$ we have performed extensive numerical simulations. However, as presented thus far, our results pertain only to the asymptotic limit of infinitely long sequences. To compare to the numerics performed at *finite* sequence lengths, it is necessary to compute the magnitude of the *corrections* to this result due to the finite sequence length which we will turn to now.

For sequences of finite length we expect a deviation from the asymptotic value $\lambda = 1$ predicted by the above considerations. In order to assess the significance of an alignment of two sequences of finite length we therefore have to characterize this deviation as well.

It was pointed out by Altschul (1991) in the context of gapless local alignment and more recently by Altschul and Gish (1996) for gapped alignment that in using the Gumbel distribution (1) for finite length sequences, one should "correct" the lengths M and N which appear in (1) by a score-dependent amount $L(\mathsf{S})$, and use instead the *effective* sequence lengths $M' = M - L(\mathsf{S})$ and $N' = N - L(\mathsf{S})$. It results from the fact that the available area to launch an island is *reduced* by the size of the island on the alignment lattice, which is the $M \times N$ square obtained by putting one of the two sequences along the $\hat{x}$ direction and the other sequence along the $\hat{y}$ direction in the plane.

As an extreme example, one notes that to have an island of the size of the entire alignment lattice, the island must be launched near the tip of the lattice; in this case, the correction term $L(\mathsf{S})$ is nearly the size of the lattice. Generally, one should take $L(\mathsf{S})$ to be the average island length[2] $\overline{\ell}(\mathsf{S})$ corresponding to the score of the maximum island peak S. Including this correction, the Gumbel statistics becomes

$$\Pr\{\mathsf{S} < x\} = \exp\left[-K \cdot (N - \overline{\ell}(x))^2\, e^{-\lambda x}\right], \qquad (28)$$

where we have used the more convenient accumulated distribution, and have taken the two sequences to be of equal length N for simplicity. Using the linear island profile $\overline{\ell}(x) = \alpha^{-1}x$ for large islands where α is the "relative entropy", Altschul *et al.* (2001) noted that the terms in Eq. (28) can be rearranged into the classic Gumbel form, i.e.,

$$\Pr\{\mathsf{S} < x\} = \exp\left[-K(N) \cdot N^2\, e^{-\lambda(N)\cdot x}\right], \qquad (29)$$

with the effective size-dependent parameter $\lambda(N) = \lambda + 2/(\alpha N)$ to leading order in $1/N$. More generally, one has the relation

$$\lambda(N) = \lambda + 2/\overline{\sigma}(N), \qquad (30)$$

where $\overline{\sigma}(\ell)$ is the inverse of the function $\overline{\ell}(\sigma)$, and gives the average score for islands of length ℓ. Note that correction formulae such as (30) are applicable as long as the number of islands in the alignment lattice is large. They should however not be applied to very small N's where the sequence lengths are of the same order as the island sizes, and the Gumbel distribution itself breaks down.

It is also possible to extend the analysis discussed above for the parameter K. Using the form $\overline{\ell}(x) = \alpha x + c$ in Eq. (28) and rearranging terms into the Gumbel form Eq. (29), we find the result

$$K(N) = K \cdot \left(1 + \frac{c}{\alpha N}\right)^2 \qquad (31)$$

[2]The island width is typically much smaller than its length and hence does not contribute to leading orders.

which is analogous to Eq. (30) for $\lambda(N)$. Unlike the case for λ, we have not yet developed a theory to compute the asymptotic value of K. It is however still possible to check the form of the correction formula (31) using the numerically obtained values of $K(N)$; see below.

Numerics

Although we presented our statistical theory here only based on the simplest linear gap function, it can be easily generalized to incorporate affine gap costs as well. Since affine gap costs are much more frequently used, we present our numerics based on affine gap costs only. In our affine gap function, we have used the symbol μ to denote the weight of gap initiation and ν to denote the weight of a gap extension. It turns out that an equation similar to (4) still exists in which the scaling constant for the transition matrix now depends on both μ and ν. The joint probability distribution $\mathcal{P}(a,b)$ is given in terms of a scoring matrix $s(a,b)$ by $\mathcal{P}(a,b) = e^{\lambda_{\rm ug} s(a,b)} p(a)p(b)$ with $\lambda_{\rm ug}$ defined by $\sum_{a,b} e^{\lambda_{\rm ug} s(a,b)} p(a)p(b) = 1$. For the $s(a,b)$, we use Dayhoff's PAM substitution matrices (Dayhoff *et al.*, 1978).

We use two sets of scoring parameters described by PAM distance $d = 120$, $\mu = 2^{-5.5}$, $\nu = 2^{-0.5}$ and PAM distance $d = 250$, $\mu = 2^{-6}$, $\nu = 2^{-0.5}$ respectively. For brevity, we refer to the first set of parameters as "PAM-120" and the second set as "PAM-250".

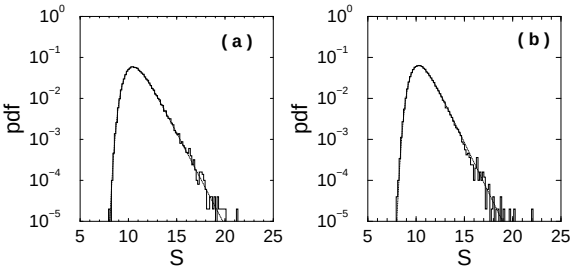

Figure 3: The pdf's for the semi-probabilistic alignment of random sequences using the two parameter sets (a) PAM-120 and (b) PAM-250. The pdf's are obtained by normalizing histograms of 50,000 pairs of random sequences of length 300 each.

We start with the numerical verification that the MLL score obeys Gumbel statistics. We use the two sets of scoring systems PAM-120 and PAM-250 satisfying the conservation condition for affine gaps. Figs. 3(a) and (b) show the pdf's of S obtained from the alignment of 50,000 pairs of random sequences of lengths 300 each, generated according to the null model (2). We see that the pdf's are well-fitted by the Gumbel distribution (1).

In order to measure $\overline{\sigma}(N)$ in a very effective way, we made use of our knowledge of the extremal ensemble. Instead of aligning random sequences and waiting for large islands, we directly generated typical island score landscapes using the correlated ensemble $Q^*_N[\mathbf{a},\mathbf{b}]$ and read off the average score for each length N. The results corresponding to the PAM-120 and PAM-250 scoring systems are shown in Fig. 4. They are well fitted by the form $\overline{\sigma} = \alpha N + c$, with statistical uncertainties in α and c well under 1%. The most striking thing about this result is that the data points in Fig. 4 were averaged over only 15 pairs of alignments and took practically no time to generate, while determining α to such precision using direct simulation or island counting will take weeks on the same computer.

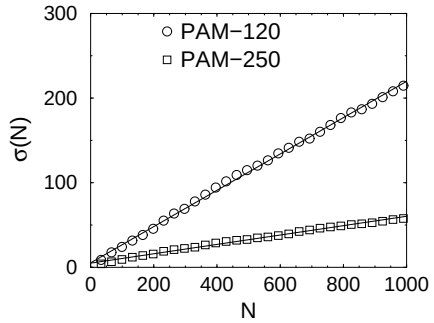

Figure 4: The circles and squares represent the alignment score of correlated sequences taken from the extremal ensemble, averaged over only 15 pairwise alignments, corresponding to the PAM-120 and PAM-250 scoring systems respectively. The lines represent the respective least-square fits to $\overline{\sigma}(N) = \alpha N + c$. The fits, which are excellent down to $N = 50$, give $\alpha = 0.0554, c = 4.85$ for PAM-250 and $\alpha = 0.2144, c = 5.22$ for PAM-120.

With the accurate determination of $\overline{\sigma}(N)$, we are now in a position to test the prediction of the sequence length dependence (30), and with it, the prediction of the asymptotic result $\lambda = 1$. The predicted expression of $\lambda(N)$ using the numerically obtained $\overline{\sigma}(N)$'s in Eq. (30) is plotted as the line in Figs. 5(a) and (b) for the PAM-120 and PAM-250 scoring systems respectively. Also plotted are the data points obtained from fitting the pdf's for sequences of different lengths to the Gumbel form (1). We find very good agreement between theory and measurements down to sequence length of $N = 75$ for PAM-120 and $N = 150$ for PAM-250. (For smaller N's, the pdf's are no longer well-described by the Gumbel distribution for reasons explained earlier.) The striking agreement found lends strong support to the theory presented.

To test the prediction of the sequence length dependence of $K(N)$, we simply plot on the verti-

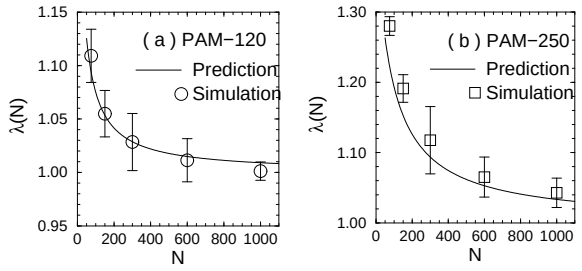

Figure 5: Direct comparisons of the numerical values of λ obtained from fitting pdf's to Gumbel form (1) and the theoretical prediction for (a) PAM-120 and (b) PAM-250 scoring systems.

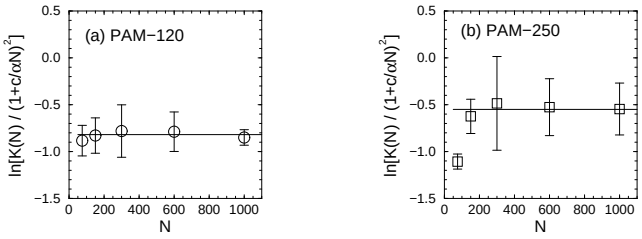

Figure 6: Dependence of $K(N)$ on the sequence length for (a) PAM-120 and (b) PAM-250 scoring systems. The horizontal lines indicate the values of the asymptotic K's obtained from data at larger N's (not shown).

cal axis $K(N)/(1 + c/(\alpha N))^2$, using the values of c and α determined from Fig. 4 for the corresponding scoring system. According to Eq. (31), this simple transformation should render the data points N-independent, and give the value of the asymptotic K. We applied this transformation to $K(N)$ separately for the PAM-120 and PAM-250 scoring systems; see Figs. 6 (a) and (b). Other than the smallest size of $N = 75$, the data points are approximately[3] N-independent, hovering around the asymptotic values indicated by the horizontal lines. The results suggest that Eq. (31) does capture the dependence of $K(N)$ on the sequence length correctly. Consequently, it is only necessary to determine $K(N)$ for one sequence length, say, at $N = 300$ by an island counting method similar to Olsen *et al.* (1999), or at $N \to \infty$ if the present theory can be extended to compute K. From this, the value of the effective K for all other N's can be deduced from the sequence length dependence formula (31).

[3]The statistical uncertainties associated with the $K(N)$'s are much larger because the actual parameter used in the Gumbel fit was $K(N)N^2$.

Summary

In this paper, we studied the extremal statistics of probabilistic sequence alignment both analytically and numerically. We find that while the statistics of straightforward probabilistic alignment is not understood, the slightly modified semi-probabilistic alignment is well described by Gumbel statistics. For the semi-probabilistic alignment, we can predict the Gumbel parameter λ, including its sequence length dependence, for different scoring functions and parameters. Moreover, for a given scoring scheme, we have characterized the corresponding extremal ensemble of most detectable sequence-pairs. This allows for an optimal choice of scoring parameters for a given search goal. Our results are verified numerically by using various PAM substitution matrices and affine gap functions.

In our study, we have not focused on the behavior of the other Gumbel parameter, K, which is more difficult to compute analytically than λ. It is however straightforward to determine K numerically by extending the island method of Olsen *et al.* (1999) to the semi-probabilistic alignment (Bundschuh *et al.*, to be published.) With the help of a precisely determined λ, and the formula for the sequence length dependence of K, it is possible to fix the value of K for all sequence lengths by counting islands from a single pairwise alignment of managable size.

Let us close with a general remark: While the numerics presented in the present study was restricted to position-independent scoring functions, this is not a prerequisite for the application of our theory. In fact, we expect that the asymptotic value of λ to remains1 as long as the probability conservation condition (4) is *locally satisfied* at each node of the alignment lattice. This can be readily accomplished for position-specific substitution and indel weights by generalizing our previous result (Yu & Hwa, 1999).

Acknowledgments

We thank Rolf Olsen for much help and many useful suggestions during the course of this study. This research is supported in part by the Beckman Foundation and the NSF through Grant No. DMR-9971456. TH further acknowledges the financial support of a Guggenheim Fellowship, and RB acknowledges a Hochschulsonderprogramm III fellowship of the DAAD. Finally, TH and YKY gratefully acknowledge the hospitality of the Center for Studies in Physics and Biology at Rockefeller University where this work was initiated.

Appendix A: Consistency of our approximations

As a simple consistency check of our assumption (23), we see that insertion of Eq. (24) into Eq. (15) recovers

the result (17), as long as α is *length-independent.* To probe the validity of the assumption more closely, let us introduce the notation $\langle ... \rangle^*$ to be the statistical average over the extremal ensemble $Q_L^* = W_L \cdot P_0$, i.e.,

$$\langle F[\mathbf{a},\mathbf{b}] \rangle^* \equiv \sum_{\{\mathbf{a},\mathbf{b}\}} F[\mathbf{a},\mathbf{b}] W[\mathbf{a},\mathbf{b}] P_0[\mathbf{a},\mathbf{b}]. \qquad (32)$$

Then, the distribution (13) can be rewritten as

$$D(h|L) = e^{-h} \langle \delta(h - \ln W_L[\mathbf{a},\mathbf{b}]) \rangle^*, \qquad (33)$$

while the result of our assumption, Eq. (24), can be re-written as

$$D(h|L) \approx e^{-h} \delta\left(h - \langle \ln W_L[\mathbf{a},\mathbf{b}] \rangle^*\right). \qquad (34)$$

Thus, the approximation made here is to replace the random variable $\ln W_L$ by its *typical* value in the extremal ensemble, $E^*[\ln W_L] \equiv \langle \ln W_L[\mathbf{a},\mathbf{b}] \rangle^*$. Approximations of this nature are poor in cases where the distribution of $\ln W_L$ is broad, e.g., if

$$\mathrm{var}(\ln W_L) \equiv \langle \ln W_L^2[\mathbf{a},\mathbf{b}] \rangle^* - E^*[\ln W_L]^2$$

is comparable to $E^*[\ln W_L]^2$. Due to the fact that the alignment of a pair of sequences of the extremal ensemble can be split into independent pieces of length ξ as discussed in the main text, we get $\mathrm{var}(\ln W_L) \approx L/\xi$ for $L \gg \xi$ according to the central-limit theorem. Since $E^*[\ln W_L] = \alpha L$, we have $\mathrm{var}(\ln W_L)/E^*[\ln W_L]^2 \to 0$ in the limit of large L. In this way, the equivalence between Eq. (33) and Eq. (34) is justified.

Appendix B: Z and W

In this appendix we show that the restricted local alignment weight $Z_{m,n}$ at fixed (m,n) is at large values statistically equivalent to $\overline{Z}$ as defined in Eq. (25). First, we note that due to the symmetry of $\mathcal{W}_{m',n';m,n}$ and the fact that random sequences are statistically equivalent to their reverse sequences, the statistics of $Z_{m,n}$ is identical to the statistics of $\widehat{Z}_{M-m,N-n}$ defined by reversing the sequences, i.e., by

$$\widehat{Z}_{m,n} \equiv 1 + \sum_{m'=1}^{M-m} \nu^{m'} + \sum_{n'=1}^{N-n} \nu^{n'} + \sum_{\substack{m+1 \le m' \le M \\ n+1 \le n' \le N}} \mathcal{W}_{m+1,n+1;m',n'}. \qquad (35)$$

Moreover, since we assume that the distribution of $Z_{m,n}$ is translationally invariant as long as we stay far enough away from the edges $m \approx 1$ and $n \approx 1$, we can without loss of generality pick $(m,n) = (M,N)$ and study the statistics of $\widehat{Z}_{0,0}$. This statistics should at most very weakly depend on the values of $\widehat{Z}_{m,n}$ for very large m and n and we can therefore extend the summations in Eq. (35) to infinity. To summarize, we expect that the statistics of $Z_{m,n}$ for a fixed (m,n) are identical to the statistics of

$$\widetilde{Z} \equiv 1 + \sum_{m'=1}^{\infty} \nu^{m'} + \sum_{n'=1}^{\infty} \nu^{n'} + \sum_{m'=1}^{\infty} \sum_{n'=1}^{\infty} \mathcal{W}_{1,1;m',n'}. \qquad (36)$$

In order to study $\widetilde{Z}$ we rewrite it as

$$\widetilde{Z} = 1 + \sum_{L=1}^{\infty} \widetilde{W}_L \qquad (37)$$

with

$$\widetilde{W}_L \equiv 2\nu^L + \sum_{m'=1}^{L-1} \mathcal{W}_{1,1;m',L} + \sum_{n'=1}^{L-1} \mathcal{W}_{1,1;L,n'} + \mathcal{W}_{1,1;L,L}. \qquad (38)$$

This quantity looks very similar to $\overline{W}$ as defined in Eqs. (26) and (5) and indeed we can easily convince ourselves using Eq. (3) that

$$\overline{W}_L \le \widetilde{W}_L \le \frac{1}{1-\nu} \overline{W}_L. \qquad (39)$$

Thus, the $\overline{Z}$ defined in Eq. (25) bounds $\widetilde{Z}$ as

$$\overline{Z} \le \widetilde{Z} \le \frac{1}{1-\nu} \overline{Z} \qquad (40)$$

and since we are only interested in the logarithms of these quantities they become statistically equivalent for large enough values of $\overline{Z}$.

Under all these transformations, sequence configurations which make $Z_{m,n}$ large for a fixed (m,n) are directly related to configurations of the (reversed !) sequences $a_m a_{m-1} \dots a_{m-L_0}$ and $b_n b_{n-1} \dots b_{n-L_0}$ which make $\overline{W}_{L_0}$ large for some L_0. Thus, these reversed sequences are drawn from the extremal ensemble $Q^*_{L_0}[a_m a_{m-1} \dots a_{m-L_0}, b_n b_{n-1} \dots b_{n-L_0}]$ derived in Eq. (22).

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.

Altschul, S.F., Bundschuh, R., Hwa, T., and Olsen, R., 2001. The estimation of statistical parameters

for local alignment score distributions. *Nucleic Acids Research* 29:351–361.

Arratia, R., Morris, P., and Waterman, M.S., 1988. Stochastic scrabbles: a law of large numbers for sequence matching with scores. *J. Appl. Prob.* 25:106-119.

Bishop, M.J., and Thompson, E.A., 1986. Maximum likelihood alignment of DNA sequences. *J. Mol. Biol.* 190:159-165.

Bundschuh, R., 2000. An Analytic Approach to Significance Assessment in Local Sequence Alignment with Gaps. *RECOMB 2000.*

Collins, J.F., Coulson, A.F.W., and Lyall, A., 1988. The significance of protein sequence similarities. *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.

Drasdo, D., Hwa, T., and Lassig, M., 1998. A Scaling Theory of Sequence Alignment with Gaps. *ISMB98*:52-58.

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

Henikoff, S., and Henikoff, J.G., 1994. Position-based Sequence Weights. *J. Mol. Biol.* 162:705-708.

Hughey, R., and Krogh, A., 1996. Hidden Markov Models for Sequence Analysis: Extension and Analysis of the Basic Method. *CABIOS* 12:95-107.

Hwa, T., and Nattermann, T., 1995. Disorder-induced depinning transition. *Phys. Rev. B* 51:455–469.

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

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.

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

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., Bundschuh, R., and Hwa, T., 1999. Rapid Assessment of Extremal Statistics for Gapped Local Alignment. *Proceedings of The Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB99).* T. Lengauer et al. eds., 211–222 (AAAI Press, Menlo Park).

Pearson, W.R., 1988. Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. USA* 85:2444-2448.

Smith, T.F., and Waterman, M.S., 1981. Identification of Common Molecular Subsequences. *J. Mol. Biol.* 147:195-197.

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

Thorne, J.L., Kishino, H., and Felsenstein, J. 1991. An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences. *J. Mol. Evol.* 33:114-124.

Thorne, J.L., Kishino, H., and Felsenstein, J., 1992. Inching toward Reality: An Improved Likelihood Model of Sequence Evolution. *J. Mol. Evol.* 34:3-16.

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.

Yu, Y.-K., and Hwa, T., 1999 Statistical Significance of Probabilistic Sequence Alignment and Related Local Hidden Markov Models. *Submitted to J. Comp. Biol.*