JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 3, 2001
Mary Ann Liebert, Inc.
Pp. 249–282

# Statistical Significance of Probabilistic Sequence Alignment and Related Local Hidden Markov Models

YI-KUO YU[1] and TERENCE HWA[2]

## ABSTRACT

**The score statistics of probabilistic gapped local alignment of random sequences is investigated both analytically and numerically. The full probabilistic algorithm (e.g., the "local" version of maximum-likelihood or hidden Markov model method) is found to have anomalous statistics. A modified "semi-probabilistic" alignment consisting of a hybrid of Smith–Waterman and probabilistic alignment is then proposed and studied in detail. It is predicted that the score statistics of the hybrid algorithm is of the Gumbel universal form, with the key Gumbel parameter $\lambda$ taking on a *fixed* asymptotic value for a wide variety of scoring systems and parameters. A simple recipe for the computation of the "relative entropy," and from it the finite size correction to $\lambda$, is also given. These predictions compare well with direct numerical simulations for sequences of lengths between 100 and 1,000 examined using various PAM substitution scores and affine gap functions. The sensitivity of the hybrid method in the detection of sequence homology is also studied using correlated sequences generated from toy mutation models. It is found to be comparable to that of the Smith–Waterman alignment and significantly better than the Viterbi version of the probabilistic alignment.**

**Key words:** sequence alignment, statistical significance, maximum likelihood, hidden Markov model.

## 1. 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. The widespread usage of these tools for database searches is closely tied to their ability to assign a $p$-value to each pairwise alignment (Karlin and Altschul, 1990). The $p$-value is much more meaningful than the alignment score itself as it gives the probability that a score could have arisen by chance.

There exist many other applications of bioinformatics where dynamic programming algorithms analogous to those of sequence alignment are extensively used. Some particularly noteworthy examples are application

[1] Department of Physics, Florida Atlantic University, Boca Raton, FL 33431-0991.
[2] Department of Physics, University of California at San Diego, La Jolla, CA 92093-0319.

of maximum likelihood methods and hidden Markov models (HMM) to protein modeling (Krogh *et al.*, 1994), gene finding (Burge and Karlin, 1997), motif search (Bucher *et al.*, 1996; Grundy *et al.*, 1997), and even sequence alignment itself (Thorne *et al.*, 1991, 1992; Hughey and Krogh, 1996); see the book by Durbin *et al.* (1998) for a recent review. These methods differ from the usual alignment algorithms, e.g., that of Smith and Waterman (1981), in that they are *probabilistic* in nature. It will be very useful if statistical characterization such as those provided by BLAST can be extended to the probabilistic methods. There are, however, two major obstacles in the way: a) unlike sequence alignment, for which the null statistics is known (either exactly or empirically) to be of the Gumbel extremal distribution (Gumbel, 1958), the *form* of the extremal statistics for probabilistic algorithms such as the HMM is not known at all; b) even for those whose null statistics are of the Gumbel form, the dependence of the two Gumbel parameters on the hundreds of model parameters is generally so complicated that it is hopeless to determine the Gumbel parameters in an efficient enough manner to render them useful.

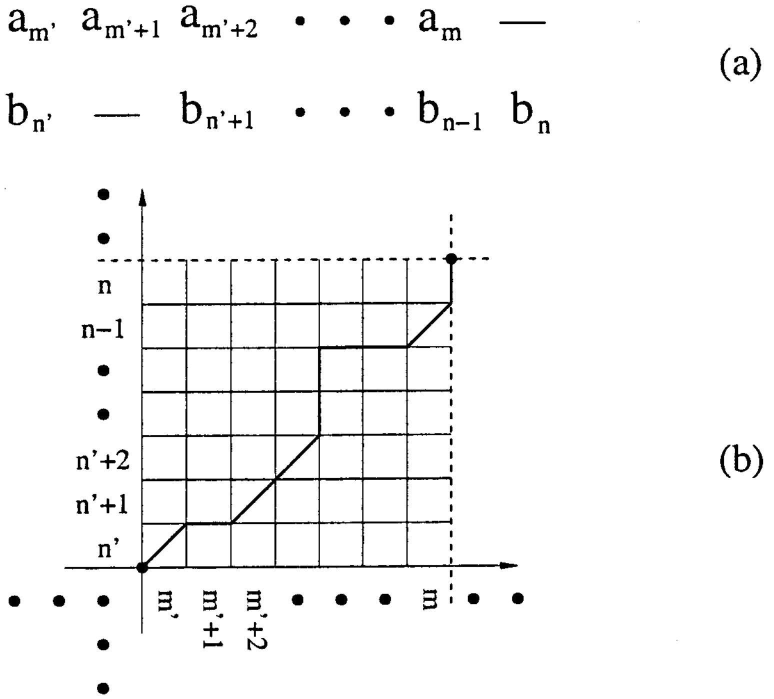
In fact, the obstacle (b) is already a problem for sequence alignment with gaps, which is known empirically to obey Gumbel statistics (Smith *et al.*, 1985; Collins *et al.*, 1988; Mott, 1992; Waterman and Vingron, 1994a, 1994b; Altschul and Gish, 1996; Olsen *et al.*, 1999). This problem is partially overcome in BLAST by precomputing the null statistics for a fixed set of scoring parameters. This, however, makes the method somewhat inflexible and especially becomes a problem for *position-specific* scoring functions (Henikoff and Henikoff, 1994) such as those used in PSI-BLAST (Altschul *et al.*, 1997). The latter is needed for detailed modeling of protein families, folds, etc. However, because of the obstacle (b), PSI-BLAST is limited presently to *uniform* gap penalty which is a very costly restriction.

In this paper, we study analytically and numerically the null statistics of probabilistic alignment as an example of the general class of maximum likelihood and HMM methods. We find numerically that the probabilistic alignments lead to anomalous statistics, with the tail of the log-odd score distribution being even *broader* than the exponential tails of the Gumbel distribution. We then propose a "semi-probabilistic" alignment which is a *hybrid* of the probabilistic and the usual Smith–Waterman-type algorithm and has the same computational complexity as the probabilistic algorithm. We focus on the extremal score statistics of the semi-probabilistic alignment. We show heuristically that its extremal statistics is of the Gumbel form and give conditions which fix the Gumbel parameter $\lambda$ for a wide range of scoring functions/parameters. Furthermore, we give a straightforward recipe for the computation of the relative entropy which characterizes the information content of an alignment. The knowledge of the relative entropy allows us to predict also the *finite-size correction* to $\lambda$, which is important for characterizing the statistics of short sequences. Comparison of our theoretical results to effective $\lambda$ values obtained from direct numerical simulation of random amino acid sequences using the PAM substitution matrices and affine gap functions shows good agreement (to within the numerical accuracy of $\sim 1\%$) for sequence lengths ranging from 100 to 1,000. We further tested the *sensitivity* of the semi-probabilistic alignment, by aligning correlated sequences generated from toy mutation models. For our simple test sequences, the sensitivity is of the same order or slightly better than that of the Smith–Waterman algorithm and can be significantly better than the Viterbi version of the full probabilistic algorithm. Taken together, we conclude that the semi-probabilistic alignment, with its well-characterized statistics, can be applied in a wide variety of contexts from simple sequence comparisons to detailed sequence modeling.

## 2. REVIEW OF ALIGNMENT AND STATISTICS

### 2.1. Alignment algorithm

Let $\mathbf{a} = [a_1, a_2, \ldots, a_M]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_N]$ be two sequences of lengths $M$ and $N$ respectively, with elements $a_i$ and $b_j$ taken from a finite character set $\chi$. Let $\hat{\mathbf{a}}_{m';m} = [a_{m'}, a_{m'+1}, \ldots, a_m]$ and $\hat{\mathbf{b}}_{n';n} = [b_{n'}, b_{n'+1}, \ldots, b_n]$ denote *subsequences* of $\mathbf{a}$ and $\mathbf{b}$ respectively, with $1 \leq m' \leq m \leq M$, and $1 \leq n' \leq n \leq N$. A restricted global alignment $\hat{\mathcal{A}}$ of the sequences $\hat{\mathbf{a}}_{m';m}$ and $\hat{\mathbf{b}}_{n';n}$ consists of an ordered set of pairings of their elements, with each other or with gaps, e.g., $\hat{\mathcal{A}} = \{(a_{m'}, b_{n'}), (a_{m'+1}, -), (a_{m'+2}, b_{n'+1}), \ldots, (a_m, b_{n-1}), (-, b_n)\}$ for the example shown in Fig. 1(a). Here, $(a_i, b_j)$ denotes the pairing of elements $a_i$ and $b_j$, and $(a_i, -)$ and $(-, b_j)$ denote pairing of an element with a *gap*; we refer to these three types of pairings as substitutions, deletions, and insertions, respectively. In typical alignment applications, pairings



$a_{m'}$ $a_{m'+1}$ $a_{m'+2}$ • • • $a_m$ —
$b_{n'}$ — $b_{n'+1}$ • • • $b_{n-1}$ $b_n$
(a)
n
n−1
n'+2
n'+1
n'
m'
m'+1
m'+2
m
(b)

FIG. 1. Example of an alignment and the corresponding directed path. (**a**) A possible global alignment of the sequences $\hat{\mathbf{a}}_{m';m}$ and $\hat{\mathbf{b}}_{n';n}$; (**b**) the directed path representation of the alignment shown in (a). For a restricted global alignment $\hat{\mathcal{A}}$ of $\hat{\mathbf{a}}_{m';m}$ and $\hat{\mathbf{b}}_{n';n}$, the corresponding directed path must have one of its ends (the "backward end") fixed at the lower left corner of the cell $(m', n')$ and the other end (the "forward end") fixed at the upper right corner of the cell $(m, n)$. The two dashed lines which mark the right and upper boundaries of the alignment region will be referred to as the "forward boundaries" in the text.

of gaps to each other are not allowed. It is also a common practice to restrict the order of insertions and deletions, e.g., to forbid insertions following deletions, in order to avoid overcounting of the same alignments.[1] With these restrictions, each alignment $\hat{\mathcal{A}}$ can be uniquely represented by the set $\mathcal{R}$ of index pairs $(i, j)$ for all paired elements $(a_i, b_j)$; e.g., $\mathcal{R} = \{)m', n'), (m' + 2, n' + 1), \ldots, (m, n - 1)\}$ for the example in Fig. 1(a). Generally, we shall use the notation

$$\mathcal{R}(m', n'; m, n) = \{(m_1, n_1), (m_2, n_2), \ldots, (m_l, n_l)\} \tag{1}$$

to denote the set of $l$ pairings in an alignment. A valid restricted global alignment $\hat{\mathcal{A}}$ for the sequences $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ is then any set $\mathcal{R}$ of index pairs satisfying the condition $m' \le m_1 < m_2 < \ldots < m_l \le m$ and $n' \le n_1 < n_2 < \ldots < n_l \le n$. $\mathcal{R}$ can also be viewed as coordinates of a *directed path* on the alignment grid, with the "backward end" of the path fixed at the lower left corner of the cell $(m', n')$ and the "forward end" fixed at the upper right corner of the cell $(m, n)$; see Fig. 1(b).

The *score* $\mathcal{S}$ of the alignment $\hat{\mathcal{A}}$ is obtained by summing up the individual pairing scores, e.g., $s(a_i, b_j)$ for the pairing of elements $a_i$ and $b_j$, and the "gap scores." For protein sequences, the frequently used pairing scores are the PAM or BLOSUM substitution scores (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992), constructed from empirical amino acid substitution frequencies. The frequently used affine-gap function assigns a cost of $\delta + \varepsilon \cdot (\ell - 1)$ for each consecutive run of $\ell$ gaps in a given sequence. An additional cost $\delta'$ can be assigned to penalize the situation where a run of gaps in one sequence is

[1] For example, the alignments $\{(a_1, -), (-, b_1), (a_2, -)\}$ and $\{(a_1, -), (a_2, -), (-, b_1)\}$ both describe the situation where the elements $a_1$ and $a_2$ are not aligned with $b_1$ and thus should not be multiply counted.

immediately followed by a run of gaps in the other sequence.[2] Let the length of the two gaps separating two pairings be $\ell_1$ and $\ell_2$, respectively. Then the gap cost function $\gamma$ can be written as

$$\gamma(\ell_1, \ell_2) = \begin{cases} 0 & \ell_1 = 0, \ell_2 = 0 \\ \delta + \varepsilon \cdot (\ell_1 - 1) & \ell_1 \geq 1, \ell_2 = 0 \\ \delta + \varepsilon \cdot (\ell_2 - 1) & \ell_1 = 0, \ell_2 \geq 1 \\ \delta' + 2\delta + \varepsilon \cdot (\ell_1 + \ell_2 - 2) & \ell_1 \geq 1, \ell_2 \geq 1. \end{cases} \tag{2}$$

Given the scoring functions $s$, $\gamma$ and the alignment path $\mathcal{R}$, the score for this alignment of $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ is uniquely determined:

$$\mathcal{S}[\mathcal{R}; \hat{\mathbf{a}}; \hat{\mathbf{b}}; s, \gamma] = \sum_{k=1}^{l} s(a_{m_k}, b_{n_k}) - \sum_{k=0}^{l} \gamma(m_{k+1} - m_k - 1, n_{k+1} - n_k - 1)], \tag{3}$$

where we used $(m_0, n_0) = (m' - 1, n' - 1)$ and $(m_{l+1}, n_{l+1}) = (m + 1, n + 1)$ to compact the notation. The alignment with the highest score (for a given sequence pair $[\hat{\mathbf{a}}, \hat{\mathbf{b}}]$ and given scoring functions) is the optimal restricted global alignment $\hat{\mathcal{A}}^*$, with score

$$S_{m',n';m,n} = \max_{\mathcal{R}(m',n';m,n)} \{\mathcal{S}[\mathcal{R}; \hat{\mathbf{a}}; \hat{\mathbf{b}}; s, \gamma]\}. \tag{4}$$

This score can be computed via a well-known dynamic programming algorithm provided by Needleman and Wunsch (1970). Below is the simplest example for the case of *linear* gap cost with $\delta = \varepsilon$, $\delta' = 0$, and no constraint in the order of occurrence of insertions and deletions. Extension to the affine gap function (2) is given in Appendix A. To compute $S_{m',n';m,n}$, one simply iterates the following recursion relation

$$S_{m',n';i,j} = \max \left\{ \begin{array}{c} S_{m',n';i-1,j-1} + s(a_i, b_j) \\ S_{m',n';i-1,j} - \varepsilon,\ S_{m',n';i,j-1} - \varepsilon \end{array} \right\} \tag{5}$$

for $i = m'$ to $m$ and $j = n'$ to $n$, with the "boundary condition"

$$S_{m',n';i,j=n'-1} = -\varepsilon \cdot [i - (m' - 1)];\ S_{m',n';i-m'-1,j} = -\varepsilon \cdot [j - (n' - 1)]. \tag{6}$$

This boundary condition enforces the anchoring of the "backward end" of the alignment path as shown in Fig. 1(b).

A *local* alignment $\mathcal{A}$ between the sequences **a** and **b** is *any* restricted global alignment of the subsequences $\hat{\mathbf{a}}_{m';m}$ and $\hat{\mathbf{b}}_{n';n}$, alignment of $\hat{\mathbf{a}}$ or $\hat{\mathbf{b}}$ with null, or the "null alignment" (i.e., no alignment at all). The optimal local alignment $\mathcal{A}^*$ is one whose score $\mathsf{S} = \mathcal{S}[\mathcal{A}^*]$ is the highest; the corresponding alignment path is denoted by $\mathcal{R}^*$. From (4), we have

$$\mathsf{S}[\mathbf{a}, \mathbf{b}; s, \gamma] = \max_{\substack{1 \leq m' \leq m \leq M \\ 1 \leq n' \leq n \leq N}} \{S_{m',n';m,n}, 0\}, \tag{7}$$

where the entry "0" in (7) selects the null alignment if alignments between all possible subsequences are below a threshold, e.g., zero. The score $\mathsf{S}$ is called the optimal local alignment score, or simply the optimal score.

Smith and Waterman (1981) developed an efficient strategy to compute the optimal score $\mathsf{S}$: First, define the "restricted" local alignment score $H_{m,n}$ to be

$$H_{m,n} = \max_{\substack{1 \leq m' \leq m \\ 1 \leq n' \leq n}} \{S_{m',n';m,n}, 0\}. \tag{8}$$

[2] In the problem considered originally by Smith and Waterman (1981), each run of gaps had to terminate with a pairing; this corresponds to the limit $\delta' = \infty$ in our scoring system. For a discussion of the generalized gap functions, see Altschul (1998).

It records the optimal local alignment between the subsequences $\hat{\mathbf{a}}_{1;m}$ and $\hat{\mathbf{b}}_{1;n}$. The $H$'s can again be computed by dynamic programming. For the simple linear gap function, it reads

$$H_{m,n} = \max \left\{ \begin{array}{c} H_{m-1,n-1} + s(a_m, b_n) \\ H_{m-1,n} - \varepsilon,\, H_{m,n-1} - \varepsilon,\, 0 \end{array} \right\}, \tag{9}$$

with the boundary condition $H_{0,n} = 0 = H_{m,0}$. The affine gap version of the algorithm is provided in Appendix A. Given $H_{m,n}$ for all $1 \le m \le M$, $1 \le n \le N$, the optimal alignment score $\mathsf{S}$ defined in (7) is obtained simply as

$$\mathsf{S}[\mathbf{a}, \mathbf{b}; s, \gamma] = \max_{\substack{1 \le m \le M \\ 1 \le n \le N}} \{H_{m,n}\}. \tag{10}$$

The combination of Equations (9) and (10) is the celebrated Smith–Waterman local alignment algorithm.

## 2.2. *Alignment score statistics*

It is important to realize that the value of the optimal score $\mathsf{S}$ does not in itself convey any meaning regarding the degree of homology between the sequences being aligned. One way to assess sequence homology is to compare the score $\mathsf{S}$ with the optimal score of aligning sequences from a null model. A frequently used null model is that of the mutually uncorrelated Markov random chains, described by the distribution function

$$P_0[\mathbf{a}, \mathbf{b}] = \prod_{\substack{1 \le m \le M \\ 1 \le n \le N}} p(a_m) \cdot p(b_n), \tag{11}$$

where $p(a)$ is the background frequency for the element $a$, with $\sum_{a \in \chi} p(a) = 1$. The probability distribution function (pdf) of optimal scores for the alignment of random sequences is

$$\mathrm{pdf}(\mathsf{S}) = \langle \delta(\mathsf{S} - \mathsf{S}[\mathbf{a}; \mathbf{b}; s, \gamma]) \rangle_0, \tag{12}$$

where $\langle \ldots \rangle_0$ denotes average over the null sequence distribution (11). The pdf (12) provides the $p$-value, that an alignment of two uncorrelated random sequences receives, an optimal score $\mathsf{S}$. It is the "holy-grail" of statistical studies of sequence alignment.

*2.2.1. Gapless alignment.* Clearly, the pdf (12) would depend generally on the sequence lengths $M$, $N$ and the scoring functions $s$ and $\gamma$. For gapless alignment, the form of the distribution function is known exactly (Arratia *et al.*, 1988; Karlin and Altschul, 1990, 1993; Karlin and Dembo, 1992) in the asymptotic limit $M, N \gg 1$. For all scoring systems satisfying the condition

$$\sum_{a,b \in \chi} p(a) p(b) s(a, b) < 0 \tag{13}$$

which includes all the PAM and BLOSUM matrices, the pdf reaches the universal form

$$D(\mathsf{S}) = KMN\lambda \exp[-\lambda \mathsf{S} - KMNe^{-\lambda \mathsf{S}}], \tag{14}$$

known as the Gumbel distribution (Gumbel, 1958). This distribution is specified completely by the two parameters $\lambda$ and $K$, with a mean $\langle \mathsf{S} \rangle_0 \equiv \mathsf{S}_0 \sim \lambda^{-1} \ln KMN$ and an exponential tail

$$D(\mathsf{S} \gg \mathsf{S}_0) = \lambda KMNe^{-\lambda \mathsf{S}}, \tag{15}$$

characterized by the parameter $\lambda$.

The theory of Karlin and Altschul provides explicit formulae for these parameters in terms of the scoring function $s$. For example, $\lambda$ can be found as the unique positive root of the equation

$$\sum_{a,b \in \chi} p(a) p(b) e^{\lambda s(a,b)} = 1. \tag{16}$$

A more complicated expression exists for the calculation of $K$, which we will not describe here. We mention instead another important characteristic of the background statistics, which we will make use of later in the text. It is the average pairwise score of the optimal alignment of random sequences,

$$\alpha = \sum_{a,b\in\chi} s(a,b)p(a)p(b)e^{\lambda s(a,b)}. \tag{17}$$

This quantity, known as the "relative entropy," is needed in the calculation of $K$. It also governs the magnitude of the "finite-size" correction of $\lambda$ and $K$ from their asymptotic values; see below.

*2.2.2. Gapped alignment.* Compared to gapless alignment, the statistics of gapped alignment for the null model (11) is much more difficult to characterize. First of all, the average optimal score $\mathsf{S}_0$ does not always have the logarithmic dependence on sequence lengths. For sufficiently small gap cost, the mean score in fact acquires a *linear* dependence on sequence length even if the condition (13) is satisfied, i.e., $\mathsf{S}_0 = v \cdot N$ (for sequences of lengths $M \approx N \gg 1$), with the proportionality factor $v \geq 0$ depending on the substitution scores and gap cost. The critical line $v = 0$ defines the loci of *phase transition* points (Waterman *et al.*, 1987; Arratia and Waterman, 1994; Bundschuh and Hwa, 1999) separating the "linear" and "logarithmic" regimes of $\mathsf{S}_0$. Various statistical properties in the vicinity of this log-linear phase transition have been characterized in several recent studies (Hwa and Lassig, 1998; Drasdo *et al.*, 1998). Also, ample empirical evidences (Smith *et al.*, 1985; Collins *et al.*, 1988; Mott, 1992; Waterman and Vingron, 1994a, 1994b; Altschul and Gish, 1996; Olsen *et al.*, 1999) suggest that the optimal score $\mathsf{S}$ of gapped alignment again obeys the Gumbel distribution (14) in the logarithmic phase. However, the functional dependence of the Gumbel parameters $\lambda$ and $K$ on the scoring functions are not known.

Recently, an efficient numerical method was developed by Olsen *et al.* (1999) to characterize the tail of the Gumbel distribution, without doing exhaustive simulation, such as shuffling. The method utilizes intermediate computational results, e.g., the restricted local alignment score $H_{m,n}$, also known as the "score landscape"; see Appendix B for details. The landscape consist of a collection of positive scoring "islands," e.g., clusters of positive $H$'s, separated by a "sea" at $H = 0$. The peak scores of the islands are found to follow Poisson-like statistics, i.e., having an exponential tail for large scores. From this, the Gumbel distribution of the optimal score $\mathsf{S}$ can be derived. In particular, the Gumbel parameters $\lambda$ and $K$ can be obtained directly from the island statistics.

The study on island statistics (as reviewed in Appendix B) indicates clearly that the key to understanding the Gumbel distribution is to characterize the probability tail of obtaining a *single* large island, the statistics of which can be more conveniently studied in the context of *global* alignment. Using the saddle point method, we give (in Appendix C) a heuristic derivation of the (Poisson-like) distribution of the large island scores. The results led to the Gumbel distribution for the optimal scores, as well as the all-important Gumbel parameter $\lambda$, in terms of the solution of the equation

$$\Omega(\lambda) \equiv \lim_{N\to\infty} \langle e^{\lambda h(N)} \rangle_0 = 1, \tag{18}$$

where $h(N) = \max_{1\leq j\leq N}\{S_{1,1;j,N}, S_{1,1;N,j}\}$, in the logarithmic phase where $\langle S_{1,1;N,N}\rangle_0 < 0$ for large $N$.

The function $\Omega(\lambda)$ contains a great deal of information and is difficult to compute in general. Only recently has it been computed (Bundschuh, 1999) for a special choice of scoring functions,[3] with

$$s(a,b) = \begin{cases} 1 & \text{if } a = b \\ -2\epsilon & \text{if } a \neq b \end{cases}$$

and linear gap cost ($\delta = \varepsilon$, $\delta' = 0$), under the (weak) approximation that the scores $s(a_i, b_j)$ are uncorrelated for different $i$'s or $j$'s. The result $\lambda(\varepsilon)$ obtained in this case is in excellent agreement with extensive numerical simulation (Bundschuh, 1999) and demonstrates the validity of Equation (18). However, the

[3] This choice of scoring functions corresponds to the problem of the Longest Common Subsequences (Chavtal & Sankoff, 1975).

computation of $\Omega(\lambda)$ for *arbitrary* scoring functions remains unsolved. Along the practical side, Mott and Tribe (1999) produced an empirical formula for $\lambda$ that works reasonably well in the large gap-cost regime. Siegmund and Yakir (2000) studied a similar limit where the maximum number of gaps is finite. Despite all of these studies, the current understanding of the statistics of gapped alignment remains very limited.

## 3. PROBABILISTIC ALIGNMENT AND STATISTICS

The Smith–Waterman algorithm (9,10) is an example of an algorithm which looks for the *optimal* solution to a combinatorial problem, the solution being in this case the optimal alignment $\mathcal{A}^*$ and the optimal score $\mathsf{S}$. An alternative approach to solving combinatorial problems such as sequence alignment is to look for a class of *probable* solutions. This approach has been taken in a number of previous studies of global alignment, e.g., the maximum-likelihood method (Thorne *et al.*, 1991, 1992), the finite-temperature method (Zhang and Marr, 1995; Kschischo and Lassig, 1999), and the hidden Markov model (Holmes and Durbin, 1998). The probabilistic approach has also been used in Smith–Waterman type local alignment: In the HMM approach as implemented in the "Sequence Alignment and Modeling" software suite (Hughey and Krogh, 1996), local alignment is accomplished by embedding probabilistic global alignment in between "free insertion modules," which allows a part of a sequence to fit to the HMM. In a different approach (Bucher and Hofmann, 1996), probabilistic Smith–Waterman is realized by normalizing the probabilistic version of global alignment against a reference with substitution weights all set to 1.

The advantage of the probabilistic approach lies in the simple interpretation of the alignment parameters and results. For example, the abstract gap cost becomes a gap insertion probability, and the optimal alignment score between two sequences becomes the overall log-likelihood of the evolutionary relation between the two sequences once the alignment weights are properly normalized; see below. However, the probabilistic approach also bears distinct disadvantages. Aside from a modest computational speed disadvantage, the probabilistic approach suffers from an ill-characterized score statistics—unlike the Smith–Waterman local alignment, for which at least the form of the optimal score distribution is known for the null model, very little is known about the distribution of the log-likelihood score of the probabilistic local alignment of random sequences. Arbitrary use of the $z$-score has been shown empirically not to produce very good results (Barret *et al.*, 1997).

In this section, we will provide a brief review of the probabilistic approach to sequence alignment. We will then present in the next section an alternative "semi-probabilistic" alignment, which combines the advantages of both the optimizational and probabilistic approach to local alignment.

### *3.1. Algorithm*

We first describe the probabilistic approach to restricted global alignment. Each restricted global alignment $\hat{\mathcal{A}}$ of the subsequences $\hat{\mathbf{a}}_{m';m}$ and $\hat{\mathbf{b}}_{n';n}$ is described by an alignment path $\mathcal{R}$ as in (1). Let each pairing $(a_i, b_j)$ contribute a "weight" $w(a_i, b_j)$ towards the net weight $\mathcal{W}$ of the alignment $\hat{\mathcal{A}}$. For the gap weights, we use

$$g(\ell_1, \ell_2) = \begin{cases} 1 & \ell_1 = 0, \ell_2 = 0 \\ \mu \cdot \nu^{\ell_1 - 1} & \ell_1 \geq 1, \ell_2 = 0 \\ \mu \cdot \nu^{\ell_2 - 1} & \ell_1 = 0, \ell_2 \geq 1 \\ \mu' \cdot \mu^2 \cdot \nu^{\ell_1 + \ell_2 - 2} & \ell_1 \geq 1, \ell_2 \geq 1 \end{cases} \tag{19}$$

where $\mu$ is the weight of gap initiation, $\nu$ is the weight of gap extension, and $\mu'$ is the additional weight for the double gap configuration. In all of our numerical work below, we will use $\mu' = 1$, which treats each run of gaps the same way. However, if one wishes to exclude the double-gap configuration as considered originally by Smith and Waterman (1981), one can simply set $\mu'$ to zero.

The net weight $\mathcal{W}$ for a given configuration of pairings $\mathcal{R}$ is just the product of the individual weight factors $w$'s and $g$'s, i.e.,

$$\mathcal{W}[\mathcal{R}; \hat{\mathbf{a}}, \hat{\mathbf{b}}; w, g] = \prod_{k=1}^{l} w(a_{m_k}, b_{n_k}) \cdot \prod_{k=0}^{l} g(m_{k+1} - m_k - 1, n_{k+1} - n_k - 1), \tag{20}$$

again with $(m_0, n_0) = (m' - 1, n' - 1)$ and $(m_{l+1}, n_{l+1}) = (m + 1, n + 1)$. The total weight for the global alignment is

$$W_{m',n';m,n} = \sum_{\mathcal{R}(m',n';m,n)} \mathcal{W}[\mathcal{R}; \hat{\mathbf{a}}, \hat{\mathbf{b}}; w, g], \tag{21}$$

where $\sum_{\mathcal{R}}$ denotes the sum over all allowed paths as defined in Section 2. This weight can be computed exactly by extending the dynamic programming algorithm of Needleman and Wunsch. For the simple linear gap function ($\mu = \nu$, $\mu' = 1$) and without any constraint in the order of occurrence of insertions and deletions, one can simply iterate the recursion relation

$$W_{m',n';i,j} = w(a_i, b_j) \cdot W_{m',n';i-1,j-1} + \nu \cdot [W_{m',n';i-1,j} + W_{m',n';i,j-1}] \tag{22}$$

for $i = m'$ to $m$ and $j = n'$ to $n$, with the boundary conditions

$$W_{m',n';i\geq m'-1,j=n'-1} = \nu^{i-(m'-1)}; \; W_{m',n';i=m'-1,j\geq n'-1} = \nu^{j-(n'-1)}. \tag{23}$$

Generalization to the affine gap function (19) is given in Appendix A.

Next, we introduce the probabilistic version of the restricted *local* alignment. The total weight of the restricted local alignment of the sequences $\hat{\mathbf{a}}_{1;m}$ and $\hat{\mathbf{b}}_{1;n}$ is

$$Z_{m,n} = 1 + \sum_{m'=1}^{m} \nu^{m'} + \sum_{n'=1}^{n} \nu^{n'} + \sum_{\substack{1\leq m'\leq m \\ 1\leq n'\leq n}} W_{m',n';m,n}, \tag{24}$$

where the first term on the right-hand side is the weight of null alignment, the second and third term are the weight of aligning a subsequence of $\hat{\mathbf{a}}_{m';m}$ or $\hat{\mathbf{b}}_{n';n}$ with the null, and the last term gives the weight of aligning the subsequence $\hat{\mathbf{a}}_{m';m}$ with $\hat{\mathbf{b}}_{n';n}$, taking the weight of "skipping" the subsequences $\hat{\mathbf{a}}_{1;m'-1}$ and $\hat{\mathbf{b}}_{1;n'-1}$ to be 1. These skipping factors accomplish exactly the task of the free insertion modules used in the HMM approach to local alignment (Hughey and Krogh, 1996). Further, using the same weighting factor of 1 for skipping the subsequences $\hat{\mathbf{a}}_{m+1;M}$ and $\hat{\mathbf{b}}_{n+1;N}$, the total weight of the local alignment between the sequences **a** and **b** becomes simply

$$\mathsf{W}[\mathbf{a}, \mathbf{b}; w, g] = 1 + \sum_{\substack{1\leq m\leq M \\ 1\leq n\leq N}} Z_{m,n} - M \cdot N, \tag{25}$$

with the last term accounting for the $M \cdot N$ redundant counts of the null alignment included in the second term. The recursion relation for $Z_{m,n}$ itself is straightforward to derive given (22), (23), and the definition (24). One finds

$$Z_{m,n} = 1 + w(a_m, b_n) \cdot Z_{m-1,n-1} + \nu \cdot [Z_{m-1,n} + Z_{m,n-1}], \tag{26}$$

with the boundary condition $Z_{0,n} = Z_{n,0} = 1$. Equations (25) and (26) define the algorithm for the probabilistic version of local alignment.

### 3.2. *Likelihood and hidden Markov model*

So far, the alignment weight $\mathcal{W}$, computed according to Equation (20) with arbitrary weight parameters $w$ and $g$, does not have any meaningful interpretation. It becomes meaningful, however, if the sum of the weights going into and out of *every* node on the alignment lattice are *equal* on average. For the linear gap problem (22), this is satisfied if the weight parameters in the bulk of the alignment lattice are chosen to obey the condition

$$\langle w(a, b)\rangle_0 + 2\nu = 1, \tag{27}$$

where $\langle \ldots \rangle_0$ denotes average over the null distribution (11) as before. Our generalization of this formula to the affine-gap function (19) is[4]

$$\langle w(a,b)\rangle_0 = \frac{(1-\nu)^2}{(1+\mu-\nu)^2+(\mu'-1)\cdot\mu^2}; \tag{28}$$

see Appendix A.2 for details. The above conditions can be satisfied by choosing the substitution weights as $w(a,b) = (1-2\nu)\mathcal{T}(b|a)/p(b)$ for the linear gap problem or

$$w(a,b) = \frac{(1-\nu)^2}{(1+\mu-\nu)^2+(\mu'-1)\mu^2}\cdot\frac{\mathcal{T}(b|a)}{p(b)} \tag{29}$$

for the affine gap function (19), where $\mathcal{T}(b|a)$ gives the transition probability of amino acid $b$ from $a$, as what was used to generate Dayhoff's PAM substitution matrices (Dayhoff *et al.*, 1978). As is true for any transition matrices, $\mathcal{T}$ has the property $\sum_{b\in\chi}\mathcal{T}(b|a) = 1$, and hence $\langle \mathcal{T}(b|a)/p(b)\rangle_0 = 1$.

Given the local conservation condition (27) or (28), the weight going into and out of any given region must also be conserved on average. For the example shown in Fig. 1(b), where the total weight going into the region is 1 [as specified by the boundary condition (23)], the total average weight $\tilde{W}_{m',n';m,n}$ leaving the region defined by the "forward boundaries" at $i = m$ and $j = n$ (the dashed lines in Fig. 1(b)) must also be 1 on average. This weight is easily computed in terms of the $W$'s. For the linear gap function, it reads

$$\tilde{W}_{m',n';m,n} \equiv [2\nu + w(a_m,b_n)]\cdot W_{m',n';m-1,n-1} + \sum_{i=m'}^{m-1}[\nu + w(a_i,b_n)]\cdot W_{m',n';i-1,n-1}$$
$$+ \sum_{j=n'}^{n-1}[\nu + w(a_m,b_j)]\cdot W_{m',n';m-1,j-1}. \tag{30}$$

The analogous quantity in the affine-gap case is given by Equation (117) in Appendix D. Generally, we can interpret $\tilde{W}_{m',n';m,n}$ at the total weight that the alignment path starts from the fixed end at $(m',n')$ and ends *anywhere* as it intersects the forward boundaries. The local conservation condition assures that

$$\langle \tilde{W}_{m',n';m,n}\rangle_0 = 1 \qquad \text{for all } m' \le m \text{ and } n' \le n. \tag{31}$$

With the existence of a conservation law for the average weight, one can interpret $\mathcal{W}[\mathcal{R};\mathbf{a},\mathbf{b};w,g]$ as the "likelihood" that the sequence **a** "evolved" into the sequence **b** according to some mutation probabilities specified by $w$ and $g$. In Appendix D, an example of such an evolution process is given in terms of a hidden Markov model, which takes a random sequence generated according to the background frequency $p(a)$, replaces element $a$ by $b$ according to the transition probability $T_c(b|a)$, and makes insertion/deletions of segments of lengths $\ell_1$, $\ell_2$ with probability $g_c(\ell_1,\ell_2)$ as specified by (19) but with parameters $\mu_c$, $\nu_c$, $\mu'_c$. We show in Appendix D that the model generates *correlated* sequences with statistics described by the joint distribution function

$$P_c[\mathbf{a},\mathbf{b};w_c,g_c] = \tilde{W}[\mathbf{a},\mathbf{b};w_c,g_c]\cdot P_0[\mathbf{a},\mathbf{b}], \tag{32}$$

with $w_c(a,b) = [(1-\nu_c)^2/((1+\mu_c-\nu_c)^2+(\mu'_c-1)\mu_c^2)]\cdot[T_c(b|a)/p(b)]$. The weights $w_c$ and $g_c$ satisfy the conservation condition (28) for affine gap functions, since the transition probability has the property

[4] Note that in the limit of the linear gap function ($\mu = \nu$ and $\mu' = 1$), Equation (28) becomes $\langle w(a,b)\rangle_0 = (1-\nu)^2$ which is different from (27). This occurs because we have excluded, via employing asymmetric recursion relations (47) and (54), the multitudes of pairing configurations involving alternating series of insertions and deletions. The asymmetry makes no difference in the Viterbi algorithm, but it is necessary in the probabilistic algorithms to avoid overcounting.

$\sum_{b\in\chi} T_c(b|a) = 1$. In the context of the hidden Markov model, we see that the average conservation law (31) can be viewed simply as the normalization condition on $P_c[\mathbf{a}, \mathbf{b}]$, i.e., $\sum_{[\mathbf{a},\mathbf{b}]} P_c[\mathbf{a}, \mathbf{b}] = \langle \tilde{W} \rangle_0 = 1$. The relation (32) will prove to be very useful in the sequel.

### 3.3. Score statistics

We now turn our attention to the main subject of this study, the statistics of probabilistic alignment, as described by the distribution of the log-likelihood score (Eddy *et al.*, 1995), e.g., pdf(ln W). This statistics is not well understood, even in comparison to the not-so-well-understood Smith–Waterman score statistics: Previously, an exponential bound on the tail of the log-odd score distribution was obtained for a simple sequence analysis problem (Milosavljevic and Jurka, 1993). Barret *et al.* (1997) applied this bound to the HMM version of local alignment and found empirically that it did not correctly account for the false positives observed. In the mean time, there is a general expectation that the distribution of ln W might still have a Gumbel form. Here, we would like to point out that there is in fact no a priori reason to expect a Gumbel distribution for log-odd scores generated by probabilistic algorithms. This is because W, computed according to Equation (25), is a *sum* of a large number of *correlated* terms, while the Gumbel distribution is typically obtained from taking the maximum of a large number of uncorrelated terms (see Appendix B.1). In the numerical study described below, we will show that the statistics of ln W is indeed *not* of the Gumbel form.

We performed a large number of probabilistic alignments of random sequences using the affine gap weights (19) specified by the two parameters $\mu$ and $\nu$ (with $\mu'$ set to 1). For the substitution weights, we use Equation (29) with

$$\mathcal{T}(b|a) = T^d(b|a) \tag{33}$$

where $T(b|a)$ is the 20 × 20 unit PAM transition matrix for 1% mutation (obtained from the NCBI website), $d$ is the so-called PAM distance, and $p(b) = T^{d\to\infty}(b|a)$ is the background amino acid distribution used to generate random sequences. The special choice of substitution weights (29) with (33) satisfies the conservation condition for the affine gap function (28); see Appendix A.2 for details. The actual algorithm used is given by Equations (54) and (55) of Appendix A, and the total weight W is obtained from Equation (25).

We tested two sets of scoring parameters described by $d = 120$, $\mu = 2^{-5.5}$, $\nu = 2^{-0.5}$ and $d = 250$, $\mu = 2^{-6}$, $\nu = 2^{-0.5}$. For brevity, we shall refer to the first set of parameters as "PAM-120" and the second set as "PM-250" scoring functions. Figs. 2(a) and (b) show respectively the distribution of ln W for the

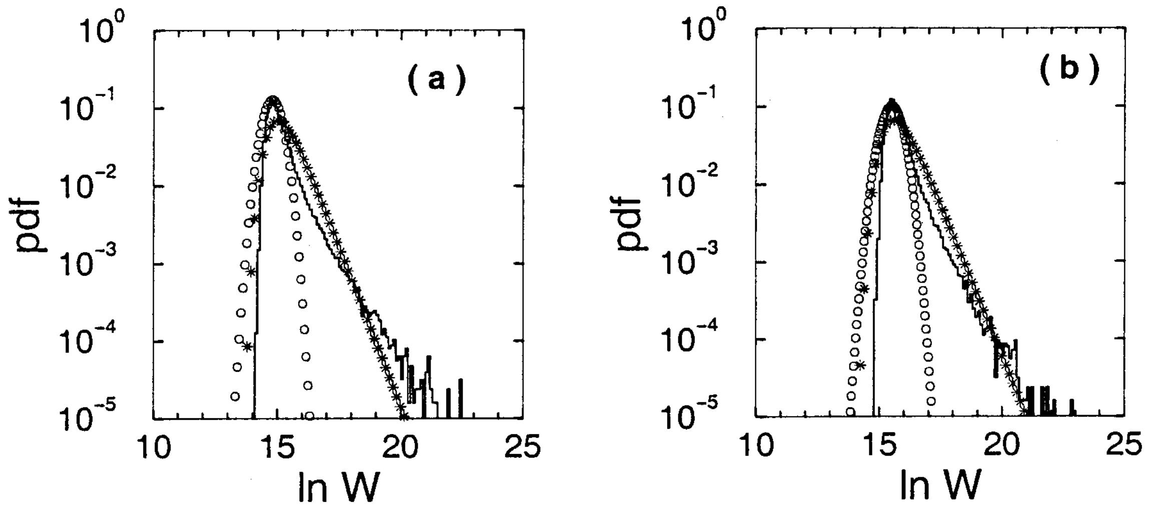


FIG. 2. Pdf's of ln W for full-probabilistic alignment. The pdf's of ln W for the two parameter sets, (a) PAM-120 and (b) PAM-250, are chosen to satisfy the average probability conservation condition (28). Each pdf (shown as a staircase) is obtained by normalizing the histogram (collected for 125,000 pairwise alignments of random sequences of length 300). The Gaussian fits are shown by open circles while the Gumbel fits are shown by star symbols.

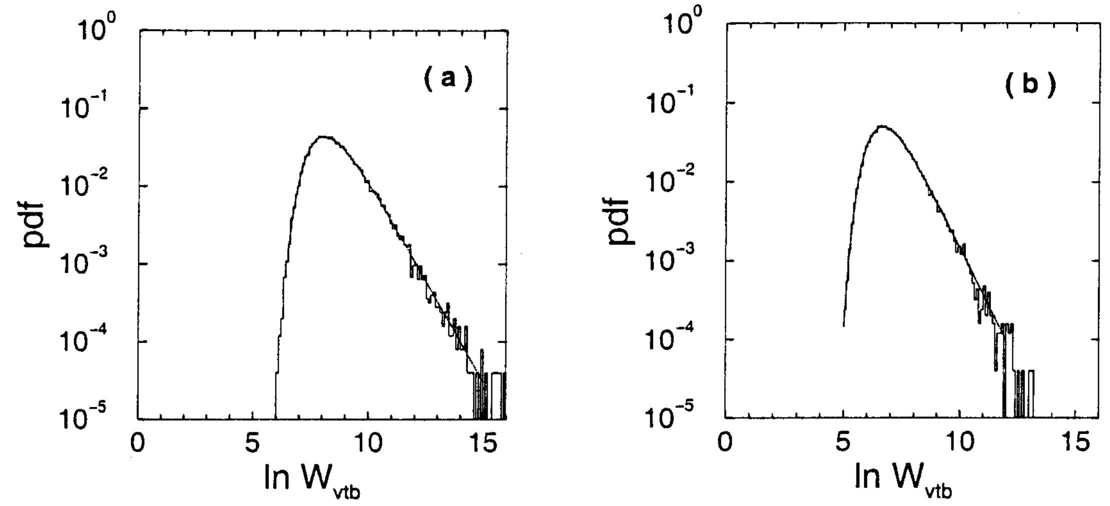


**FIG. 3.** Pdf's for the Viterbi score $\ln W_{\text{vtb}}$. The pdf's for the Viterbi score $\ln W_{\text{vtb}}$ for the two parameter sets, (**a**) PAM-120 and (**b**) PAM-250, are chosen to satisfy the conservation condition (28). Each pdf (drawn as staircase) is obtained from normalizing the histogram of 25,000 pairwise alignments of random sequences of length 300. The lines are least-square fits to the Gumbel distribution.

PAM-120 and PAM-250 scoring functions for 125,000 alignments of Markov sequence pairs of length 300 each. From the figures, it is clear that the tails of the distribution functions are neither exponential nor Gaussian. The best least-square fits to the Gumbel and Gaussian distributions, shown as the dashed line and long-dashed lines, respectively, are not satisfactory at all.

We then repeated the above alignments using the Viterbi algorithm, which keeps only the alignment path with the maximum weight $\mathcal{W}$. It is essentially the Smith–Waterman algorithm with affine gaps, as specified in Equation (45) of Appendix A, with the gap costs[5] $\delta = \ln \mu$, $\delta' = 0$, and $\varepsilon = \ln \nu$. However the substitution score of the Viterbi algorithm is shifted from the PAM score $s_d \equiv \ln[T^d(b|a)/p(b)]$ by an amount

$$\Delta s(\mu, \nu) = \ln \frac{(1-\nu)^2}{(1+\mu-\nu)^2} < 0 \qquad (34)$$

due to the multiplicative term in Equation (29) (with $\mu' = 1$). Thus, $s(a, b) = \Delta s(\mu, \nu) + s_d(a, b)$ for the Viterbi algorithm.

As expected for alignments of the Smith–Waterman type, the pdf's of the Viterbi score $\ln W_{\text{vtb}}$ are well described by the Gumbel distribution (14); see Figs. 3(a) and (b). The values of $\lambda$ as obtained from the least-square fit[6] are shown in Fig. 4(a) for various sequence lengths $N$. The $\lambda$'s have significant sequence length dependence; also, their asymptotic values are different for the different scoring systems used. For comparison, we plot in Fig. 4(b) the $\lambda$'s obtained from the straight Smith–Waterman alignment with $\Delta s = 0$. Strong length and scoring function dependences are found for this case also, as expected. There is an overall increase in $\lambda$ (see Figs. 3(a) and (b)) when going from Smith–Waterman to the Viterbi alignment. This results from the negative shift $\Delta s$, which pushes the Viterbi further away from the log-linear phase transition line of this system. However, as already stated in Section 2, there is so far no detailed understanding of these parameter dependences for alignments of the Smith–Waterman type.

[5] On the scale of the PAM scoring system used by BLAST, where the substitution scores are defined as $s_d(a, b) = 2 \cdot \log_2[T^d(b|a)/p(b)]$ for PAM-120 and $s_d(a, b) = 3 \cdot \log_2[T^d(b|a)/p(b)]$ for PAM-250, our values of $\mu$ and $\nu$ translate to gap costs of $\delta = 11$, $\varepsilon = 1$, for the PAM-120 scoring system and $\delta = 18$, $\varepsilon = 1.5$, for the PAM-250 scoring system.

[6] The statistical error of $\lambda$ was estimated in the following way: For a given sequence length, we ran 125,000 pair-wise alignments of random sequences. We then divide the data into five sets of 25,000 alignments each and fit each set to the Gumbel distribution to obtain five $\lambda$ values. We then take twice the maximum difference between these $\lambda$'s as the size of the error bar.

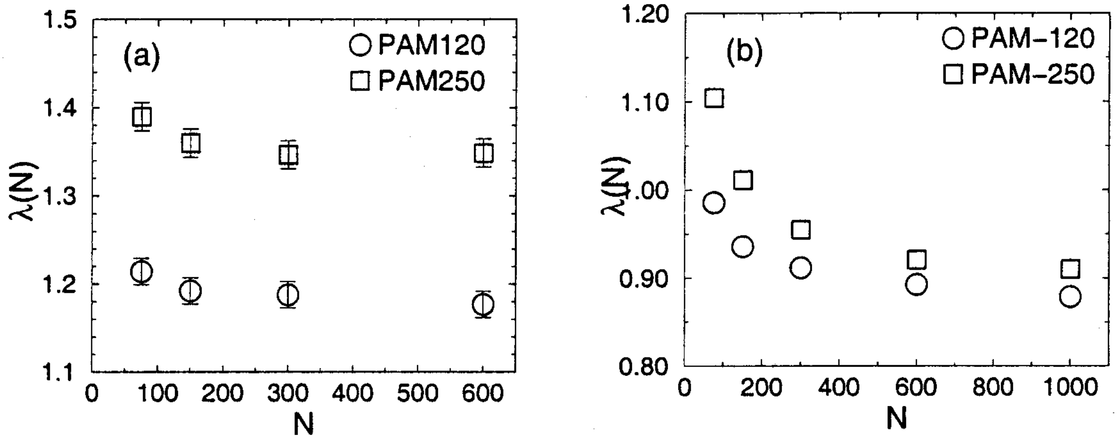


**FIG. 4.** The Gumble parameter λ versus system size. The λ values at different sequence-pair lengths are shown for the two parameter sets, PAM-120 and PAM-250, for (**a**) the Viterbi version of the probabilistic alignment and (**b**) the Smith–Waterman alignment without the shift (34) to the PAM substitution scores. It is evident that the values of λ depend on the scoring functions and sequence lengths in nontrivial ways.

## 4. SEMI-PROBABILISTIC LOCAL ALIGNMENT

In this section, we present a slight modification of the probabilistic local alignment, at no extra cost to computational complexity compared to the fully probabilistic version discussed in Section 3. This modification will allow us to develop a theory for the score statistics of the resulting alignments.

We will take the total weight $Z_{m,n}$ for the restricted local alignment as defined in Equation (24) and computed according to the recursion relation (26) for linear gap function or according to (54) and (55) for the affine gap function. However, instead of obtaining the total weight W for the probabilistic local alignment by summing over all the $Z$'s as in (25), we follow the optimizational approach (10) and construct the maximum log-likelihood (MLL) score

$$\Phi[\mathbf{a}, \mathbf{b}; w, g] = \max_{\substack{1\le m\le M \\ 1\le n\le N}} \{\ln Z_{m,n}\}. \tag{35}$$

The MLL score is manifestly a *hybrid* of both the probabilistic and optimizational approaches to local alignment. We refer to alignment based on the MLL score as "semi-probabilistic alignment" and refer to this algorithm as the "hybrid algorithm."

As we will see, the advantage of the hybrid algorithm is that the MLL score statistics can be much better characterized than both the log-likelihood score ln W of the probabilistic approach and the optimal score S of the optimizational approach, without sacrificing the sensitivity of the alignment. Just as for the statistics of Smith–Waterman alignment discussed in Section 2, the statistics of the MLL score can be characterized by studying the corresponding *global* alignment problem defined by (22) or its affine gap version (47). In particular, the statistical properties of the log-likelihood score ln $W_{m',n';m,n}$ for probabilistic global alignment is very much analogous to the properties of the optimal global alignment score $S_{m',n';m,n}$. Repeating the considerations based on islands and score profiles (Appendix B) and the saddle point calculation (Appendix C) for probabilistic global alignment, we again find Poisson-like statistics for the islands, implying the Gumbel distribution for the MLL score. The corresponding Gumbel parameter λ can again be obtained by solving an equation similar to (18) with $h(N) = \ln \tilde{W}_{1,1;N,N}$, e.g.,

$$\lim_{N\to\infty} \langle [\tilde{W}_{1,1;N,N}(w,g)]^{\lambda}\rangle_0 = 1 \tag{36}$$

with $\lambda > 0$. A simple expression is also available for the relative entropy,

$$\alpha = \lim_{N\to\infty} N^{-1}\langle \ln \tilde{W}(w,g)\cdot[\tilde{W}(w,g)]^{\lambda}\rangle_0. \tag{37}$$

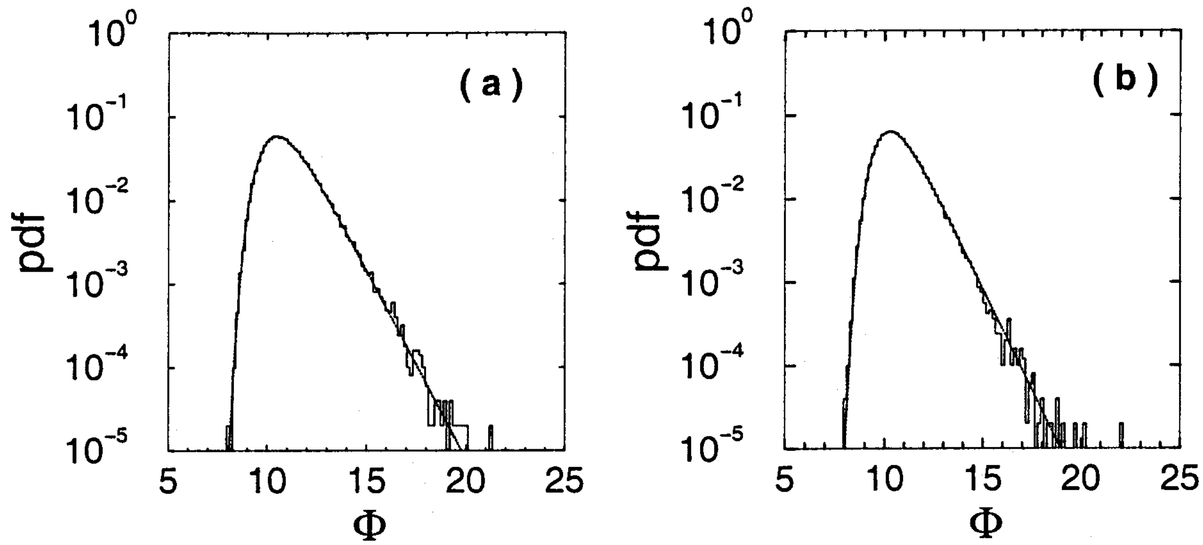


**FIG. 5.** Pdf's for the semi-probabilistic alignment of random sequences. The pdf's are obtained by normalizing histograms of 50,000 pairs of random sequences of length 300 each, using the two parameter sets, (**a**) PAM-120 and (**b**) PAM-250, chosen to satisfy the conservation condition (28).

### *4.1. The magic λ*

Equation (36) is difficult to solve in general, just like its Smith–Waterman counterpart (18). However, for semi-probabilistic alignment, there does exist a special solution of (36) which we will exploit here. We note that for the special value of $\lambda = 1$, the condition (36) reduces to $\langle \tilde{W} \rangle_0 = 1$. But the latter condition is just (31), which we already found to hold if the individual weight parameters satisfy the average probability conservation condition, e.g., Eq. (27) for linear gaps or (28) for affine gaps. Since equations of the form (36) admit only one positive solution of λ, we expect that *as long as one chooses weight parameters respecting the conservation condition, then the MLL score of the semi-probabilistic alignment will always have Gumbel statistics with* $\lambda = 1$, in the asymptotic limit of long sequence lengths $M$, $N \gg 1$. This is a very powerful result, analogous to Equation (16) of gapless alignment, which fixes $\lambda = 1$ for any choice of log-odd substitution score $s(a, b)$.

We next examine numerically the validity of the prediction that a) the MLL score of the hybrid algorithm obeys Gumbel statistics and b) $\lambda = 1$ for weight parameters satisfying the conservation condition. We again use the two sets of scoring systems PAM-120 and PAM-250 which satisfy the conservation condition (28) for affine gaps. We use the affine gap algorithm (54) and (55) described in Appendix A and calculate the MLL score using (35). Figures 5(a) and (b) show the pdf's of Φ obtained from the alignment of 50,000 pairs of random sequences of lengths 300 each, generated according to the null model (11). We see that the pdf's are well fitted by the Gumbel distribution (14). This should be compared to the fully probabilistic alignment, whose score ln W exhibits anomalous statistics as shown in Fig. 2.

We plotted in Fig. 6(a) the λ values obtained from fitting the pdf's for sequences of different lengths $N$ to the Gumbel form (14). (For the sake of completeness, we also included in Fig. 6(b) the values of the other Gumbel parameter $K$ obtained from the same fit. We shall, however, be focused mostly on λ in the following.) Unlike the Viterbi case described earlier, where the λ's for the PAM-120 and PAM-250 scoring systems approach different values upon increasing $N$ as shown in Fig. 4(a), here we see clearly a tendency for the λ's of both systems to converge towards a common value close to 1, as expected from the above theoretical considerations.[7] However, the rates of convergence to the asymptotic values are different for the two scoring systems. In order to test more quantitatively our prediction that the asymptotic value is indeed $\lambda = 1$, it is necessary to characterize precisely the finite-size correction to λ.

[7] We have independently verified that, away from the probability conservation condition (28), the pdf of Φ is still of the Gumbel form, but the asymptotic values of λ can deviate significantly from 1.

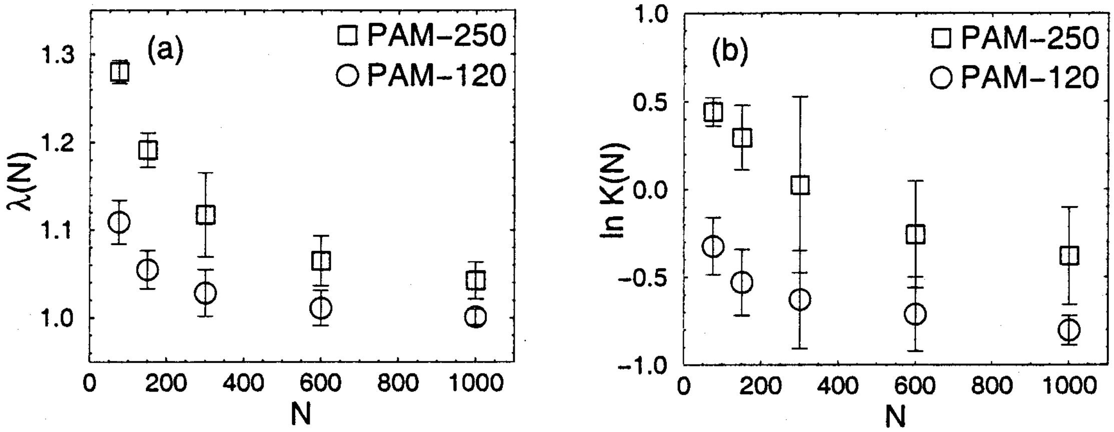


**FIG. 6.** Gumbel parameters versus system size for semi-probabilistic alignments. The values of (**a**) $\lambda$ and (**b**) $K$ are obatined from least-square fits of pdf's, such as those in Fig. 5 to the Gumbel form (14), for different sequence lengths. The circles and squares correspond to the PAM-120 and PAM-250 scoring systems, respectively, chosen to satisfy the conservation condition (28).

### *4.2. The finite-size correction*

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 (14) for finite length sequences, one should "correct" the lengths $M$ and $N$ which appear in (14) by a score-dependent amount $L(\mathsf{S})$ and use instead the *effective* sequence lengths $M' = M - L(\mathsf{S})$ and $N' = N - L(\mathsf{S})$. The origin of the correction term is easy to see from the score landscape picture described in Appendix B: It results from the fact that the available area on the alignment lattice to launch an island is *reduced* by the size of the island itself. 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[8] $\overline{\ell}(\mathsf{S})$ corresponding to the score of the maximum island peak $\mathsf{S}$; see Appendix C. Including this correction, the Gumbel statistics becomes

$$\Pr(\mathsf{S} < x) = \exp[-K \cdot (N - \overline{\ell}(x))^2 e^{-\lambda x}], \tag{38}$$

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 $\alpha$ is the "relative entropy," Altschul *et al.* (2000) noted that the terms in Equation (38) can be rearranged into the classic Gumbel form, i.e.,

$$\Pr(\mathsf{S} < x) = \exp[-K(N) \cdot N^2 e^{-\lambda(N)\cdot x}], \tag{39}$$

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

$$\lambda(N) = \lambda + 2/\overline{\sigma}(N), \tag{40}$$

where $\overline{\sigma}(\ell)$ is the inverse of the function $\overline{\ell}(\sigma)$ and gives the average score for islands of length $\ell$. Note that finite-size correction formulae such as (40) 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.

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

*4.2.1. The relative entropy.* To utilize the finite-size correction formulae, we need to obtain the relative entropy $\alpha$, and also, more generally, the subleading terms. Direct measurement of $\alpha$, however, is time consuming even by the island method because the occurrences of large islands are rare. However, great simplification takes place at the magic $\lambda$ value, and the expression (37) for $\alpha$ can be evaluated directly for *arbitrary* scoring parameters satisfying the conservation condition (28).

By Equation (32), the right-hand side of (37) can be related to the *correlated* ensemble $P_c$, i.e.,

$$\alpha = \lim_{N\to\infty} N^{-1} \sum_{[\mathbf{a},\mathbf{b}]} \ln \tilde{W}_{1,1;N,N}[\mathbf{a}, \mathbf{b}; w, g] P_c[\mathbf{a}, \mathbf{b}; w, g]. \tag{41}$$

This equation suggests a very simple recipe for the computation of $\alpha$: First, generate the correlated sequences **a** and **b** using the evolution model in Appendix D, taking the transition probability to be $T_c(b|a) = T^d(b|a)$, and indel probabilities as described by $\mu_c = \mu$, $\nu_c = \nu$, and $\mu'_c = \mu'$. Then, apply the probabilistic *global* alignment (47) to these correlated sequences using the *same* scoring parameters and compute $\tilde{W}$ for this sequence pair using Equation (117). Quantity $\alpha$ is readily obtained by averaging over the ensemble of correlated sequences generated according to this procedure.

This approach can be extended also to the computation of other quantities. We note that Equation (41) relates $\alpha$, which is a property of the rare large islands, to the average property of the corresponding quantity in the correlated ensemble $P_c$. In other words, *typical sequences generated according to $P_c$ mimic those rare subsequences of the random sequence ensemble that gives rise to the large islands.* This correspondence can therefore be exploited to measure other properties of the large islands. For example, the average island peak score $\overline{\sigma}(N)$ can be obtained as

$$\overline{\sigma}(N) = \langle \ln \tilde{W}_{1,1;N,N} \rangle_c \tag{42}$$

where $\langle \ldots \rangle_c$ refers to the average over the "correlated ensemble" weighted by $P_c$. In this way, Equation (41) becomes merely $\alpha = \lim_{N\to\infty} \overline{\sigma}(N)/N$, which is the definition of the relative entropy.

We measured $\overline{\sigma}(N)$ using the correlated ensemble as described above. The result corresponding to the PAM-120 and PAM-250 scoring systems are shown in Fig. 7. They are well fitted by the form $\overline{\sigma} = \alpha N + c$, with statistical uncertainties in $\alpha$ and $c$ well under 1%. The most striking thing about this result is that the data points in Fig. 7 were averaged over only 15 pairs of alignments and took practically no time to generate, while determining $\alpha$ to such precision using direct simulation or island counting will take weeks on the same computer.

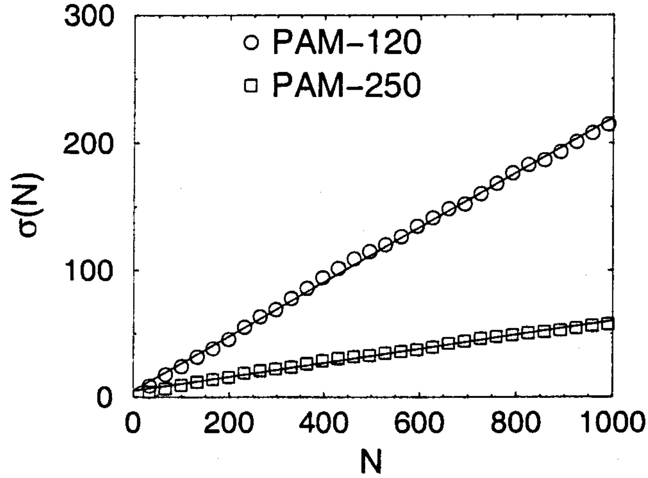


**FIG. 7.** Fast assessment of relative entropy. The circles and squares represent the alignment score of correlated sequences averaged over only 15 pairwise alignments corresponding to the PAM-120 and PAM-250 scoring systems, respectively. The lines represent the respective least-square fits to $\overline{\sigma}(N) = \alpha N + c$. The fits, which are excellent down to $N = 50$, give $\alpha = 0.0554$ and $c = 4.85$ for PAM-250 and give $\alpha = 0.2144$ and $c = 5.22$ for PAM-120.

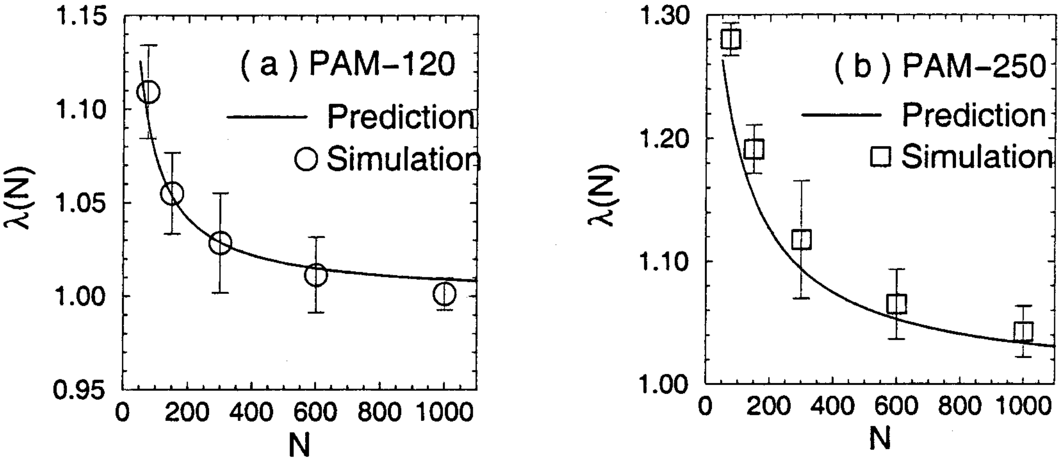


**FIG. 8.** Finite size dependence of Gumbel parameter λ. Direct comparisons of the numerical values of λ as shown in Fig. 6(a) and the theoretical prediction for (**a**) PAM-120 and (**b**) PAM-250 scoring systems.

*4.2.2. Direct comparisons.* With the accurate determination of $\overline{\sigma}(N)$, we are now in a position to test the finite-size prediction (40), 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 Equation (40) is plotted as the line in Figs. 8(a) and (b) for the PAM-120 and PAM-250 scoring systems respectively. Also plotted are the data points shown already in Fig. 6(a). 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.

It is also possible to extend the finite-size analysis discussed above for the parameter $K$. Using the form $\overline{\ell}(x) = \alpha x + c$ in Equation (38) and rearranging terms into the Gumbel form Equation (39), we find the result

$$K(N) = K \cdot \left(1 + \frac{c}{\alpha N}\right)^2 \tag{43}$$

which is analogous to Equation (40) for $\lambda(N)$. Unlike the case for $\lambda$, 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 finite-

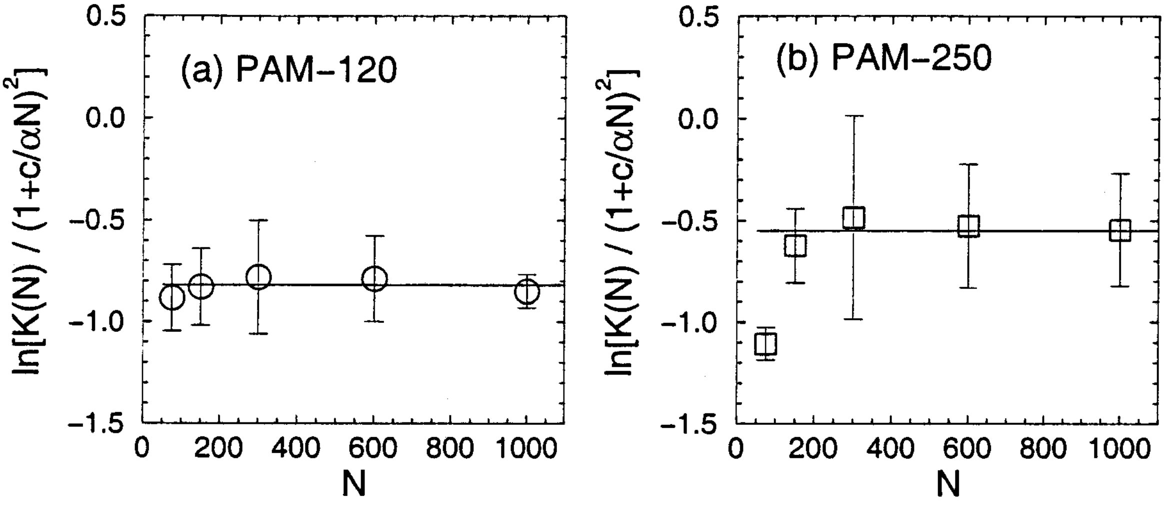


**FIG. 9.** Finite size dependence of Gumbel parameter $K(N)$. For (**a**) PAM-120 and (**b**) PAM-250 scoring systems, the horizontal lines indicate the values of the asymptotic $K$'s obtained from data at larger $N$'s (not shown).

size correction formula (43) using the numerically obtained values of $K(N)$ shown in Fig. 6(b). To do so, we simply divide the vertical axis of Fig. 6(b) by the factor $(1 + c/(\alpha N))^2$, using the values of $c$ and $\alpha$ determined from Fig. 7 for the corresponding scoring system. According to Equation (43), 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. 9(a) and (b). Other than the smallest size of $N = 75$, the data points are approximately[9] $N$-independent, hovering around the asymptotic values indicated by the horizontal lines. The results suggest that Equation (43) does capture the correct finite-size effect. 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 effective $K$ for all other $N$'s can be deduced from the finite size formula (43).

## 5. HOMOLOGY DETECTION

Characterization of the score distribution is only a step towards the final goal of sequence alignment: homology detection. Even though we have shown that the semi-probabilistic alignment of random sequences yields scores obeying a well-characterized distribution at the probability conservation condition, the result will not be of much use in practice if it gives low scores to homologous sequences, i.e., if it has low sensitivity. In this section, we examine the sensitivity of homology detection by the semi-probabilistic alignment and compare it to the usual Smith–Waterman alignment and the Viterbi version of the full-probabilistic alignment at the probability conservation condition. As mentioned already, the Viterbi and Smith–Waterman differ only by a shift $\Delta s(\mu, \nu)$ in the substitution scores used; see Equation (34). We will not include the full probabilistic alignment itself in our comparison because its statistics is not well understood (even numerically). Furthermore, since the tail of the pdf for $\ln \mathsf{W}$ is broader than the exponential (see Fig. 2), we expect that it will give a larger p-value (and hence a smaller significance) than the Viterbi for the high-scoring alignments.

Ideally, we want to perform the detection comparison by using related biological sequences, e.g., those from the database SwissProt. As a first step, we will use correlated sequences generated from the toy evolution model described in Appendix D. These should be sufficient for our purpose of evaluating qualitatively whether the semi-probabilistic alignment might suffer a substantial loss in sensitivity compared to the standard methods.

We generated correlated sequences according to the model of Appendix D, using substitution probabilities given by $T^{d_c}(b|a)$ and indel probabilities given by $\mu_c$, $\nu_c$, and $\mu'_c = 1$. Two sets of 1,500 correlated sequence pairs of lengths 300 were generated, corresponding to the mutation parameters $\mu_c = 2^{-6}$, $\nu_c = 2^{-0.5}$, $d_c = 250$, and $\mu_c = 2^{-5.5}$, $\nu_c = 2^{-0.5}$, $d_c = 120$. Each of these correlated sequence pairs is then aligned using the hybrid, Smith–Waterman, and Viterbi algorithms, each for a set of a dozen or so scoring parameters characterized by different PAM distance $d$'s and initiation costs ($\delta$ for Smith–Waterman, or $\mu$ for the hybrid or Viterbi algorithms). From these alignments, we obtain the average correlated score $\mathsf{S}_c$ for each of the scoring parameter settings. Next, we align 35,000 pairs of random sequences to determine the Gumbel parameters $\lambda$ and $K$ for each of the dozen or so scoring parameter settings used. The significance $\sum$ of an alignment is then determined by integrating the tail of the Gumbel distribution (15) to the score $\mathsf{S}_c$. In "bits", it is, i.e.,

$$\sum \equiv -\log_2(KN^2 e^{-\lambda \mathsf{S}_c}). \tag{44}$$

In Fig. 10, we plot the significance $\sum$ obtained from the different scoring parameters for correlated sequences generated with $\mu_c = 2^{-5.5}$, $\nu_c = 2^{-0.5}$, $d_c = 120$. Fig. 10(a) shows the dependence on the gap initiation parameter $\mu$ for the probabilistic (or $\delta = 2\log_2 \mu$ for Smith–Waterman) methods, with PAM distance fixed at $d = d_c = 120$ and gap extension parameter fixed at $\nu = 2^{0.5}$ (or $\epsilon = 2\log_2 \nu$ for

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

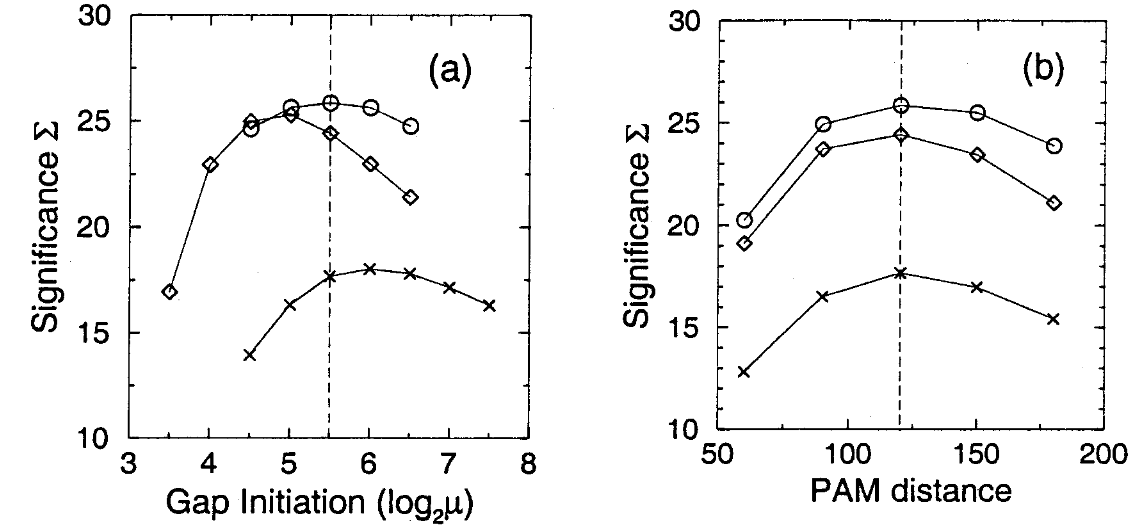


**FIG. 10.** Comparison of significance estimate of the hybrid algorithm (circles), the Smith–Waterman algorithm (diamonds), and the Viterbi version of the full-probabilistic method (crosses). The significance $\sum$ is derived from Eq. (44), using the average score obtained from 1,500 correlated sequence pairs as $S_c$. The correlated sequences are generated from the toy model of Appendix D, with mutation parameters $\mu_c = 2^{-5.5}$, $\nu_c = 2^{-0.5}$, $d_c = 120$. In (**a**), the gap extension cost is fixed at $\nu = \nu_c$ (or $\varepsilon = 2\log_2 \nu_c$ for the Smith–Waterman alignment) and the substitution score is fixed at $d = d_c$, i.e., at PAM-120. In (**b**), the gap extension cost is still fixed at $\nu = \nu_c$; the gap initiation cost is also fixed at $\mu = \mu_c$ (or $\delta = 2\log_2 \mu_c$ for the Smith–Waterman alignment).

Smith–Waterman.) In Fig. 10(b), we instead fix the gap initiation to $\mu = \mu_c = 2^{-5.5}$ and vary the PAM distance $d$. We see that the performance of the hybrid algorithm is comparable to Smith–Waterman and is significantly better than Viterbi. The same plots are repeated in Figs. 11(a) and (b) for the sequences generated with the other set of mutation parameters ($\mu_c = 2^{-6}$, $\nu_c = 2^{-0.5}$, $d_c = 250$). In this case, there is an overall degradation of the significance compared to the ones shown in Fig. 10, as the mutation is much stronger than in the case with $d_c = 120$. However, the relative performances of the three methods remain the same, with the hybrid method being comparable to Smith–Waterman and significantly better than Viterbi.

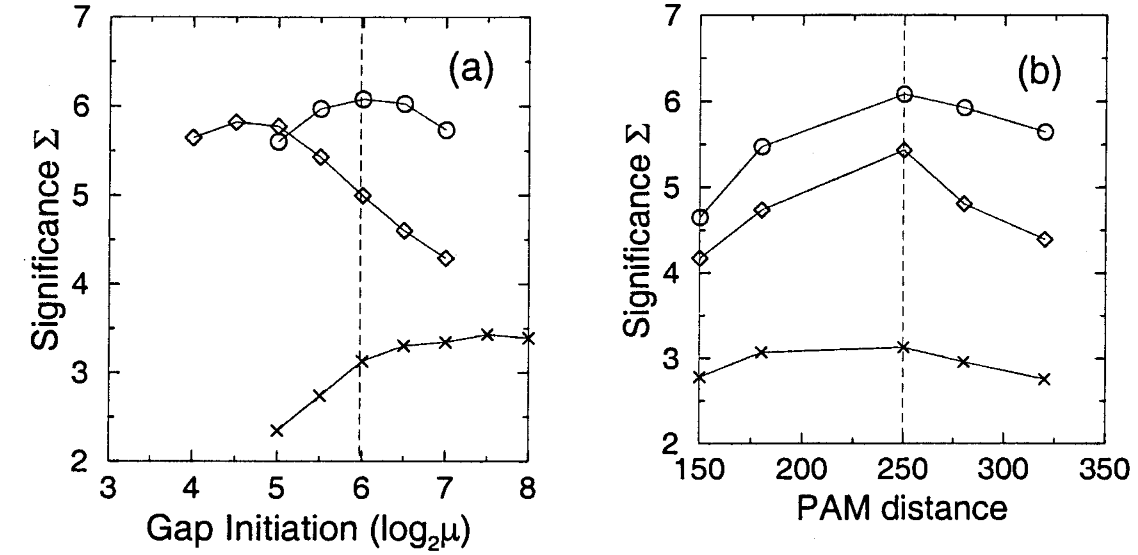


**FIG. 11.** Same comparison as in Fig. 10, except the correlated sequences are generated with the mutation parameters $\mu_c = 2^{-6}$, $\nu_c = 2^{-0.5}$, $d_c = 250$.

We should caution that conclusions obtained from this simple sensitivity comparison is based on very crude homology models. It should not be too surprising that the hybrid method performed well, since it is the closest to the generative model. Also, we do not expect that Smith–Waterman alignment will always outperform the Viterbi, even within the class of correlated sequences generated by our mutation model. In particular, if the degree of homology between the sequence pairs is much higher, then the Viterbi is likely to outperform Smith–Waterman because its larger $\lambda$ value will make the significance $\sum$ larger. Nevertheless, we can conclude from this comparison that the hybrid method does not suffer any unexpected generic problems which would make it significantly less sensitive. Detailed studies on biological sequences are obviously needed to determine the actual degree of sensitivity of the hybrid method.

## 6. SUMMARY AND OUTLOOK

In this paper, we studied the extremal statistics of probabilistic sequence alignment both analytically and numerically. We find that while the straightforward probabilistic alignment gives rise to anomalous score statistics, the slightly modified semi-probabilistic alignment is well described by Gumbel statistics. For the semi-probabilistic alignment, we can predict the Gumbel parameter $\lambda$, including its finite size dependence, for different scoring functions and parameters. Our results are verified numerically by using various PAM substitution matrices and affine gap functions. We further studied the sensitivity of the new hybrid method by aligning correlated sequences generated from toy mutation models. We find the sensitivity to be comparable to that of the Smith–Waterman alignment and significantly better than the Viterbi version of probabilistic alignment.

In our study, we have not focused on the behavior of the other Gumbel parameter, $K$, which is more difficult to compute than $\lambda$ analytically. It should, however, be straightforward to determine $K$ numerically by extending the island method of Olsen *et al.* (1999) to the semi-probabilistic alignment. With the help of precisely determined $\lambda$ and the finite-size correction formula for $K$, it is possible to fix the value of $K$ for all sequence lengths by counting islands from a few pairwise alignments (R. Olsen, private communication, 1999).

Let us close with a general remark: Although we restricted the numerics presented in the present study 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 $\lambda$ to remain 1 as long as the probability conservation condition, such as (27) or (28), is *locally satisfied* at each node of the alignment lattice. This can be readily accomplished for position-specific substitution and indel weights by generalizing the formula (29). Similarly, our formula for the relative entropy $\alpha$ is good also for position-specific scores, and the recipe for the calculation of $\alpha$ given in Section 4.2.1 remains valid. Thus we expect our method of predicting $\lambda(N)$ to apply generally to position-specific scoring functions. This will be tested by extensive numerical studies and will be reported elsewhere. If verified, then the hybrid method will be of much use to a variety of maximum-likelihood or HMM based applications, by providing detailed statistical characterization they currently lack. It should also be of use to database search tools, such as PSI-BLAST, by allowing for position-specific gap functions. On the short side, the semi-probabilistic approach shares with any other probabilistic approach a problem of reduced speed due to the floating-point operations required in its execution. However, the speed reduction is moderate,[10] and we believe it should be outweighed by the overwhelming advantage of a well-characterized null statistics which is not currently available to most of these methods.

## ACKNOWLEDGMENTS

We are grateful to Daniel Fisher for suggesting the saddle-point approach to gapless local alignment. We are also indebted to Ralf Bundschuh for critical discussions which led eventually to the simple recipe for calculating the relative entropy. TH wishes to thank Stephen Altschul for invaluable guidance towards

[10] The reduction factor is machine dependent. On the SGI workstation, we found the reduction to be within a factor of 2; but it can be as much as a factor of $\sim 4$ on PC's which are optimized for integer operations.

relevant problems in this subject, and YKY wishes to thank Ralf Bundschuh and Rolf Olsen for much help and many 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 YKY acknowledges the hospitality of the Center for Studies in Physics and Biology at Rockefeller University where this work was initiated.

## APPENDIX A: THE AFFINE GAP FUNCTION

### *A.1. Algorithms*

The discussion in the main text is carried on in the context of the linear gap function for simplicity. All of the numerics performed in this study are done instead with the more useful affine gap function, with the constraint that deletions cannot immediately follow a series of insertions. In this appendix, we will provide the details of the affine gap algorithm used.

We start with the affine gap version of the Smith–Waterman local alignment. Consider the affine gap function (2) parameterized by the gap opening cost $\delta$, gap extension cost $\varepsilon$, and extra double-gap cost $\delta'$. It will be convenient to introduce three auxiliary quantities $H^S_{m,n}$, $H^D_{m,n}$, and $H^I_{m,n}$, which are defined through the recursion relations

$$\begin{aligned}
H^S_{m,n} &= \max\left\{\begin{array}{l} H^S_{m-1,n-1} + s(a_m, b_n), \\ H^D_{m-1,n-1} + s(a_m, b_n),\ H^I_{m-1,n-1} + s(a_m, b_n) \end{array}\right\}, \\
H^D_{m,n} &= \max\{H^S_{m-1,n} - \delta,\ H^D_{m-1,n} - \varepsilon\}, \qquad (45) \\
H^I_{m,n} &= \max\left\{\begin{array}{l} H^S_{m,n-1} - \delta,\ H^I_{m,n-1} - \varepsilon \\ H^D_{m,n-1} - \delta - \delta' \end{array}\right\},
\end{aligned}$$

supplemented by the boundary conditions $H^S_{0,n} = H^D_{0,n} = H^I_{0,n} = 0$ and $H^S_{m,0} = H^D_{m,0} = H^I_{m,0} = 0$. Then, $H_{m,n}$ as defined in (8) can be obtained simply as

$$H_{m,n} = \max\{0, H^S_{m,n}, H^D_{m,n}, H^I_{m,n}\}. \qquad (46)$$

The optimal score $\mathsf{S}$ is still given in terms of the $H$'s according to (10).

Next, we describe the algorithm for probabilistic global alignment with affine gap weights of the form (19). To compute $W_{m',n';m,n}$ as defined in (21), we need to introduce the auxiliary quantities $W^S$, $W^D$, and $W^I$, which can be computed from the recursion relation

$$\begin{aligned}
W^S_{m',n';m,n} &= w(a_m, b_n) \cdot [W^S_{m',n';m-1,n-1} + \mu^D_1 \cdot W^D_{m',n';m-1,n-1} + \mu^I_1 \cdot W^I_{m',n';m-1,n-1}], \\
W^D_{m',n';m,n} &= \mu^D_2 W^S_{m',n';m-1,n} + \nu W^D_{m',n';m-1,n}, \qquad (47) \\
W^I_{m',n';m,n} &= \mu^I_2 W^S_{m',n';m,n-1} + \nu W^I_{m',n';m,n-1} + \mu^I_2 \mu' \mu^D_1 W^D_{m',n';m,n-1},
\end{aligned}$$

with $m \geq m'$ and $n \geq n'$. In (47), the parameters $\nu$ is the gap extension weight, and $\mu'$ is the extra weighting factor associated with the double-gap configuration. The additional parameters $\mu^D_1(\mu^I_1)$ and $\mu^D_2(\mu^I_2)$ can be interpreted respectively as the weight of terminating a deletion (insertion) and the weight of creating a deletion (insertion). Note that for gap weights of the form (19), which only penalize gap opening and extension (but not specifically gap creation and termination), one can choose arbitrary values of the parameters $\mu^{D,I}_{1,2}$ as long as they satisfy the condition

$$\mu^D_1 \cdot \mu^D_2 = \mu = \mu^I_1 \cdot \mu^I_2, \qquad (48)$$

where $\mu$ is the gap opening weight. The total weight (22) becomes

$$W_{m',n';m,n} = W^S_{m',n';m,n} + W^D_{m',n';m,n} + W^I_{m',n';m,n}, \qquad (49)$$

and the boundary conditions (23) now take on the form

$$\begin{aligned} W^S_{m',n';m=m'-1,n\geq n'} &= 0 = W^S_{m',n';m\geq m',n=n'-1}, \\ W^D_{m',n';m=m'-1,n\geq n'-1} &= 0 = W^I_{m',n';m\geq m'-1,n=n'-1}, \\ W^D_{m',n';m\geq m',n=n'-1} &= \mu_2^D \nu^{m-m'}, \\ W^I_{m',n';m=m'-1,n\geq n'} &= \mu_2^I \nu^{n-n'}, \\ W^S_{m',n';m=m'-1,n=n'-1} &= 1. \end{aligned} \tag{50}$$

Finally, we describe probabilistic local alignment. The generalization of Equation (24) to the affine gap case is

$$Z_{m,n} = 1 + \sum_{m'=0}^{m-1} \mu_2^D \nu^{m'} + \sum_{n'=0}^{n-1} \mu_2^I \nu^{n'} + \sum_{\substack{1\leq m'\leq m \\ 1\leq n'\leq n}} W_{m',n';m,n}. \tag{51}$$

In light of the expression (49) for $W_{m',n';m,n}$ and the recursion relations (47), one can derive analogous relations for $Z_{m,n}$ in terms of the auxiliary quantities $Z^{S,D,I}_{m,n}$. Let

$$\begin{aligned} Z^S_{m,n} &= 1 + \sum_{\substack{1\leq m'\leq m \\ 1\leq n'\leq n}} W^S_{m',n';m'n}, \\ Z^D_{m,n} &= \sum_{\substack{1\leq m'\leq m \\ 1\leq n'\leq n+1}} W^D_{m',n';m,n}, \\ Z^I_{m,n} &= \sum_{\substack{1\leq m'\leq m+1 \\ 1\leq n'\leq n}} W^I_{m',n';m,n}. \end{aligned} \tag{52}$$

Then these auxiliary quantities can be computed recursively as

$$\begin{aligned} Z^S_{m,n} &= 1 + w(a_m, b_n)[Z^S_{m-1,n-1} + \mu_1^D \cdot Z^D_{m-1,n-1} + \mu_1^I \cdot Z^I_{m-1,n-1}], \\ Z^D_{m,n} &= \mu_2^D Z^S_{m-1,n} + \nu Z^D_{m-1,n}, \\ Z^I_{m,n} &= \mu_2^I Z^S_{m,n-1} + \nu Z^I_{m,n-1} + \mu_2^I \mu' \mu_1^D Z^D_{m,n-1} \end{aligned} \tag{53}$$

and $Z_{m,n}$ is obtained as

$$Z_{m,n} = Z^S_{m,n} + Z^D_{m,n} + Z^I_{m,n}. \tag{54}$$

From $_{m,n}$, the total weight W of the probabilistic local alignment is obtained according to (25), and the Maximum Log-Likelihood score of the semi-probabilistic local alignment is obtained according to (35).

## A.2. *Conservation conditions*

As described in Section 3, the alignment weight $W$ can be interpreted as "likelihood" if the local weight parameters satisfy the condition that the total weights going in and out of each node is equal on average. For the probabilistic affine gap problem (47) and (49), this condition can be easily satisfied by exploiting the freedom (48) in choosing the weight parameters $\mu^{D,I}_{1,2}$. They lead to the following additional constraints on these weight parameters:

$$\begin{aligned} \langle w(a,b)\rangle_0 + \mu_2^D + \mu_2^I &= 1, \\ \mu_1^I \langle w(a,b)\rangle_0 + \nu &= 1, \\ \mu_1^D \langle w(a,b)\rangle_0 + \nu + \mu_2^I \mu' \mu_1^D &= 1. \end{aligned} \tag{55}$$

These constraints, together with (48), uniquely determine all of these $\mu_{1,2}$'s, as well as a constraint relating $\langle w\rangle_0$, $\mu$, $\nu$ and $\mu'$. The result is:

$$\mu_1^D = \mu/\mu_2^D = [(1+\mu-\nu)^2 + (\mu'-1)\mu^2]/(1+\mu'\mu-\nu),$$
$$\mu_1^I = \mu/\mu_2^I = [(1+\mu-\nu)^2 + (\mu'-1)\mu^2]/(1-\nu), \tag{56}$$

and

$$\langle w(a,b)\rangle_0 = \frac{(1-\nu)^2}{(1+\mu-\nu)^2 + (\mu'-1)\mu^2}. \tag{57}$$

Equation (57) is the only condition that the real weight parameters $w(a,b)$, $\mu$, $\nu$, and $\mu'$ have to satisfy. For position-specific scoring systems, this condition needs to be satisfied at each node of the alignment lattice.

## APPENDIX B: SCORE ISLANDS AND THE GUMBEL DISTRIBUTION

### *B.1. Gapless local alignment*

In this appendix, we review the concepts of "score landscape" and "islands" which are central to our theoretical analysis. We start by examining gapless local alignment, for which the recursion relation (9) simplifies to

$$H_{m,n} = \max\{H_{m-1,n-1} + s(a_m, b_n), 0\}. \tag{58}$$

We are interested in the statistics of the optimal score

$$\mathsf{S} = \max_{\substack{1\le m\le M\\ 1\le n\le N}} \{H_{m,n}\} \tag{59}$$

for random sequences **a** and **b** described by the distribution $p(a)$. Because the sequences being aligned are random, we can take $m = n$ in (58) without loss of generality. Equation (58) then becomes a discrete Langevin equation with

$$H_{n,n} \equiv H(n) = \max\{H(n-1) + \eta(n), 0\}, \tag{60}$$

where the "noise" $\eta(n) \equiv s_{a_n,b_n}$ is uncorrelated and given by the distribution

$$\rho_0(\eta) = \sum_{a,b\in\chi} \delta(\eta - s(a,b))p(a)p(b). \tag{61}$$

Due to the construction of the scoring system $s(a,b)$, it turns out that the average value of $\eta$ is negative, i.e.,

$$\sum_{a,b} s(a,b)p(a)p(b) < 0. \tag{62}$$

The "dynamics" of the evolution equation (60) are, qualitatively, as follows: The score $H(n)$ starts at zero. If the next local score $\eta(n+1)$ is negative—which is the more typical case due to the Equation (62)—then $H$ remains zero. But if the next local score is positive, then $H$ will increase by that amount. Once it is positive, $H(n)$ performs a "random walk" with independent increments $\eta(n)$. Due to the condition in (62), there is a *negative drift* which forces $H(n)$ to eventually return to zero. After it is reset to zero, the whole process starts over again. The qualitative "temporal" behavior of the score $H(n)$ is depicted in Fig. 12.

From the figure, it is clear that the "score landscape" can be divided into a series of *islands* of positive scores separated by "seas" defined by $H = 0$. Each such island originates from a single jump out of

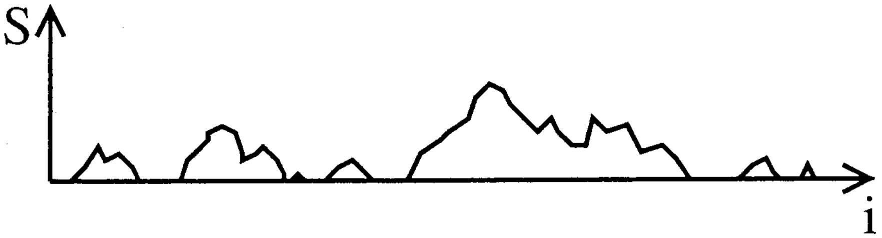


**FIG. 12.** Sketch of the total score as a function of sequence position in gapless local alignment.

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 $\eta(n)$, the islands are *statistically independent* of each other. The same statistical independence applies to the maxima of different islands. Let the total number of islands at $n = N$ be $\kappa(N)$, and let the maximal score of the $i^{\text{th}}$ island be $\sigma_i$. The global optimal score S in (59) can be alternatively written as

$$\mathsf{S} = \max\{\sigma_1, \sigma_2, \dots, \sigma_\kappa\}. \tag{63}$$

As will be shown in Appendix C, the island peak score $\sigma_i$ obeys a Poisson-like distribution for large $\sigma_i$'s, i.e.,

$$\Pr(\sigma_i > x) = (\text{constant}) \cdot e^{-\lambda x} \qquad \text{for } \sigma \gg \lambda^{-1}. \tag{64}$$

Then, according to (63),

$$\Pr(\mathsf{S} < x) = \prod_{i=1}^{\kappa} [1 - \Pr(\sigma_i > x)] \xrightarrow{\kappa\to\infty} \exp[-\kappa e^{-\lambda x}]. \tag{65}$$

The Gumbel distribution shown in Equation (14) is obtained as the derivative of (65).

### *B.2. Gapped local alignment*

Recently, Olsen *et al.* (1999) generalized the above island picture to gapped local alignment: By construction of the Smith–Waterman algorithm (9) or (45), many points on the alignment lattice have score $H = 0$ in the logarithmic regime. As for gapless alignment, a positive score will be generated out of this "sea" of zeroes if a good match occurs by chance. This positive score can then lead to further positive scores via the recursion relation (9). To every positive scoring point on the lattice, i.e., for every $H_{m,n} > 0$, we can associate a path $\mathcal{R}^\dagger(m, n)$ which is the optimal local alignment path given that its forward end is fixed at $(m, n)$. An island is defined to be the collection of points $(m, n)$'s linked together by their respective optimal paths $\mathcal{R}^\dagger(m, n)$. By this definition, every lattice point with a positive score belongs to exactly one island (up to degeneracy of optimal paths which mainly blurs the island boundaries and does not change the distribution of large islands scores). For details, see Olsen *et al.* (1999).

Each of the islands has a maximum score which we denote by $\sigma_i$ as we did in the gapless case. Thus, the optimal score S is given by (63) again, with the total number of islands $\kappa$ depending on the lattice size $M \cdot N$. Since the large islands are well separated by a sea of zero scores, they are statistically independent objects. Thus, their peak scores $\sigma_i$ are again independent and identically distributed random variables, and Gumbel statistics can again be recovered via Equation (65) once the distribution of $\sigma$'s is specified.

The same island picture applies to the case of semi-probabilistic alignment. The score landscape is now obtained by replacing $H_{m,n}$ by $\ln Z_{m,n}$, where $Z_{m,n}$ is computed according to the recursion relation (26). In the logarithmic regime, where the corresponding probabilistic global alignment (22) has exponentially small weights, e.g.,

$$\lim_{N\to\infty} \frac{1}{N} \langle \ln W_{1,1;N,N} \rangle_0 < 0,$$

the majority of lattice points have $\ln Z = 0^+$. In contrast to Smith–Waterman local alignment, the probabilistic algorithm generates a sea with small "ripples" slightly above zero. This makes the boundaries of the individual islands somewhat fuzzy. It does not, however, affect the assignment of *large* islands. It is again straightforward to identify the peak island score ($\ln Z$) for each island (R. Olsen, private communication). We denote them by $\sigma_i$, and the MLL score, as defined in (35), is again of the form (63). Since the large islands are still widely separated by the sea of zeroes, the $\sigma$'s are again uncorrelated. Thus, the distribution of the MLL score $\Phi$ is again given by the distribution of the island peak score.

## APPENDIX C: STATISTICS OF LARGE ISLANDS

In this appendix, we give a heuristic derivation of the probability distribution of the maximum island score $\sigma$ used in Appendix B. To explain our approach, we shall first rederive the exact results of Karlin and Altschul (1990) for gapless alignment in a heuristic manner, by making some reasonable assumptions which dramatically simplify the key calculation. We shall then apply the same assumptions to the gapped alignment case where no exact results exists. These assumptions will simplify the gapped calculation, allowing us to derive the form of the Poisson-like distribution of the island score (64) as well as the condition for the parameter $\lambda$.

### *C.1. Gapless alignment*

In gapless alignment, the score profile $H(n)$ of a *single* island is

$$H(n) = \sum_{j=1}^{n} \eta(j), \qquad \text{with } H(n) > 0 \qquad \text{for all } n \geq 1, \tag{66}$$

where $\eta(j) = 2_{a_j,b_j}$ is again described by the distribution function $\rho_0$ in (62), with $j = 1$ taken to be the island initiation position. The peak island score $\sigma$ is

$$\sigma = \max_{1 \leq n < \infty} H(n),$$

occurring at some position $n = \ell$ such that $H(\ell) = \sigma$. Various island statistics can be derived from the joint probability

$$Q(\sigma, \ell) = \left\langle \theta\left(\sum_{j=1}^{\ell} \eta(j) - \sigma\right) \cdot \prod_{n \neq \ell} \theta\left(\sigma - \sum_{j=1}^{n} \eta(j)\right) \theta\left(\sum_{j=1}^{n} \eta(j)\right) \right\rangle_0, \tag{67}$$

where the factors following $\prod_{n \neq \ell}$ enforce the condition that $H(n)$ is bounded between 0 and $\sigma$, except at $n = \ell$ where $H(n)$ exceeds $\sigma$. From $Q(\sigma, \ell)$, the probability distribution function of $\sigma$ is obtained as

$$\mathrm{pdf}(\sigma) = \sum_{\ell=1}^{\infty} Q'(\sigma, \ell), \tag{68}$$

where $Q'(\sigma, \ell) \equiv \partial Q / \partial \sigma$. The average island length $\overline{\ell}$ is

$$\overline{\ell}(\sigma) = \sum_{\ell=1}^{\infty} \ell \cdot Q'(\sigma, \ell). \tag{69}$$

What make the calculation of $Q(\sigma, \ell)$ difficult are the restriction factors in (68). Motivated by the exact result (Karlin and Altschul, 1990) that

$$\lim_{\sigma \to \infty} \overline{\ell}(\sigma) = \alpha^{-1} \cdot \sigma, \tag{70}$$

where $\alpha$ is known as the "relative entropy," we hypothesize that the removal of the restriction factors in (68) does not change the *leading* behavior of the probability distribution function in the limit of large $\sigma$, e.g., $\sigma \gg \sigma_0$, where $\sigma_0$ is the typical island score (which will turn out to be of the order $\lambda^{-1}$). This leads us to consider the unrestricted probability

$$\tilde{Q}(\sigma|\ell) = \left\langle \theta\left(\sigma - \sum_{j=1}^{\ell} \eta(j)\right)\right\rangle_0 , \tag{71}$$

where $\theta(x)$ is the Heaviside unit step function with $\theta(x) = 1$ if $x > 0$ and with $\theta(x) = 0$ if $x < 0$. Alternatively, one can consider its differentiated form

$$\tilde{Q}'(\sigma|\ell) = \left\langle \delta\left(\sigma - \sum_{j=1}^{\ell} \eta(j)\right)\right\rangle_0 , \tag{72}$$

which describes the probability that the corresponding *global* alignment score

$$S(\ell) \equiv S_{1,1;\ell,\ell} = \sum_{j=1}^{\ell} \eta(j) \tag{73}$$

reaches the value $\sigma$ after $\ell$ steps. In term of $\tilde{Q}'$, our hypothesis can be expressed simply as

$$\lim_{\sigma \gg \sigma_0} \mathrm{pdf}(\sigma) \approx \sum_{\ell=1}^{\infty} \tilde{Q}'(\sigma|\ell). \tag{74}$$

The computation of $\tilde{Q}'$ is straightforward. From Equation (72), we have

$$\tilde{Q}'(\sigma|\ell) = \int dk e^{-ik\sigma} \cdot \prod_{j=1}^{\ell} \int d\eta(j) e^{ik\eta(j)} \rho_0(\eta_j) = \int dk e^{-ik\sigma} \cdot [\hat{\rho}_0(k)]^{\ell} \tag{75}$$

where

$$\hat{\rho}_0(k) \equiv \int d\eta \rho_0(\eta) e^{ik\eta} = \sum_{a,b\in\chi} e^{iks(a,b)} p(a) p(b) \tag{76}$$

is the Fourier transform of $\rho_0(\eta)$. Using this expression in (74) and replacing the sum over $\ell$ by an integral, we find the following for the pdf of interest:

$$\lim_{\sigma \gg \sigma_0} \mathrm{pdf}(\sigma) \approx \int_1^{\infty} d\ell \int_{-\infty}^{\infty} dk e^{-ik\sigma} \cdot [\hat{\rho}_0(k)]^{\ell}. \tag{77}$$

It is convenient to change the integration variable $\ell$ to $u = \ell/\sigma$. The double integral (77) can then be evaluated in the limit of large $\sigma$ by applying the saddle-point method *twice.* Integration over $k$ yields the result

$$\tilde{Q}'(\sigma|u\sigma) = e^{-ik^*\sigma + u\sigma \ln \hat{\rho}(k^*)} \tag{78}$$

together with the saddle-point condition

$$-i + u\frac{d}{dk} \ln \hat{\rho}_0(k)\bigg|_{k=k^*(u)} = 0 \tag{79}$$

which also provides the implicit function $k^*(u)$. The second integration over $u$ can be evaluated at the second saddle-point condition:

$$\left[-i + u\frac{d}{dk}\ln\hat{\rho}_0(k)\right]_{k=k^*} \frac{dk^*}{du}\bigg|_{u=u^*} + \ln\hat{\rho}_0(k^*(u^*)) = 0, \tag{80}$$

which is reduced to

$$\ln\hat{\rho}_0(k^*(u^*)) = 0, \tag{81}$$

given the first saddle-point (80). Finally, from Equations (78) and (81), we have

$$\lim_{\sigma\gg\sigma_0} \mathrm{pdf}(\sigma) \approx \tilde{Q}'(\sigma|u^*\sigma) \approx e^{-ik^*(u^*)\sigma}. \tag{82}$$

The results (82) and (81) can be written in a more familiar form by introducing $\lambda(u) \equiv ik^*(u)$. We then have

$$\lim_{\sigma\gg\sigma_0} \mathrm{pdf}(\sigma) \approx e^{-\lambda(u^*)\sigma}, \tag{83}$$

where $\lambda(u^*)$ is given by $\ln[\hat{\rho}_0(-i\lambda(u^*))] = 0$. Recalling the definition (76), the condition for $\lambda(u^*)$ can be expressed as

$$\sum_{a,b\in\chi} e^{\lambda s_{a,b}} p(a)p(b) = 1, \tag{84}$$

which is the exact result obtained by Karlin and Altschul.

Next, we note that the length scale $\ell^*(\sigma)$ selected by the saddle point approximation is given by $\ell^*(\sigma) = u^*\sigma$, where $u^*$ is defined by the two saddle-point conditions (79) and (81). Expecting that $\ell^*(\sigma) \approx \overline{\ell}(\sigma)$ in the limit of large $\sigma$, we have

$$\alpha = u^{*-1} = -i\frac{d\hat{\rho}_0}{dk}\bigg|_{k^*(u^*)}, \tag{85}$$

or more explicitly

$$\alpha = \sum_{a,b\in\chi} s(a,b)e^{\lambda s(a,b)} p(a)p(b), \tag{86}$$

using the definition of (76). The expression for the relative entropy given by Karlin and Altschul is recovered as $\lambda \cdot \alpha$.

The above saddle point approximation will be asymptotically exact provided that appropriate conditions on the second derivative of the exponent of the integrand are satisfied, i.e., provided that

$$\sigma\frac{\partial^2}{\partial(ik)^2}[-iku + u\ln\hat{\rho}_0(k)]_{k=k^*(u^*)} \xrightarrow{\sigma\to\infty} -\infty, \tag{87}$$

$$\sigma\frac{\partial^2}{\partial u^2}[-ik + u\ln\hat{\rho}_0(k)]_{k=k^*(u^*)} \xrightarrow{\sigma\to\infty} -\infty. \tag{88}$$

The first condition (87) comes with the first saddle point equation (79) and the second condition comes with the second saddle point equation (80). For the first condition, $u$ is held as a positive constant and the condition is satisfied through the use of the Schwarz inequality

$$\left[\sum_{a,b\in\chi} e^{\lambda s_{a,b}} p(a)p(b)\right] \cdot \left[\sum_{a,b\in\chi} s_{a,b}^2 e^{\lambda s_{a,b}} p(a)p(b)\right] > \left[\sum_{a,b\in\chi} s_{a,b}e^{\lambda s_{a,b}} p(a)p(b)\right]^2. \tag{89}$$

The second condition, Equation (88), can also be easily verified through differentiating Equation (79) with respect to $u$ and applying the Schwarz inequality (89).

## *C.2. Gapped local alignment*

The single most important ingredient in the above heuristic approach is the linear dependence of island score $\sigma$ on island length $\ell$. Since this linear relation is again expected in gapped alignment, it is reasonable to extend the above approach to the gapped case. In particular, we conjecture that the island peak score distribution, e.g., the gapped analog of Equations (67) and (68), can again be obtained by their unrestricted counterpart, e.g., the gapped version of Equation (71), which involves only the gapped *global* alignment score, $S_{1,1;m,n}$.

Given an island score profile $H_{m,n}$ for gapped local alignment, there are now *two* coordinates (e.g., $\ell_1$ and $\ell_2$) for the position of the island peak position. (The island initiation position is taken here to be (1,1).) The peak island score $\sigma$ is now specified by the joint probability

$$Q(\sigma, \ell_q, \ell_2) = \left\langle \theta(\sigma - H_{\ell_1,\ell_2}) \prod_{\substack{m\neq\ell_1 \\ n\neq\ell_2}} \theta(H_{m,n} - \sigma) \right\rangle_0 , \tag{90}$$

whose derivative $Q'(\sigma, \ell_1, \ell_2) \equiv \frac{d}{d\sigma} Q(\sigma, \ell_1, \ell_2)$ specifies the probability distribution

$$\text{pdf}(\sigma) = \sum_{\ell_1,\ell_2} Q'(\sigma, \ell_1, \ell_2). \tag{91}$$

Again, we hypothesize that for large $\sigma$ the inclusion of the restriction factors in (90) are not important. We first remove the lower restriction at score 0 by replacing $H_{m,n}$ by the global alignment score $S_{1,1;m,n}$. Thus we conjecture that

$$\lim_{\sigma\gg\sigma_0} Q(\sigma, \ell_1, \ell_2) \approx \left\langle \theta(\sigma - S_{1,1;\ell_1,\ell_2}) \prod_{\substack{m\neq\ell_1 \\ n\neq\ell_2}} \theta(S_{1,1;m,n} - \sigma) \right\rangle_0 \equiv Q_1(\sigma, \ell_1, \ell_2). \tag{92}$$

Next, we observe that the upper restriction at score $\sigma$ represented by $\prod_{m\neq\ell_1,n\neq\ell_2}$ in (93) can be compactified by first considering the auxiliary quantity

$$\tilde{S}_{1,1;n,n} \equiv \max_{1\leq j\leq n} \{S_{1,1;j,n}, S_{1,1;n,j}\}, \tag{93}$$

which is the optimal global alignment score for an alignment path with the backward end fixed at (1, 1) and the other end free to be anywhere along the outer boundary of $i = m$ or $j = n$. It then follows[11] that

$$\sum_{j=1}^{\ell} [Q_1(\sigma, j, \ell) + Q_1(\sigma, \ell, j)] = \left\langle \theta(\sigma - \tilde{S}_{1,1;\ell,\ell}) \cdot \prod_{n\neq\ell} \theta(\tilde{S}_{1,1;n,n} - \sigma) \right\rangle_0 \equiv Q_2(\sigma, \ell). \tag{94}$$

Finally, we relax the upper score restriction and assume that

$$\lim_{\sigma\gg\sigma_0} Q_2(\sigma, \ell) \approx \langle \theta(\sigma - \tilde{S}_{1,1;\ell,\ell}) \rangle_0 \equiv \tilde{Q}(\sigma|\ell). \tag{95}$$

[11] We have not included here the occurrence of rare cases where multiple maxima exist among $S_{1,1;m,\ell}$ and $S_{1,1;\ell,n}$; in any case, they should not affect the leading behavior of the pdf for large $\sigma$.

Combining Equations (91), (92), (94), and (95), we obtain

$$\lim_{\sigma \gg \sigma_0} \mathrm{pdf}(\sigma) \approx \sum_{\ell=1}^{\infty} \tilde{Q}'(\sigma|\ell) \tag{96}$$

with

$$\tilde{Q}'(\sigma|\ell) = \langle \delta(\sigma - \tilde{S}_{1,1;\ell,\ell}) \rangle_0. \tag{97}$$

Equation (96) is the generalization of the hypothesis (74) to gapped alignment.

The calculation for $\tilde{Q}'$ closely follows the gapless case. We have

$$\tilde{Q}'(\sigma|\ell) = \int dk e^{ik\sigma} \langle \exp[ik\tilde{S}_{1,1;\ell,\ell}] \rangle_0. \tag{98}$$

Unlike the gapless case, it is no longer possible to decompose $\langle e^{ikS} \rangle_0$ into a product of $\ell$ independent terms. Nevertheless, we expect the result to be of the form

$$\langle \exp[ik\tilde{S}_{1,1;\ell,\ell}] \rangle_0 = [\hat{\rho}(k)]^{\ell} \tag{99}$$

for large $\ell$, with a nontrivial function $\hat{\rho}(k)$. We can now follow exactly the analysis described above for the gapless case, with the substitution of $\hat{\rho}_0(k)$ by $\hat{\rho}(k)$. We again find a Poisson-like distribution

$$\lim_{\sigma \gg \sigma_0} \mathrm{pdf}(\sigma) \sim (\text{constant}) \cdot e^{-\lambda\sigma},$$

with the parameter $\lambda$ given by

$$\hat{\rho}(-i\lambda) = 1 \tag{100}$$

and the relative entropy

$$\alpha = -i \frac{d}{dk} \ln \hat{\rho}(k) \bigg|_{k=-i\lambda}. \tag{101}$$

With the result (100), Equation (99) becomes

$$\lim_{\ell \to \infty} \langle \exp[\lambda \cdot \tilde{S}_{1,1;\ell,\ell}] \rangle_0 = 1, \tag{102}$$

which is the gapped alignment generalization of the Karlin–Altschul solution (84). From (100) and (102), we also obtain an explicit formula for the relative entropy:

$$\alpha = \lim_{\ell \to \infty} \ell^{-1} \langle \tilde{S}_{1,1;\ell,\ell} e^{\lambda \cdot \tilde{S}_{1,1;\ell,\ell}} \rangle_0. \tag{103}$$

The above approach generalizes straightforwardly to the semi-probabilistic alignment algorithm presented in Section 4, with the substitution of $\tilde{S}_{1,1;\ell,\ell}$ by $\ln \tilde{W}_{1,1;\ell,\ell}$.

## APPENDIX D: HIDDEN MARKOV MODEL OF SEQUENCE EVOLUTION

In this appendix, we describe a hidden Markov model $\mathcal{M}$ which mimics simple evolution processes and generates pairs of *correlated* sequences $\mathbf{a}_{1,m}$ and $\mathbf{b}_{1,n}$. We will find that the joint probability distribution

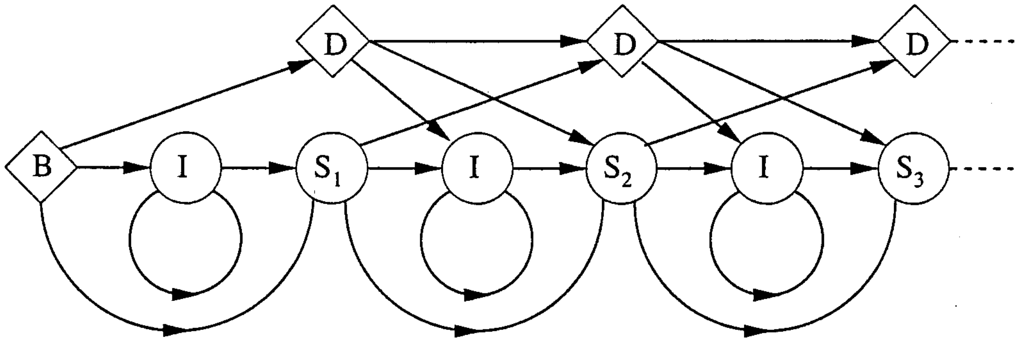


**FIG. 13.** Schematics of the hidden Markov mode $\mathcal{M}$ for sequence evolution. The different states are $B$ for the "begin" state, $I$ for the "insertion" states, $D$ for the "deletion" states, and $S_i$ for the substitution of $i^{\text{th}}$ element of the input sequence $\mathbf{a}$. The arrows indicate the allowed transitions between the states, with transition probabilities given in the text. Sequence elements are "emitted" in the states denoted by circles ($I$ and $S_i$): An element $b$ is emitted with probability $p(b)$ every time the state $I$ is visited and emitted with probability $T_c(b|a_i)$ every time a state $S_i$ is visited. The evolution process terminates either when the new sequence generated reaches a specified length or the input sequence is exhausted.

$P_c[\mathbf{a}, \mathbf{b}]$ specified by $\mathcal{M}$ corresponds directly to the alignment weights $\tilde{W}[\mathbf{a}, \mathbf{b}]$ described in Sec. 3. This connection is used in Sec. 4 to compute the relative entropy of the semi-probabilistic alignment.

First a random sequence $\mathbf{a}_{1;m}$ (the ancestor sequence) is generated according to the amino acid frequencies[12] $p(a)$ for each element $a$. Then, one mutates the sequence $\mathbf{a}$ according to the mutation model illustrated schematically in Fig. 13: Starting from the "begin" state $B$, the model visits a series of "states" $I$, $D$, or $S_i$ (respectively the "insertion", "deletion", and "substitution" states) sequentially by following the arrows in a stochastic manner, with probability $q(Y|X)$ for the transition from state $X$ to $Y$,[13] with $X, Y \in \{S, D, I\}$ and $q(D|I) = 0$. The transition probabilities obey the conservation conditions

$$\sum_{Y} q(Y|X) = 1 \tag{104}$$

for each state $X$. Every time a circled state ($I$ or $S_i$) is visited, a new sequence element $b$ is "emitted", with probability $p(b)$ in state $I$ and probability $T_c(b|a_i)$ in state $S_i$. The elements $b$'s derived from the execution of the model are labeled in order, as $b_1$, $b_2$, etc. The model stops when the length of the derived sequence $\mathbf{b}$ reaches $n$, or when all of the elements in $\mathbf{a}_{1,m}$ are exhausted. In the latter case, random elements are generated according to $p(b)$ and added to the sequence $\mathbf{b}$ until it reaches the length $n$.

Each sequence pair $[\mathbf{a}, \mathbf{b}]$ generated according to the mutation model $\mathcal{M}$ can be described by the set $\mathcal{R}_c(1, 1; m, n) = \{(m_1, n_1), (m_2, n_2), \ldots, (m_l, n_l)\}$ analogous to the alignment path described in Section 2, with the index pairs $(m_k, n_k)$ denoting the substitution of $a_{m_k}$ by $b_{n_k}$. Let the coordinates of the last substitution event be $i = m_l \leq m$ and $j = n_l \leq n$. The probability $\mathcal{P}_c$ of obtaining a sequence pair $[\hat{\mathbf{a}}_{1,i}, \hat{\mathbf{b}}_{1,j}]$, with substitutions specified by $\mathcal{R}_c$ is

$$\begin{aligned} \mathcal{P}_c[\hat{\mathbf{a}}_{1,i}, \hat{\mathbf{b}}_{1,j}; \mathcal{R}_c] = {} & \prod_{k=1}^{l} [T_c(b_{n_k}|a_{m_k})p(a_{m_k})] \\ & \cdot \prod_{k=0}^{l-1} \left[ g'_c(m_{k+1} = m_k = 1, n_{k+1} - n_k - 1) \cdot \prod_{i_k=m_k+1}^{m_{k+1}-1} p(a_{i_k}) \cdot \prod_{j_k=n_k+1}^{n_{k+1}-1} p(b_{j_k}) \right], \end{aligned} \tag{105}$$

[12] The frequency $p(a)$ is chosen as the largest eigenvector (with eigenvalue 1) of the transition matrix $T_c(b|a)$.

[13] For simplicity, we use here $q(Y|B) = q(Y|S)$, although the transition probability for the first state can in principle be different.

with $(m_0, n_0) = (0, 0)$ and

$$g'_c(\ell_D, \ell_I) = \begin{cases} q(S|S) & \ell_D = 0, \ell_I = 0 \\ q(D|S) \cdot [q(D|D)]^{\ell_D - 1} \cdot q(S|D) & \ell_D \geq 1, \ell_I = 0 \\ q(I|S) \cdot [q(I|I)]^{\ell_I - 1} \cdot q(S|I) & \ell_D = 0, \ell_I \geq 1 \\ q(D|S) \cdot [q(D|D)]^{\ell_D - 1} \cdot q(I|D) & \\ \cdot [q(I|I)]^{\ell_1 - 1} \cdot q(S|I) & \ell_D \geq 1, \ell_I \geq 1. \end{cases} \tag{106}$$

The expression $g'_c$ is close to the form of the affine gap function (19). Let us express the transition probabilities in the following way:

$$\begin{aligned} q(D|S) \cdot q(S|D) &= \mu_c^D \cdot q(S|S) \\ q(I|S) \cdot q(S|I) &= \mu_c^I \cdot q(S|S) \\ q(D|S) \cdot q(I|D) \cdot q(S|I) &= \mu'_c \cdot \mu_c^D \cdot \mu_c^I \cdot q(S|S) \end{aligned} \tag{107}$$

and further restrict the parameter space to

$$\mu_c^D = \mu_c = \mu_c^I$$

$$q(D|D) = \nu_c = q(I|I). \tag{108}$$

Then we have $g'_c(\ell_D, \ell_I) = q(S|S) \cdot g_c(\ell_D, \ell_I)$, where

$$g_c(\ell_D, \ell_I) = \begin{cases} 1 & \ell_D = 0, \ell_I = 0 \\ \mu_c \cdot \nu_c^{\ell_D - 1} & \ell_D \geq 1, \ell_I = 0 \\ \mu_c \cdot \nu_c^{\ell_I - 1} & \ell_D = 0, \ell_I \geq 1 \\ \mu'_c \cdot \mu_c^2 \nu_c^{\ell_D + \ell_I - 2} & \ell_D \geq 1, \ell_I \geq 1. \end{cases} \tag{109}$$

Equation (109) is of the same form as the affine gap weight (19). The transition probabilities can be expressed in terms of $\mu_c$, $\mu'$ and $\nu_c$ using the definitions (107) and (108) and the conservation condition (104). We find

$$\begin{aligned} q(S|S) &= \frac{(1 - \nu_c)^2}{(1 + \mu_c - \nu_c)^2 + (\mu'_c - 1)\mu_c^2}, \\ q(D|S) &= \frac{(1 + \mu'_c \mu_c - \nu_c)\mu_c}{(1 + \mu_c - \nu_c)^2 + (\mu'_c - 1)\mu_c^2}, \\ q(I|S) &= \frac{(1 - \nu_c)\mu_c}{(1 + \mu_c - \nu_c)^2 + (\mu'_c - 1)\mu_c^2}, \\ q(S|D) &= \frac{(1 - \nu_c)^2}{1 + \mu'_c \mu_c - \nu_c}, \\ q(I|D) &= \frac{(1 - \nu_c)\mu'_c \mu_c}{1 + \mu'_c \mu_c - \nu_c}, \\ q(S|I) &= 1 - \nu_c. \end{aligned} \tag{110}$$

Using Equations (109) and (110) in Equation (105) and comparing the expression (20) for the weight $\mathcal{W}$ of an alignment path, we see that

$$\mathcal{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1,j}; \mathcal{R}_c] = P_0[\hat{\mathbf{a}}_{1,i}, \hat{\mathbf{b}}_{1,j}] \cdot w_c(a_i, b_j) \cdot \mathcal{W}[\mathcal{R}_c; \hat{\mathbf{a}}_{1;i=1}, \hat{\mathbf{b}}_{1,j-1}; w_c, g_c] \tag{111}$$

with

$$w_c(a,b) = \frac{(1-\nu_c)^2}{(1+\mu_c-\nu_c)^2 + (\mu'_c - 1)\mu_c^2} \cdot \frac{T_c(b|a)}{p(b)} \tag{112}$$

The restricted total probability $\hat{P}_c$ for the sequence pair $[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1;j}]$ with $(a_i, b_j)$ paired is obtained by summing over all allowed set of substitutions $\mathcal{R}_c$, subject to the constraint that the last substitution is the pair $(a_i, b_j)$. Using Equation (111) together with (21), we find

$$\begin{aligned}\hat{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1;j}] &= \sum_{\mathcal{R}_c(1,1;i-1,j-1)} \mathcal{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1,j}; \mathcal{R}_c] \\ &= P_0[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1;j}] \cdot w_c(a_i, b_j) \cdot W_{1,1;i-1,j-1}[\hat{\mathbf{a}}_{1;i-1}, \hat{\mathbf{b}}_{1;j-1}; w_c, g_c].\end{aligned} \tag{113}$$

To find the unrestricted total probability of generating the *entire* sequence pair $[\mathbf{a}_{1;m}, \mathbf{b}_{1,n}]$, we need to account for how the model $\mathcal{M}$ terminates after the last substitution event $(a_i, b_j)$. According to the description above, if either $i = m$ or $j = n$; then the other sequence is simply completed by generating random elements using the background frequency $p$. However, if $i < m$ and $j < n$, then one of the following occurs to complete use of the sequences before the other sequence is completed with random elements.

- Insertion occurs after the substitution $(a_i, b_j)$, completing the sequence $\mathbf{b}$ (i.e., for $b_{j+1} \ldots b_n$) with random elements. The associated probability is

$$\tilde{P}_c[\hat{\mathbf{a}}_{1;i<m}, \hat{\mathbf{b}}_{1,n}] = \hat{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1,j}] \cdot [q(I|S)/\nu] \cdot \prod_{j'=j+1}^{n} [\nu p(b_{j'})].$$

- Deletion occurs after the substitution $(a_i, b_j)$: the remainder of the sequence $\mathbf{a}$ (i.e., elements $a_{i+1}, \ldots, a_m$) is completely deleted. The associated probability is

$$\tilde{P}_c[\hat{\mathbf{a}}_{1;m}, \hat{\mathbf{b}}_{1,j<n}] = \hat{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1,j}] \cdot [q(D|S)/\nu] \cdot \prod_{i'=i+1}^{m} [\nu p(a_{i'})].$$

- A deletion occurs from elements $a_{i+1}, \ldots, a_k$ with $k < m$, followed by an insertion completing the sequence $\mathbf{b}$. The associated probability is

$$\tilde{P}'_c[\hat{\mathbf{a}}_{1;k<m}, \hat{\mathbf{b}}_{1,n}] = \hat{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1,j}] \cdot [q(D|S)/\nu] \cdot \prod_{i'=i+1}^{k} [\nu p(a_{i'})] \cdot [q(I|D)/\nu] \cdot \prod_{j'=j+1}^{n} [\nu p(b_{j'})].$$

The total probability of obtaining the sequence pair [**a**, **b**] is then the sum of the above possibilities. We have

$$\begin{aligned}P_c[\mathbf{a}, \mathbf{b}] &= \hat{P}_c[\hat{\mathbf{a}}_{1;m}, \hat{\mathbf{b}}_{1;n}] \\ &+ \sum_{i=1}^{m-1} \prod_{i'=i+1}^{m} p(a_{i'}) \cdot \{\hat{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1;n}] + \tilde{P}_c[\hat{\mathbf{a}}_{1;i}, \hat{\mathbf{b}}_{1;n}]\} \\ &+ \sum_{j=1}^{n-1} \prod_{j'=j+1}^{m} p(b_{j'}) \cdot \{\hat{P}_c[\hat{\mathbf{a}}_{1;m}, \hat{\mathbf{b}}_{1;j}] + \tilde{P}_c[\hat{\mathbf{a}}_{1;m}, \hat{\mathbf{b}}_{1;j}]\} \\ &+ \sum_{k=1}^{m-1} \prod_{i'=k+1}^{m} p(a_{i'}) \tilde{P}'_c[\hat{\mathbf{a}}_{1;k}, \hat{\mathbf{b}}_{1;n}].\end{aligned} \tag{114}$$

The above can be written compactly as

$$P_c[\mathbf{a}, \mathbf{b}] = P_0[\mathbf{a}, \mathbf{b}] \cdot \tilde{W}_{1,1;m,n}[\mathbf{a}, \mathbf{b}; w_c, g_c]. \tag{115}$$

$\tilde{W}$ is most transparent when expressed in terms of the auxiliary quantities $W^S$, $W^D$, and $W^I$ introduced in the dynamic programming calculation of $W$ [Equation (47)] in Appendix A,

$$\tilde{W}_{1,1;m,n} = \sum_{i=1}^{m-1} [W^S_{1,1;i,n} + W^I_{1,1;i,n}] + \sum_{j=1}^{n-1} [W^S_{1,1;m,j} + W^D_{1,1;m,j}] + W^S_{1,1;m,n} + W^I_{1,1;0,n} + W^D_{1,1;m,0}, \tag{116}$$

with the parameters of Equation (47) determined by the weight functions $w_c$ and $g_c$. Equation (116) is just the affine gap version of Equation (30) describing the total weight entering the forward boundaries at $i = m$ and $j = n$ (i.e., the dashed lines of Fig. 1(b)).

Equation (115) together with the efficient formula (116) for computing $\tilde{W}$ are very useful results which are exploited in the main text to compute the relative entropy with minimal effort. Here, we make another application of (115) on the average log-likelihood score of the alignment of correlated sequence pairs. Suppose the sequences **a** and **b** are generated by the evolution model $\mathcal{M}$, with mutation probabilities $w_c$ and $g_c$. Let us align these two sequences using probabilistic local alignment (25) and (26), with weight functions $w$ and $g$. We have

$$\phi(w, g; w_c, g_c) \equiv \sum_{[\mathbf{a},\mathbf{b}]} \ln \tilde{W}[\mathbf{a}, \mathbf{b}; w, g] P_c[\mathbf{a}, \mathbf{b}; w_c, g_c] = \langle \ln \tilde{W}[w, g] \cdot \tilde{W}[w_c, g_c]\rangle_0, \tag{117}$$

due to the relation (115).

An important issue in homology detection is to determine the alignment weights $(w, g)$ which maximizes $\phi$. Given the normalization condition $\langle \tilde{W}[w, g]\rangle_0 = 1 = \langle \tilde{W}[w_c, g_c]\rangle_0$ and using the inequality $\ln x \le x - 1$ for all $x > 0$, it is straightforward to show that

$$\phi(w, g; w_c, g_c) \le \phi(w_c, g_c; w_c, g_c). \tag{118}$$

Thus the optimal alignment weights $w^*$ and $g^*$ needed to obtain maximum $\phi$ are given by

$$g^*(\ell) = g_c(\ell) \qquad \text{and} \qquad w^*(a, b) = w_c(a, b). \tag{119}$$

The relations (118) and (119) form the basis of the maximum likelihood approach to the parameter selection problem encountered in any optimization problems (Durbin *et al.*, 1998).

Relations similar to those described above between the alignment weights and correlated ensembles can be established for local alignment, provided we expand the mutation model to include free insertion modules (Hughey and Krogh, 1996) at the beginning and end to generate random background sequences. Additional probability conservation conditions are needed to specify the relative abundance and length of the correlated substrings. We will not delve into this case here as it will not be directly relevant to the subject matter of this study.

## REFERENCES

Altschul, S.F. 1991. Substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 119, 555–565.

Altschul, S.F. 1998. Generalized affine gap costs for protein sequence alignment. *Proteins* 32, 88–96.

Altschul, S.F., Bundschuh, R., Hwa, T., and Olsen, R. 2000. The estimation of statistical parameters for local alignment score distributions. *Nucl. Acids Res.* 29, 351–361.

Altschul, S.F., and Gish, W. 1996. Local alignment statistics. *Methods Enzymol.* 266, 460–480.

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., 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. *Nucl. Acids Res.* 25, 3389–3402.

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

Arratia, R., and Waterman, M.S. 1994. A phase transition for the score in matching random sequences allowing deletions. *Annals of Applied Probability* 4, 200–225.

Barret, C., Hughey, R., and Karplus, K. 1997. Scoring hidden Markov models. *CABIOS* 13, 191–199.

Bucher, P., and Hofmann, K. 1996. A sequence similarity search algorithm based on a probabilistic interpretation of an alignment scoring system. *ISMB-96*, 44–50.

Bucher, P., Karplus, K., Moeri, N., and Hoffman, K. 1996. A flexible motif search technique based on generalized profiles. *Comput. Chem.* 20, 3–24.

Bundschuh, R. 1999. An analytic approach to significance assessment in local sequence alignment with gaps, to appear in *RECOMB 2000*.

Bundschuh, R., and Hwa, T. 1999. An analytical study of the phase transition line in local sequence alignment with gaps. *RECOMB 99*, 70–76.

Burge, C., and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94.

Chvátal, V., and Sankoff, D. 1975. Longest common subsequences of two random sequences. *J. Appl. Probab.* 12:306–315.

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. *Atlas of Protein Sequence and Structure* 5 supp. 3, 345–358.

Drasdo, D., Hwa, T., and Lassig, M. 1998. A scaling theory of sequence alignment with gaps. *ISMB 98*, 52–58.

Durbin, R., Eddy, S., Krogh, A., and Mitchinson, G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, England.

Eddy, S., Mitchison, G., and Durbin, R. 1995. Maximum discrimination hidden Markov models of sequence consensus. *J. Comp. Biol.* 2, 9–23.

Grundy, W.N., Bailey, T.L., Elkan, C.P., and Baker, M.E. 1997. Meta-MEME: Motif-based hidden Markov models of protein families. *Computer Appl. Biosci.* 13, 397–406.

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

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

Henikoff, S., and Henikoff, J.G. 1994. Position-based sequence weights. *J. Mol. Biol.* 162, 705–708.

Holmes, I., and Durbin, R. 1998. Dynamic programming alignment accuracy. *RECOMB 98*, 102–108.

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 Lassig, M. 1998. Optimal detection of sequence similarity by local alignment. *RECOMB 98*, 109–116.

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

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

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* 235: 1501–1531.

Kschischo, M., and Lassig, M. 1999. Finite-temperature sequence alignment, to appear in *PSB 2000*.

Milosavljevic, A., and Jurka, J. 1993. Discovering simple DNA sequences by the algorithmic similarity method. *CABIOS* 9, 407–411.

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

Mott, R., and Tribe, R. 1999. Approximate statistics of gapped alignment. *J. Comp. Biol.* 6, 91–112.

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 *in Proc. 7th Int. Conf. Intelligent Systems for Molecular Biology*, 211–222, AAAI Press, Menlo Park.

Pearson, W.R. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.

Siegmund, D., and Yakir, B. 2000. Approximate P-values for sequence alignments. *Annals of Stat.* 28, 657–680.

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. *Nucl. Acids Res.* 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., Gordon, L., and Arratia, R. 1987. Phase transitions in sequence matches and nucleic acid structure. *Proc. Natl. Acad. Sci. USA* 84, 1239–1243.

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. USA* 91: 4625–4628.

Zhang, M.Q., and Marr, T.G. 1995. Alignment of molecular sequences seen as random path analysis. *J. Theo. Biol.* 174, 119–129.

Address correspondence to:
*Yi-Kuo Yu*
*Department of Physics*
*Florida Atlantic University*
*777 Glades Road*
*Boca Raton, FL 33431-0991*

*E-mail:* yyu@fau.edu