JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 1, Number 2, 1994
Mary Ann Liebert, Inc., Publishers
Pp. 93–104

Approximations to Profile Score Distributions

LARRY GOLDSTEIN[1] and MICHAEL S. WATERMAN[1,2]

ABSTRACT

Profiles, which are summaries of multiple alignments of a sequence family, are used to find new instances of the family in databases. In this paper, we study the maximum score M obtained when the profile is aligned without indels at all possible positions of a random sequence. The main theorem gives an approximation to the distribution function of M with an explicit bound on the error. This theorem implies that M has a limiting extreme value distribution.

INTRODUCTION

DATABASE SEARCHES ARE NOW ROUTINE IN MOLECULAR BIOLOGY. A newly determined DNA sequence is compared to nucleic acid databases to discover similar sequences that have already been studied. Often it is easier to find protein similarities by comparing the amino acid sequences encoded in the DNA sequences. Therefore, a putative gene sequence may be translated into an amino acid sequence and then compared to a protein sequence database. Often, gene locations are unknown and translation into amino acid sequence is done in all six reading frames. The results of the protein sequence comparisons can be very important to an understanding of the biology of the new sequence. The famous discovery of a striking similarity between human platelet-derived growth factor (PDG-F) and the cancer-related virus v-*sis* oncogene product was the result of a computer search (Doolittle *et al*, 1983). Similarly, many other discoveries have been made, and every new sequence is analyzed in this manner.

Often, biologically significant comparisons will be fairly weak due to the time since divergence from a common ancestor because evolutionary changes may have accumulated and obscured the ancestral relationship. The ability to detect common evolutionary history is frequently improved by considering a set of related sequences. Often this is done by making a multiple alignment of the sequences. To illustrate this we present a multiple alignment of $N = 7$ DNA sequences of length $m = 8$.

	Sequence position							
Sequence	1	2	3	4	5	6	7	8
1	T	T	A	C	T	A	T	C
2	A	A	C	G	T	C	T	C
3	A	A	G	C	A	G	C	C
4	C	T	C	T	T	T	G	C
5	A	T	G	A	A	A	A	C
6	G	C	G	C	T	T	T	C
7	A	G	G	A	T	G	C	C

Departments of [1]Mathematics and of [2]Molecular Biology, University of Southern California, Los Angeles, CA 90089-1113.

No pair of these sequences has a strong similarity. Sequences 4 and 5 match in only 2 of 8 positions, for example. However in every position except 6 there is a majority letter, so that the alignment might be summarized as the "consensus" sequence ATGCT{A,T,G}TC, where the tie between A, T, and G in position 6 is represented as {A,T,G}.

Multiple alignment is an area that began early (Sankoff, 1975; Waterman *et al.*, 1976) and is still under active development (Carrillo and Lipman, 1988; Pevzner, 1992; Gusfield, 1993; Wang and Jiang, in preparation). With the variety of available methods, it remains true that most multiple alignments are made by merging pairwise alignments, often by a greedy algorithm where the most closely related sequences are merged first. This immediately brings up the problem of how to align a sequence to two or more sequences already in an alignment.

In Waterman and Perlwitz (1984), some mathematical aspects of this problem of merging alignments are studied. The idea is to take the positions of the aligned sequences

	Sequence position			
Sequence	1	2		m
1	l_{11}	l_{12}	...	l_{1m}
2	l_{21}	l_{22}	...	l_{2m}
⋮				
N	l_{N1}	l_{N2}	...	l_{Nm}

and to summarize the statistics of the letters in positions i, $1 \le i \le m$ to form a 'weighted average sequence.' For example, for letter α, the quantity

$$f_{i\alpha} = \sum_{k=1}^{N} 1\{a_{ki} = \alpha\}/N$$

is the fraction of α's in column i. For sequences composed of letters from an alphabet $\mathcal{A}$ of size d, set $\mathbf{f}_i = (f_{i1}, \ldots, f_{id})$. Above, for example, $f_6 = (f_{6,A}, f_{6,G}, f_{6,T}, f_{6,C}) = (2/7, 2/7, 2/7, 1/7)$. This allows us to make use of more information about the letters in a given position of an alignment and not be restricted to a consensus letter. All that is required to score an alignment of a sequence with a weighted average sequence is a measure of similarity $P_i(l)$ between the letter l and the statistics of position i. Then the weighted average sequence $\mathbf{F} = \mathbf{f}_1\mathbf{f}_2 \ldots \mathbf{f}_m$ can be aligned to a sequence by any of the standard algorithms.

The most popular and useful implementation of these ideas is known as profile analysis (Gribskov *et al.*, 1987). The position-dependent profile score, denoted by $P_i(l)$, depends on the letter l and the distribution $\mathbf{f}_i$ of letters in position i. The score $P_i(l)$ is high when letter l is often found in position i in the alignment. The score X for a particular alignment of letters can then be given by summing up the scores $P_i(l)$ over all the letters in the alignment. We may represent the profile $\mathbf{P} = \{P_i(l)\}$ as an array, with the i^{th} column given by $P_i(l)$, as l ranges over the letters in the sequence alphabet $\mathcal{A}$.

One simple measure can be derived from individual pairwise substitution weights, where aligning letter x with letter y receives score $s(x,y)$. The $P_i(l)$ can be defined as the average score of l under $\mathbf{f}_i$ by $P_i(l) = \sum_k s(l,k)f_{ik}$. In many applications of profile analysis, the Smith–Waterman (Smith and Waterman, 1981) dynamic programming algorithm for local alignments is used to find significant matches to all or part of the profile. The statistical distribution of Smith–Waterman scores is well studied. See Arratia *et al.* (1988) and Karlin and Altschul (1990) for statistical results when indels are not allowed. Waterman and Vingron (1994) numerically extend the Poisson approximation to allow indels. However, most profiles are developed for specific motifs, and it is frequently desirable to determine where in a sequence the entire profile best fits. Then the score for aligning a profile $\mathbf{P}$ with a sequence $\mathbf{l} = l_1 l_2 \ldots l_{n+m-1}$ is the maximum profile score over all sets of m consecutive letters $l_j l_{j+1} \ldots l_{j+m-1}$, that is,

$$N_n = \max_{1 \le j \le n} X_j,$$

where

$$X_j = \sum_{i=1}^{m} P_i(l_{i+j-1}) \quad j = 1, 2, \ldots, n$$

Although the usual dynamic programming algorithms allow indels to be included if desired, for the statistical results presented in this paper, indels are not considered.

To evaluate the statistical significance of M, the score N properly normalized, it is first necessary to understand the distribution of the n individual profile scores X_j. Here is an easy heuristic. If $L_1L_2 \ldots L_{n-m+1}$ are independent letters, then each X_j is the sum of independent random variables. Hence, if the profile $\mathbf{P}$ is well behaved, each X_j will be approximately normally distributed by the central limit theorem. Furthermore, if $|j-k| \geq m$, X_j and X_k are determined by sets of disjoint letters and are therefore independent. In other words, X_1, $X_2, \ldots, X_n$ is an m-dependent sequence of random variables, each with a distribution close to normal. Because the maximum of an m-dependent normal sequence, suitably standardized, has an asymptotic extreme value distribution with distribution function

$$G(x) = \exp(-\exp(-x))$$

it is reasonable to conjecture that the score M has this limiting extreme value distribution as well.

There are a number of technical difficulties in proving this conjecture. First, to invoke the central limit theorem, each X_j must be the sum of a growing number of terms $m \to \infty$. Further, to obtain the asymptotic extreme value distribution, it is necessary to take the maximum of a growing number $n \to \infty$ of profile scores X_j. Therefore, we need to consider scores $X_1, X_2, \ldots, X_n$ constructed from a profile table with m columns as m, $n \to \infty$; we will achieve this behavior by taking m as a function of n. Hence, for each n, $X_1, X_2, \ldots, X_n$ is an m-dependent sequence where $m = m_n$ depends on n.

With the number of columns m now large, one must insure that the columns are not too correlated. In biological profiles, typical columns will usually be only slightly correlated; however, it may be the case that some columns will be highly correlated for functional or structural reasons. Although the technical condition Equation (10) in the next section that the maximum absolute column correlation η is strictly less than 1 is always satisfied in practice for any finite table with no two columns identical, it is still of interest to compute η for a given table. For the immunoglobin table of Gribskov *et al.*, (1987), the maximum column correlation η equals 0.94, below the upper bound of 1. In the next section, we present our model for the profile problem, including a simple set of conditions that we require our sequence of tables to satisfy, thus making precise our notion of 'well behaved' mentioned above.

Theorem 1 in the section "Convergence to Extreme Value Distribution" establishing the convergence in distribution of the maximum profile score to the extreme value is proved using a version of Chen–Stein Poisson approximation. Here is a sketch of the argument. First, with constants a_n, c_n given in Lemma 4, we show that for the test level $u = u_n = x/a_n + c_n$, the probability that a standardized profile score will exceed u is well approximated by $\psi(u)$, the probability that a standard normal variate exceeds u. Hence, the average number of exceedences will be close to $\mu_n = n\psi(u)$. As the number of times X_j exceeds u will be approximately Poisson, the probability there are no exceedences, which happens if and only if the maximum does not exceed u, is approximately the same as $e^{-\mu_n}$, the probability that a Poisson variable with mean μ_n takes the value zero. As $\mu_n \to e^{-x}$, the probability that the maximum is bounded by u tends to $\lim_{n\to\infty} e^{-\mu_n} = e^{-e^{-x}}$. In addition, Theorem 1 provides a bound on the rate of convergence to this limit by the Chen–Stein Poisson approximation method; this bound gives information on the quality of the approximation.

Necessary lemmas are presented in the section "Lemmas." In the section "Results of Simulation and Database Search," we study the behavior of a specific profile on real biological data and consider several factors that affect the fit to the extreme value distribution by simulation experiments. Some needed technical results appear in the Appendix, as well as a result indicating the necessity of a feature of our profile model.

PROFILE MODEL

So that the distribution of profile scores can be approximated by the normal using the central limit theorem, each profile score needs to be represented as a sum with a growing number of terms m. Therefore, we consider a sequence of problems indexed by n, the number of profile scores, with $n \to \infty$, and the number of columns m depending on n, with $m = m_n$ also tending to ∞.

Let $L_1, L_2, \ldots, L_{n+m-1}$ be independent identically distributed letters over an alphabet $\mathcal{A}$. For given n, we consider a profile table with $m = m_n$ columns represented as the array $\mathbf{P}^{(n)} = \{P_i^{(n)}\}_{\{i:1\leq i\leq m\}}$, where each $P_i^{(n)}$ is a real valued function on $\mathcal{A}$. As each profile score is a sum of m terms, to apply the central limit theorem we are required to have $\lim_{n\to\infty} m_n = \infty$.

We form the profile score at position j by calculating the 'moving average'

$$X_j^{(n)} = \sum_{i=1}^{m} P_i^{(n)}(L_{i+j-1}),\ j = 1, 2, \ldots, n. \tag{1}$$

For each n, the distribution of $(X_j^{(n)}, X_{j+\delta}^{(n)})$ does not depend on j for $1 \le j, j+\delta \le n$. It follows that $X_j^{(n)}$ are identically distributed and that the covariance $\sigma_{\delta,n}^2 = \mathrm{Cov}(X_j^{(n)}, X_k^{(n)})$ depends only on n and $\delta = |k-j|$. Note that $X_j^{(n)}, X_k^{(n)}$ are independent for $\delta \ge m$, and so $\mathrm{Cov}(X_j^{(n)}, X_k^{(n)}) = 0$ for these δ.

As the $X_j^{(n)}$ are identically distributed, we can find their common mean β_n by $EX_1^{(n)}$; hence

$$\beta_n = \sum_{i=1}^{m} EP_i^{(n)}(L_i) = \sum_{i=1}^{m} EP_i^{(n)}(L),$$

where L is a letter with the common letter distribution on $\mathcal{A}$. For $0 \le \delta < m$, we may calculate the covariance of two scores from sequence segments δ apart by

$$\sigma_{\delta,n}^2 = \mathrm{Cov}(X_1^{(n)}, X_{1+\delta}^{(n)}) = E\left\{\sum_{i=1}^{m} [P_i^{(n)}(L_i) - EP_i^{(n)}(L)] \sum_{j=1}^{m} [P_j^{(n)}(L_{j+\delta}) - EP_j^{(n)}(L)]\right\}.$$

Using that the letters are independent, and that terms of the form $P_i^{(n)}(L_i) - EP_i^{(n)}(L)$ have mean zero, we see that

$$\sigma_{\delta,n}^2 = \sum_{i=1}^{m-\delta} E\{[P_{i+\delta}^{(n)}(L_{i+\delta}) - EP_i^{(n)}(L)][P_i^{(n)}(L_{i+\delta}) - EP_{i+\delta}^{(n)}(L)]\}$$

$$= \sum_{i=1}^{m-\delta} \mathrm{Cov}(P_{i+\delta}^{(n)}(L_{i+\delta}), P_i^{(n)}(L_{i+\delta})); \tag{2}$$

in particular, the common variance of the scores $X_j^{(n)}$ is given by

$$\sigma_n^2 = \sigma_{0,n}^2 = \sum_{i=1}^{m} \mathrm{Var}(P_i^{(n)}(L)).$$

If the alphabet $\mathcal{A}$ is the set of d letters $\alpha_1, \alpha_2, \ldots, \alpha_d$ with frequencies $f_1, f_2, \ldots, f_d$, then

$$\beta_n = \sum_{i=1}^{m} \sum_{k=1}^{d} P_i^{(n)}(\alpha_k) f_k, \tag{3}$$

and

$$\sigma_n^2 = \sum_{i=1}^{m} \mathrm{Var}(P_i^{(n)}(L_i)) = \sum_{i=1}^{m} \left\{\sum_{k=1}^{d} [P_i^{(n)}(\alpha_k)]^2 f_k - \left[\sum_{k=1}^{d} P_i^{(n)}(\alpha_k) f_k\right]^2\right\}. \tag{4}$$

We standardize $X_j^{(n)}$ in order to have variables with mean zero and variance 1:

$$Y_j^{(n)} = (X_j^{(n)} - \beta_n)/\sigma_n; \tag{5}$$

as $X_j^{(n)} - \beta_n = \sum_{i=1}^{m} [P_i^{(n)}(L_{i+j-1}) - EP_i(L)]$, we may in what follows assume without loss of generality that the profile table columns P_i have mean zero with respect to the common letter distribution on $\mathcal{A}$.

Define the correlation ·

$$\rho_\delta^{(n)} = \mathrm{Cov}(Y_j^{(n)}, Y_k^{(n)}) = (\sigma_{\delta,n}^2/\sigma_{0,n}^2) \text{ for } \delta = |k-j|.$$

Note that $Y_j^{(n)}, Y_k^{(n)}$ are independent for $\delta \ge m$, and so $\rho_\delta^{(n)} = 0$ for these δ.

We study the distribution of

$$M_n = \max_{1 \le j \le n} Y_j^{(n)}. \tag{6}$$

Define the norm of a profile table column by $\|P\| = \sup_{x \in \mathcal{A}} |P(x)|$, and assume that the arrays $\mathbf{P}^{(n)}$ satisfy

$$\sup_{n,1\le i\le m} \|P_i^{(n)}\| = K < \infty. \tag{7}$$

Assume further that there exists $A > 0$ such that

$$\mathrm{Var}P_i^{(n)}(L) \ge A \text{ for all } n \text{ and } 1 \le i \le m, \tag{8}$$

and that the maximum column correlation is bounded strictly by 1:

$$\rho = \sup_{1\le n<\infty} \{|\rho_\delta^{(n)}| : 1 \le n < \infty \text{ and } 1 \le \delta < m\} < 1. \tag{9}$$

As condition (9) may be difficult to verify, we present a condition easier to check that insures condition (9); in particular, condition (9) is satisfied whenever the maximum correlation

$$\eta = \sup_{n,1\le i\ne j\le m} |\mathrm{Cor}(P_i^{(n)}(L), P_j^{(n)}(L))| < 1. \tag{10}$$

To see that (10) implies (9), begin with $\sigma_{\delta,n}^2$ as given by (2), to see that, with $a_i^2 = \mathrm{Var}\, P_i^{(n)}(L)$ and $\delta \ne 0$,

$$|\sigma_{\delta,n}^2| \le \sum_{i=1}^{m-\delta} |\mathrm{Cov}(P_{i+\delta}^{(n)}(L_{i+\delta}), P_i^{(n)}(L_{i+\delta}))| \le \sum_{i=1}^{m-\delta} \eta^{a_i + \delta^{a_i}};$$

but as $\sum_{i=1}^{m-\delta} a_i a_{i+\delta} \le \sum_{i=1}^{m} a_i^2 = \sigma_{0,n}^2$ by the Cauchy–Schwartz inequality, we have that $|\sigma_{\delta,n}^2| \le \eta\sigma_{0,n}^2$, and hence $\rho_\delta^{(n)} \le \eta$ as desired.

The generality of considering an array of functions indexed by n may at first appear unnecessary. Indeed, in the Appendix we give an example, in the case where $\mathcal{A} = [0, 1]$, of a table $\mathbf{P}^{(n)} = \{P_i\}_{i:1\le i\le m}$ constructed from the first m of a collection of fixed functions $P_1, P_2, \ldots$ that satisfies condition (10) and therefore (9). However, Proposition (1) in the Appendix shows that it is not possible to construct a profile table $\mathbf{P}^{(n)}$ that satisfies condition (10) using the first n of a collection of functions, even in the simple case of a uniform distribution over the finite alphabet $\{1, 2, \ldots, d\}$ with d fixed. This difficulty can be avoided if we are allowed to consider an array constructed from functions $P_i^{(n)}$ which depends on n, defined on $\{1, 2, \ldots, d\}$ where d may also depend on n, for then it is not difficult to construct examples where (10) and therefore (9) are satisfied.

In what follows, we write $a_k \asymp b_k$ when $0 < \liminf_{k\to\infty}|a_k/b_k| \le \limsup_{k\to\infty}|a_k/b_k| < \infty$, $a_k = O(b_k)$ if $\limsup_{k\to\infty}|a_k/b_k| < \infty$ and $a_k = o(b_k)$ if $\limsup_{k\to\infty}|a_k/b_k| = 0$. Constants will be denoted $C_1, C_2, \ldots$, each not necessarily the same at each occurrence. We drop the superscript n when there is no danger of confusion.

LEMMAS

Let Z be a standard normal random variable; denote the density of Z by $\phi(u)$ and $P(Z > u)$ by $\Psi(u)$. Recall that $X_1, X_2, \ldots, X_n$ and therefore $Y_1, Y_2, \ldots, Y_n$ are identically distributed, that both the X and Y variates are defined as the sum of m terms, and that we may assume without loss of generality that the functions P_i have mean zero with respect to the common letter distribution on $\mathcal{A}$. The first lemma shows that the tail probabilities of Y_i are asymptotic to the tail probabilities of a standard normal even for moderate test levels.

Lemma 1 *For Y_1 as in (5) and $v_m = o(m^{1/6})$,*

$$\frac{P(Y_1 > v_m)}{\Psi(v_m)} \to 1 \text{ as } m \to \infty.$$

Proof: For each n, we apply Theorem (2) with $q = m$ to the mean zero random variables $R_i = P_i^{(n)}(L_i)$. The variables are bounded by assumption (7), so we may set $M = K$, and condition (8) implies that condition (16) is satisfied with $B = A$. As $v_m = o(m^{1/6})$, the error term f of Theorem (2) tends to zero as $m \to \infty$. □

The next lemma gives a bound on the tail probabilities of the joint distribution of (Y_j, Y_k).

Lemma 2 *Let $Y_1, Y_2, \ldots, Y_n$ be defined as in (5), and $0 \le \rho < 1$ as in condition (9). Then there exists a constant C such that if $v_m = o(m^{1/6})$, then for all $1 \le |j - k| < m$, $1 \le j, k \le n$,*

$$p_{jk} = P(Y_j > v_m, Y_k > v_m) \le C[\Psi(v_m)]^{\frac{2}{1+\rho}} v_m^{\frac{1-\rho}{1+\rho}}.$$

Proof: Note that

$$P(Y_j > v_m,\, Y_k > v_m) \le P(Y_j + Y_k > 2v_m).$$

With $\delta = |k - j|$, note that $\mathrm{Var}(Y_j + Y_k) = 2(1 + \rho_\delta^{(n)})$. Using the bound $|\rho_\delta^{(n)}| \le \rho < 1$ on the correlations given by condition (9), we obtain the bound

$$P(Y_j + Y_k > 2v_m) \le P\left(\sqrt{\frac{Y_j + Y_k}{2(1 + \rho_\delta^{(n)})}} > \sqrt{\frac{2}{1+\rho}} v_m \right) = r_{jk}, \text{ say.} \tag{11}$$

We will apply Theorem (2) with $q = m + \delta$. Assuming without loss of generality that $j < k$, let

$$R_i^{(n)} = \begin{cases} P_i^{(n)}(L_{i+j-1}) & \text{for } 1 \le i \le \delta, \\ P_i^{(n)}(L_{i+j-1}) + P_{i-\delta}^{(n)}(L_{i+j-1}) & \text{for } \delta + 1 \le i \le m, \\ P_{i-\delta}^{(n)}(L_{i+j-1}) & \text{for } m + 1 \le i \le m + \delta, \end{cases}$$

so that $Y_j + Y_k = \sum_{i=1}^{m+\delta} R_i^{(n)}$. We note that $ER_i^{(n)} = 0$, as the profile table rows have mean zero with respect to the letter distribution on $\mathcal{A}$; furthermore, these variates are bounded by assumption (7), and we may set $M = 2K$ in Theorem (2). Using (8) we have

$$\mathrm{Var}(Y_j + Y_k) \ge 2A[\delta + (m - \delta)(1 - \rho)],$$

and hence condition (16) is satisfied with $B = 2A(1 - \rho)$. As $v_m = o(m^{1/6})$, Theorem 2 yields that r_{jk} is asymptotically uniformly bounded by a constant times (and is asymptotic to) $\Psi(\sqrt{\frac{2}{1+\rho}}\, v_m)$. Using the bound

$$[u^{-1} - u^{-3}]\phi(u) < \Psi(u) < u^{-1}\,\phi(u),$$

holding for all $u > 0$ (Feller, Chapter VII, Lemma 2) we see that $\Psi(\sqrt{\frac{2}{1+\rho}}\, v_m)$ is bounded by

$$\phi\left(\sqrt{\frac{2}{1+\rho}}\, v_m \right) v_m^{-1} \le C_1[\phi(v_m)]^{\frac{2}{1+\rho}}\, v_m^{-1} \le C_2[\Psi(v_m)]^{\frac{2}{1+\rho}}\, v_m^{\frac{1-\rho}{1+\rho}},$$

as required. □

The following lemma is a consequence of Theorem 1 of Arratia *et al.* (1989),

Lemma 3 *Let $I = \{1, 2, \ldots, n\}$ and for each $j \in I$, let B_j be a Bernoulli random variable with $p_j \equiv P(B_j = 1) = 1 - P(B_j = 0) \in (0, 1)$. Let*

$$W_n \equiv \sum_{j \in I} B_j, \text{ and } \lambda_n \equiv EW_n = \sum_{j \in I} p_j. \tag{12}$$

For each $j \in I$, suppose there is a set of dependence for B_j, $N_j \subset I$, with $j \in N_j$, such that

$$B_j \text{ is independent of } \{B_k : k \notin N_j\}.$$

Define

$$b_1 \equiv \sum_{j \in I} \sum_{k \in N_j} p_j p_k \text{ and}$$

$$b_2 \equiv \sum_{j \in I} \sum_{j \ne k \in N_j} p_{jk}, \text{ where } p_{jk} \equiv E(B_j B_k).$$

Then

$$|P(W_n = 0) - e^{-\lambda_n}| \le b_1 + b_2.$$

Corollary 1 *If $\lambda_n \to \lambda$, and $b_1 + b_2 \to 0$ as $n \to \infty$, then $P(W_n = 0) \to e^{-\lambda}$.*

CONVERGENCE TO EXTREME VALUE DISTRIBUTION

Let

$$M_n = \max_{1 \le j \le n} Y_j^{(n)},$$

$$a_n = (2 \log n)^{1/2}, \tag{13}$$

$$c_n = (2 \log n)^{1/2} - 1/2(2 \log n)^{-1/2} (\log \log n + \log 4\pi). \tag{14}$$

The calculation demonstrating the next lemma is standard and can be found in Galambos (1987).

Lemma 4 *For given x, let*

$$u_n = \frac{x}{a_n} + c_n;$$

then

$$\lim_{n \to \infty} n\Psi(u_n) = e^{-x}.$$

Our main result proving convergence to the extreme value distribution, with bounds on the rate of convergence, appears next.

Theorem 1 *Let* $\mathbf{L}$ *be a sequence* $L_1, L_2, \ldots, L_{n+m-1}$ *composed of independent and identically distributed letters over an alphabet* $\mathcal{A}$. *Suppose that the profile tables* $\mathbf{P}^{(n)}$ *satisfy conditions (7), (8), and (9) above. Let* M_n *be the maximum profile score, as given in equation (6), and for given x, let*

$$\lambda_n = nP(Y_1 > u_n) \text{ with } u_n = x/a_n + c_n,$$

with a_n, c_n *as in* (13) *and* (14). *With* $0 \le \rho < 1$ *by* (9), *suppose that* $m \asymp n^{\kappa}$ *where* $\kappa \in (0, \frac{1-\rho}{1+\rho})$. *Then,*

$$|P(a_n(M_n - c_n) \le x) - e^{-\lambda_n}| = o(n^{-\gamma}) \text{ for every } \gamma \in (0, \tfrac{1-\rho}{1+\rho} - \kappa).$$

Proof: For $j \in I = \{1, 2, \ldots, n\}$, let

$$B_j = I\{Y_j > u_n\} \text{ and } W_n = \sum_{j=1}^{n} B_j.$$

Note

$$\{a_n(M_n - c_n) \le x\} = \{W_n = 0\}.$$

With

$$\mu_n = n\psi(u_n), \tag{15}$$

we have

$$\lim_{m \to \infty} \lambda_n/\mu_n = 1$$

using Lemma 1 and that $u_n = o(m^{1/6})$. As $\mu_n \to e^{-x}$ by Lemma 4,

$$\lambda_n \to e^{-x} = \lambda \text{ as } n \to \infty.$$

Taking $N_j = \{k : |k - j| < m\}$, the independence condition of Lemma 3 is satisfied. Now

$$b_1 = \sum_{j \in I} \sum_{k \in N_j} p_j p_k = |I||N_j| p_j^2 \le 2mn[P(Y_1 > u_n)]^2 = 2m\lambda_n^2/n = O(n^{\kappa-1}).$$

Now, using Lemma 2,

$$b_2 = \sum_{j \in I} \sum_{j \ne k \in N_j} p_{jk} \le C_1 |I||N_j| [\Psi(u_n)]^{\frac{2}{1+\rho}} u_n^{\frac{1-\rho}{1+\rho}} \le C_2 n^{1+\kappa-\frac{2}{1+\rho}} u_n^{\frac{1-\rho}{1+\rho}} \mu_n^{\frac{2}{1+\rho}}.$$

Note first that μ_n converges to a constant, and so does not affect the order of b_2. Noting that $b_1 = o(b_2)$ completes the proof. □

We have the immediate

Corollary 2

$$P(a_n(M_n - c_n) \leq x) \to \exp(-e^{-x}).$$

Proof: We have that $\lambda_n \to \lambda$ and $b_1 + b_2 \to 0$ as $n \to \infty$; hence the corollary follows from Corollary (1).

RESULTS OF SIMULATION AND DATABASE SEARCH

Theorem 1 shows that the score M for a random sequence of length $n + m - 1$ has a distribution close to that of the maximum of n normal variates, and therefore close to the extreme value distribution. The theorem is an asymptotic result, and so the quality of the extreme value approximation should be explored for finite samples. The practical fit to this theoretical distribution is studied in the $q - q$ plot (Fig. 1). In Fig. 1, the immunoglobulin (Ig) profile of length $m = 49$ of Gribskov *et al.* (1987) was applied to Newat, a database of unique sequences assembled by Doolittle (1981).

To test this approximation for finite sequence lengths and finite profile table size, the Ig profile was applied

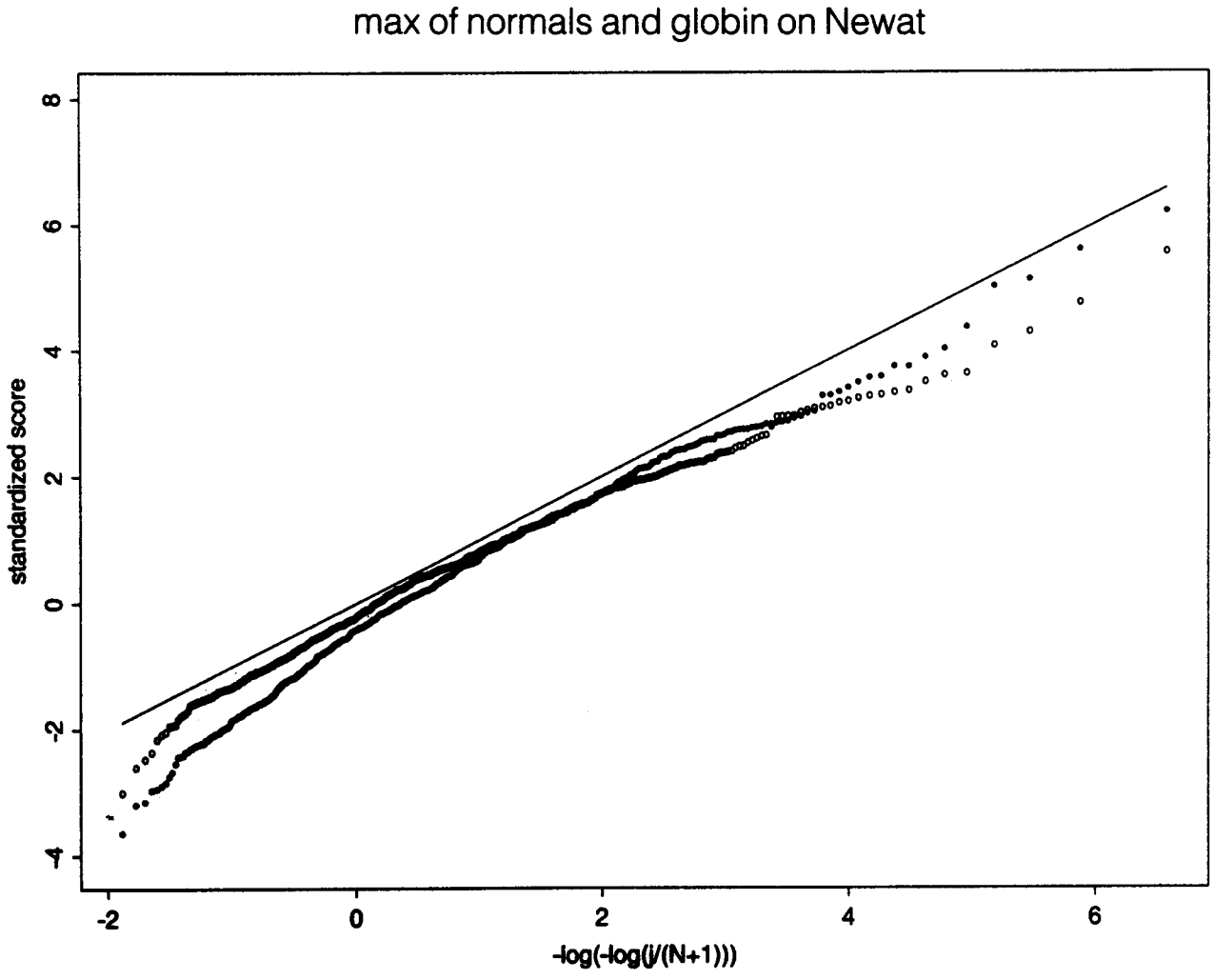

FIG. 1. $q - q$ Plots. The closed circles (●) are obtained by applying the Ig profile table to the Newat database. For each database sequence, there is an open circle (○) obtained by taking the maximum of a corresponding number of independent normals.

to the $N = 714$ sequences in the Newat database of length 150 or more; we refer to this subset of the database as the Newat database in what follows. The mean and variance of the Ig profile table was calculated using formulas (3) and (4), respectively, using letter frequencies f_k that of the Newat database. The scores X_j, Y_j, and M were calculated according to equations (1), (5), and (6), respectively. Each score M was then scaled by the constants a_n, c_n given in (13) and (14) to yield the standardized score $a_n(M - c_n)$. For example, a sequence of length 682 yields $n = 682 - 49 + 1 = 634$ profile scores X_j, and so is scaled with this value of n. The collection of standardized scores for the Newat database was computed and ordered; from the theorem, the j^{th} largest of the N Newat scores should be approximately equal to the $j/(N + 1)$ quantile of the extreme value distribution. Each closed circle (•) in Fig. 1 corresponds in this way to a sequence in the Newat database; if the closed circle represents, say, the j^{th} largest standardized score, the vertical axis gives the value of this score, and the horizontal axis the $j/(N + 1)$ quantile of the extreme value distribution. For comparison, we note that points in the figure generated from scores drawn from the extreme value distribution would lie on the line $y = x$ in the figure.

For the Ig profile applied to the Newat database, there are two factors that affect the fit to the theoretical distribution. First, there is an error incurred in approximating the profile scores X by the normal distribution, and next, an error incurred by approximating the distribution of the maximum of normals by its asymptotic limit, the extreme value distribution.

To study these two factors, the following simulation experiment was performed. For each sequence in the Newat database, a corresponding 'ideal' standardized score was generated by taking the maximum of a number of normal variates appropriate for that sequence. In particular, a sequence of length $n + m - 1$, has maximum score M, which is the maximum of the scores $X_j, j = 1, \ldots, n$ each of which is approximated theoretically by independent normal variates with mean β and variance σ^2. Correspondingly, one can generate n independent normals with mean β and variance σ^2, and consider the 'ideal' score M obtained by taking the maximum of these normal variates. Such an ideal standardized score can be generated for each sequence in the database, and the resulting collection of standardized scores then ordered. Each open circle (◦) in Fig. 1 corresponds in this way to the ideal score of a sequence in the Newat database. If the open circle represents, say, the j^{th} largest standardized ideal score, the vertical axis gives the value of this score, and the horizontal axis the $j/(N + 1)$ quantile of the extreme value distribution.

Hence, the discrepancy between the graph of open circles and the $y = x$ line demonstrates the error incurred by approximating the maximum of a finite number of normals by the extreme value distribution. This convergence is known to be slow (see Hall, 1980), and we cannot expect the distribution of profile scores to be any better approximated by the extreme value distribution than is the distribution of the maximum of a corresponding number of independent normal variates.

However, one can observe in Fig. 1 that the graphs of closed circles and open circles are somewhat close; in other words, there is only some little discrepancy between the maximum value of the profile scores and the maximum value of independent normals with the same mean and variance.

We note that even when the profile scores are well approximated by the maximum of independent normals, such as in Fig. 1, the extreme value distribution, represented by the line $y = x$, is not yet attained. This lack of fit is due to the slow rate of convergence of the distribution of the maximum of independent normals to the extreme value, and so is improved only when scoring longer sequences.

However, one may avoid the difficulty due to the slow rate of convergence to the extreme value distribution, even when the sequences are not long, by approximating the distribution of profile scores by the maximum of independent normals directly. The extreme value distribution is attained in the limit when $n \to \infty$ as $P(a_n(M_n - c_n) \leq x)$ is approximately $\exp(-\lambda_n)$, where $\lambda_n = nP(Y_1 > u_n)$ and λ_n is asymptotic to $\mu_n = n\Psi(u_n)$, and $\mu_n \to e^{-x}$. From Fig. 1, it is clear that a better approximation to $P(a_n(M_n - c_n) \leq x)$ would be obtained by $\exp(-\mu_n)$, since this quantity more directly approximates the event that the maximum of independent normals lie below the test level u_n. These issues are explored in more detail in Arratia *et al.* (1990).

Each comparison of a profile with a sequence produces a score. Without a result like that of Theorem 1 to approximate p values, comparisons must be ranked by score. Because long sequences have more opportunity to achieve good matches to the profile, and therefore high scores, by chance alone, ranking by scores not adjusted for length can be misleading. In Table 1 we show the 25 sequences from Newat with Ig profile scores with the smallest p-values. Notice that the sixth smallest p-value of 0.010 is obtained by ORPE with a score of 171, which is smaller than the next 5 scores, each with a larger p-value. In fact, the score of 171 is also obtained from an *E. coli* potassium transport protein with a p-value of 0.049; this sequence has length 682, while ORPE has length 120. We see therefore that the approximate p-values given account for the fact that a short sequence is less likely to match the profile well than a longer one by chance alone.

TABLE 1. Ig PROFILE COMPARISON IN NEWAT

org	score	length	p-value	sequence description
TCHR	196	312	0.001180	HUMAN T-CELL-SPECIFIC PROTEIN
TUAP	193	451	0.002149	PIG BRAIN TUBULIN, ALPHA
LOAV	193	863	0.003338	HTLV-III ENV-LOR
IGDH	189	512	0.003886	HUMAN IMMUNOGLOBULIN DELTA CHAIN
RH11	183	460	0.007665	RHINOVIRUS 14 (COMMON COLD)
ORPE	171	120	0.010093	TORPEDO CALIFORNICA ACETYLCH. RECEPTOR
APEC	180	449	0.010956	E. COLI ALKALINE PHOSPHATASE
OMPA	177	345	0.012805	E. COLI OMPA MAJOR SURFACE PROTEIN
G6PD	179	482	0.013155	HUMAN GLUCOSE-6-PHOSPHATE DEHYDROGENASE
ATPC3	175	293	0.014217	RABBIT MUSCLE ATPASE-CALCIUM
CC4H	171	188	0.015351	HUMAN COMPLEMENT C4
KDPC	171	190	0.015500	E. COLI POTASSIUM TRANSPORT PROTEIN, KDP-C
EKEM	178	569	0.017123	MOUSE EPIDERMAL KERATIN
PGAP	171	225	0.018074	PSEUDOMONAS PUTIDA 2-KTO-3-DXY-6-PHSPH ALD
PRSN	172	278	0.019423	SERRATIA ZINC PROTEASE
ATYE	174	423	0.022003	E.COLI AMINO ACYL-tRNA SYNTHETASE, TYROS.
INMG	166	155	0.022292	MOUSE IMMUNE TYPE INTERFERON
TCRH	171	306	0.023838	HUMAN THYMOCYTE T-CELL RECEPTOR
TCRA	170	273	0.024217	CRUCIFER FAMILY CRAMBIN, PLANT SEED PROTEIN
EFGE	164	144	0.025702	E. COLI ELONGATION FACTOR G
CY2R	161	116	0.027828	RHODOPSEUDOMONAS CAPSULATA CYTOCHROME C2
FALH	174	610	0.030259	HUMAN FIGRINOGEN ALPHA CHAIN
ACRB	172	493	0.032275	TORPEDO CALIFORNICA ACETYLCH. RECEPTOR
APRM	164	180	0.032381	MOUSE ADENINE PHOSHORIBOSYL TRANSFERASE
MPDD	172	508	0.033147	DROSOPHILA MYSTERY PROTEIN, "DUEY", C31

APPENDIX

The following Theorem is an adaptation of Feller's Theorem XVI.7,3 on large deviations with a uniform bound under the assumption of bounded variates. Because the proof of Theorem (2) is obtained by a minor modification of Fellers proof, it is omitted below.

Theorem 2 *Let $R_1, R_2, \ldots$ be real valued, independent mean zero random variables bounded by a constant M, and set*

$$S_q = \sum_{i=1}^{q} R_i \text{ and } s_q^2 = \sum_{i=1}^{q} \text{Var}(R_i).$$

Suppose that there exists $B > 0$ such that

$$s_q^2 \geq Bq. \tag{16}$$

Then there exists a constant C depending only on M and B such that

$$P(S_q/s_q > x) = \Psi(x)(1 + f) \text{ where } |f| \leq Cx^3/\sqrt{q},$$

for all sequences $x = x_q \to \infty$ such that $x_q/s_q \to 0$ as $q \to \infty$, where Ψ denotes the upper tail of the standard normal distribution.

As discussed in the Introduction, the generality of considering $\mathbf{P}^{(n)}$ as an array of functions indexed by n may at first appear unnecessary. Indeed, this generality is not required if we were to consider the case where L is distributed uniformly on the 'alphabet' [0, 1] and $P_1^{(n)} = P_i$ is the i^{th} nonconstant element of a bounded, orthonormal system on $L^2[0, 1]$. If we take, say, the Rademacher functions then $EP(L) = 0$, $\text{Var}P(L) = 1$ (so we may take $C_1 = C_2 = 1$), and, in particular, $\rho = 0$ as $|\text{Cor}(P_i, P_j)| = 0$ for $i \neq j$ by orthogonality. The necessity arises when we consider functions defined on an alphabet $\{1, 2, \ldots, d\}$ of *fixed* size; if we insist that the distribution

of L is, say, uniform over this finite set and that again $EP(L) = 0$ and $\mathrm{Var}P(L) = 1$ then that condition (10) cannot be likewise satisfied is demonstrated in the following proposition.

Proposition 1 *Let* $L \in \{1, 2, \ldots, d\}$, *and let* $P_1, P_2, \ldots,$ *be such that* $EP_i(L) = 0$ *and* $0 < C_1 \le \mathrm{Var}(P_i(L)) \le C_2$ *for* $i = 1, 2, \ldots$. *Then for any* $m = m_n \to \infty$,

$$\lim_{n\to\infty} \max_{1\le i\ne j\le m} |\mathrm{Cor}(P_i(L), P_j(L))| = 1.$$

Proof: Let $f_k = P(L = k)$, $k = 1, 2, \ldots, d$. We may assume without loss of generality that all f_k are positive. For $\mathbf{x},\mathbf{y} \in \mathbf{R}^d$ define the inner product and norm

$$\langle \mathbf{x},\mathbf{y}\rangle = \sum_{k=1}^{d} x_k y_k f_k, \quad \|\mathbf{x}\|^2 = \langle \mathbf{x},\mathbf{x}\rangle,$$

and let $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbf{R}^d$. Define

$$\mathcal{C} = \{\mathbf{x} \in \mathbf{R}^d : \langle \mathbf{x}, \mathbf{1}\rangle = 0, \|\mathbf{x}\|^2 \in [C_1, C_2]\}.$$

The condition that $P_i(L)$ have mean zero and variance in the interval $[C_1, C_2]$ is equivalent to the condition that the corresponding vector $\mathbf{x}$ with components $x_k = P(k)$ lie in $\mathcal{C}$. Furthermore, if $\mathbf{x}$ and $\mathbf{y}$ are the vectors that correspond to the functions P_i, P_j, then $\mathrm{Cor}(P_i(L), P_j(L)) = \langle \mathbf{x},\mathbf{y}\rangle/(\|\mathbf{x}\|\,\|\mathbf{y}\|)$.

Take $0 < \varepsilon < \frac{1}{2}$ arbitrary and define for $x \in \mathcal{C}$

$$A(x) = \left\{ y \in \mathcal{C} : \left| \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \right| > 1 - \varepsilon \right\}.$$

For all $\mathbf{x} \in \mathcal{C}$, $A(\mathbf{x})$ is open and $\mathbf{x} \in A(\mathbf{x})$; hence the union of the sets $A(\mathbf{x})$ over $\mathbf{x} \in \mathcal{C}$ is an open cover of $\mathcal{C}$. Since $\mathcal{C}$ is compact, there exists $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N$ such that $A(\mathbf{v}_i)$, $i = 1, 2, \ldots, N$ cover $\mathcal{C}$. Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N+1}$ be the vectors in $\mathcal{C}$ corresponding to $P_1, P_2, \ldots, P_{N+1}$. Two of these vectors, say $\mathbf{v}_1$, $\mathbf{v}_2$ must lie in the same set, say, A_k. But $\mathrm{Cor}(P_1(L), P_k(L)) > 1 - \varepsilon$ and $\mathrm{Cor}(P_2(L), P_k(L)) > 1 - \varepsilon$ implies that $\mathrm{Cor}(P_1(L), P_2(L)) > 1 - 2\sqrt{2\varepsilon}$; since ε is arbitrary, the result follows. □

ACKNOWLEDGMENTS

This work supported by grants from the National Institutes of Health (M.S.W.) and the National Science Foundation (L.G. and M.S.W.). The authors would like to thank Professor Sam Karlin for stimulating discussion.

REFERENCES

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

Arratia, R., Goldstein, L., and Gordon, L. 1989. Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* 17, 9–25.

Arratia, R., Goldstein, L., and Gordon, L. 1990. Poisson approximations and the Chen–Stein method. *Statist. Sci.* 5, 403–434.

Carrillo, H., and Lipman, D. 1988. The multiple sequence alignment problem in biology. *SIAM J. Applied Math.* 48, 1073–1082.

Doolittle, R.F. 1981. Similar amino acid sequences: chance or common ancestry? *Science* 214, 149–159.

Doolittle, R.F., Hunkapiller, M.W., Hood, L.E., Devare, S.G., Robbins, K.C., Aaronson, S.A., and Antoniades, H.N. 1983. Simian sarcoma virus *onc* gene, v-*sis*, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science* 221, 275.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, Inc., Canada.

Galambos, J. 1987. *The Asymptotic Theory of Extreme Order Statistics.* Robert E. Krieger Publishing Company, Inc., Malabar, FL.

Gribskov, M., McLachlan, A.D., and Eisenberg, D. 1987. *Proc. Natl. Acad. Sci. USA* 84, 4355–4358.

Gusfield, P. 1993. Efficient methods for multiple sequence alignment with guaranteed error. *Bull. Math. Biol.* 55, 141–154.

Hall, P. 1980. Estimating probabilities for normal extremes. *Adv. Appl. Probab.* 12, 491–500.
Karlin, S., and Altschul, S.F. 1900. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87, 2264–2268.
Pevzner, P. 1992. Multiple alignment, communication cost, and graph matching. *SIAM J. Applied Math.* 56, 1763–1779.
Sankoff, D. 1975. Minimal mutation trees of sequences. *SIAM J. Applied Math.* 28, 35–42.
Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195.
Wang, L., and Jiang, T. On the complexity of multiple sequence alignment. *J. Comp. Biol.*, in press.
Waterman, M.S., and Perlwitz, M.D. 1984. Line geometrics for sequence comparisons. *Bull. Math. Biol.* 46, 567–577.
Waterman, M.S., and Vingron, M. 1994. Rapid and accurate estimates of statistical significance for sequence database searches. *Proc. Natl. Acad. Sci. USA* 91, 4625–4628.
Waterman, M.S., Smith, T.F., and Beyer, W.A. 1976. Some biosequence metrics. *Adv. Math.* 20, 367–387.

Address reprint requests to:
Dr. Larry Goldstein
Department of Mathematics
1042 W. 36th Place, DRB 155
University of Southern California
Los Angeles, CA 90089-1113

Received for publication December 30, 1993; accepted March 29, 1994.