Estimating statistical significance of sequence alignments

MICHAEL WATERMAN

Departments of Mathematics and Molecular Biology, University of Southern California, Los Angeles, California 90089-1113, U.S.A.

SUMMARY

Algorithms that compare two proteins or DNA sequences and produce an alignment of the best matching segments are widely used in molecular biology. These algorithms produce scores that when comparing random sequences of length n grow proportional to n or to $\log(n)$ depending on the algorithm parameters. The Azuma–Hoeffding inequality gives an upper bound on the probability of large deviations of the score from its mean in the linear case. Poisson approximation can be applied in the logarithmic case.

1. INTRODUCTION

Sequence comparison algorithms are widely applied to produce aligned amino acid and nucleotide sequences. The DNA databases, DDBJ, EMBL and GenBank, contain about 180×10^6 basepairs as of Spring 1994 and they double in size every two years. New DNA sequences are compared to the DNA databases and translations of DNA sequences into amino acid sequences are compared to the protein databases. These database searches find relationships of newly determined sequences to known sequences, providing hypotheses as to the evolution and function of the new sequences. (Barker & Dayhoff 1982; Doolittle *et al.* 1983). This is one of the ways that computation is essential to the practice of modern biology.

The sequence comparison algorithms produce scores representing the similarity of the molecules. If x and y are two aligned letters, $s(x, y)$ is the associated similarity score. A gap of k letters receives a score $-g(k)$. Thus with $s(x, y) = \mathbb{I}(x = y)$ and $g(k) = \delta k$, the alignment $\mathbb{A}$

goo d
ga –d

has score $1 + 0 - \delta + 1 = 2 - \delta = S(\mathbb{A})$. Note that there is a deletion of the second 'o' in good (or an insertion of 'o' between a and d in gad). The problem of sequence alignment is to find the highest scoring alignments. Insertions and deletions (indels) make alignment a hard computational problem.

As there are thousands of sequences in a database, it is not possible to look at each comparison. Instead the scores should be screened by estimates of statistical significance so that the scientist only examines the most statistically significant alignments.

In the next section a commonly used alignment algorithm is presented. When applied the random sequences of length n, the scores grow with n either proportional to n or proportional to $\log(n)$. The algorithm parameters determine this behaviour. In the linear growth region, the Azuma–Hoeffding inequality gives an upper bound for $\mathbb{P}(S - \mathbb{E}(S) > \gamma n)$. In the logarithmic region, Poisson approximation can be applied to give good estimates for the probability of large scores. Numerical studies are performed for both these approximations.

2. ALGORITHM

Our sequences will be $\mathbf{x} = x_1 x_2 \ldots x_n$ and $y = y_1 y_2 \ldots y_m$ for deterministic letters x_i and y_j and $\mathbf{X} = X_1 X_2 \ldots X_n$ and $\mathbf{Y} = Y_1 Y_2 \ldots Y_m$ for *iid* letters X_i and Y_j. For ease of exposition we take $s(x, y)$ to be the score of aligned letters and $g(k) = k\delta$ to be the penalty of a k letter indel. This makes a penalty of δ per deleted letter. The first algorithm is for global alignment, where all of $\mathbf{x}$ must be aligned with all of $\mathbf{y}$. The algorithm is an application of dynamic programming which solves the alignment problem by building up solutions to subproblems. Set $S_{i,j} = S(x_1 x_2 \ldots x_i, y_1 y_2 \ldots y_j)$. An alignment achieving score $S_{i,j}$ must end in one of these ways

$$\begin{matrix} x_i \\ - \end{matrix} \quad \text{or} \quad \begin{matrix} x_i \\ y_j \end{matrix} \quad \text{or} \quad \begin{matrix} - \\ y_j \end{matrix},$$

because $\begin{matrix} - \\ - \end{matrix}$, aligning deletions, is not valid in alignment. Optimality requires the alignment preceding the final aligned letters to be optimal if the overall alignment is. Therefore

$$S_{i,j} = \max\{S_{i-1,j} - \delta, S_{i-1,j-1} + s(x_i, y_j), S_{i,j-1} - \delta\}.$$

To begin the recursion set $S_{0,j} = -\delta j$ and $S_{i,0} = -\delta i$. This algorithm takes $O(nm)$ time.

Alignments are determined by tracing back from the optimal score $S(\mathbf{x}, \mathbf{y}) = S_{n,m}$ to determine the steps from $(0, 0)$ to (n, m).

Sequences that are known to be related by descent from a common ancestor should be aligned by global alignment. Many sequences have one or more

Phil. Trans. R. Soc. Lond. B (1994) **344**, 383–390
Printed in Great Britain

segments or intervals that align well but can be otherwise unrelated. In this case local alignment algorithms are recommended. The following local alignment algorithm of Smith & Waterman (1981) is a modification of the global alignment algorithm. Define

$$H(\mathbf{x},\mathbf{y}) = \max\{0; S(x_k \dots x_i, y_l \dots y_j): 1 \leqslant k \leqslant i \leqslant n, 1 \leqslant l \leqslant j \leqslant m\}.$$

While this definition requires solving

$$\binom{n+1}{2}\binom{m+1}{2}$$

separate alignment problems there is an $O(nm)$ algorithm for this problem too. Set

$$H_{i,j} = \max\{0; S(x_k \dots x_i, y_l \dots y_j): 1 \leqslant k \leqslant i, 1 \leqslant l \leqslant j\}.$$

Then the recursion is

$$H_{i,j} = \max\{H_{i-1,j} - \delta, H_{i-1,j-1} + s(x_i, y_j), H_{i,j-1} - \delta, 0\},$$

with $H_{i,j} = 0$ if either i or j are 0. The score $H(\mathbf{x},\mathbf{y}) = \max\{H_{i,j} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$ is the largest value of $H_{i,j}$.

In table 1*a* we show a simple local alignment example with $\mathbf{x}$ = TCTGACAAAGGCAAC, $\mathbf{y}$ = CGTCCAATAGCCAAT, $s(x,y) = +1$ if $x = y$, $s(x,y) = -1$ if $x \neq y$ and $\delta = 1$. The optimal local alignment has score $H = 6$ and the traceback in boxes yields the alignment

CAATAGCCAA
CAA–AGGCAA.

There may well be several local alignments of interest. Many intersect the set of optimal local alignments, differing in small ways from the optimal alignments. These are not of the most interest, at least in an initial look at the sequence comparison. Instead we ask if there are any other distinct alignments of interest. Define an alignment clump to be the set of alignments sharing one or more pair of aligned letters with a given alignment. When the first optimal alignment is found, the matrix can be declumped by removing the effect of all alignments in the clump. Then the largest remaining score is the size of the second best alignment clump. This procedure can be continued as long as desired (Waterman & Eggert 1987). In table 1*b* the above example is delumped (outlined by lighter lines) and the second best clump and alignment highlighted. The corresponding alignment of score 3 is

CAA
CAA.

We close this section by noting that costs of $g(k) = \alpha + \beta k$ for indels of length k are commonly used and that there is a simple $O(nm)$ algorithm to compute alignments.

3. A PHASE TRANSITION

Now let the sequences $\mathbf{X} = X_1 X_2 \dots X_n$ and $\mathbf{Y} = Y_1 Y_2 \dots Y_n$ have *iid* letters. The random variable of interest is $H(\mathbf{X},\mathbf{Y})$ where we wish to estimate tail probabilities. First consider the mean or dominant

Table 1. *(a) Best local alignment; (b) declumping for the second best alignment*

(*a*)

-	C	G	T	C	C	A	A	T	A	G	C	C	A	A	T
T	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1
C	1	0	0	2	1	0	0	0	0	0	1	1	0	0	0
T	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1
G	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
A	0	0	0	0	0	1	1	0	1	0	0	0	1	1	0
C	1	0	0	1	1	0	0	0	0	0	1	1	0	0	0
A	0	0	0	0	0	2	1	0	1	0	0	0	2	1	0
A	0	0	0	0	0	1	3	2	1	0	0	0	1	3	2
A	0	0	0	0	0	1	2	2	3	2	1	0	1	2	2
G	0	1	0	0	0	0	1	1	2	4	3	2	1	1	1
G	0	1	0	0	0	0	0	0	1	3	3	2	1	0	0
C	1	0	0	1	1	0	0	0	0	2	4	4	3	2	1
A	0	0	0	0	0	2	1	0	1	1	3	3	5	4	3
A	0	0	0	0	0	1	3	2	1	0	2	2	4	6	5
C	1	0	0	1	1	0	2	2	1	0	1	3	3	5	5

(*b*)

-	C	G	T	C	C	A	A	T	A	G	C	C	A	A	T
T	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1
C	1	0	0	2	1	0	0	0	0	0	1	1	0	0	0
T	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1
G	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
A	0	0	0	0	0	1	1	0	1	0	0	0	1	1	0
C	1	0	0	1	0	0	0	0	0	0	1	1	0	0	0
A	0	0	0	0	0	0	1	0	1	0	0	0	2	1	0
A	0	0	0	0	0	1	0	0	1	0	0	0	1	3	2
A	0	0	0	0	0	1	2	1	0	0	0	0	1	2	2
G	0	1	0	0	0	0	1	1	0	0	0	0	0	1	1
G	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
C	1	0	0	1	1	0	0	0	0	0	2	1	0	0	0
A	0	0	0	0	0	2	1	0	1	0	1	1	0	1	0
A	0	0	0	0	0	1	3	2	1	0	0	0	2	1	0
C	1	0	0	1	1	0	2	2	1	0	1	1	1	1	0

term $H(\mathbf{X},\mathbf{Y})$ as $n \to \infty$. We take the simple scoring parameter $s(x,y) = +1, x = y$; $s(x,y) = -1, x \neq y$ and $g(k) = \delta k$.

The global alignment score is subadditive in distribution. Set $S_n = S(X_1 \dots X_n, Y_1 \dots Y_n)$:

$$S_{n+m} \geqslant S_n + S(X_{n+1} \dots X_{n+m}, Y_{n+1} \dots Y_{n+m}),$$

where $S(X_{n+1} \dots X_{n+m}, Y_{n+1} \dots Y_{n+m})$ equals S_m in distribution. Subadditive ergodic theory implies the existence of a constant $a(\mu,\delta)$ such that

$$\lim_{n\to\infty} \frac{S_n}{n} = a(\mu,\delta), \text{ almost surely and } L_1. \qquad (1)$$

A famous version of this problem is called the longest common subsequence problem where $\delta = 0$. Chvátal & Sankoff (1975) show the existence of the constant $a(\infty, 0)$ but even for $\mathbb{P}(X_i = 0) = 1 - \mathbb{P}(X_i = 1) \in (0,1)$ the constant remains unknown.

It is clear that $S_n \leqslant H_n$ and

$$\frac{S_n}{n} \leqslant \frac{H_n}{n} \leqslant 1.$$

Thus the asymptotics of H_n are between $na(\mu,\delta)$ and n. When $a(\mu,\delta) > 0$ it can be proved that

$$\lim_n \frac{H_n}{n} = a(\mu,\delta),$$

in probability.

When $a(\mu, \delta) < 0$ the situation is very different. Positive scores of local alignments are rare events. It can be proved that for a certain constant b, for all $\epsilon > 0$,

$$\mathbb{P}\left((1-\epsilon)b < \frac{H_n}{\log(n)} < (2+\epsilon)b\right) \to 1,$$

and it is conjectured that $\lim H_n/\log n \to 2b$. A heuristic for this result goes as follows. Let $s(x, y) = -\infty$ when $x \neq y$ and $\delta = \infty$. H_n is then the longest exactly matching region. Set $p = \mathbb{P}(X = Y)$. Neglecting end effects there are n^2 places to start an alignment of length m so the expected number is about $n^2 p^m$. Solving $1 = n^2 p^m$ yields $m = 2\log_{1/p}(n)$.

So we have learned that H_n has linear growth for $\{(\mu, \delta): a(\mu, \delta) > 0\}$ while H_n has logarithmic growth for $\{(\mu, \delta): a(\mu, \delta) < 0\}$. The graph of $a(\mu, \delta) = 0$ appears in figure 1. It is the location of a phase transition between linear and logarithmic growth of score as sequence length $n \to \infty$. These results appear in Arratia & Waterman (1994).

The linear and logarithmic growth regions have quite distinct statistical behaviour. In the next two sections we will give some theory and numerical simulations to illustrate this.

4. THE LINEAR REGION

For (μ, δ) such that $a(\mu, \delta) > 0$, we have

$$\lim_{n\to\infty} \frac{H_n}{n} = \lim_{n\to\infty} \frac{S_n}{n} = a(\mu, \delta).$$

Therefore, divided by n, H_n and S_n behave the same way. This holds because the average score per pair of letters is positive and it is always advantageous to extend to an essentially global alignment.

Now some results are given for alignments with $a(\mu, \delta) > 0$. First we give a lemma that deserves to be well known.

Lemma 1 (Azuma–Hoeffding) *Let $Z_0 = 0, Z_1, Z_2, \ldots$ be a martingale relative to $\{\mathcal{F}_n\}$ so that $Z_{n-1} = \mathbb{E}(W|\mathcal{F}_{n-1}), n > 1$. If there is a sequence of positive constants c_n such that*

$$|Z_n - Z_{n-1}| \leqslant c_n \text{ for } n \geqslant 1,$$

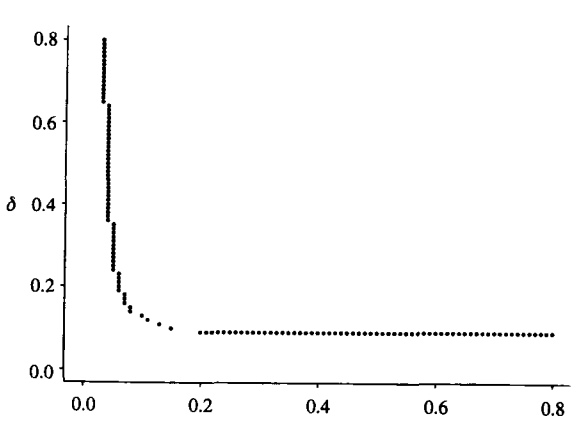

Figure 1. Phase transition boundary.

then

$$\mathbb{E}(e^{\beta Z_n}) \leqslant \exp\left[\beta^2/(2\sum_{i=1}^{n} c_k^2)\right].$$

An outline of the proof is given in Williams (1991).

To apply this to sequence alignment let $s^* = \max\{s(x, y)\}, s_* = \min\{s(x, y)\}$, and indel penalty $g(k) = \alpha + \beta k$. Then set

$$c = \max\{\min\{2s^* + 4g(1), 2s^* - 2s_*\}, 0\}.$$

We can obtain with some work (Arratia & Waterman 1994):

$$\mathbb{P}(S_n - \mathbb{E}(S_n) \geqslant \gamma n) \leqslant e^{-\gamma^2 n/(2c^2)}. \quad (2)$$

This bound will be examined with data below, but it gives exponential decay of deviations from $\mathbb{E}(S_n)$. In addition equation (2) can be extended to H_n in the linear region.

Steele (1986) has a result on the variance for non-symmetric statistics that has a similar sytle upper bound. Applied to alignment we obtain

$$Var(S_n) \leqslant n(1-p)c, \quad (3)$$

where $p = \mathbb{P}(X_1 = Y_1)$.

Equations (2) and (3) give us bounds on important quantities for alignments that are generally interesting in the linear region and for global alignments. We are interested in several questions. Does Azuma-Hoeffding provide useful bounds on the tail probabilities? Does Steele's result provide a useful bound on the variance? How do these results compare between global alignments and local alignments in the linear region?

To explore this we first look at the LCS problem ($s(x, y) = \mathbb{I}(z = y)$ and $\delta = 0$) for $\mathbb{P}(X = A) = 1 - \mathbb{P}(X = B) \in (0, 1)$. In figure 2 we give histograms of 1000 scores for $n = 250, 500, 750, 1000$ along with graphs of the corresponding estimates $\mathbb{P}(S_n - \mathbb{E}S_n \geqslant \gamma n)$ versus γ (dotted line) and the bound $e^{-\gamma^2 n/(2c^2)}$ (solid line) where $c + 2$. Clearly these bounds are not useful.

Corresponding graphs appear in figure 3 where global alignment with the Dayhoff PAM250 matrix with $g(k) = 5 + k$. Here $c = 58$. The decay of the tail probabilities is slower but Azuma–Hoeffding is not useful. Repeating this analysis for local alignment in figure 4 does not change the curves very much. These linear region local alignments are truly global alignments.

For each of these three situations (LCS, global PAM250 and local PAM250), the mean of S_n is known to grow like $a \cdot n$ and $Var(S_n) \leqslant n(1-p)c$. The means and variances are shown in figure 5. Certainly $Var(S_n)$ looks linear in n for all three cases.

5. THE LOGARITHMIC REGION

When we move to the case where

$$\lim_{n\to\infty} \frac{S_n}{n} = a(\mu, \delta) < 0,$$

the positive scoring local alignments are rare events and the statistical behaviour of local alignment scores $H_n = H(X_1 \ldots X_n, Y_1 \ldots Y_n)$ is very different from that obtained with parameters in the linear region. In recent years Poisson approximations have been given

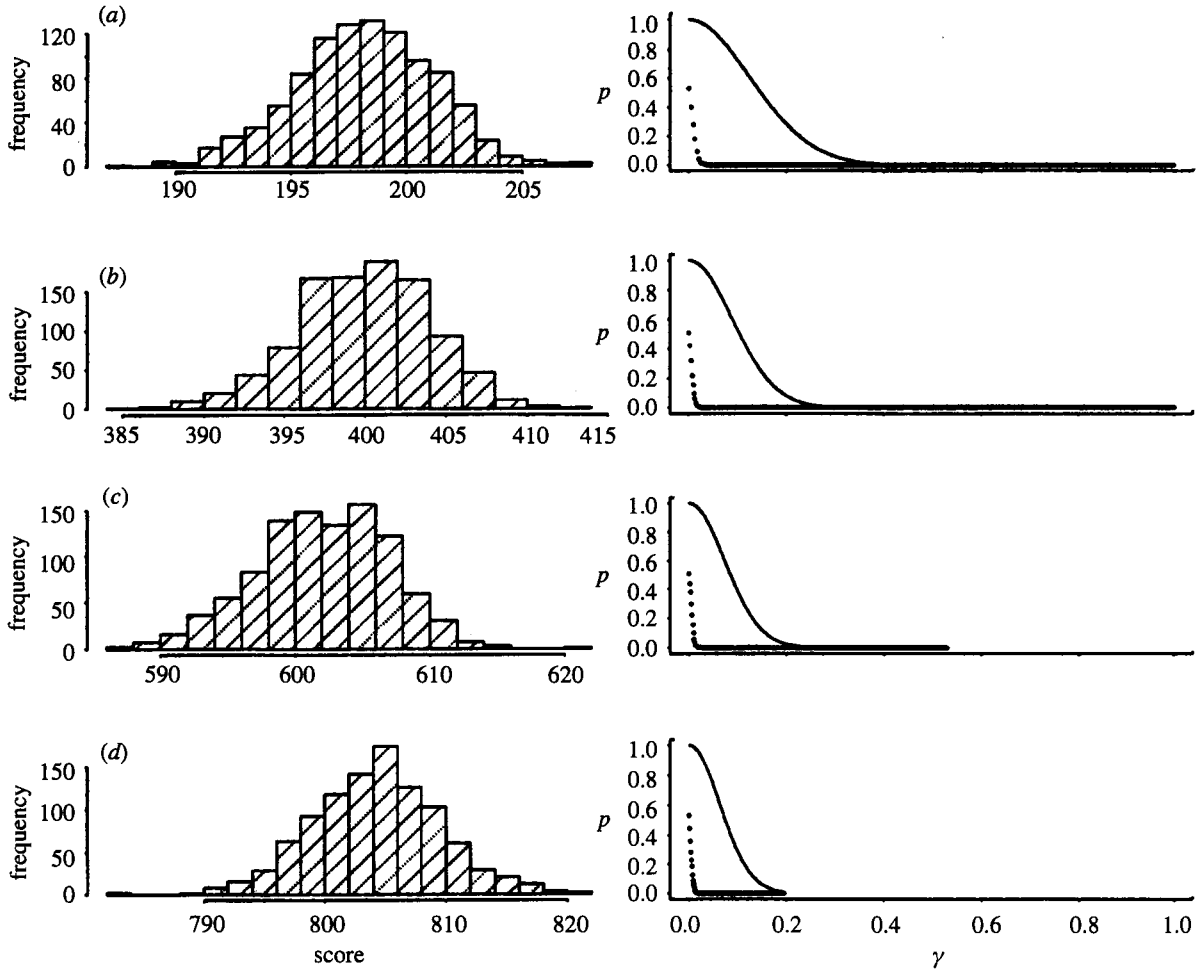

Figure 2. Histograms and tail probabilities for LCS: (*a*) $n = 250$; (*b*) $n = 500$; (*c*) $n = 750$; and (*d*) $n = 1000$.

for special cases of H_n and related statistics. For the longest alignment with up to a given fraction of mismatches (Arratia *et al.* 1990) or a given fraction of indels (Neuhauser 1994) the Chen-Stein method (Arratia *et al.* 1989) has been applied to give Poisson approximations. Our problem requires handling scoring and weighted indels.

Let $s(x,y)$ be a scoring function so that $\max_{x,y} s(s,y) > 0$ and $\mathbb{E}(s(X,Y)) < 0$, and let $g(k) = \infty$ for all $k \geqslant 1$ so there are no indels. Define $\xi \in (0,1)$ to be the largest root of $1 - \mathbb{E}(\lambda^{-s(X,Y)}) = 0$. Then

$$\lim_{n\to\infty} (H_n / \log_{1/\xi}(n^2)) = 1, \tag{4}$$

in probability. When we compare sequences of length n and m the divisor becomes $\log_{1/\xi}(nm)$, which we call the centre of the distribution of H_n. For random sequences there is a constant γ that can be determined numerically (by solving an equation) such that

$$\mathbb{P}\{H(\mathbf{X},\mathbf{Y}) > t = \log_{1/\xi}(mn) + c\} \sim 1 - \mathrm{e}^{-\gamma mn\xi^t}. \tag{5}$$

The first result equation (4) obtaining the centre $\log_{1/\xi}(nm)$ for scoring was given in Arratia *et al.* (1988). Later Karlin & Altschul (1990) extended the result to the more general scoring schemes described above and presented equation (5) which is a Poisson approximation. The idea of the Poisson approximation is that the number of clumps exceeding the centre by c, with $t = \log_{1/\xi}(mn) + c$, is Poisson with mean

$$\lambda \equiv \gamma mn\xi^t = \gamma\xi^c.$$

A Poisson with mean $\gamma mn\xi^t$ has no scores as large as t with probability $\mathrm{e}^{-\gamma mn\xi^t}$.

To put this style of Poisson approximation in context we refer to the Poisson clumping heuristic (Aldous 1989). In this model clumps are located according to a Poisson process, and them clump sizes are assigned independently to the clumps. For our sequence comparison problem, the number of (alignment) clumps with score exceeding a test value $t = \text{centre} + c$ has an approximate Poisson distribution with mean λ. The probability that at least one score exceeds t is $1 - \mathbb{P}(\text{no score exceeds } t) = 1 - \mathrm{e}^{-\lambda}$. This model has only been rigorously established for the case described in the preceding paragraph. None the less we provide numerical evidence that Poisson clumping model holds in the entire logarithmic region. Alignment clumps are marked by the end (i,j) of optimal local alignments and the score $H_{i,j}$ is the clump size.

There is an obvious way to estimate ξ and γ by using equation (5). Simulate N scores $H(\mathbf{X},\mathbf{Y})$ for sequences of length n and m. The distribution function $F_H(t)$ is $\mathrm{e}^{-\gamma nm\xi^t}$ from the Poisson assumption. Taking a log–log transformation of the empirical distribution

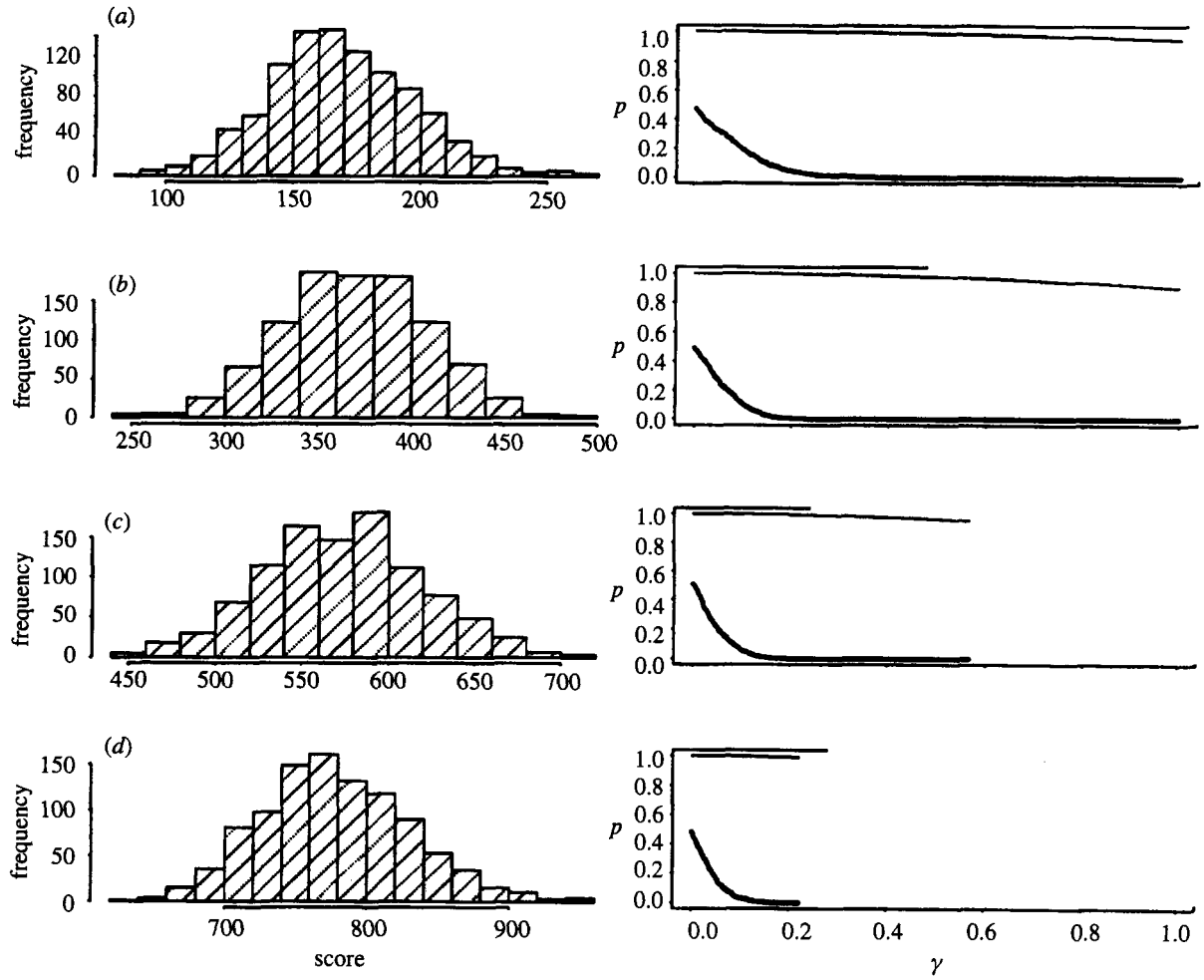

Figure 3. Histograms and tail probabilities for global alignment using PAM250; (*a*) $n = 250$; (*b*) $n = 500$; (*c*) $n = 750$; and (*d*) $n = 1000$.

function, we expect $\log(\gamma) + \log(nm) + t\log(\xi)$ and by fitting this curve γ and ξ can be estimated. As we use a straightforward sample of N comparisons, we call this method direct estimation. A drawback is the time required to do $N = 1000$ (say) comparisons with $n = m = 900$.

A more efficient method of estimation that tests the Poisson clumping model is as follows. Let $H_{(i)}$ denote the size of the i-th largest clump. Using our computational algorithm to declump we produce scores $H_{(1)} \geqslant H_{(2)} \geqslant \ldots H_{(N)}$ for the first N clumps. Scores $\geqslant t$ should, by the clumping heuristic, constitute a random sample of $H(\mathbf{X}, \mathbf{Y})$ that exceed t. Using one comparison, then, we can declump and obtain a sample to estimate ξ and γ as above. This procedure is called declumping estimation.

Now that we have described two methods to estimate ξ and γ that are accomplished by simulation. Earlier we have shown that the distribution fits independently simulated data very well even when mn is changed (Waterman & Vingron 1994). Here we will explore the fit to the Swisprot database with $N = 14\,642$ protein sequences.

The test of fit is done as follows. A query sequence of length m is compared with N database sequences. Score H_i comes from comparison of the query sequence with a database sequence of length n_i. The model is

$$\mathbb{P}(H_i \leqslant t) = \mathrm{e}^{-\gamma m n_i \xi^t}. \tag{6}$$

To obtain *iid* random variables, the random variables

$$U_i = \mathrm{e}^{-\gamma m n_i \xi^{H_i}},$$

are all uniform (0,1). While the N sequences in Swisprot are not independent, we proceed as if they are. Ordering $U_{(1)} \geqslant U_{(2)} \geqslant \ldots \geqslant U_{(N)}$, recall that $\mathbb{E}(U_{(i)}) = i/(N+1)$. Letting the i-th score in this list be H_I^*,

$$U_{(i)} = \mathrm{e}^{-\gamma m n_{(i)} \xi^{H_i^*}},$$

and

$$-\log\left(-\log\left(\frac{i}{N+1}\right)\right) + \log(mn_i) = -\log\gamma - H_i^* \log p. \tag{7}$$

A comparison of human α hemoglobin with the Swisprot database was performed. Using both direct and declumping estimates of γ and ξ, equation (7) was used in the following way to obtain figure 6*a* (direct) and figure 6*b* (declumped). We simulated 1000 pairs sequences of length 900, *iid* with letter

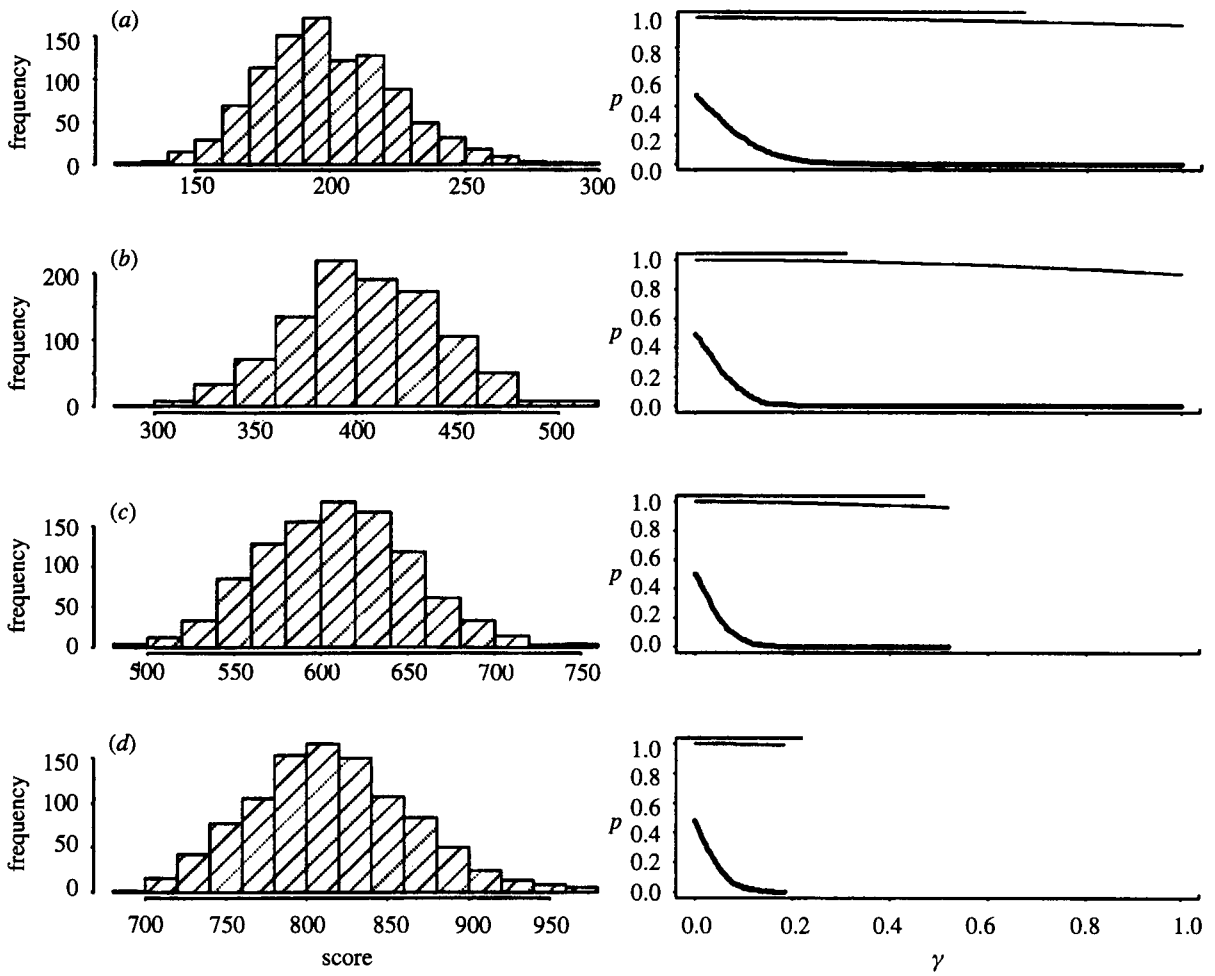

Figure 4. Histograms and tail probabilities for local alignment using PAM250; (*a*) $n = 250$; (*b*) $n = 500$; (*c*) $n = 750$; and (*d*) $n = 1000$.

frequencies of hemoglobin and of the database. If the fit were perfect the data would lie on the solid line at 45°. Interestingly both direct and declumped estimates give about the same fit, giving a good test of the Poisson clumping heuristic.

Because sequences are known to have dependent

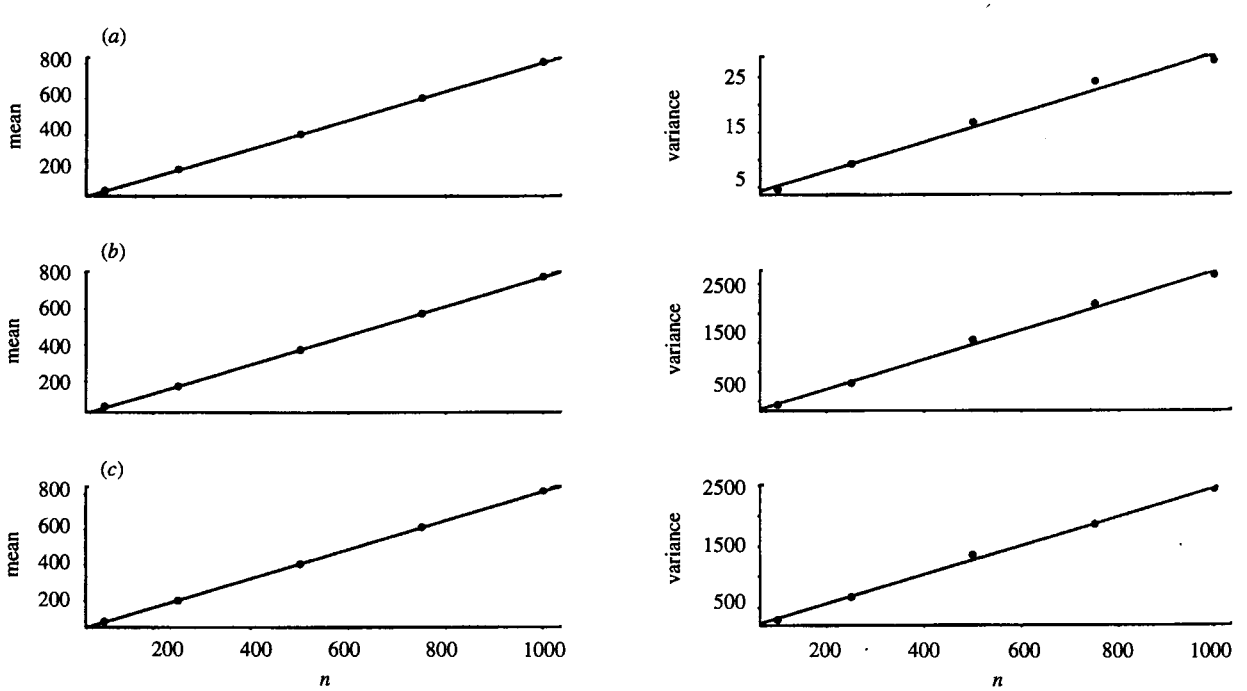

Figure 5. Means and variances as a function of n for: (*a*) LCS; (*b*) global PAM250; and (*c*) local PAM250.

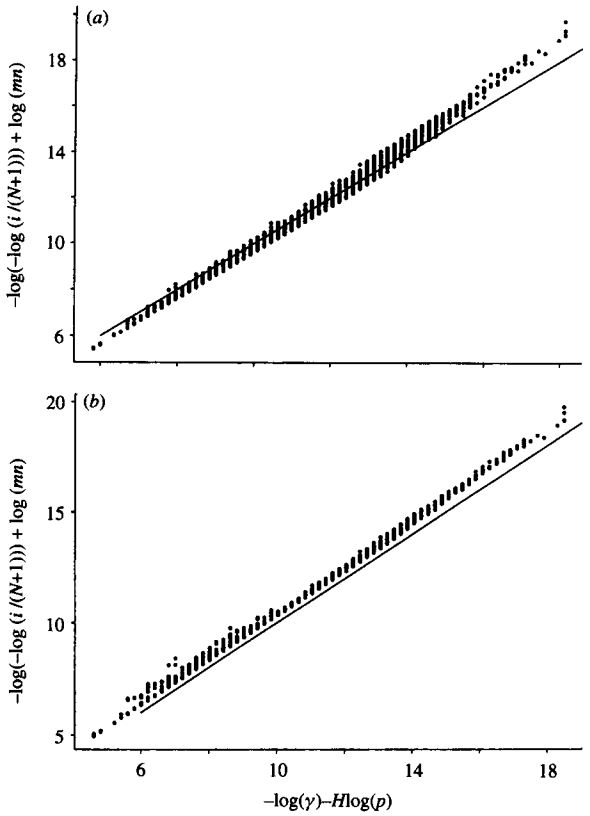

Figure 6. (*a*) Direct estimates; (*b*) declumping estimates.

letters, the lack of fit may simply be due to database sequences not having *iid* letters. In figure 7 we show the fit from simulating sequences by a Markov chain (with estimated transition probabilities from Swissprot). The small improvement is consistent with early work of Smith *et al.* (1985).

In figure 8 direct estimation is used with $m = 142$ (the length of α hemoglobin) and $n = 350$ (approximately the median database sequence length). The fit is better than that of figure 6, almost entirely due to an improved estimate of γ, which scales the 'clump volume' γmn.

Notice that we have been fitting the distribution of scores from a database comparison by estimating parameters by simulation, not by using the data to estimate parameters. We now develop a maximum

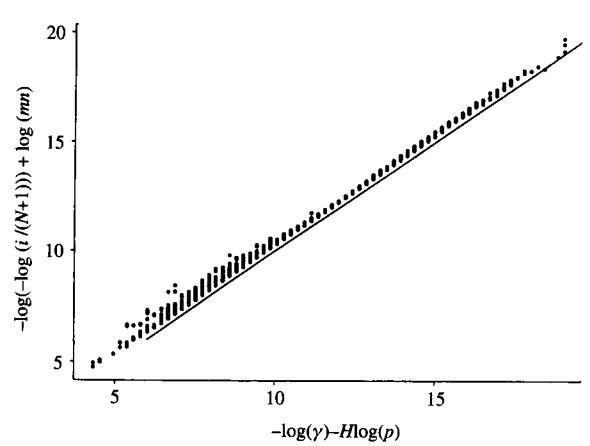

Figure 7. Estimates from Markov sequences.

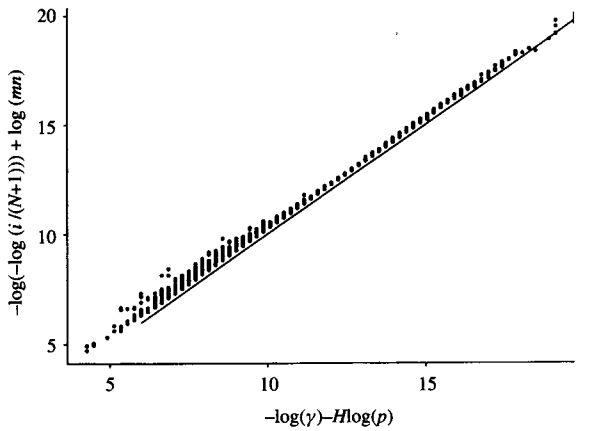

Figure 8. Direct estimates with $m = 142$ and $n = 350$.

likelihood model. Our approach is closely related to that of Mott (1992). He understood the implications of the phase transition results of Arratia & Waterman, and used the extreme value form of the distribution function, which is another face of Poisson approximation. Mott used a four-parameter model. Using equation (6), the density function for H is

$$f(t) = \gamma mn \log(1/\xi)\xi^t e^{-\gamma mn\xi^t}.$$

The likelihood of $H_1, H_2, \ldots$ is then

$$\mathbb{L} = (\gamma m \log(1/\xi))^N \left(\prod_{i=1}^{N} n_i\right) \xi^{\sum_{i=1}^{N} H_i} e^{-\gamma m \sum_{i=1}^{N} n_i \xi^{H_i}},$$

and

$$\log \mathbb{L} = \log(\gamma m)^N + N \log\log(1/\xi) + \sum_{i=1}^{N} \log n_i + \left(\sum_{i=1}^{N} H_i\right) \log \xi - \gamma m \sum_{i=1}^{N} n_i \xi^{H_i}.$$

The equations

$$\frac{\partial \log \mathbb{L}}{\partial \xi} = 0$$

and

$$\frac{\partial \log \mathbb{L}}{\partial \gamma} = 0$$

become

$$\sum_{i=1}^{N} H_i + \frac{N}{\log(\xi)} - \gamma m \sum_{i=1}^{N} n_i H_i \xi^{H_i} = 0, \tag{8}$$

and

$$\gamma m \sum_{i=1}^{N} n_i \xi^{H_i} - N = 0, \tag{9}$$

the maximum likelihood equations.

There is a quick application. In figure 6*a* ξ appears to be quite good. Using that value ξ^* we use equation (9) to re-estimate γ:

$$\gamma = \frac{N}{m \sum_{i=1}^{N} n_i (\xi^*)^{H_i}}.$$

The fit shown in figure 9 is not as good as the earlier fits. Finally we solve the maximum likelihood equations (8) and (9) to obtain the fit in figure 10, unfortunately of about the same quality as figure 9.

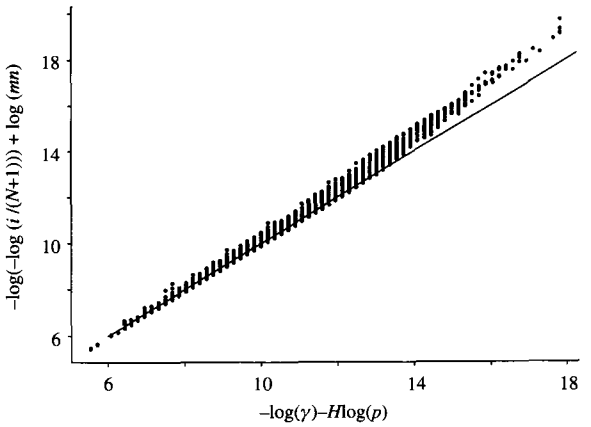

Figure 9. Maximum likelihood γ.

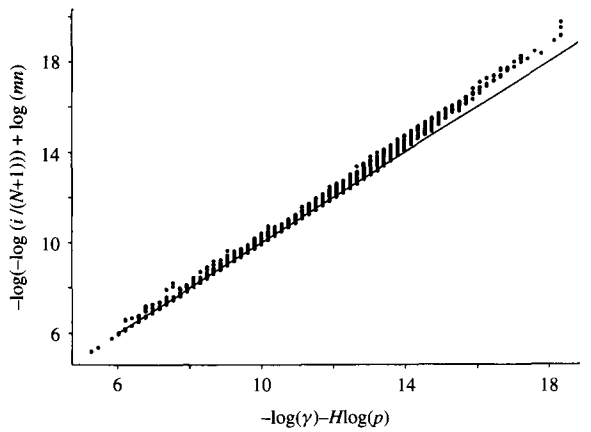

Figure 10. Maximum likelihood estimates.

In Waterman & Vingron (1993) we applied these ideas to Newat, a database assembled by R. Doolittle (1981) that removes closely related sequences. In that database the sequences are more likely to be independent. Our maximum likelihood fit of Newat scores is better than we achieve here with Swisprot. It remains to be seen whether with a 'single copy' version of Swisprot will give improved fits.

Software is available by anonymous ftp from hto-e.usc.edu. This research was supported by grants from the National Institutes of Health and the National Science Foundation.

REFERENCES

Aldous, D. 1989 *Probability approximations via the Poisson clumping heuristic*. New-York: Springer-Verlag.

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

Arratia, R., Gordon, L. & Waterman, M.S. 1990 The Erdös-Rényi Law in distribution, for coin tossing and sequence matching. *Ann. Statist.* **18**, 539–570.

Arratia, R., Morris, P. & Waterman, M.S. 1988 Stochastic Scrabble: a law of large numbers for sequence matching with scores. *J. appl. Prob.* **25**, 106–119.

Arratia, R. & Waterman, M.S. 1994 A phase transition for the score in matching random sequences allowing deletions. *Ann. appl. Prob.* **4**, 200–225.

Barker, W.C. & Dayhoff, M.O. 1982 Viral *src* gene product are related to the catalytic chain of mammalian cAMP-dependent protein kinase. *Proc. natn. Acad. Sci. U.S.A.* **79**, 2836.

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

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

Doolittle, R.F. *et al.* 1983 *Simian sarcome* virus *onc gene, v-sis*, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science, Wash.* **221**, 275.

Karlin, S. & Altschul, S.F. 1990 Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. natn. Acad. Sci. U.S.A.* **87**, 2264–2268.

Mott, R. 1992 Maximum-likelihood estimation of the statistical distribution of Smith-Waterman local sequence similarity scores. *Bull. Math. Biol.* **54**, 59–76.

Neuhauser, C. 1994 A Poisson approximation for sequence comparisons with insertions and deletions. *Ann. Stat.* (In the press.)

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

Smith, T.F., Burks, C. & Waterman, M.S. 1985 The statistical distribution of nucleic acid similarities. *Nuclear Acid Res.* **13**, 645–656.

Steele, J.M. 1986 An Efron-Stein inequality for nonsymmetric statistics. *Ann. Stat.* **14**, 753–758.

Waterman, M.S. & Eggert, M. 1987 A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *J. molec. Biol.* **197**, 723–728.

Waterman, M.S. & Vingron, M. 1994 Sequence comparison significance and Poisson approximation. *Stat. Sci.* (In the press.)

Williams, D. 1991 *Probability with martingales*. Cambridge University Press.

Discussion

R. A. ELTON (*Medical Statistics Unit, University of Edinburgh, U.K.*). Are there examples where comparisons of biologically unrelated sequences using Professor Waterman's method give quite high probabilities? This would provide reassuring negative evidence for the validity of the theory. Also, how does the fact that real DNA is not composed of random sequences of equally common bases affect the distribution of his statistics?

M. WATERMAN. The primary issue is to be able to detect weak alignments that are biologically significant from alignments that score high simply due to the large number of sequences and alignments. There are at least two difficulties with running human α hemoglobin versus the Swisprot database. The first comes from the non-independence of the sequences in the database. As mentioned in the paper, another test in Vingron & Waterman (1994) applies the theory to Newat, a 'single copy' database. The fit is very good there with the right number of non-homologous sequences with high scores. The second problem comes from the fact that real protein sequences do not have *iid* letters. As noted in Smith *et al.* (1985) there is a lack of sensitivity of score distribution on the dependencies in biological sequences. In any case dependencies that can be modeled can easily be included in our simulations. We have not found this to be necessary in practice.