Lectures on Mathematics in the Life Sciences
Volume 17, 1986

Probability Distributions for DNA Sequence Comparisons

Michael S. Waterman

Dedicated to the memory of Stanislaw Ulam

ABSTRACT. Recently DNA sequence comparisons have focused on finding long matching segments between two sequences, rather than matching the entire sequences. Generalizations of the celebrated Erdos-Renyi law give laws of large numbers and extreme value distributions for random variables equal to the length of the longest exact match and longest approximate match between the sequences. The cases of independent, identically distributed sequences and of Markov chains are presented. In the final section, simulated sequences and sequences from bacteriophage lambda are analyzed in light of these theoretical results.

1. INTRODUCTION AND DNA SEQUENCE COMPARISONS

As the nucleic acid sequence data accumulate at an increasing rate, comparison of the sequences becomes increasingly central. In the late 1960's, macromolecular sequence comparison meant calculating a matrix of pairwise distances between sequences of a protein, such as cytochrome - C, taken from a number of organisms. Then, with a variety of algorithms, attempts were made to reconstruct the evolutionary history of the protein. The paper by Fitch and Margoliash (1967) is an excellent early example of such a study. In general, such sequence comparisons involve finding a minimal evolutionary pathway by which one of the molecules can be transformed into another by events like mutations, deletions, and insertions.

Recently, Sanger (1977) and Maxam and Gilbert (1977) independently discovered methods for rapid sequencing of DNA. Sanger and Gilbert were awarded Nobel prizes in 1980 for this work. Their procedures created a revolution in molecular biology. Among the revelations from the data is a new, dynamic picture of the genome with elements called transposons which relocate themselves and with viruses which move genetic material within and between genomes. For these and other reasons, our picture of the genome emerges as a mosaic of variously sized blocks of DNA sequences. Unexpected relationships have been discovered between viral DNA and host DNA (e.g., see Doolittle et al., 1983, Naharro et al., 1984 and Weiss, 1983). These discoveries have implications for the study of cancer. In these situations, instead of entire sequences with high similarity, it is contiguous subsequences or segments with high similarity that are found.

Sequence comparison problems have motivated algorithms and associated mathematics (see Waterman, 1984, for a review). In this article, I will survey some recent results concerning probability distributions appropriate for the "highly similar segments" problems of sequence comparison. However, before discussion of the probabilistic aspects of the problem, I present algorithms which may be used to analyze Monte Carlo and biological sequence data, as in Section 4 of this paper.

Suppose the two DNA sequences being compared are $\mathcal{X} = x_1x_2\dots x_n$ and $\mathcal{Y} = y_1y_1\dots y_m$ where x_i and $y_i \in \{A,T,G,C\}$. The random variable of interest here is

$$\begin{aligned} M_n(k) = \max\{m: \; & x_{i+\ell} = y_{i+\ell} \text{ for } \ell = 1 \text{ to } m \text{ except for} \\ & \text{at most } k \text{ failures, for some } 0 \le i \le n-m\}. \end{aligned}$$

Thus $M_n(k)$ is the longest match interrupted by at most k mismatches. By moving along diagonals of constant $j-i$ and storing the positions of the (at most) k mismatches, it is easy to see that $M_n(k)$ can be computed in time nm and storage k. It is possible, for example, to find $M_n(0)$, the longest exact match, in time $(n+m)\log(n+m)$ but the concern here is not with finding the most efficient algorithms. It is more difficult to improve the algorithm given here for $M_n(k)$, $k \neq 0$. For example, with

$$\begin{aligned} \mathcal{X} &= \text{A G T C T G A A G C A C A A C T G T}, \\ \mathcal{Y} &= \text{T A T C T T T G A A G C C C A T T T} \end{aligned}$$

$M_{18}(0) = 6$ for the match of TGAAGC beginning at x_5 and y_7. Also $M_{18}(2) = 11$ for these sequences.

As mentioned above, DNA sequences evolve by deletion and insertion as well as mutation of letters. That is, spaces can be placed in either sequence. A random variable of interest here is

$$H = \max_{\substack{I \subset \mathcal{X} \\ J \subset \mathcal{Y}}} \{\#\text{ matches} - \mu\#\text{ mismatches} - \lambda\#\text{ insertion/deletions}\},$$

where $I \subset \mathcal{X}$, $J \subset \mathcal{Y}$ ranges over all contiguous segments in the indicated sequences. Smith and Waterman (1981a,b) solved this problem with an algorithm which uses time proportional to nm. Let $H_{i,j}$ be the maximum score (# matches − μ# mismatches − λ# insertion/deletions) of two segments that end in x_i and y_j, or zero, whichever is larger. It is shown that, if $H_{i,0} = H_{0,j} = 0$ for $0 < i < n$ and $1 < j < m$, then

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

where

$$s(x,y) = \begin{cases} 1 & \text{if } x = y \\ -\mu & \text{if } x \neq y \end{cases} .$$

Of course,

$$H = \max\{H_{i,j} : 1 < i < n,\ 1 < j < m\}.$$

Figure 1 is an example of $(H_{i,j})$ for a specific problem. The value of H is 3.10 corresponding to the aligned segments

```
C T G C G
C T G G G .
```

		G	T	C	C	G	C	T	G	C	G
	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
T	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
C	0.0	0.0	0.0	2.0	1.0	0.0	1.0	0.0	0.1	1.0	0.0
T	0.0	0.0	1.0	0.0	1.1	0.1	0.0	2.0	0.0	0.0	0.1
G	0.0	1.0	0.0	0.1	0.0	2.1	0.1	0.0	3.0	1.0	1.0
G	0.0	1.0	0.1	0.0	0.0	1.0	1.2	0.0	1.0	2.1	2.0
G	0.0	1.0	0.1	0.0	0.0	1.0	0.1	0.3	1.0	0.1	3.1
A	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.1	1.1
A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0

Figure 1. Similarity matrix where matches receive weight 1, mismatches receive weight -.9, and deletion/insertions receive weight -2.0.

Earlier work on probability distributions for sequence alignments is motivated by best matching of entire sequences. Chvatal and Sankoff (1975) studied the length L_n of the longest subsequence common to two random sequences with a finite alphabet and equally likely letters. Deken (1979) noted that there exists a c such that

$$P(\lim_{n\to\infty} \frac{L_n}{n} = c) = 1.$$

Steele (1982) showed $\mathrm{Var}(L_n) = O(n)$. In this paper, I consider probability distributions for the "highly similar segments" problem.

In the following sections, I survey results principally obtained in collaboration with my colleagues Richard Arratia and Louis Gordon at the University of Southern California. Section 2 presents laws of large numbers for a large class of random variables including $M_n(k)$, $k > 0$. Section 3 contains much finer distributional results for $M_n(k)$, $k > 0$, related to the extreme value distribution. Finally, in Section 4, Monte Carlo and biological sequences are examined in light of the probabilistic results. Even the complex random variable H displays evidence of an extreme value distribution.

2. LAWS OF LARGE NUMBERS

In this section, I present laws of large numbers for the asymptotic behavior of the longest match between two random DNA sequences. Random here means either independent and identically distributed or Markov, although similar laws for m-dependent processes can be obtained. While the laws of large numbers only give order of magnitude results, these estimates are surprisingly good. The extreme value results of Section 3 give much more precise results and allow comparison with the more easily obtained results of this section.

The first problem considered here will be the length R_n of the longest match between two independent identically distributed (i.i.d.) DNA sequences of length n which are in a fixed alignment. Let

$$p = P(\text{Match}) = p_A^2 + p_T^2 + p_G^2 + p_C^2 ,$$

where p_i is the probability of base i. The random variable R_n is only of interest in the cases $p \in (0,1)$. The problem posed here can be restated as the longest run of heads in n coin tosses where $p = P(\text{Heads})$. Erdös and Rényi (1970) presented results which contain the following theorem. I give an outline of a proof which prepares the way for generalization.

Theorem 2.1 Let $X_1, X_2, \ldots, Y_1, Y_2 \ldots$ be independent and identically distributed and let $0 < p \equiv P(X_1 = Y_1) < 1$. Define $R_n = \max\{m: X_{i+k} = Y_{i+k}$ for $k = 1$ to m, for some $0 < i < n-m\}$. Then

$$P\left(\lim_{n\to\infty} \frac{R_n}{\log_{1/p}(n)} = 1\right) = 1.$$

Proof. Let $A_i = \{X_{i+k} = Y_{i+k}$ for $k = 1$ to $m\}$ where $0 < i < n-m$. Now, for $\varepsilon > 0$, let $m = [(1+\varepsilon)\log_{1/p}(n)]$. Since $P(A_i) = p^m$, we have

$$p\, n^{-(1+\varepsilon)} < P(A_i) = p^m < n^{-(1+\varepsilon)}.$$

For $n = n_k = [(1/p)^k]$,

$$P(\ _{k}\ _{i}\ A_i(n_k)) < \sum_{k=1}^{\infty} \sum_{i=0}^{n_k - m} P(A_i(n_k))$$

$$< \sum_{k=1}^{\infty} \frac{p^{\varepsilon k}}{(1-p^k)^{\varepsilon}} < \infty.$$

Therefore, by the Borel-Cantelli lemma (Chung, 1968),

$P(A_i(n_k)$ occurs infinitely often$) = 0$.

Since longest match length increases with n, it follows that

$$P(\overline{\lim_{n}}\ R_n/\log_{1/p}(n) < 1) = 1.$$

This establishes half of the result.

To obtain a corresponding lower bound, recall that the Borel-Cantelli lemma has a converse if the events are independent. To create non-overlapping head runs, let

$$B_i \equiv A_{mi} = \{X_{mi+k} = Y_{mi+k} \quad \text{for} \quad k = 1,\dots,m\},$$

and

$$S = \sum_i I(B_i).$$

This time, with $m = [(1 - \varepsilon) \log_{1/p}(n)]$,

$$E(S) > ([\Sigma \frac{n}{m}] - 1)p^m \approx \frac{n}{m} pn^{-(1-\varepsilon)} = \frac{p}{m} n^{\varepsilon}.$$

Therefore

$$\lim_n E(S) = \infty$$

and

$$P(\lim_n S > 0) = 1.$$

It follows that

$$P(\underline{\lim}_n R_n/\log_{1/p}(n) > 1) = 1.$$

Next I take up a problem of more direct interest to molecular biology, the length $M_n(0)$ of the longest match between two sequences when shifts are allowed. Allowing shifts gives n^2 choices for (i,j), the starting position of a match run. Above there were only n starting positions. This naive approach might suggest that $M_n(0)$ grows like $\log_{1/p}(n^2) = 2 \log_{1/p}(n)$. This turns out to be correct and is formalized in the next theorem, due to Arratia and Waterman (1984a).

Theorem 2.2. Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be independent and identically distributed and let $0 < p \equiv P(X_1 = Y_1) < 1$. Define

$$M_n(0) = \max\{m: X_{i+k} = Y_{j+k} \text{ for } k = 1 \text{ to } m, \text{ for some } 0 \leq i, j \leq n-m\}.$$

Then

$$P(\lim_{n\to\infty} \frac{M_n(0)}{\log_{1/p}(n)} = 2) = 1.$$

Proof. The upper bound is established as in Theorem 2.1, but the lower bound is more difficult. Define $A_{ij} = \{X_{i+k} = Y_{j+k}$ for $k = 1$ to $m\}$ and let $m = [(2-\varepsilon) \log_{1/p}(n)]$. Let $B_{ij} = A_{mi,mj}$ and $S = \sum_{i,j} I(B_{ij})$. As in Theorem 2.1, $E(S) \approx m^{-2} n^{\varepsilon} \to \infty$. To handle the dependence introduced by shifting, it is sufficient to show

$$\mathrm{Var}(S)/(E(S))^2 \to 0.$$

See, e.g., Problem 10, Section 4.3 of Chung (1968). Using the above formula for S,

$$\begin{aligned} \mathrm{Var}(S) &= E(\sum_{i,j}(I(B_{ij}) - P(B_{ij}))^2 \\ &\leq \sum P(B_{ij} \cap B_{k\ell}) \\ &= \sum_{\substack{i=k\\ j=\ell}} + \sum_{\substack{i\neq k\\ j\neq \ell}} + 2\sum_{\substack{i=k\\ j\neq \ell}} P(B_{ij} \cap B_{k\ell}). \end{aligned}$$

In the last line the first sum is less than $E(S)$ while the second is 0. The third sum has approximately $2(n/m)^3$ terms. Let $p = \sum_a (P(X = a))^2 = \sum_a q_a^2$, i.e., q_a is the probability distribution on the atoms of X. Now, using a version of Holder's inequality (Hardy, Littlewood, and Polya (1934), formula 2.10.3) for $j \neq k$,

$$P(B_{ij} \cap B_{ik}) = [\sum_a (q_a)^3]^m < [\sum_a q_a^2]^{3m/2} = (P(B_{ij}))^{3/2}.$$

Combining these estimates yields,

$$\mathrm{Var}(S) < E(S) + 2(E(S))^{3/2},$$

and

$$\frac{\mathrm{Var}(S)}{(E(S))^2} < \frac{m^{-3} n^{3\varepsilon/2}}{(m^{-2} n^{\varepsilon})^2} = mn^{-\varepsilon/2} \to 0.$$

This completes the proof of the theorem.

Notice that the effects introduced by shifting make the theorem more difficult to prove. It is possible to obtain results for Markov chains as well (Arratia and Waterman, 1985a).

Theorem 2.3. Let $X_1, X_2, \ldots,$ and $Y_1, Y_2, \ldots$ be two independent Markov chains on a finite alphabet S which are irreducible and aperiodic and have transition probabilities (p_{ij}) $i,j \in S$. Let $p \in (0,1)$ be the largest eigenvalue of the substochastic matrix $((p_{ij})^2)$ $i,j \in S$. Then

$$P(\lim_n \frac{M_n(0)}{\log_{1/p}(n)} = 2) = 1.$$

Our interest in establishing Theorem 2.3 is the fact that (first-order) nearest neighbor effects in DNA sequences are statistically significant (Smith et al., 1983). Another feature of DNA evolution that concerns us is substitutions, insertions, and deletions of letters as well as inversions. We ask how many letters can be "removed" from the sequences to lengthen the match and still retain the $2 \log_{1/p}(n)$ behavior. The result is surprisingly strong.

Theorem 2.4. Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ and p be as in Theorem 2.2 or 2.3. Let $M_n^*(k)$ be the longest match between $X_1 \ldots X_n$ and $Y_1 \ldots Y_n$ allowing shifts and removal of k single letters, i.e.,

$$\begin{aligned} M_n^*(k) \equiv \max\{m:\ & X_{i(a)} = Y_{j(a)} \text{ for } a = 1 \text{ to } m, \text{ for some integers} \\ & 1 \le i(1) < \ldots < i(m) \le n \text{ and } 1 \le j(1) < \ldots < j(m) \le n \\ & \text{where } i(m) - i(1) < m+k \text{ and } j(m) - j(1) < m+k\}. \end{aligned}$$

Then for any constant k or any deterministic sequence $k = k(n)$ where

$$k = o(\log(n)/\log \log(n)),$$

it follows that

$$\lim_n M_n^*(k)/\log_{1/p}(n) = 2$$

in probability.

Another feature of DNA sequences Arratia and I (1985b) considered was that all sequences do not have the same distribution. We obtained results for Markov sequences as well as i.i.d. sequences but only the i.i.d. results are discussed here. The surprising discovery is that M_n can still have $2 \log_{1/p}(n)$ behavior even when the marginal distributions are quite different.

Theorem 2.5. Let $X_1, X_2, \ldots$ be distributed as μ, $Y_1, Y_2, \ldots$ be distributed as ν with all letters independent and $p = P(X_1 = Y_1) \in (0,1)$. Then there is a constant $C(\mu,\nu) \in [1,2]$ such that

$$P(\lim_{n\to\infty} M_n/\log_{1/p}(n) = C(\mu,\nu)) = 1.$$

In addition

$$C(\mu,\nu) = \sup_{\gamma \in Pr(S)} \min\left\{\frac{\log(1/p)}{H(\gamma,\mu)}, \frac{\log(1/p)}{H(\gamma,\nu)}, \frac{2\log(1/p)}{\log(1/p) + H(\gamma,\alpha)}\right\}$$

where $\alpha_a \equiv \mu_a \nu_a / p$, $H(\alpha,\nu) = \sum_a \alpha_a \log(\alpha_a/\nu_a)$, and γ ranges over the probability distributions on the state space S. Here log can be to any base. Also $C(\mu,\nu) = 2$ if and only if

$$\max\{H(\alpha,\nu), H(\alpha,\mu)\} < (\tfrac{1}{2}) \log(1/p).$$

The set of μ such that $C(\mu,\nu) = 2$ for a fixed ν has positive diameter. Of course, $C(\nu,\nu) = 2$ by Theorem 2.2. That a large set of μ satisfies $C(\mu,\nu) = 2$ is another indication of the strength of the "2 log(n)" law.

3. EXTREME VALUE DISTRIBUTIONS

Next I give some results for the "exact" distribution of R_n and $M_n(k)$. For coin tossing the first results were given in a paper by Erdös and Révész (1975) and improved by Guibas and Odlyzko (1980). More recently Gordon, Schilling, and I (1984) gave a probabilistic analysis which is easily motivated and extends the earlier results.

I begin by again considering R_n, the length of the longest head run where $p = P(\text{Heads})$. Each head run is preceded by a tail and has length m with probability qp^m, $q = 1-p$. There are approximately nq tails in n trials so that

$$R_n \approx \max_{1<i<nq} Z_i$$

where Z_i is geometric and $P(Z_i = m) = qp^m$. Also $Z_i = [W_i]$ where W_i are i.i.d. exponential random variables with mean $1/\lambda$, $\lambda = \ln(1/p)$, and [ ] is the usual greatest integer function. Therefore it follows that

$$R_n \approx [\max_{1<i<nq} W_i].$$

The maximum of i.i.d. exponential random variables is an extreme value random variable. Letting V denote a random variable such that $P(V < t) = \exp(-e^{-t})$ (that is, the standard extreme value distribution), R_n then should satisfy

$$R_n \approx [\ln(nq)/\lambda + V/\lambda].$$

Hence

$$\begin{aligned} E(R_n) &\approx \ln(nq)/\lambda + E(V)/\lambda - 1/2 \\ &= \ln(nq)/\lambda + \gamma/\lambda - 1/2 , \\ &= \log_{1/p}(n) + \log_{1/p}(q) + \gamma/\lambda - 1/2 , \end{aligned}$$

where $E(V) = \gamma$ is the Euler-Mascheroni constant and the $1/2$ is Sheppard's continuity correction. For $p = q = 1/2$, the approximation is $(\ln(n) + \gamma)/\lambda - 3/2$, exactly the leading terms found by Guibas and Odlyzko (1980). Applying the same approach to the variance, $\mathrm{Var}(R_n) \approx \pi^2/6\lambda^2 + 1/12$. Here the $1/12$ is Sheppard's correction for variance and $\mathrm{Var}(V) = \pi^2/6$.

The next theorem appears in Gordon, Schilling, and Waterman (1984).

Theorem 3.1. Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be independent and identically distributed and let $0 < p = P(X_1 = Y_1) < 1$. Let

$$R_n(k) = \max\{m\colon X_{i+\ell} = Y_{i+\ell} \text{ for } \ell = 1 \text{ to } m \text{ except for at most k failures, for some } 0 \le i \le n-m\}.$$

Then for $\lambda = \ln(1/p)$,

$$\begin{aligned} E(R_n(k)) = {} & \log_{1/p}(n) + k \log_{1/p}\log_{1/p}(n) \\ & + (k+1)\log_{1/p}(q) - \log_{1/p}(k!) + k \\ & + \gamma/\lambda - 1/2 + r_1(n) + o(1), \end{aligned}$$

and

$$\mathrm{Var}(R_n(k)) = \pi^2/6\lambda^2 + 1/12 + r_2(n) + o(1),$$

where, for $\theta = \pi^2/\lambda$,

$$|r_1(n)| < (2\pi)^{-1}\theta^{1/2} e^{-\theta}(1-e^{-\theta})^{-2}$$

and

$$|r_2(n)| < (1.1 + .7\theta)(2\theta^{1/2} e^{-\theta}(1-e^{-\theta})^{-3}).$$

Notice that the bounds are about equal to 1.6×10^{-6} (or 3.45×10^{-4}) for the mean and 6×10^{-5} (or 2.64×10^{-2}) for the variance when $p = 1/2$ (or $1/4$). The striking feature of the variance being approximately constant with n is derived from the extreme value distribution.

The next question of interest to DNA sequence analysts is whether these results carry over to matching with shifts. This is answered in the affirmative by the next theorem which is proved in Arratia, Gordon, and Waterman (1984). Karlin et al. (1984) announced a similar result for $M_n(0)$, the longest match with no mismatch. Their result as stated gives no error estimates and differs from the one given here for $k = 0$ by minor constants.

Theorem 3.2. Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be independent and identically distributed and let $0 < p = P(X_1 = Y_1) < 1$. Let

$$M_n(k) = \max\{m\colon X_{i+\ell} = Y_{j+\ell} \text{ for } \ell = 1 \text{ to } m \text{ fails at most } k \text{ times, for some } 0 \le i,\ j \le n-m\}.$$

Then

$$E(M_n(k)) = \log_{1/p}(n^2) + k \log_{1/p}\log_{1/p}(n^2) + (k+1)\log_{1/p}(q) - \log_{1/p}(k!) + k + \gamma/\lambda - {}^1\!/_2 + r_1(n) + o(1),$$

and

$$\mathrm{Var}(M_n(k)) = \pi^2/6\lambda^2 + 1/12 + r_2(n) + o(1).$$

The functions $r_1(n)$ and $r_2(n)$ are bounded by the corresponding functions of θ in the statement of Theorem 3.1. DNA sequences do not always have equal lengths, and in Arratia, Gordon, and Waterman (1984) a more general theorem appears with n^2 replaced by n_1n_2, the product of the lengths of the sequences. The necessary condition is $\log(n_1)/\log(n_2) \to 1$.

It is possible to present these results in the case of Markov chains or even m-dependence. However, the analysis of DNA sequences seems only to require the i.i.d. case. The next section examines data supporting this observation.

4. DATA ANALYSIS.

In this final section, I test the previous theory by applying the algorithms of Section 1 to Monte Carlo and DNA sequences. The DNA sequences come from a bacteriophage named lambda. (Bacteriophages are viruses that infect bacteria). The complete sequence of lambda is now known and has 48,502 base pairs (Sanger et al., 1982). Beginning at base 1 of lambda, I chose $2^7, 2^8, \ldots, 2^{12}$ bases in a nonoverlapping manner. Then I repeated this process until the remaining sequence was less than $2^7 + 2^8 + \ldots + 2^{12}$ bases. Therefore I have six sequences 2^7 long, six sequences 2^8 long,..., and six sequences 2^{12} long. The value of p for lambda is essentially 1/4. Similar Monte Carlo data was generated with all bases equally likely so that $p = 1/4$.

Assuming the simplest $\log(n^2)$ law, $y = a \log(n^2) + b = ax + b$, the scores $M_n(0), M_n(1), \ldots, M_n(6)$, and H were plotted vs. $\log_{1/p}(n^2)$ where all pairs of sequences of equal length are compared. The results are shown in Figures 2 through 9, where (a) is Monte Carlo and (b) is lambda. In $M_n(1)$ through $M_n(6)$ the slope or coefficient of $\log(n^2)$ is not 1. This is accounted for by the $k \log_{1/p}\log_{1/p}(n^2)$ term of Theorem 3.2. In the range of 2^7 to 2^{12}, $k \log_4\log_4(n^2) \approx (.1)k \log_4(n^2) + c$. When this approximately linear effect of the loglog term is accounted for, the coefficient of the $\log_{1/p}(n^2)$ term is approximately 1 as predicted. The constant terms, b, are not equal to those given by the formula for $E(M_n(k))$ in Theorem 3.2 and require additional terms in the expansion. See Arratia et al. (1984) for details. The DNA scores have a slightly larger variance but the agreement is good.

Smith, Waterman, and Burks (1985) calculate H for over 40,000 pairs of vertebrate sequences in an empirical study. That work was the motivation for the theoretical results reported in this paper. The fit to the data there is

$$H = 2.55 \log_{1/p}(n_1 n_2) - 8.99$$

where n_1 and n_2 are the length of sequences being compared. The fits in Figure 9 are reasonably close to this one. What this all suggests is that the distribution of maximum segments scores is not very dependent on the biological details of the sequence genomes. Alignments with similarity scores differing significantly from the expected scores should be carefully examined by the sequence analyst.

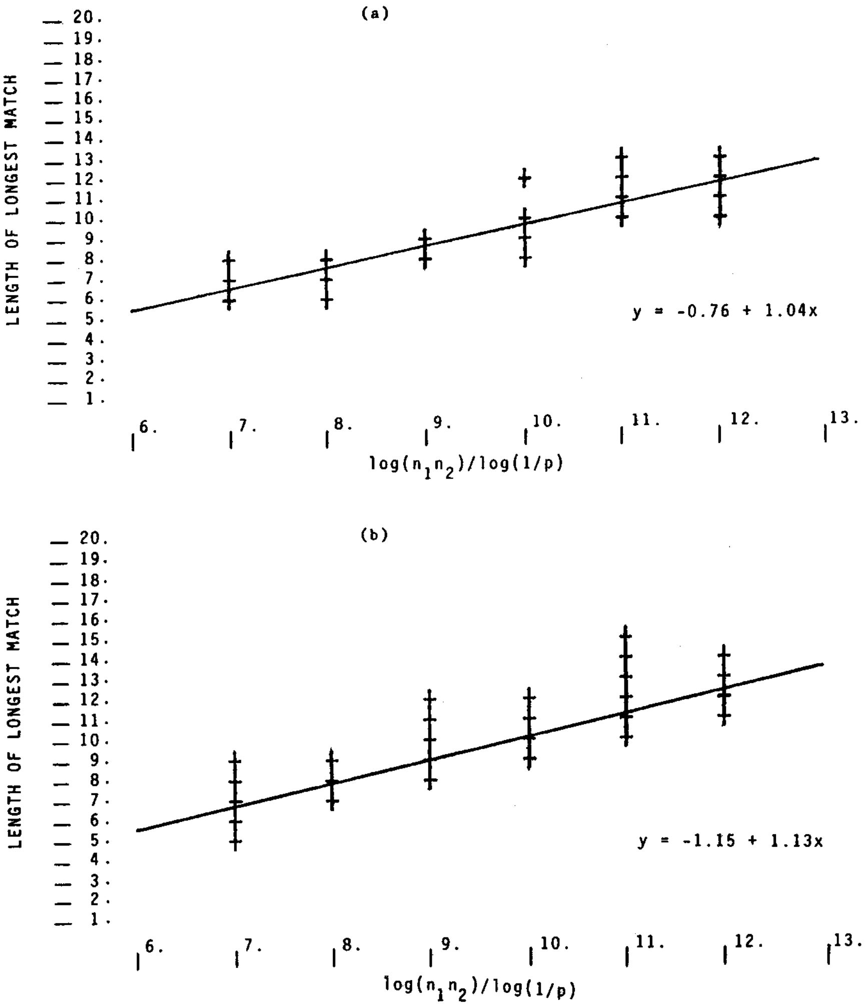

Fig. 2. k = 0.

Figures 2-8. Length of longest matches with k = 0,1,...,6 mismatches. For each value of $x = \log_{1/p}(n_1n_2) = \log_4(n_1n_2)$ there are 15 sequence comparisons. (a) Monte Carlo sequences and (b) Lambda sequences. The best linear fits of y = longest match length with k mismatches are given by y = a + bx.

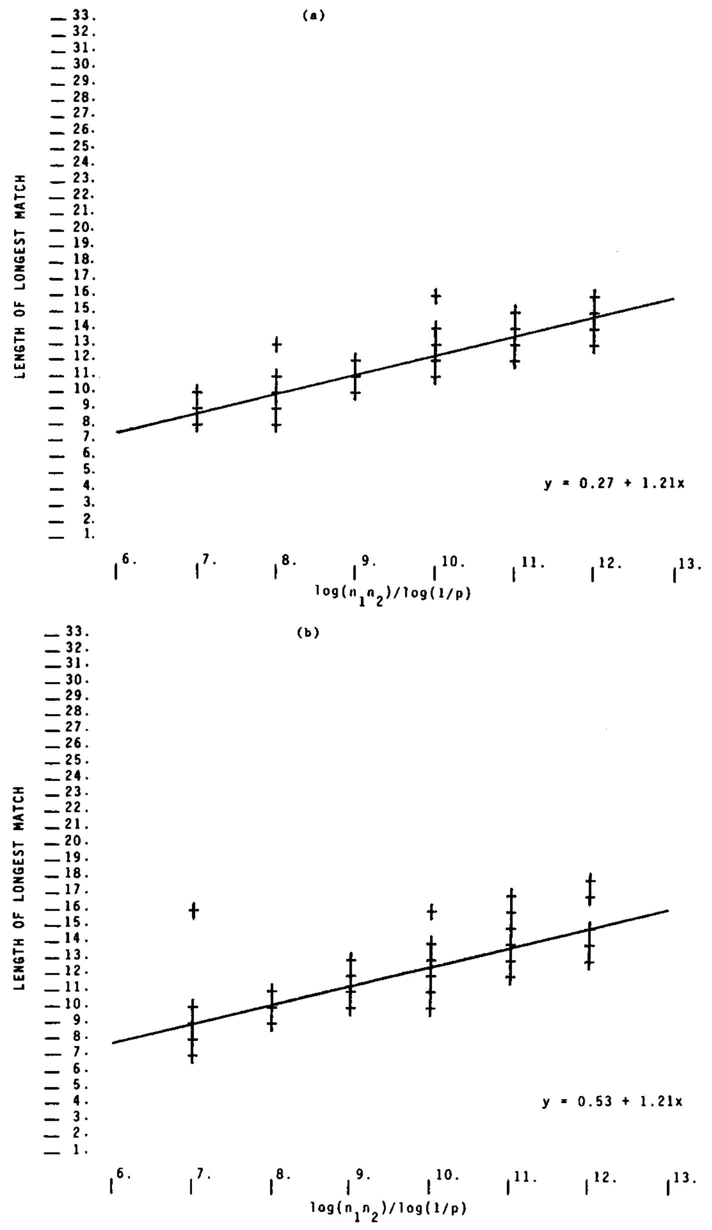
(a)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 0.27 + 1.21x
6.
7.
8.
9.
10.
11.
12.
13.
log(n_1n_2)/log(1/p)
(b)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 0.53 + 1.21x
6.
7.
8.
9.
10.
11.
12.
13.
log(n_1n_2)/log(1/p)

Figure 3. k = 1.

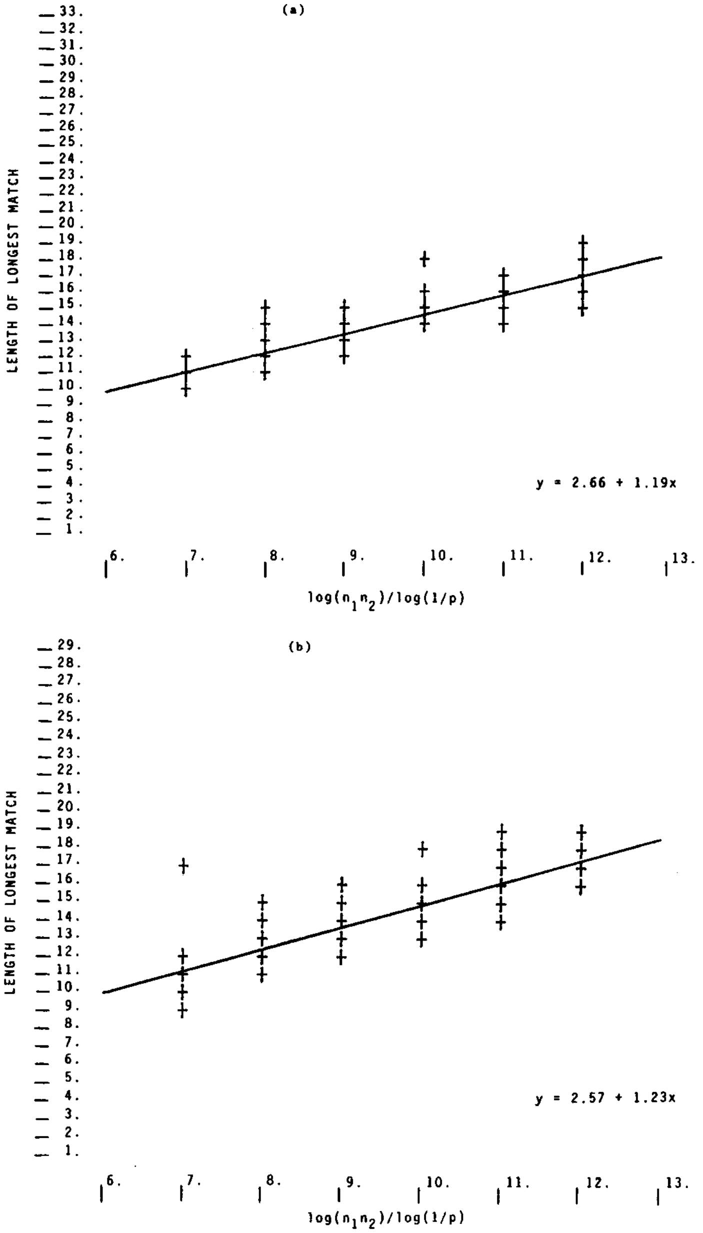
(a)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 2.66 + 1.19x
6.
7.
8.
9.
10.
11.
12.
13.
log(n1n2)/log(1/p)
(b)
LENGTH OF LONGEST MATCH
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 2.57 + 1.23x
6.
7.
8.
9.
10.
11.
12.
13.
log(n1n2)/log(1/p)

Figure 4. k = 2.

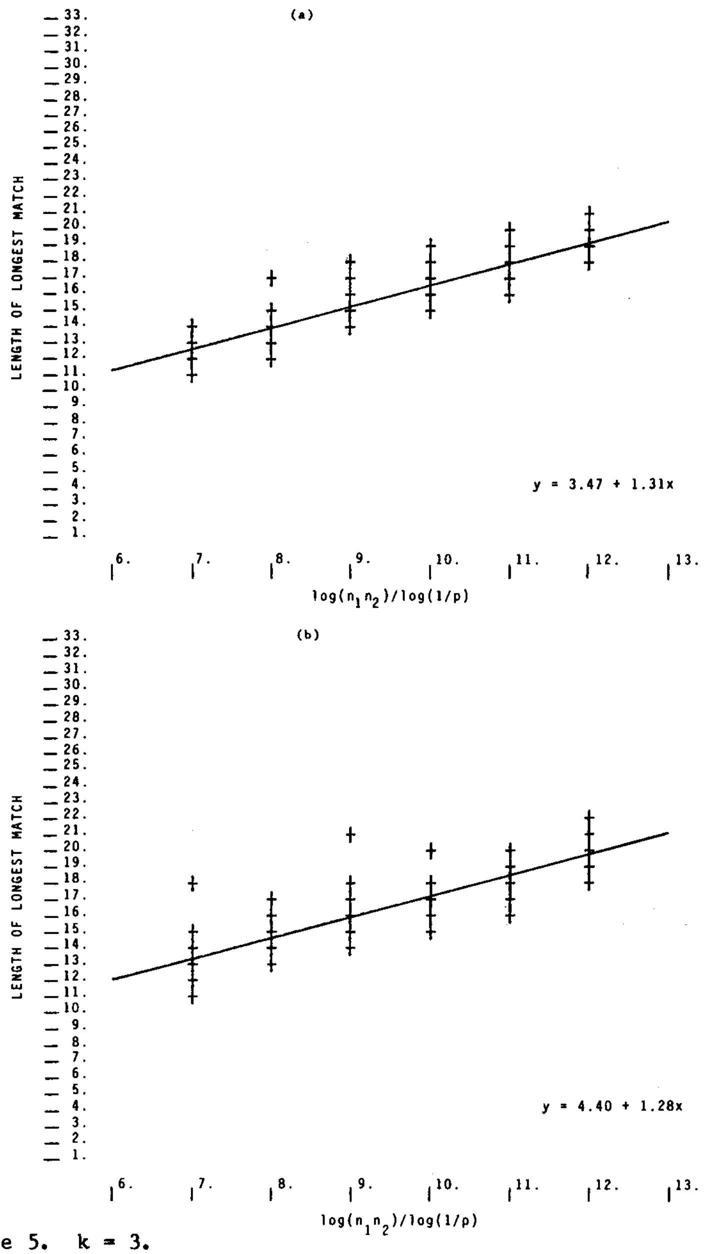
(a)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 3.47 + 1.31x
6.
7.
8.
9.
10.
11.
12.
13.
log(n_1n_2)/log(1/p)
(b)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 4.40 + 1.28x
6.
7.
8.
9.
10.
11.
12.
13.
log(n_1n_2)/log(1/p)

Figure 5. k = 3.

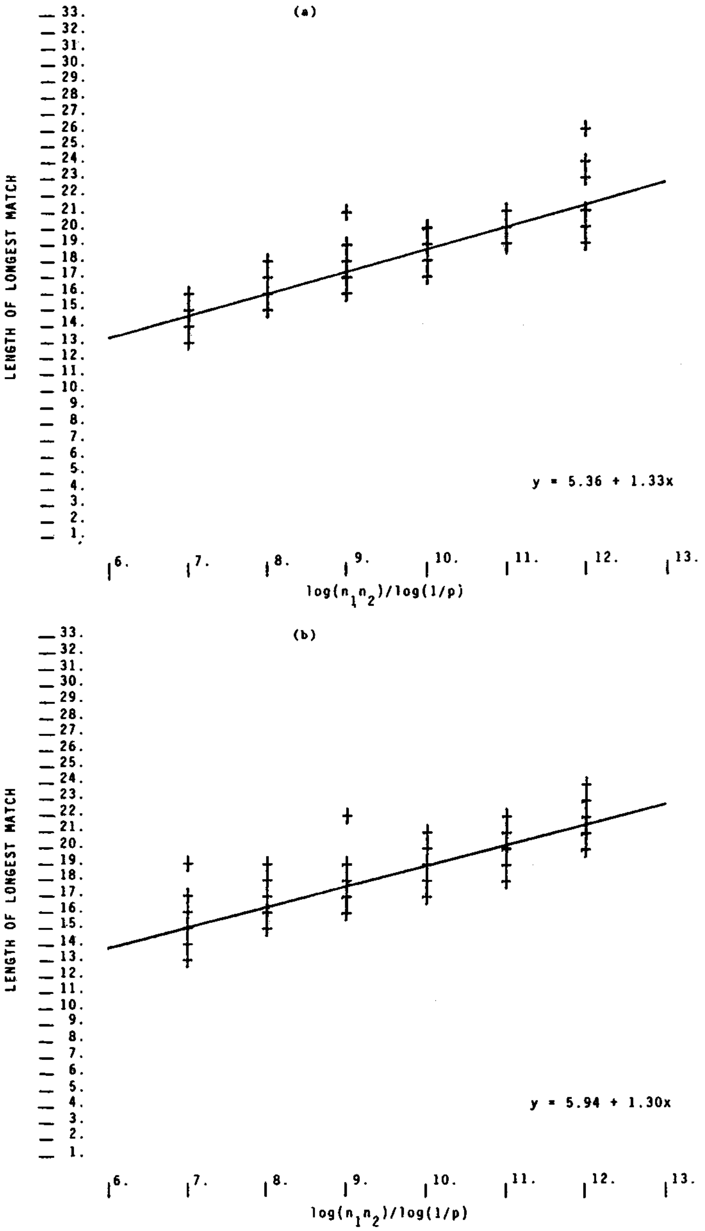
(a)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 5.36 + 1.33x
6.
7.
8.
9.
10.
11.
12.
13.
log(n1n2)/log(1/p)
(b)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 5.94 + 1.30x
6.
7.
8.
9.
10.
11.
12.
13.
log(n1n2)/log(1/p)

Figure 6. k = 4.

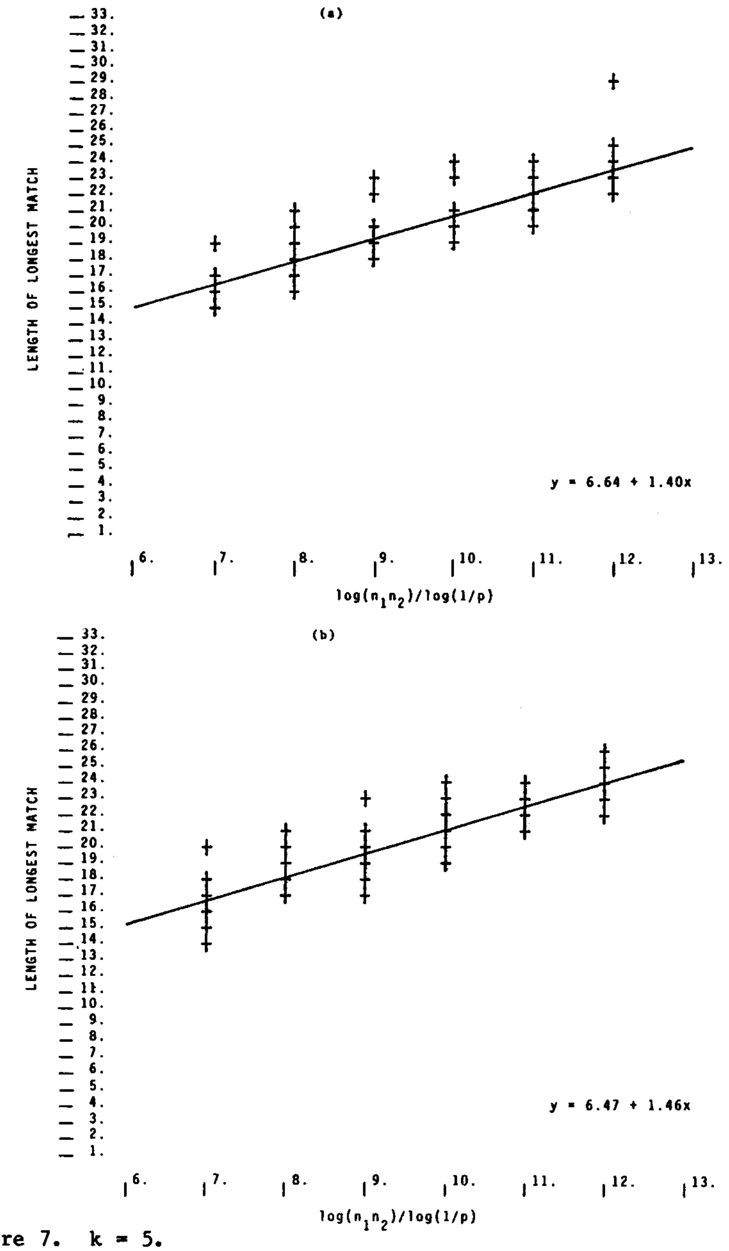
(a)
LENGTH OF LONGEST MATCH
33. 32. 31. 30. 29. 28. 27. 26. 25. 24. 23. 22. 21. 20. 19. 18. 17. 16. 15. 14. 13. 12. 11. 10. 9. 8. 7. 6. 5. 4. 3. 2. 1.
y = 6.64 + 1.40x
6. 7. 8. 9. 10. 11. 12. 13.
$\log(n_1 n_2)/\log(1/p)$
(b)
LENGTH OF LONGEST MATCH
33. 32. 31. 30. 29. 28. 27. 26. 25. 24. 23. 22. 21. 20. 19. 18. 17. 16. 15. 14. 13. 12. 11. 10. 9. 8. 7. 6. 5. 4. 3. 2. 1.
y = 6.47 + 1.46x
6. 7. 8. 9. 10. 11. 12. 13.
$\log(n_1 n_2)/\log(1/p)$

Figure 7. k = 5.

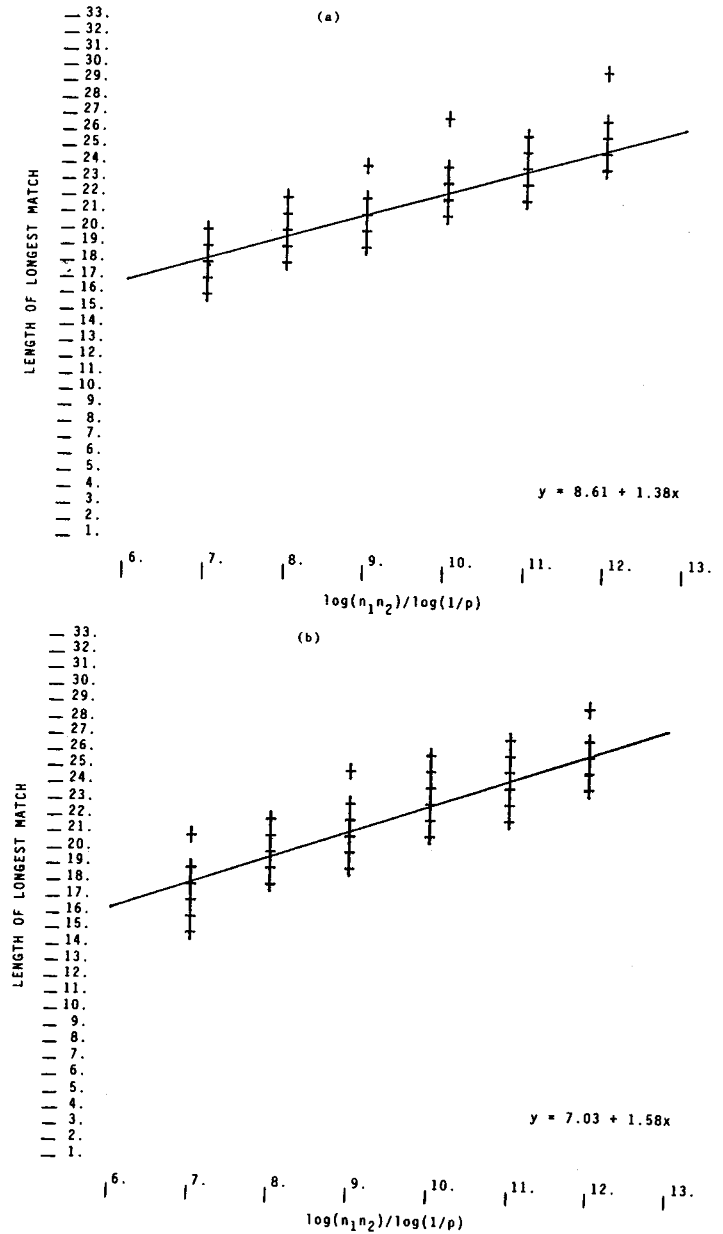
(a)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 8.61 + 1.38x
6.
7.
8.
9.
10.
11.
12.
13.
log(n1n2)/log(1/p)
(b)
LENGTH OF LONGEST MATCH
33.
32.
31.
30.
29.
28.
27.
26.
25.
24.
23.
22.
21.
20.
19.
18.
17.
16.
15.
14.
13.
12.
11.
10.
9.
8.
7.
6.
5.
4.
3.
2.
1.
y = 7.03 + 1.58x
6.
7.
8.
9.
10.
11.
12.
13.
log(n1n2)/log(1/p)

Figure 8. k = 6.

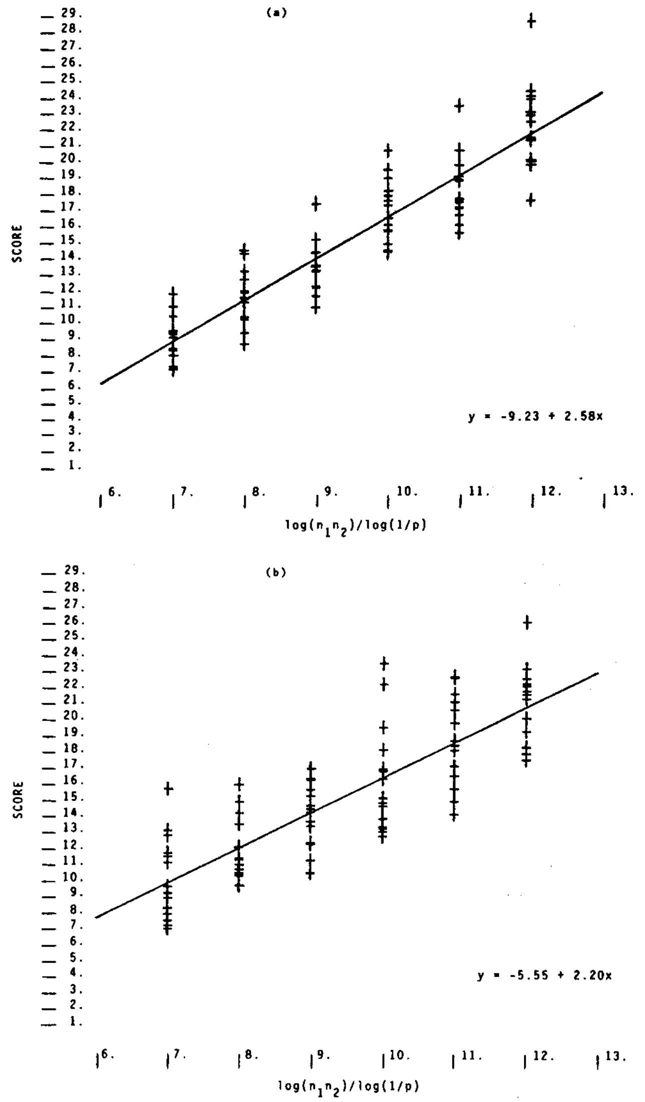

Figure 9. Best similarity scores. For each value of $x = \log_{1/p}(n_1 n_2) = \log_4(n_1 n_2)$ there are 15 sequence comparisons. (a) Monte Carlo sequences and (b) lambda sequences. The best linear fits of y = score = max (# matches - .9# mismatches - 2.0# deletions/insertions) are given by y = a + bx.

ACKNOWLEDGEMENT

This work was supported by the System Development Foundation.

BIBLIOGRAPHY

1. Arratia, R., Gordon, L., and Waterman, M.S., "An extreme value distribution for sequence matching", Manuscript, (1984).

2. Arratia, R. and Waterman, M.S., "An Erdös-Rényi law with shifts", Adv. in Math., 55(1985,a), 13-23.

3. Arratia, R. and Waterman, M.S., "Critical Phenomena in sequence matching", Ann. Prob. In press. (1985,b).

4. Chung, K.L., A Course in Probability Theory. Harcourt, Brace, and World, Inc., New York, 1968.

5. Chvatal, V. and Sankoff, D., "Longest common subsequences of two random sequences", J. Appl. Prob., 12(1975), 306-315.

6. Deken, J.G., "Some limit results for longest common subsequences", Discrete Math., 26(1979), 17-31.

7. Doolittle, R.F., Hunkapiller, M.W., Hood, L.E., Devare, S.G., Robbins, K.C., Aaronson, S.A., and Antoniades, H.M., "Simian sarcoma viruses and gene v-sis is derived from the gene (or genes) encoding platelet-derived growth factors", Science, 211(1983), 275-276.

8. Erdös, P. and Rényi, A., "On a new law of large numbers", J. Analyse Math. 22, 103-111. Reprinted in Selected Papers of Alfred Rényi, Vol. 3, 1962-1970, Akademiai Kiado, Budapest, 1976.

9. Erdös, P. and Révész, P., "On the length of the longest head-run". Topics in Information Theory, Colloquia Math. Soc. J. Bolyai 16, Keszthely (Hungary), (1975), 219-228.

10. Fitch, W.M. and Margoliash, E., "Construction of polygenetic trees", Science, 155(1967), 279-284.

11. Gordon, L., Schilling, M.F., and Waterman, M.S., "An extreme value theory for long head runs". Manuscript, (1984).

12. Guibas, L.J., and Odlyzko, A.M., "Long repetitive pattern in random sequences", Z. Warscheinlickeitstheorie verw. Gebiete. 53(1980), 241-262.

13. Hardy, Littlewood, and Polya. Inequalities, Cambridge University Press, Cambridge, 1934.

14. Karlin, S., Ghandour, G., and Foulser, D.E., and Korn, L.J., "Comparative analysis of human and bovine papillomarviruses". Mol. Biol. and Evol., 1(1984), 357-370.

15. Maxam, A.M., and Gilbert, W., "A new method for sequencing DNA", Proc. Natl. Acad. Sci., 74(1977), 560-564.

16. Naharro, G., Robbins, K.C., and Reddy, E.P., "Gene product of v-fgr onc: hybrid protein containing a portion of actin and tyrosin-specific protein kinase", Science, 223(1984), 63-66.

17. Sanger, F.S., Nicklen, F.S., and Coulson, A.R., "DNA sequencing with chain terminating inhibitors", Proc. Natl. Acad. Sci., 74(1977), 5463-5467.

18. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B., "Nucleotide sequence of bacteriophage λ DNA", J. Mol. Biol. 162(1982), 729-773.

19. Smith, T.F., and Waterman, M.S., "Identification of common molecular subsequences", J. Mol. Biol., 147(1981,a), 195-197.

20. Smith, T.F., and Waterman, M.S., "Comparison of biosequences", Adv. Appl. Math., 2(1981,b), 482-489.

21. Smith, T.F., Waterman, M.S., and Burks, C., "The statistical distribution of nucleic acid similarities". Nucl. Acids Res., 13(1985), 645-656.

22. Smith, T.F., Waterman, M.S., and Sadler, J.R., "Statistical characterization of nucleic acid sequence functional domains", Nucleic Acids Res., 11(1983), 2205-2220.

23. Steele, M.J., "Long common subsequences and the proximity of two random strings", SIAM J. Appl. Math., Vol. 42, No. 4(1982), 731-737.

24. Waterman, M.S., "General methods of sequence comparison", Bull. Math. Biol., 46(1984), 473-500.

25. Weiss, R., "Oncogenes and growth factors", Nature 304(1983), 12.

DEPARTMENTS OF MATHEMATICS AND MOLECULAR BIOLOGY
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113