Bulletin of Mathematical Biology Vol. 54, No. 5, pp. 785–812, 1992.
Printed in Great Britain.

0092–8240/9255.00+0.00
Pergamon Press Ltd

POISSON, COMPOUND POISSON AND PROCESS APPROXIMATIONS FOR TESTING STATISTICAL SIGNIFICANCE IN SEQUENCE COMPARISONS*

■ LARRY GOLDSTEIN and MICHAEL S. WATERMAN
Department of Mathematics,
University of Southern California,
Los Angeles, CA 90089–1113, U.S.A.

DNA and protein sequence comparisons are performed by a number of computational algorithms. Most of these algorithms search for the alignment of two sequences that optimizes some alignment score. It is an important problem to assess the statistical significance of a given score. In this paper we use newly developed methods for Poisson approximation to derive estimates of the statistical significance of k-word matches on a diagonal of a sequence comparison. We require at least q of the k letters of the words to match where $0<q\leqslant k$. The distribution of the number of matches on a diagonal is approximated as well as the distribution of the order statistics of the sizes of clumps of matches on the diagonal. These methods provide an easily computed approximation of the distribution of the longest exact matching word between sequences. The methods are validated using comparisons of vertebrate and *E. coli* protein sequences. In addition, we compare two HLA class II transplantation antigens by this method and contrast the results with a dynamic programming approach. Several open problems are outlined in the last section.

1. Introduction. International sequence databases provide rapid and easy access to increasing DNA and protein data. These databases are extremely important to the progress of biology. Databases can provide pointers to the scientific literature associated with a specific sequence or a family of sequences, as well as provide the information base to test hypotheses about new or existing sequences. One of the most common uses of sequence databases is to screen the database with a new sequence in order to find homologous sequences. There are some striking examples where unexpected sequence similarities were discovered by these database searches. Platelet derived growth factor and the *v-sis* oncogene product are highly similar and the similarity was found by a database search. Because of this similarity, it is now believed that the *v-sis* gene encodes a growth factor. Another example of a computer discovery is the similarity between bovine cyclic AMP dependent kinase and the Rousavian and Maloney murine sarcoma virus *src* proteins. This similarity supports the origination of the *src* genes in host genomes. In addition to more dramatic findings, sequence comparisons are used routinely to create or test hypotheses

* This work was supported by grants DMS 90-05833 from NSF and GM 36230 from NIH.

about the function of a protein or DNA sequence or about the membership of a sequence in a family.

Many algorithms have been developed to find similarities between two sequences. The most widely used family of programs for database searches has been developed by Wilbur and Lipman (1983), Lipman and Pearson (1985) and Pearson and Lipman (1988) and includes the programs FASTN, FASTP, FASTA and LFASTA. These useful and rapid programs are based on locating regions which have an unusual degree of similarity. After these regions are located, a more refined analysis is applied to produce the best alignments. The recent paper by Pearson (1990) gives a nice review of these methods. Among other methods developed to compare sequences, dynamic programming is the method in widest use. Beginning with Needleman and Wunsch (1970), many refinements and improvements have been developed (for a review, see Waterman, 1984, 1989). In fact the final alignments in the FASTP, FASTA and LFASTA use these dynamic programming algorithms, restricted to a specific region to decrease computation time.

One of the key steps in any database search is evaluation of scores, or numerical measures of sequence similarity for sequence relationships that warrant further examinations. When tens of thousands of comparisons are made, it is not possible to manually examine them all; statistical methods are usually employed. Moreover, it is useful to know the statistical distribution of alignment scores under a null distribution. This allows the scientist to accurately estimate p-values and to test hypotheses about the relatedness of sequences. Of course the null distribution of alignment scores is useful for sequence comparisons that are not part of full database searches. The purpose of the present paper is to apply and extend some recent results in probability theory to a class of problems in sequence comparisons.

In Arratia *et al.* (1989) a general approach is presented for the Poisson approximation of dependent events. This approach has the advantages of being simple to use and intuitive, as well as providing bounds for the error incurred in making the approximation itself. This general method was motivated by our interests in sequence comparisons, in particular the distribution of the Smith and Waterman (1981) dynamic programming score functions. In Smith *et al.* (1985), a CRAY was used to compute a large number of Smith–Waterman scores by comparing DNA sequences from GenBank. The central term in the growth of score with separate lengths is $O(\log(nm))$ where n and m are the lengths of the two sequences being compared. A number of mathematical results were motivated by this study. The probability distribution of the longest exact match between two random sequences is known as well as the longest match with k mismatches (Arratia *et al.*, 1986). Also see Karlin (1983) for an announcement of related results. The extreme value distribution is used to approximate scores which exceed the central term. With Arratia *et al.* (1989) and Arratia *et al.* (1990b), the Poisson approximation was developed and even more difficult problems could be approached. Generalization to sequence matching with scores was begun by Arratia *et al.* (1988). In Karlin and Altschul (1990) much more general scoring schemes are considered. Their Poisson approximation formula is used to assess statistical significance in the new database search method BLAST (Altschul *et al.*, 1990). Finally Arratia *et al.* (1990a) and Arratia and Waterman (1989) study the distribution of the longest region with at least $a \times 100\%$ matches, where $a > P$ (two random letters match). Poisson approximation is key to proving the results in Arratia *et al.* (1990a).

In this paper we study variants of the following problem. Choose a word size k. For sequences $A_1A_2 \ldots A_{n_1}$, $B_1B_2 \ldots B_{n_2}$ consider diagonals (A_i, B_j) with $i-j$ constant. For various quality matching, we study the k-word matches on the diagonals. Many of our results are derived for one diagonal. In the examples and discussion (Sections 7 and 8), we consider all the diagonals. No insertions or deletions are allowed in our analysis. In return for this limitation we obtain many results that are computable and in closed form. If we require q of k letters to be identical in order to declare a k-word match, each match has quality $a = q/k$. In contrast with Arratia *et al.* (1990a), where they study long regions of quality $a = q/k$, we study the statistical properties of k-word matches, where k is not required to be large. Of course when $q = k = 1$, we are simply looking at single letter matches on a diagonal.

The simplicity and power of these Poisson approximation techniques are not too widely known, so we have included simplified versions of the two basic theorems in Section 2. The second theorem is a process version for Poisson approximation which is necessary to prove some of our results. The outline of the paper is:

Section 2. Two Poisson approximation theorems.
Section 3. Clump size distribution.
  3.1. Perfect matching.
Section 4. Poisson approximation for quality q/k matches.
  4.1. Perfect matching.
    4.1.1. $k = 1$, $s = 0$.
    4.1.2. $k > 1$, $s = 1$.
Section 5. Approximation by the compound Poisson distribution.
  5.1. Perfect matching.
Section 6. Maximum clump size distribution.
Section 7. Data analysis.
Section 8. Discussion: open problems.

Section 3 studies the "clumps" of 1's along the diagonal, where 1 denotes a quality q/k match. This clumping occurs because there is dependence between overlapping k-words. Clump is rigorously defined, and in Theorem 3 reflection principal techniques are applied to approximate the clump size distribution.

The case of perfect matching, $q=k$, has a nice geometric clump size distribution and is given in Section 3.1. While the size of the clumps are not equivalent to other scores such as LFASTA and Smith–Waterman obtain, they are analogous.

In Section 4 we turn to studying the number of clumps on a diagonal, which is approximated by a Poisson distribution. Bounds on the approximation can be given for one choice of clump definition. When $q=k$ more precise information is presented in Section 4.1. $k=1$ gives the binomial distribution (4.1.1) while $k>1$ is handled in (4.1.2).

In Section 5, we study the total number of quality q/k k-word matches on a diagonal. As we will see, there are approximately a Poisson number of clumps, and the clump size distribution is also known. Therefore the total number of q/k matches on a diagonal should be a compound Poisson distribution. Again we can give bounds on the approximation. The perfect matching case of $q=k$ has an easily computable form. In spite of the "local" nature of the FASTA family of programs, many investigators have used the diagonal/hashing technique described in Wilbur and Lipman (1983) to compute total hits on a diagonal. In fact a recursive scheme to approximate the statistical significance of total number of perfect matches when $q=k$ has been given by Mott *et al.* (1989). Their recursions do not have an analytical solution and must be solved numerically.

When scanning a dot matrix for regions of biological significance, the maximum clump size distribution is of much interest. Section 6 studies this relevant random variable. In fact we study the order statistics for the clump size along a diagonal and obtain nice formulas for their distribution. In this case the maximum clump size for all diagonals can be studied. As is expected in this problem, the largest clump, in the case $q=k$, has an integerized extreme value distribution.

Section 7 presents some data analysis.

Finally Section 8 discusses some open problems. Our treatment deals with a diagonal in isolation, while real sequence comparisons must simultaneously consider all diagonals. The dependence has been studied for other, related problems. Sequences have unequal lengths and distributions. If the differences are not too great, the same results should hold. Handling non-independent letters is also a problem of interest. Our approximate matching k-words do not allow insertions and deletions, a feature that would be very difficult to include in the analysis. Finally, the Poisson approximation theorems give error bounds for total variation distance while relative error would be very valuable.

2. Two Poisson Approximation Theorems. We first present the Poisson approximation theorems that we apply below. Somewhat more general versions of these theorems appear in Arratia *et al.* (1989) and Arratia *et al.* (1990a).

Let I be an index set, and assume that for each $\alpha\in I$ there is a Bernoulli random variable $X_\alpha\in\{0,1\}$. The random variable X_α denotes the occurrence ($X_\alpha=1$) or non-occurrence ($X_\alpha=0$) of an event. Let $p_\alpha=EX_\alpha$:

$$W=\sum_{\alpha\in I} X_\alpha,$$

and $\lambda=EW$. W is the total number of occurrences of events; λ is the expected number. Should each p_α be small, $|I|$, the size of index set large, and the X_α not too dependent, then we should expect W to be approximately equal in distribution to Z, a Poisson random variable with parameter λ. This fact is well known when the X_α are independent. The theorems that follow demonstrate that the approximation is valid in cases where there is dependence, and they provide a bound on the error in the approximation.

In many examples, and in particular all those that concern us below, the dependence between the X_α can be confided in a "neighborhood of dependence". More precisely, suppose that for each α we can find a set B_α such that:

$$X_\alpha \text{ is independent of } \{X_\beta\}, \quad \beta\notin B_\alpha. \tag{1}$$

Define:

$$b_1\equiv\sum_{\alpha\in I}\sum_{\beta\in B_\alpha} p_\alpha p_\beta$$

and

$$b_2\equiv\sum_{\alpha\in I}\sum_{\alpha\neq\beta\in B_\alpha} p_{\alpha\beta}, \quad \text{where } p_{\alpha\beta}\equiv E(X_\alpha X_\beta). \tag{2}$$

Let Z denote a Poisson random variable with mean λ, so that for $k=0,1,2,\ldots, P(Z=k)=e^{-\lambda}\lambda^k/k!$. For h a real valued function let $\|h\|\equiv\sup_k|h(k)|$. We denote the total variation distance between the distributions of X and Y by:

$$\|X-Y\|\equiv\sup_{\|h\|=1}|Eh(X)-Eh(Y)|=2\sup_A|P(X\in A)-P(Y\in A)|.$$

A more general version of the following theorems appears in Arratia *et al.* (1989).

THEOREM 1. *Let* W *be the number of occurrences of dependent events, and let* Z *be a Poisson random variable with* $EZ=EW=\lambda$. *Then, under condition* (1):

$$\tfrac{1}{2}\|W-Z\|\leq(b_1+b_2)\frac{1-e^{-\lambda}}{\lambda}\leq(b_1+b_2).$$

To approximate the distributions occurring below that depend on the entire process of indicators $\{X_\alpha\}_{\alpha\in I}$ we will use the following theorem. We note that the set B_α used in the calculation of b_1 and b_2 is not necessarily the same in Theorems 1 and 2.

THEOREM 2. *For each* $\alpha\in$I, *let* Z_α *be a Poisson random variable with mean* p_α, *with the* Z_α *mutually independent. Then, under condition* 1, *the total variation distance between the process* $\mathbf{X}=\{X_\alpha\}_{\alpha\in I}$ *and the Poisson process* $\mathbf{Z}=\{Z_\alpha\}_{\alpha\in I}$ *satisfies:*

$$\tfrac{1}{2}\,\|\mathbf{X}-\mathbf{Z}\|\leqslant(2b_1+2b_2).$$

3. Clump Size Distribution. Consider two sequences $\mathbf{A}=A_1A_2\ldots A_{n_1}$ and $\mathbf{B}=B_1B_2\ldots B_{n_2}$ from an alphabet $\{l_1,\ldots,l_d\}$ of size d. Suppose that both sequences are composed of independent letters with distribution $P(A=l_i)=r_i$, $P(B=l_i)=s_i$. The probability that two letters match is:

$$p=\sum_{i=1}^{d} r_is_i.$$

Recall that *diagonals* are made up of (A_i, B_j) pairs with $j-i$ constant; that is, a sequence comparison at fixed offset. By reindexing the portion of both sequences that are involved in the comparison if necessary, this segment taken to be of length n, we see that without loss of generality all diagonals we study may be considered as comparisons between two sequences of equal length:

$$\begin{array}{l}A_1\,A_2\ldots A_n\\ B_1\,B_2\ldots B_n.\end{array}$$

Consider a window of length k beginning at position $\alpha\in I$, where:

$$I=\{\alpha\colon 1\leqslant\alpha\leqslant n-k+1\};$$

this window is made up of the portion of the above two sequences at positions $\alpha,\alpha+1,\ldots,\alpha+k-1$, that is the k-words from the two sequences. In such a window we look for a "quality q/k match", that is, whether we have q or more of the k (A, B) pairs of letters in the window agree or are identical.

More precisely, let:

$$D_\alpha=\mathbf{1}(A_\alpha=B_\alpha),$$

that is, D_α is 1 if the letters in the αth position agree, and zero otherwise. The window of length k that begins at position α has:

$$M_\alpha=D_\alpha+D_{\alpha+1}+\cdots+D_{\alpha+k-1}$$

matching letters. We say we have a quality q/k match begin at position α if:

$$Y_\alpha=\mathbf{1}(M_\alpha\geqslant q)$$

takes the value 1.

The locations where there is a quality q/k match, that is, the colleciton of α for which $Y_\alpha=1$, tend to occur in "clumps". We will see that by defining a "clump" appropriately, the total number of clumps W will be approximately Poisson. This is an example of the Poisson clumping heuristic of Aldous (1989). In what follows, the clump size, that is, the number of indices α in a clump for which $Y_\alpha=1$, is of central importance. We now define the clump distribution.

The Bernoulli variable X_α will indicate that a clump begins at position α. The random variables X_α will be obtained from Y_α by:

$$X_\alpha=Y_\alpha(1-Y_{\alpha-1})\ldots(1-Y_{\alpha-s})=Y_\alpha\prod_{j=1}^{s}(1-Y_{\alpha-j}).$$

The parameter s will depend on the quality q and the word size k. For example, if $q=k>1$, we see below that $s=1$ is the appropriate choice. We will write $q/k/s$ when referring to these parameters. We take $Y_\alpha=0$ for $\alpha\leqslant 0$. In words, we say a new clump begins at position α if the window at position α contains a quality q/k match, but there were no quality q/k matches at position $\alpha-1,\alpha-2,\ldots,\alpha-s$. Let C be the random variable that counts the number of q/k matches in a clump. More precisely, suppose a clump begins at position α, that is, suppose that $X_\alpha=1$. The clump that begins at position α ends at the index $\beta=\beta_\alpha$ where:

$$\beta=\min\{\gamma\geqslant\alpha\colon Y_\gamma=1,\ Y_{\gamma+1}=0,\ldots,Y_{\gamma+s}=0\}. \tag{3}$$

Hence, for such β, C is the random variable with distribution given by:

$$P(C=m)=P\left(\sum_{\gamma=\alpha}^{\beta}Y_\gamma=m\,\middle|\,X_\alpha=1\right)\quad m=1,2,\ldots. \tag{4}$$

For the purpose of defining C, we take the sequence of indicators D_α to be doubly infinite.

Consider a window of length k starting at position α containing exactly q matches. The window at position $\alpha+1$ shares all but one of its entries with the window at α; it has either $q-1$, q, or $q+1$ matches. In particular, $M_{\alpha+1}=M_\alpha+(D_{\alpha+k}-D_\alpha)$; the increment $E_\alpha=D_{\alpha+k}-D_\alpha$ is either -1, 0 or 1. Defining:

$$E_t=D_{t+k}-D_t$$

we see that if $X_\alpha=1$ and $\gamma>\alpha$:

$$Y_\gamma = 1 \text{ if and only if } \sum_{t=\alpha+1}^{\gamma} E_t \geqslant 0.$$

In what follows, it is useful to visualize the collection of the above partial sums as "paths" in the sense of Feller (1968, Vol. I, Ch. 3). The height of the path V_α at α corresponds to the number of matches in the k long window starting at α and either gain one unit of height, lose one unit of height, or stay constant when moving, i.e. sliding the window, one step to the right.

LEMMA 3. *For* $j \geqslant 1$ *the number of paths from* $(0, j)$ *to* $(l, 0)$, *which are strictly positive at times* $1, 2, \ldots, l-1$ *and have* u *negative steps is equal to:*

$$\frac{j}{u}\binom{l-1}{u-j,\, u-1,\, l-2u+j}, \quad \text{if } j \leqslant u \leqslant j + \left\lfloor \tfrac{l-j}{2} \right\rfloor.$$

More precisely, this expression counts the number of paths:

$$V_\alpha = j + \sum_{t=1}^{\alpha} E_t$$

such that $V_\alpha > 0$ *for* $1 \leqslant \alpha \leqslant l-1$, $V_l = 0$ *and* $\sum_{t=1}^{l} 1\{E_t = -1\} = u$.

This is the same as the number of paths which begin at $(0, 0)$, *end at* $(l, -j)$, *are strictly negative at times* $1, 2, \ldots l-1$ *and have* u *negative steps.*

For $j = 0$, *the number of paths* $V_\alpha = \sum_{t=1}^{\alpha} E_t$ *such that* $V_\alpha > 0$ *for* $1 \leqslant \alpha \leqslant l-1$, $V_l = 0$ *and* $\sum_{t=1}^{\alpha} 1\{E_t = -1\} = u$ *is:*

$$\frac{1}{u}\binom{l-2}{u-1,\, u-1,\, l-2u}, \quad \text{if } 1 \leqslant u \leqslant \left\lfloor \tfrac{l}{2} \right\rfloor.$$

Proof. By the reflection principle, the first expression is the number of paths from $(0, j)$ to $(l-1, -1)$, with u negative steps, subtracted from the number of paths from $(0, j)$ to $(l-1, 1)$ with $u-1$ negative steps:

$$\binom{l-1}{u-j,\, u-1,\, l-2u+j} - \binom{l-1}{u-j-1,\, u,\, l-2u+j}.$$

That this is the same as the number of paths of the second type is true by reflection, time reversal and translation. The argument for $j = 0$ is similar. ■

Feller (1968, Vol. I, Ch. 3) considers paths which go only up or down; that is, they cannot remain constant. In particular, Feller's formula 3.7 follows as a special case of the above expression for $j = 0$ when $2u = l$.

Under an independent increment approximation, the clump size distribution $P(C = m)$ is derived by an application of Theorem 4 following the proof.

THEOREM 4. *Let:*

$$P(E = 1) = p_+, \quad P(E = -1) = p_-, \quad P(E = 0) = p_0$$

and suppose $E_1, E_2, \ldots$ *are independent each with the distribution of* E. *For* $j \geqslant 0$ *let:*

$$V_\alpha = j + \sum_{t=1}^{\alpha} E_t.$$

Define:

$$\beta = \inf\{\alpha: V_\alpha = 0, V_{\alpha+1} < 0, \ldots, V_{\alpha+s} < 0\},$$
$$R^+ = \inf\{l: V_\alpha > 0, 1 \leqslant \alpha \leqslant l-1, V_l = 0\},$$
$$f_{j,l} = P(R^+ = l),$$
$$\tilde{C} = \left|\{\alpha: 1 \leqslant \alpha \leqslant \beta, V_\alpha \geqslant 0\}\right|,$$

and finally:

$$q_m = P(\tilde{C} = m) \text{ with } j = 0.$$

Then:

$$f_{0,l} = \begin{cases} j = 1 f_{0,j} & \text{if } l = 1 \\ \sum_{u=1}^{[l/2]} \frac{1}{u}\binom{l-2}{u-1,\, u-1,\, l-2u} [p_+ p_-]^u p_0^{l-2u} & \text{if } l \geqslant 2. \end{cases} \tag{5}$$

For $j \geqslant 1$, $f_{j,l} = 0$ *if* $l < j$. *Otherwise, for* $l \geqslant j$:

$$f_{j,l} = \sum_{u=j}^{j+[(l-j)/2]} \frac{j}{u}\binom{l-1}{u-j,\, u-1,\, l-2u+j} p_+^{u-j} p_-^u p_0^{l-2u+j}. \tag{6}$$

Furthermore, letting:

$$b_i = \begin{cases} p_0 + \sum_{i=2}^{s} f_{0,i} & \text{if } i = 1 \\ f_{0,i} & \text{if } i \geqslant 2, \end{cases} \tag{7}$$

we have:

$$q_m = \begin{cases} \sum_{t=1}^{s} f_{t,s} & \text{if } m=1 \\ \sum_{i=1}^{m-1} b_i q_{m-i} & \text{if } m \geq 2. \end{cases} \tag{8}$$

A solution to the above recursion yields an explicit formula for q_m, *given for* $m=1$ *in the equation above and otherwise by:*

$$q_{m+1} = q_1 \sum_{\mathbf{a} \in \mathscr{A}} \binom{|\mathbf{a}|}{a_1, a_2, \ldots, a_m} \prod_{i=1}^{m} b_i^{a_i} \quad \text{if } m \geq 1 \tag{9}$$

where:

$$\mathbf{a} = (a_1, a_2, \ldots, a_m), \quad \mathscr{A} = \left\{ \mathbf{a} : a_i \geq 0, \sum_{i=1}^{m} i a_i = m \right\} \quad \text{and } |\mathbf{a}| = \sum_{i=1}^{m} a_i.$$

Proof. Consider first $f_{0,l}$. Clearly, $f_{0,1} = P(R^+ = 1) = p_0$. Otherwise, $f_{0,l} = P(R^+ = l)$ is calculated by using Lemma 3 to count the number of paths that have first return time at l which use u negative steps, multiplying by their probability, and summing. The probability $f_{j,l}$ is calculated similarly. In order for $\tilde{C}=1$, the path must go from $(0, 0)$ to $(1, -1)$ and end at $(s, -t)$ for $t = 1, 2, \ldots, s$ without touching or hitting the axis. Ending at a given t has probability $f_{t,s}$, by Lemma 3, summing yields q_1. To derive the recursion for q_m we argue as follows. The first increment E_1 is either -1, 0 or 1. If $E_1 > 0$ then $R^+ \leq m-1$, else $\tilde{C} > m$. If $R^+ = i$ for $2 \leq i \leq m-1$, an event of probability $f_{0,i}$, then the portion of the path between time 0 and i leaves us to collect $m-i$ more such indices on the new clump that begins at i, and this occurs with probability q_{m-i}. *If* $E_1 = 0$, which occurs with probability p_0, we collect a total of m indices j for which $V_j \geq 0$ if the new clump starting at $(1, 0)$ has a total of $m-1$ such indices. This occurs with probability q_{m-1}. Defining R^- in the obvious way, if $E_1 < 0$ then $R^- \leq s$, else $\beta = 0$. For $E_1 < 0$ and a return time at i, $2 \leq i \leq s$, we need to collect $m-1$ more such indices, which occurs with probability q_{m-1}. ∎

Although the increments of the path that records q/k matches are not independent, one may obtain a good first order approximation to the clump distribution by replacing the increments of the path by independent increments. This is especially of interest since the actual clump distribution based on the exchangeability of the increments of the true path is complex. This analysis will appear elsewhere.

To use an independent increment approximation, consider that given $X_\alpha = 1$,

a Bernoulli variable that makes up the window beginning at α has probability $a = q/k$ of being 1; the Bernoulli variables outside this window are 1 with probability p. Hence, we should have a net loss of 1 with probability $a(1-p)$ in sliding this window over one position: we lose a 1 inside the window with probability a and gain a zero outside the window with probability $1-p$. Similar remarks apply to obtain the probabilities of having a net loss of 0 or net gain of 1. Note that the Bernoulli variables inside the window, though exchangeable, are not quite independent; their sum is constrained to add to q, for example.

Hence, with:

$$p_+ = p(1-a), \quad p_- = a(1-p), \quad p_0 = ap + (1-a)(1-p),$$

the probabilities q_m from the above equations give good approximations to $P(C=m)$. For example, writing $q/k/s$ for our parameters, for 6/12/3 matching we have simulated and calculated values of $E[C]$ and $E[\tilde{C}]$ of 4.57 and 4.92.

3.1. Perfect matching. $q=k>1$, $s=1$. When $s=1$ it is easy to see that $q_1 = p_- = (1-p)$. Furthermore, when $q=k$ we have all Bernoulli variables in a window where $X_\alpha = 1$ are 1. Hence, the clump propagates when shifting the window one unit if and only if the Bernoulli variable outside the window at position $\alpha + k$ is 1. Therefore, the clump is of size at least x with probability p^x. In this case, the distributions of C and $\tilde{C}$ coincide; one may verify that equation (9) reduces to the geometric when $a=1$ and $s=1$: $P(C=m) = P(\tilde{C}=m) = (1-p)p^m$.

4. Poisson Approximation for Number of Clumps. Suppose the quality q of a window of length k is chosen so that there are only few windows on the average that have quality q/k matches. In this case the total number of clumps should be approximately Poisson.

The total number of clumps on a diagonal is:

$$W = \sum_{\alpha=1}^{n-k+1} X_\alpha.$$

We note that W is the sum of *dependent* indicator random variables. With proper choice of B_α so that condition (1) is satisfied, Theorem 1 yields the following corollary on how far the distribution of W is from that of the Poisson random variable Z with the same mean.

COROLLARY 5. *Let* Z *be Poisson with mean* $\lambda = \mathrm{EW}$. *Then:*

$$\tfrac{1}{2} \|\mathrm{W} - \mathrm{Z}\| \leq (b_1 + b_2)(1 - e^{-\lambda})/\lambda,$$

where b_1 *and* b_2 *are given by equation* (2) *where:*

$$B_\alpha = \{\beta: |\beta - \alpha| \leqslant k+s-1\}.$$

We estimate the quantities above as follows. First, ignoring boundary effects we have that λ is approximately $(n-k+1)p_\alpha$. Now:

$$p_\alpha = EX_\alpha = P(X_\alpha = 1) = P(X_\alpha = 1 \,|\, Y_\alpha = 1)P(Y_\alpha = 1).$$

Easily:

$$P(Y_\alpha = 1) = P(Bin(k, p) \geqslant q) = \sum_{j=q}^{k} \binom{k}{j} p^j (1-p)^{k-j}.$$

Furthermore $P(X_\alpha = 1 \,|\, Y_\alpha = 1)$ is the average proportion of time a q/k match starts a clump, and therefore the same as the average number of clumps per matches, or, in other words, the reciprocal of the average clump size, that is, $P(X_\alpha = 1 \,|\, Y_\alpha = 1) = 1/E[C]$. Hence:

$$\lambda \doteq (n-k+1)P(\mathrm{Bin}(k, p) \geqslant q)/E[C].$$

For a bound on the error, first take $s = k$; this choice asymptotically will drive the error bound to zero as $n \to \infty$ (Arratia *et al.*, 1989). Furthermore, it can be shown that with $s = k$:

$$a - p \leqslant P(X_\alpha = 1 \,|\, Y_\alpha = 1) \leqslant (a-p) + 2(1-a)e^{-kH(a,p)}$$

where $H(a, p)$ is defined by:

$$H(a, p) = a \log(a/p) + (1-a) \log((1-a)/(1-p)). \tag{10}$$

However, for small window sizes or cases with a close to p, $(a-p)$ may not yield a good approximation to $1/E[C]$.

For our choice of B_α:

$$b_1 = |I||B_\alpha|p_\alpha^2 = \lambda^2 \frac{|B_\alpha|}{|I|} < 2k/(n-k).$$

Taking $\beta > \alpha$, note that $X_\alpha X_\beta = 0$ for $\beta \in B_\alpha$ and $\alpha < \beta \leqslant \alpha + k$. Otherwise, for $\alpha + k < \beta \leqslant 2k$ we have $p_{\alpha\beta} = E[X_\alpha X_\beta] \leqslant E[X_\alpha Y_\beta] = E[X_\alpha]Y_b] = E[C]p_\alpha^2$. Hence, $b_2 < E[C]b_1$, and we may bound the error by $(1 + E[C])2k/(n-k)$.

We are not constrained to take $s = k$. In fact, we have found through simulation that for "small" (non-asymptotic) word sizes, the total number of clumps follow a Poisson distribution more closely for other choices of s. The value of s that gives the best fit to the Poisson may be found in Table 1. This table was constructed by simulation of 1 000 diagonals of length $n = 1\,000$ with a matching probability of $p = 0.25$. For these s, we would expect the error in the

fit to the Poisson to be less; hence the above calculation with $s = k$ may be used as an upper bound. In the table, entries of $\sim$ appear where $q/k \leqslant 0.25$, and entries of * appear where there were less than five hits expected in the 1 000 simulations. We should remark that the larger values of s vary somewhat with the simulation, with very little difference in the fit to the Poisson.

4.1. Perfect matching: q = k. In the case $q = k$ more precise information is available.

4.1.1. k = 1, s = 0. When $q = k = 1$ and $s = 0$ the sum W is the total number of matches on the diagonal and has a simple binomial distribution $B(n, p)$. In this instance the formula:

$$P(W > an) = \sum_{j=[an]+1}^{n} \binom{n}{j} p^j (1-p)^{n-j}$$

is exact but often computationally unfeasible. One may easily apply Theorem 1 to obtain a Poisson approximation. With $B_\alpha = \{\alpha\}$ condition (1) is satisfied, and Theorem 1 yields the following:

$k \backslash q$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	1	1																		
3	2	1	1																	
4	4	1	1	1																
5	~	1	1	5	1	1														
6	~	1	2	3	3	1	1													
7	~	1	1	2	3	2	3	1												
8	~	1	1	1	2	6	6	3	*											
9	~	~	1	1	2	2	3	4	1	*										
10	~	~	1	1	2	3	4	2	6	*	*									
11	~	~	1	1	2	3	5	4	6	2	*	*								
12	~	~	1	1	2	3	4	4	6	4	1	*	*							
13	~	~	~	1	1	2	4	6	9	7	4	*	*	*						
14	~	~	~	1	1	2	4	5	7	6	4	3	*	*	*					
15	~	~	~	1	1	2	3	5	12	5	4	3	*	*	*	*				
16	~	~	~	1	1	1	2	4	5	15	7	3	2	*	*	*	*			
17	~	~	~	~	1	2	2	4	5	14	5	9	4	*	*	*	*	*		
18	~	~	~	~	1	1	2	4	5	7	10	7	3	2	*	*	*	*	*	
19	~	~	~	~	1	1	2	4	4	6	19	7	6	2	*	*	*	*	*	*
20	~	~	~	~	1															

Table 1. Optimal values of s for q/k

COROLLARY 6.

$$\tfrac{1}{2}\|W - Z\| \leqslant np^2,$$

where Z *is Poisson with mean* np.

Hence, $P(W>an)$ is approximated by $P(Z>an)$ to within an error of np^2.

Other useful bounds and approximations using the methods of large deviations are given in Arratia and Gordon (1989) and mentioned here for completeness. These approximations are not only very accurate but intuitively informative as well.

THEOREM 7. *For* $0<p<a<1$, *large deviation theory for the binomial distribution yields that:*

$$P(W\geqslant an)\leqslant e^{-nH},$$

and

$$P(W\geqslant an)\sim\frac{1}{1-r}\left(\frac{1}{\sqrt{2\pi a(1-a)n}}\right)e^{-nH},$$

where $H=H(a,p)$ *and is defined in* (10) *and*

$$r=p(1-a)/(a(1-p)).$$

We write $f(x)\sim g(x)$ *to mean* $\lim_{x\to\infty} f(x)/g(x)=1$.

4.1.2. $k>1$, $s=1$. For the case of perfect matching $q=k>1$ and $s=1$, the value of λ and upper bounds on b_1 and b_2 can be made explicit (see Arratia *et al.*, 1989, for a more complete treatment). Here, all the X_α have the same distribution for $\alpha\geqslant 2$. We have:

$$\lambda=p^k\{1+(n-k)(1-p)\} \tag{11}$$

exactly. Take:

$$B_\alpha=\{\beta:|\beta-\alpha|\leqslant k\}.$$

By breaking up the sum for b_1 into terms which include $\alpha=1$ and those which do not:

$$b_1<\lambda^2(2k+1)/(n-k+1)+2\lambda p^k. \tag{12}$$

Due to the term $(1-Y_{\alpha-1})$, a clump must begin with a mismatch. Hence, for $\beta\in B_\alpha$, $\beta\neq\alpha$ we have $X_\alpha X_\beta=0$. Hence, $b_2=0$.

In the special case $q=k>1$ and a uniform distribution on the letters in the alphabet, matches in different diagonals are independent. We use Theorem 1 to obtain an approximation to the distribution of the length of the longest matching word. Let $N=N(n_1,n_2,k)$ be the maximum number of matches between all words of length k from sequence **A** compared to all words of length k from **B**, of lengths n_1, n_2, respectively. Then the probability $P(N<k)$, that is, the probability that the best match between any k word in **A** and any k word in **B** is less than k can be approximated using the equivalence:

$$\{N<k\}=\{W=0\},$$

where W is a sum of indicator variables, and hence approximately Poisson by Theorem 1. In this case, in analogy to equation (11) (see Goldstein, 1990 for details):

$$\lambda=p^k\{(n_1+n_2-2k+1)+(n_1-k)(n_2-k)(1-p)\},$$

and we have the closed form approximation with error bound:

$$|P(N<k)-\exp(-\lambda)|<(b_1+b_2).$$

Since the uniformity of the alphabet gives independence between matches on different diagonals, we have by analogy with the above and equation (12) $b_2=0$ and an error bound of no more than:

$$b_1<\lambda^2(2k+1)/((n_1-k+1)(n_2-k+1))+2\lambda p^k.$$

For example, in matching two sequences of DNA, one of length $n_1=103$ and the other of length $n_2=154$, a common word of length $k=9$ corresponds to a $\lambda=0.04$, and hence a probability of $\exp(-\lambda)=0.039$ with an error of less than 2.5×10^{-6}. Mott *et al.* (1990) consider a recursive formula to approximate the probability of such a match, in their case considerably more computation is required and error bounds are not available.

Arratia *et al.* (1990a) gives bounds on the Poisson approximation for the length of the longest match for arbitrary alphabet distribution.

5. Approximation by the Compound Poisson Distribution. We now obtain an approximation for the total number of q/k matches on a diagonal:

$$S=\sum_{\alpha=1}^{n-k+1} Y_\alpha.$$

First, we have shown that the total number of clumps on a diagonal is approximately equal in distribution to the Poisson random variable Z. Let $C_1, C_2, \ldots$ be independent copies of C, and define:

$$C^{*j}=C_1+C_2+\cdots+C_j \quad j=1,2,\ldots$$

If there are, say, j clumps on a diagonal, then S counts a total of m matches of quality q/k if $C_1+C_2+\cdots+C_j=m$, that is, if $C^{*j}=m$.

For $m=1, 2, \ldots$ the distribution of C^{*j} can be found from the distribution of C by the simple recursive relations:

$$P(C^{*j}=m)=\begin{cases} P(C=m) & j=1 \\ \sum_{i=1}^{m-1} P(C^{*(j-1)}=m-i)P(C=i) & j\geq 2. \end{cases} \tag{13}$$

With $P(C=m)=q_m$, the above recursion yields the solution:

$$P(C^{*j}=m)=\sum_{i_1+i_2+\cdots+i_j=m} \prod_{k=1}^{j} q_{i_k}. \tag{14}$$

Since there are about Z clumps, we see that the distribution of S is close to the distribution of:

$$\hat{S}=\sum_{i=1}^{Z} C_i, \tag{15}$$

where Z, C_1, $C_2, \ldots$ are independent. The distribution of $\hat{S}$ is called a compound Poisson distribution, or more precisely, the Poisson distribution compounded by the distribution of C.

By conditioning on $Z=j$, we find that for $m\neq 0$:

$$P(\hat{S}=m)=\sum_{j=1}^{m} P(\hat{S}=m|Z=j)P(Z=j),$$

or

$$P(\hat{S}=m)=\begin{cases} P(Z=m) & \text{if } m=1 \\ \sum_{j=1}^{m} P(C^{*j}=m)P(Z=j) & \text{if } m\geq 1. \end{cases} \tag{16}$$

By equation (15), the probability generating function $\phi_S(x)=E[x^S]$ is obtained by composing the generating function of the Poisson with that of C:

$$\phi_S(x)=\phi_Z(\phi_C(x))=\exp\{\lambda(\phi_C(x)-1)\}. \tag{17}$$

One may approximate the significance $P(S>a)$ by $P(\hat{S}>a)$; bounds for the quality of this approximation using Theorem 2 are given in Theorem 8. For small a the significance probability may be approximated using equation (16) as:

$$P(\hat{S}>a)=1-\sum_{m=0}^{[a]} P(\hat{S}=m). \tag{18}$$

For the case of large a one may apply the Chernoff bound (see Billingsley, 1986) to obtain:

$$P(\hat{S}>a)\leq \min_{y\geq 1} y^{-a}\,\phi_S(y). \tag{19}$$

We now apply Theorem 2 to obtain an error bound when approximating the distribution of S by $\hat{S}$. Recall that Theorem 2 is the process version of Poisson approximation for dependent Bernoulli random variables.

THEOREM 8. *The total variation distance between* S *and* $\hat{\mathrm{S}}$ *satisfies:*

$$\tfrac{1}{2}\|\mathrm{S}-\hat{\mathrm{S}}\|\leq(2\mathrm{b}_1+2\mathrm{b}_2)+2\lambda\mathrm{P}(\mathrm{C}>\mathrm{k}),$$

where b_1 *and* b_2 *are given in equation* (2) *for:*

$$\mathrm{B}_\alpha=\{\beta:|\beta-\alpha|\leq 2(\mathrm{k}-1)+\mathrm{s}\}.$$

Proof. We follow Arratia *et al.* (1990a). For $i=1, 2, \ldots, k-1$ let:

$$X_{\alpha,i}=(1-Y_{\alpha+i})Y_{\alpha+i-1}\cdots Y_\alpha \prod_{j=1}^{s}(1-Y_{\alpha-j}),$$

and

$$X_{\alpha,k}=Y_{\alpha+k-1}\cdots Y_\alpha \prod_{j=1}^{s}(1-Y_{\alpha-j}).$$

We have expanded the index set so that we may keep track of the *type* or size of the clump occurring at α. Here, a clump is of type i, for $i<k$, if it is of size i, and of type k if it is of size k or greater. It is clear that:

$$X_\alpha=\sum_{i=1}^{k} X_{\alpha,i}. \tag{20}$$

Furthermore, since for $i=1, 2, \ldots, k$, $X_{\alpha,i}$ is a function of the independent Bernoulli random variables $\{D_{\alpha-s}, \ldots, D_{\alpha+2k-2}\}$, we see that by taking:

$$B_\alpha=\{\beta:|\beta-\alpha|\leq 2(k-1)+s\}$$

condition (1) is satisfied. Using the partition structure (20) as in Arratia *et al.* (1990a) we have that b_1 and b_2 of Theorem 2 may be calculated from

equation (2) with the unpartitioned random variables, that is, with $p_\alpha = EX_\alpha$ and $p_{\alpha\beta} = EX_\alpha X_\beta$. For a subset $A \subset \{0, 1, \ldots\}$, define a function h on $\mathbf{X}$ and $\mathbf{Z}$ by:

$$h(\mathbf{X}) = \mathbf{1}_A\left(\sum_{i=1}^{k} iX_{\alpha,i}\right) \quad h(\mathbf{Z}) = \mathbf{1}_A\left(\sum_{i=1}^{k} iZ_{\alpha,i}\right).$$

An application of theorem 2 now yields that the total variation distance between $h(\mathbf{X})$ and $h(\mathbf{Z})$ is at most $2(2b_1 + 2b_2)$. Now note that $S = h(\mathbf{X})$ and $\hat{S} = h(\mathbf{Z})$ except when there is a clump of size of more than k. For $\mathbf{X}$, the probability that there exists such a clump may be bounded using the Bonferroni inequality, with $\beta = \beta_\alpha$ as in equation (3):

$$P\left(\exists\alpha\colon X_\alpha = 1, \sum_{\gamma=\alpha}^{\beta} Y_\gamma \geqslant k\right) \leqslant \sum_\alpha P\left(X_\alpha = 1, \sum_{\gamma=\alpha}^{\beta} Y_\gamma \geqslant k\right)$$

$$= \sum_\alpha P\left(\sum_{\gamma=\alpha}^{\beta} Y_\gamma \geqslant k \,\middle|\, X_\alpha = 1\right) P(X_\alpha = 1) = P(C > k) \sum_\alpha p_\alpha$$

$$= \lambda P(C > k).$$

By construction $P(X_\alpha = 1) = P(Z_\alpha = 1)$ and a similar computation for $\mathbf{Z}$ completes the proof. ■

Estimates for the values of λ, b_1, and b_2 may be obtained as in the previous section, taking into account that B_α has been enlarged.

5.1. Perfect matching: q = k, s = 1. As noted above, for the case of perfect matching with $q = k$ and $s = 1$ the distribution of the clump size C is geometric, that is:

$$P(C = m) = (1-p)p^{m-1} \quad m = 1, 2, \ldots.$$

Using equation (14) we have:

$$P(C^{*j} = m) = \binom{m-1}{j-1}(1-p)^j p^{m-j} \quad m = j, j+1, \ldots,$$

and that therefore, by equation (16):

$$P(\hat{S} = m) = \begin{cases} e^{-\lambda} & \text{if } m = 0 \\ e^{-\lambda}\displaystyle\sum_{j=1}^{m} \binom{m-1}{j-1} \frac{(\lambda(1-p))^j p^{m-j}}{j!} & \text{otherwise.} \end{cases} \tag{21}$$

This equation that approximates the distribution for perfect matching appears in Karlin and Ost (1987), Theorem 2.4.

We establish a corresponding Chernoff bound for the tail of this distribution, and obtain an upper bound on the total variation distance of the Poisson approximation. To obtain the Chernoff bound (19), note that since:

$$\phi_C(x) = \frac{x - px}{1 - px},$$

by relation (17):

$$\phi_S(x) = \exp\left\{-\lambda \frac{1-x}{1-px}\right\}.$$

Equation (19) now yields:

$$P(\hat{S} > a) \leqslant y^{-a}\phi_S(y)$$

where $y = \max(1, x)$ and:

$$x = \frac{\lambda(1-p) + 2ap - \sqrt{\lambda^2(1-p)^2 + 4ap\lambda(1-p)}}{2ap^2}.$$

We can obtain an upper bound to the total variation distance between S and $\hat{S}$ by applying Theorem 8; this yields the following.

COROLLARY 9. *For any set* A:

$$|\mathrm{P}(\mathrm{S} \in \mathrm{A}) - \mathrm{P}(\hat{\mathrm{S}} \in \mathrm{A})| \leqslant 4\lambda^2(4\mathrm{k}-1)/(\mathrm{n}-\mathrm{k}) + 5\lambda\mathrm{p}^{\mathrm{k}} \tag{22}$$

where the value of λ *is given in equation* (11).

Proof. Let B_α be as in Theorem 8. By Theorem 8, we need only calculate b_1, b_2 and $P(C > k)$. The latter equals p^k. To calculate b_1, break up the sum $\sum_\alpha \sum_{\beta \in B_\alpha} p_\alpha p_\beta$ into two parts, depending on whether or not p_1 appears. This yields the bound:

$$b_1 < \lambda^2(4k-1)/(n-k) + 2\lambda p^k.$$

In order to bound b_2 note that if $|\alpha - \beta| \leqslant k$, $\alpha \neq \beta$ then $X_\alpha X_\beta = 0$, since at least one of the indices is not 1, that clump of perfect matches begins with a mismatch. For $|\alpha - \beta| > k$, X_α is independent of X_β is independent of X_β and so $EX_\alpha X_\beta = EX_\alpha EX_\beta$. Hence $b_2 < b_1$. ■

Hence, when using probabilities computed from the distribution of $\hat{S}$, one makes an error of no more than the above when approximating the probability

of the same event for S. It will be seen below that this bound is conservative and the approximation performs especially well in the region of interest.

6. Maximum Clump Size Distributions. We now consider the distribution of the number of q/k matches in the longest clump. Again, let $W=\sum_\alpha X_\alpha$. For each α such that $X_\alpha=1$, we have a total of:

$$C_\alpha=\sum_{\gamma=\alpha}^{\beta} Y_\alpha$$

q/k matches, where $\beta=\beta_\alpha$ is as in definition (3). Let:

$$M_{(1)}\geqslant M_{(2)}\geqslant\cdots\geqslant M_{(W)}$$

be the order statistics of the C_α for the W indices α such that $X_\alpha=1$. In particular:

$$M_{(1)}=\max\{C_\alpha: X_\alpha=1\}.$$

In the case $q=k$, $M_{(1)}$ corresponds to the length of the longest exact matching word. While Mott *et al.* (1990) derive heuristic recursion formulas for the probability of long exact matching words, in the last few years there has been a good deal of work that provides computable approximations with error bounds. See Arratia *et al.* (1900a) and Arratia *et al.* (1990b). It is remarkable that the distribution of *all* the order statistics can be approximated by these Poisson methods.

Again, since W is approximately Poisson, we may approximate the above collection by:

$$\tilde{M}_{(1)}\geqslant\tilde{M}_{(2)}\geqslant\cdots\geqslant\tilde{M}_{(Z)},$$

the order statistics of:

$$\{C_j: j=1,2,\ldots,Z\},$$

where $Z, C_1, C_2, \ldots$ are independent, Z is Poisson λ and C_j have distribution C. By the usual, elementary independent Bernoulli thinning of a Poisson argument, we have that the sum:

$$U=\sum_{j=1}^{Z}\mathbf{1}\{C_j>x\}$$

is Poisson with parameter $\lambda P(C>x)$. By the equivalence:

$$\{\tilde{M}_{(j)}\leqslant x\}=\{U\leqslant j-1\},$$

we immediately derive:

$$P(\tilde{M}_{(j)}\leqslant x\}=e^{-\lambda P(C>x)}\sum_{i=0}^{j-1}\frac{(\lambda P(C>x))^i}{i!}.$$

In particular, note that when $s=1$ and $q=k$ we have for x a non-negative integer $P(C\geqslant x)=p^x$ and therefore $\tilde{M}_{(1)}$ has an integerized extreme value distribution.

We now obtain error bounds. In the notation of the last section, for $x<k$ we have the equivalence between events:

$$\{M_{(j)}\leqslant x\}=\left\{\sum_{\alpha\in I,i>x}X_{\alpha,i}\leqslant j-1\right\}. \tag{23}$$

Hence, setting:

$$h(\mathbf{X})=\mathbf{1}_A\left(\sum_{\alpha\in I,i>x}X_{\alpha,i}\right),\quad h(\mathbf{Z})=\mathbf{1}_A\left(\sum_{\alpha\in I,i>x}Z_{\alpha,i}\right)$$

and arguing as in the proof of Theorem 8 using the process version of Poisson approximation we obtain the following

THEOREM 10. *The total variation distance between* $\mathrm{M}_{(j)}$ *and* $\tilde{\mathrm{M}}_{(j)}$ *satisfies:*

$$\tfrac{1}{2}\left\|\mathrm{M}_{(j)}-\tilde{\mathrm{M}}_{(j)}\right\|\leqslant(2\mathrm{b}_1+2\mathrm{b}_2)+2\lambda\mathrm{P}(\mathrm{C}>\mathrm{k}),$$

where b_1 *and* b_2 *are given in equation* (2) *for:*

$$\mathrm{B}_\alpha=\{\beta:\left|\beta-\alpha\right|\leqslant 2(\mathrm{k}-1)+\mathrm{s}\}.$$

7. Data Analysis. Even though DNA sequences can be better described by a second order Markov chain than a sequence of independent letters (Tavaré and Giddings, 1989), the independence assumption for the purposes of deriving the distribution of scores fits the data quite well (see Smith *et al.*, 1985, for a study of DNA sequences from GenBank). Our theoretical results are based on quality q/k matches with match probability p. DNA sequences often have $p\approx 0.25$ while protein sequences often have p a little larger than 0.05. Therefore, results about the number of matches depend on q/k, n, and p, as well as independence. Not surprisingly then, we find that an approximation to the theoretical distribution computed under the assumption of independence sufficiently describes the null distribution for quality q/k matches.

To test the fit of an approximate theoretical distribution against an empirical distribution given by matching unrelated protein sequences, we have studied the distribution of S, the total number of q/k matches on a diagonal. 123 protein sequences from vertebrates were matched against 104 from *E. coli*, yielding 12 792 pairwise comparisons. Here $q=k=2$ was chosen. Only the first $n=200$ letters of each sequence were considered in order to make all sequences the same length so that one could make a comparison with the theoretical distribution. For the given collection of sequences $p=0.0589$, only slightly larger than what would be the matching probability of 1/20 for the uniform distribution over the 20 amino acids. For the matches of length $k=2$ considered, equation (11) yields $\lambda=0.6499$. Comparison of the two distributions is given in Fig. 1, and is seen to be quite close. A graph of the difference between the distribution of the actual data and what is predicted by equation (21) is plotted in Fig. 2. All differences are well within statistical fluctuation and the nonrandom error bound given by equation (22) above. For example, the largest difference of 0.01756 between the two distributions occurs for the case of no quality q/k matches at all, and has a standard error of 0.004. Equation (22) yields that differences in probabilities between the true distribution of S and the distribution 21 in this instance are no more than 0.0710. It is clear from Fig. 2 that this bound is conservative and well accounts for differences between the two distributions for values close to zero.

For an application of our methods to related sequences, we consider two protein sequences, both human leukocyte class II histocompatibility antigens (HLA), one an SB alphachain precursor of length 261, the other a DR-1 betachain precursor of length 266, with PIR accession numbers A24283 and A24431, respectively. The dot matrix for matches with $q=k=2$ is shown in Fig. 3. For these sequences, $p=0.053$. We of course look at each of the diagonals, as is appropriate in biological sequence comparison. Although not striking visually, one diagonal of length $n=259$ contains a total of $S=11$ $q=k=2$ matches. For the given values of n, p, and k, equation (11) yields $\lambda=0.68$; calculating significance according to equations (21) and (18) we obtain an approximate p-value of 3.2×10^{-8}. This p-value only takes into account the length $n=259$ diagonal. When we consider all diagonals, we note that a diagonal sum of 11 or larger is the union of the events E_i that diagonal i has a sum of 11 or larger. The most elementary Bonferroni inequality is:

$$P\left(\bigcup_i E_i\right) \leqslant \sum_i P(E_i),$$

and this is our approach to finding a p-value for the two sequence comparisons. Using the same method for all diagonals i, we obtain the overall significance estimate $\sum_i P(E_i) \approx 2.5\times10^{-6}$. This does not appear to us to be an optimal way

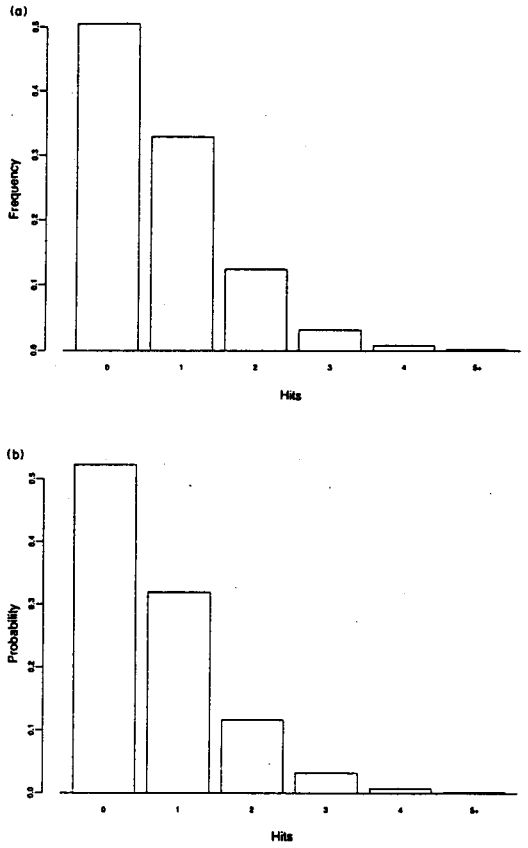

Figure 1. Comparison of protein sequence scores with the theoretical distribution from equation (3). (a) A: histogram for 13 792 vertebrate–*E. Coli* comparisons. (b) The values of $P(\hat{S}=m)$ for comparison.

to combine the statistical tests on each diagonal. $S=11$ on a diagonal of length 20 is a very different result for $S=11$ on a diagonal of length 259, and this should be taken into account. One way of finding an overall significance level is:

$$\prod_i \hat{p}_i = \prod_i P(S \geqslant s_i)$$

where the product is overall diagonals i, and s_i is the observed diagonal sum for diagonal i. The method of combining p-values implicitly used above is min $\hat{p}_i$

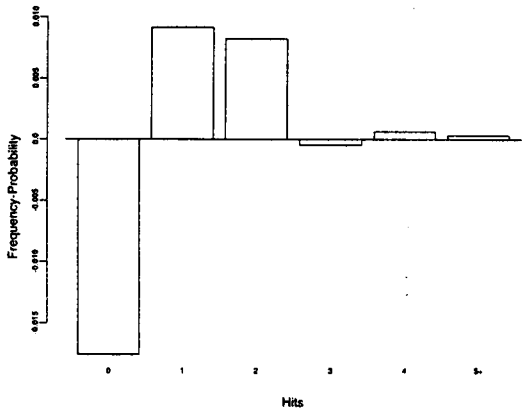

Figure 2. The values of data-theory from Fig. 1.

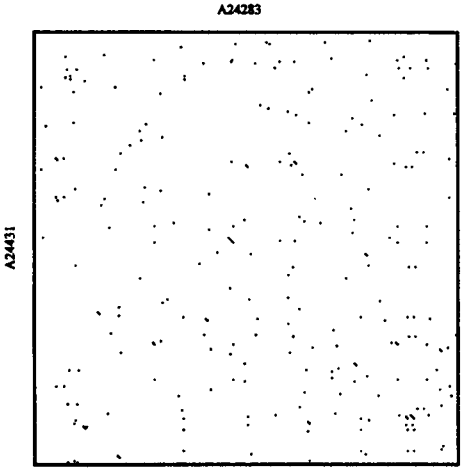

Figure 3. Dot matrix for the comparison of two human leukocyte class II histocompatibility antigens, with $q=k=2$.

not $\prod_i \hat{p}_i$. See Oosterhoff (1969) for a monograph on combination of one-sided tests.

To apply the results for the case $q \neq k$, we consider $q=3$ and $k=4$. Recall that $p=0.053$ for our two HLA proteins. We must simulate to obtain the value $s=2$

that gives the best fit to a Poisson number of clumps where $\lambda \approx 0.1$. For $q=3$, $k=4$, $s=2$ we use equation (9) from Section 3 to obtain:

$$P(C=m) \approx q_m.$$

Then, the convolution formula (16) yields the approximation $P(\hat{S}=m)$. Finally the p-values for diagonal sums are arrived at with equation (18). The most significant diagonal is again the diagonal of length 259 with a total of $S=9$ matches. The approximate p-value for the diagonal is 5×10^{-5}.

Next we consider the maximum clump size $M_{(1)}$ distribution. Since clump size is not restricted to a diagonal as is diagonal sum, we do not have the delicate difficulties of combination of one-sided tests discussed above. Instead we view the problem as having one long diagonal with $n=(261)(266)$, the product of sequence lengths. The $q=k=2$ case has maximum clump size $M_{(1)}=4$ and $\lambda=185$ from equation (11). Also $P(C \geqslant 4)=0.000148$. Therefore the maximum clump p-value is estimated by:

$$\begin{aligned} P(M_{(1)} \geqslant 4) &\approx P(\tilde{M}_{(1)} \geqslant 4) \\ &= 1 - e^{-\lambda P(C \geqslant 4)} \\ &= 0.027. \end{aligned}$$

For the $q=3$, $k=4$ case we have maximum clump $M_{(1)}=5$ and using the approximation for λ given in Section 4.1, $\lambda=27.7$. Also $P(C \geqslant 5)=0.0132$ so that the p-value is estimated by:

$$\begin{aligned} P(M_{(1)} \geqslant 5) &\approx 1 - e^{-\lambda P(C \geqslant 5)} \\ &= 0.3. \end{aligned}$$

Clearly the maximum clump size of 4 in the $q=2$ search could be considered significant, while the value of 5 in the $q=3$, $k=4$ search could not. In either case, the diagonal sums are more significant than the individual clump sizes.

We also compared these two sequences using the more sensitive local dynamic programming algorithm of Smith and Waterman (1981). For this algorithm, we weighted gaps of length k by $w(k)=5+3k$. The weight matrix T of Taylor (1986) and Waterman and Jones (1990) was used for weighting pairs of amino acids. This matrix counts the number of shared properties between two amino acids and has entries ranging from 0 to 8. Figure 4a shows the alignment obtained by using the matrix with (i,j)th entry $T_{i,j}-4.5$ for weighting pairs of amino acids. Notice that the alignment contains a 122 amino acid segment of the diagonal found by our statistical method. Relaxing the weighting by using a matrix with entries $T_{i,j}-4.0$ gives the longer alignment of Fig. 4b which, although it has three length 1 gaps, is also principally comprised of the same diagonal.

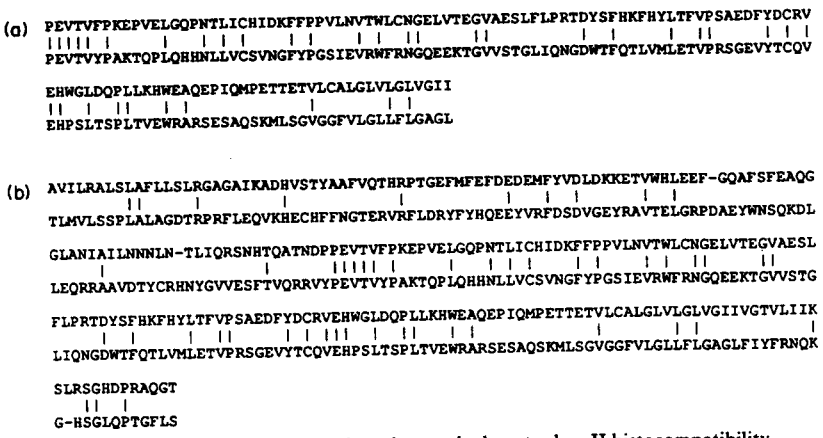

Figure 4. Best local alignments of two human leukocyte class II histocompatibility antigens with gap penalty function $w(k)=5+3k$. (a) Alignment with $(T_{ij}-4.5)$. (b) Alignment with $(T_{ij}-4.0)$.

8. Discussion: Open Problems. In this paper we have given computable, closed form approximations to random variables of interest for comparing two independent sequences by studying quality q/k k-word matches on a diagonal. The proofs of our results would not be as manageable without the Poisson approximation, Theorems 1 and 2. There are of course a number of extensions and problems that remain unsolved. Next we discuss several groups of such open problems.

8.1. All diagonals. When we perform sequence comparison, we do not study a diagonal in isolation. Of course the dependence between diagonals greatly complicates the analysis. In Arratia *et al.* (1986) and Arratia *et al.* (1990b), related quite technical analyses are carried out for the longest match with k mismatches and the longest quality a match. Their results indicate that the correct approximation in our case for the largest clump could be obtained by replacing the diagonal length n by the product of sequence lengths $(n_1-t+1)(n_2-t+1)$, where t is the test matching length. That is, the comparison is treated as if there is one long diagonal of length $(n_1-t+1)(n_2-t+1)$. Of course the error bounds will be different, and the proof much harder. When extremal properties of diagonals, such as the diagonal sum S, are studied, the simple Bonferroni approach of Section 7 can be used, but more work on combination of statistical tests would be valuable.

8.2. Unequal sequence lengths and distributions. All sequences are not created equal, usually they are not of identical length nor have identical distributions. Still the conjectures described in Section 8.1 should hold when the differences are not too great. In Arratia *et al.* (1986), the effective diagonal length is shown to be n_1n_2 if $1 \geqslant \log(n_1)/\log(n_2) \to 1 > 0$. A more general sufficient condition is given in Arratia *et al.* (1990a).

8.3. Non-independence. DNA and protein sequences are not well modeled by independent letters. However, we have described in Section 7 an example where the scores from biological data are well approximated by independence. There may well be deep reasons for these observations (Haiman, 1987). Nonetheless it is of interest to study these problems when the sequences are Markov or m-dependent.

8.4. Insertions and deletions. Of course real biological sequence evolution often involves insertion or deletion of sequence letters. This greatly increases the difficulty. The only general results are those of Waterman *et al.* (1987) where some results on growth rates of Smith–Waterman scores are derived.

8.5. Relative error. The Poisson approximation theorems we use give error bounds in terms of total variation distance. It would be useful and informative to have results for relative error. If p is the correct tail probability, such as $p=P(S>m)$, and $\hat{p}$ is the probability from our formula, $|p-\hat{p}|/p$ is the error in $\hat{p}$ relative to p. This is another area for future research.

The authors are grateful to Tim Hunkapillar for suggesting the HLA protein sequences for comparison. They also wish to thank Mark Eggert; his help on this project was especially critical.

LITERATURE

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman. 1990. Basic local alignment search tool. *J. molec. Biol.* **214**, 1–8.

Aldous, D. J. 1989. *Probability Approximations via the Poisson Clumping Heuristic*. New York: Springer-Verlag.

Arratia, R., L. Gordon and M. S. Waterman. 1986. An extreme value theory for sequence matching. *Ann. Statist.* **14**, 971–993.

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

Arratia, R. and L. Gordon. 1989. Tutorial on large deviations for the binomial distribution. *Bull. math. Biol.* **51**, 125–131.

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

Arratia, R. and M. S. Waterman. 1989. The Erdös-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann Prob.* **3**, 1152–1169.

Arratia, R., L. Goldstein and L. Gordon. 1990a. Poisson approximation and the Chen-Stein method. *Stat. Sci.* **5**, 403–423.

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

Billingsley, P. 1986. *Probability and Measure*. New York: John Wiley and Sons.

Feller, W. 1968. *An Introduction to Probability Theory and its Applications*, Vol. I, 3rd Edn. New York: John Wiley and Sons.

Goldstein, L. 1990. Poisson approximation and DNA sequence matching. *Communs Stat. Theory Meth.* **19**, 4167–4179.

Haiman, G. 1987. Étude des extrêmes d'une suite stationnaire *m*-dépendante avec une application relative aux accroissements du processus de Wiener. *Ann. Inst. Henri Poincaré*, **23**, 425–258.

Karlin, S., G. Ghandour, F. Ost, S. Tavare and L. J. Korn. 1983. New approaches for computer analysis of nucleic acid sequences. *Proc. natn. Acad. Sci. U.S.A.* **80**, 5660–5664.

Karlin, S. and F. Ost. 1987. Counts of long aligned word matches among random letter sequences. *Adv. appl. Prob.* **19**, 293–351.

Karlin, S. and S. F. Altschul. 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.

Lipman, D. J. and W. R. Pearson. 1985. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.

Mott, R. F., T. B. L. Kirkwood and R. N. Curnow. 1989. A test for the statistical significance of DNA sequence similarities for application in databank searches. *CABIOS* **5**, 123–131.

Mott, R. F., T. B. L. Kirkwood and R. N. Curnow. 1990. An accurate approximation to the distribution of the length of the longest matching word between two random DNA sequences. *Bull. math. Biol.* **6**, 773–784.

Needleman, S. B. and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. molec. Biol.* **48**, 444–453.

Oosterhoff, J. 1969. *Combination of one-sided statistical tests*. Mathematical Centre Tracts, No. 28, Mathematical Centre, Amsterdam, The Netherlands.

Pearson, W. R. and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. natn. Acad. Sci. U.S.A.* **85**, 2444–2448.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. In: *Methods in Enzymology*, Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences. R. F. Doolittle (Ed.). Vol. 183, pp. 63–98. New York: Academic Press.

Smith, T. F. and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. molec. Biol.* **147**, 195–197.

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

Tavaré, S. and B. Giddings. 1989. Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences*, M. S. Waterman (Ed.). Florida, U.S.A.: CRC Press.

Taylor, W. R. 1986. The classification of amino acid conservation. *J. theor. Biol.* **119**, 205–218.

Waterman, M. S. 1984. General methods of sequence comparions. *Bull. math. Biol.* **46**, 473–500.

Waterman, M. S., L. Gordon and R. Arratia. 1987. Phase transitions in sequence matches and nucleic structure. *Proc. natn. Acad. Sci.* **84**, 1239–1243.

Waterman, M. S. 1989. *Mathematical Methods for DNA Sequences*. M. S. Waterman (Ed.). Florida, U.S.A.: CRC Press.

Waterman, M. and R. Jones. 1990. Consensus methods for DNA and protein sequence alignments. In *Methods in Enzymology*, Vol. 183, R. Doolittle (Ed.). New York: Academic Press.

Wilbur, W. J. and D. Lipman. 1983. Rapid similarity searches of nucleic acid and protein databanks. *Proc. natn. Acad. Sci. U.S.A.* **80**, 726–730.

Received 3 March 1991