Bulletin of Mathematical Biology, Vol. 59, No. 3, pp. 483–496, 1997
Elsevier Science Inc.

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00096-1

MATCHING AMONG MULTIPLE RANDOM SEQUENCES

■ JOSEPH I. NAUS
Department of Statistics,
Rutgers, The State University,
Piscataway, NJ 08855, U.S.A.

(*E.mail: naus@rci.rutgers.edu*)

■ KE-NING SHENG
Roberts Pharmaceutical Corporation,
Eatontown, NJ 07724, U.S.A.

In searching for strong homologies between multiple nucleic acid or protein sequences, researchers commonly look at fixed-length segments in common to the sequences. Such homologies form the foundation of segment-based algorithms for multiple alignment of protein sequences. The researcher uses settings of "unusualness of multiple matches" to calibrate the algorithms. In applications where a researcher has found a multiple matching word, statistical significance helps gauge the unusualness of the observed match. Previous approximations for the unusualness of multiple matches are based on large sample theory, and are sometimes quite inaccurate. Section 2 illustrates this inaccuracy, and provides accurate approximations for the probability of a common word in R out of R sequences. Section 3 generalizes the approximation to multiple matching in R out of S sequences. Section 4 describes a more complex approximation that incorporates exact probabilities and yields excellent accuracy; this approximation is useful for checking the simpler approximations over a range of values. © 1997 Society for Mathematical Biology

1. Introduction. The unusualness of matching common words in multiple sequences is important in studying homologies between protein and nucleotide sequences, and plays a role in a variety of multiple alignment algorithms. Sobel and Martinez (1986) describe their multiple alignment program which uses an algorithm to locate common segments among multiple sequences. The significance levels are approximated by scrambling sequences. The computational algorithm of Waterman *et al.* (1984) is based on looking for common words or subsequences. Waterman (1986) discusses the multiple sequence alignment algorithm based on matching "consensus" words of user specified length (and allowed mismatches). He looks for a common word of length k. (To restrict the amount of shifting and reduce the comparisons, he focuses within a window of width W.) He notes the importance of evaluating statistical significance and describes a direct approach. Karlin and Ost (1987, 1988) develop very general asymptotic

results for such matching, and we will compare the approximations based on their results with new approximations developed in this paper.

Leung *et al.* (1991) survey a variety of methods for finding similar segments in multiple sequences. They develop a computational algorithm for finding multiple matches or repeats. They note: "The method achieves great speed by requiring the segments of all reported alignments to share a 'core block' of identical letters exceeding some minimum length. It is recommended that this 'core' parameter be chosen with reference to the statistical properties of maximal length common words among random letter sequences." For this purpose, they refer to Karlin and Ost (1988).

We seek to evaluate the accuracy of existing asymptotic formulae that are used as approximations for the probability of a common word appearing in multiple protein sequences. For the case of two sequences and the probability of a common word in the same positions in the two aligned sequences, highly accurate approximations, exact results and tight bounds are known, both for the case of perfect or imperfect matches; see Glaz and Naus (1991).

For the case of a common word anywhere in two sequences, Mott *et al.* (1990) and Sheng and Naus (1994) give highly accurate approximations. Even the asymptotic formula of Karlin and Ost (1988), usually gives excellent approximations for two sequences, though care must be taken to use the right modification of the asymptotic formula (see equation (4) below). For the case of a common word in the same position in multiple aligned sequences accurate approximations (similar to those for two aligned sequences) as well as asymptotic results (Karlin and Ost, 1987) exist.

For the case of the probability of a word in common to (but in any position in) multiple sequences, Karlin and Ost (1988) give an asymptotic approximation. One might suspect, based on the excellence of the modified asymptotic result for two sequences, that this approximation would give excellent accuracy for more than two sequences. Our results indicate that the accuracy of the existing asymptotic approximations deteriorates as the number of sequences increase.

2. Comparison of Asymptotic Formula for Probability of a Word in Common to Multiple Sequences. Given S independent sequences, each of T letters, we seek the probability $P_{R,S}(w,T)$ that there exists a word of length w that is common to at least R of the S sequences. Karlin and Ost (1988) derive asymptotic results for $P_{R,S}(w,T)$ for the case where each sequence consists of i.i.d. letters, as well as the more general case of dependent sequences of letters. For simplicity in our comparisons in this section, we will begin with the case of independent and identically distributed sequences of i.i.d. letters: Sequence 1: $X_{11}, X_{12}, \ldots, X_{1T}$; Sequence 2: $X_{21}, X_{22}, \ldots, X_{2T}; \ldots$; Sequence S: $X_{S1}, X_{S2}, \ldots, X_{ST}$, where all the X_{ij} are

mutually independent and identically distributed with $P(X_{ij} = k) = p_k$, for $k = 1, 2, \ldots, m$, where m is the number of different letters in the alphabet.

Let $\lambda = \sum_{k=1}^{m} p_k^R$. For the special case of an equally likely m-letter alphabet, $p_k = 1/m$, and $\lambda = (1/m)^{R-1}$. Karlin and Ost's (1988) asymptotic result for the case of independent and identically distributed sequences of i.i.d. letters states that for T sufficiently large, and w of order $R \log T / (-\log \lambda)$ that

$$P_{R,S}(w, T) \sim 1 - \exp\{-(1 - \lambda) T^R C_R^S \lambda^w\} \tag{1}$$

where $C_R^S = S!/R!(S - R)!$. More precisely, $w = [\{R \log T/(-\log \lambda)\} + 1 + x]$, where $[y]$ denotes the largest integer less than or equal to y.

In this section, we will be making comparisons and interpretations for the special case $S = R$, where the common word of length w appears in all S sequences. For this case, Karlin and Ost's (1988) asymptotic result reduces to

$$P_{R,R}(w, T) \sim 1 - \exp\{-(1 - \lambda) T^R \lambda^w\}. \tag{2}$$

The asymptotic results (1) and (2) are based on mathematically rigorous conditions, which assure the result in the limit. Whether these results provide useful approximations depends on the rate of convergence to the limit, together with the range of sequence lengths and word lengths of interest to biologists.

Karlin *et al.* (1985) state that the asymptotic result (2) is an excellent approximation for the probability of the longest matching word. (See their comment on p. 37 in reference to their equation (1) which, for the independence model, reduces to result (2).) In their paper, they focus on the longest words in common to three immunoglobulin sequences, and use high significance levels. For this case of three long sequences and highly unusual long matches, approximation (2) is reasonable, although (4) below is better. For more or shorter sequences approximation, (2) is less accurate. The discussion that follows is from the perspective of developing approximations and assessing their accuracy for a practical range of word lengths and sequence lengths.

Consider first the situation where the R sequences are aligned one above the other. λ is the probability that the first letter in each of the R sequences is the same. Consider the R words consisting of the first w letters in each of the R sequences; λ^w is the probability that the R words are the same word. There are approximately T^R ways that we can pick the starting letter of each word in the R sequences, and the expected number of these ways where all R words match is approximately $T^R \lambda^w$. For small values of $P_{R,R}(w, T)$, the probability that any particular set of starting

positions for the R words (one from each sequence) leads to a common R-way match is very small, and this leads via a Poisson-type approximation to

$$P_{R,R}(w,T) \sim 1 - \exp\{-T^R\lambda^w\}. \tag{3}$$

Note that when an R-sequence $w+1$ letter match occurs, this is double counted as two w-letter word matches. Subtracting $T^R\lambda^{w+1}$ from the expectation used in (3) gives (2). Equivalently, we can view the declumping as requiring the R-way w-letter word matches to also satisfy that at least one of the letters preceding the R words differs. Either view gives the $(1-\lambda)$ term in (2), which handles one type of dependence in the matches. A somewhat better approximation is gotten by taking $(T-w+1)^R$ as the number of possible starting positions for the R words (one from each sequence). In a similar way, a modified Karlin and Ost's approximation is

$$P_{R,R}(w,T) \sim 1 - \exp\{-(1-\lambda)(T-w+1)^R\lambda^w\}. \tag{4}$$

While approximation (4) works better than (2), and for the case of $R=2$ sequences is an excellent approximation, it appears to deteriorate for even a moderate number of sequences. We note there is a strong dependence between the multiple matching in different combinations of positions. For example, consider $R=9$ sequences with $w=4$. Consider a match for two of the sets of starting positions considered in $(T-w+1)^R$ or (T^R). In set one, let $X_{1j}=X_{2j}=\cdots=X_{9j}$, for $j=1,2,3,4$; the first four letters in each of the nine sequences match. In set two, let $X_{1j}=X_{2j}=\cdots=X_{8j}=X_{9,j+6}$, for $j=1,2,3,4$. These two sets of matches are highly dependent. We conjecture that this type of dependence has more impact on the inaccuracy of the approximations (2)–(4) as the number of sequences increases.

Our approach mitigates this type of dependence by starting with a word in a given starting position in the first sequence, and approximating the probability that that word appears anywhere within each of the other $R-1$ sequences. The same is done for each other word in the $(T-w+1)$ other starting positions in the first sequence. The approximation is refined further to take into account that not all of the $(T-w+1)$ words are expected to be distinct. There still remains dependence because of the overlapping nature of words within the first sequence. Section 4 describes a more complex approximation to handle some of this dependence. However, based on comparisons with both the more complex approximation and simulation studies, the easy to compute approximation (8) appears quite accurate.

2.1. *Our simpler approximation.* Consider the first w letters of sequence 1 as a random word. Let R_w denote the probability that this random word

appears somewhere in each of the other $R-1$ sequences. There are $(T-w+1)$ words in sequence 1. The probability that any particular one of them appears in all $R-1$ other sequences is R_w. The expected number of such matches is $(T-w+1)R_w$. We can adjust somewhat for overlapping in counting sets of w-word matches by requiring that when a particular word in sequence 1 appears in all $R-1$ other sequences, the letter preceding the word in sequence 1 differs from the letter preceding the matching word in at least one of the $R-1$ other sequences. This has the effect of multiplying the expectation by $(1-p^{R-1}) = 1-\lambda$. The expected number of such decoupled matches is $(T-w+1)R_w(1-\lambda)$. In section 4, we discuss finding R_w exactly, and more complex approximations for $P_{R,R}(w,T)$. In the present section, we approximate R_w as follows. Let δ_i denote the probability that the random word of w letters (the first w letters in sequence 1) appears somewhere in sequence i. From the independence of the $R-1$ sequences, R_w is approximately $\prod_{i=2}^{R}\delta_i$, and for the i.i.d. sequences case, R_w is approximately δ^{R-1}. Approximately (as opposed to exactly) because even though the sequences are independent, and the occurrence of a fixed word in sequence i is independent of the occurrence of that word in sequence j, it is not true that the occurrence of a random word in sequence i is independent of a random word in sequence j. This is because certain word patterns which can occur in overlapping positions have somewhat different probabilities of occurring in a sequence than other patterns that cannot overlap. The more complex approximations and bounds of Section 4 take into account the types of patterns that can occur. In the present section, we take the approximation $R_w \approx \delta^{R-1}$, and use this in the approximations for $P_{R,R}(w,T)$. For the comparisons studied, this simplification did not lose much accuracy. We approximate the expected number of decoupled matches by $(T-w+1)\delta^{R-1}(1-\lambda)$. A Poisson-type approximation gives, for the probability that at least one of the $T-w+1$ words in sequence 1 appears in common to all R sequences,

$$P_{R,R}(w,T) \sim 1-\exp\{-(T-w+1)\delta^{R-1}(1-\lambda)\}. \tag{5}$$

Note that we can also approximate δ, the probability that the first w (random) letters of sequence 1 appear somewhere in sequence 2. There are $(T-w+1)$ starting positions for the matching word in sequence 2, and the probability that in at least one of these positions the two words match is

$$\delta \sim 1-\exp\{-(T-w+1)p^w\} \tag{6}$$

where $p=\sum_{k=1}^{m}p_k^2$. For the case of an equally likely m-letter alphabet, $p=1/m$.

By expanding the right-hand side of (6), we see that when $(T-w+1)p^w$ is sufficiently small, then it approximates δ. Substituting this into (5), and

noting that $(p^w)^{R-1} = \lambda^w$ shows a situation where approximation (5) with (6) has a similar limit as (4). One can compute and compare these and other approximations such as the refinement (8) to study when they differ and when they are similar. This, together with simulations, provides information on the accuracy of the approximations.

The approximation based on (5) and (6) gives quite good accuracy. Table 1 compares approximations (5) and (6) with a refined but simple to compute approximation (8) discussed below, with a more complex approximation (14) discussed in Section 4, with the Karlin and Ost (1988) approximation (2) and its modified form (4), and with simulations based on 100,000 trials. Table 2 gives additional comparisons for five-letter words. In Table 1 (Table 2), the lengths of the sequences are chosen so that the probability of a four-letter (five-letter) word in common to all sequences is as close to 0.01 and 0.05 as possible. For shorter sequences, the probabilities will be smaller.

The comparisons illustrate several points. We see that as the number of sequences increase, the Karlin and Ost (1988) approximation deteriorates even with the modified form (4). The new approximations represents a substantial improvement in accuracy over the Karlin and Ost (K and O) approximations, with the improvement most dramatic as the number of sequences increases. However, we also see from Table 1 that our approxi-

Table 1. Probability of a four-letter word in common to R out of R i.i.d. sequences each of T letters drawn from an equally likely four-letter alphabet

No. of seqs. (R)	Length of seqs. (T)	Probability of a word in common					
		K and O	K and O (4)	Our (5) and (6)	Our (8)	Our (14)	Simulation
2	5	0.071	0.012	0.012	0.012	0.015	0.015
3	12	0.024	0.010	0.010	0.010	0.010	0.010
4	24	0.019	0.011	0.010	0.010	0.010	0.010
5	39	0.021	0.014	0.011	0.010	0.010	0.010
6	55	0.025	0.018	0.011	0.010	0.010	0.011
7	71	0.032	0.024	0.011	0.010	0.010	0.010
8	87	0.045	0.034	0.011	0.010	0.010	0.010
9	103	0.068	0.053	0.012	0.010	0.010	0.010
10	118	0.105	0.082	0.012	0.010	0.010	0.010
2	7	0.134	0.046	0.045	0.045	0.052	0.052
3	19	0.093	0.057	0.054	0.052	0.052	0.051
4	36	0.094	0.067	0.056	0.053	0.052	0.053
5	54	0.101	0.077	0.053	0.048	0.048	0.047
6	74	0.139	0.110	0.057	0.050	0.051	0.050
7	93	0.192	0.156	0.059	0.050	0.051	0.051
8	111	0.274	0.227	0.060	0.049	0.050	0.051
9	128	0.393	0.332	0.060	0.048	0.049	0.050
10	145	0.581	0.506	0.063	0.049	0.051	0.050

Table 2. Probability of a five-letter word in common to R out of R i.i.d. sequences each of T letters drawn from an equally likely four-letter alphabet

No. of seqs. (R)	Length of seqs. (T)	Probability of a word in common				
		K and O	K and O-modif. (4)	Our (5) and (6)	Our (8)	Our (14)
2	7	0.035	0.007	0.009	0.009	0.008
3	27	0.017	0.011	0.011	0.011	0.011
4	63	0.014	0.011	0.010	0.010	0.010
5	112	0.016	0.013	0.011	0.010	0.010
6	166	0.018	0.016	0.011	0.010	0.010
7	223	0.024	0.021	0.011	0.010	0.010
8	280	0.031	0.028	0.011	0.010	0.010
9	337	0.045	0.041	0.012	0.010	0.010
10	392	0.067	0.061	0.012	0.010	0.010
2	12	0.100	0.046	0.060	0.060	0.049
3	43	0.069	0.052	0.053	0.052	0.049
4	95	0.072	0.061	0.054	0.052	0.051
5	156	0.080	0.071	0.054	0.050	0.050
6	222	0.101	0.091	0.055	0.050	0.050
7	289	0.136	0.124	0.057	0.050	0.050
8	354	0.189	0.174	0.058	0.049	0.050
9	418	0.275	0.256	0.060	0.050	0.050
10	479	0.402	0.376	0.062	0.050	0.050

mation based on (5) and (6) seems to overestimate somewhat the probability of a common match. The following modification of approximations (5) and (6) appears more accurate.

2.2. *Refining the approximation* (5) *and* (6). In developing equation (5), we note that there are about $(T-w+1)$ words in sequence 1. We then approximate the probability that at least one of these words appears in all the other $R-1$ sequences. However, this approach assumes that the $(T-w+1)$ words are all distinct. This would not be true if $(T-w+1) > 4^w$, the total possible number of w-letter words from a four-letter alphabet. This is not a problem for any of the examples in Table 1. Of greater importance, we do not expect that all the $(T-w+1)$ words in sequence 1 are distinct. Let $V = C^w$ denote the number of possible w-letter words from a C-letter alphabet. Given that $(T-w+1)$ words are picked at random (with replacement) from the V possible words, let T^* denote the expected number of distinct words:

$$T^* = V\left(1 - \{(V-1)/V\}^{T-w+1}\right). \tag{7}$$

This result is the expected number of V cells that are occupied when $(T-w+1)$ balls are distributed at random. The proof follows by letting $Z_i = 1$ iff cell i is occupied, 0 otherwise, $i = 1,2,\ldots,V$. The number of occupied cells is the sum of the Z_is. $E(Z_i) = P(Z_i = 1) = 1 - \{(V-1)/V\}^{T-w+1}$. Our modified approximation replaces $(T-w+1)$ by T^* in equation (5), but not in equation (6). The approximation is

$$P_{R,R}(w,T) \sim 1 - \exp\{-(T^*)\delta^{R-1}(1-\lambda)\} \tag{8}$$

where T^* is given by equation (7), and δ is given by (6).

Approximation (8) involves a two-stage Poisson approximation. The first stage approximates δ using equation (6). The second stage incorporates this value into equation (8) to find the match probability. We would expect the Poisson approximation for δ in the first stage to be reasonably accurate when $(T-w+1)$ is large and p^w is small. We would expect the second stage Poisson approximation (8) to be accurate when δ^{R-1} is small and T is large.

Tables 1 and 2 study the accuracy of the approximations for small probabilities (0.05, 0.01) of R-way matching. We anticipate and find that the Poisson approximations (8) and approximation (14) are accurate for this case. Table 3 studies the accuracy of the approximations for w and R fixed and letting T increase. This leads to higher values for the probabilities. While approximation (4) deteriorates as T increases (particularly for larger R), approximations (8) and (14) remain highly accurate.

Table 4 relates the size of unusual match word, number of sequences, and common sequence length for the case of a four-letter equally likely alphabet and independent sequences. For a given w and R, approximation (8) is used to find the sequence length T for which the match probability is 0.010. Bold terms indicate cases where approximations (8) and Karlin and Ost (4) are within 0.001. For $R = 2,3,4$, the approximations give close results, and similarly for $R = 5$ for $w \geqslant 8$.

We see from Tables 1 and 2 that for match probabilities less than 0.05, T does not have to be very large for the approximation to the match probability to be very accurate. For the case of very short sequences (for example $T = 5$, 7 and $w = 4$), approximation (8) loses accuracy. In these cases, R is typically small. If T is small, $R = 2$, one should use the more accurate approximations of Mott *et al.* (1990) or Sheng and Naus (1994). For T very small, and R small and greater than 2, one can use approximation (14) or even simulations. For all of the situations listed in Table 4, for $R \geqslant 3$, we anticipate approximation (8) to be accurate. For match probabilities of 0.05 (or greater; see Table 3), the sequences will be even larger (than for the cases for 0.01), and we anticipate approximation (8) to be accurate.

Table 4 illustrates how the sequence size grows with word size and number of sequences for a match probability of 0.01. For six or more

Table 3. Probability of a four-letter word in common to R out of R i.i.d. sequences each of T letters drawn from an equally likely four-letter alphabet

No. of seqs. (R)	Length of seqs. (T)	Probability of a word in common			
		K and O (4)	Our (8)	Our (14)	Simulation
3	12	0.010	0.010	0.010	0.010
	19	0.057	0.052	0.052	0.051
	23	0.108	0.097	0.097	0.096
	30	0.245	0.214	0.213	0.212
	40	0.515	0.443	0.438	0.440
	50	0.774	0.678	0.671	0.675
	60	0.929	0.852	0.846	0.849
4	24	0.011	0.010	0.010	0.010
	36	0.067	0.053	0.052	0.053
	50	0.249	0.181	0.179	0.178
	60	0.462	0.330	0.326	0.322
	70	0.693	0.508	0.501	0.503
	80	0.873	0.683	0.674	0.672
	90	0.965	0.824	0.816	0.819
5	39	0.014	0.010	0.010	0.010
	54	0.077	0.048	0.048	0.047
	60	0.130	0.078	0.078	0.078
	70	0.269	0.152	0.152	0.151
	80	0.466	0.261	0.259	0.259
	90	0.685	0.398	0.395	0.394
	100	0.864	0.549	0.545	0.543
8	87	0.034	0.010	0.010	0.010
	111	0.227	0.049	0.050	0.051
	125	0.494	0.104	0.106	0.107
	135	0.722	0.163	0.166	.0168
	145	0.899	0.242	0.245	0.245
	155	0.981	0.339	0.343	0.342
	165	0.999	0.451	0.455	0.455
	175	1.000	0.569	0.573	0.574

Table 4. Sequence lengths T that give w-word R sequence match probability = 0.01 by approximation (8) for a four-letter equally likely alphabet

Word size W	Number of sequences: R								
	2	3	4	5	6	7	8	9	10
4	**5**	**12**	**24**	39	55	71	87	103	118
5	**7**	**27**	**63**	112	166	223	280	337	392
6	**12**	**62**	**171**	327	511	710	914	1119	1321
7	**21**	**149**	**473**	972	1591	2280	3006	3745	4485
8	**37**	**366**	**1322**	**2915**	4987	7373	9943	12609	15310
9	**67**	**913**	**3716**	**8775**	15693	23927	33011	42598	52442
10	**128**	**2287**	**10477**	**26477**	49494	77838	109873	144287	180112

Bold terms also give close to 0.01 under Karlin and Ost (4), close being ± 0.001.

Table 5. Probability of a w-letter word in common to R out of R i.i.d. sequences each of T letters drawn from an equally likely four-letter alphabet

No. of seqs. (R)	Word length (w)	Length of seqs. (T)	K and O (4)	Our (8)	(100,000) Simul.	Bonferroni Bd. (15)
4	6	172	0.0111	0.0102	0.0103	0.0107
4	6	257	0.0561	0.0498	0.0493	0.0536
4	7	474	0.0107	0.0101	0.0096	0.0105
4	7	713	0.0544	0.0500	0.0508	0.0533
6	4	74	0.110	0.050	0.050	0.060
6	5	222	0.091	0.050	0.050	0.057
6	6	682	0.080	0.050	—	0.056
10	4	145	0.506	0.049	0.050	0.068
10	5	479	0.376	0.050	—	0.064
10	6	1605	0.287	0.050	—	0.062

sequences, the sequence lengths are such as to make simulations difficult for more than short word lengths. For match probabilities of 0.05, the sequence sizes are even longer. Our previous results indicate that the differences between approximations (4) and (8) will be clearest for six or more sequences, and easier to detect as match probabilities increase. This point is illustrated by the entries in Table 5. This is why our simulations are limited to small word sizes. Section 4, equation (15) gives $(T-w+1)R_w$ as an exact Bonferroni upper bound for $P_{R,R}(w,T)$. Table 5 computes the bound, and compares it with approximations (4) and (8), and simulations for $w=4,5,6,7$. The bound is useful in demonstrating for longer (and more difficult to simulate) sequences the slow convergence of the asymptotic approximation (4) for six or more sequences.

2.3. *Generalization to different sequence lengths and compositions.*

Case a: Independently but Not Identically Distributed Sequences of i.i.d. Letters. Given R sequences $\{X_{1j}\},\{X_{2j}\},\ldots,\{X_{Rj}\}$, where for $\{X_{ij}\}$, $j=1,2,\ldots,T_i$ for $i=1,2,\ldots,R$. Let $P(X_{ij}=k)=p_{i,k}$ for $k=1,2,\ldots,m$ for an m-letter alphabet. $T_1,T_2,\ldots,T_R$ are the lengths of the different sequences, and $p_{i,k}$ allows for different letter likelihoods in the different sequences. Following the reasoning leading to approximation (8), we find its generalization:

$$P_{R,R}(w;T_1,T_2,\ldots,T_R)\approx 1-\exp\left\{-T^*\prod_{i=2}^{R}\delta_i\left(1-\prod_{h=2}^{R}P_h\right)\right\} \tag{9}$$

where

$$\delta_i=1-\exp\{-(T_i-w+1)P_i^w\}$$

and

$$P_i = \sum_{k=1}^{m} p_{1,k} p_{i,k}.$$

Karlin *et al.* (1989) follow a warning of Arratia and Waterman that there are problems with the asymptotic approximations when the $p_{i,k}$ differ sharply from sequence to sequence. They note that for two independent sequences of independent letters, a sufficient condition for convergence of the approximation is for $\max_{1 \leqslant k \leqslant m}(p_{1,k}, p_{2,k}) \leqslant \sqrt{P_2}$. They note that if the sequences have the same probabilities, the condition holds. Even if this is not the case, but one sequence has an equally probable alphabet, then in practice the condition is likely to be met. In applying our approximation (9), choose for sequence 1 the sequence that has closest to equal probabilities. This will help in the approximation of the δ_i. The pairwise condition would be $\max_{1 \leqslant k \leqslant m}(p_{1,k} p_{i,k}) \leqslant \sqrt{P_i}$, for $i = 2, 3, \ldots, R$. Karlin *et al.* (1989) give the sufficient condition for the convergence of their approximation for R independent sequences, that $\max_{i,k}(p_{i,k}) \leqslant (\sum_{k=1}^{m}(\prod_{i=1}^{R} p_{i,k}))^{1/R}$. On p. 152, they apply their approximation to the following example of a matching w-word in four out of four sequences of lengths 2165, 1985, 1648 and 3915 with respective probabilities (P_T, P_C, P_A, P_G) of (0.327, 0.203, 0.261, 0.210); (0.318, 0.198, 0.275, 0.208); (0.264, 0.203, 0.263, 0.270); (0.274, 0.181, 0.307, 0.238). Using their criterion (10), they find that a $w = 9$ is significant at the 0.01 level. Using their formula (10) gives probability of a $w \geqslant 8 = 0.20$, and probability of a $w \geqslant 9$ of 0.004. Using our approximation (9), and taking the third sequence ($T = 1648$) as our first sequence since it has the most nearly equal p_is gives probabilities for a $w \geqslant 8$ of 0.11 and for a $w \geqslant 9$ of 0.002, which leads to the same conclusion as Karlin *et al.* If we had taken a different sequence as our sequence 1, the conclusion would be the same, although the probabilities would vary (for $w = 8$ between 0.113 and 0.148; for $w = 9$ between 0.0021 and 0.0029).

Case b: Independently but Not Identically Distributed Sequences of Dependent (but Stationary) Distributed Letters. Note that for case a, P_i^w is the probability of a word match between the first w letters of sequence 1 and the first w letters of sequence i. Let $G_i(w)$ denote the probability of a match between the first w letters of sequence 1 and the first w letters of sequence i. This probability can be computed as the sum over all possible w-words of the product of the probabilities of the word appearing in sequence 1 and sequence i. Approximation (9) is generalized to case b by replacing P_i^w by $G_i(w)$ in the approximation for δ_i.

3. Generalizing the Approximation to Perfect Matches in R out of S Sequences. This section gives two approximations for the probability of a perfect w-letter word match in at least R out of the S sequences. Consider the $V = C^w$ possible w-letter words from a C-letter alphabet. Use (6) to approximate δ, the probability that a given w-letter word appears in a random sequence of T letters. The probability that a particular (fixed) w-letter word appears in at least R out of S independent sequences, each of length T letters, is the sum of binomial probabilities:

$$G(R|S;\delta) = \sum_{i=R}^{S} C_i^S \delta^i (1-\delta)^{S-i}. \tag{10}$$

Order the V possible w-letter words in alphabetical order; let E_i denote the event that the ith word appears in at least R out of the S sequences. The probability that at least one of the V possible words appears in exactly R out of S sequences is $P_{R,S}(w,T) = P(\bigcup E_i)$. Using the Bonferroni lower bound and Hunter upper bound

$$\sum P(E_i) - \sum P(E_i \cap E_j) \leqslant P(\bigcup E_i) \leqslant \sum P(E_i) - \sum P(E_i \cap E_{i+1})$$

gives the approximate bounds

$$VG(R|S;\delta) - C_2^V G^2(R|S;\delta) \leqslant P_{R,S}(w,T) \leqslant VG(R|S;\delta) - (V-1)G^2(R|S;\delta). \tag{11}$$

The second approximation writes $P(\bigcup E_i) = 1 - P(\bigcap E_i^c)$, and approximates $P(E_i^c)$ by $1 - G(R|S;\delta)$. Taking the further approximation $P(\bigcap E_i^c) \approx \prod P(E_i^c)$ gives

$$P_{R,S}(w,T) \approx 1 - [1 - G(R|S;\delta)]^V \tag{12}$$

where δ can be approximated as in (6) or by

$$\delta^* = 1 - \{(V-1)/V\}^{T-w+1}. \tag{13}$$

Table 6 compares the approximation (12), approximate bounds (11), with the Karlin and Ost (1) and modified form, and simulations based on 100,000 trials. For cases where we were able to simulate, our approximation appeared more accurate than the Karlin and Ost approximations. For many cases, approximation (12) gave the closest approximation.

4. Incorporating Exact Results into the Approximation. Naus and Sheng (1996) derive results to measure the unusualness of perfect or almost

Table 6. Probability of a w-letter word in common to at least R out of S i.i.d. sequences each of T letters drawn from an equally likely four-letter alphabet

Number of			Length of seqs.	Probability of a word in common					
w	R	S	(T)	K and O (1)	K and O (1*)[a]	Our (11) LB	Our (11) UB	Our (12)	Simul.[b]
4	5	6	20	0.0044	0.0020	0.0016	0.0016	0.0016	0.0015
			30	0.033	0.020	0.014	0.014	0.014	0.014
			40	0.133	0.092	0.058	0.060	0.059	0.058
			50	0.353	0.273	0.156	0.171	0.162	0.159
			60	0.661	0.567	0.324	0.405	0.336	0.326
			70	0.904	0.847	0.482	0.801	0.556	0.537
5	7	12	100	0.066	0.050	0.025	0.025	0.025	
	8	12	100	0.0042	0.0030	0.0015	0.0015	0.0015	
8	10	20	5000	0.078	0.077	0.027	0.027	0.027	

[a] K and O (1*) uses approximation (1), but replaces T by $(T-w+1)$.
[b] 100,000 trials. (12) is computed using δ^* defined in (13).

perfect matches in multiply aligned sequences; they also find the probability of high scoring segments in all possible alignments of two sequences for various scoring systems. Theorem 2 of Section 6 of that paper gives an approximation for the probability that there exists a word of length w that is common to all of R random sequences. The approach is to condition on various pattern types for the word, and to find component probabilities that are used in the approximation.

The approximation is developed as follows. Let E_1 denote the event that the first w letters of the random sequence 1 appear somewhere in each of the other $R-1$ sequences. Let R_w denote $P(E_1)$, where the probability is evaluated over the distribution of all possibilities for the first w letters of sequence 1. Let E_t denote the event that the w-letter word consisting of letters t through $t+w-1$ in sequence 1 appears somewhere in each of the other $R-1$ sequences. Let R_{w+1} denote $P\{(E_1 \cap E_2)$, where the probability is computed over the distribution of all possibilities for the first $w+1$ letters of sequence 1. The approach is to get exactly the component probabilities R_w and R_{w+1}, and incorporate them into the following approximation:

$$\begin{aligned} P_{R,R}(w,T) = P(\cup E_i) &\approx 1 - P(E_1^c)P(E_2^c|E_1^c)P(E_3^c|E_2^c)P(E_4^c|E_3^c)\cdots \\ &= 1 - P(E_1^c)\{P(E_2^c|E_1^c)\}^{T^*-1} \\ &= 1 - (1-R_w)\{(1-2R_w+R_{w+1})/(1-R_w)\}^{T^*-1} \end{aligned} \tag{14}$$

where, in Naus and Sheng (1996), $T^* = T-w+1$. Here, T^* is defined as in equation (7).

We can also develop upper bounds for $P_{R,R}(w,T)$ using R_w. The Bonferroni upper bound is

$$P_{R,R}(w,T) = P(\cup E_i) \leqslant \sum P(E_i) = (T-w+1)R_w. \tag{15}$$

Table 5 illustrates the usefulness of this bound.

Naus and Sheng (1996) generalize the approach in Sheng and Naus (1994) to get the exact component probabilities R_w and R_{w+1}. We have since developed a difference equation approach that extends the range of computation of R_w and R_{w+1}. Programs and details are available from the authors.

The approximation (14) is still computationally more complex than the simple approximations (5) and (6) and (8) developed above. For many cases, approximations (8) and (14) give similar results. We used the more complex approximation (14) as a benchmark against which to compare the simpler approximations, particularly in the case of many long sequences where simulations were impractical. For practical applications, we would use approximation (8) or its generalizations.

REFERENCES

Glaz, J. and J. I. Naus. 1991. Tight bounds and approximations for scan statistic probabilities for discrete data. *Ann. Appl. Prob.* **1**, 306–318.

Karlin, S., G. Ghandour and D. Fousler. 1985. DNA sequence comparisons of human, mouse, and rabbit immunoglobulin Kappa gene. *Mol. Biol. Evol.* **2**, 35–52.

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

Karlin, S., F. Ost and B. E. Blaisdell. 1989. Patterns in DNA and amino acid sequences and their statistical significance. In *Mathematical Methods for DNA Sequences*, M. S. Waterman (Ed), ch. 6. Boca Raton, FL: CRC Press Inc.

Karlin, S. and F. Ost. 1988. Maximal length of common words among random sequences. *Ann. Prob.* **16**, 535–563.

Leung, M. Y., B. E. Blaisdell, C. Burge and S. Karlin. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *J. Mol. Biol.* **221**, 1367–1378.

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.* **52**, 773–784.

Naus, J. and K. N. Sheng. 1996. Screening for unusual matched segments in multiple protein sequences. *Commun. in Statist., Simulation and Computation* **25**, 937–952.

Sheng, K. N. and J. Naus. 1994. Pattern matching between two non-aligned random sequences. *Bull. Math. Biol.* **56**, 1143–1162.

Sobel, E. and H. M. Martinez. 1986. A multiple sequence alignment program. *Nucleic Acids Res.* **14**, 363–374.

Waterman, M. S. 1986. Multiple sequence alignment by consensus. *Nucleic Acids Res.* **14**, 9095–9102.

Waterman, M. S., R. Arratia and D. J. Galas. 1984. Pattern recognition in several sequences; consensus and alignment. *Bull. Math. Biol.* **46**, 515–527.

Received 19 August 1996
Revised version accepted 21 October 1996