The Annals of Statistics
1986, Vol. 14, No. 3, 971–993

AN EXTREME VALUE THEORY FOR SEQUENCE MATCHING

BY RICHARD ARRATIA,[1,2] LOUIS GORDON AND MICHAEL WATERMAN[2]

University of Southern California

Consider finite sequences $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ where the letters $\{X_t\}$ and $\{Y_t\}$ are chosen i.i.d. on a countable alphabet with $p = P\{X_1 = Y_1\} \in (0,1)$. We study the distribution of the longest contiguous run of matches between the X's and Y's, allowing at most k mismatches. The distribution is closely approximated by that of the maximum of $(1-p)mn$ i.i.d. negative binomial random variables. The latter distribution is in turn shown to behave like the integer part of an extreme value distribution. The expectation is approximately

$$\log(qmn) + k \log\log(qmn) + k \log(q/p) - \log(k!) + \gamma \log(e) - \tfrac{1}{2},$$

where $q = 1-p$, log denotes logarithm base $1/p$, and γ is the Euler constant. The variance is approximated by $(\pi \log(e))^2/6 + \frac{1}{12}$. The paper concludes with an example in which we compare segments taken from the DNA sequence of the bacteriophage lambda.

0. Introduction. DNA sequences can be represented as finite sequences over the four-letter alphabet $\{A, C, G, T\}$. Such a sequence corresponds to successive appearances of the nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T). One of the impressive accomplishments of molecular biology is the facility with which the sequences corresponding to actual genetic material are determined. Much effort is currently invested in determining the DNA sequences belonging to the chromosomes of various organisms. See for example the book *Nucleotide Sequences 1984* [Anderson et al. (1984)], which is an atlas of such representations. By mid-1985, DNA sequences with a total length of approximately 5×10^6 were known, and sequencing was proceeding at an approximate rate of 10^6 letters per year.

Sequences belonging to seemingly unrelated organisms have been found to possess long contiguous subsequences which are practically identical. Doolittle et al. (1983) report an unexpected relationship of this kind between viral DNA and host DNA. The identification and interpretation of such shared contiguous subsequences are of substantial interest to biologists. See Waterman (1984) for a review of these methods.

These aspects of matching between sequences lead us to ask the following mathematical question: for two independently generated random sequences, what is the distribution of the length of the longest run of contiguous matches? Evolution of nucleotide sequences proceeds by substitution, insertion, and deletion of nucleotides. Substitutions motivate us to study the distribution of the length of the longest contiguous run of matches allowing for a fixed number k of

Received October 1984; revised October 1985.
[1]Supported by NSF grant DMS-8402590.
[2]Supported by a grant from the System Development Foundation.
AMS 1980 *subject classifications*. Primary 62E20; secondary 62P10.
Key words and phrases. Extreme value, matching, Poisson process, inclusion–exclusion, DNA sequences.

mismatches. We do not obtain corresponding results for the more difficult case of insertions and deletions.

In Smith et al. (1985), a data analysis shows real, unrelated DNA sequences to have the same distribution of matching scores as independent sequences of the same composition. This supports the claim that the distributional results of this paper have genuine biological interest. In Section 7, the sequence of the virus lambda is shown to fit the distribution very well.

The question of the longest run of consecutive matches is closely related to the study of runs in a single sequence. Erdös and Révész (1975) consider the length of the longest run of heads in a sequence of heads and tails generated by fair coin tossing. Using a combinatorial approach, they provide almost sure upper and lower boundaries for the longest head run, and for runs interrupted by at most k tails.

Guibas and Odlyzko (1980) use generating function methods to provide deep results on problems related to those of Erdös and Révész. They look at the longest run of repetitions of a specified pattern. Among many other results, they compute the expectation and variance of the length of the longest run of repetitions, and they make the intriguing remark that the length of the longest run of repetitions has no limiting distribution.

Several results from Guibas and Odlyzko are generalized to the case of biased coin tossing, and runs with at most k interruptions, in Gordon, Schilling, and Waterman (1986). In that paper, a sequence of coin tosses is represented in terms of independent geometric random variables, and then analyzed using inclusion–exclusion and counting methods in a way similar to Watson (1954).

This paper gives results for the approximate large sample distribution, mean and variance for the length of the longest run of consecutive matches, allowing a fixed number $k = 0, 1, 2\ldots$ of mismatches, between two sequences with all letters independently drawn from some given distribution. The results in their most useful form are collected and proved in Theorems 1 and 2 of Section 4. Karlin et al. (1983) report limiting variances and expectations for pure runs of matches of a sequence with itself, similar to those we obtain in Theorem 2. Strong laws of large numbers for the length of the longest match between two or more sequences were given in Arratia and Waterman (1985a) and (1985b).

Define the length of the longest match starting within the first m letters of the sequence $X_1, X_2, \ldots$ and the first n letters of $Y_1, Y_2, \ldots$ to be

$$M(m, n) \equiv \max\{u\colon X_{i+1} \cdots X_{i+u} = Y_{j+1} \cdots Y_{j+u}, \text{ for some } 0 \le i < m, 0 \le j < n\}.$$

Throughout this paper we assume that the letters $\{X_i\}$ and $\{Y_i\}$ are all chosen independently from a countable alphabet S according to a nontrivial distribution μ. Let p represent the probability that two different letters match, so that

$$p \equiv \sum_{a \in S} (\mu_a)^2 \in (0,1). \tag{0.1}$$

The definition of $M(m, n)$ directly suggests an analysis according to position. Let $A(i, j)$ be the length of the match which begins after position (i, j),

so that $A(i, j) \geq u$ iff $X_{i+1} \cdots X_{i+u} = Y_{j+1} \cdots Y_{j+u}$. Thus $M(m, n) = \max_{0 \leq i < m,\ 0 \leq j < n} A(i, j)$, expressing the length of the longest match as the maximum of mn random variables $A(i, j)$, where each $A(i, j)$ is geometrically distributed with parameter p. These random variables $A(i, j)$ are dependent; it should not be intuitively obvious that their maximum is comparable in distribution to the maximum of mn independent copies of $A(i, j)$.

Note that the distribution of the family $\{A(i, j)\}$ is strictly stationary with respect to translations in each of the two parameters. A general theory of maxima for stationary sequences is well developed, but does not seem directly applicable here; see for example the book by Leadbetter, Lindgren and Rootzén (1983). Another situation involving the maximum of a multiple-parameter stationary family occurs in Darling and Waterman (1985).

Our main result, which is stated in Theorem 2, can be loosely described as: if m and $n \to \infty$ with $(\log m)/(\log n) \to 1$, then $M(m, n)$ has approximately the same distribution as the maximum of $(1 - p)mn$ independent geometric(p) random variables. Thus $M(m, n) - \log_{1/p}((1 - p)mn)$ has approximately an integerized extreme value distribution: $P(M(m, n) - \log_{1/p}((1 - p)mn) < c)$ is uniformly approximated by $\exp(-p^c)$, for c such that $c + \log_{1/p}((1 - p)mn)$ is an integer. Theorem 2 also states results for the length of the longest match run with at most k mismatches, which behaves like the maximum of $(1 - p)mn$ independent negative binomial random variables.

1. Survey of the analysis. The mutual dependence of the random variables $A(i, j)$ shows up in two qualitatively different ways. First of all, for each (i, j), the events $\{A(i, j) \geq u\}$ and $\{A(i + 1, j + 1) \geq u\}$ are positively correlated. Given the event $\{A(i, j) \geq u\}$ that we observe a match of length at least u after (i, j), the expected number of consecutive events $\{A(i, j) \geq u\}$, $\{A(i + 1, j + 1) \geq u\}, \{A(i + 2, j + 2) \geq u\}, \ldots$ associated with that match is $1 + p + p^2 + \cdots = (1 - p)^{-1}$. The same phenomenon, clusters of average size $(1 - p)^{-1}$, occurs in the analysis of the longest head run in a sequence of tosses of a p-biased coin.

We need, therefore, to compensate for this clustering when approximating the distribution of $M(m, n)$. This is achieved by introducing the random variables $A'(i, j) \equiv A(i, j)1(X_i \neq Y_j)$, where $1(E)$ is the indicator function for the event E. In contrast to the positive correlation of analogous events defined in terms of the random variables $A(i, j)$, the events $\{A'(i, j) \geq u\}, \{A'(i + 1, j + 1) \geq u\}$, $\{A'(i + 2, j + 2) \geq u\}, \ldots$ are negatively correlated. The distribution of each $A'(i, j)$ is a mixture, with weights $(1 - p)$ and p, respectively, of the geometric(p) distribution and the unit mass on zero. If there were no other dependence, then $M(m, n)$ would be like the maximum of mn independent copies of $A'(i, j)$, which in turn is like the maximum of $(1 - p)mn$ independent copies of $A(i, j)$.

The second type of dependence occurs whenever the distribution μ of the letters is not the uniform distribution, $\mu_a = 1/|S|$ for all $a \in S$. The simplest aspect of this dependence is that a match at (i, s) is positively correlated with a match at (i, t) for all $s \neq t$: $P(X_i = Y_s \text{ and } X_i = Y_t) = P(X_i = Y_s = Y_t) = \sum_{a \in S}(\mu_a)^3 \geq (\sum_{a \in S}(\mu_a)^2)^2 = P(X_i = Y_s)P(X_i = Y_t)$, with equality if and only if

μ is uniform. It requires calculation, which we carry out in Sections 2–4, to prove that if m and $n \to \infty$ with $(\log m)/(\log n) \to 1$, then this second type of dependence does not have a significant effect on $M(m, n)$. In contrast, it is shown in Arratia and Waterman (1985b) that there exists a critical constant $\theta_{\text{cr}} \in [0, 1)$ depending on μ, with $\theta_{\text{cr}} = 0$ iff μ is uniform, such that if m and $n \to \infty$ with $\log(m)/\log(n) \to \theta \in (0, \infty)$, then

$$M(m, n)/\left(\log_{1/p}(mn)\right) \to_p 1 \quad \text{iff } \theta \in \left[\theta_{\text{cr}}, 1/\theta_{\text{cr}}\right]. \tag{1.1}$$

See Section 5 for further discussion of the situation in which $\log(n)/\log(m) \to \theta \neq 1$.

We now describe, for general k, the framework that will be used to prove our theorems about the length of the longest k-interrupted match run between two sequences. The discussion above has been about the special case, $k = 0$. Fix an integer u to serve as a test level.

We say that a k-interrupted run of length u is witnessed at (i, j) if $X_i \neq Y_j$ and $X_{i+s} = Y_{j+s}$ for all but at most k values of s, $1 \leq s \leq u$. We write $R_k(i, j; u)$ for the indicator of this event, which involves $2(u + 1)$ randomly chosen letters. When $k = 0$, we also speak of a (pure) run of u matches, witnessed at (i, j). We write $N_k(m, n; u)$ for the total number of witnesses at level u, and $M_k(m, n; u)$ for the length of the longest k-interrupted match, starting within the first m letters of one sequence and the first n letters of the other. Formally, we define, for integers $0 \leq k \leq u$, the random variables

$$\begin{aligned} R_k(i, j; u) &\equiv 1(X_i \neq Y_j) 1\Big(k \geq \sum_{1 \leq s \leq u} 1(X_{i+s} \neq Y_{j+s})\Big), \\ N_k \equiv N_k(m, n; u) &\equiv \sum_{0 \leq i < m,\, 0 \leq j < n} R_k(i, j; u), \end{aligned} \tag{1.2}$$

and

$$M_k \equiv M_k(m, n) \equiv \max\{u\colon N_k(i, j; u) > 0\}.$$

In the special case of pure matching, $k = 0$, we omit the subscript k. Note that the event $\{M_k \geq u\}$ involves $m + u$ X's and $n + u$ Y's. We show in Section 6 that the natural alternate definition involving exactly m X's and n Y's is negligibly different from the above definition.

Bonferroni's (1936) inclusion–exclusion formulae, as used in Watson (1954), are the basis of our calculation that $P(N_k > 0) \approx \exp(-EN_k)$. Write $B \equiv B(m, n; u) \equiv \{b = (i, j; u)\colon 0 \leq i < m,\ 0 \leq j < n\}$ for the set of indices associated with potential witnesses to $\{M_k(m, n) \geq u\}$. For finite $C \subset B$, write $R_k(C) = \prod_{b \in C} R_k(b)$ for the indicator that for each $b = (i, j; u) \in C$, $X_i \neq Y_j$, and $X_{i+1} \cdots X_{i+u}$ matches $Y_{j+1} \cdots Y_{j+u}$ with at most k mismatches. The rth term in the inclusion–exclusion series for $P(N_k > 0)$ is the expectation of the random variable

$$S_k^{(r)} \equiv S_k^{(r)}(m, n; u) \equiv \sum_{C \subset B\colon |C| = r} R_k(C). \tag{1.3}$$

Truncation of the inclusion–exclusion series gives lower and upper bounds on the

random variable $1(N_k > 0)$: for $\tau = 0, 1, 2, \ldots$

$$\sum_{1 \le r \le 2\tau} (-1)^{r+1} S_k^{(r)} \le 1(N_k > 0) \le \sum_{1 \le r \le 2\tau+1} (-1)^{r+1} S_k^{(r)}. \tag{1.4}$$

Furthermore, for each $j = 0, 1, 2, \ldots$, inclusion–exclusion (using $S_k^{(0)} \equiv 1$) gives lower and upper bounds on the random variable $1(N_k = j)$: for $\tau = 0, 1, 2, \ldots$

$$\sum_{0 \le r \le 2\tau - 1} (-1)^r \binom{j+r}{j} S_k^{(j+r)} \le 1(N_k = j) \le \sum_{0 \le r \le 2\tau} (-1)^r \binom{j+r}{j} S_k^{(j+r)}. \tag{1.5}$$

These truncated inclusion–exclusion bounds allow us to prove, in Theorems 1 and 1′, that the number N_k of locations at which k-interrupted matches of length at least u begin is approximately Poisson, and that the locations at which these matches occur are approximately a Poisson process. To prove Theorem 1, we need to evaluate $ES_k^{(r)}$, for all positive integers r. For $r = 1$, $S_k^{(1)} = N_k$. For fixed $r = 2, 3, \ldots$, and any fixed t, Lemma 5 implies that as m and $n \to \infty$ with $\log(m)/\log(n) \to 1$, uniformly over integers u such that $EN_k \le t$,

$$ES_k^{(r)} - (EN_k)^r / r! \to 0. \tag{1.6}$$

Taking expectations in (1.5) and using (1.6), we obtain the result $P(N_k = j) \to \exp(-EN_k)(EN_k)^j/j!$, uniformly in integers u such that $EN_k \le t$. Thus the Poisson distribution approximates the distribution of the number of witnesses to a match.

For arbitrary m and n, there is the trivial upper bound, $P(M_k(m, n) \ge u) = P(N_k > 0) \le EN_k$, which should be used when EN_k is small. Theorem 1 shows that, for m and n large with $\log(m)/\log(n)$ near 1,

$$P(M_k(m, n) \ge u) = P(N_k > 0) \approx 1 - \exp(-EN_k) = EN_k\left(1 - EN_k/2 + (EN_k)^2/6 - \cdots\right),$$

so that the relative error in using the trivial upper bound is also small when the bound is small.

At the level of establishing a limit distribution for the number of witnesses, $N = N(m, n; u)$, the inclusion–exclusion framework described above is equivalent to the method of moments. To see this, note that for each d, the two sets of random variables, $\{S^{(1)}, S^{(2)}, \ldots, S^{(d)}\}$ and $\{N, N^2, \ldots, N^d\}$, both have the same d-dimensional linear span. Thus, calculating finite limits for $ES^{(1)}, ES^{(2)}, \ldots$ is equivalent to calculating finite limits for the moments of N. However, the inclusion–exclusion framework directly gives lower and upper bounds on $P(N = 0)$, and thus could be used to get rate-of-convergence estimates for the distribution of $M(m, n) - \log_{1/p}((1-p)mn)$.

2. Combinatorial aspects of comparisons. In this section we discuss certain combinatorial aspects of matching with shifts. There are two levels of graphs to consider. At the lower level, *comparison graphs* describe which comparisons are made between individual sites and their associated letters. At the higher level, *adjacency graphs* describe the overlap among bundles of comparisons.

Recall from (1.2) that a run of matches must start with a mismatch. Therefore, it is notationally convenient to think of the sequences as beginning with one additional letter, which we index by 0. We shall see in Section 6 that the mathematically convenient definition of maximum run of matches is distributionally equivalent to the more intuitive definition of longest run of matches.

In the rest of the paper, we consider two sequences of characters, $X_0, X_1, X_2, \ldots$ and $Y_0, Y_1, Y_2, \ldots$, located in the Euclidean plane at the lattice sites $Z^+ \times \{1,2\}$. The process of comparing X_i with Y_j is represented as an undirected edge between $(i,1)$ and $(j,2)$. We usually call this the edge or *comparison* (i, j), instead of using the more correct but awkward notation $\{(i,1),\{j,2)\}$.

A set of $u+1$ consecutive parallel comparisons is called a *bundle*, denoted by the triple $(i, j; u) \equiv \{(i+k, j+k)\colon k = 0,1,\ldots,u\}$. The definition of bundle anticipates the study of the random variables $R_k(i, j; u)$ introduced in (1.2). A set C of r bundles, all sharing the same value of u, is called a *configuration* or an r-configuration. Two bundles, $b_s = (i_s, j_s; u)$ and $b_t = (i_t, j_t; u)$ are said to be *x-adjacent*, denoted $b_s \approx_x b_t$, if $b_s \neq b_t$ and $|i_s - i_t| \leq u$. They are said to be *y-adjacent*, denoted $b_s \approx_y b_t$, if $b_s \neq b_t$ and $|j_s - j_t| \leq u$. The bundles are said to be *adjacent*, denoted $b_s \approx b_t$, if either $b_s \approx_x b_t$, or $b_s \approx_y b_t$, or both. Note that, by definition, a bundle is not adjacent to itself. The two bundles, b_s and b_t, are said to be *parallel* if $j_s - i_s = j_t - i_t$. They are said to be *doubly adjacent* if both $b_s \approx_x b_t$ and $b_s \approx_y b_t$. If two bundles are both adjacent and parallel, then they are doubly adjacent.

The reader may wish to construct the following figure. The comparison of X_i with Y_j corresponds to the lattice point (i, j) in Z^2. A bundle is then a set of $u+1$ points lying on a 45-degree line. Parallel bundles overlap on these 45-degree lines. Adjacent bundles have overlapping x or y coordinates. Biologists perform visual analysis of matching by using such representations in what they call "dot matrix analysis." See, for example, Waterman (1984).

Associated with an r-configuration C are two graphs. The first is the *comparison graph*, denoted $\cup(C)$, because it is the union of the bundles in C. The comparison graph $\cup(C)$ is a bipartite graph on $Z^+ \times \{1,2\}$ having at most $r(u+1)$ edges (comparisons). Each comparison connects an x-site in $Z^+ \times \{1\}$ with a y-site in $Z^+ \times \{2\}$.

The second graph is the *adjacency graph*, denoted $A(C)$, which is the restriction to C of the adjacency relation. Thus, the vertices of $A(C)$ are bundles, and the edge $\{b_s, b_t\}$ is in $A(C)$ if and only if $b_s \in C$, $b_t \in C$, and $b_s \approx b_t$. The inclusion of edge $\{b_s, b_t\}$ in $A(C)$ indicates that some of the comparisons which constitute the bundles b_s and b_t share common sites.

The comparison graph $\cup(C)$ has a particularly simple structure if the adjacency graph $A(C)$ is atomic. This case corresponds to a set of r mutually independent runs of matches, because none of the bundles in C overlap. The

comparison graph $\cup(C)$ then contains exactly $r(u + 1)$ disjoint edges, each corresponding to some one comparison in some one bundle. We get our lower bound on $ES^{(r)}$ quite easily in Lemma 4, just by considering the contribution from configurations of this form.

There are some configurations C of bundles whose comparison graph $\cup(C)$ is quite complicated; see Guibas and Odlyzko's (1981) discussion of periods in strings for a treatment of matching a single sequence to itself. In order to put an upper bound on the contribution $ER(C)$ to $ES^{(r)}$ from such a configuration, we select a reduced subset $C^* \subset C$, so that $ER(C) \leq ER(C^*)$. The reduced configuration C^* must be chosen small enough so that $\cup(C^*)$ is simple, but large enough so that $ER(C^*)$ is a useful upper bound.

We now define the reduction operation, which reduces an r-configuration C to the $(r - 1)$-configuration $C^* = C - \{b_r\}$. A *reduction* removing b_r from C is allowed if C contains two bundles b_s and b_t such that:

(a) $b_r \approx_x b_s$ and $b_r \approx_y b_t$, and
(b) $C - \{b_r\}$ contains a pair of bundles which is adjacent but not parallel.

Note that b_s and b_t need not be distinct. A configuration which cannot be reduced is called *irreducible*.

Not every configuration can be obtained by reduction from a larger configuration. Neither can a particular configuration be obtained by reduction from very many other configurations. These two consequences of the definition are formalized in the lemma below.

LEMMA 1. *Let C^* be a given r-configuration and let $t > r$. The number of t-configurations, which yield C^* after some series of $t - r$ reduction operations, is less than $[(2u + 1)t]^{2(t-r)}$. If none of the bundles in C^* are adjacent, or if all bundles in C^* are parallel, the number of such t-configurations is zero.*

PROOF. Consider the adjacency graph $A(C)$ of a configuration which is about to be reduced by deleting the bundle $b = (i, j; u)$ from the configuration C. From condition (a) in the definition of reduction, there must exist bundles b_2 and b_3 in C, such that $b_2 \approx_x b$ and $b_3 \approx_y b$. Hence, at each of the $t - r$ stages of reduction, there are fewer than $[(2u + 1)t]^2$ choices for (i, j), these choices being determined by the remaining undeleted bundles. Reduction never produces a configuration of mutually nonadjacent or mutually parallel bundles because of condition (b) in the definition. □

The next proposition tells us that the comparison graph of an irreducible configuration has a relatively simple structure.

LEMMA 2. *Let $r > 1$ and $C = \{b_1, \ldots, b_r\}$ be an irreducible r-configuration having a connected adjacency graph $A(C)$. Exactly one of the following three cases must occur.*

Case 1. *No pair of bundles in C is doubly adjacent. In this case, every connected component of the comparison graph $\cup(C)$ is a tree having r or fewer edges, and $\cup(C)$ contains exactly $r(u + 1)$ edges.*

Case 2. *C contains a pair of bundles which are doubly adjacent but not parallel. In this case, C consists of $r = 2$ bundles, every connected subgraph of $\cup(C)$ is a simple path, and $\cup(C)$ contains exactly $2(u + 1)$ edges.*

Case 3. *C contains a pair of overlapping (i.e., doubly adjacent and parallel) bundles. In this case, all bundles in C are parallel, every connected subgraph of $\cup(C)$ consists of a single edge, and there are fewer than $r(u + 1)$ edges in $\cup(C)$.*

PROOF. Case 1. Assume that no two bundles in C are doubly adjacent. It follows from this that for each comparison $c \in \cup(C)$ there is a *unique* bundle $b \in C$ such that $c \in b$, and so $\cup(C)$ has exactly $r(u + 1)$ edges.

Assume, in order to obtain a contradiction, that $\cup(C)$ contains a cycle. Choose such a cycle having minimal length τ. Since $\cup(C) \subset Z^+ \times \{1, 2\}$ is a bipartite graph, τ must be even, with $\tau \geq 4$, and we can label the sites along the cycle as $(s(1), 1), (s(2), 2), (s(3), 1), \ldots, (s(\tau), 2)$. Recall that (i, j) is our notation for the undirected edge $\{(i, 1), (j, 2)\}$. The edges that form this cycle are the comparisons

$$\begin{aligned} c_k &= (s(k), s(k + 1)) \qquad \text{if } k \text{ is odd,} \\ &= (s(k + 1), s(k)) \qquad \text{if } k \text{ is even,} \end{aligned}$$

for $k = 1$ to τ, with the understanding that $s(\tau + 1) \equiv s(1)$. For $k = 1$ to τ, let b_k be the unique bundle in C such that $c_k \in b_k$. Now $b_1 \approx_x b_\tau$, $b_1 \approx_y b_2$, and (b_2, b_3) is a pair of bundles which are adjacent but not parallel, so that a reduction removing b_1 is possible. This contradicts the irreducibility of C and proves that $\cup(C)$ contains no cycles.

Now assume, in order to obtain a contradiction, that $\cup(C)$ contains a connected subgraph G having more than r edges. By the pigeonhole principle, there is some pair of edges in G which both belong to the same bundle in C. Since G is connected, there is a path of edges in G which begins with one of these two edges and ends with the other. Choose such a path in $\cup(C)$, $c_1, c_2, \ldots, c_\tau$, to have minimal length τ. For $k = 1$ to τ let $b_k \in C$ with $c_k \in b_k$, so that $b_1 = b_\tau$. It is not possible that $\tau = 3$, for then b_1 and b_2 would be doubly adjacent. The pairs (b_1, b_2), (b_2, b_3), and (b_3, b_4) are alternately x-adjacent and y-adjacent. By hypothesis b_3 and b_4 are not doubly adjacent, so they are not parallel. Hence a reduction removing b_2 from C is possible. This contradicts the irreducibility of C and proves that no connected component of $\cup(C)$ has more than r edges.

Case 3. Assume that C contains a pair of doubly adjacent, parallel bundles, b_t and b_s. Any bundle adjacent to b_s must be parallel (and hence doubly adjacent) to b_s, since otherwise C could be reduced by the reduction removing b_t. Similarly, any bundle adjacent to b_t must be parallel to b_t. Since $A(C)$ is connected, it follows by induction that C consists of r mutually parallel bundles.

All comparisons in $\cup(C)$ are parallel, so no two distinct comparisons can intersect, and hence every connected subgraph of $\cup(C)$ consists of a single comparison. There will be fewer than $r(u + 1)$ comparisons in $\cup(C)$ because b_s and b_t share some comparisons.

Case 2. At this point, we have disposed of any irreducible C having no doubly adjacent pair of bundles in the proof of Case 1. We have disposed of any

irreducible C having a pair of adjacent and parallel bundles in the proof of Case 3. Therefore, we may assume that C contains a pair of doubly adjacent bundles and that no pair of adjacent bundles in C is parallel.

Let b_s and b_t be two bundles in C which are doubly adjacent but not parallel. C cannot contain a third bundle, adjacent to b_s (and hence not parallel to b_s), since otherwise C could be reduced by the reduction removing b_t. Similarly, C cannot contain a third bundle, adjacent to b_t. Since $A(C)$ is connected, it follows that $r = 2$ and $C = \{b_s, b_t\}$.

No comparison can belong to two nonparallel bundles, so $\cup(C)$ contains exactly $2(u + 1)$ comparisons. Label the two bundles in C as $(i, i + k; u)$ and $(i', i' + k'; u)$. A site $(s, 1)$ in $\cup(C)$ can only share an edge with sites of the form $(s + k, 2)$ or $(s + k', 2)$, and a site $(s, 2)$ in $\cup(C)$ can only share an edge with sites of the form $(s - k, 1)$ or $(s - k', 1)$. From this it follows that the only connected subgraphs of $\cup(C)$ are simple paths. □

3. Estimates using Jensen's inequality. Recall that the letters $X_0, X_1, \ldots, Y_0, Y_1, \ldots \in S$ are assumed to be i.i.d. (μ), for some nontrivial distribution μ. Let p_r be the probability that r different letters match, for $r = 1, 2, \ldots$, so

$$p_r = \sum_{a \in S} (\mu_a)^r, \tag{3.1}$$

and $p \equiv p_2$. Jensen's inequality [Hardy, Littlewood and Pólya (1934), formula 2.10.3] says that for $0 < r < s$, $(p_s)^{1/s} < (p_r)^{1/r}$. Thus for $s = 3, 4, \ldots$, $p_s < p^{s/2}$. Let $\delta = \delta(\mu)$ be defined by $(p_3)^{1/3} = p^{\delta} p^{1/2}$, so that $p^2 = [\Sigma_a(\mu_a)^2]^2 = [E(\mu_X)]^2 \le E[(\mu_X)^2] = \Sigma_a(\mu_a)^3 = p_3 < p^{3/2}$ implies that $0 < \delta \le \frac{1}{6}$. Thus Jensen's inequality, $p_s \le (p_3)^{s/3}$ for $s = 3, 4, \ldots$, yields

$$p_s \le p^{s/2} p^{\delta s} \quad \text{for } s = 3, 4, \ldots . \tag{3.2}$$

LEMMA 3. *Given a nontrivial distribution μ on a countable alphabet S, let $p \in (0, 1)$ and $\delta \in (0, \frac{1}{6}]$ be defined as in the discussion preceding* (3.2). *For $r \ge 2$, let C be an irreducible r-configuration having connected adjacency graph $A(C)$. The following estimates hold, for $k = 0, 1, 2, \ldots$.*

Case 1. *If no two bundles in C are doubly adjacent, then*

$$E\{R_k(C)\} \le p^{-rk} u^{rk} p^{((r+1)/2)u} p^{\delta u}.$$

Case 2. *If C consists of $r = 2$ bundles which are doubly adjacent but not parallel then*

$$E\{R_k(C)\} \le p^{-2k} u^{2k} p^{u} p^{\delta u}.$$

Case 3. *If all bundles in C are parallel, and $C = \{(i + a_s, j + a_s; u)$: $s = 1$ to $r\}$, with $0 = a_1 < a_2 < \cdots < a_r = a$, then*

$$E\{R_k(C)\} \le p^{-2k} u^{2k} p^{2u} \quad \text{if } a > u,$$

$$E\{R_k(C)\} = 0 \quad \text{if } a \le u \text{ and } r > k + 1,$$

and

$$E\{R_k(C)\} \le u^{k-r+1}a^k p^{u+a-2k} \quad \text{if } a \le u \text{ and } r \le k+1.$$

PROOF. For any bipartite graph $G \subset Z^+ \times \{1,2\}$, define the event $E_G \equiv \{X_i = Y_j$ whenever the sites $(i,1)$ and $(j,2)$ are connected in $G\}$. Let t denote a connected component of G, and write $|t|$ for the number of vertices in that component. Since all letters are i.i.d. (μ),

$$P(E_G) = \prod_t (p_{|t|}). \tag{3.3}$$

Cases 1 and 2. Lemma 2 says that $\cup(C)$ has exactly $r(u+1)$ edges, and each connected subgraph of $\cup(C)$ is a tree.

Choose and fix, in each of the r bundles, some k comparisons after the first comparison. There are $\binom{u}{k}^r \le u^{rk}$ ways to do this. Delete these edges, together with the first comparison from each of the r bundles in C, from the graph $\cup(C)$, to form a subgraph G which has exactly $r(u-k)$ edges. In order for the indicator $R_k(C)$ to be 1, there must be some such choice of G, with $X_i = Y_j$ whenever the sites $(i,1)$ and $(j,2)$ are connected in G, hence

$$E\{R_k(C)\} \le \sum_G P(E_G) \le u^{rk} \max_G P(E_G). \tag{3.4}$$

Let t be a connected component of G, so that t is a tree with $|t|$ vertices and $|t|-1$ edges. Let

$$\tau \equiv \sum_{t:\,|t|\ge 3} 1 \quad \text{and} \quad \theta \equiv \sum_{t:\,|t|\ge 3} (|t|-1),$$

so that τ is the number of connected components which have 2 or more edges, and θ is the total number of edges contained in these τ trees. Since G has $r(u-k)$ edges, G has exactly $r(u-k)-\theta$ components t having $|t|=2$. From (3.2) and (3.3),

$$P(E_G) = \prod_t p_{|t|} \le p^{r(u-k)-\theta} p^{(\theta+\tau)/2} p^{(\theta+\tau)\delta}. \tag{3.5}$$

The first factor above comes from trees consisting of a single comparison. The last two terms come from inequality (3.2), applied to the τ trees consisting two or more comparisons, with a total of θ edges connecting $\theta+\tau$ sites.

In Case 1, every connected component of $\cup(C)$ has at most r edges. Thus every component t of G has at most r edges, so $\tau \ge \theta/r$. Substitute this into (3.5) to get the first of the two inequalities below:

$$\begin{aligned} P(E_G) &\le p^{r(u-k)-\theta} p^{(\theta+(\theta/r))/2} p^{(\theta+(\theta/r))\delta} \\ &\le p^{-rk} p^{((r+1)/2+\delta)u}. \end{aligned} \tag{3.6}$$

The second inequality is valid because $\theta \in [0, ru]$. To check this inequality, take logarithms to get an inequality which is linear in θ, and thus only needs to be checked at the endpoints. At the endpoint $\theta = 0$, we use $\delta \le \frac{1}{2}$, and at the endpoint $\theta = ru$, we use $\delta \ge 0$. Combining (3.4) and (3.6) establishes Case 1.

In Case 2, substitute $r = 2$ and $\tau \geq 0$ into (3.5) to get the first of the inequalities below:

$$P(E_G) \leq p^{2(u-k)-\theta} p^{\theta/2} p^{\theta\delta} \leq p^{-2k} p^{(1+\delta)u}. \tag{3.7}$$

As in Case 1, the last inequality is valid because $\theta \in [0, 2u]$; it checks at $\theta = 0$ using $\delta \leq 1$, and it checks at $\theta = 2u$ using $\delta \geq 0$. Combining (3.4) and (3.7) establishes Case 2.

Case 3. If $a > u$, then the bundles b_1 and b_r are not adjacent, so $R_k(b_1)$ and $R_k(b_r)$ are independent. Hence $ER_k(C) \leq ER_k(\{b_1, b_r\}) = [ER_k(b_1)]^2 \leq [u^k p^{u-k}]^2$ and the bound is established.

If $r > k + 1$ and $a \leq u$, then $R_k(C)$ is identically zero, because to have $R_k(C) = 1$ requires that there are nonmatches at $(i + a_s, j + a_s)$ for $s = 2$ to r, which forces $r - 1 \geq k + 1$ nonmatches between $X_{i+1} \cdots X_{i+u}$ and $Y_{j+1} \cdots Y_{j+u}$.

Finally, consider the case $r \leq k + 1$ and $a \leq u$. To have k-runs witnessed at bundles $b_2, \ldots, b_r$, there must be nonmatches at comparisons $(i + s, j + s)$ for $s \in \{a_2, \ldots, a_r\}$. Thus to have a k-run at b_1, there can be at most $k - (r - 1)$ nonmatches at comparisons $(i + s, j + s)$ for $s \in \{1, 2, \ldots, u\} - \{a_2, \ldots, a_r\}$. To have a k-run at b_r, there can be at most k nonmatches at comparisons $(i + s, j + s)$ for $s \in \{u + 1, \ldots, u + a\}$. Choose a subset of size $k - r + 1$ from $\{1, 2, \ldots, u\} - \{a_2, \ldots, a_r\}$, and a subset of size k from $\{u + 1, \ldots, u + a\}$, and require matches at $(i + s, j + s)$ for the $u + a - 2k$ values s in $\{1, 2, \ldots, u + a\} - \{a_2, \ldots, a_r\}$ and the two chosen subsets. The two subsets can be chosen in $\leq (u - r + 1)^{k-r+1} a^k \leq u^{k-r+1} a^k$ ways. Thus

$$E\{R_k(C)\} \leq u^{k-r+1} a^k p^{u+a-2k}. \quad \square$$

4. Approximate distribution of the longest match run. In this section we compute the approximate distribution of the longest k-interrupted run of matches. We combine the combinatorial work on reduction of Section 2 and the probability bounds of Section 3 with Watson's (1954) formulation of the inclusion–exclusion principle to obtain our main result, Theorem 1. Using (1.4) and (1.6), we need only show, in an appropriate sense, that $E\{S_k^{(r)}(m, n; u)\}$ converges to $[E\{S_k^{(1)}(m, n; u)\}]^r / r!$. We obtain the uniform lower and upper bounds for this in Lemmas 4 and 5, respectively.

Throughout this section, we adopt the convention that $m \leq n$. We write $q = 1 - p$, where p, the probability of a match on a given comparison, is defined in (1.1). Let $G_k(u) = E\{R_k((0, 0; u))\}/q$ = the probability that there are no more than k nonmatches among the u comparisons $(1, 1), (2, 2), \ldots, (u, u)$. Since $G_k(u)$ is the probability of at least $u - k$ "failures" (matches, each with probability p,) before the $(k + 1)$th "success" (nonmatches, each with probability q), $G_k(u)$ is the upper tail probability of a negative binomial distribution. It is easily

seen that

$$G_k(u) = \sum_{t \geq u} \binom{t}{k} p^{t-k} q^{k+1} = q^{k+1}(d/dp)^k \{ p^u/(1-p) \}/k! \tag{4.1}$$

for integers $u \geq k$. We use the final expression above to define $G_k(u)$ for all real u, and note that

$$G_k(u) \sim (qu/p)^k p^u/k! \quad \text{as } u \to \infty. \tag{4.2}$$

LEMMA 4. *If* $2u > r$ *and* $n \geq m > 4ru$, *then*

$$E\{S_k^{(r)}(m, n; u)\} > [1 - 2ru(m+n)/(mn)]^r [qmnG_k(u)]^r/r!.$$

PROOF. $S_k^{(r)}(m, n; u)$ was defined in (1.3). We can bound $E\{S_k^{(r)}(m, n; u)\}$ below by only summing over those configurations C consisting of r mutually nonadjacent bundles in $B(m, n; u)$, the set of bundles whose initial comparison connects one of the first m x-sites with one of the first n y-sites. For these C, $R_k(C)$ is the product of r-independent indicator random variables, each with expectation $qG_k(u)$, so $ER_k(C) = (qG_k(u))^r$. In choosing such a configuration C one bundle at a time, the added bundle can start at any site at least u sites away from any of the previously chosen bundles, allowing at least $[m - (r-1)(2u+1)][n - (r-1)(2u+1)] > mn - 2ur(m+n)$ choices for the additional bundle. Hence there are at least $[mn - 2ur(m+n)]^r/r!$ such r-configurations C. □

LEMMA 5. *Choose and fix* $k \geq 0$, $r \geq 1$, *and* $\varepsilon > 0$. *Let* $m, n \to \infty$ *in such a manner that* $\log(m)/\log(n) \to 1$. *Let*

$$U(mn) = \{u: mnu^k p^u \in [\varepsilon, \varepsilon^{-1}]\}. \tag{4.3}$$

Then

$$\sup_{u \in U(mn)} \left| \{(qmnG_k(u))^r/r!\}^{-1} E\{S_k(m, n; r, u)\} - 1 \right| \to 0.$$

PROOF. The crucial observation is that the random variable $R_k(C)$ is a product of mutually independent indicators which correspond to connected components in the adjacency graph $A(C)$. Further, the reduction operations let us study components whose graphs in $\cup(C)$ are of particularly simple form.

We assume that $m \leq n$ and that $u > 1$. We write $\eta, \eta_0, \ldots$ for constants which depend on p, k, ε, and r, but not on m, n, or u, and whose exact value is not of interest.

Consider a particular r-configuration C. If C is reducible, then it may be reduced to some irreducible configuration $C^* \subset C$; if C is irreducible, we take $C^* = C$. Let $C_1, \ldots, C_\tau$ be the connected components of $A(C^*)$, so that the adjacency graphs $A(C_s)$ are individually connected, but no pair of comparison graphs $\cup(C_s)$ and $\cup(C_t)$ share any sites. Hence the indicators $R_k(C_s)$, $s = 1$ to τ, are functions of disjoint sets of independent random variables, and so are

independent. Thus

$$ER_k(C) \le ER_k(C^*) = \prod_{1 \le s \le \tau} ER_k(C_s).$$

Since C^* is irreducible, so is each C_s. To make use of Lemma 1, let $\theta(C^*) = 1$ if $|C^*| = r$ and $\theta(C^*) = \eta u^{2r}$, with $\eta = (4r)^{2r}$ if $|C^*| < r$. Write S_k for $S_k^{(r)}(m, n; u)$. We have

$$ES_k \le \sum_{\{C^* \text{ irreducible}\}} \theta(C^*) \prod_{s \le \tau} E\{R_k(C_s)\}. \tag{4.4}$$

The sum is taken over all irreducible configurations C^* which are r-configurations, or which can be obtained by reduction from an r-configuration. The product is taken over the τ subsets C_s of C^* which correspond to connected components in the adjacency graph $A(C^*)$.

Note that if $\theta(C^*) > 1$, then C^* was obtained from an r-configuration by reduction, so there exists some pair of bundles in C^* which are adjacent but not parallel, and hence at least one of the C_s belongs to Cases 1 or 2 in Lemma 2. For $j = 1, 2, 3$ let σ_j be the sum of $ER_k(C^*)$ over all ρ-configurations C^* belonging to Case j of Lemma 2, with C^* irreducible, $A(C^*)$ connected, and $2 \le \rho \le r$. Proceeding from (4.4), and assuming $u \in U$ in (4.3), we have

$$ES_k \le \left\{qmnG_k(u) + r!\eta_0\left[\eta u^{2r}(\sigma_1 + \sigma_2) + \sigma_3\right]\right\}^r / r!. \tag{4.5}$$

From (4.2) and (4.3), $qmnG_k(u)$ is bounded away from zero and infinity. The factor η_0 appears in (4.5) with each σ_j to compensate for possibly spurious multiplication by factors $qmnG_k(u)$ when taking the rth power. Because $u \in U(mn)$, $\eta_0 < \varepsilon^{-r}$. The multiplication within braces by $r!$ cancels the effect of division by the final $r!$. It is used to count r-configurations in which no two bundles are adjacent. The factor ηu^{2r} which multiplies σ_1 and σ_2 corresponds to those C^* with $\theta(C^*) > 1$, and is obtained from Lemma 1.

We now analyze the contributions to (4.5) of the sums σ_1, σ_2, and σ_3.

Case 1. Because we have assumed that $m \le n$, there are at most $\rho! mn([m + n][2u + 1])^{\rho - 1} \le \eta_1 mn^\rho u^\rho$ ρ-configurations C_s, with $\rho \le r$, in which no two bundles are doubly adjacent, but $A(C_s)$ is connected. Using Lemma 3 for the first inequality and (4.3) for the second inequality, we get

$$\begin{aligned}
\eta u^{2r}\sigma_1 &< \sum_{\rho \le r} \eta_2 mn^\rho u^{2r} u^{\rho + \rho k} p^{u((\rho+1)/2+\delta)} \\
&< \sum_{\rho \le r} \eta_3 mn^\rho u^{3r+kr} (mn)^{-((\rho+1)/2+\delta)} \\
&= \sum_{\rho \le r} \eta_3 (n/m)^{(\rho-1)/2} (mn)^{-\delta} u^{3r+kr}.
\end{aligned}$$

Since $\log(n)/\log(m) \to 1$, it follows that $(n/m)^{(\rho-1)/2} n^{-\delta/2} \to 0$. Further, (4.3) implies that $u \sim \log_{1/p}(mn)$, hence $u^{3r+kr}(mn)^{-\delta/2} \to 0$, and so $u^{2r}\sigma_1 \to 0$, uniformly over $u \in U$.

Case 2. There are fewer than $mn\,(2u + 1)^2$ configurations consisting of two bundles which are doubly adjacent but not parallel. Using this and Lemma 3,

$\sigma_2 < \eta_4 mnu^2 u^{2k} p^{u+u\delta}$. We now employ (4.3) to obtain $u^{2r}\sigma_2 < \eta_5 u^{2+2r+2k} p^{u\delta} \to 0$, uniformly over $u \in U(mn)$.

Case 3. Let $C^* = \{(i + a_s, j + a_s; u)\colon s = 1, 2, \ldots, \rho\}$ with $2 \le \rho \le r$ and $0 = a_1 < a_2 < \cdots < a_\rho = a$. Assume also that $A(C^*)$ is connected. Because $A(C^*)$ is connected, there are fewer than $mn[\min(a, u)]^{\rho-1}$ such ρ-configurations. Use Lemma 3 for the first line below, where the first sum corresponds in Lemma 3 to the sub-case $r > k + 1$ or $a > u$, and the second term corresponds to the sub-case $r \le k + 1$ and $a \le u$, and then use (4.3),

$$\begin{aligned}
\sigma_3 &\le \eta_6 \left[\sum_{2 \le \rho \le r} mnu^{\rho-1} u^{2k} p^{2u} + \sum_{2 \le \rho \le r} \sum_{1 \le a \le u} mna^{\rho-1+k} u^{k-\rho+1} p^{u+a} \right] \\
&< \eta_7 \left[\sum_{2 \le \rho \le r} u^{k+\rho-1} p^u + \sum_{2 \le \rho \le r} \sum_{1 \le a} a^{\rho-1+k} u^{-\rho+1} p^a \right] \\
&< \eta_8 \left[u^{k+r-1} p^u + u^{-1} \right] \to 0.
\end{aligned}$$

Combining the results for Cases 1, 2, and 3, it follows that the upper bound (4.5) is of form $\{qmnG_k(u) + o(1)\}^r/r!$, uniformly over the set U in (4.3). This upper bound, together with the lower bound from Lemma 4, establish Lemma 5. □

It is now an easy matter to prove our main theorem about large sample approximations to the distribution of $M_k(m, n)$. We write

$$\begin{aligned}
v_k(n) = \{ &\log_{1/p}(n) + k \log_{1/p}\log_{1/p}(n) \\
&+ k \log_{1/p}(q/p) - \log_{1/p}(k!) \}.
\end{aligned} \tag{4.6}$$

The solution of the equation $nG_k(v) = 1$ is almost $v_k(n)$. Because of (4.2), $nG_k(v_k(n)) \to 1$ as $n \to \infty$. We write $\lfloor \cdot \rfloor$ for the greatest integer function, and $\beta_1(x) = x - \lfloor x \rfloor$.

THEOREM 1. *Let $m, n \to \infty$ in such a way that $\log(m)/\log(n) \to 1$. Write $\lambda \equiv \lambda(k, m, n; u) \equiv qnmG_k(u)$. Then*

$$\sup_{u \in Z} \left| P\{M_k(m, n) \ge u\} - \{1 - \exp(-\lambda)\} \right| \to 0. \tag{4.7}$$

In addition, for every fixed t, and $j = 0, 1, 2, \ldots$,

$$\sup_{\{u \in Z\colon |u - v_k(mn)| \le t\}} \left| P\{N_k(m, n; u) = j\} - \exp(-\lambda)(\lambda)^j/j! \right| \to 0. \tag{4.8}$$

PROOF. First fix t, and note that $\{mnu^k p^u\colon |u - v_k(mn)| \le t\}$ is bounded away from zero and infinity. Thus for each $\tau = 0, 1, \ldots$, the expectations of (1.4) and (1.5) can be evaluated in the limit as $m, n \to \infty$, uniformly over integers u such that $|u - v_k(mn)| \le t$, by means of Lemma 5. By taking τ sufficiently large, (4.8) is proved. For the special case $j = 0$ of (4.8), both $P\{M_k(m, n) \ge u\}$ and $\{1 - \exp[-\lambda]\}$ tend to 0 [or 1] as $u - v_k(mn)$ tends to $+\infty$ [or $-\infty$]. This establishes (4.7). □

The next theorem says that the locations (i, j) along the two sequences, at which long matches occur, are distributed approximately independently and uniformly throughout their possible range $[1, m] \times [1, n]$. Formally, we define the random measures $\xi \equiv \xi_k(m, n; u)$ on $[0,1]^2$ by

$$\xi \equiv \sum_{i<m,\, j<n} \delta_{i/m,\, j/n} R_k(m, n; u),$$

where $\delta_{x,y}$ denotes unit mass at the point (x, y). The total mass of ξ is $\xi([0,1]^2) \equiv N_k(m, n; u)$, which by Theorem 1 is approximately Poisson in distribution, with parameter λ. Theorem 1′ says that the point processes ξ are uniformly close to the corresponding constant intensity Poisson processes on $[0,1]^2$.

THEOREM 1′. *Let $B_1, \ldots, B_d$ be disjoint rectangles in $[0,1]^2$. Let a_i be the area of B_i and let $\lambda = qnmG_k(u)$. Fix t and let $j_1, \ldots, j_d$ be nonnegative integers. If m and $n \to \infty$ with $\log(m)/\log(n) \to 1$, then*

$$\sup_{u:\, |u - v_k(mn)| \le t} \left| P(\xi(B_i) = j_i \text{ for } i = 1 \text{ to } d) - \prod \exp(-\lambda a_i)(\lambda a_i)^{j_i} i/j_i! \right| \to 0.$$

PROOF. For each i, a formula like (1.5) gives upper and lower bounds on the indicator function $1(\xi(B_i) = j_i)$ in terms of sums of indicators $R_k(C)$ with index sets C of size $|C| \le 2\tau$. After multiplying out, this gives upper and lower bounds on $1(\xi(B_i) = j_i$ for $i = 1$ to $d)$ in terms of sums of indicators $R_k(C)$ with index sets C of size $|C| \le 2\tau d$. For the expectations of these bounds, Lemma 5 and the proof of Lemma 4, applied with $r = 1, \ldots, 2\tau d$, show that the combined contribution from those C corresponding to dependent events is uniformly negligible, and that the expectations for these bounds are uniformly close to what they would be if all the indicators $R_k(i, j; u)$ were mutually independent. By taking τ sufficiently large, the theorem is proved. □

In Theorem 2, we establish the most useful results about the approximate distribution of $M_k(m, n)$. Although the main result is a consequence of Theorem 1 of Ferguson (1984), we have chosen to follow the constructive approach of Gordon, Schilling and Waterman (1986) for three reasons. First, that approach makes clear the choice of centering constant $v_k(\cdot)$. Second, the representation of assertion (c) makes clear the relationship of the various limiting distributions along the appropriate subsequences. Third, the representation (c) and its use of $\beta_1(x) = x\lfloor x \rfloor$ suggests the Fourier methods which enable us to compute the expectations in assertion (e).

Before stating Theorem 2, we note as in Gordon, Schilling and Waterman (1986) that $(d/du)G_k(u)$ is the product of p^u times a polynomial in u whose leading coefficient is negative. Hence, for fixed p, there exists some constant u_0 such that $G_k(u)$ is continuous and strictly decreasing for all real $u > u_0$.

We denote by $Z_1, Z_2, \ldots$ an i.i.d. sequence of absolutely continuous random variables such that $P\{Z_j > u\} = 1 - G_k(u)$ for all real $u \ge u_0$. We denote by W

a standard extreme value random variable having cumulative distribution function $\exp(-e^{-u})$. Recall that W has expectation γ, the Euler–Mascheroni constant 0.577... and has variance $\pi^2/6$. We write $l = \log(1/p)$. The bounds on the remainder terms of assertions (e) and (f) are very similar to those obtained in Boyd (1972).

THEOREM 2. *Let $m, n \to \infty$ in such a manner that $\log(n)/\log(m) \to 1$. We may conclude that:*

(a) $P\{M_k(m, n) > u\} - P\{\lfloor \max\{Z_s: s \le qmn\} \rfloor > u\} \to 0$, *uniformly in u, for all real u.*

(b) $P\{M_k(m, n) > u\} - P\{\lfloor W/l + v_k(qnm) \rfloor > u\} \to 0$, *uniformly in u, for all real u.*

(c) $P\{M_k(m, n) > u + v_k(qnm)\} - P\{\lfloor W/l + \beta_1(v_k(qmn)) \rfloor - \beta_1(v_k(qmn)) > u\} \to 0$, *uniformly in u, for all real u.*

(d) *The random variables $\{M_k(m, n) - v_k(qnm)\}^2$ are uniformly integrable.*

(e) $E\{M_k(m, n)\} = v_k(qnm) + \gamma/l - \frac{1}{2} + r_1(m, n) + o(1)$, *where* $\theta = \pi^2/l$ *and* $|r_1(s)| \le (2\pi)^{-1}\theta^{1/2}e^{-\theta}(1 - e^{-\theta})^{-2}$.

(f) $\operatorname{Var}\{M_k(m, n)\} = \pi^2/(6l^2) + \frac{1}{12} + r_2(m, n) + o(1)$, *where* $|r_2(s)| \le 2(1.1 + 0.7\theta)\theta^{1/2}e^{-\theta}(1 - e^{-\theta})^{-3}$.

PROOF. Assertion (a) follows as an immediate consequence of Watson (1954). The random variables Z_j are independent, and $sG_k(v_k(s + u)) \to p^u$ for all real u as $s \to \infty$. Let $L_s = \max\{Z_j: j \le s\}$. We conclude from Watson (1954) that $L_s - v_k(s)$ converges in distribution to W/l and that $P\{L_{qnm} \ge t\}$ is approximated by $\exp(-qnmG_k(t))$ uniformly for all integers t.

Fix $\tau > 0$. Let $\{u_s\}$ be any sequence of integers for which $|u_{qnm} - v_k(qnm)| < \tau$. Theorem 1 tells us that $P\{M_k(m, n) \ge u_{qnm}\}$ is approximated by $1 - \exp(-qnmG_k(u_{qnm}))$, uniformly over such sequences $\{u_s\}$. For any integer t, $P\{L_{qnm} \ge t\} = P\{\lfloor L_{qnm} \rfloor \ge t\}$. Hence, $P\{M_k(m, n) \ge t\} - P\{\lfloor L_{qnm} \rfloor \ge t\} \to 0$ uniformly in integers t for which $|t| < \tau$. Because both $M_k(m, n)$ and $\lfloor L_s \rfloor$ have as support the integers, it follows that $P\{M_k(m, n) \ge u\} - P\{\lfloor L_{qnm} \rfloor \ge u\} \to 0$ uniformly over all real u.

Assertions (b) and (c) follow from (a) because $L_s - v_k(s)$ converges in distribution to W/l.

We next prove (d) which asserts the uniform integrability of the random variables $M_k(m, n)^2$. In order to simplify notation, we write $\eta, \eta_1, \eta_2, \ldots$ for constants whose exact values are not material to the proof. We also write v for $v_k(qnm)$.

Because the function $s^k p^s$ is bounded, we obtain as a consequence of (4.2) that

$$P\{M_k(m, n) \ge v + t\} \le nmG_k(v + t) \le \eta t^k p^t \quad \text{for any } t > 1. \tag{4.9}$$

Choose ζ to be a constant between 0 and 1, which we will specify later. We now use Chebyshev's inequality to bound $P\{M_k(m, n) < v - t\}$ for $1 < t < (1 - \zeta)v$. Note that $P\{M_k(m, n) < v - t\} = P\{S_k^{(1)}(m, n; v - t) = 0\}$ and that the latter probability involves a sum of indicator random variables most pairs of which are

independent. We assume without loss of generality that $m \le n$. An argument identical to that in the proof of Lemma 5 shows that

$$\operatorname{Var}\{S_k^{(1)}(m, n; v - t)\} \le nmG_k(v - t) + \eta\big(mn^2v\sigma_1 + nmv\sigma_2 + nm\sigma_3\big),$$

where the first term is attributable to the variances of nm Bernoulli random variables, and the terms σ_1, σ_2, and σ_3 are contributions of products of indicators. These products correspond to Cases 1, 2, 3 of Lemmas 2 and 3, as applied to 2-configurations which are, respectively, singly adjacent, doubly adjacent but not parallel, and both parallel and doubly adjacent.

From Lemma 3, these terms may be bounded:

$$\begin{aligned}\sigma_1 &\le \eta_1 v^{2k} p^{(3/2+\delta)(v-t)} \\ &\le \eta_2 v^{k/2}(mn)^{-(3/2+\delta)} p^{-(3/2+\delta)t}, \\ \sigma_2 &\le \eta_3 (v-t)^{2k} p^{(1+\delta)(v-t)} \\ &\le \eta_4 v^k (mn)^{-(1+\delta)} p^{-(1+\delta)t},\end{aligned}$$

and

$$\sigma_3 \le \eta_5 v^{-2}(mn)^{-1}.$$

By hypothesis, $E\{S_k^{(1)}(m, n; v - t)\} = nmG_k(v - t) \ge \eta p^{-t}$. Hence, because $\delta \le \frac{1}{4}$,

$$P\{M_k(m, n) < v - t\} \le \eta p^{t/4} \quad \text{for } 1 < t < (1 - \zeta)v. \tag{4.10}$$

Now break the sets $\{X_i\}$ and $\{Y_j\}$ into $m/(\zeta v)$ disjoint pieces to obtain:

$$\begin{aligned}P\{M_k(m, n) < \zeta v\} &< (1 - G_k(\zeta v))^{m/(\zeta v)} \\ &< \exp(-mG_k(\zeta v)/[\zeta v]) \\ &< \exp\big(-mv^{-1}(nm)^{-\zeta}/\eta\big) < \exp\big(-m^{1-3\zeta}/\eta\big).\end{aligned} \tag{4.11}$$

The final inequality follows because $\log(n)/\log(m) \to 1$. Now choose and fix $\zeta = \frac{1}{4}$. Combining (4.9), (4.10), and (4.11) gives

$$E\Big\{|M_k(m, n) - v|^2 1\big(|M_k(m, n) - v| > \tau\big)\Big\} < v^2 p^{m^{1/4}/\eta} + \sum_{\{t>\tau\}} \eta t^2 p^{t/4}$$

and so we have established the uniform integrability of $|M_k(m, n) - v_k(qnm)|^2$.

Assertion (e) follows exactly as in Gordon, Schilling and Waterman (1986), because assertion (d) lets us approximate the mean and variance of $M_k(m, n)$ by the corresponding moments of $\eta + \lfloor W/l - \eta \rfloor$. □

5. The relative growth of the two sequence lengths must be constrained. Here is a calculation that shows why the hypothesis "$\log(m)/\log(n) \to 1$" is needed to state Theorems 1 and 2 in a form with minimal conditions on μ. This calculation shows that if $(\log m)/(\log n)$ does not converge to 1, then a suitable choice of μ makes the second moment of $N = N_0(m, n; u)$ blow up. $N_0(m, n; u)$, the count of witnesses to matches having length at least u, was defined in (1.2).

Fix $\theta \in (0,1)$ and take $m, n \to \infty$ with $\log(m)/\log(n) \to \theta$. Consider only configurations C of the form $\{(i, r; u), (i, s; u)\}$ with $r \neq s$, $i \leq m$, and $r, s \leq n$. The number of these configurations is $\binom{n}{2}m$. Denote by $\lceil \cdot \rceil$ the least integer function. Take $u = \lceil \log_{1/p}((1-p)mn) \rceil$ so that $c \equiv u - \log_{1/p}((1-p)mn) \in [0, 1)$ and $EN = p^c \in (p, 1]$. Note that $\log_{1/p}(\binom{n}{2}m) \sim \log_{1/p}(mn^2) \sim (\log_{1/p}(mn))(2+\theta)/(1+\theta) \sim u(2+\theta)/(1+\theta)$. We have $R(C) = 1(X_t \neq Y_r)1(X_t \neq Y_s)1(X_{i+t} = Y_{j+t} = Y_{k+t}$ for $t = 1$ to $u)$, and $ER(C) = (\Sigma_a \mu_a(1-\mu_a)^2)(p_3)^u$ where $p_3 = \Sigma_a(\mu_a)^3$. Thus $\log(ER(C)) \sim u \log(p_3)$. The combined contribution to $ES^{(2)}$ from these $\binom{n}{2}m$ configurations is $\binom{n}{2}mER(C)$ with $\log_{1/p}(\binom{n}{2}mER(C)) \sim u[(2+\theta)/(1+\theta) - \log_p(p_3)]$. Jensen's inequality says that $p_3 \geq p^{3/2}$.

Given any $\theta < 1$, if $\log_p(p_3)$ is sufficiently close to $\frac{3}{2}$, then the coefficient of u in the expression for $\log_{1/p}(\binom{n}{2}mER(C))$ is positive. In that case $ES^{(2)} \to \infty$.

For fixed $p \in (0,1)$, there are examples of μ having $\Sigma_{a \in S}(\mu_a)^2 = p$ and $\log_p(p_3)$ arbitrarily close to Jensen's bound, $\frac{3}{2}$, although the alphabet S may also have to be large. However, for any particular μ, the condition "$(\log m)/(\log n) \to 1$" might not be necessary for the conclusion of Theorems 1 and 2 to hold.

Consider the above argument in more detail. Let $\theta_h \equiv [2 - \log_p(p_3)]/[\log_p(p_3) - 1]$. Elementary manipulation shows that $\theta_h \geq 0$, with equality iff μ is uniform. In the case μ is nonuniform, and m and $n \to \infty$ with $(\log m)/(\log n) \to \theta < \theta_h$, the above argument shows that the number N of witnesses to a match of length at least $c + \log_{1/p}((1-p)mn)$ is bounded in L^1 and unbounded in L^2; however, this argument fails to resolve whether the family of random variable $\{M(m,n) - \log_{1/p}(mn)\}$ is tight. We conjecture that for each μ there exist centering constants $u(m,n)$, namely $u(m,n) = E(M(m,n))$, such that the family $\{M(m,n) - u(m,n)\colon m, n \in Z^+\}$ is tight.

In this paper, we have analyzed $M(m,n)$ by the position at which the match appears. In contrast, an analysis of $M(m,n)$ by the pattern which is matched is carried out in Arratia and Waterman (1985b). In order to describe the result of that analysis by pattern, consider the distribution α defined by $\alpha_a = P(X_1 = a | X_1 = Y_1) = (\mu_a)^2/p$, and note that $\mu = \alpha$ iff μ is uniform. Let $H(\alpha, \mu)$ be the relative entropy, $H(\alpha,\mu) = \Sigma_{a \in S}\alpha_a \log(\alpha_a/\mu_a)$, so that $H(\alpha,\mu) \geq 0$, with equality iff μ is uniform. Formula (1.1) in the introduction to this paper gives necessary and sufficient conditions for $M(m,n)/(\log_{1/p}(mn)) \to_P 1$ when m grows like n^θ; the critical value there is given by $\theta_{\text{cr}} = H(\alpha,\mu)/(\log(1/p) - H(\alpha,\mu))$.

6. Boundary effects are negligible. To be strictly accurate, we note that $M_k(m,n)$, defined in (1.2), is not really the length of the longest k-interrupted matching between $X_1 \cdots X_m$ and $Y_1 \cdots Y_n$. Instead, $M_k(m,n)$ is the maximum of an m by n block of random variables from a two-parameter stationary family, and the event $\{M_k(m,n) \geq u\}$ involves $m + n + 2u$ letters, X_0 through X_{m+u-1} and Y_0 through Y_{n+u-1}. Given actual data, $X_1 \cdots X_m$ and $Y_1 \cdots Y_n$, the length $M_k^d(m,n)$ of the longest k-interrupted matching is defined below; note that it is

a nontrivial function of all $m + n$ letters. Here for comparison are the two definitions:

$$M_k^d \equiv M_k^d(m, n) \equiv \max\{u: X_{i+1} \cdots X_{i+u} \text{ matches } Y_{j+1} \cdots Y_{j+u} \text{ with at most } k \text{ mismatches, for some } 0 \le i \le m - u, 0 \le j \le n - u\},$$

$$M_k \equiv M_k(m, n) \equiv \max\{u: X_{i+1} \cdots X_{i+u} \text{ matches } Y_{j+1} \cdots Y_{j+u} \text{ with at most } k \text{ mismatches, and } X_i \ne Y_j, \text{ for some } 0 \le i < m, 0 \le j < n\}.$$

In Lemma 6 we show that $M_k^d(m, n)$ is on the order of $\log(mn)$, provided that each sequence is at least this long. Theorem 3 essentially says that boundary effects are negligible if and only if each sequence is many times longer than $M_k^d(m, n)$.

LEMMA 6. *Suppose the letters* $\{X_i\}$ *and* $\{Y_i\}$ *are i.i.d.* (μ) *from a finite alphabet* S. *Let* $H = -\sum_{a \in S} \mu_a \log(\mu_a) > 0$ *be the entropy of* μ, *and let* $p = \sum_{a \in S} (\mu_a)^2$. *Fix* $k > 0$. *For any* $\varepsilon > 0$, *as* m *and* $n \to \infty$,

$$P\big(\min(m, n, (1 - \varepsilon)\log(\max(m, n))/H) \le M_k^d(m, n) < (1 + \varepsilon)\log_{1/p}(mn)\big) \to 1.$$

PROOF. The upper bound comes from Chebyshev's inequality, as discussed following formula (1.5), and using (4.2) in case $k > 0$. For the lower bound, it suffices to consider the case $k = 0$, since $M^d(m, n) \equiv M_0^d(m, n) \le M_k^d(m, n)$ for all m, n, and k. Without loss of generality we may assume that $m \le n$.

Let $t \equiv \min(m, \lfloor (1 - \varepsilon)\log(n)/H \rfloor)$. Write W for the random word $W \equiv X_1 \cdots X_t \in S^t$. We will show that with probability tending to 1, the word W appears somewhere in $Y_1 \cdots Y_n$, so that $M^d(m, n) \ge t$. Since $(1/t)\log(\mu^t(W)) \to -H$ in probability, $P(\mu^t(W) > \exp(-(1 + \varepsilon)Ht)) \to 1$. For large n, $\exp(-(1 + \varepsilon)Ht) \ge \exp(-(1 - \varepsilon^2)\log n) > t^2/n$. Now in n/t independent trials, each with success probability $> t^2/n$, the probability of no successes is bounded above by $(1 - t^2/n)^{n/t} < e^{-t} \to 0$. Hence

$$P\big\{Y_{jt+1} \cdots Y_{jt+t} = W \quad \text{for some} \quad 0 \le j < n/t | \mu^t(W) > t^2/n\big\} \to 1,$$

and the lower bound is proved. □

THEOREM 3. *Suppose the letters* $\{X_i\}$ *and* $\{Y_i\}$ *are i.i.d.* (μ) *from a finite alphabet* S, *with* $0 < \mu_a < 1$ *for all* $a \in S$. *Fix* $k \ge 0$, *and let* m *and* $n \to \infty$ *along a given sequence* (m_i, n_i). *The following conditions are equivalent*:

(a) $(\log n)/m \to 0$ *and* $(\log m)/n \to 0$,
(b) $P(M_k^d(m, n) = M_k(m, n)) \to 1$.

PROOF. Fix any $\varepsilon > 0$, and let $t \equiv t(m, n) \equiv \lfloor (1 + \varepsilon)\log_{1/p}(mn) \rfloor$, so that by Lemma 6, $P(M_k^d \ge t) \to 0$. We place the letters $X_0 X_1 \cdots X_{m+t-1}$ around a circle of length $m + t$, and similarly place the letters $Y_0 Y_1 \cdots Y_{n+t-1}$ around a

circle of length $n + t$. Let L_{ij} be the length of the k-interrupted match following position (i, j) on this pair of circles: $L_{ij} \equiv \max\{u\colon X_{i+1} \cdots X_{i+u}$ matches $Y_{j+1} \cdots Y_{j+u}$ with at most k mismatches$\}$, where the indices for X are taken modulo $(m + t)$ and those for Y are taken modulo $(n + t)$.

We show that (a) implies (b). Let $A \equiv Z^2 \cap [0, m + t) \times [0, n + t)$, and let $B \equiv Z^2 \cap (t, m - t) \times (t, n - t)$. Let $E \equiv E(m, n)$ be the event $\{\max_{(i,j)\in B} L_{ij} < \max_{(i,j)\in A-B} L_{ij}\}$. Since the two-parameter family (L_{ij}) is stationary, $P(E) \le |A - B|/|A|$, and thus condition (a) implies that $P(E) \to 0$. Note that $\{M_k^d \ne M_k\} \subset E \cup \{M_k^d \ge t\}$, so that $P(M_k^d \ne M_k) \to 0$, which is condition (b).

Now we assume that (a) fails, and show that (b) fails. Without loss of generality we may assume that $m \le n$ and $(\log n)/m$ is bounded away from 0. We need to show that $P(M_k^d \ne M_k)$ is bounded away from zero. Let $\tilde{L}_{ij} = (L_{ij})1(X_i \ne Y_j)$, still taking the index i modulo $(m + t)$ and the index j modulo $(n + t)$. Let $s = \min(m - 1, (1 - \varepsilon)\log(n)/H)$. Let D be the event $\{s < M_k^d < t\}$. By Lemma 6, $P(D) \to 1$. Let $I \equiv Z^2 \cap [1, m) \times [1, n)$, and let $J \equiv Z^2 \cap [m - s, m) \times [1, n)$. Let F be the event $\{\max_{I-J} \tilde{L}_{ij} \le \max_J \tilde{L}_{ij}\}$. Since the two-parameter family $(\tilde{L}_{ij})$ is stationary, $P(F) \ge |J|/|I| = s/m$. By the assumptions that (a) fails and that $m \le n$, s/m is bounded away from zero, and hence $P(F)$ is bounded away from zero.

Let G be the event $\{\max_{(i,j)\in I-J} \tilde{L}_{ij} < \max_{(i,j)\in J} \tilde{L}_{ij}\}$. Note that $G \cap D \subset \{M_k^d \ne M_k\}$. Since $P(D) \to 1$, we need only show that $P(G)$ is bounded away from zero in order to conclude that (b) fails. Let $c \equiv \min_{a\in S} \mu_a/(1 - \mu_a)$, so that $c \in (0, 1]$. Consider the map from $F \cap G^c \cap D$ into G defined informally as follows. Let $u = \max_{I-J} \tilde{L}_{ij} = \max_J \tilde{L}_{ij}$, and find the match $X_{i+1} \cdots X_{i+u} = Y_{j+1} \cdots Y_{j+u}$ with $(i, j) \in J$ which minimizes i and j. Then change the letter X_{i+u+1} so that it agrees with Y_{j+u+1}. This map f is many-to-one, but $P(B) \ge cP(f^{-1}(B))$ for any $B \subset \text{range}(f)$. Hence $P(G) \ge c(P(F) - P(G) - P(D^c))$, so $\liminf(P(G)/P(F)) \ge c/(1 + c)$. This shows that $P(G)$ is bounded away from zero. □

7. A biological example. The complete DNA sequence of the virus lambda can be represented as a word consisting of 48,502 letters from the alphabet $\{A, C, G, T\}$, with respective relative frequencies $\{0.2543, 0.2343, 0.2643, 0.2471\}$. See Sanger et al. (1982). The virus is a bacteriophage that infects the bacterium *Escherichia coli*. As of mid-1984, the lambda sequence is the longest completely known genetic sequence.

Because viral sequences do not seem to contain repeated or redundant regions, we do not expect disjoint segments to contain long matching regions. As a test of the utility of the asymptotic theory, we started at the beginning of the lambda sequence and took disjoint segments of length $2^7, 2^8, \ldots, 2^{12}$, repeating the segmentation for 6 complete cycles. In all, 48,378 of the 48,502 letters were grouped into 36 disjoint segments.

Each segment of length 2^m was matched against every other segment of length 2^m, and the longest k-interrupted match run was determined for each value of $k = 0, 1, 2, \ldots, 6$. This matching scheme accounted for 630 determinations of longest match runs $M_k(2^m, 2^m)$: $\binom{6}{2}$ pairings with 6 levels of m by 7 levels of k.

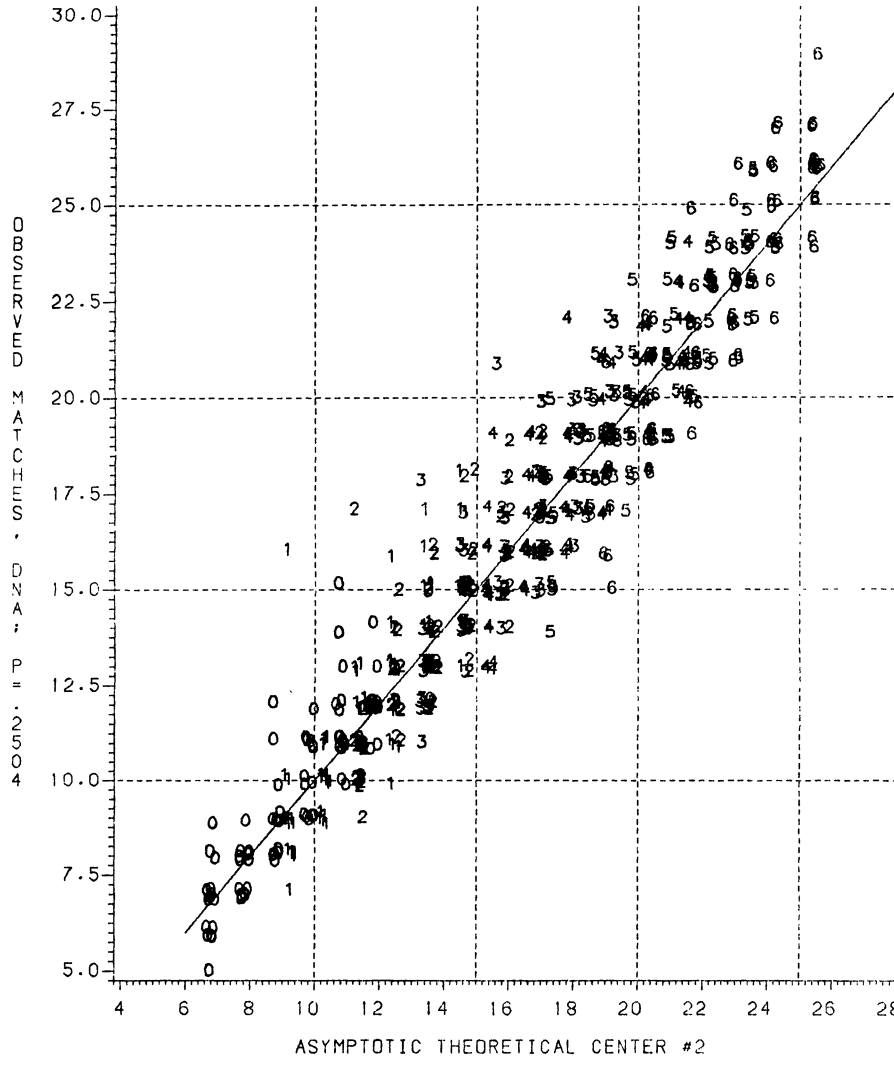

FIG. 1.

For the value of $p = 0.2505 = 0.2543^2 + 0.2343^2 + 0.2643^2 + 0.2471^2$, we computed the exact solution $u_k(m)$ to the equation

$$2^{-2m} = (1 - p)G_k(u), \tag{7.1}$$

where $G_k(u)$ is defined in (4.1). If the order of letters in the lambda sequence were well approximated by an i.i.d. scheme with the cited frequencies, we would expect as a consequence of Theorems 1 and 2 to observe longest match runs consisting of about $u_k(m) + \gamma/l - \frac{1}{2}$ letters.

Presented in Figure 1 is a scatterplot of the 630 actual values of longest k-interrupted match runs plotted against their asymptotic theoretical values

TABLE 1
Comparison of two asymptotically equivalent forms for centering constant

m	7	7	7	10	10	10	12	12	12
k	0	3	6	0	3	6	0	3	6
$v_k(q2^{2m})$	6.8	12.0	15.1	9.8	15.8	19.7	11.8	18.2	22.5
$u_k(q2^{2m})$	6.8	13.4	19.0	9.8	16.9	22.9	11.8	19.2	25.4
$u_k - v_k$	0.0	−1.4	−3.9	0.0	−1.1	−3.2	0.0	−1.0	−2.9

$u_k(m) + \gamma/l - \frac{1}{2}$. The digits $0, 1, \ldots, 6$ signify the value of k for the particular k-interrupted run. Because the longest match run must take on integer values, the points themselves are randomly jittered by a small uniformly distributed value in order to display the density of deviations. The data appear well centered about the theoretical value.

The reader will note that we have used u_k, and not the asymptotically equivalent value v_k defined in (4.6) and used in Theorem 2. That such accommodation to small sample properties is necessary is evident from Table 1 (and our initial less satisfactory attempts at plotting the longest match run against v_k).

Given the remarkable fit of predicted location, the next natural question is the quality of approximation for the predicted dispersion. Here the answer is less satisfactory. Spread is greater than predicted by the asymptotic theory, especially for larger values of k. We believe that a large portion of the lack of fit is attributable to slowness of convergence. This suspicion is supported by the data from the bacteriophage lambda, and from a simulation in which the segments' letters were truly generated by an i.i.d. mechanism, with uniform distribution upon an alphabet of four letters.

As a consequence of Theorem 2(c), $P\{M_k(2^m, 2^m) - \lfloor u_k(q2^{2m}) \rfloor = j\}$ should be well approximated by $\exp(-\eta p^{j+1}) - \exp(-\eta p^j)$, where $\log_{1/p}\eta = b_1(u_k(q2^{2m}))$, the fractional part of $u_k(q2^{2m})$. Presented in Table 2 is a comparison of

TABLE 2
Observed and expected deviations from asymptotic centers for the longest match run

j	lambda (p = 0.2505) observed	lambda (p = 0.2505) expected	simulation (p = 0.2500) observed	simulation (p = 0.2500) expected
−4	1	0.0	0	0.0
−3	3	0.0	2	0.0
−2	26	1.6	27	1.2
−1	103	93.1	152	87.1
0	192	277.9	217	277.3
1	173	177.7	148	182.1
2	79	58.6	60	60.5
3	32	15.8	15	16.3
4	15	4.0	5	9.1
5	3	1.0	3	1.0
6	2	0.3	1	0.3
7	1	0.1	0	0.1

deviations from $\lfloor u_k(q2^{2m}) \rfloor$. The columns for expected values are computed as discussed above for the biological sequence with $p = 0.2505$, and for the simulated biological sequence with $p = 0.2500$. The "observed" columns correspond to the biological data plotted in Figure 1, and for simulated data from a uniform four-letter alphabet with segments of identical length.

We tentatively conclude from this simulation and from other simulation work that the degree of lack of fit from asymptotic prediction could be attributable to small sample properties of the distribution of the maximum k-interrupted match run length. A greater understanding of rates of convergence to the integerized extreme value distribution is clearly needed.

Acknowledgment. The authors thank S. Karlin for calling to our attention the question of the locations of long matches, addressed here in Theorem 1′.

REFERENCES

ANDERSEN, J. S. et al. (1984). *Nucleotide Sequences 1984: A Compilation from the GenBank and EMBL Data Libraries* **1, 2.** IRL Press, Oxford.

ARRATIA, R. and WATERMAN, M. S. (1985a). An Erdös–Rényi law with shifts. *Adv. in Math.* **55** 13–23.

ARRATIA, R. and WATERMAN, M. S. (1985b). Critical phenomena in sequence matching. *Ann. Probab.* **13** 1236–1249.

BONFERRONI, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubbl. Istit. Sup. Sci. Econ. Commerciali Firenze* **8** 1–62.

BOYD, D. W. (1972). Losing runs in Bernoulli trials. Unpublished manuscript.

DARLING, R. and WATERMAN, M. S. (1985). Matching rectangles in d-dimensions: algorithms and laws of large numbers. *Adv. in Math.* **55** 1–12.

DOOLITTLE, R. F. et al. (1983). Simian sarcoma viruses one gene v-sis is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science* **221** 275–276.

ERDÖS, P. and RÉVÉSZ, P. (1975). On the length of the longest head-run. Topics in Information Theory. *Colloquia Math. Soc. J. Bolyai* **16** 219–228. Keszthely, Hungary.

FERGUSON, T. S. (1984). On the distribution of Max and Mex. Preprint.

GORDON, L., SCHILLING, M. F. and WATERMAN, M. S. (1986). An extreme value theory for long head runs. *Probab. Theory Rel. Fields* **72** 279–288.

GUIBAS, L. J. and ODDLYZKO, A. M. (1980). Long repetitive patterns in random sequences. *Z. Wahrsch. verw. Gebiete* **53** 241–262.

GUIBAS, L. J. and ODLYZKO, A. M. (1981). Periods in strings. *J. Combin. Theory Ser. A* **30** 19–42.

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1934). *Inequalities.* Cambridge University Press.

KARLIN, S., GHANDOUR, G., OST, F., TAVARE, S. and KORN, L. J. (1983). New approaches for computer analysis of nucleic acid sequences. *Proc. Nat. Acad. Sci. U.S.A.* **80** 5660–5664.

LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

SANGER, F., COULSON, A. R., HONG, G. F., HILL, D. F. and PETERSON, G. B. (1982). Nucleotide sequence of bacteriophage λ DNA. *J. Molecular Biol.* **162** 729–773.

SMITH, T. F., WATERMAN, M. S. and BURKS, C. (1985). The statistical distribution of nucleic acid similarities. *Nucleic Acids Research* **13** 645–656.

WATERMAN, M. S. (1984). General methods of sequence comparison. *Bull. Math. Biol.* **46** 473–500.

WATSON, G. S. (1954). Extreme values in samples from m-dependent stationary stochastic processes. *Ann. Math. Statist.* **25** 798–800.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089