Reprinted from ADVANCES IN MATHEMATICS
All Rights Reserved by Academic Press, New York and London

Vol. 55, No. 1, January 1985
Printed in Belgium

An Erdös–Rényi Law with Shifts

RICHARD ARRATIA*

AND

MICHAEL S. WATERMAN†

*Department of Mathematics, University of Southern California,
Los Angeles, California 90089-1113*

Motivated by the comparison of DNA sequences, a generalization is given of the result of Erdös and Rényi on the length R_n of the longest run of heads in the first n tosses of a coin. Consider two sequences, $X_1X_2 \cdots X_n$ and $Y_1Y_2 \cdots Y_n$. The length of the longest matching consecutive subsequence, allowing shifts, is $M_n \equiv \max\{m: X_{i+k} = Y_{j+k} \text{ for } k = 1 \text{ to } m, \text{ for some } 0 \leqslant i, j \leqslant n-m\}$. Suppose that all the "letters" are independent and identically distributed. The length of the longest match without shifts has the same distribution as R_n, the length of the longest head run for a biased coin with $p = P(X_i = Y_i)$, described by the Erdös–Rényi law: $P(\lim_{n\to\infty} R_n/\log_{1/p}(n) = 1) = 1$. For matching with shifts, our result is: $P(\lim_{n\to\infty} M_n/\log_{1/p}(n) = 2) = 1$. Loosely speaking, allowing shifts doubles the length of the longest match. The case of Markov chains is also handled. © 1985 Academic Press, Inc.

1. Introduction

Erdös and Rényi [3] considered the length R_n of the longest run of heads in the first n tosses of a coin. One result, from the book of Rényi [8], is that

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

Here, $p \in (0, 1)$ is the probability of heads for a single toss. More refined estimates on the behavior of R_n are given in Erdös and Révész [4] and Guibas and Odlyzko [5]; a survey appears in Révész [9].

In this paper, we consider a problem motivated by the comparison of DNA sequences, which are taken from an alphabet of four letters:

* This author supported by NSF Grant MCS-8301960.
† This author supported by a grant from the System Development Foundation.

0001-8708/85 $7.50

A = adenine, G = guanine, C = cytosine, and T = thymine. Given two such sequences, such as

$$X_1X_2\cdots X_n\cdots = A\ G\ T\ C\ T\ G\ A\ A\ G\ C\ A\ C\ A\ A\ G\ T\ G\ T\cdots$$
$$Y_1Y_2\cdots Y_n\cdots = T\ A\ T\ C\ T\ T\ T\ G\ A\ A\ G\ C\ C\ C\ A\ T\ T\ T\cdots$$

the degree of relationship between the sequences may be measured by the length of the longest matching consecutive subsequence. For two sequences of length n, this length is

$$M \equiv M_n \equiv \max\{m\colon X_{i+k} = Y_{j+k} \text{ for } k = 1 \text{ to } m, \text{ for some } 0 \leqslant i, j \leqslant n-m\}. \tag{2}$$

Smith and Waterman [10] give an efficient algorithm for finding such matches, which allow all possible shifts of one sequence relative to the other. The results of this paper were suggested by a data analysis of DNA sequences by Smith, Waterman, and Burks [11].

Suppose that all the "letters" $X_1, X_2, \ldots, Y_1, Y_2, \ldots$, are independent and identically distributed. If no shifting between the two sequences is allowed, then the data may be reduced to a single sequence of heads and tails, with a head reported for the ith toss when $X_i = Y_i$. Thus the length of the longest match without shifts is like R_n, the length of the longest head run for a biased coin with $p = P(X_i = Y_i)$, described by the Erdös–Rényi law, Eq. (1). For matching with shifts, Theorem 2 below implies that

$$P(\lim_{n\to\infty} M_n/\log_{1/p}(n) = 2) = 1. \tag{3}$$

Loosely speaking, allowing shifts doubles the length of the longest match. To see that M_n should grow like $2\log_{1/p}(n)$ note that a match of length $m = 2\log_{1/p}(n)$ starting from X_i and Y_j occurs with probability $p^m \approx n^{-2}$, which balances against $\approx n^2$ choices for (i, j).

For the example above, with $n = 18$, we have $R_n = 3$, corresponding to the matching subsequence "$T\ C\ T$" in the third position, while $M_n = 6$, corresponding to "$T\ G\ A\ A\ G\ C$" starting at $i = 5$ in the X's and $j = 7$ in the Y's.

2. The Longest Matching Subsequence

For the remainder of this paper, assume that the two sequences $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ have the same distribution and are independent of each other. Let $p = P(X_1 = Y_1)$; any unqualified log will denote $\log_{1/p}$.

In the next three theorems, three different structures are taken for the distribution of the individual sequences. In all three cases, it is easy to get an upper bound on M:

$$\exists c \text{ such that } \forall b, n, \qquad P(M_n - 2\log_{1/p}(n) > b) < cp^b. \tag{4}$$

In Theorem 1, we get a lower bound of the same form, under the assumption that all letters are independent and *uniformly distributed* over a finite alphabet: for some constant $c > 0$, for all n and $b > 0$, $P(|M_n - 2\log_{1/p}(n)| > b) < cp^b$. In Theorem 2, we assume that the letters are i.i.d., and show that

$$\forall \varepsilon > 0, \quad P((M - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n) \in (-2-\varepsilon, \varepsilon)) \to 1 \quad \text{as } n \to \infty. \tag{5}$$

In Theorem 3, assuming that each sequence is a Markov chain with transition matrix $[p_{ij}]$, the result is (5) again, using $p =$ the largest eigenvalue of the matrix $[(p_{ij})^2]$.

In each of the above cases, it seems reasonable to conjecture that the family of random variables $\{M - 2\log_{1/p}(n)\}$ is tight, with limits in distribution which are derived from the extreme value distribution by scaling and rounding. The motivation for this conjecture is the following result for R_n, the longest head run in n tosses of a p-biased coin, taken from Guibas and Odlyzko [5] and Louis Gordon (private communication): with $q = 1 - p$:

If $n \to \infty$ along a subsequence such that $((\log_{1/p}(qn)) \bmod 1) \to \alpha \in (0, 1)$, then

$$R_n - \lfloor \log_{1/p}(qn) \rfloor \to \lfloor Z/\ln(1/p) + \alpha \rfloor, \qquad \text{where } P(Z < c) = \exp(-e^{-c}).$$

THEOREM 1. *Let $X_1, X_2, ..., Y_1, Y_2, ...,$ be independent letters, each uniformly distributed over the finite alphabet $\{1, 2, ..., a\}$, so that $p \equiv P(X_1 = Y_1) = 1/a \leqslant \frac{1}{2}$. Then for all $b \in R$, and for $n = 1, 2, ...,$*

$$P(M - 2\log_{1/p}(n) \geqslant b) \leqslant p^b, \tag{6}$$

and

$$P(M - 2\log_{1/p}(n) < b) \leqslant (1+p)/(1-p)(1 - (2/n)\log_{1/p}(n))^{-2} p^{-b-1}. \tag{7}$$

Proof. Fix a parameter b and let

$$m \equiv m(n, n) \equiv \lceil 2\log n + b \rceil.$$

Let

$$A_{ij} \equiv A_{ij}(m) \equiv \{X_{i+k} = Y_{j+k} \text{ for } k = 1 \text{ to } m\}$$

so that

$$n^{-2}p^{b+1} < P(A_{ij}) = p^m \leqslant n^{-2}p^b.$$

Consider the random variable

$$T \equiv T(n, m) \equiv \sum_{0 \leqslant i,j \leqslant n-m} 1(A_{ij})$$

so that, as events, $\{M \geqslant m\} = \{T > 0\}$. The upper bound (6) follows from

$$P(M \geqslant m) = P(T > 0) \leqslant ET = \sum P(A_{ij}) = (n-m+1)^2 p^m \leqslant p^b. \quad (8)$$

To establish (7), we need an upper bound on var(T). Any pair of events A_{ij} and $A_{i'j'}$, for which both $|i-i'| \geqslant m$ and $|j-j'| \geqslant m$, is independent, just because independent blocks of letters are involved. Under the assumption that all letters are equally likely, there are several more independent pairs A_{ij} and $A_{i'j'}$. Those pairs for which $|i-i'| < m$ and $|j-j'| \geqslant m$, or $|i-i'| \geqslant m$ and $|j-j'| < m$, are independent; this follows easily by conditioning on the entire sequence $X_1 \cdots X_n$, or on $Y_1 \cdots Y_n$. For those pairs in which $|i-i'| < m$, $|j-j'| < m$ and $k \equiv (i-i') - (j-j') \neq 0$, it requires calculation to see that the pair of events is independent; this result may seem surprising when k is small, because the event $A_{ij} \cap A_{i'j'}$ in this case implies the existence of a repeating subpattern of length k.

In the expansion

$$\operatorname{var}(T) = \sum_{0 \leqslant i,j,i',j' \leqslant n-m} \operatorname{cov}(1(A_{ij}), 1(A_{i'j'})),$$

the only nonzero terms are those for which $k \equiv i' - i = j' - j \in (-m, m)$. For these terms,

$$\operatorname{cov}(1(A_{ij}), 1(A_{i+k,j+k})) = p^{m+|k|} - p^{2m}.$$

For each k, there are at most $(n-m+1)^2$ choices for (i, j), so that

$$\begin{aligned}&(n-m+1)^{-2} \operatorname{var}(T)\\ &\quad \leqslant \sum_{k \in (-m,m)} (p^{m+|k|} - p^{2m}) < \sum p^{m+|k|} < (1+p)/(1-p)\, p^m,\end{aligned}$$

and hence

$$\operatorname{var}(T) < (1+p)/(1-p)\, ET.$$

We have

$$P(M < m) = P(T = 0) \leqslant \operatorname{var}(T)/(ET)^2 < ((1+p)/(1-p))/ET$$
$$< ((1+p)/(1-p))(n-m+1)^{-2}n^2p^{-b-1} \sim ((1+p)/(1-p))p^{-b-1}.$$

This establishes relation (7). ▮

THEOREM 2. *Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots,$ be independent and identically distributed and let $0 < p \equiv P(X_1 = Y_1) < 1$. Then (6) holds, and for any $\varepsilon > 0$,*

$$P((M - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n) \in (-4-\varepsilon, 1+\varepsilon) \text{ eventually}) = 1.$$

Proof. The argument leading up to (8) in Theorem 1 remains valid under the present hypothesis, so that (6) holds here. Fix a parameter b and let

$$m \equiv m(n, b) \equiv \lceil 2\log n + b \log\log n \rceil.$$

From (6), we get $P(M \geqslant m) \leqslant p^{b\log\log n} = (\log n)^{-b}$, which tends to zero as $n \to \infty$, provided that $b > 0$. To get an almost sure result via the Borel–Cantelli lemma, exploit the fact that $\forall\omega$, $M = M(n, \omega)$ is nondecreasing in n, while $m(n) \sim 2\log n$ increases slowly. Along the subsequence $n \equiv n(k) \equiv \lceil (1/p)^k \rceil$ we have $\forall b > 1$, P ($M \geqslant m$ for infinitely many k) $= 0$, and hence $\forall b > 1$, P ($M \geqslant m$ i.o.) $= 0$.

To establish lower bounds on M, consider matches between blocks of length m, each starting with an offset that is a multiple of m. Let

$$B_{ij} \equiv A_{mi,mj} \equiv \{X_{mi+k} = Y_{mj+k} \text{ for } k = 1 \text{ to } m\},$$

and let S be the sum of the indicators of these events:

$$S \equiv S(n, m) \equiv \sum_{0 \leqslant i,j \leqslant (n/m)-1} 1(B_{ij}).$$

As events, $\{M < m\} \subset \{S = 0\}$. Our choice of m yields $pn^{-2}(\log n)^{-b} < p^m = P(B_{ij})$, so that

$$ES \sim (n/m)^2 p^m > p(\log n)^{-2-b}.$$

In the expansion for $\operatorname{var}(S)$ as a sum of covariances of indicator functions, there are $\sim(n/m)^2$ diagonal terms, whose combined contribution is $<ES$, and there are $\sim(n/m)^4$ terms which are all zero, corresponding to the independent pairs of events B_{ij} and B_{kl} with $i \neq k$ and $j \neq l$. Finally, there are $\sim 2(n/m)^3$ nontrivial terms for pairs of events such as B_{ij} and B_{ik} with $j \neq k$, or B_{ij} and B_{kj} with $i \neq k$. Let q_a be the weights of the atoms in the distribution of X_1,

so that $p = \sum_a (q_a)^2$. We have, using Holder's inequality (Hardy, Littlewood, and Polya [6, Formula 2.10.3]) that for $j \neq k$,

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

so that the combined contribution to $\operatorname{var}(S)$ from these $\sim 2(n/m)^3$ nontrivial pairs is $< 2(ES)^{3/2}$. Combining these estimates, we have $\operatorname{var}(S) < ES + 2(ES)^{3/2}$. Using Chebyshev's inequality,

$$\begin{aligned} P(M < m) \leqslant P\,(S = 0) \leqslant \operatorname{var}(S)/(ES)^2 &< 1/ES + 2(ES)^{-1/2} \\ &< 2p^{1/2}(\log n)^{1+b/2}. \end{aligned}$$

Thus, to conclude that $P(M < m) \to 0$, we need $b < -2$. To get an almost sure result, consider again the skeleton $n \equiv n(k) \equiv \lceil (1/p)^k \rceil$; it follows that if $b < -4$, then $P(M < m \text{ i.o.}) = 0$. ∎

THEOREM 3. *Let $X_1, X_2, \ldots,$ and $Y_1, Y_2, \ldots,$ be independent Markov chains on a finite alphabet S, each irreducible and aperiodic, with transition probabilities $[p_{ij}]_{i,j \in S}$. Let $p \in (0, 1)$ be the largest eigenvalue of the substochastic matrix $[(p_{ij})^2]$, so that there exist constants $0 < c_0 < c_1 < \infty$, such that with any initial state for $X_1 = Y_1$, and for all n,*

$$c_0\, p^n \leqslant P(X_k = Y_k \text{ for } k = 1 \text{ to } n) \leqslant c_1\, p^n. \tag{9}$$

Then

$$P(M - 2\log_{1/p}(n) \geqslant b) \leqslant c_1\, p^b, \tag{10}$$

and for any $\varepsilon > 0$,

$$P((M - 2\log_{1/p}(n))/\log_{1/p}\log_{1/p}(n) \in (-4 - \varepsilon, 1 + \varepsilon) \text{ eventually}) = 1.$$

Proof. The argument leading up to (8) in Theorem 1 is valid here, if the factor c_1 is inserted appropriately, so that (10) holds. Fix a parameter b and let

$$m \equiv m(n, b) \equiv \lceil 2 \log n + b \log\log n \rceil.$$

From (10), $P(M > n) < c_1(\log n)^{-b}$. Consider the skeleton of times $n \equiv n(k) \equiv \lceil (1/p)^k \rceil$, and use the Borel–Cantelli lemma to conclude that if $b > 1$, then $P(M \geqslant m \text{ i.o.}) = 0$.

To give a lower bound on M, the strategy is to find independent blocks of letters within the two sequences $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$. These blocks must have length at least m, and there must be at least $c\; n/m$ such blocks among the first n letters of each process, for some constant $c > 0$. We apply Doeblin's method (Chung [1]). Fix any letter $a \in S$ and let $\tau(X, k)$ be the time of the kth visit to state a by the process X,

$$\tau(X,0) \equiv 0; \qquad \tau(X, k+1) \equiv \min\{i > \tau(X, k) \colon X_i = a\}, \qquad \text{for} \qquad k = 0, 1, \ldots .$$

Let $\tau(Y, k)$ be the time of the kth visit to a by the process Y. Note that the excursions $(X_i; \tau(X, k) \leqslant i < \tau(X, k+1))$ for $k = 1, 2, \ldots,$ are mutually independent, and in each excursion the letter a appears exactly once, in the first position.

Let $A_{ij} \equiv A_{ij}(m) \equiv \{X_{i+k} = Y_{j+k} \text{ for } k = 1 \text{ to } m\}$, and for $i, j \geqslant 1$, let

$$B_{ij} \equiv A_{\tau(X,mi)-1,\tau(Y,mj)-1},$$

so that each event B_{ij} involves m consecutive excursions of X and m consecutive excursions of Y. Denote the equilibrium distribution for X and Y by $(\pi(i), i \in S)$. The expected length of each excursion is $E[\tau(X, k+1) - \tau(X, k)] = 1/\pi(a)$. Let

$$E \equiv \{\tau(X, \lceil n\pi(a)/2 \rceil) < n - m\} \cap \{\tau(Y, \lceil n\pi(a)/2 \rceil) < n - m\}.$$

By the weak law of large numbers, $P(E) \to 1$ as $n \to \infty$. On the event E, the events B_{ij} for $1 \leqslant i, j \leqslant n\pi(a)/(2m)$ involve only $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$. Let S be the sum of the indicators of these events:

$$S \equiv \sum_{1 \leqslant i,j \leqslant n\pi(a)/(2m)} 1(B_{ij}).$$

On the event $E \cap \{S > 0\}$, we have $(M - 2\log n)/\log\log n \geqslant b$. Write $P^a(A_{11})$ for $P(A_{11} \mid X_1 = Y_1 = a)$, and observe that $\forall i, j,\ P(B_{ij}) = P^a(A_{11})$. Note that $P^a(A_{11}) > c_0 n^{-2}(\log n)^{-b}$, and hence if $b < -2$,

$$ES = (\lfloor n\pi(a)/(2m) \rfloor)^2 P^a(A_{11}) > c_0(\pi(a)/2)^2(\log n)^{-2-b} \to \infty \quad \text{as } n \to \infty.$$

In the expansion for var(S) as a sum of covariances of indicator functions, there are $\sim(n\pi(a)/(2m))^2$ diagonal terms, whose combined contribution is $< ES$, and there are $\sim(n\pi(a)/(2m))^4$ terms which are all zero, corresponding to the independent pairs of events, B_{ij} and B_{kl}, with $i \neq k$ and $j \neq 1$. Finally, there are $\sim 2(n\pi(a)/(2m))^3$ nontrivial terms for pairs of events such as B_{ij} and B_{ik} with $j \neq k$, or B_{ij} and B_{kj} with $i \neq k$.

For each "word" $w \in S^m$, let $p_w \equiv P((X_1, \ldots, X_m) = w \mid X_1 = a)$ so that $P(B_{ij}) = \sum_w (p_w)^2$. We have, using Holder's inequality, that for $j \neq k$

$$P(B_{ij} \cap B_{ik}) = \left[\sum_w (p_w)^3\right] < \left[\sum_w (p_w)^2\right]^{3/2} = P(B_{ij})^{3/2},$$

so that the combined contribution to var(S) from these $\sim 2(n\pi(a)/(2m))^3$ nontrivial pairs is $<2(ES)^{3/2}$. As in Theorem 2,

$$P(S=0) \leqslant \text{var}(S)/(ES)^2 < 1/ES + 2(ES)^{-1/2} = O((\log n)^{1+b/2}).$$

We have $P(M<m) \leqslant P(S=0) + P(E^c) \to 0$ provided $b<-2$. The strong law of large numbers for the return times to state a yields $P(E^c \text{ i.o.}) = 0$. An argument using the skeleton of times $n \equiv n(k) \equiv \lceil (1/p)^k \rceil$ shows that if $b<-4$, then $P(S=0 \text{ i.o.}) = 0$, and hence $P(M<m \text{ i.o.}) = 0$. ∎

3. Substitutions, Insertions, Deletions, and Inversions

In addition to the transposition of segments of DNA, the transformations of molecular evolution include single letter substitutions and insertions, deletions, or inversions of one or more letters. Thus, we consider the length of the longest match allowing shifts as well as a given number of substitutions, insertions, and inversions.

In the following two patterns of length $n=20$, the length of the longest match, allowing shifts, is 3, and there are nine places where this longest match appears—at the occurrences of "$A\ A\ A$" in both patterns:

$$X_1 \cdots X_{20} = T\ T\ T\ T\ T\ T\ T\ A\ A\ A\ T\ A\ A\ A\ C\ A\ A\ A\ C\ C,$$

$$Y_1 \cdots Y_{20} = G\ A\ A\ A\ G\ A\ A\ A\ G\ A\ A\ A\ G\ C\ C\ C\ G\ G\ G\ G.$$

If a single substitution is allowed to correct a mismatch, then by changing Y_5 from G to T, the match "$A\,A\,A\,T\,A\,A\,A$" of length 7 can be found. If one deletion is allowed, then deletion of $Y_{13} = G$ produces the match "$A\,A\,A\,C\,C$" of length 5. If three corrections are allowed, via either substitution or deletion, then there is the match "$A\,A\,A\,T\,A\,A\,A\,C\,A\,A\,A\,C\,C$" of length 13.

Only deletions are considered in Theorem 4 below, because the effects of a number of substitutions, insertions, and inversions can be "covered" by a comparable number of deletions. In detail, a pair of deletions, one for each sequence, can be used instead of a substitution to correct a single mismatch; the match produced after deletion will be one shorter than the match produced by substitution. In the above example, delete $X_{11} = T$ and $Y_5 = G$ to produce the match "$A\,A\,A\,A\,A\,A$" of length 6, instead of the match "$A\,A\,A\,T\,A\,A\,A$" of length 7. The effect of an insertion in one sequence could also be "covered" by a deletion in the other sequence, although the deletion produces a match which is shorter by one. In the example above, insertion of the letter G between X_{18} and X_{19} produces the match "$A\,A\,A\,G\,C\,C$" of length 6, versus the match "$A\,A\,A\,C\,C$" of length 5 achieved by deleting $Y_{13} = G$. The effect of an inversion can be duplicated by two substitutions; the correspondence between deletions and substitutions has already been discussed.

THEOREM 4. *Let $X_1, X_2, \ldots, Y_1, Y_2, \ldots,$ and p be as given in either Theorem 2 or Theorem 3. Let $M(k) = M_n(k)$ be the longest match between $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ allowing shifts and allowing at most k single letter deletions in each sequence*:

$$\begin{aligned} M_n(k) \equiv M(k) \equiv \max\{m\colon X_{i(a)} = Y_{j(a)} \textit{ for } a = 1 \textit{ to } m, \textit{ for some integers} \\ 1 \leqslant i(1) < \cdots < i(m) \leqslant n \textit{ and } 1 \leqslant j(1) < \cdots < j(m) \leqslant n \\ \textit{with } i(m) - i(1) < m + k \textit{ and } j(m) - j(1) < m + k\}. \end{aligned}$$

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

$$k = o(\log(n)/\log\log(n)), \tag{11}$$

we have, as $n \to \infty$,

$$M(k)/\log_{1/p}(n) \xrightarrow{P} 2.$$

Proof. The length of the longest match allowing shifts from Theorems 2 and 3 is $M = M(0)$, and for any pair of sequences, $M(0) \leqslant M(1) \leqslant \cdots$, so the lower bounds given in Theorems 2 and 3 for M also serve here for $M(k)$.

To establish an upper bound on $M(k)$, fix $\varepsilon > 0$ and let $m = m(n, \varepsilon) = \lceil (2 + \varepsilon) \log n \rceil$. Consider the event $A_{k;ij}$ that there is a match of length at least m, starting at the ith position among the X's and the jth position among the Y's, and allowing at most k single letter deletions,

$$A_{k;ij} \equiv \bigcup \{X_{i(a)} = Y_{j(a)} \text{ for } a = 1 \text{ to } m\},$$

where the union is taken over $i(1), \ldots, i(m)$, $j(1), \ldots, j(m)$ such that $i \leqslant i(1) < \cdots < i(m) \leqslant n$ and $j \leqslant j(1) < \cdots < j(m) \leqslant n$ with $i(m) - i(1) < m + k$ and $j(m) - j(1) < m + k$. Let $S_k = \sum_{0 < i,j < n-m+1} 1(A_{k;ij})$ so that $S_k = 0$ implies $M(k) < (2 + \varepsilon) \log n$. With $S \equiv S_0$, $ES = (n - m + 1)^2 p^m = O(n^{-\varepsilon})$.

For each choice of i, the number of choices for indices $i(1) < \cdots < i(m)$ with $i(1) \geqslant i$ and $i(m) - i(1) < m + k$ is $(m + k)!/(m!k!)$. In the i.i.d. setup of Theorem 2, $P(A_{k;11}) \leqslant [(m + k)!/(m!k!)]^2 P(A_{11}) < (m + k)^{2k} P(A_{11}) < P(A_{11})(3 \log n)^{2k}$ for large enough n, so that $ES_k < ES(3 \log n)^{2k} = O(n^{-\varepsilon}(3 \log n)^{2k})$. Thus, to show that $\forall \varepsilon > 0$, $ES_k \to 0$, we need $\forall \varepsilon > 0$, $(\log n)^k < n^\varepsilon$ eventually, or by taking logs, we need $\forall \varepsilon > 0$, $k \log\log n < \varepsilon \log n$ eventually, which is precisely (11).

In the Markov setup of Theorem 3, we claim that for each allowable choice of the indices $i(1),\ldots, j(m)$, $P(X_{i(a)} = Y_{j(a)}$ for $a = 1$ to $m) \leqslant cp^{m-2k}$, where p is specified at (9) and c is a constant depending only on our Markov transition mechanism. To prove this, let $T(a)$ be the matrix with $T(a)_{rs} = (P^{i(a+1)-i(a)})_{rs}(P^{j(a+1)-j(a)})_{rs}$, so that $P(X_{i(a)} = Y_{j(a)}$ for $a = 1$ to m, $X_{i(m)} = s \mid X_{i(1)} = Y_{j(1)} = r) = [\prod_{1 < a < m-1} T(a)]_{rs}$. Our claim is proved by noting that at least $m - 2k$ of the matrices $T(a)$ are the matrix $[(p_{rs})^2]$ whose largest eigenvalues is p, while the rest of the matrices $T(a)$ are substochastic. By counting the possibilities for $i, j,\ldots, j(m)$ it follows from our claim that $P(M(k) \geqslant (2 + \varepsilon) \log n) \leqslant [(m + k)!/(m!k!)]^2(n - m + 1)^2 cp^{m-2k} = o(1)$ as $n \to \infty$. ∎

The case $k = \infty$, i.e., no limit on the number of deletions allowed, corresponds exactly to the "longest common subsequence of two random sequences" studied in Chvatal and Sankoff [2].

Corollary 1. *Let $X_1, X_2,\ldots, Y_1, Y_2,\ldots,$ and p be as given in either Theorem 2 or Theorem 3. Let $M(k)^* = M_n(k)^*$ be the longest match between $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ allowing shifts and allowing at most k deletions, insertions, corrections, or transpositions. If $k = o(\log n/\log \log n)$, then $M(k)^*/\log_{1/p}(n) \to^P 2$.*

Proof. As seen from the discussion preceding Theorem 4, a match of length m achieved with at most k deletions, insertions, corrections, or transpositions corresponds to a match of length at least $m - 2k$ achievable with at most $2k$ deletions, as counted by $M(2k)$ in Theorem 4, i.e., $M(k)^* \leqslant M(2k) + 2k$. Trivially, $M(k) \leqslant M(k)^*$. Now if $k = o(\log(n)/\log \log n)$, then by Theorem 4, $M(2k)/\log n \to 2$ and also $(M(2k) + 2k)/\log n \to 2$. ∎

4. Extensions

Various extensions to Theorems 2 and 3, such as (12) and (13), can be obtained with minor modification of the proofs already given. We omit the details of the proofs.

Consider matches between two sequences of not necessarily equal lengths, say $n_1 \leqslant n_2$. That is, consider $M \equiv M(n_1, n_2) \equiv \max\{m: X_{i+k} = Y_{j+k}$ for $k = 1$ to m, for some $0 \leqslant i \leqslant n_1 - m$, $0 \leqslant j \leqslant n_2 - m\}$, where $X_1, X_2,\ldots, Y_1, Y_2,\ldots,$ and p are as given in Theorem 2. Define $x = \log_{1/p}(\sum_\alpha (q_\alpha)^3) \in (3/2, 2]$. If $x \neq 2$ and $(2 - x)/(x - 3/2) \leqslant \lim(\log n_1/\log n_2)$, or if $x = 2$ and $0 < \lim(\log n_1/\log n_2)$, then

$$P(\lim_{n \to \infty} M/(\log_{1/p}(n_1 n_2)) = 1) = 1. \tag{12}$$

Consider $r \geqslant 2$ independent sequences. For fixed r, the length $M = M_{n,r}$ of the longest consecutive subsequence common to the first n letters of all r sequences satisfies

$$P(\lim_{n\to\infty} M_{n,r}/\log_{1/p}(n) = r) = 1, \tag{13}$$

where $p = (\sum_a (p_a)^r)$ in the i.i.d. setup of Theorem 2, or $p =$ the largest eigenvalue of the substochastic matrix $[(p_{ij})^r]$, in the Markov setup of Theorem 3.

Acknowledgments

Louis Gordon supplied several helpful discussions. The second author benefited from discussions with Paul Erdos. Since the above work was completed, the authors became aware of a similar result: Karlin *et al.* [7] announced the mean and approximate variance of the length of the longest repeat within a single sequence.

References

1. K. L. Chung, "Markov Chains with Stationary Transition Probabilities," Springer-Verlag, Berlin/New York, 1960.
2. V. Chvatal and D. Sankoff, Longest common subsequence of two random sequences, *J. Appl. Prob.* 12 (1975), 306–315.
3. P. Erdös and A. Rényi, On a new law of large numbers, *J. Anal. Math.* 22 (1970), 103–111; reprinted in "Selected Papers of Alfred Rényi, Vol. 3, 1962–1970," Akademiai Kiado, Budapest, 1976.
4. P. Erdős and P. Révész, On the length of the longest head-run, *in* "Topics in Information Theory, Colloquia Math. Soc. J. Bolyai, No. 16. pp. 219–228, Keszthely, Hungary, 1975.
5. L. J. Guibas and A. M. Odlyzko, Long repetitive patterns in random sequences, *Z. Wahrsch. Verw. Gebiete* 53 (1980), 241–262.
6. Hardy, Littlewood, and Polya, "Inequalities," Cambridge University Press, London/New York, 1934.
7. S. Karlin, G. Ghandour, F. Ost, S. Tavare, and L. J. Korn, New approaches for computer analysis of nucleic acid sequences, *Proc. Natl. Acad. Sci. U.S.A.* 80 (1983), 5660–5664.
8. A. Rényi, "Probability Theory," Akademiai Kiado, Budapest, 1970.
9. P. Révész, Strong theorems on coin tossing, *in* "Proceedings of the International Congress of Mathematicians, Helsinki," pp. 749–754, 1978.
10. T. F. Smith and M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* 147 (1981), 195–197.
11. T. F. Smith, M. S. Waterman, and C. Burks, The statistical distribution of nucleic acid similarities, Preprint, 1983.

Printed by the St. Catherine Press Ltd., Tempelhof 41, Bruges, Belgium