The Annals of Probability
1985, Vol. 13, No. 4, 1236–1249

CRITICAL PHENOMENA IN SEQUENCE MATCHING

By Richard Arratia[1] and Michael S. Waterman[2]

University of Southern California

Dedicated to the memory of Mark Kac.

We give a generalization of the result of Erdös and Rényi on the length R_n of the longest head run in the first n tosses of a coin. Consider two independent sequences, $X_1X_2 \cdots X_m$ and $Y_1Y_2 \cdots Y_n$. Suppose that $X_1, X_2, \ldots$ are i.i.d. μ, and $Y_1, Y_2, \ldots$ are i.i.d. ν, where μ and ν are possibly different distributions on a common finite alphabet S. Let $p \equiv P(X_1 = Y_1) \in (0, 1)$. The length of the longest matching consecutive subsequence is $M_{m,n} \equiv \max\{k: X_{i+r} = Y_{j+r} \text{ for } r = 1 \text{ to } k, \text{ for some } 0 \le i \le m - k, 0 \le j \le n - k\}$. For m and $n \to \infty$ with $\log(m)/\log(mn) \to \lambda \in (0,1)$, our result is that there is a constant $K \equiv K(\mu, \nu, \lambda) \in (0, 1]$ such that $P(\lim M_{m,n}/\log_{1/p}(mn) = K) = 1$. The proof uses large deviation methods. The constant K is determined from a variational formula involving the Kullback–Liebler distance or relative entropy. A simple necessary and sufficient condition for $K = 1$ is given. For the case $m = n$ ($\lambda = 1/2$) and $\mu = \nu$, $K = 1$. The set of (μ, ν, λ) for which $K = 1$ has nonempty interior. The boundary of this set is the location of a phase transition. The results generalize to more than two sequences and to Markov chains. A strong law of large numbers is given for the proportion of letters within the longest matching word; the limiting proportion exhibits critical behavior, similar to that of K.

1. Introduction. This paper gives a generalization of the result of Erdös and Rényi on the length of the longest run of heads in the first n tosses of a coin. Our motivation is the comparison of DNA sequences, which are sometimes modeled as sequences of i.i.d. letters, or as letters of a Markov chain, with different distributions used for different sequences; see Smith, Waterman, and Sadler (1983).

Consider two sequences of length n, $X_1X_2 \cdots X_n$ and $Y_1Y_2 \cdots Y_n$. The length of the longest consecutive match, without shifts, is

$$(1) \quad R_n \equiv \max\{m: X_{i+k} = Y_{i+k} \text{ for } k = 1 \text{ to } m, \text{ for some } 0 \le i \le n - m\}.$$

The length of the longest consecutive match, allowing shifts, is

$$(2) \quad M_n \equiv \max\{m: X_{i+k} = Y_{j+k} \text{ for } k = 1 \text{ to } m, \text{ for some } 0 \le i, j \le n - m\}.$$

Suppose that the two sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are independent, with all letters chosen from a common finite alphabet S. Assume that $X_1, X_2, \ldots$

Received March 1984; revised December 1984.
[1]Supported by NSF Grant MCS-8301960.
[2]Supported by a grant from the System Development Foundation.
AMS 1980 *subject classifications*. Primary 60J10; secondary 68G10, 94A17.
Key words and phrases. Entropy, Kullback–Liebler distance, large deviations, sequence matching.

are i.i.d. (μ), and $Y_1, Y_2, \dots$ are i.i.d. (ν), where μ and ν are probability distributions on S. Let $p \equiv P(X_1 = Y_1) = \Sigma_{a \in S}(\mu_a \nu_a)$, and assume that $p \in (0,1)$.

To compute the length R_n of the longest match without shifts, the two sequences of letters may first be reduced to a single sequence of "heads" and "tails," with a "head" reported for the ith toss when $X_i = Y_i$. Thus R_n is the length of the longest head run in the first n tosses of a p-biased coin, described by the Erdös–Rényi law [Rényi (1970)]:

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

For the length M_n of the longest match with shifts, in the case $\mu = \nu$, it is shown in Arratia and Waterman (1985) that $P(\lim_{n\to\infty} M_n/\log_{1/p}(n) = 2) = 1$, so that

$$P\Big(\lim_{n\to\infty} M_n/R_n = 2\Big) = 1. \tag{4}$$

Loosely speaking, allowing shifts between two independent sequences *with the same distribution* doubles the length of the longest match. To see that M_n might grow like $2\log_{1/p}(n)$, note that a match of length $m = \lceil 2\log_{1/p}(n) \rceil$ 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). However, if μ and ν are not "close," in a sense to be made precise later, then allowing shifts will not double the length of the longest match, i.e., (4) does not hold.

For a class of examples in which we can explicitly determine when allowing shifts doubles the length of the longest match, let $X_1, X_2, \dots$ be a sequence of fair coin tosses, and let $Y_1, Y_2, \dots$ be an independent sequence of biased coin tosses, with $\theta = P(Y_1 = \text{heads}) \in [0,1]$. For all θ, $p = \frac{1}{2}$, so by (3), R_n grows like $\log_2(n)$. In the case $\theta = \frac{1}{2}$, the two sequences have the same distribution, so that M_n satisfies (4). In the case $\theta = 1$, the Y sequence is pure heads, so that $M_n \equiv R_n$ is the length of the longest head run in $X_1 X_2 \cdots X_n$, i.e., allowing shifts has no effect on the length of the longest match, and (4) does not hold. What happens for intermediate cases, when one sequence represents a fair coin and the other sequence represents a biased but nondegenerate coin? Part of the answer, given by Theorem 1, is that (4) holds iff $\theta \in [x, 1 - x]$, where $x = 0.11002786\dots$ is the smaller solution of $(x)\log(x) + (1 - x)\log(1 - x) = -(\log 2)/2$.

Theorem 1 states that if $X_1, X_2, \dots$ is i.i.d. (μ) and $Y_1, Y_2, \dots$ is i.i.d. (ν), with all letters independent and $p = P(X_1 = Y_1) \in (0,1)$, then there exists a constant $C = C(\mu, \nu) \in [1,2]$ such that

$$P\Big(\lim_{n\to\infty} M_n/R_n = C\Big) = 1. \tag{5}$$

[In the notation used in the summary and in Section 6, $C(\mu, \nu) \equiv 2K(\mu, \nu, 1/2)$.]

Let $\alpha \equiv \alpha(\mu, \nu)$ be the distribution on S corresponding to matching a single pair of letters:

$$\alpha_a \equiv (\mu_a \nu_a)/p = P(X_1 = Y_1 = a \mid X_1 = Y_1). \tag{6}$$

A necessary and sufficient condition for $C = 2$ is that

$$\begin{aligned} &\sum (\mu_a \nu_a / p) \log(\mu_a) \le (\log p)/2 \\ \text{and } &\sum (\mu_a \nu_a / p) \log(\nu_a) \le (\log p)/2, \end{aligned} \tag{7}$$

or equivalently, after a little manipulation,

$$H(\alpha, \nu) \le (1/2)\log(1/p) \quad \text{and} \quad H(\alpha, \mu) \le (1/2)\log(1/p). \tag{7'}$$

Here $H(\cdot,\cdot)$ is the relative entropy or Kullback–Leibler distance: $H(\alpha, \nu) \equiv \Sigma \alpha_a \log(\alpha_a/\nu_a) \ge 0$, with $H(\alpha, \nu) = 0$ iff $\alpha = \nu$.

Let $H(\alpha) \equiv -\Sigma \alpha_a \log(\alpha_a) \ge 0$ be the entropy of α. Note that $H(\alpha, \mu) + H(\alpha, \nu) = -H(\alpha) + \log(1/p)$, so that if $H(\alpha, \mu) = H(\alpha, \nu)$ (in particular if $\mu = \nu$) then $H(\alpha, \mu) = H(\alpha, \nu) = [-H(\alpha) + \log(1/p)]/2 \le (1/2)\log(1/p)$ so by (7'), $C(\mu, \nu) = 2$. Furthermore, if $\nu = \mu$ and μ is nontrivial, then α is nontrivial, so $H(\alpha) > 0$, and $H(\alpha, \mu) = H(\alpha, \nu) < (1/2)\log(1/p)$. It follows from (7') that for a fixed nontrivial distribution μ, $C(\mu, \nu) = 2$ for all distributions ν in some neighborhood of μ.

2. Further discussion. For any distributions μ and ν, it is very easy to get an upper bound on $M_n - 2\log_{1/p}(n)$, as follows. For $m = 1, 2, \ldots,$ define the event

$$A_{ij} \equiv \{X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m}\} \tag{8}$$

that some "witness" to the event $\{M \ge m\}$ appears at positions i in the X sequence and j in the Y sequence. Note that $P(A_{ij}) = p^m$ for each choice of i and j. Thus if $m = 2\log_{1/p} n + b$ is an integer, so that $p^m = n^{-2}p^b$, we have $P(M \ge m) = P(\bigcup_{0 \le i, j \le n-m} A_{ij}) < n^2 p^m = p^b$. Write $\lfloor x \rfloor$ for the greatest integer $\le x$, and write $\lceil x \rceil$ for the least integer $\ge x$. Using $m = \lceil (2+\varepsilon)\log_{1/p} n \rceil$ yields $P(M_n/(\log_{1/p} n) > 2 + \varepsilon) < n^{-\varepsilon}$, and an argument using the Borel–Cantelli lemma along a skeleton of times $n_k \equiv \lceil p^{-k} \rceil$ implies that $1 = P(\limsup(M_n/(\log_{1/p} n)) \le 2)$.

The idea behind the proof of Theorem 1 is contained in the following calculation, which shows directly how condition (7) arises. Let $m = \lceil 2\log_{1/p}(n) \rceil$, so that $n^2 p^m \in (p, 1]$. For each "word" $w \in S^m$ let E_w be the event that w appears within both $X_1 X_2 \cdots X_n$ and $Y_1 Y_2 \cdots Y_n$:

$$\begin{aligned} E_w &\equiv E_{w,n} \\ &\equiv \{w = X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m} \text{ for some } 0 \le i,\ j \le n - m\}. \end{aligned} \tag{9}$$

From the independence of the two sequences,

$$\begin{aligned} &P(E_w) \\ &\quad = P\Big(\bigcup_{0 \le i \le n-m} \{w = X_{i+1} \cdots X_{i+m}\}\Big) P\Big(\bigcup_{0 \le j \le n-m} \{w = Y_{j+1} \cdots Y_{j+m}\}\Big) \\ &\quad < [(n\mu^m(w)) \wedge 1][(n\nu^m(w)) \wedge 1]. \end{aligned} \tag{10}$$

Now with unions and sums taken over $w \in S^m$,

$$P(M \geq m) = P\Big(\bigcup_w E_w\Big) \leq \sum_w P(E_w) \tag{11}$$

$$< \sum_w [(n\mu^m(w)) \wedge 1][(n\nu^m(w)) \wedge 1]$$

$$\leq \sum_w n\mu^m(w) n\nu^m(w) \tag{12}$$

$$= n^2 p^m \in (p, 1],$$

using $p = \Sigma_a(\mu_a\nu_a)$ to get the final equality. By the weak law of large numbers, most of the contribution to the sum at (12) comes from words w in which the proportions of letters are approximately those of the distribution α at (6). The condition (7) is that for words w with proportions α, both $n\mu^m(w)$ and $n\nu^m(w)$ are not larger than $np^{m/2} \in (p^{1/2}, 1]$, so that the truncations "$\wedge 1$" in the line preceding (12) have a negligible effect on the sum.

In the general case, $C = C(\mu, \nu) \in [1,2]$ is defined by the requirement that with $m = \lceil C \log_{1/p} n \rceil$, the sum in (11) is ≈ 1, in the sense that $0 = \lim_{n\to\infty}((1/m)\log[\Sigma_{w\in S^m} P(E_w)])$. To show that $M_n/R_n \to C$ in probability, only minor modification of the above calculation is needed. The upper bound on M, $\forall\ \varepsilon > 0$ $P(M_n/\log_{1/p} n > C + \varepsilon) \to 0$, is easily proved; it suffices to use $m = \lceil (C + \varepsilon)\log_{1/p} n \rceil$, and show that $P(M \geq m) = P(\cup E_w) \leq \Sigma P(E_w) \to 0$ as $n \to \infty$. To get the lower bound, $\forall\ \varepsilon > 0$ $P(M_n/\log_{1/p} n > C - \varepsilon) \to 1$, is more difficult; a bound on correlations is needed. For each word $w \in S^m$ consider the event G_w that w appears at a multiple of m within both $X_1X_2 \cdots X_n$ and $Y_1Y_2 \cdots Y_n$:

$$G_w \equiv G_{w,n} \equiv \bigcup_{0 \leq i,j \leq (n/m)-1} \{w = X_{im+1} \cdots X_{(i+1)m} = Y_{jm+1} \cdots Y_{(j+1)m}\}, \tag{13}$$

so that $\cup_w G_w \subset \{M \geq m\}$. For $m = \lceil (C - \varepsilon)\log_{1/p} n \rceil$, calculation shows that $\Sigma_w P(G_w) \to \infty$. For $w \neq v \in S^m$, the events G_w and G_v are negatively correlated (Lemma 1), so that $\Sigma P(G_w) \to \infty$ implies $P(\cup_w G_w) \to 1$.

3. Distinguishing matches by the proportions of letters involved. Let $\mathrm{Pr}(S) = \{\gamma \in R^d\colon \gamma_a \geq 0, \Sigma\gamma_a = 1\}$ be the set of probability measures on our finite alphabet $S \equiv \{1, 2, \ldots, d\}$, and for $m = 1, 2, \ldots,$ for any word $w \in S^m$, let $L(w) \in \mathrm{Pr}(S)$ be the vector whose ath component is the proportion of letter a among the letters of w:

$$\text{for } a \in S, \qquad L(w)_a \equiv (1/m) \sum_{1 \leq i \leq m} 1(w_i = a).$$

For $U \subset \mathrm{Pr}(S)$ define the length $M_{n,U}$ of the longest match between $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ with proportions in U:

$$M_{n,U} \equiv \max\{m\colon X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m} = w \text{ for some } w \text{ with } L(w) \in U, \text{ for some } 0 \leq i,\ j \leq n - m\}. \tag{14}$$

Given μ and ν with $p \equiv \Sigma\mu_a\nu_a \in (0,1)$, for $\gamma, \beta \in \Pr(S)$ and $c > 0$ define (with the convention that $\log 0 = -\infty$, but $0 \log 0 = 0$)

$$
\begin{aligned}
b(\gamma,\beta,c) &\equiv (1/c)\log(1/p) + \sum \gamma_a \log \beta_a, \\
f(\gamma,c) &\equiv H(\gamma) + 0 \wedge b(\gamma,\mu,c) + 0 \wedge b(\gamma,\nu,c), \qquad (15) \\
g(\gamma) &\equiv \inf\{c\colon f(\gamma,c) < 0\}.
\end{aligned}
$$

Informally, $f(\gamma, c)$ represents $1/m$ times the log of the contribution to the sum in the line before (12), from words w having $L(w)$ near γ, if $m = \lceil c \log_{1/p} n \rceil$. Note that $f(\gamma, \cdot)$ is nonincreasing, and if $H(\gamma) > 0$ and $f(\gamma, c) = 0$, then $f(\gamma, \cdot)$ is strictly decreasing in a neighborhood of c. Thus for nontrivial $\gamma \in \Pr(S)$, $g(\gamma)$ is the unique value for c for which $f(\gamma, c) = 0$. If $\gamma = \delta_a$ is the point mass on the letter $a \in S$ and $q \equiv \min(\mu_a, \nu_a) > 0$, then $g(\gamma) = (\log p)/(\log q) \in (0,2)$.

The expressions for f and g in (15) allow a remarkable degree of simplification. Let $f_0(\gamma,c) \equiv H(\gamma)$; $f_1(\gamma,c) \equiv H(\gamma) + b(\gamma,\mu,c)$; $f_2(\gamma,c) \equiv H(\gamma) + b(\gamma,\nu,c)$; and $f_3(\gamma,c) \equiv H(\gamma) + b(\gamma,\mu,c) + b(\gamma,\nu,c)$; so that $f \equiv \min(f_0, f_1, f_2, f_3)$. For $i = 1,2,3$, define $g_i(\gamma)$ by the requirement that $f_i(\gamma, g_i(\gamma)) = 0$, so that $g(\gamma) = \min_{1 \le i \le 3} g_i(\gamma)$. Now

$$
f_1(\gamma,c) \equiv -\sum \gamma_a \log \gamma_a - (\log p)/c + \sum \gamma_a \log \mu_a = -H(\gamma,\mu) - (\log p)/c,
$$

so that $g_1(\gamma) = \log(1/p)/H(\gamma,\mu)$ and $g_2(\gamma) = \log(1/p)/H(\gamma,\nu)$. Also

$$
\begin{aligned}
f_3(\gamma,c) &\equiv -\sum \gamma_a \log \gamma_a - 2(\log p)/c + \sum \gamma_a \log \mu_a \nu_a \\
&= \sum \gamma_a \log(\mu_a \nu_a/(p\gamma_a)) - (2-c)(\log p)/c \\
&= -H(\gamma,\alpha) - (2-c)(\log p)/c,
\end{aligned}
$$

so that $g_3(\gamma) = (2\log(1/p))/(\log(1/p) + H(\gamma,\alpha))$. Thus

$$
g(\gamma) = \min\left\{ \frac{\log(1/p)}{H(\gamma,\mu)}, \frac{\log(1/p)}{H(\gamma,\nu)}, \frac{2\log(1/p)}{\log(1/p) + H(\gamma,\alpha)} \right\}. \qquad (16)
$$

THEOREM 1. *If $X_1, X_2, \ldots$ are i.i.d. (μ) and $Y_1, Y_2, \ldots$ are i.i.d. (ν), with all letters independent and $p = P(X_1 = Y_1) \in (0,1)$, then for any open $U \subset Pr(S)$, $M_{n,U}/(\log_{1/p} n)$ converges a.s. to $\sup_{\gamma \in U} g(\gamma)$. In particular, $1 = P(\lim_{n\to\infty} M_n/\log_{1/p} n = C(\mu,\nu))$ where*

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

and $C(\mu,\nu) = 2$ if and only if both $H(\alpha,\nu)$, $H(\alpha,\mu) \le (1/2)\log(1/p)$.

PROOF. Fix an open nonempty set $U \subset \Pr(S)$ and let $c = \sup_{\gamma \in U} g(\gamma)$.

First we prove the lower bound, that $P(M_{n,U} > (c - \varepsilon)\log_{1/p} n$ eventually$) = 1$. If $c = 0$ (which occurs iff there is some letter $a \in S$ with $\alpha_a = 0$ and $\gamma_a > 0$ $\forall$ $\gamma \in U$,) then the lower bound is automatic. Assume that $c > 0$. Let $\varepsilon > 0$ be given; we may assume that $\varepsilon < c$. Fix a particular nontrivial $\beta \in U$ for which $g(\beta) > c - \varepsilon$. From the strict monotonicity of $f(\beta, \cdot)$ in a neighborhood of $g(\beta)$, it follows that $f(\beta, c-\varepsilon) > 0$. Let $\delta = f(\beta, c-\varepsilon)/5$. Fix an open set V with $\beta \in V \subset U$ for which the final two terms in expression (15) for $f(\cdot, c-\varepsilon)$ vary by at most δ from their values at β, so that $\forall\ \gamma \in V,\ 0 \wedge b(\gamma, \mu, c-\varepsilon) \geq 0 \wedge b(\beta, \mu, c-\varepsilon) - \delta$, and similarly with ν in place of μ.

The number of words w of length m with proportions $L(w)$ in V is at least $\exp(m(H(\beta) - \delta))$, if m is sufficiently large, by Lemma 2. Let

$$T \equiv T(V, n, m) \equiv \sum_{w \in S^m:\ L(w) \in V} 1(G_{w,n}),$$

so that with $m = \lceil (c-\varepsilon)\log_{1/p} n \rceil$,

$$\{T \neq 0\} = \bigcup_{w \in S^m:\ L(w) \in V} G_{w,n} \subset \{M_{n,V} > (c-\varepsilon) log_{1/p} n\}.$$

Using Lemma 3, for sufficiently large n we have

$$\begin{aligned}(1/m)\log(ET) &\geq H(\beta) - \delta \\ &\qquad + 0 \wedge b(\beta, \mu, c-\varepsilon) - \delta + 0 \wedge b(\beta, \nu, c-\varepsilon) - \delta - \delta \\ &= f(\beta, c-\varepsilon) - 4\delta = \delta > 0,\end{aligned}$$

so that $ET > \exp(m\delta)$ for large n. Using Chebyshev's inequality and then Lemma 1 to get $\operatorname{var}(T) < ET$,

$$\begin{aligned}P(M_{n,V} > (c-\varepsilon)\log_{1/p} n) &\geq P(T \neq 0) \\ &> 1 - \operatorname{var}(T)/\{E(T)\}^2 \\ &> 1 - 1/E(T) \\ &> 1 - \exp(-m\delta).\end{aligned}$$

A Borel–Cantelli argument along the skeleton of times $n_k \equiv \lceil p^{-k} \rceil$ implies that $1 = P(M_{n,V} > (c-\varepsilon)\log_{1/p} n$ eventually). Hence $1 = P(M_{n,U} > (c-\varepsilon)\log_{1/p} n$ eventually).

Now we prove the upper bound. For each $\gamma \in U$, $c \geq g(\gamma)$ implies $f(\gamma, c + \varepsilon/2) < 0$. Hence at least one of the two terms $b(\gamma, \mu, c+\varepsilon/2), b(\gamma, \nu, c+\varepsilon/2)$ is < 0, and not controlled by the truncation with 0. With $\delta = (1/5)\log(1/p)$ $[(c+\varepsilon/2)^{-1} - (c+\varepsilon)^{-1}]$, it follows that for all $\gamma \in U$, $f(\gamma, c+\varepsilon) \leq -5\delta < 0$.

Each of the three terms in expression (15) for f is continuous, and $\Pr(S)$ is compact, so that we can pick a finite collection $\{\gamma_i, V_i\}$ such that $U \subset \bigcup_i V_i$, and for each $i, \gamma_i \in V_i \subset U$, and for all $\gamma \in V_i$, $H(\gamma) < H(\gamma_i) + \delta, 0 \wedge b(\gamma, \mu, c+\varepsilon) < 0 \wedge b(\gamma_i, \mu, c+\varepsilon) + \delta$, and $0 \wedge b(\gamma, \nu, c+\varepsilon) < 0 \wedge b(\gamma_i, \nu, c+\varepsilon) + \delta$.

The number of words $w \in S^m$ with proportions $L(w) \in V_i$ is less than $\exp[m(H(\gamma_i) + 2\delta)]$, for sufficiently large m, by Lemma 2. Let

$$T_i \equiv T(V_i, n, m) \equiv \sum_{w \in S^m:\ L(w) \in V(i)} 1(E_{w,n}),$$

so that with $m = \lceil (c+\varepsilon)\log_{1/p} n \rceil$, $\{M_{n,V(i)} \geq (c+\varepsilon)\log_{1/p} n\} \subset \{T_i \neq 0\}$. Using the upper bound on $P(E_{w,n})$ from Lemma 3, for large n we have

$$\begin{aligned}(1/m)\log(ET_i) &\le H(\gamma_i) + 2\delta + b(\gamma_i, \mu, c+\varepsilon) + \delta + b(\gamma_i, \nu, c+\varepsilon) + \delta \\ &= f(\gamma_i, c+\varepsilon) + 4\delta \le -\delta < 0,\end{aligned}$$

so that $ET_i < \exp(-m\delta)$ for large n.

A Borel–Cantelli argument with $n_k \equiv \lceil p^{-k} \rceil$ implies that for each i, $0 = P(M_{n,V(i)} > (c+\varepsilon)\log_{1/p} n$ infinitely often). Hence $1 = P(M_{n,U} < (c+\varepsilon)\log_{1/p} n$ eventually). □

LEMMA 1. *Let $X_1, X_2, \dots, Y_1, Y_2, \dots$ be independent S-valued variables, let integers m and n be fixed, and for any two distinct $w, v \in S^m$, consider the events G_w and G_v defined in* (13). *The events G_w and G_v are negatively correlated.*

PROOF. Writing $k \equiv \lfloor n/m \rfloor$, we have

$$\begin{aligned}P\big((G_w)^c \cap (G_v)^c\big) &= (1 - \mu^m(w) - \mu^m(v))^k (1 - \nu^m(w) - \nu^m(v))^k \\ &\le (1-\mu^m(w))^k (1-\mu^m(v))^k (1-\nu^m(w))^k (1-\nu^m(v))^k \\ &= P\big((G_w)^c\big) P\big((G_v)^c\big).\end{aligned}$$ □

LEMMA 2. *Let $S = \{1, 2, \dots, d\}$ and let $U \subset \Pr(S)$ be an open subset of the set of probability measures on S. The number of words of length m with proportions in U grows like $\exp(m \sup_{\gamma \in U} H(\gamma))$, i.e.,*

$$\lim_{m\to\infty} (1/m)\log\big(|\{w \in S^m \colon L(w) \in U\}|\big) = \sup_{\gamma \in U} H(\gamma).$$

PROOF. This result is contained in the theory of large deviations of sums of independent R^d-valued random vectors, as in Bahadur (1971). We present a simple proof, in order to prepare the way for Lemma 4 and to keep this paper self-contained. Now $|\{w \in S^m\colon (L(w) \in U\}| = \sum m!/(m_1! \cdots m_d!)$, where the sum is taken over integers $m_1, \dots, m_d$ for which $\sum m_i = m$ and $\gamma = (m_1/m, \dots, m_d/m) \in U$. From $n \log n - n + 1 < \log(n!) < (n+1)\log(n+1) - n$ it follows that

$$\begin{aligned}H(\gamma) - m^{-1}\textstyle\sum(1 + \log(m_i + 1)) &< m^{-1}\log(m!/[m_1! \cdots m_d!]) \\ &< H(\gamma) + m^{-1}\log m,\end{aligned}$$

where $\gamma = (m_1/m, \dots, m_d/m) \in \Pr(S)$. The lower bound on $(1/m)\log(|\{w \in S^m\colon L(w) \in U\}|)$ is demonstrated by taking a single choice of $(m_1, \dots, m_d)$ with proportions $\gamma = (m_1/m, \dots, m_d/m)$ whose entropy $H(\gamma)$ approximates $\sup_{\gamma\in U} H(\gamma)$. For the upper bound, note that the number of terms in the sum is $\le m^d$, and $(1/m)\log(m^d) \to 0$ as $m \to \infty$. □

LEMMA 3. *Suppose $X_1, X_2, \dots$ are i.i.d. (μ) and $Y_1, Y_2, \dots$ are i.i.d. (ν), with all letters independent. Let $c > 0$ and $p \in (0,1)$ be given and let $m \equiv$*

$m(n,c) \equiv \lceil c\log_{1/p} n \rceil$. *Let* $w \in S^m$ *have proportions* $L(w)$ *such that* $L(w_a) = 0$ *whenever* $\mu_a \nu_a = 0$. *Then the function* f *defined in* (15) *and the events* $E_{w,n}$ *and* $G_{w,n}$ *defined in* (9) *and* (13) *satisfy*

$$f(L(w),c) - H(L(w)) - \varepsilon < (1/m)\log P(G_{w,n}) < (1/m)\log P(E_{w,n}) < f(L(w),c) - H(L(w)),$$

where $\varepsilon = (2/m)[\log(4m) + \log(1/p)/c] \to 0$ *as* $n \to \infty$.

Proof. To see the upper bound note that

$$(1/m)(\log n) \le 1/(c\log_{1/p} n)(\log n) = \log(1/p)/c$$

and

$$(1/m)\log(\mu^m(w)) = (1/m)\sum_{1\le i\le m}\log(\mu_{w(i)}) = \sum_{a\in S} L(w)_a \log(\mu_a),$$

so that

$$(1/m)\log[1 \wedge (n\mu^m(w))] \le 0 \wedge b(L(w),\mu,c).$$

Thus

$$P(E_w) = P\Big(\bigcup_{0\le i\le n-m}\{w = X_{i+1}\cdots X_{i+m}\}\Big)P\Big(\bigcup_{0\le j\le n-m}\{w = Y_{j+1}\cdots Y_{j+m}\}\Big)$$
$$< [(n\mu^m(w)) \wedge 1][(n\nu^m(w)) \wedge 1],$$

and hence $(1/m)\log P(E_{w,n}) < 0 \wedge b(L(w),\mu,c) + 0 \wedge b(L(w),\nu,c) = f(L(w),c) - H(L(w))$.

For the lower bound, by independence

$$P(G_w) = P\Big(\bigcup_{0\le i\le n/m-1}\{w = X_{mi+1}\cdots X_{mi+m}\}\Big)$$
$$\cdot P\Big(\bigcup_{0\le j\le n/m-1}\{w = Y_{mj+1}\cdots Y_{mj+m}\}\Big)$$
$$= \big[1-(1-\mu^m(w))^{\lfloor n/m\rfloor}\big]\big[1-(1-\nu^m(w))^{\lfloor n/m\rfloor}\big].$$

For all $z \in [0,1]$ and $n = 0,1,2,\ldots,$ $1-(1-z)^n \ge (1/2)(nz \wedge 1)$, so that

$$P(G_w) \ge (1/4)(\lfloor n/m\rfloor \mu^m(w) \wedge 1)(\lfloor n/m\rfloor \nu^m(w) \wedge 1)$$
$$\ge 1/(4m^2)(n\mu^m(w) \wedge 1)(n\nu^m(w) \wedge 1).$$

Since $m - 1 \le c\log_{1/p} n$, $(1/m)\log(n) \ge \log(1/p)/c - \log(1/p)/(mc)$. Thus for all n and w,

$$(1/m)\log P(G_w) \ge f(L(w),c) - H(L(w)) - (2/m)[\log(2m) + \log(1/p)/c].$$

□

4. A strong law of large numbers. Informally, Theorem 1 says that for the longest consecutive match between $X_1X_2\cdots X_n$ and $Y_1Y_2\cdots Y_n$ with proportions near a given distribution γ, the length relative to $\log_{1/p} n$ tends almost surely to $g(\gamma)$. Now the function g: $\Pr(s) \to [0, 2]$ is continuous, and we will prove that g achieves its maximum $C(\mu, \nu)$ at a unique distribution β. It then follows easily from Theorem 1 that for any neighborhood U of β, the longest match with proportions in U will be longer than the longest match with proportions not in U, almost surely as $n \to \infty$. Thus the proportions of *all* matching words of maximal length tend almost surely to β, as $n \to \infty$.

Depending on μ and ν, the distribution α of letters in a simple match may or may not be the distribution β which maximizes g. For the coin tossing example discussed in Section 1, $\mu = (0.5, 0.5)$ and $\nu = \alpha = (1 - \theta, \theta)$, any case having $0 < H(\nu) < (1/2)\log 2$ gives an example with $\beta \neq \alpha$.

THEOREM 2. *In the setup of Theorem* 1, *there is a unique* $\beta \in \Pr(S)$ *such that*

$$g(\beta) = C(\mu, \nu) = \sup_{\gamma \in \Pr(S)} g(\gamma).$$

If $C(\mu, \nu) = 2$ (*in particular, if* $H(\alpha, \mu) = H(\alpha, \nu)$,), *then* $\beta = \alpha$. *As* $n \to \infty$, *the proportions of letters, in all words of maximal length common to both* $X_1 X_2 \cdots X_n$ *and* $Y_1 Y_2 \cdots Y_n$, *tend almost surely to* β:

$$1 = P\Big(0 = \limsup_{n \to \infty} \big\{|\beta - L(w)|: w = X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m}$$
$$\text{for some } 0 \le i, j \le n - m, \text{ with } m = M_n\big\}\Big).$$

PROOF. To see that g achieves its maximum at a unique distribution β, consider the expression for g in (16): $g = \min_{1 \le i \le 3} g_i$. Since g_1 and g_2 have no local maxima in the interior of $\Pr(S)$, g achieves its maximum either at α, where g_3 has its unique maximum, or else on one of the surfaces $g_i = g_j$. A maximum for g on the surface $g_1 = g_2$ is easily ruled out, since $g_1(\gamma) = g_2(\gamma) = c > 0$ implies $f_1(\gamma, c) = f_2(\gamma, c) = 0$ and thus $f_3(\gamma, c) = -H(\gamma) < 0$, so that $g_3(\gamma) < c$.

If $g_1(\gamma) = g_3(\gamma) = c > 0$ then $0 = f_1(\gamma, c) = f_3(\gamma, c)$ so that $f_2(\gamma, c) = H(\gamma) > 0$ and hence $g_2(\gamma) > c$ so that $g(\gamma) = c = g_3(\gamma)$. On the surface $\{\gamma: g_1(\gamma) = g_3(\gamma)\} \equiv \{\gamma: \log(1/p) + H(\gamma, \alpha) = 2H(\gamma, \mu)\}$, g_3 is maximized by minimizing $H(\gamma, \alpha)$. It follows from the strict convexity of $H(\cdot, \alpha)$ and of $H(\cdot, \mu)$ that there is a unique γ_μ which achieves this. Similarly, there is a unique γ_ν which maximizes $g(\gamma)$ given the constraint $g_2 = g_3$. It remains to show that $g(\gamma_\mu) \neq g(\gamma_\nu)$. If $g(\gamma_\mu) = g(\gamma_\nu)$, then $\gamma_\mu \neq \alpha$ so $2 > g_3(\gamma_\mu) = g_1(\gamma_\mu)$ and hence $H(\alpha, \mu) \ge H(\gamma_\mu, \mu) > (1/2)\log(1/p)$. The same argument yields $H(\alpha, \nu) > (1/2)\log(1/p)$, which is impossible, since $H(\alpha, \mu) + H(\alpha, \nu) = -H(\alpha) + \log(1/p) \le \log(1/p)$. We have shown that there exist a distribution β such that $g(\beta) > g(\gamma)$ for all $\gamma \neq \beta$.

Since $g(\gamma) \le (2\log(1/p))/(\log(1/p) + H(\gamma, \alpha))$ by (16), and $H(\gamma, \alpha) \ge 0$ with equality iff $\gamma = \alpha$, it follows that if $C(\mu, \nu) = 2$, then for $\gamma \neq \alpha$, $g(\gamma) < 2 = g(\alpha)$.

Given $\varepsilon > 0$, let $U = \{\gamma \in \Pr(S): |\gamma - \beta| > \varepsilon\}$. Let $\delta = (1/2)(g(\beta) - \sup_{\gamma \in U} g(\gamma))$; $\delta > 0$ since $\Pr(S)$ is compact and g is continuous. By Theorem 1, there is a random N which is almost surely finite, such that for all

$n > N$, $M_n/\log_{1/p} n > g(\beta) - \delta$ and $M_{n,U}/\log_{1/p} n < \sup_{\gamma \in U} g(\gamma) + \delta = g(\beta) - \delta$. Thus $n > N$ implies that $|\beta - L(w)| \le \varepsilon$, for all w with $w = X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m}$ for some $0 \le i,\ j \le n - m$, with $m = M_n$. □

5. Matching between two different Markov processes. In this section we generalize Theorem 1 to the situation in which $X_1 X_2 \cdots X_n$ and $Y_1 Y_2 \cdots Y_n$ are independent sequences of letters governed by two different Markov transition mechanisms on the finite alphabet $S = \{1, 2, \dots, d\}$.

It is necessary to keep track of the proportions of *pairs* of letters that appear in adjacent positions. Note that for any word $w \in S^m$ and letter $i \in S$, the number of adjacent pairs in w that begin with i is equal to the number of pairs that end in i, provided that the word is wrapped around a circle so that the pair (last letter, first letter) is counted as one of the m pairs. Thus we define:

$$\text{for } w \in S^m, \qquad \tilde{L}(w)_{ij} = (1/m) \sum_{1 \le k \le m} 1(w_k w_{k+1} = ij); \qquad i, j \in S,$$

with the understanding that w_{m+1} is evaluated as w_1. Let

$$B \equiv \left\{ q \in \Pr(S^2) \colon \forall\, i, j \in S, q_{ij} > 0 \text{ and } \sum_{k \in S} q_{ik} = \sum_{k \in S} q_{ki} \right\},$$

be the set of strictly positive *balanced* proportions of pairs, so that for any word w, $\tilde{L}(w) \in \overline{B}$. For $q, r \in \Pr(S^2)$ define

$$\tilde{H}(q) \equiv - \sum_{i, j \in S} q_{ij} \log\left(q_{ij} \Big/ \left(\sum_{k \in S} q_{ik} \right) \right)$$

and

$$\tilde{H}(q, r) \equiv \sum_{i, j} q_{ij} \log\left(\left(q_{ij} \Big/ \left(\sum_k q_{ik} \right) \right) \Big/ \left(r_{ij} \Big/ \left(\sum_k r_{ik} \right) \right) \right).$$

Note that if π and σ are the marginals of q and r, respectively, so that $\pi_i = \sum_k q_{ik}$ and $\sigma_i = \sum_k r_{ik}$, then $[q_{ij}/\pi_i]$ is the stochastic matrix governing a Markov process, and if $q \in B$ then (π_i) is the invariant measure: $\sum_i \pi_i (q_{ij}/\pi_i) = \sum_i q_{ij} = \pi_j$. Also $\tilde{H}(q, r) \ge 0$, with equality iff $q = r$, since $\tilde{H}(q, r) = \sum_i \pi_i [H(q_{i(\cdot)}/\pi_i, r_{i(\cdot)}/\sigma_i)]$. Note that $\tilde{H}(q, r) \le H(q, r)$, with equality iff $\pi = \sigma$, since $\tilde{H}(q, r) = H(q, r) - H(\pi, \sigma)$.

LEMMA 4. *Let $S = \{1, 2, \dots, d\}$ and let $U \subset \Pr(S^2)$ be open. The number of words of length m with "proportions of pairs" in U grows like $\exp(m \sup_{q \in U \cap B} \tilde{H}(q))$, i.e.,*

$$\lim_{m \to \infty} (1/m) \log\left(|\{ w \in S^m \colon \tilde{L}(w) \in U \}| \right) = \sup_{q \in U \cap B} \tilde{H}(q).$$

PROOF. We give an elementary proof, but note that this result could also be proved by applying the large deviation theory in Donsker and Varadhan (1975) to the *two-step* Markov chain with state space S^2 and transition probabilities $p_{(i,j),(k,l)} = (1/d)\delta_{jk}$.

Let integers $m_{ij} > 1$, $i, j \in S$, be given, with the property that for each $i \in S$, $\Sigma_j m_{ij} = \Sigma_j m_{ji}$. Let $m = \Sigma_{ij} m_{ij}$ and let $m_i = \Sigma_j m_{ij}$, for each $i \in S$. Let $q_{ij} = m_{ij}/m$, for $i, j \in S$, so that $q \in B$. Elementary analysis of multinomial coefficients, as in Lemma 2, will complete the proof, once it is shown that

$$\prod_{i \in S} \left((m_i - 1)!/(m_{i1}! \cdots (m_{i,i+1} - 1)! \cdots m_{id}!)\right)$$
$$\leq |\{w \in S^m\colon \tilde{L}(w) = q\}| \leq |S| \prod_{i \in S} (m_i!/(m_{i1}! \cdots m_{id}!)),$$

with $d + 1$ identified as 1 in the lower bound. [The question of counting $\{w \in S^m\colon \tilde{L}(w) = q\}$ exactly is addressed in Billingsley (1961), Baum and Eagon (1966), and Zaman (1984).] A given word $w \in S^m$ with $\tilde{L}(w) = q$ determines, for each $i \in S$, a partition of the set $\{1, 2, \ldots, m_i\}$ into subsets $S_{i1}, \ldots, S_{id}$, with $|S_{ij}| = m_{ij}$ under the condition that $k \in S_{ij}$ if the kth appearance of letter i in the word is immediately followed by letter j. The word can be reconstructed from its starting letter w_1 and these partitions; this proves the upper bound.

The lower bound is the number of words satisfying the additional conditions that the last appearance of letter 1 is followed by letter 2, the last appearance of 2 is followed by a $3, \ldots,$ with the word ending in letter d. Let $n_{ij} = m_{ij} - \delta_{i,i+1}$, with the index $d + 1$ replaced by 1, so that $n_i \equiv \Sigma_j n_{ij} = \Sigma_j n_{ji}$; i.e., $[n_{ij}]$ also satisfies the balance equations. Partition the set $\{1, 2, \ldots, n_i\}$ into subsets $S_{i1}, \ldots, S_{id}$, with $|S_{ij}| = n_{ij}$. These partitions determine a word w with $\tilde{L}(w) = q$, via the recipe: for $k \leq n_i$, the kth appearance of letter i is followed by letter j, iff $k \in S_{ij}$. The word begins with letter 1. When letter i appears for the $(1 + n_i)$th time, all n_i pairs ending in i have been used up, and we put down a letter $i + 1$ and then continue to follow the partitions. This happens first with letter 1, then letter $2, \ldots,$ then letter d, at which point the word is completed. □

THEOREM 3. *Let $X_1 X_2 \cdots$ and $Y_1 Y_2 \cdots$ be independent Markov chains on $S = \{1, 2, \ldots, d\}$. Let $P = [p_{ij}]$ and $Q = [q_{ij}]$ be the transition matrices governing X and Y, respectively, with $p_{ij} > 0$ and $q_{ij} > 0$ for all $i, j \in S$. Let π and σ be the equilibrium distributions for X and Y, and define μ and $\nu \in B \subset \Pr(S^2)$ by*

$$\mu_{ij} = \pi_i p_{ij}, \qquad \nu_{ij} = \sigma_i q_{ij}, \qquad i, j \in S.$$

Consider the substochastic matrix $R = [r_{ij}] \equiv [p_{ij} q_{ij}]$, and let p, (r_i), and (l_i) be its principal eigenvalue and corresponding left and right positive eigenvectors, normalized so that $\Sigma l_i r_i = 1$. Since $[r_{ij} r_j/(p r_i)]$ is a stochastic matrix which governs a Markov process with equilibrium $(l_i r_i)$, we define $\alpha \in B$ by

$$\alpha_{ij} = l_i r_{ij} r_j / p, \qquad i, j \in S.$$

Define $\tilde{g}\colon \Pr(S^2) \to [0, 2]$, using (16) with H replaced by $\tilde{H}$. Then for any open $U \subset \Pr(S^2)$, $M_{n,U}/(\log_{1/p} n)$ converges a.s. to $\sup_{\gamma \in U \cap B} \tilde{g}(\gamma)$. In particular, $1 = P(\lim_{n \to \infty} M_n/\log_{1/p} n = C(P, Q))$, where

$$C(P, Q) = \sup_{\gamma \in \Pr(S^2) \cap B} \min\{\log(1/p)/\tilde{H}(\gamma, \mu), \log(1/p)/\tilde{H}(\gamma, \nu),$$
$$(2 \log(1/p)/(\log(1/p) + \tilde{H}(\gamma, \alpha))\},$$

and $C(P,Q) = 2$ *if and only if both* $\tilde{H}(\alpha,\nu)$, $\tilde{H}(\alpha,\mu) \le (1/2)\log(1/p)$. *Furthermore, there is a unique* $\beta \in B$ *such that* $\tilde{g}(\beta) = C(P,Q)$ *and*

$$1 = P\Big(0 = \limsup_{n\to\infty} \big\{|\beta - \tilde{L}(w)|: w = X_{i+1} \cdots X_{i+m} = Y_{j+1} \cdots Y_{j+m}$$

$$\text{for some } 0 \le i, j \le n - m, \text{ with } m = M_n\big\}\Big).$$

If $C(P,Q) = 2$, *then* $\beta = \alpha$.

Proof. The proof follows those of Theorems 1 and 2, with minor changes such as the substitution of Lemma 4 in place of Lemma 2. In place of the events G_w involving nonoverlapping blocks of m letters, we apply Doeblin's method: Fix a letter $a \in S$ and consider blocks involving m successive returns to letter a. Details of this method in the context of matching with shifts are given in Arratia and Waterman (1985). The remaining modifications are routine. □

6. Sequences with different lengths; more than two sequences. Comparison of DNA sequences often involves two sequences with very different lengths, such as 200 and 4000. Consider the length $M(m,n)$ of the longest consecutive matching between two sequences of lengths m and n, say $X_1 \cdots X_m$ and $Y_1 \cdots Y_n$. Even in the case where all $m + n$ letters are i.i.d., the limit of the ratio $(\log m)/(\log n)$ can have a critical role in determining first, whether or not $M(m,n)$ grows asymptotically like $\log_{1/p}(mn)$, and second, the composition of the best matching word.

Proceeding as in Section 3, we analyze $M(m,n)$ according to the proportions $L(w)$ of letters within the matching word w. Thus, for $U \subset \Pr(S)$ let

$$M_U(m,n) \equiv \max\{t: X_{i+1} \cdots X_{i+t} = Y_{j+1} \cdots Y_{j+t} = w$$

$$\text{for some } w \text{ with } L(w) \in U, \text{ for some } 0 \le i \le m - t, 0 \le j \le n - t\},$$

so that when $U = \Pr(S)$, $M_U(m,n) \equiv M(m,n)$.

Theorem 4. *Assume that* $X_1, X_2, \dots$ *are i.i.d.* (μ) *and* $Y_1, Y_2, \dots$ *are i.i.d.* (ν), *with all letters independent and* $p = P(X_1 = Y_1) \in (0,1)$. *Define* $\alpha \in \Pr(S)$ *by* $\alpha_a = \mu_a \nu_a / p$. *Assume that* m *and* $n \to \infty$, *with* $(\log m)/(\log(mn)) \to \lambda \in (0,1)$. *For* $\lambda \in (0,1)$ *and* $\gamma \in \Pr(S)$ *define*

$$G(\gamma,\lambda) \equiv \min\{\lambda \log(1/p)/H(\gamma,\mu), (1-\lambda)\log(1/p)/H(\gamma,\nu), \log(1/p)/(\log(1/p) + H(\gamma,\alpha))\}. \tag{17}$$

Then for any open $U \subset \Pr(S)$, $M_U(m,n)/(\log_{1/p}(mn))$ *converges a.s. to* $\sup_{\gamma \in U} G(\gamma,\lambda)$. *In particular, with* $K(\mu,\nu,\lambda) \equiv \sup_{\gamma \in \Pr(S)} G(\gamma,\lambda) \in (0,1]$, *we have*

$$1 = P\Big(\lim_{n\to\infty} M(m,n)/\log_{1/p}(mn) = K(\mu,\nu,\lambda)\Big), \text{ and}$$

$$K(\mu,\nu,\lambda) = 1 \quad \text{iff both } H(\alpha,\mu) \le \lambda \log(1/p) \tag{18}$$

$$\text{and } H(\alpha,\nu) \le (1-\lambda)\log(1/p).$$

PROOF. The proof is very similar to the proof of Theorem 1. In place of f and g as defined at (15), we now use

$$F(\gamma,c,\lambda) \equiv H(\gamma) + 0 \wedge b(\gamma,\mu,c/\lambda) + 0 \wedge b(\gamma,\nu,c/(1-\lambda)),$$

with the idea that $F(\gamma,c,\lambda)$ represents $1/t$ times the log of the contribution to $\Sigma_w[(m\mu^t(w)) \wedge 1][(n\nu^t(w)) \wedge 1]$, from words $w \in S^t$ having $L(w)$ near γ, when $t = \lceil c\log_{1/p}(mn)\rceil$. Elementary manipulation shows that $G(\gamma,\lambda) = \inf\{c: F(\gamma,c,\lambda) < 0\}$. The correspondence with the notation of Theorem 1 is that $F(\gamma,c,\frac{1}{2}) = f(\gamma,2c)$, $2G(\gamma,\frac{1}{2}) = g(\gamma)$, and $2K(\mu,\nu,\frac{1}{2}) = C(\mu,\nu)$. □

In the special case $\mu = \nu$, Theorem 4 says that if $(\log m)/(\log(mn)) \to \lambda \in (0,1)$, then $M(m,n)$ is asymptotic to $\log_{1/p}(mn)$ iff $\lambda \in [\lambda_{cr}, 1-\lambda_{cr}]$, where $\lambda_{cr} \equiv H(\alpha,\mu)/\log(1/p) \in [0,\frac{1}{2})$. Note that in this case, with $\mu = \nu$, the following are equivalent: $\lambda_{cr} = 0$; $H(\alpha,\mu) = 0$; $\alpha = \mu$; μ is the uniform distribution on S.

If $\beta \equiv \beta(\mu,\nu,\lambda)$ is the unique distribution on S for which $G(\beta,\lambda) = \sup_{\gamma\in \Pr(S)} G(\gamma,\lambda)$, then as in Theorem 2, there is a strong law of large numbers for the composition of the best matching word: If m and $n \to \infty$ with $(\log m)/(\log(mn)) \to \lambda \in (0,1)$, then with probability one, the proportions $L(w)$ of letters within any longest matching word w common to $X_1 \cdots X_m$ and $Y_1 \cdots Y_n$ tends to β. There are examples in which β varies nontrivially with λ, even with $\mu = \nu$, such as any biased coin tossing example, with $\mu = \nu = (1-\theta,\theta)$, and $\theta \ne \frac{1}{2}$.

Theorem 1 can also be generalized to the case of $r \ge 2$ independent sequences, allowing r different distributions and r different lengths. As in Theorem 3, all of this can also be done for r independent Markov chains, allowing r different transition matrices. In either the i.i.d. or the Markov case, the expressions corresponding to F and G in the statement of Theorem 4 become quite complicated—F becomes the sum of $H(\gamma)$ plus r terms, each involving relative entropy and truncation, and the formula corresponding to (16) and (17) expresses G as a minimum of $2^r - 1$ smooth terms. The one result which remains reasonably simple is the necessary and sufficient condition for the length of the longest match to be asymptotic to $\log_{1/p}$ of the number of positions in which such a match might occur. This result is given, for the i.i.d. case, in Theorem 5.

THEOREM 5. *Suppose that for $j = 1$ to r, the letters $X_1^j, X_2^j, \ldots$ are i.i.d. (μ_j), where $\mu_1,\ldots,\mu_r$ are probability distributions on a finite alphabet S. Let $p \equiv \Sigma_{a\in S}\mu_1(a)\cdots\mu_r(a)$, and assume $p \in (0,1)$. Define $\alpha \in \Pr(S)$ by $\alpha(a) = \mu_1(a)\cdots\mu_r(a)/p$. Define the length $M \equiv M(n_1,\ldots,n_r)$ of the longest word* *appearing, for* $j = 1$ *to* r, *within the first* n_j *letters of the* jth *sequence*:

$$M \equiv \max\left\{m: \phi \neq \bigcap_{j=1 \text{ to } r} \left\{w \in S^m: X^j_{i+1} \cdots X^j_{i+m} = w, \right.\right.$$

$$\left.\left. \textit{for some } 0 \leq i \leq n_j - m\right\}\right\}.$$

Suppose that $n_1, \dots, n_r \to \infty$ *with* $(\log n_j)/(\log(n_1 \cdots n_r)) \to \lambda_j > 0$, *for* $j = 1$ *to* r. *Then there is a constant* $K \equiv K(\mu_1, \dots, \mu_r; \lambda_1, \dots \lambda_r) \in (0, 1]$ *such that*

$$1 = P\left(M/\log_{1/p}(n_1 \cdots n_r) \to K\right)$$

and

$$K = 1 \quad \textit{iff} \quad H(\alpha, \mu_j) \leq \lambda_j \log(1/p) \textit{ for } j = 1 \textit{ to } r.$$

PROOF. The argument is essentially the same as that for Theorems 1 and 4. □

Acknowledgment. Guidance and motivation for this paper came from the work of Donsker and Varadhan (1975 and 1983).

REFERENCES

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

BAHADUR, R. R. (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia.

BAUM, L. E. and EAGON, J. A. (1966). The number of circular patterns compatible with a pseudo-symmetric connected graph. *Canad. J. Math.* **18** 237–239.

BILLINGSLEY, P. (1961). Statistical methods in Markov chains. *Ann. Math. Statist.* **31** 12–40.

DONSKER, M. D. and VARADHAN, S. R. S. (1975). Asymptotic evaluation for certain Markov process expectations for large time, I. *Comm. Pure Appl. Math.* **28** 1–47.

DONSKER, M. D. and VARADHAN, S. R. S. (1983). Asymptotic evaluation for certain Markov process expectations for large time, IV. *Comm. Pure Appl. Math.* **36** 183–212.

RÉNYI, A. (1970). *Foundations of Probability.* Holden-Day, San Francisco.

SMITH, T. F., WATERMAN, M. S. and SADLER, J. R. (1983). Statistical characterization of nucleic acid sequence functional domains. *Nucleic Acids Res.* **11** 2205–2220.

ZAMAN, A. (1984). Urn models for Markov exchangeability. *Ann. Probab.* **12** 223–229.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113