The Annals of Probability
1989, Vol. 17, No. 3, 1152–1169

THE ERDÖS–RÉNYI STRONG LAW FOR PATTERN MATCHING WITH A GIVEN PROPORTION OF MISMATCHES

BY R. ARRATIA[1] AND M. S. WATERMAN[2]

University of Southern California

Consider two random sequences $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$ of i.i.d. letters in which the probability that two distinct letters match is $p > 0$. For each value a between p and 1, the length of the longest contiguous matching between the two sequences, requiring only a *proportion* a of corresponding letters to match, satisfies a strong law analogous to the Erdös–Rényi law for coin tossing. The same law applies to matching between two nonoverlapping regions within a single sequence $X_1 \cdots X_n$, and a strong law with a smaller constant applies to matching between two overlapping regions within that single sequence. The method here also works to obtain the strong law for matching between multidimensional arrays, between two Markov chains and for the situation in which a given proportion of *mis*matches is required.

1. Informal introduction. The Erdös–Rényi law [Erdös and Rényi (1970)] for coin tossing is a strong law for the behavior of the length of long success rich runs. For the length R_n of the longest run of pure heads in n tosses, the result is that with probability 1, $\lim_{n\to\infty} R_n/\log_{1/p}(n) = 1$, where $p = P(\text{heads}) > 0$. A more general result is quoted as formula (1) in Section 2 for R_n^a, the longest head rich run in which the fraction of heads is at least $a > p$. Note that $R_n \equiv R_n^1$. In many practical situations, such as manufacturing or roulette, observations are taken sequentially and each can be classified as success or failure. For these cases, it is possible to use the Erdös–Rényi law to test the hypothesis that the success probability is p. From another point of view, the Erdös–Rényi law can be used to recognize patterns of unusually long runs of successes (or failures).

Our interest is in the recognition of unusually long patterns or words common to two random sequences. The patterns are unknown prior to an examination of the sequences. The motivation for this work is the comparison of DNA sequences, which can be modeled as sequences of i.i.d. or Markov distributed letters. Evolution operates to conserve, although imperfectly, patterns important to biological function. It is a task of biology to discover these patterns and their function. In earlier work, we generalized the Erdös–Rényi law for pure head runs ($a = 1$) to exact matching patterns between two sequences. In this paper we extend those results to include matchings of quality a between p and 1.

The examples below illustrate the natural analogs of R_n^a studied in this paper. Two words form a "quality a matching" if they have the same length and the

Received September 1987.

[1] Supported by NIH Grant GM 36230 and NSF Grant DMS-88-15106. Work done in part at the IMA in Minneapolis.

[2] Supported by NIH Grant GM 36230 and a grant from the System Development Foundation.

AMS 1980 *subject classifications.* 62E20, 62P10.

Key words and phrases. Matching, large deviations, Ising model, Potts model, Hamming distance, DNA sequences, protein sequences.

fraction of matches among the pairs of letters in corresponding positions is at least a. A word in a sequence of letters is any finite contiguous subsequence. Starting with two sequences of length n, we define M_n^a, in Section 2, as the length of the longest quality a matching pair of words, one chosen from each sequence. Starting with a single sequence of length n, we define D_n^a (respectively, S_n^a), in Section 5, as the length of the longest quality a matching pair of words, chosen from distinct nonoverlapping (respectively, overlapping) blocks of positions in the single sequence.

If $X_1X_2 \cdots X_{17}$ is "we love matmatics" (blank is one letter) and $Y_1Y_2 \cdots Y_{17}$ is "statistics is fun", then

$M_{17}^1 = 4$: using $X_{14} \cdots X_{17} = Y_7 \cdots Y_{10} =$ "tics",

$M_{17}^{0.75} = 8$: $X_{10} \cdots X_{17} =$ "atmatics" matches 6/8 of $Y_3 \cdots Y_{10} =$ "atistics"

and

$M_{17}^{0.65} = 9$: $X_9 \cdots X_{17} =$ "matmatics" matches 6/9 of $Y_2 \cdots Y_{10} =$ "tatistics".

This first example, using "matmatics" instead of "mathematics", is meant to emphasize a serious limitation of the theory of approximate sequence matching in this paper—letters can be changed but not deleted. The book by Kruskal and Sankoff (1983) presents the case for considering insertions and deletions along with single letter substitutions. Some results which allow a proportion of insertions and deletions are announced in Waterman, Gordon and Arratia (1987).

If $X_1X_2 \cdots X_{37}$ is "banana probabilists statistics banana", then

$D_{37}^1 = 6$: $X_2 \cdots X_6 =$ "banana" $= X_{32} \cdots X_{37}$,

$S_{37}^1 = 3$: $X_2 \cdots X_4 =$ "ana" $= X_4 \cdots X_6$,

$S_{37}^{0.5} = 6$: $X_{17} \cdots X_{22} =$ "sts st" matches 3/6 of $X_{19} \cdots X_{24} =$ "s stat",

$S_{37}^{0.41} = 12$: $X_{11} \cdots X_{22}$ matches 5/12 of $X_{13} \cdots X_{24}$

and

$S_{37}^{0.38} = 21$: $X_2 \cdots X_{22}$ matches 8/21 of $X_4 \cdots X_{24}$.

This particular sequence was almost composed by a monkey at a typewriter [Feller (1968), page 202]—the genomic DNA of humans and chimpanzees differ by about 2% [Sibley and Alquist (1984)].

Here is an overview of this paper. The main results are Theorem 1, about matching two independent sequences, and Theorem 4, about self-overlapping repeats in a single sequence of i.i.d. letters. In Section 3 we prove the easy half of Theorem 1, the upper bound, using the natural "analysis by position." In Section 4, we explain why analysis by position fails to prove the lower bound in *some* cases, and then present a proof of the lower bound which works in all cases, using an "analysis by pattern." Theorems 3 and 4 are given in Section 5—analysis by position works easily, in Theorem 3, to prove both the lower and upper bounds for a strong law involving a nonexplicit constant. The hard work then remains, in

Theorem 4, in establishing something interesting about the constant; to give an exact analysis of a large deviation rate for a Markov chain (the one-dimensional Ising or Potts model) and to compare that rate with the large deviation rate function for coin tossing. In Sections 6, 7 and 8 we present easy extensions of Theorem 1 to multidimensional arrays of letters, to Markov chains and to matching *requiring* a given proportion of mismatches.

For more than two sequences, the notion of approximate matching can be generalized in several different ways from the case of only two sequences. Given $r \geq 2$ words of the same length t, to say that the r words form a quality a matching might reasonably be defined by any of the following four requirements:

1. For some choice of at least at of the positions $1, \dots, t$, all r words agree at those positions.
2. Each of the $\binom{r}{2}$ pairs of words forms a quality a matching.
3. For some tree connecting the labels $1, \dots, r$, each of the $r - 1$ pairs of words corresponding to an edge forms a quality a matching.
4. There exists a "consensus word" of length t which forms a quality $a/2$ matching with each of the r given words.

The fourth definition, involving a consensus pattern, is discussed further in Waterman, Galas and Arratia (1984). All four of the above definitions coincide in the case $r = 2$, except for a discrepancy with the fourth definition in cases where $\lceil at \rceil \neq 2\lceil at/2 \rceil$. Consider M_n^a, the length of the longest quality a matching common to r sequences of length n, with all rn letters i.i.d., using each of the four definitions above. In every case, it should be possible to obtain a strong law of the form as $n \to \infty$, $M_n^a/\log(n) \to K$. With the first definition, the constant is $K = r/H(a, p(r))$, where $p(r) \equiv \Sigma(\mu_l)^r$, and this strong law is provable by the method used in this paper for the special case $r = 2$. With the other three definitions, it is not easy to give an explicit formula for the K. The constant K must be determined by considering, as part of an analysis by pattern, the proportions of the various matching and nonmatching r-tuples from the alphabet, in the spirit of definition (7) below.

2. Formal introduction. Let $a \in [0, 1]$ be given. The length M_n^a of the "longest matching of quality a, allowing shifts" between two sequences $X_1 X_2 \cdots X_n$ and $Y_1 Y_2 \cdots Y_n$ is defined by

$$M_n^a \equiv \max\left\{t \colon \exists\, i, j \in [0, n - t],\, a \leq t^{-1} \sum_{1 \leq k \leq t} 1(X_{i+k} = Y_{j+k})\right\}.$$

For the sake of comparison, we also consider the length $R_n \equiv R_n^a$ of the "longest head run of quality a," in a sequence $Z_1, Z_2, \dots, Z_n$ of $\{0, 1\}$-valued random variables:

$$R_n^a \equiv \max\left\{t \colon \exists\, i \in [0, n - t],\, a \leq t^{-1} \sum_{1 \leq k \leq t} Z_{i+k}\right\}.$$

By taking $Z_i \equiv 1(X_i = Y_i)$, we have that R_n^a is the length of the "longest matching of quality a, *not* allowing shifts" between $X_1 X_2 \cdots X_n$ and $Y_1 Y_2 \cdots Y_n$.

Let $S = \{1, 2, \dots, d\}$ be a finite alphabet with $d \geq 2$ and let μ be a probability distribution on S with $\mu_l > 0$ for all $l \in S$. Assume that all letters $X_1, X_2, \dots, Y_1, Y_2, \dots$ are mutually independent, with distribution μ. Let $p \equiv P(X_1 = Y_1) = \Sigma_{l \in S}(\mu_l)^2$. We will usually take $a \in (p, 1]$.

The Erdös–Rényi law [Erdös and Rényi (1970)] for a sequence of independent tosses $Z_1, Z_2, \dots$ of a p-coin, applicable here with $Z_i \equiv 1(X_i = Y_i)$, is a description of an almost-sure growth rate for R_n^a:

$$(1) \qquad \forall\, a \in (p, 1], \qquad 1 = P\big(R_n^a/\log(n) \to 1/H(a, p)\big),$$

where

$$(2) \qquad H(a, p) \equiv (a)\log(a/p) + (1 - a)\log((1 - a)/(1 - p))$$

is the relative entropy between a p-coin and an a-coin [so that $H(1, p) = \log(1/p)$]. In this paper, we derive the analogous description for M_n^a:

THEOREM 1. $\forall\, a \in (p, 1], 1 = P(M_n^a/\log(n) \to 2/H(a, p))$.

The proof of this theorem is given in Sections 3 and 4. This result, combined with the Erdös–Rényi law, implies that $M_n^a/R_n^a \to 2$ almost surely. Loosely speaking, for each fixed quality $a > p$, allowing shifts doubles the length of the longest match.

3. Upper bound: Analysis by position. Consider the event that a matching of quality a and length t is found after positions i and j in the two sequences,

$$(3) \qquad G_{i,j}^{a,t} \equiv \Big\{ ta \leq \sum_{1 \leq k \leq t} 1\big(X_{i+k} = Y_{j+k}\big) \Big\}.$$

For all positive integers t, the elementary large deviation bound for tossing a p-coin yields, for fixed i, j, that $P(G_{i,j}^{a,t}) \leq \exp(-tH(a, p))$. Since the event $\{M_n^a \geq t\}$ is a union of no more than n^2 events of this form, we have the upper bound

$$\forall\, t \geq 1, \qquad \forall\, a \in (p, 1], \qquad P(M_n^a \geq t) \leq n^2 \exp(-tH(a, p)).$$

In particular this implies that $P(M_n^a \geq (1 + \varepsilon)\log(n^2)/H(a, p)) \leq n^{-2\varepsilon} \to 0$ for all $\varepsilon > 0$. Using the Borel–Cantelli lemma along an exponentially increasing skeleton of times, such as $n_k = 2^k$, we obtain the almost sure result

$$\forall\, \varepsilon > 0, \quad \forall\, a \in (p, 1], \quad 1 = P\big(M_n^a \leq (1 + \varepsilon)\log(n^2)/H(a, p) \text{ eventually}\big).$$

4. Lower bound: Analysis by pattern. Let $\varepsilon > 0$ be given and let

$$(4) \qquad t = \left\lceil \frac{(1 - \varepsilon)\log(n^2)}{H(a, p)} \right\rceil.$$

Our goal is to show that $P(M_n^a \geq t) \to 1$.

The "natural" way to attempt this, which is carried out in Arratia and Waterman (1985a) for the extreme case, $a = 1$ is to continue the "analysis by position" which gave us the lower bound of the previous section. We use the "first and second moments method" together with nonoverlapping blocks. In detail, consider the random variable $N = N(n, t, a)$ defined by

$$N = \sum_{0 \le i,\, j \le (n-t)/t} 1\left(G^{a,t}_{it,\, jt}\right),$$

where the event $G^{a,t}_{i,j}$ is defined at (3) above. We have $EN \to \infty$ and $\{N > 0\} \subset \{M^a_n \ge t\}$, and we would like to show that $\operatorname{var}(N)/(EN)^2 \to 0$ in order to conclude that $P(N > 0) \to 1$. Thanks to the use of nonoverlapping blocks, most of the asymptotically $(n/t)^4$ terms in the expansion of $\operatorname{var}(N)$ are zero, but there are still asymptotically $2(n/t)^3$ positive off-diagonal terms, from indices (i, j) and (i', j') with $i = i'$ or else $j = j'$. In the extreme case $a = 1$, it is true that $\operatorname{var}(N)/(EN)^2 \to 0$, because $p_3 \equiv \Sigma(\mu_l)^3 < p^{3/2}$, which is a consequence of Jensen's inequality. (See Section 8 below for a discussion of the case corresponding to the opposite extreme, $a = 0$.) However, there exist cases of the three parameters, a, p and p_3 with $0 < p < a < 1$, in which for all sufficiently small $\varepsilon > 0$, $\operatorname{var}(N)/(EN)^2 \to \infty$. In these cases, the "natural" strategy fails. For $0 < p < a < 1$, the necessary and sufficient condition for this failure to occur is that

$$\lim\left\{\log P\left(G^{a,t}_{0,0} \cap G^{a,t}_{0,t}\right)/\log P\left(G^{a,t}_{0,0}\right)\right\} < 3/2,$$

which is equivalent to

$$\inf_b \left\{ aH\left(b, \frac{p_3}{p}\right) + (1 - a)H\left(\frac{a(1-b)}{1-a}, \frac{p - p_3}{1 - p}\right)\right\} \le \frac{H(a, p)}{2},$$

where the infimum is taken over $b \in [p_3/p, 1]$ such that $a(1 - b)/(1 - a) \in [(p - p_3)/(1 - p), 1]$. In all cases where the natural strategy succeeds, the analysis by position can be refined to approximate the distribution of M^a_n; this is carried out in Arratia, Gordon and Waterman (1988). The framework for getting distributional results when first and second moments can be controlled is presented in Arratia, Goldstein and Gordon (1989).

Instead of the natural analysis by position described in the paragraph above, in this section we establish the lower bound using an analysis by pattern. For the case $a = 1$, such an analysis is used in Arratia and Waterman (1985b) in order to derive strong laws for M^1_n under the added complexity of different lengths or distributions for the two sequences; Markov chains and more than two sequences are also handled there at no additional cost. Because the modifications required to make the analysis by pattern work for the case $a < 1$ are subtle and orthogonal to the techniques of Arratia and Waterman (1985b), we confine this paper to the simplest setup: two sequences having the same length and distribution.

For words $w, z \in S^t$, let $A_w \equiv A(w, n, t)$ (respectively, B_z) be the event that word w (respectively, z) occurs within the first n letters of the sequence X

(respectively, Y) following a position which is a multiple of t, i.e.,

$$A_w \equiv \bigcup_{0 \le i \le (n-t)/t} \{w = X_{ti+1} \cdots X_{ti+t}\},$$

$$B_z \equiv \bigcup_{0 \le i \le (n-t)/t} \{z = Y_{ti+1} \cdots Y_{ti+t}\}.$$

Thanks to the use of nonoverlapping blocks in the definition of A_w, distinct events A_w and $A_{w'}$ are negatively correlated: For $w \ne w' \in S^t$, $P(A_w \cap A_{w'}) \le P(A_w)P(A_{w'})$. Define

$$(5) \qquad E_{w,z} \equiv A_w \cap B_z,$$

so that $E_{w,z}$ is the event that for $X_1 \cdots X_n$ and $Y_1 \cdots Y_n$, each blocked off into words of length t, the words w and z are found in X and Y, respectively. The sequences X and Y are independent, so that $P(E_{w,z}) = P(A_w)P(B_z)$. For $w, w', z, z' \in S^t$, if $w \ne w'$ and $z \ne z'$, the events $E_{w,z}$ and $E_{w',z'}$ are negatively correlated. However, there is very strong positive correlation between distinct events $E_{w,z}$ and $E_{w',z'}$ with $w = w'$ or else $z = z'$. Specifically, in the case $w = w'$ and $z \ne z'$ we have $P(E_{w,z}|E_{w,z'})/P(E_{w,z}) = P(B_z|B_{z'})/(P(A_w)P(B_z)) \sim 1/P(A_w)$.

Let $I \equiv I(t, a)$ be the set of pairs of words of length t which match each other with quality a, i.e., $I \equiv \{(w, z) \in (S^t)^2 \colon ta \le \Sigma_{1 \le i \le t} 1(w_i = z_i)\}$. The motivation here is that $\bigcup_{(w,z) \in I} E_{w,z} \subset \{M_n^a \ge t\}$, so it suffices to show that $P(\bigcup_{(w,z) \in I} E_{w,z}) \to 1$. Our choice of t, specified at (4), yields $\Sigma_{(w,z) \in I} P(E_{w,z}) \to \infty$, so we would be finished if we could show that the events $E_{w,z}$ are not too much positively correlated with each other, but we cannot do even this.

Here is an outline of our strategy to get past the difficulty of positive correlation between events $E_{w,z}$ and $E_{w',z'}$ with $w = w'$ or else $z = z'$. We consider a subset $J \subset I$ of the induces, which makes the dominant contribution to $\Sigma_{(w,z) \in I} P(E_{w,z})$, so that $\Sigma_{(w,z) \in J} P(E_{w,z}) \to \infty$. The set J is symmetric, i.e., $(w, z) \in J$ iff $(z, w) \in J$. All words w such that $(w, z) \in J$ for some z have the same composition, so that for all $(w, z) \in J$, $P(E_{w,z}) = P(A_w)P(B_z) = (P(A_w))^2$ has the same value. Thus for $(w, z) \ne (w', z') \in J$ with $w = w'$ or else $z = z'$, $P(E_{w,z} \cap E_{w',z'}) = (P(A_w))^3 = (P(E_{w,z}))^{3/2}$. The set J is homogeneous, in the sense that $|\{z \colon (w, z) \in J\}|$ has the same value for every w such that $(w, z) \in J$ for some z. The symmetry and homogeneity of J imply that $|\{((w, z), (w', z')) \in J^2 \colon w = w' \text{ or else } z = z'\}| \le 2(|J|)^{3/2}$.

To see that the above strategy results in an acceptable amount of positive correlation, define the random variable

$$T \equiv T(n, t, a) \equiv \sum_{(w,z) \in J} 1(E_{w,z}),$$

so that $\{T > 0\} \subset \{M_n^a \ge t\}$. Consider the expansion

$$\operatorname{var}(T) = \sum_{(w,z),(w',z') \in J} \operatorname{cov}(1(E_{w,z}), 1(E_{w',z'})).$$

The contribution from the diagonal terms is less than ET. Each term in which $w \neq w'$ and $z \neq z'$ is negative. There are fewer than $2(|J|)^{3/2}$ terms in which $w = w'$ or else $z = z'$, and each of these terms is less than $(P(E_{w,z})^{3/2}$, so the net contribution from these terms is less than $2(ET)^{3/2}$. Thus $\operatorname{var}(T) \leq ET + 2(ET)^{3/2}$ and by Chebyshev's inequality,

$$(6) \qquad P(M_n^a < t) \leq P(T = 0) \leq \operatorname{var}(T)/(ET)^2 = O\big((ET)^{-1/2}\big) \to 0.$$

We proceed to define $J \equiv J(t, a)$. The distribution μ of letters in our two sequences determines a probability distribution α on S, and a probability distribution γ on S^2, corresponding to simple matches and simple mismatches, as follows: For $b, c \in S$,

$$\alpha_b \equiv P(X_1 = b | X_1 = Y_1) = (\mu_b)^2/p,$$

$$\gamma_{bc} \equiv P(X_1 = b, Y_1 = c | X_1 \neq Y_1) = 1(b \neq c)\mu_b \mu_c/(1 - p).$$

Define

$$(7) \qquad J \equiv \Big\{(w, z) \in (S^t)^2 \colon \sum_{1 \leq i \leq t} 1(b = w_i = z_i) = \lfloor at\alpha_b \rfloor, \text{ for } b \in S,\ b \neq d, \text{ and } \sum_{1 \leq i \leq t} 1(b = w_i,\ c = z_i) = \lfloor (1 - a)t\gamma_{bc} \rfloor, \text{ for } b \neq c \in S \Big\}.$$

Thus, each pair of words $(w, z) \in J$ is required to match in the same number $s \equiv t - \Sigma_{b \neq c \in S} \lfloor (1 - a)t\gamma_{bc} \rfloor$ of places, with $s \geq at$ and $s/t \to a$ as $t \to \infty$. The letters in those s places are required to have a fixed empirical distribution, close to α. At positions which do not match, the pairs of letters which appear are required to have a fixed empirical distribution, close to γ.

First we find the growth rate for $|J|$. Let $H(\nu) = -\Sigma \nu_b \log \nu_b$ denote the entropy of a probability distribution ν and let $H(a) = -a \log a - (1 - a)\log(1 - a)$ denote the symmetric entropy for $a \in [0, 1]$. We have, by Stirling's formula, that

$$(8) \qquad t^{-1} \log(|J|) \to H(a) + aH(\alpha) + (1 - a)H(\gamma).$$

In this limit, the first term is $H(a) = \lim t^{-1} \log\binom{t}{s}$, which corresponds to choosing which positions will match. The second term involves a multinomial coefficient which corresponds to choosing which letters to assign to these matching positions, and the last term corresponds to choosing which pairs of letters appear in each of the nonmatching positions.

Next we find the decay rate for $P(E_{w,z})$. Let $(w, z) \in J$, so that $P(E_{w,z}) = (P(A_w))^2$. Let $k = \lfloor n/t \rfloor$, so that $P(A_w) = 1 - (1 - (\mu^t(w))^k)$. Since $[1 - (1 - z)^k]/[1 \wedge kz] \in [1/2, 1]$ for $k = 1, 2, \ldots$ and for all $z \in [0, 1]$, we have $\lim t^{-1}\log P(A_w) = 0 \wedge \lim t^{-1}[\log(n/t) + \log(\mu^t(w))]$. Thus

$$\begin{aligned}
&\lim t^{-1}\log P(E_{w,z})\\
&\quad = 0 \wedge \lim t^{-1}\big[\log(n^2) + \log\big(\mu^t(w)\mu^t(z)\big)\big]\\
&\quad = 0 \wedge \Big[H(a,p)/(1-\varepsilon) + a\sum \alpha_b \log\big((\mu_b)^2\big) + (1-a)\sum \gamma_{bc}\log(\mu_b\mu_c)\Big]\\
&\quad = 0 \wedge \big[H(a,p)/(1-\varepsilon) + a(\log(p) - H(\alpha))\\
&\qquad\qquad + (1-a)(\log(1-p) - H(\gamma))\big]\\
&\quad = 0 \wedge \big[H(a,p)/(1-\varepsilon) - H(a,p) - H(a) - aH(\alpha) - (1-a)H(\gamma)\big].
\end{aligned}$$

For sufficiently small positive ε, the expression in brackets is negative, so the truncation with 0 may be ignored:

$$\begin{aligned}
&\lim t^{-1}\log P(E_{w,z})\\
&\quad = \varepsilon H(a,p)/(1-\varepsilon) - H(a) - aH(\alpha) - (1-a)H(\gamma).
\end{aligned}\tag{9}$$

Now $ET = |J|P(E_{w,z})$, so combining (8) with (9) yields, for sufficiently small positive ε,

$$\lim_{n\to\infty} t^{-1}\log(ET) = \varepsilon H(a,p)/(1-\varepsilon) > 0.$$

Using (6) and the Borel–Cantelli lemma along an exponentially increasing skeleton of times, such as $n_k = 2^k$, we obtain the almost sure result

$$\forall\, \varepsilon > 0, \quad \forall\, a \in (p,1], \quad 1 = P\big(M_n^a \ge (1-\varepsilon)\log(n^2)/H(a,p) \text{ eventually}\big).$$

This completes the proof of Theorem 1. □

5. Repeats, self-overlapping and otherwise, within a single sequence of i.i.d. letters. The length D_n^a of the "longest matching of quality a" (not allowing self-overlap) within a single sequence $X_1X_2\cdots X_n$ is defined by

$$\begin{aligned}
D_n^a \equiv \max\Big\{t\colon &\exists\, i,j \in [0, n-t],\ |i-j| \ge t,\\
&a \le t^{-1}\sum_{1\le k\le t} 1\big(X_{i+k} = X_{j+k}\big)\Big\}.
\end{aligned}\tag{10}$$

This section was inspired by the innocent and natural question: With i.i.d. letters, in giving a strong law for $D_n^a/\log(n)$, would it make a difference if D_n^a had been defined with the restriction $|i-j| > 0$ instead of $|i-j| \ge t$? The answer, which turns out to be "no," is a consequence of Theorem 4. Since there are $\approx n^2$ places to locate a long matching, versus $\approx n$ places to locate a long self-overlapping matching, the question boils down to whether the large deviation rate for matching independent sequences is twice as large as the large deviation rate for self-overlapping matching.

THEOREM 2. $\forall\, a \in (p,1]$, $1 = P(D_n^a/\log(n) \to 2/H(a,p))$.

PROOF. The proof is almost identical to that of Theorem 1, with the sequence X playing the roles of both sequences X and Y. This gives us, instead of (5), that $E_{w,z} \equiv A_w \cap A_z$, with $P(E_{w,z}) < P(A_w)P(A_z)$ (instead of the equality in the proof of Theorem 1), but we still have, for $(w,z) \neq (w',z') \in J$, that $P(E_{w,z}) \sim P(A_w)^2$ and $P(E_{w,z} \cap E_{w',z'}) < P(E_{w,z})^{3/2}$. If also $w \neq w'$ and $z \neq z'$, then the events $E_{w,z}$ and $E_{w',z'}$ are negatively correlated. □

The length S_n^a of the "longest self-overlapping matching of quality a," within a single sequence $X_1X_2 \cdots X_n$ is defined by

$$S_n^a \equiv \max\Big\{t\ \exists\, i,j \in [0,n-t], 0 < |i-j| < t, \qquad (11)$$
$$a \le t^{-1} \sum_{1 \le k \le t} 1(X_{i+k} = X_{j+k})\Big\}.$$

THEOREM 3. $\forall\, a \in (p,1]$, $1 = P(S_n^a/\log(n) \to 1/r(a,\mu))$, *where* $r(a,\mu)$ *is the large deviation rate characterized by*

$$\forall\, a \in [p,1], \quad r(a,\mu) \equiv \lim(-1/t)\log\Big\{P\Big(at \le \sum_{1 \le i \le t} 1(X_i = X_{i+1})\Big)\Big\}, \qquad (12)$$
$$\forall\, a \in [0,p], \quad r(a,\mu) \equiv \lim(-1/t)\log\Big\{P\Big(at \ge \sum_{1 \le i \le t} 1(X_i = X_{i+1})\Big)\Big\}.$$

PROOF. The lower bound is easily established using nonoverlapping blocks, as follows. Let $\varepsilon > 0$ be given, let $t = \lfloor (1-\varepsilon)\log(n)/r(a,\mu) \rfloor$ and let

$$N = \sum_{0 \le i \le (n-t-1)/(t+1)} 1\Big(ta \le \sum_{1 \le k \le t} 1\big(X_{(t+1)i+k} = X_{(t+1)i+1+k}\big)\Big),$$

so that $\{S_n^a/\log(n) < (1-\varepsilon)/r(a,\mu)\} \subset \{N = 0\}$. We have $EN \to \infty$ as $n \to \infty$, where N counts the number of independent events that occur, so $P(N=0) < e^{-EN}$ and the Borel–Cantelli lemma implies

$$P(S_n^a/\log(n) < (1-\varepsilon)/r(a,\mu) \text{ i.o.}) = 0.$$

The upper bound is not as straightforward as Section 2 because the rate $r(a,\mu)$ corresponds directly only to those cases with shift $|i-j| = 1$ in the definition (11). Define the number of matches between a block of length t and the same length block shifted by m,

$$U(t,m) \equiv \sum_{k=1}^{t} 1(X_k = X_{m+k}), \qquad (13)$$

so that the specification (12) of the rate r states, for $a \in (p,1]$, that $r(a,\mu) = \lim\{-t^{-1}\log[P(ta \le U(t,1))]\}$. We will show that

$$\forall\, t, m \ge 1, \qquad P(ta \le U(T,m)) \le e^{-tr(a,\mu)}, \qquad (14)$$

and hence

$$\forall\, t \geq 1, \qquad a \in (p,1], \qquad P(S_n^a \geq ta) \leq 2tne^{-tr(a,\mu)}.$$

From this it follows, as in Section 2, that

$$\forall\, \varepsilon > 0, \qquad a \in (p,1], \qquad 1 = P(S_n^a \leq (1+\varepsilon)\log(n)/r(a,\mu) \text{ eventually}).$$

To prove (14), we use the Laplace transform. For $\beta \in R$, define the "transfer matrix" $M(\beta)$,

$$(15) \qquad \forall\, i, j \in \{1,\dots,d\}, \qquad M(\beta)_{ij} \equiv \sqrt{\mu_i \mu_j}\exp(\beta 1(i=j))$$

and let $\lambda(\beta)$ denote its spectral radius,

$$(16) \qquad \lambda(\beta) \equiv |||M(\beta)|||.$$

Let $\mathbf{u}$ be the d-dimensional unit vector with components $\mathbf{u}_i \equiv \sqrt{\mu_i}$. We have

$$(17) \quad \forall\, \beta \in R, \qquad \forall\, t \geq 1, \qquad Ee^{\beta U(t,1)} = \sum_{i,j} \mathbf{u}_i [M^t(\beta)]_{ij} \mathbf{u}_j \leq \lambda^t(\beta).$$

Now for $t, m \geq 1$, the random variable $U(t,m)$ can be expressed as a sum of $m \wedge t$ independent random variables, each of which has the same distribution as $U(s,1)$ for some $s \geq 1$, and the values of s that occur sum to t. For example,

$$U(7,3) = \{1(X_1 = X_4) + 1(X_4 = X_7) + 1(X_7 = X_{10})\}$$
$$+ \{1(X_2 = X_5) + 1(X_5 = X_8)\} + \{1(X_3 = X_6) + 1(X_6 = X_9)\}$$

so

$$Ee^{\beta U(7,3)} = Ee^{\beta U(3,1)}\{Ee^{\beta U(2,1)}\}^2.$$

Thus, using (17) repeatedly,

$$(18) \qquad \forall\, \beta \in R, \qquad \forall\, t, m \geq 1, \qquad Ee^{\beta U(t,m)} \leq \lambda^t(\beta).$$

The above inequality means that the usual exponential upper bound applies *uniformly* in the amount m of shift: $\forall\, a \in (p,1]$, $\forall\, t, m \geq 1$,

$$P(ta \leq U(t,m)) \leq \inf_\beta e^{-t\beta} Ee^{\beta U(t,m)} \leq \inf_\beta e^{-t\beta}\lambda^t(\beta) = e^{-tr(a,\mu)},$$

which proves (14). The final equality above is discussed further at (22)–(24) below. □

As a prelude to Theorem 4, we observe that for the case of perfect matching, i.e., $a = 1$, non-self-overlapping repeats grow faster than self-overlapping repeats, because $p \equiv \Sigma\mu_l^2 > (\max \mu_l)^2$, and hence $r(1,\mu) = -\log(\max \mu_l) > -\log\sqrt{p} = \frac{1}{2}H(1,p)$.

THEOREM 4. $\forall\, a \in (p,1]$,

$$(19) \qquad r(a,\mu) > \tfrac{1}{2}H(a,p).$$

Hence as $n \to \infty$,

$$(20) \qquad 1 = P(\lim(D_n^a/S_n^a) = 2r(a,\mu)/H(a,p) > 1),$$

$$(21) \qquad 1 = P(D_n^a > S_n^a \textit{ eventually}).$$

PROOF. Using Theorems 2 and 3, (20) is a consequence of (19) and obviously (20) implies (21). Here is an overview of our proof of (19). Inequality (19) for a single value of a is too difficult to prove directly, except in the special case $d = 2$—which is possible but messy; try it for yourself. However, it is relatively easy to prove that $\inf_{a \in (p,1]}\{r(a,\mu) - \frac{1}{2}H(a,p)\} \geq 0$, because each large deviation rate function, r and $\frac{1}{2}H$, is the Legendre transform of a corresponding free energy function, and then Fenchel's duality relation equates the infimum of the difference with the infimum of the opposite difference of the transforms. That this opposite difference is nonnegative can be verified and a little more argument gets the strict inequality for (19).

Write $U(t)$ [$\equiv U(t,1)$ in formula (13)] for the number of matches (i.e., the "energy") in an interval of $t+1$ letters and t bonds, and for $\beta \in R$, write $Z(t,\beta)$ for the corresponding "partition function",

$$(22) \qquad U(t) \equiv \sum_{k=1}^{t} 1(X_k = X_{+k}), \qquad Z(\beta,t) \equiv Ee^{\beta U(t)}.$$

Define the "free energy function" f,

$$(23) \qquad \forall\, \beta \in R, \quad f(\beta) \equiv \lim t^{-1} \log Z(\beta,t).$$

The setup above is called the one-dimensional Ising model, in case $d = 2$, and the one-dimensional Potts model, in case $d \geq 2$. Standard statistical mechanics, or large deviation theory if you prefer, asserts that f and r are convex and each other's Legendre transform: $f = r^*$ and $r = f^*$, i.e., $\forall\, a \in (0,1)$,

$$(24) \qquad r(a,\mu) = \sup_{\beta \in R}\{a\beta - f(\beta)\} = a\beta_a - f(\beta_a),$$

where β_a is the solution of $f'(\beta) = a$. We observe that the one-to-one correspondence $a \leftrightarrow \beta_a$ has $0 \leftrightarrow -\infty$, $p \leftrightarrow 0$ and $1 \leftrightarrow \infty$. From (15)–(17) we see that $f(\beta) = \log(\lambda(\beta))$.

Considerations like the above, with the indicators $1(X_k = X_{1+k})$ replaced by *independent* p-coins, show that the large deviation rate function for tossing a p-coin, $H(\cdot) \equiv H(\cdot, p)$, defined at (2), belongs to a Legendre transform pair: $G = H^*$ and $H = G^*$, i.e.,

$$H(a,p) = \sup_{\beta}\{a\beta - G(\beta)\}, \quad \text{where } G(\beta) \equiv \log(pe^{\beta} + 1 - p).$$

We are interested in $\frac{1}{2}H(a,p) = \frac{1}{2}\sup_\beta\{a\beta - G(\beta)\} = \sup_\beta\{a(\beta/2) - \frac{1}{2}G(2(\beta/2))\} = \sup_\beta\{a\beta - \frac{1}{2}G(2\beta)\}$, so we define functions h on $(0,1)$ and g on R which form a Legendre transform pair: $h = g^*$, $g = h^*$, where

$$h(a) \equiv \tfrac{1}{2}H(a,p), \qquad g(\beta) \equiv \tfrac{1}{2}G(2\beta) = \log\sqrt{pe^{2\beta} + 1 - p}\,.$$

Fenchel's duality relation states that

$$(25)\quad \inf_{a\in(0,1)}\{r(a,\mu)-\tfrac{1}{2}H(a,p)\} \equiv \inf_{a\in(0,1)}\{f^*-g^*\} = \inf_{\beta\in R}\{g-f\}.$$

The matrix $M(\beta)$ is real and symmetric, so all its eigenvalues are real and hence $\lambda^2(\beta) \le \mathrm{Tr}(M^2(\beta))$. We compute

$$M^2(\beta)_{ii} = \sum_j M_{ij}M_{ji} = \sum_j \mu_i\mu_j + (e^{2\beta}-1)\mu_i^2 = \mu_i + (e^{2\beta}-1)\mu_i^2,$$

so that $\mathrm{Tr}(M^2(\beta)) = \Sigma_i(\mu_i + (e^{2\beta}-1)\mu_i^2) = 1 + (e^{2\beta}-1)p = e^{2g(\beta)}$. Thus

$$e^{2f(\beta)} = \lambda^2(\beta) \le \mathrm{Tr}(M^2(\beta)) = e^{2g(\beta)},$$

hence $\forall\ \beta,\ f(\beta) \le g(\beta)$. Since $f(0) = 0 = g(0)$, the common value of the infimum in (25) is zero.

To prove the strict inequality (19), assume that $a_0 \in (0,1]$ satisfies $r(a_0,\mu) - \frac{1}{2}H(a_0,p) = 0$. We observed, just before stating Theorem 4, that $a_0 \ne 1$, so a_0 is an interior extremum of $r - h$, and hence $r'(a_0) = h'(a_0)$; call the common value β_0. Since $f = r^*$ and $g = h^*$, we have $f(\beta_0)$ $\sup_a\{\beta_0 a - r(a,\mu)\} = \beta_0 a_0 - r(a_0,\mu)$, and similarly $g(\beta_0) = \beta_0 a_0 - h(a_0)$. Subtracting yields $f(\beta_0) - g(\beta_0) = h(a_0) - r(a_0,\mu) = 0$, hence $f(\beta_0) = g(\beta_0)$, hence $\lambda^2(\beta_0) = \mathrm{Tr}(M^2(\beta_0))$. This implies that $M(\beta_0)$ has rank 1, hence $\beta_0 = 0$, hence $a_0 = p$. This proves (19). □

6. Matching multidimensional arrays of i.i.d. letters. All of the results of the previous sections generalize easily to the case of multidimensional arrays of i.i.d. letters, which shows that the geometry of the index set does not play a significant role. Even the multidimensional version of Theorem 3, which appears to involve overlapping cubes, really only involves matches along one-dimensional chains of sites, regardless of the dimension of the cubes.

Fix an integer $\nu \ge 1$ to serve as the number of dimensions; the case $\nu = 2$ corresponds to matching discretized pictures. Assume that all letters $X_{\mathbf{i}}, Y_{\mathbf{i}}$ for $\mathbf{i} \in Z^\nu$ are mutually independent, with distribution μ as before. Use boldface to denote vectors in Z^ν, so that $\mathbf{1} \equiv (1,1,\dots,1)$, and take the usual coordinatewise partial order on Z^ν, so that the interval $[\mathbf{1}, t\mathbf{1}]$ is a cube in Z^ν containing t^ν sites. We generalize the definitions of M_n^a, R_n^a, D_n^a and S_n^a, by taking all of the indices and endpoints for intervals to be elements of Z^ν, taking n to be the *length* of the side of the large cubes over which we look for the largest quality a matching subcubes and measuring these small subcubes in terms of their *volume*. Thus we define R_n^a (respectively, M_n^a), the volume of the largest quality a matching cube, not allowing shifts (respectively, allowing shifts) between the two cubes of side n, $\{X_{\mathbf{i}}: \mathbf{i} \in [\mathbf{1}, n\mathbf{1}]\}$ and $\{Y_{\mathbf{i}}: \mathbf{i} \in [\mathbf{1}, n\mathbf{1}]\}$,

$$R_n^a \equiv \max\Big\{t^\nu: \exists\ \mathbf{i} \in [0,(n-t)\mathbf{1}],\ a \le t^{-\nu} \sum_{\mathbf{1}\le\mathbf{k}\le t\mathbf{1}} 1(X_{\mathbf{i}+\mathbf{k}} = Y_{\mathbf{i}+\mathbf{k}})\Big\},$$

$$M_n^a \equiv \max\Big\{t^\nu: \exists\ \mathbf{i},\mathbf{j} \in [0,(n-t)\mathbf{1}],\ a \le t^{-\nu} \sum_{\mathbf{1}\le\mathbf{k}\le t\mathbf{1}} 1(X_{\mathbf{i}+\mathbf{k}} = Y_{\mathbf{j}+\mathbf{k}})\Big\}.$$

Similarly we define D_n^a (respectively, S_n^a), the volume of the largest quality a matching pair of nonoverlapping (respectively, overlapping) cubes, inside the cube of side n, $\{X_{\mathbf{i}}: \mathbf{i} \in [\mathbf{1}, n\mathbf{1}]\}$,

$$D_n^a \equiv \max\Big\{t^\nu: \exists\, \mathbf{i},\mathbf{j} \in [0,(n-t)\mathbf{1}],$$

$$\|\mathbf{i}-\mathbf{j}\| \geq t,\ a \leq t^{-\nu} \sum_{\mathbf{1} \leq \mathbf{k} \leq t\mathbf{1}} 1(X_{\mathbf{i}+\mathbf{k}} = X_{\mathbf{i}+\mathbf{k}})\Big\},$$

$$S_n^a \equiv \max\Big\{t^\nu: \exists\, \mathbf{i},\mathbf{j} \in [0,(n-t)\mathbf{1}],$$

$$0 < \|\mathbf{i}-\mathbf{j}\| < t,\ a \leq t^{-\nu} \sum_{\mathbf{1} \leq \mathbf{k} \leq t\mathbf{1}} 1(X_{\mathbf{i}+\mathbf{k}} = X_{\mathbf{i}+\mathbf{k}})\Big\}.$$

In the above definitions, $\|\cdot\|$ is the sup norm; $\|\mathbf{i}-\mathbf{j}\| = \max_{1 \leq \mathbf{k} \leq \nu}|i_k - j_k|$.

The heuristic which easily suggests that $(M_n^a/\log(n) \to 2\nu/H(a, p))$ is still an analysis by position is stated as follows: The event $\{M_n^a \geq t^\nu\}$ is a union of $\approx n^{2\nu}$ (which counts the number of choices for the locations $\mathbf{i},\mathbf{j}$ of the subcubes) not too dependent simple events, each of probability $\approx \exp(-t^\nu H(a, p))$. Thus the expected number of simple events has order 1, rather than zero or infinity, iff $1 \approx n^{2\nu}\exp(-t^\nu H(a, p))$. Taking logarithms, this condition becomes $2\nu\log(n) \sim t^\nu H(a, p)$. The discussion in the second paragraph of Section 4 shows that the notion "not too dependent" can be made the key to a rigorous proof that $(M_n^a/\log(n) \to 2\nu/H(a, p))$ in probability in some but not all cases of the parameters a and μ.

The multidimensional analog of the Erdös–Rényi law,

$$\forall\, a \in (p,1], \qquad 1 = P\big(R_n^a/\log(n) \to \nu/H(a, p)\big),$$

is proved in Darling and Waterman (1985). By combining this with Theorem 5 below, we get a statement in which the dimension ν does not appear: $\forall\, a \in (p,1]$, $1 = P(M_n^a/R_n^a \to 2)$. Loosely speaking, allowing shifts doubles the volume of the largest quality a matching for each $a > p$.

Theorem 5. *Theorems* 1–4 *for i.i.d. letters generalize to* $\nu \geq 1$ *dimensions*: $\forall\, a \in (p,1]$,

$$1 = P\big(M_n^a/\log(n) \to 2\nu/H(a, p)\big),$$

$$1 = P\big(D_n^a/\log(n) \to 2\nu/H(a, p)\big),$$

$$1 = P\big(S_n^a/\log(n) \to \nu/r(a, \mu)\big)$$

and hence

$$1 = P\big(D_n^a/S_n^a \to 2r(a, \mu)/H(a, p) > 1\big).$$

Proof. There are no essential changes from the proof given in the one-dimensional case. There are notational changes: t now becomes t^ν and a "word" $w \in S^t$, as in the definition (50), becomes a "pattern" of letters arranged on the cube of side t: $w \in S^{[\mathbf{1}, t\mathbf{1}]}$. In Section 4, the quantity $U(t, m)$ defined at (13) is now generalized to the number of matches between a cube of side t and the same cube shifted by $\mathbf{m} \in Z^{\nu}$:

$$U(t,\mathbf{m}) \equiv \sum_{\mathbf{k}\in[\mathbf{1},\,t\mathbf{1}]} 1(X_{\mathbf{k}} = X_{\mathbf{m}+\mathbf{k}}).$$

Now $U(t,\mathbf{m})$ can be expressed as a sum of independent random variables, each of which has the same distribution as the *one-dimensional* $U(s,1)$ for some s. The values of s that occur sum to t^{ν} (each one is at most t), so that (17) can be used repeatedly to get the bound analogous to (18). For example, with $\nu = 2$, $t = 3$ and $\mathbf{m} = (1,1)$, we have

$$Ee^{\beta U(3,(1,1))} = Ee^{\beta U(3,1)}\{Ee^{\beta U(2,1)}\}^2\{Ee^{\beta U(1,1)}\}^2 \le \lambda^9(\beta). \qquad \square$$

7. Matching between two Markov chains. It is possible to generalize Theorems 1 and 2, but not Theorems 3 and 4, to the case of Markov chains. Assume, as in Sections 2–5, that the alphabet is $S = \{1,2,\dots,d\}$ and that the sequences $X_1, X_2,\dots$ and $Y_1, Y_2,\dots$ are independent, but generalize and take each sequence to be governed by an irreducible and aperiodic d by d stochastic matrix $P \equiv [P_{lm}]$, with equilibrium distribution μ.

As before, $p \equiv \Sigma(\mu_l)^2$ gives the value of $P(X_i = Y_j)$, provided that both chains are stated in equilibrium. For $a \in (p,1]$, the large deviation rate for the two chains to match with quality a is

$$(26) \qquad H(a;[P]) \equiv \lim_{t\to\infty} -t^{-1}\log P\Bigl(a \le t^{-1}\sum_{1\le k\le t} 1(X_k = Y_k)\Bigr).$$

Theorems 1 and 2 hold for Markov chains, provided that $H(a,p)$ is replaced by $H(a;[P])$.

The quantity $H(a;[P])$ is really a large deviation rate for the Markov chain with state space S^2, whose d^2 by d^2 transition matrix Q is given by $[Q_{(i,j),(k,l)}] = [P_{ik}P_{jl}]$. The equilibrium for Q puts mass p on the diagonal $D \equiv \{(l,m)\colon l = m\} \subset S^2$, and we are considering large deviations in which the occupation measure for the chain governed by Q puts mass $a > p$ on the diagonal D of the state space.

In the special case that each Markov chain is an i.i.d. sequence, this reduces to the relative entropy $H(a,p)$ for coin tossing, defined at formula (2). In general, large deviation theory tells us that the limit defining $H(a;[P])$ exists, that its value for $a \in (p,1]$ is strictly positive and finite and that the function $H(\cdot;[P])$ is convex. There is a variational formula for H, from which the value of H may be numerically computed, but it seems to us that for $a < 1$, *except* in the cases of i.i.d. sequences or the symmetric 2 by 2 transition matrix P, it is not possible to give an explicit formula for $H(a;[P])$. For $a = 1$, $H(1;[P]) = -\log(\lambda)$, where λ is the spectral radius of the Schur product of P with itself, i.e., the substochastic d by d matrix $[P_{lm}^2]$. Properties of the Schur powers of a stochastic matrix are discussed in Karlin (1985).

THEOREM 6. *For the irreducible aperiodic Markov chain described above,*

$$\forall\, a \in (p,1], \quad 1 = P(M_n^a/\log(n) \to 2/H(a;[P])),$$
$$\forall\, a \in (p,1], \quad 1 = P(D_n^a/\log(n) \to 2/H(a;[P])).$$

PROOF. The proof of the upper bound is just like Section 3. Care must be taken in proving the statement about D_n^a since nonoverlapping segments within a single Markov chain are not independent. However, this only affects the upper bound by a constant factor, since the event $H_{i,j}^{a,t} \equiv \{at \le \sum_{1 \le k \le t} 1(X_{i+k} = X_{j-k})\}$, which is the same as the event $G_{i,j}^{a,t}$ defined at (3) but with X in place of Y, satisfies $P(H_{i,j}^{a,t})/P(G_{i,j}^{a,t}) \le 1/\min\{\mu_l\}$, whenever $|i-j| \ge t$.

The proof of the lower bound is the argument from Section 4, with some modifications, as follows. We still use a set $J \subset (S^t)^2$ consisting of *some* of the quality a matching pairs (w, z) of words of length t. Instead of specifying the empirical distribution of the t letter pairs (w_i, z_i) in the definition (7), we must now specify the empirical distribution of the t doublet pairs $((w_i, w_{i+1}), (z_i, z_{i+1}))$, taking the indices modulo t. Instead of the definition (5) of $E_{w,z}$ which uses blocks of t consecutive letters to provide negative correlations, we must now use Doeblin's method, taking blocks of t consecutive excursions from some fixed letter back to itself, and requiring that the specified words w and z appear at the start of such a block. See Arratia and Waterman (1985a, b) for the use of Doeblin's method in a similar setup. □

It is easy to see why Theorems 3 and 4, describing self-overlapping repeats, do not generalize directly to Markov chains: The relation between two letters a fixed offset m apart, say X_i and X_{i+m}, depends strongly on the value of m. Even the value p in the quantification $\forall\, a \in (p,1]$, which should be the average quality of matching, depends on the offset m:

$$p(m) \equiv \sum_{l \in S} \mu_l [P^m]_{ll} = (\text{a.s.}) \lim_{t \to \infty} U(t,m)/t,$$

where $U(t, m)$, defined at (13), is the number of matches between two blocks of length t at offset m from each other. Notice that $p(m) \to p$ as $m \to \infty$. For example, the Markov chain with $d = 2$ and transition matrix $P_{11} = P_{22} = 0.1$ and $P_{12} = P_{21} = 0.9$ has equilibrium $\mu = (\frac{1}{2}, \frac{1}{2})$, $p = \frac{1}{2}$, $p(1) = 0.1$, $p(2) = 0.82$ and $p(3) = 0.244$.

The generalization to Markov chains of Theorem 3 for self-overlapping repeats should take the form $\forall\, a \in (p^*, 1]$, $1 = P(S_n^a/\log(n) \to 1/r(a;[P]))$, where $p^* \equiv \sup_{m \ge 1} p(m)$. A version of Theorem 4 would make sense only if $p = p^*$; otherwise it would be like comparing grapes and watermelons. To see this in more detail, observe that $S_n^a = \max_{m \ge 1} S_n^a(m)$, where

$$S_n^a(m) \equiv 0 \vee \max\Big\{t > m \colon \exists\, i, j \in [0, n-t],$$
$$i - j = m,\ ta \le \sum_{1 \le k \le t} 1(X_{i+k} = X_{j+k})\Big\}.$$

Now for each $m \geq 1$, there is a version of Theorem 3, of the form $\forall\ a \in (p(m), 1]$, $1 = P(S_n^a(m)/\log(n) \to 1/r(a; [P], m))$, where $r(a; [P], m) \equiv \lim -t^{-1} \log P(at \leq U(t, m))$. The proper generalization of Theorem 3 should combine all of the these versions.

There are many interesting questions that naturally arise:

1. Does $p = p^*$ imply that the Markov chain is actually an i.i.d. sequence?
2. Does there exist a finite value m^* with $p^* = p(m^*)$?
3. As a function of d, what upper bound can be given for the minimum solution m^* of $p^* = p(m^*)$?
4. If m^* exists, does it follow that $r(a; [P]) = \lim -t^{-1} \log P(at \leq U(t, m^*))$?

8. Requiring a given proportion of mismatches. Let $a \in [0,1]$ be given. The length $M_n^{\leq a}$ of the "longest nonmatching of quality a, allowing shifts" between two sequences $X_1X_2 \cdots X_n$ and $Y_1Y_2 \cdots Y_n$ is defined by

$$M_n^{\leq a} \equiv \max\left\{t \colon \exists\ i, j \in [0, n-t],\ a \geq t^{-1} \sum_{1 \leq k \leq t} 1(X_{i+k} = Y_{j+k})\right\}.$$

The only difference between this and the definition of M_n^a is in the direction of the inequality. For the sake of comparison, we also consider the length $R_n^{\leq a}$ of the "longest head-free run of quality a," in a sequence $Z_1, Z_2, \ldots, Z_n$ of $\{0,1\}$-valued random variables,

$$R_n^{\leq a} \equiv \max\left\{t \colon \exists\ i \in [0, n-t],\ a \geq t^{-1} \sum_{1 \leq k \leq t} Z_{i+k}\right\}.$$

By taking $Z_i \equiv 1(X_i = Y_i)$, we have that $R_n^{\leq a}$ is the length of the "longest nonmatching of quality a, not allowing shifts" between $X_1X_2 \cdots X_n$ and $Y_1Y_2 \cdots Y_n$.

As in Sections 2–5, assume that all letters $X_1, X_2, \ldots, Y_1, Y_2, \ldots$ are mutually independent elements of $S = \{1, 2, \ldots, d\}$, with distribution μ. Let $p \equiv P(X_1 = Y_1) = \sum_{l \in S}(\mu_l)^2$; we will usually take $a \in [0, p)$.

Can limit laws for $M_n^{\leq a}$ be derived from laws for M_n^a, perhaps by complementation? The answer is a surprising, but definite "no." First, observe that since $H(a, p) = H(1-a, 1-p)\ \forall\ a, p \in [0,1]$, the Erdös–Rényi law (1), applied to $\{1 - Z_i\}$, directly implies that if $Z_1, Z_2, \ldots$ are p-coins, then $R_n^{\leq a}/\log(n) \to 1/H(a, p)$ almost surely. However, even in the case $d = 2$ where the sequences being compared represent coin tossing, there is no way to derive strong laws for $M_n^{\leq a}$ from Theorem 1, the strong law for M_n^a.

The absence of duality between matching and nonmatching, allowing shifts, can be seen most clearly by considering the extreme cases, $a = 0$, with $M_n^{\leq 0}$ being the length of the longest perfect nonmatching, and $a = 1$, with M_n^1 being the length of the longest perfect matching. If for example, $X_1X_2 \cdots = Y_1Y_2 \cdots = 0101\ldots$, then $M_n^{\leq 0} = n - 1$ and $M_n^1 = n$.

We consider further the two extreme cases, $a = 0$ and $a = 1$. The analysis by position using first and second moments, described in the second paragraph of Section 4, works for M_n^1 for all μ. This is because the condition needed, namely

$P(X_1 = Y_1 = Y_2) < \{P(X_1 = Y_1)\}^{3/2}$, is always valid, thanks to Jensen's inequality—see Arratia and Waterman (1985a) for details. With $a = 0$, the analysis by position for $M_n^{\leq 0}$ works if and only if $P(X_1 \neq Y_1$ and $X_1 \neq Y_2) > \{P(X_1 \neq Y_1)\}^{3/2}$. This last condition is valid for some but not all μ. For example, if $d = 2$ and $\mu = (\theta, 1-\theta)$, then $P(X_1 \neq Y_1$ and $X_1 \neq Y_2) = \theta(1-\theta)$ and $P(X_1 \neq Y_1) = 2\theta(1-\theta)$, so the first and second moments method works if and only if $1 < 8\theta(1-\theta)$. However, the analysis by pattern works in all cases, just as easily for $a \in [0, p)$ as it did for $a \in (p, 1]$.

The length $D_n^{\leq a}$ (respectively, $S_n^{\leq a}$) of the "longest nonmatching of quality a," not allowing (respectively, allowing) self-overlap, within a single sequence $X_1 X_2 \cdots X_n$ is defined by

$$D_n^{\leq a} \equiv \max\left\{t \colon \exists\, i, j \in [0, n-t],\ |i-j| \geq t,\ a \geq t^{-1} \sum_{1 \leq k \leq t} 1(X_{i+k} = X_{j+k})\right\},$$

$$S_n^{\leq a} \equiv \max\Bigg\{t \colon \exists\, i, j \in [0, n-t],$$

$$0 < |i-j| < t,\ a \geq t^{-1} \sum_{1 \leq k \leq t} 1(X_{i+k} = X_{j+k})\Bigg\}.$$

Theorem 7. *Theorems* 1–4 *for i.i.d. letters generalize to nonmatching*: $\forall$ $a \in [0, p)$,

$$a = P\left(M_n^{\leq a}/\log(n) \to 2/H(a, p)\right),$$
$$1 = P\left(D_n^{\leq a}/\log(n) \to 2/H(a, p)\right),$$
$$1 = P\left(S_n^{\leq a}/\log(n) \to 1/r(a, \mu)\right)$$

and hence

$$1 = P\left(D_n^{\leq a}/S_n^{\leq a} \to 2r(a, \mu)/H(a, p) > 1\right).$$

Proof. There are only a few minor changes from the proofs given for Theorems 1–4. The biggest change is that the definition (7) of J is modified as

$$J \equiv \Bigg\{(w, z) \in (S^t)^2 \colon \sum_{1 \leq i \leq t} 1(b = w_i = z_i) = \lfloor at\alpha_b \rfloor, \text{for } b \in S,\ b \neq d, \text{and}$$

$$\sum_{1 \leq i \leq t} 1(b = w_i,\ c = z_i) = \lceil (1-a)t\gamma_{bc} \rceil, \text{for } b \neq c \in S\Bigg\}.$$

The only change was to replace one occurrence of floor $\lfloor \cdot \rfloor$, with ceiling $\lceil \cdot \rceil$, so that each pair of words $(w, z) \in J$ is now required to match in a number $s \equiv t - \sum_{b \neq c \in S} \lfloor (1-a)t\gamma_{bc} \rfloor$ of places, with $s \leq at$ instead of $s \geq at$. □

The multidimensional and Markov generalizations also can be easily extended to corresponding theorems about nonmatching.

Acknowledgments. We are grateful to Barry Simon, Francis Comets and Yun-Shyong Chow for their help in the proof of Theorem 4, and to the IMA in Minneapolis for bringing us all together.

REFERENCES

ARRATIA, R., GOLDSTEIN, L. and GORDON, L. (1989). Two moments suffice for Poisson approximations: The Chen–Stein method. *Ann. Probab.* **17** 9–25.

ARRATIA, R., GORDON, L. and WATERMAN, M. S. (1988). The Erdös–Rényi law in distribution, for coin tossing and sequence matching. Preprint.

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.

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

ERDÖS, P. and RÉNYI, A. (1970). On a new law of large numbers. *J. Analyse Math.* **22** 103–111. Reprinted (1976) in *Selected Papers of Alfred Rényi* **3** 1962–1970. Akadémiai Kiadó, Budapest.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

KARLIN, S. and OST, F. (1985). Some monotonicity properties of Schur powers of matrices and related inequalities. *Linear Algebra Appl.* **68** 47–65.

KRUSKAL, J. B. and SANKOFF, D. (1983). *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison.* Addison-Wesley, Reading, Mass.

SIBLEY, C. G. and ALQUIST, J. E. (1984). The phylogeny of the humanoid primates, as indicated by DNA-DNA hybridization. *J. Mol. Evol.* **20** 2–15.

WATERMAN, M. S., GALAS, D. and ARRATIA, R. (1984). Pattern recognition in several sequences: Consensus and alignment. *Bull. Math. Biol.* **46** 515–527.

WATERMAN, M. S., GORDON, L. and ARRATIA, R. (1987). Phase transitions in sequence matches and nucleic acid structure. *Proc. Nat. Acad. Sci. U.S.A.* **84** 1239–1243.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
DRB 306, UNIVERSITY PARK
LOS ANGELES, CALIFORNIA 90089-1113