The Annals of Statistics
1990, Vol. 18, No. 2, 539–570

THE ERDÖS–RÉNYI LAW IN DISTRIBUTION, FOR COIN TOSSING AND SEQUENCE MATCHING[1]

BY R. ARRATIA, L. GORDON AND M. S. WATERMAN

University of Southern California

We study approximations to the distribution of counts of matches in the best matching segment of specified length when comparing two long sequences of i.i.d. letters. The key tools used are large-deviation inequalities and the Chen–Stein method of Poisson approximation. The origin of the problem in molecular biology is indicated.

1. Introduction. A strand of DNA can be represented as a long string of letters from the four-letter alphabet $\{a, c, g, t\}$. Currently, a large amount of laboratory effort is being expended in the determination and subsequent compilation of genetic information from various organisms. This information consists of listings of these long strings. A natural question arises from comparison of two or more such strings, by biologists' efforts to determine when a comparison detects an unusual congruence shared among the compared strings. Such statistical problems are naturally cast in the usual hypothesis-testing context, in which we need to compute the tail probability (the biologists' p-value) for a seemingly unusual event.

The work we report here is motivated by the scientific desire to compute the sort of tail probabilities of interest to molecular biologists in their evaluation of closely matching regions of different biological sequences. Until recently, the standard tool used in computing tail probabilities was a probabilistic use of the Bonferroni inequalities as pioneered in Watson (1954). Such calculations essentially establish a Poisson approximation for the distribution of counts of weakly dependent rare events. See, for example, the moment calculations in Karlin and Ost (1987) and the discussion in Karlin, Ghandour Ost, Tavare and Korn (1983). Use of the Bonferroni inequalities requires computation of moments of arbitrarily large order; the task is always tedious and frequently technically demanding.

A promising alternative to using Bonferroni methods to establish the Poisson approximation for dependent events is to use methods developed in Chen (1975) and Stein (1986). In Arratia, Goldstein and Gordon (1989), the Chen–Stein method of Poisson approximation is generalized to a multivariate context, and various examples relevant to sequence matching are presented. Indeed, the realization that the results of Arratia, Gordon and Waterman (1986) can be obtained without the high-order moment calculations required by Bonferroni methods has enabled us to cope successfully with problems

Received November 1988; revised November 1989.

[1] Supported by NIH grant 36230, NSF grant DMS-88-15106 and a grant from the System Development Foundation.

AMS 1980 *subject classifications*. Primary 62F05; secondary 92A10.

Key words and phrases. Moving average, scan statistics, sequence matching, large deviations.

whose difficulty we previously thought insurmountable. The mathematical motivation for this paper is to present the tools needed to apply the Chen–Stein method to sequence comparison problems.

In this paper, we consider the simplest problem of possible statistical interest, matching segments from two independent sequences of independent identically distributed letters. Surprisingly, in Section 7, we shall see that even such a naive formulation might be useful in a biological context. Our main results, Theorems 3 and 4 of Section 6.3, give the asymptotic distribution of unusually rich matches between independent sequences. These approximations are described in terms of corresponding distributional results for unusually head-rich regions found in a single random sequence of i.i.d. coin tosses. The latter distributions are given in Theorem 1 of Section 4 and in Theorem 2 of Section 5. Our notation and a discussion of the applicability of the Poisson approximation in the context of unusually head-rich regions for coin tossing appear in the rest of this introduction, and continue in the following four sections.

The method for analyzing the probability of occurrences of long runs of heads or matches involves a Poisson approximation. There are two distinct issues: The expected number of events (λ or λ') must be approximated, and the dependence among the events being counted must be controlled. The first issue, handled in part by Lemma 1, is the same for coin tossing and for matching independent, i.i.d. sequences. The second issue is handled easily in the case of coin tossing (Theorems 1 and 2) and is very complicated in the case of sequence matching (Theorems 3 and 4).

Let $0 < p < a \leq 1$ and let $\dots, Z_{-1}, Z_0, Z_1, Z_2, \dots$ be an i.i.d. sequence with $p = P(Z_i = 1) = 1 - P(Z_i = 0)$. Let $R_n \equiv R_n^a$ be the length of the longest consecutive run, contained within the first n tosses, in which the fraction of 1's is at least a. We refer to R_n as the length of the longest quality a head run. Erdös and Rényi (1970) proved that

$$(1) \qquad \frac{R_n}{\log(n)} \to \frac{1}{H(a,p)} \quad \text{almost surely,}$$

where

$$(2) \qquad H(a,p) = a\log\left(\frac{a}{p}\right) + (1-a)\log\left(\frac{1-a}{1-p}\right)$$

is the relative entropy of a and p, with $H(1,p) = \log(1/p)$. Deheuvels, Devroye and Lynch (1986) prove a refinement of this which, in the case of coin tossing, essentially says that for $p < a < 1$,

$$(3) \qquad \frac{1}{\log\log(n)}\left(R_n - \frac{\log(n)}{H(a,p)}\right) \to \frac{-1}{2H(a,p)} \quad \text{in probability.}$$

In this paper we attempt to analyze the distribution of R_n. Define centering constants $c_n \equiv c_n^a$,

$$(4) \qquad c_n \equiv \begin{cases} (\log(n) - \frac{1}{2}\log\log(n))/H(a,p) & \text{if } p < a < 1, \\ \log(n)/\log(1/p) & \text{if } a = 1. \end{cases}$$

Our result implies that the family $\{R_n - c_n\}$ of random variables is tight and yields explicit bounds on the tails of the distribution of $R_n - c_n$. The distribution of the longest pure head run, which is the case $a = 1$, was analyzed by Goncharov (1942).

For fixed positive integers t and n, let $S_{n;t}$ be the maximum number of heads occurring among t consecutive tosses starting within the first n tosses. Theorem 1 gives a simple approximation to the distribution of $S_{n;t}$: for integers s with $p < s/t \leq 1$,

$$P(S_{n;t} < s) \text{ is close to } \exp\left(-n\left(\frac{s}{t} - p\right)P(Z_1 + \cdots + Z_t = s)\right),$$

together with explicit bounds on the error in this approximation. This approximation is useful when $P(S_{n;t} < s)$ is not extremely close to 0 or 1, which for large n requires that t have order of magnitude $\log n$ and that s be not far from at, where $a \in (p, 1]$ satisfies $t = \log(n)/H(a,p)$. Similar distributional approximations are given for more general summands, but without the explicit location constant [corresponding to the factor $((s/t) - p)$] and without rates of convergence, in Deheuvels and Devroye [(1987), Theorem 6]. See also Kómlos and Tusnády (1975) and Naus (1979, 1982) for related results. Having rates of convergence lets us prove a sharp version of the law of the iterated logarithm for $S_{n;t}$. See Corollary 2 of Section 4.

The two families, $\{R_n^a : a \in (p, 1]\}$ and $\{S_{n;t} : 1 \leq t \leq n\}$ are closely related. At the level of statements with a normalizing factor that tends to infinity, such as (1) and (3), results as $n \to \infty$ for R_n^a with $a \in (p, 1]$ are equivalent to results for $S_{n;t}$ as $n, t \to \infty$ with $t/\log(n) \to c$ and $1/c \in (0, \log(1/p)]$. At the level of distributional results, however, there is a significant difference between the two families. In particular, knowledge of the values of $P(S_{n;t} < s)$ for all n, t and s is not enough to yield a good approximation to the distribution of R_n^a, although it is enough to imply our tightness result for $\{R_n^a - c_n^a\}$. The reader is urged to pause and react: Is there a paradox here?

There is a paradox, if one is guided by the study of pure head runs, corresponding to the extreme case $a = 1$. In sharp contrast to the situation for pure head runs, with $a < 1$ and t a positive integer, it is possible for there to be a quality a run of length *t or greater*, but none of length exactly t. For example, with $a = \frac{1}{2}$ and $t = 6$, there is a quality a run of length $t + 2$, but none of length t, in the sequence $\cdots$ 00000**11000011**00000 . . . . By inserting a region of length l with $\lceil al \rceil$ ones into the middle of the above example, we get an example with a run of length $t + l + 2$, but none of length $t + l$. For example, with the inserted region *010101* of length $l = 6$, the new example . . . 00000**1100***010101***0011**00000 . . . has a quality $\frac{1}{2}$ run of length $12 + 2$ but none of length 12.

Such examples have a substantial chance of happening at random: if n, t and a are such that $P(R_n < t)$ is not close to 0 or 1, then there is a substantial chance that the longest quality a run has length slightly less than t. By placing a few extra ones outside each end of this run, we can create a run of length greater than t without creating a run of length exactly t.

The method for analyzing the distribution of $S_{n;t}$ is the following. The event $\{S_{n;t} < at\}$ is the event that, starting in the first n tosses, there is no run of quality a and length *exactly* t. Consider the places at which there is a quality a head run of length t. Such places occur in clumps, and we define a random variable W which represents the number of clumps that occur in the first n tosses. Using the Chen–Stein method, as presented in Arratia, Goldstein and Gordon (1989), we show that the distribution of W is close to the Poisson distribution with parameter $\lambda \equiv EW$ and, in particular, $P(S_{n;t} < at)$ is close to $P(W = 0)$, which is close to $\exp(-\lambda)$. We then give a simple approximation to λ, namely, $\lambda^* \equiv n(a-p)P(Z_1 + \cdots + Z_t = \lceil at \rceil)$. The net result is that $P(S_{n;t} < at)$ is close to $\exp(-\lambda^*)$, with an explicit bound on the error.

There is a nice heuristic, in the language of Aldous (1989), to explain the value of λ^*, as follows. Every clump of windows of length t and quality a starts with a window having exactly $\lceil at \rceil$ heads, and $nP(Z_1 + \cdots + Z_t = \lceil at \rceil)$ is the expected number of such windows. An argument using the ballot theorem shows that the average number of such windows per clump is approximately $1/(a-p)$, so that the expected number λ of clumps is approximately $\lambda^* = n(a-p)P(Z_1 + \cdots + Z_t = \lceil at \rceil)$.

The method for analyzing the distribution of R_n^a is similar to that for $S_{n;t}$. Fix a test length t and consider the places in the first n tosses at which there is a quality a head run of length *t or greater*. Such places occur in clumps, so we define W' to be the number of clumps that occur. Using the Chen–Stein method, we show that the distribution of W' is close to the Poisson distribution with parameter $\lambda' \equiv EW'$, so that, in particular, $P(R_n < t)$ is close to $\exp(-\lambda')$. Unfortunately, we can't find a simple asymptotic formula for λ'—it is easy to approximate the expected number of windows of quality a and length t or greater, but we don't know how to approximate the average number of such windows per clump, corresponding to the factor $1/(a-p)$ in the case of $S_{n;t}$ which involves windows of length exactly t. Given the Poisson approximation that $P(R_n < t)$ is close to $\exp(-\lambda')$, the tightness of the family $\{R_n - c_n\}$ is equivalent to having lower and upper bounds on λ' which imply that as $n, t \to \infty$, $\lambda' \to \infty$ if $t - c_n \to -\infty$ and $\lambda' \to 0$ if $t - c_n \to \infty$.

The results for sequence matching can best be expressed in terms of the result for coin tossing. Let $A_1, A_2, \ldots, A_m$ and $B_1, B_2, \ldots, B_n$ be independent integer-valued random "letters," with distribution μ for the A's and ν for the B's. Let $p \equiv \sum_{l \in Z} \mu_l \nu_l$, so that $\forall\ i, j$, $p = P(A_i = B_j)$. A single sequence of p-coin tosses, whose length is the product mn, *may or may not* mirror the matchings that occur between the $m + n$ letters. Whether this is the case depends on μ, ν and a, as well as on the relative magnitudes of m and n.

Let $a \in (p, 1]$. Let $M^a_{m,n}$ be the length of the longest "quality a" matching consecutive segment common to the two sequences $A_1 \dots A_m$ and $B_1 \dots B_n$. Let $N_{m,n;t}$ be the maximum number of matches occurring between two segments of length t, one segment from each sequence. We show that for large m and n, the distribution of $M^a_{m,n}$ is close to that of R^a_{mn}, and $P(N_{m,n;t} < at)$ is close to $P(S_{mn,t} < at)$, provided that a, μ, ν and the ratio $\log(m)/\log(n)$ satisfy a certain sufficient condition, which is summarized by (33); we do not believe that this condition is necessary. The condition is esentially the requirement of no blowup of the second moment of the number of places in the two sequences where the highly matching segments occur. For the important special case $\mu = \nu$ and $m = n$, this condition is *not* satisfied for all cases of a and μ, although for each μ it is satisfied in a neighborhood of $a = 1$.

The almost sure limit of $M^a_{m,n}/[\log(mn)/H(a,p)]$, as $m, n \to \infty$ with $\log n/\log(mn) \to \rho \in (0,1)$, always exists; if the distribution of $M^a_{m,n}$ is close to that of R^a_{mn}, then this almost sure limit must be 1. In Arratia and Waterman (1985), it is shown how to identify the almost sure limit for the case $a = 1$; the limit is a continuous but not analytic function of (μ, ν, ρ). There are cases, both with $a = 1$, $\mu = \nu$, $m \neq n$ and with $a = 1$, $m = n$, $\mu \neq \nu$, in which the almost sure limit is not 1, so the distributional approximation of $M^a_{m,n}$ by R^a_{mn} must fail. That 1985 paper uses "analysis by pattern," in contrast to the "analysis by position" in this paper. For the relatively crude task of deriving strong laws, analysis by pattern is quite robust, handling general μ, ν, m, n and extending easily to the case of Markov chains, but for the more delicate task of deriving a distributional approximation such as Theorem 2, analysis by pattern would be much more complex than analysis by position. A paper by Arratia and Waterman (1989) extends the analysis by pattern to the case $p < a < 1$, although for simplicity only the case $\mu = \nu$, $m = n$ is handled there, and in these cases, the almost sure limit always is 1. Thus there are cases, with $a < 1$, $\mu = \nu$, and $m = n$, in which the method of this paper fails to establish the distributional approximation, but the failure is not detected by the almost sure limit, which is still 1.

2. The expected number of clumps of head runs. The indicator $Y_\alpha \equiv Y(\alpha, s, t)$ of the event that a run of length t containing s heads begins at position α in the sequence of coin tosses is defined by the formula

$$Y_\alpha \equiv 1(s = Z_\alpha + Z_{\alpha+1} + \cdots + Z_{\alpha+t-1}). \tag{5}$$

The indicators Y_α and Y_β for $|\alpha - \beta| < t$ are highly correlated, so that 1's in the sequence Y tend to occur in clumps. To get a Poisson approximation, we must count the clumps, so we define the indicator that a clump begins at α to be

$$X_\alpha \equiv Y_\alpha(1 - Y_{\alpha-1})(1 - Y_{\alpha-2}) \cdots (1 - Y_{\alpha-t}). \tag{6}$$

In spirit, X_α is like the indicator of a renewal at α, but there are two differences. First, the indicator X_α is measurable with respect to a block of $2t - 1$ coins, which makes it easier to work with, compared to the indicator of a renewal, which depends on the entire past. Second, there is a slight difference between EX_α and the probability of a renewal at α, arising from situations like the following. In the case $a = 1$, a run of heads of length several t following a long tail run is counted as several renewals spaced exactly t apart; but only the first of these is counted by an X_α.

Our goal in this section is to show that, with $s = \lceil at \rceil$, we have, as $t \to \infty$, $EX_1/P(Z_1 + \cdots + Z_t = s) \to a - p$; this result together with error bounds is Lemma 1. There are two intuitively appealing ways to describe this result in the language of Aldous (1989). First, in terms of occurrences of windows of length t with *exactly* $\lceil at \rceil$ heads, the expected clump size tends to $\lim EY_1/EX_1 = 1/(a-p)$. Second, in terms of occurrences of windows of length t and quality a, i.e., with $\lceil at \rceil$ *or more* heads, the expected clump size tends to $(a - ap)/(a - p)^2$, since elementary computation gives $P(Z_1 + \cdots + Z_t = s)/P(Z_1 + \cdots + Z_t \geq s) \to (a-p)/(a-ap)$. In the special case $a = 1$, we observe that the two notions of clump coincide, and also Lemma 1 applies with $s = t$ so that the lower and upper bounds are equal.

Lemma 1. *Let s and t be positive integers with $s \leq t$, and let $a \equiv s/t$. Let X_1 and Y_1 be the indicators defined by* (5) *and* (6), *let the Z_i be p-coins and let $H(a,p)$ be the relative entropy of p-coins and a-coins, as defined by* (2). *Then*

$$a - p \leq \frac{EX_1}{EY_1} \leq a - p + 2(1-a)P\left(\sum_{j=1}^{t} Z_j > s\right) \tag{7}$$

$$\leq a - p + 2(1-a)e^{-tH(a,p)}.$$

Proof. For motivation, we think of a window of length t, placed initially over $[1,t]$ and then slid backward one step at a time, with D_i being the decrease at step i in the number of heads having indices inside the window. Let $D_i \equiv -Z_{1-i} + Z_{t+1-i}$ and $S_k \equiv D_1 + \cdots + D_k$. On the event $Y_1 = 1$, $X_1 = 1$ $(S_k \neq 0$ for $k = 1,\ldots,t)$.

Let U be the number of positive terms among $D_1, \ldots, D_t$, and let V be the number of negative terms. By the ballot theorem [Feller (1968), volume 1, chapter 3],

$$P(S_k \neq 0 \text{ for } k = 1,\ldots,t|U,V,Y_1) = \frac{|U-V|}{U+V}\,\frac{U+V}{t} = \frac{|U-V|}{t}.$$

The last factor, $(U+V)/t$, corresponds to the requirement $S_1 \neq 0$. The ballot theorem is applicable because the family $\{(Z_{1-i}, Z_{t+1-i})\colon i = 1,\ldots,t\}$ conditional on Y_1 is exchangeable. Notice that $D_i \in \{-1, 0, 1\}$ and that if $D_i = 1$, then $Z_{t+1-i} = 1$, while if $D_i = -1$, then $Z_{t+1-i} = 0$. Thus, conditional on the event $\{Y_1 = 1\}$, the distributions of U and V are binomial$(s, 1-p)$ and binomial$(t-s, p)$, respectively and independently. In particular,

$$E(U - V|Y_1 = 1) = s(1-p) - (t-s)p = at - pt.$$

Hence,

$$P(X_1 = 1|Y_1 = 1) = \frac{1}{t}E(|U - V||Y_1 = 1) \geq a - p.$$

Our notation for the positive and negative parts is $x = x^+ - x^-$. Note that $|U - V| = U - V + 2(U - V)^-$. For an upper bound on $E((U - V)^-|Y_1 = 1))$, we observe that the maximum possible value of $V - U$ on the event $Y_1 = 1$ is $(t - s)$, so that $E((U - V)^-|Y_1 = 1)) \leq (t - s)P(V - U > 0|Y_1 = 1)$. Let Q be binomial(t, p) in distribution. Now, conditional on $Y_1 = 1$, the distribution of $V + s - U$ is that of Q, so $P(V - U > 0|Y_1 = 1) = P(Q > s) \leq \exp(-tH(a, p))$ is the usual exponential upper bound from large deviation theory. □

3. The Chen–Stein method for Poisson approximations.

To keep this paper self-contained, we present briefly the Chen–Stein method for establishing Poisson approximations, as given in Arratia, Goldstein and Gordon (1989). See also Stein (1986) or Barbour (1982) and Barbour and Holst (1989) for more on this method.

Let I be an arbitrary index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with $P(X_\alpha = 1) = 1 - P(X_\alpha = 0) > 0$. Let

$$W \equiv \sum_{\alpha \in I} X_\alpha, \quad \text{and} \quad \lambda \equiv EW.$$

We assume that $\lambda \in (0, \infty)$. Denote by Z a Poisson(λ) random variable.

For each $\alpha \in I$, suppose we have chosen $B_\alpha \subset I$, with $\alpha \in B_\alpha$. We think of B_α as a "neighborhood of dependence" for α, such that X_α is independent or nearly independent of all of the X_β for β not in B_α. Define

$$b_1 \equiv \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} EX_\alpha EX_\beta,$$

$$b_2 \equiv \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} E(X_\alpha X_\beta),$$

$$b_3 \equiv \sum_{\alpha \in I} s_\alpha, \quad \text{where } s_\alpha \equiv E\Big|E\{X_\alpha - (EX_\alpha)\big|\sigma(X_\beta\colon \beta \in I - B_\alpha)\}\Big|.$$

Loosely speaking, when b_1, b_2 and b_3 are all small, then the total number W of events is approximately Poisson, the locations of the dependent events approximately form a Poisson process and the dependent events are almost indistinguishable from a collection of independent events having the same marginal probabilities. In this setup, b_1 measures the neighborhood size, b_2 measures the expected number of neighbors of a given occurrence and b_3 measures the dependence between an event and the occurrences outside its neighborhood.

In many applications, such as those in this paper, the natural choice of B_α makes X_α independent of $\sigma(X_\beta\colon \beta \in I - B_\alpha)$, so that $b_3 = 0$, and this can be verified without performing any calculations. In these situations, if the neighborhoods are small, as measured by b_1, then checking that b_2 is small is essentially equivalent to checking that the second moment of W is well behaved, since $b_2 - b_1 = E(W^2) - \lambda - \lambda^2 = E(W^2) - E(Z^2)$. Two recent studies of sequence matching, Arratia, Gordon and Waterman (1986), and Karlin and Ost (1987), established Poisson approximations by the method of inclusion–exclusion (which essentially requires that all moments of W are well behaved) and thus required stronger restrictions on the distributions of the sequences being matched.

We denote the total variation distance between the distributions of W and Z by

$$\begin{aligned}\|\mathscr{L}(W) - \mathscr{L}(Z)\| &\equiv \sup_{\|h\|=1} |Eh(W) - Eh(Z)| \\ &= 2\sup_A |P(W \in A) - P(Z \in A)|.\end{aligned}$$

LEMMA 2. *Let W be the number of occurrences of dependent events, and let Z be a Poisson random variable with $EZ = EW = \lambda$. Then*

$$\begin{aligned}\|\mathscr{L}(W) - \mathscr{L}(Z)\| &\le 2\left[(b_1 + b_2)\frac{1 - e^{-\lambda}}{\lambda} + b_3(1 \wedge 1.4\lambda^{-1/2})\right] \\ &\le 2(b_1 + b_2 + b_3),\end{aligned}$$

and

$$|P(W = 0) - e^{-\lambda}| \le \frac{(b_1 + b_2 + b_3)(1 - e^{-\lambda})}{\lambda} < (1 \wedge \lambda^{-1})(b_1 + b_2 + b_3).$$

For $\alpha \in I$, let Y_α be a random variable whose distribution is Poisson with mean EX_α, with the Y_α mutually independent. The total variation distance between the dependent Bernoulli process $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$ and the Poisson process $\mathbf{Y}$ on I with the same intensity, $\mathbf{Y} \equiv (Y_\alpha)_{\alpha \in I}$, satisfies

$$\|\mathscr{L}(\mathbf{X}) - \mathscr{L}(\mathbf{Y})\| \le 2(2b_1 + 2b_2 + b_3).$$

For $\alpha \in I$, let X'_α have the same distribution as X_α, with the X'_α mutually independent. The total variation distance between the dependent Bernoulli process $\mathbf{X} \equiv (X_\alpha)_{\alpha \in I}$ and the independent Bernoulli process $\mathbf{X}' \equiv (X'_\alpha)_{\alpha \in I}$ having the same marginals satisfies

$$\|\mathscr{L}(\mathbf{X}) - \mathscr{L}(\mathbf{X}')\| \le 2(2b_1 + 2b_2 + b_3) + 2\sum (EX_\alpha)^2.$$

Observe that the total variation distance $\|\mathscr{L}(\mathbf{X}) - \mathscr{L}(\mathbf{X}')\|$ can be interpreted as twice the minimum value of $P(\mathbf{X} \ne \mathbf{X}')$ over all realizations of both processes on the same probability space.

4. The Erdös–Rényi law for coin tossing: The best region of a given length t. Let Z_i be p-coins, i.e., an independent sequence with $p = P(Z_i = 1) = 1 - P(Z_i = 0)$, and let $S_{n;t}$ be the maximum number of heads occurring in a window of length t, starting within the first n tosses: $S_{n;t} \equiv \max_{1 \le i \le n}(Z_i + \cdots + Z_{i+t-1})$. In this section we approximate the distribution of $S_{n;t}$.

For integer α and positive integers $s \leq t$, define indicators

$$Y_\alpha \equiv Y(\alpha, s, t) \equiv 1\left(s = \sum_{k=0}^{t-1} Z_{\alpha+k}\right),$$

and

$$X_\alpha \equiv X(\alpha, s, t) \equiv Y_\alpha \prod_{j=1}^{t} (1 - Y_{\alpha-j}).$$

Let $I \equiv \{1, 2, \dots, n\}$, and define

$$W \equiv W(n, s, t) \equiv \sum_{\alpha \in I} X_\alpha. \tag{8}$$

Informally, Y_α is the indicator that a window of length t containing s heads starts at α. If two such windows overlap, we say that they belong to the same clump, and X_α is the indicator that a clump starts at α. The random variable W is the number of clumps that begin within the first n tosses. Apart from "boundary effects," the event $\{S_{n;t} < s\}$ agrees with the event $\{W = 0\}$. The error in this approximation can be controlled by observing that $\{W \neq 0\} \subset \{S_{n;t} \geq s\}$, and

$$\{S_{n;t} \geq s, W = 0\} \subset \{Y_1 + \cdots + Y_t > 0\} \cup \{Z_1 + \cdots + Z_t > s\}.$$

Hence,

$$0 \leq P(W = 0) - P(S_{n;t} < s) \leq tEY_1 + P(Z_1 + \cdots + Z_t > s). \tag{9}$$

We now establish a Poisson approximation for W using the Chen–Stein method. Let $\lambda \equiv EW$. The indicator random variable X_α is measurable with respect to the $2t$ coins Z_j at $\alpha - t, \dots, \alpha + t - 1$. Thus, we let

$$B_\alpha \equiv \{\beta \in I : |\alpha - \beta| < 2t\} \quad \text{for } \alpha = 1 \text{ to } n,$$

so that $b_3 = 0$ and $b_1 < (4t - 1)\lambda EX_1$. [It would be $b_1 = (4t - 1)\lambda EX_1$ if the index set were a circle.] If $|\alpha - \beta| \leq t$, then $E(X_\alpha X_\beta) = 0$, but if $t < |\alpha - \beta| < 2t$, then we can only conclude that $E(X_\alpha X_\beta) \leq (EX_\alpha)EY_\beta$, so that $b_2 < 2t\lambda EY_1$.

Using the Chen–Stein method as presented in Lemma 2, we have

$$\begin{aligned} |P(W = 0) - e^{-EW}| &\leq (b_1 + b_2)(1 \wedge 1/\lambda) \\ &\leq 2t\lambda(2EX_1 + EY_1)(1 \wedge 1/\lambda) \\ &\leq 6tEY_1. \end{aligned} \tag{10}$$

Combining the upper bound (9) on boundary effects with the result (10) of the Chen–Stein method, and spelling out EY_1 in terms of the underlying p-coins Z_i, we have

$$\begin{aligned} |P(S_{n;t} < s) - e^{-EW}| \leq 7tP(Z_1 + \cdots + Z_t = s) \\ + P(Z_1 + \cdots + Z_t > s). \end{aligned} \tag{11}$$

The other main ingredient in approximating the distribution of $S_{n;t}$ is provided by the elementary relation $\lambda \equiv \lambda(n, s, t) \equiv EW = nEX_1$ combined with the clump size bounds from Lemma 1,

$$(12)\quad \frac{s}{t} - p \leq \frac{EW}{nP\left(\sum_{j=1}^{t} Z_j = s\right)} \leq \frac{s}{t} - p + 2\left(1 - \frac{s}{t}\right)P\left(\sum_{j=1}^{t} Z_j > s\right).$$

To summarize these findings, we present the following theorem.

THEOREM 1. *Let Z_i be p-coins, and let $S_{n;t} \equiv \max_{1 \leq i \leq n}(Z_i + \cdots + Z_{i+t-1})$. For all positive integers n, s, t, with $s \leq t$ and $s/t > p$,*

$$(13)\quad P(S_{n;t} < s) \textit{ is approximately } \exp\left(-n\left(\frac{s}{t} - p\right)P(Z_1 + \cdots + Z_t = s)\right),$$

with the error in this approximation controlled by (11) *and* (12).

From the explicit error bounds in Theorem 1 it is easy to extract various less explicit statements about the rate of convergence, such as the following two: For fixed $\varepsilon > 0$, as $n \to \infty$,

$$\sup_{s,t:\, \varepsilon < EW < 1/\varepsilon,\, s/t \geq p+\varepsilon} \left| P\{S_{n;t} < s\} - \exp\left[-n\left(\frac{s}{t} - p\right)P(Z_1 + \cdots + Z_t = s)\right]\right|$$
$$= O\left(\frac{\log(n)}{n}\right).$$
$$\sup_{s,t:\, s/t \geq p+\varepsilon} \left| P\{S_{n;t} < s\} - \exp\left[-n\left(\frac{s}{t} - p\right)P(Z_1 + \cdots + Z_t = s)\right]\right|$$
$$= O\left(\frac{(\log(n))^2}{n}\right).$$

The sharpness of the error bound $O(n^{-1} \log n)$ in the first of the above statements can be shown by using inclusion–exclusion to analyze $P(W = 0)$ in the case $s = t = \lfloor \log_{1/p}(n) \rfloor$.

Along the same lines as the remark preceding Lemma 1, Theorem 1 could be stated as follows: with $a \equiv s/t$,

$$P(S_{n;t} < s) \text{ is approximately } \exp\left(-n\frac{(a-p)^2}{a - ap}P\left(\sum_{j=1}^{t} Z_j \geq s\right)\right),$$

which can be interpreted as saying that $S_{n;t}$ is distributed like the maximum of $n(a-p)^2/(a - ap)$ independent copies of $Z_1 + \cdots + Z_t$, where $a \in (p, 1]$ satisfies $n = \exp(tH(a, p))$.

Theorem 1 may be the cleanest way to summarize our understanding of the distribution of $S_{n;t}$. In order to get tightness and convergence in distribution statements out of Theorem 1, it only remains to use asymptotics for

$P(Z_1 + \cdots + Z_t = s)$. Furthermore, by using the error bounds of Theorem 1, it is easy to derive sharp versions of strong laws as corollaries. We carry this out below for the case $t, n \to \infty$ with $t/\log(n) \to c > 1/\log(1/p)$.

If both s and $t - s$ are large, then a good approximation to $P(Z_1 + \cdots + Z_t = s)$ is provided by Stirling's formula [Feller (1968)]: with $a \equiv s/t$,

$$(14)\quad \exp(c_1) < P(Z_1 + \cdots + Z_t = s)e^{tH(a,p)}\sqrt{2\pi a(1-a)t} < \exp(c_2),$$

where

$$c_1 = \frac{1}{12t+1} - \frac{1}{12s} - \frac{1}{12(t-s)}, \quad c_2 = \frac{1}{12t} - \frac{1}{12s+1} - \frac{1}{12(t-s)+1}.$$

Observe that the derivative with respect to a of the relative entropy $H(a,p)$ may be expressed, for $0 < p$, $a < 1$, in terms of the odds ratio,

$$r \equiv r(a,p) = p(1-a)/(a(1-p)).$$

Note that $\partial H(a,p)/\partial a = -\log r(a,p)$. When s/t is close to $a \in (p,1)$, increasing s by 1 decreases λ by a factor close to $r(a,p)$.

Define centering constants $b(n,t) \equiv b(n,t;a,p)$ by

$$(15)\quad \begin{aligned} b(n,t) \equiv{} & \frac{a\log(n)}{H(a,p)} - \frac{1}{2}\log_{1/r}\log(n) - \frac{1}{2}\log_{1/r}\left(\frac{2\pi a(1-a)}{H(a,p)}\right) \\ & + \log_{1/r}(a-p). \end{aligned}$$

If $a \in (p,1)$ satisfies $t = \log(n)/H(a,p)$, then $b(n,t)$ gives the location of the "center" of the distribution of $S_{n;t}$, in a sense made precise by Corollary 1. Crudely stated, if $s = b(n,t)$ is an integer, then among n tosses of the p-coins Z_i, the expected number of clumps of windows of length t having s heads, which is approximately $n(a-p)P(Z_1 + \cdots + Z_t = s)$, is close to 1, thanks to Stirling's approximation. To see this, observe first that if we only used the first term of $b(n,t)$, taking $s = a\log(n)/H(a,p)$ would yield $s/t = a$, and the principal factor of Stirling's approximation to $P(Z_1 + \cdots + Z_t = s)$ would be $\exp(-tH(s/t,p)) = n^{-1}$. The remaining terms of $b(n,t)$ involve $\log_{1/r}$ and should be viewed as the change in s needed to compensate for the factor of $(a-p)$ and the remaining factors in Stirling's approximation, in order that $n(a-p)P(Z_1 + \cdots + Z_t = s)$ be close to 1.

COROLLARY 1. *For t and n with $t > \log_{1/p}(n)$ define $a \equiv a(n,t)$ by the requirements that*

$$n = e^{tH(a,p)}, \qquad p < a < 1.$$

Then for each $\varepsilon > 0$, uniformly as $t, n \to \infty$ with $1 + \varepsilon \le t/\log_{1/p}(n) \le 1/\varepsilon$,

$$\sup_x \left|P(S_{n;t} - b(n,t) < x) - \exp(-r^x)\right| \to 0,$$

where the supremum is taken over $x \in R$ such that $x + b(n,t) \in Z$.

PROOF. This corollary follows from elementary manipulation of Theorem 1 together with the asymptotics (14) for large deviations in the binomial distribution. □

Let V have the extreme value distribution, $P(V < x) = \exp(-e^{-x})$. Then $P(V/\log(1/r(a,p)) < x) = \exp(-r^{x})$ for all real x, so the above says that $S_{n;t} - b(n,t)$ converges in distribution to $-f + \lfloor f + V/\log(1/r(a,p))\rfloor$ if n, $t \to \infty$ with $t \sim \log n/H(a,p)$ and $b(n,t) \pmod 1) \to f \in [0,1)$ for fixed $a \in (p,1)$.

The strong laws for the case $a = 1$, pure head runs, with form similar to Corollary 2, appear in Erdös and Révész (1975) and Révész (1978).

COROLLARY 2. *Fix $a \in (p,1)$. With $t \equiv t(n) \equiv \lfloor \log(n)/H(a,p)\rfloor$ and $r \equiv r(a,p)$, as $n \to \infty$,*

$$1 = P\left(\liminf\left(S_{n;t} - b(n,t) + \log_{1/r}\log\log n\right) = -1\right),$$

$$1 = P\left(\limsup \frac{S_{n;t} - b(n,t)}{\log_{1/r}\log n} = 1\right),$$

and

$$1 = P\left(\limsup \frac{S_{n;t} - b(n,t) - \log_{1/r}\log n}{\log_{1/r}\log\log n} = 1\right).$$

More generally, for any monotone function f with $f(n) \sim \log_{1/r}\log n$,

$$P\left(\left(S_{n;t} \geq b(n,t) + f(n) \textit{ infinitely often}\right)\right) = 0 \ (\textit{respectively}, 1),$$

according as

$$\sum_{k} r(a,p)^{f(\exp(k))} < \infty \ (= \infty).$$

PROOF. This is a straightforward application of the Borel–Cantelli lemmas together with the approximate distribution of $S_{n;t}$ given in Theorem 1.

To prove the left tail result, for positive integers s and the corresponding positive real constant c define the event

$$E(n,s) \equiv \left\{S_{n;t} < s \equiv b(n,t) - \frac{\log(c\log\log n)}{\log(1/r)}\right\}.$$

Uniformly for $\log c$ in any compact interval, as $n \to \infty$ we have $\lambda \equiv \lambda(n,t,s) \sim c\log\log n$ and, by Theorem 1, the error in approximating the distribution of $S_{n;t}$ is $|P(E(n,s)) - \exp(-\lambda)| = O((\log n)(\log\log n/n)$. Since we always argue via the Borel–Cantelli lemmas along an exponentially increasing skeleton, such as $n = n(k) = \exp(\varepsilon k)$, the crucial property is that the sum over k of these errors in the approximation is finite.

To prove the lower bound for the lim inf, fix a small $\varepsilon > 0$. Let $n \equiv n(k) \equiv \lfloor \exp(\varepsilon k)\rfloor$ and choose $s = s(n)$ such that $\log_{1/r}(c) \in [\varepsilon, 1 + \varepsilon)$. Since $c \geq$

$(1/r)^{\varepsilon} > 1$, as $k \to \infty$ we have $\lambda \sim c \log(\varepsilon k) \sim c \log k$ and $\sum_k \exp(-\lambda) < \infty$. Thus, $\sum_k P(E(n(k), s)) < \infty$, so the Borel–Cantelli lemma implies $0 = P(E(n(k), s) \text{ i.o. } (k))$. The behavior of $S_{n;t} - b(n, t)$ as $n \to \infty$ is controlled by the behavior along the skeletons $n = n(k)$ because $S_{n;t}$ is nondecreasing as n increases, and $b(n(k+1), t) - b(n(k), t) \to \varepsilon a/H$, which is small for small ε. This proves that $\liminf(\cdots) \geq -1$ almost surely.

To prove the upper bound for the lim inf is harder since we need independence to use the converse Borel–Cantelli lemma. Given $\varepsilon > 0$, we can choose, for all sufficiently large k, a value $n \equiv n(k) \in [\exp(k^{1+\varepsilon}), \exp(k^{1+\varepsilon} + 2H(a, p)/a)]$, and an integer $s = s(k)$ such that $s = b(n, t) - \log_{1/r}(c \log\log n)$ with $c \in [1 - 3\varepsilon, 1 - 2\varepsilon]$. Note that we have a superexponential time skeleton, with $n(k-1)/n(k) \to 0$. In order to get independence, we replace the event $E(n(k), s(k))$ by a modification $E^*(n, s)$ in which $S_{n;t}$ is replaced by $S^*_{n,t} \equiv \max_{n(k-1) < i \leq n(k) - t}(Z_{i+1} + \cdots + Z_{i+t})$, with $t = t(n(k))$. Note that $S^*_{n(k),t}$ involves coins Z_i for $i \in (n(k-1), n(k)]$, and is distributed like $S_{n(k) - n(k-1) - t;t}$, so, effectively, λ is decreased by a factor of $1 - [n(k-1) + t + 1]/n(k) \to 1$. The events $E^*(n(k), s)$ for $k \geq 1$ are mutually independent. As $k \to \infty$, we have $\lambda \sim c \log\log n \sim c \log(k^{1+\varepsilon}) \sim (1 + \varepsilon) c \log k$. Since $(1 + \varepsilon)c < 1 - \varepsilon$ for all k, we have $\sum_k \exp(-\lambda) = \infty$. Thus, $\sum_k P(E^*(n(k), s)) = \infty$, so the Borel–Cantelli lemma implies $1 = P(E^*(n(k), s) \text{ i.o. } (k))$. Now $s(k) - b(n(k-1), t(n(k-1)))$ is asymptotic to $(a/H)(k^{1+\varepsilon} - (k-1)^{1+\varepsilon})$, which grows like k^{ε}, while $\log\log n$ grows like $\log k$, so that $0 = P((S_{n(k-1); t(n(k))} \geq s(k)) \text{ i.o. } (k))$. Thus $1 = P(E(n(k), s) \text{ i.o. } (k))$. Now for integers s, $\{S_{n;t} < s\} = \{S_{n;t} \leq s - 1\}$, and ε arbitrarily small means that c is close to 1, so that $\log_{1/r} c$ is close to 0. Thus, we have proved that $\liminf(\cdots) \leq -1$ almost surely.

To prove the right tail results is much easier in each direction. For positive integers s and the corresponding real constant f, define the event

$$G(n, s) \equiv \{S_{n;t} \geq s \equiv b(n, t) + f\}.$$

Uniformly for $f/(\log\log n)$ bounded away from zero and infinity, as $n \to \infty$ we have $\lambda \equiv \lambda(n, t, s) \sim r^f$, and by Theorem 1, $|P(G(n, s)) - (1 - \exp(-\lambda))| = O((\log n)/n)$. As $\lambda \to 0$, the right tail probability is $1 - \exp(-\lambda) \sim \lambda$. Take $n = n(k) = \lfloor \exp(k) \rfloor$, so that as $k \to \infty$ we have $P(G(n, s)) \sim \lambda \sim r^f$. The result for $\sum_k r^{f(\exp(k))} < \infty$ follows from the Borel–Cantelli lemma, since $S_{n;t}$ is nondecreasing in n and $b(n(k+1), t) - b(n, k)$ is bounded. To get the result for $\sum_k r^{f(\exp(k))} = \infty$, replace the event $G(n, s)$ by $G^* \equiv \{\max_{n/2 \leq i \leq n} Z_{i+1} + \cdots + Z_{i+t} \geq b(n, t) + f\}$. Observe that $G^*(n, s) \subset G(n, s)$, and for large k, the events $G^*(n(k), s)$ are mutually independent. The effect of replacing G by G^* is to reduce λ by a factor close to $\frac{1}{2}$, so that $P(G^*)/P(G) \to \frac{1}{2}$. Thus $\sum_k r^{f(\exp(k))} = \infty$ implies $\sum_k P(G^*(n(k), f)) = \infty$, and the Borel–Cantelli lemma implies $P(G^*(n(k), f) \text{ i.o.}) = 1$, so that $1 = P(S_{n;t} \geq b(n, t) + f(n) \text{ i.o. }) = 1$. □

The distribution of long head runs interrupted only by a fixed number k of tails is described by the case $s = t - k$ of Theorem 1. Such runs are also discussed in Guibas and Odlyzko (1980), Karlin and Ost (1987) and in Gordon, Schilling and Waterman (1986), which has the law of the iterated logarithm corresponding to Corollary 3.

COROLLARY 3. *For $k = 0, 1, 2, \ldots,$ consider the length $R_n(k)$ of the longest "k-interrupted head run" starting in the first n tosses of p-coins Z_i. Formally,*

$$\{R_n(k) < t\} = \{S_{n;t} < t - k\}.$$

Define centering constants $c_n(k)$ by

$$c_n(k) \equiv \log_{1/p}(n) + k \log_{1/p} \log_{1/p}(n) - \log_{1/p}(k!) + \log_{1/p}\left((1-p)^{k+1} p^{-k}\right). \tag{16}$$

Then for each fixed k, the family $\{R_n(k) - c_n(k) | n \geq 1\}$ is tight and as $n \to \infty$,

$$\sup_x |P\{R_n(k) - c_n(k) < x\} - \exp(-p^x)| = O(n^{-1}\log(n)),$$

where the supremum is taken over x for which $x + c_n(k)$ is an integer.

It is natural to view the above case of $R_n(k)$ as interpolating, via $a = 1 - k/n$, between the cases R_n^1 and R_n^a for fixed $p < a < 1$. In particular, it seems surprising that the coefficient of the log log n term is positive in (16) for $c_n(k)$, while the coefficient is negative in (4) for c_n^a. However, there is no paradox; the normalizing factor for the log n term, $H(a, p)$, has infinite derivative at $a = 1$, so that small increases in a—as $a = 1 - k/\log_{1/p}(n) \to 1$ with k held fixed and $n \to \infty$—effect compensating changes in the $\log(n)/H(a, p)$ term of (4). This unification of current results for constant quality a with the previous results of Arratia, Gordon and Waterman (1986) is one benefit of using Poisson approximations like (13) as a conceptual foundation, in place of an extreme-value approximation. A second practical benefit is that numerical approximation is frequently better if one uses the Poisson approximation. Such a realization, in the context of maxima of i.i.d. normal variates is the subject of Hall (1980).

5. The Erdös-Rényi law for coin tossing: The longest quality *a* run.

Let $0 < p < a \leq 1$ and $\ldots, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ be p-coins. Let $R_n \equiv R_n^a$ be the length of the longest consecutive run, contained within the first n tosses, in which the fraction of heads is at least a. In this definition of R_n, if the phrase "contained in" were changed to "starting in," then the event $\{R_n < t\}$ would not be measurable with respect to $\sigma(Z_1, \ldots, Z_{n+l})$ for any finite l.

In terms of $S_{n;t}$, which is the maximum number of heads in a run of t tosses *starting* within the first n tosses,

$$\{R_n^a \geq t\} = \bigcup_{l \geq t} \{S_{n-l;l} \geq al\}. \tag{17}$$

Since we do not understand the relation between $S_{n;l}$ for $l = t, t+1, t+2, \ldots,$ we cannot completely approximate the distribution of R_n, even though we have a satisfactory approximation for the distribution of $S_{n;l}$ for each pair (n, l).

From (17) we get the crude lower and upper bounds

$$P(S_{n-t;t} \geq at) \leq P(R_n^a \geq t) \leq \sum_{l \geq t} P(S_{n-l;l} \geq al). \tag{18}$$

Now Theorem 1 says that $P(S_{n-t;t} \geq at)$ is approximately

$$1 - \exp(-(n-t)P(Z_1 + \cdots + Z_t = \lceil at \rceil)).$$

From this and the large deviation expansion (14) of $P(Z_1 + \cdots + Z_t = \lceil at \rceil)$, it follows that as $n, t \to \infty$, we have $P(S_{n-t;t} \geq at)$ tends to 1 (respectively, 0) if $t - \{\log(n) - \frac{1}{2}\log\log(n)\}/H(a,p)$ tends to minus infinity (respectively, infinity). When $P(S_{n-t;t} \geq at)$ is small, the sum over l in (18) is comparable to a convergent geometric series with ratio $\exp(-H(a,p))$, and the first term of this sum is $P(S_{n-t;t} \geq at)$, so the sum tends to zero when $P(S_{n-t;t} \geq at)$ tends to zero. Write $R_n^* \equiv R_n - \{\log(n) - \frac{1}{2}\log\log(n)\}/H(a,p)$. We have just shown that the family $\{R_n^*\}$ is tight.

Roundoff is a significant consideration in the analysis of R_n^a. If $Z_{\alpha+1} + \cdots + Z_{\alpha+l} \geq al$, and $\lceil al \rceil = \lceil al + a \rceil$, then regardless of the values of Z_α and $Z_{\alpha+l+1}$, there are quality a windows of length $l+1$ starting at α. Thus we define the set of "admissable" values for quality a,

$$A_a \equiv \{t \in Z : at \leq s < at + a \text{ for some } s \in Z\}.$$

The only way that R_n^a can have a value outside the admissable set A_a is with $R_n^a = n$.

We now work out a Poisson approximation for the number of clumps of runs underlying the event $\{R_n^a \geq t\}$. Even though we can't evaluate the expected number of clumps, this approximation lets us compare matches in sequence comparisons with runs in coin tossing. Furthermore, we believe that this clump analysis is the proper way to understand the relation between the dependent random lengths $S_{n;l}$ for $l \geq t$.

For t a positive integer and $a \in [0,1]$, use the indicators $Y(\alpha, s, t) \equiv 1(s = \sum_{k=0}^{t-1} Z_{\alpha+k})$ from the previous section to define indicators

$$Y'_\alpha \equiv Y'(\alpha, t) \equiv \max\{Y(\alpha, \lceil at \rceil, t), Y(\alpha, \lceil at + a \rceil, t+1), \ldots, Y(\alpha, \lceil 2at \rceil, 2t)\}$$

$$X'_\alpha \equiv X'(\alpha, t) \equiv Y'(\alpha, t) \prod_{j=1}^{2t} (1 - Y'(\alpha - j, t)).$$

Informally, Y'_α is almost the indicator that a quality a window of length t or greater begins at position α, and X'_α is the indicator that a clump of such windows begins at position α.

In the definition of Y'_α, we use the upper bound $2t$, since when the fraction of heads in a window is at least a, the same must be true in the first or second half of the window. If there is a quality a window of length t or greater within the first n tosses, there must be a quality a window of length between t and $2t$ inclusive, contained within the first n tosses. Among all such windows consider those of minimal length, in particular the leftmost of these, to see that

$$\{R_n \geq t\} \subset \left\{1 = \max_{1 \leq \alpha \leq n-t} Y'_\alpha\right\} \cup \{Z_1 + \cdots + Z_t > \lceil at \rceil\}.$$

Let $I \equiv \{1, 2, \ldots, n-t\}$, and in analogy to (8), the definition of W in the previous section, define

$$W' \equiv W'(n-t, a, t) \equiv \sum_{\alpha \in I} X'(\alpha, t).$$

Apart from boundary effects, W' is the number of clumps within the first n tosses of "quality a, length t or greater" head runs, so the event $\{R_n < t\}$ can be approximated by the event $\{W' = 0\}$. We have

$$\{R_n < t\} \subset \{W' = 0\} \cup \{Y'_{n-2t+1} + \cdots + Y'_{n-t} > 0\}$$

and

$$\{R_n \geq t, W' = 0\} \subset \{Y'_1 + \cdots + Y'_{2t} > 0\} \cup \{Z_1 + \cdots + Z_t > \lceil at \rceil\},$$

so that

$$\begin{aligned} |P(W' = 0) - P(R_n < t)| &\leq 3tEY'_1 + P(Z_1 + \cdots + Z_t > \lceil at \rceil) \\ &\leq (3t+1)e^{-tH(a,p)}. \end{aligned} \tag{19}$$

The upper bound

$$EY'_1 \leq e^{-tH(a,p)} \tag{20}$$

can be proved by Cramer's argument: compute expectations with respect to the probability Q under which $\ldots, Z_{-1}, Z_0, Z_1, Z_2, \ldots$ are a-coins, and observe that on the event $\{Y'_\alpha = 1\}$, the Radon–Nikodym derivative satisfies $dP/dQ \leq e^{-tH(a,p)}$.

We now establish a Poisson approximation for W' using the Chen–Stein method. Let $\lambda' \equiv EW'$. Let $B_\alpha \equiv \{\beta \in I: |\alpha - \beta| < 4t\}$ for $\alpha = 1$ to n, so that $b_3 = 0$. If $|\alpha - \beta| \leq 2t$ then $E(X'_\alpha X'_\beta) = 0$, but if $2t < |\alpha - \beta| < 4t$, then we can only conclude that $E(X'_\alpha X'_\beta) \leq (EX'_\alpha)EY'_\beta$, so that $b_2 < 4t\lambda' EY'_\alpha$ and $b_1 < 8t\lambda' EX'_\alpha$.

Using the Chen–Stein method as presented in Lemma 2, we have

$$\begin{aligned} |P(W' = 0) - e^{-\lambda'}| &\leq (b_1 + b_2)(1 \wedge 1/\lambda') \\ &\leq t\lambda'(8EX'_1 + 4EY'_1)(1 \wedge 1/\lambda') \\ &\leq 12tEY'_1. \end{aligned} \tag{21}$$

Combine the upper bound (19) on boundary effects, the upper bound (20) on EY'_1, and the result (21) of the Chen–Stein method, to obtain the following theorem.

THEOREM 2.

$$|P(R_n < t) - e^{-\lambda'}| \leq 16te^{-tH(a,p)}, \tag{22}$$

where $R_n \equiv R_n^a$, defined formally by (17), is the length of the longest quality a head run contained within n tosses of a p-coin, with $0 < p < a \leq 1$.

The relation (22) expresses a Poisson approximation in terms of the expected number of clumps, $\lambda' \equiv EW' = (n-t)EX_1'$. To calculate $EX_1' \equiv EX'(1,t)$ is not easy when $a < 1$. It is easy to calculate $EY_1' \equiv EY'(1,t)$, so essentially our problem is that we can't compute the average clump size EY_1'/EX_1'. In Section 2 we successfully computed the average clump size EY_1/EX_1, so to see why the method of Section 2 breaks down, we outline one way to calculate EY_1'.

Define a stopping time τ with values in $[t,\infty]$,

$$\tau \equiv \inf\{l \geq t\colon Z_1 + \cdots + Z_l = \lceil al \rceil\},$$

so that we have $Y_1' \equiv 1(\tau \leq 2t)$ and $Y_1 \equiv 1(\tau = t)$, and we can effectively calculate each term of $EY_1' = \Sigma_{t \leq l \leq 2t} P(\tau = l)$. With this approach, one reason that the method of Section 2 breaks down is that conditional on $\{\tau = l\}$, the coins $Z_1, \dots, Z_l$ are exchangeable only in the case $l = t$.

The simplest upper bound on R_n^a is just that the probability of seeing a quality a window of length t or greater contained in the first n tosses is no larger than the expected number of quality a windows of length t or greater starting in the first n tosses,

$$P(R_n^a \geq t) \leq nP(\tau < \infty). \tag{23}$$

This bound has the same order of decay as the upper bound in (18) and is simpler in that it avoids the issues of boundary effects, clumping, and Poisson approximation.

6. The Erdös–Rényi law for sequence matching. Let $\dots, A_{-1}, A_0, A_1, A_2, \dots$ and $\dots, B_{-1}, B_0, B_1, B_2, \dots$ be independent integer valued random "letters," say with distribution μ for the A_i's and distribution ν for the B_i's. Let

$$p \equiv p(\mu,\nu) \equiv \sum_{l \in Z} \mu_l \nu_l,$$

so that $\forall\, i, j$, $p = P(A_i = B_j)$, and assume that $0 < p < 1$. Let $a \in (p, 1]$. We are interested in the distribution, for large m and n, of the length $M_{m,n}^a$ of the longest quality a matching between two words, one from the sequence $A_1 \dots A_m$, and the other, of equal length, from the sequence $B_1 \dots B_n$. We are also interested in $N_{m,n;t} \equiv N(m,n;t)$, which is the maximum number of matches between two words of length t, one from $A_1 \dots A_m$ and the other from $B_1 \dots B_n$.

Let Z_{ij} be the indicators that there is a match between positions i in the first sequence and j in the second sequence.

$$Z_{ij} \equiv 1(A_i = B_j),$$

so that $p = P(Z_{ij} = 1) = 1 - P(Z_{ij} = 0)$ for each i, j. The Z_{ij}'s collectively are *not* a collection of p-coins, however, since distinct indicators Z_{ij} and Z_{kl} are *strictly* positively correlated if $i = k$ or else $j = l$, unless both sequences of letters A and B are *uniform* in distribution over a common finite alphabet.

However, for each fixed i, j, the sequence $Z_{ij}, Z_{i+1,j+1}, Z_{i+2,j+2}, \dots$ is a sequence of p-coins, i.e., these random variables are independent. Thanks to this independence, for each s, t, the indicators Y_α and X_α defined below, with a two-dimensional spatial index α, have the same expectations as the indicators Y_α and X_α of the previous sections, with a one-dimensional spatial index.

Here is the setup for a Poisson approximation to analyze the distribution of $N(m, n; t)$ along the same lines as the analysis of $S_{n;t}$ in Section 4. For positive integers $s \le t$, define indicators

$$(24)\qquad \begin{aligned} Y_{ij} &\equiv Y(i,j,s,t) \equiv 1\left(s = \sum_{k=0}^{t-1} Z_{i+k,j+k}\right), \\ X_{ij} &\equiv Y_{ij} \prod_{l=1}^{t} (1 - Y_{i-l,j-l}). \end{aligned}$$

Let $I \equiv \{1, 2, \dots, m - t + 1\} \times \{1, 2, \dots, n - t + 1\}$ and

$$W \equiv W(m,n,s,t) \equiv \sum_{\alpha \in I} X_\alpha, \qquad \lambda \equiv EW = (m - t)(n - t)EX_\alpha.$$

The random variable W represents the number of clumps of locations at which closely matching pairs of words occur. Since we are using the Chen–Stein method, we can establish a Poisson approximation for W if the second moment of W is well behaved. Whether the second moment of W blows up or is well behaved depends on a large deviation rate depending on a and the distributions μ and ν, as summarized by formula (33).

In the next two sections we analyze the strong dependence between two matching events which share letters. Define the indicator that the words of length t at position i in the first sequence and j in the second sequence match in at least s places,

$$V_{ij} \equiv V_{(i,j)} \equiv V(i,j,s,t) \equiv 1\left(s \le \sum_{0 \le k < t} 1(A_{i+k} = B_{j+k})\right).$$

We have, for each $\alpha = (i, j)$, that $X_\alpha \le Y_\alpha \le V_\alpha$, so that $E(X_\alpha X_\beta) \le E(V_\alpha V_\beta)$. It is easier to deal with the indicators V rather than X or Y, since V is increasing in the match indicators $Z_{ij} \equiv 1(A_i = B_j)$. The goal is to show that there is not too much positive correlation between X_α and X_β when $\alpha \equiv (i, j)$ and $\beta \equiv (i', j')$ are "adjacent," i.e., when $|i - i'| < t$ or $|j - j'| < t$.

6.1. *Doubly adjacent, nonparallel bundles of comparisons, a.k.a. accordions, bent spaghetti.* In this section we handle the case where $\alpha \equiv (i, j)$ and $\beta \equiv (i', j')$ are "doubly adjacent, but not parallel," i.e., $|i - i'| < t$ and $|j - j'| < t$ but $j - i \ne j' - i'$. The parallel case, $i - i' = j - j'$ and $\alpha \ne \beta$, requires no effort since in that case $X_\alpha X_\beta \equiv 0$. Essentially, we have already dealt with the strong dependence between adjacent, parallel bundles of comparisons, by going from the indicators Y to the indicators X, and computing a clump size $1/(a - p)$ in Section 2. Our result for the nonparallel case is that the exponential rate of decay, relative to t, of $E(X_\alpha X_\beta)$, is faster than the exponential rate of decay for EX_α. This is stated below formally as $2I(a) > H(a,p)$. In the situation of interest for a distributional approximation, we have two sequences of lengths m and n, and there are about mnt^2 doubly adjacent pairs α, β. Since t grows like $\log(mn)/H(a,p)$, our result says that the total correlation arising from doubly adjacent pairs of events is negligible.

Consider the exemplary doubly adjacent, nonparallel case $\alpha = (0,0)$, $\beta = (1,0)$, so that V_α is the indicator of at least s matches between letters A_k and B_k, and V_β is the indicator of at least s matches between letters A_{k+1} and B_k, for $k = 1,\dots,t$. There are a total of $2t$ comparisons involved here, which go back and forth between the two sequences A and B: A_1 is compared to B_1, which is compared to A_2, which is compared to B_2, which is compared to A_3 and so on. The pattern of these comparisons may be visualized as an accordion or a piece of bent spaghetti. With $s = \lfloor at \rfloor$, we are interested in the large deviation rate $I(a) \equiv I(a,\mu,\nu)$ defined by

$$(25) \qquad I(a) = \lim_{t\to\infty} \frac{-1}{2t} \log E(V_{00}V_{01}) = \lim_{t\to\infty} \frac{-1}{2t} \log E(X_{00}X_{01}).$$

The comparisons in the general doubly adjacent, nonparallel case can be expressed in terms of independent accordions, of individual length 1 or more, and total length $2t$. For example, with $t = 5$ and $\alpha = (0,2)$, $\beta = (1,0)$, there are three independent accordions: $A_0B_2A_3B_5$ with length 3, $B_0A_1B_3A_4B_6$ with length 4 and $B_1A_2B_4A_5$ with length 3.

We use the language of statistical mechanics for the following large deviation analysis. For the special case $\mu = \nu$, this analysis was carried out in Arratia and Waterman (1989) in the context of self-overlapping repeats within a single sequence, so here we stress the new features arising from $\mu \neq \nu$. See Ellis (1985) or Varadhan (1984) for expositions of large deviation theory. Consider the number $U(l)$ of matches, i.e., the energy, in an accordion of length l, starting in the sequence A (and noting that accordions starting in the sequence B will satisfy the same inequalities):

$$U(l) \equiv \sum_{0\le k<l/2} 1(A_k = B_k) + \sum_{0<k<l/2} 1(B_k = A_{k+1}).$$

Note, for l odd, that the first sum has one more term than the second sum. To understand the Laplace transform, $E\exp(\gamma U(l))$ for $\gamma \in R$, define the nonnegative real "transfer matrix" $M \equiv M(\gamma)$, indexed by the alphabet

$$(26) \qquad M_{ij} = \sqrt{\mu_i \nu_j}\,\exp(\gamma 1(i=j)).$$

Since M is symmetric only if the two distributions μ and ν are equal, we are interested in the growth rate $\lambda \equiv \lambda(\gamma)$, with "free energy" $f(\gamma) \equiv \log(\lambda)$, defined by

$$(27) \qquad \exp(2f(\gamma)) \equiv \lambda^2(\gamma) \equiv |||MM'|||.$$

That is, λ^2 is the spectral radius of the symmetric matrix MM', where M' is the transpose of M.

With column unit vectors $\mathbf{u} \equiv (\sqrt{\mu_i})_{i \in Z}$ and $\mathbf{v} \equiv (\sqrt{\nu_i})$, we have, with the matrix product having l factors, alternately M and M', that

$$E\exp(\gamma U(l)) = \mathbf{u}'MM'M \cdots M'\mathbf{u} \le \lambda^l(\gamma) \quad \text{for even } l,$$
$$E\exp(\gamma U(l)) = \mathbf{u}'MM'M \cdots M\mathbf{v} \le \lambda^l(\gamma) \quad \text{for odd } l.$$

The first of the above two inequalities, for the case with l even, is true virtually by definition. For the case with l odd, observe that, with I the identity matrix, the matrix $I - \mathbf{v}\mathbf{v}'$ is positive semidefinite, so that

$$\begin{aligned}(Ee^{\gamma U(l)})^2 &= \mathbf{u}'MM' \cdots M\mathbf{v}(\mathbf{u}'MM' \cdots M\mathbf{v})' \\ &= \mathbf{u}'MM' \cdots M\mathbf{v}\mathbf{v}'M'M \cdots M'\mathbf{u}' \le \mathbf{u}'(MM')^{2l}\mathbf{u} \le \lambda^{2l}.\end{aligned}$$

By decomposing the $2t$ comparisons involved in V_α and V_β in the nonparallel case into independent accordions of total length $2t$, and applying the inequalities above, we have that $\lambda(\gamma)^{2t} = \exp(2tf(\gamma))$ dominates the expected value of the exponential of γ times the number of matches achieved. For α and β not parallel, the event indicated by $V_\alpha V_\beta$ requires at least $2s \equiv 2at$ matches, so for each $\gamma \in R$ we have the exponential bound,

$$E(V_\alpha V_\beta) \le \exp(-2t(a\gamma - f(\gamma))).$$

The optimal exponential rate, i.e., coefficient of $2t$ in the upper bound above, is obtained by maximizing $a\gamma - f(\gamma)$, using the value $\gamma \equiv \gamma_a$ which satisfies $f'(\gamma_a) = a$. Large deviation theory shows that this rate is $I(a)$ and that the functions I and f are each other's Legendre transform. For the computation of upper bounds, the key relation is that I is the Legendre transform of f:

$$\forall\, a \in (p,1), \quad I(a) = \sup_{\gamma \in R}(a\gamma - f(\gamma)) = a\gamma_a - f(\gamma_a), \qquad \text{where } f'(\gamma_a) = a; \tag{28}$$

and $I(1) = \log(1/p)$.

The argument that $2I(a) > H(a,p)$ for all $a \in (p,1]$ uses Fenchel's duality relation for Legendre transforms, and is given in detail in Arratia and Waterman (1989). To display the relation between $I(a)$ and $\frac{1}{2}H(a,p)$, compute the square of the Hilbert–Schmidt norm of M, trace$(MM') = \sum_{i,j} M_{ij}M_{ij} = \sum_{i,j}\mu_i\nu_j + \sum_i \mu_i\nu_i(e^{2\gamma} - 1) = 1 - p + pe^{2\gamma}$, which is greater than or equal to λ^2, so that $f(\gamma) \ge \frac{1}{2}\log(1 - p + pe^{2\gamma})$. We observe that, as a function of a, $H = H(a,p)$ is the Legendre transform of $\log(1 - p + pe^{\gamma})$ as a function of γ, and $\frac{1}{2}H$ is the Legendre transform of $\frac{1}{2}\log(1 - p + pe^{2\gamma})$.

Lemma 3 summarizes these findings.

LEMMA 3. *With $s \ge at$, $a \in (p,1]$ and $\alpha = (i,j)$, $\beta = (i',j')$ with $|i - i'| < t$ and $|j - j'| < t$, and $\alpha \ne \beta$,*

$$E(X_\alpha X_\beta) \le e^{-2tI(a)},$$

where the large deviation rate I, characterized by relations (25), (26), (27), *and* (28), *satisfies $2I(a) > H(a,p)$.*

6.2. *Singly adjacent bundles of comparisons, a.k.a. crabgrass.* In this section we handle the case where $\alpha \equiv (i, j)$ and $\beta \equiv (i', j')$ are "singly adjacent," i.e., $|i - i'| < t$ *or* $|j - j'| < t$, but *not* both. We analyze the exponential decay rate for $E(X_\alpha X_\beta)$. The overall result is somewhat surprising: Even in the simplest case, $\mu = \nu$ (so that both sequences have the same distribution) and $m = n$ (so that both sequences have the same length), there are cases with $a \in (p, 1)$ with so much positive correlation between indicators X_α and X_β that the second moment of W blows up.

In the exemplary case $\alpha = (0, 0)$, $\beta = (0, t)$, the event indicated by $V_\alpha V_\beta$ involves $2t$ comparisons between pairs of letters. The dependence graph for these comparisons can be viewed as t independent accordions of length 2, namely, $B_k A_k B_{t+k}$ for $0 \le k < t$. Drawn with the sequence A on the bottom, each accordion resembles the letter v and the entire picture of overlapping v's resembles crabgrass. In the general case, say, with $\alpha = (i, j)$, $\beta = (i + d, j')$ with $0 \le d < t$ and $|j - j'| \ge t$, exactly d of the v's split into pairs of independent single edges.

Consider the conditional probability c that three independent letters match, given that the first two match, say, with the first letter chosen with distribution μ and the other two with distribution ν:

$$c \equiv c(\mu, \nu, \nu) \equiv P(B_2 = A_1 | A_1 = B_1) = \frac{1}{p} \sum_{l \in Z} \mu_l \nu_l \nu_l = \frac{\sum_{l \in Z} \mu_l \nu_l \nu_l}{\sum_{l \in Z} \mu_l \nu_l}.$$

Note that $cp = P(A_1 = B_1 = B_2)$ and

$$\frac{p - cp}{1 - p} = P(B_2 = A_1 | A_1 \ne B_1).$$

Since $p = E(\nu(A_1))$ and $cp = E(\nu(A_1)^2) \ge p^2$, we have $c \ge p$, with equality if and only if ν restricted to the support of μ is constant. In the special case $\mu = \nu$, we have from Jensen's inequality that $cp < p^{3/2}$, i.e., $c < \sqrt{p}$.

Let $G(a; \mu, \nu)$ be the large deviation rate defined by

$$G(a; \mu, \nu) = \lim_{t \to \infty,\, s/t \to a} -\frac{1}{t} \log E(V_{0t} | V_{00} = 1).$$

Note that $G(a; \mu, \nu) = -H(a, p) + \lim(-1/t) \log E(V_{00} V_{0t})$. For $a = 1$, we have simply $G(1; \mu, \nu) = \log(1/c)$. In the case $c = 1$, which is allowed since we have only assumed that $0 < p < 1$, we will have simply that $G(a; \mu, \nu) = 0$ for all a.

We now deal with the remaining cases. Assume for this paragraph that $c < 1$ and $p < a < 1$. Consider the indices $i \in [1, t)$ for which $A_i = B_i$ and let $b \in (c, 1)$ denote the fraction of these for which $A_i = B_{t+i}$. Roughly speaking, this accounts for abt matches out of at comparisons, so that we still need $(a - ab)t$ matches $A_i = B_{i+t}$ from the $(1 - a)t$ positions i for which $A_i \ne B_i$. Thus, b is such that $(a - ab)/(1 - a) \in ((p - cp)/(1 - p), 1)$. We see that

$G(a;\mu,\nu)$ depends on μ and ν only through the values of p and c, and that

$$(29)\qquad G(a;\mu,\nu) = \inf_b\left[aH(b,c) + (1-a)H\left(\frac{a-ab}{1-a},\frac{p-cp}{1-p}\right)\right].$$

The derivative with respect to b of the expression in brackets is

$$(30)\quad a\left[\log\left(bp(1-c)^2(1-2a+ab)\right) - \log\left(ca(1-b)^2(1-2p+cp)\right)\right].$$

Therefore, the critical value of b is a root of a quadratic equation in b. Recall that $c \ge p$. The equation can be written as $Ab^2 + Bb + C$, with $A = a(c-p) \ge 0$, $B = -2a(c-p) - p(1-c)^2 < 0$ and $C = ac(1-2p+cp) \ge 0$. Hence, we may solve explicitly for the critical value b_0. Specifically, $b_0 = a$ when $c = p$, and

$$(31)\qquad b_0 = 1 - \frac{p(1-c)^2}{2a(c-p)}\left(\left(1 + \frac{4a(1-a)(c-p)}{p(1-c)^2}\right)^{1/2} - 1\right),$$

when $c > p$. In the latter case, (31) is the only critical value less than 1 and the minimum is attained there. The concavity of the square root implies that $a < b_0$ when $c > p$.

It is also convenient to express the large deviation rate $G(a;\mu,\nu)$ relative to the rate $H(a,p)$, so we define

$$(32)\qquad \theta \equiv \theta(a;\nu,\mu) \equiv \frac{G(a;\mu,\nu)}{H(a,p)} = \lim \frac{\log E(V_{0t}|V_{00}=1)}{\log E(V_{00})}.$$

The limit is taken as $s,t \to \infty$ with $s/t \to a$, and we recall that the indicators V_{ij} involve matching two segments of length t and requiring s or more matches. Observe that by interchanging the roles of the two sequences, we have

$$\theta(a;\nu,\mu) \equiv \frac{G(a;\mu,\nu)}{H(a,p)} = \lim \frac{\log E(V_{t0}|V_{00}=1)}{\log E(V_{00})}.$$

Informally, the parameter θ measures how hard it is to obtain a second quality a matching choosing fresh letters for only one of the sequences, relative to how hard it is to obtain the first quality a matching. The two extreme cases should be pointed out. If ν restricted to the support of μ is constant, so that $c = p$, then $\theta(a;\mu,\nu) = 1$ for all a, which can be seen either directly or by checking that $b = a$ is the critical value of b. If ν is unit mass on a letter in the support of μ, then $\theta(a;\mu,\nu) = 0$ for all a.

LEMMA 4. *If $\alpha = (i,j)$ and $\beta = (i',j')$ with $|i-i'| < t$ and $|j-j'| \ge t$, and $s \ge at$ with $p < a \le 1$, then*

$$E(V_\alpha V_\beta) \le \exp(-t(1+\theta)H(a,p)),$$

where $\theta \equiv \theta(a;\mu,\nu)$ is given by (29), (31) *and* (32).

PROOF. Consider first the case $\alpha = (0,0)$, $\beta = (0,t)$. Let

$$U \equiv \sum_{0 \le k < t} 1(A_k = B_k), \quad U' \equiv \sum_{0 \le k < t} 1(A_k = B_{t+k})$$

denote the numbers of matches involved in the events indicated by V_α and V_β, so that $V_\alpha V_\beta = 1(U \ge s)1(U' \ge s) \le 1(U + U' \ge 2s)$. We compute, for real γ, $\exp(tF(\gamma)) \equiv Ee^{\gamma(U+U')} = (cpe^{2\gamma} + 2(p - cp)e^{\gamma} + 1 - 2p + cp)^t$. For each γ, there is the exponential upper bound $P(U + U' \ge 2s) \le \exp(-t(2a\gamma - F(\gamma)))$. Large deviation theory for the two-dimensional vectors $(1(A_k = B_k), 1(A_k = B_{t+k})) \in R^2$ implies, for fixed $a > p$, with $s = \lfloor at \rfloor$, that $E(X_\alpha X_\beta)$, $E(V_\alpha V_\beta)$ and $P(U + U' \ge 2s)$ all have the same exponential rate of decay relative to t, which we have labeled as $G(a; \mu, \nu) + H(a,p) = (1 + \theta)H(a,p)$, and that this rate is also realized by optimizing the exponential upper bound. Thus, $(1 + \theta)H(a,p) = \sup_\gamma(2a\gamma - F(\gamma)) = 2a\gamma_a - F(\gamma_a)$, with $\gamma_a > 0$.

Now changing α, β to the general case, so that U and U' become $U \equiv \Sigma_{0 \le k < t} 1(A_{i+k} = B_{j+k})$, $U' \equiv \Sigma_{0 \le k < t} 1(A_{i'+k} = B_{j'+k})$, changes $Ee^{\gamma(U+U')}$ by replacing some of the factors of $(cpe^{2\gamma} + 2(p - cp)e^{\gamma} + 1 - 2p + cp)$, corresponding to accordions of length 2, with factors of $(pe^{\gamma} + 1 - p)^2$, corresponding to pairs of independent edges. To check that

$$\left(cpe^{2\gamma} + 2(p - cp)e^{\gamma} + 1 - 2p + cp\right) \ge \left(pe^{\gamma} + 1 - p\right)^2 \quad \text{for } \gamma \ge 0,$$

simply check for equality at $\gamma = 0$ and verify the appropriate inequality for the derivative with respect to γ, using $c \ge p$ for this. Thus, the exponential upper bound, with $\gamma = \gamma_a$, proves this lemma. □

To interpret the above lemma, consider the case where $m, n \to \infty$ with

$$\log(n)/\log(mn) \to \rho > 0, \qquad \log(m)/\log(mn) \to 1 - \rho > 0.$$

We calculate exponential growth rates, writing $\approx$ to show that the logarithms of two quantities are asymptotic. In order to have $\lambda \equiv EW$ bounded away from zero and infinity, we must have $mn \approx e^{tH(a,p)}$. There are asymptotically $2tmn^2 \approx mn^2 \approx (mn)^{1+\rho}$ terms $E(X_\alpha X_\beta)$ with $\alpha = (i,j)$ and $\beta = (i', j')$, with $|i - i'| < t$ and $|j - j'| \ge t$, so the total contribution to $E(W^2)$ from these terms is $\approx (mn)^{1+\rho}\exp(-t(1 + \theta)H) \approx (mn)^{1+\rho-(1+\theta)} = (mn)^{\rho-\theta}$, with $\theta = \theta(a; \mu, \nu)$.

The preceding calculation shows that the second moment of W blows up, due to pairs of events which share letters in the sequence A but not B, if $\rho > \theta(a; \mu, \nu)$, and not if $\rho < \theta(a; \mu, \nu)$. Similarly, the second moment blows up due to pairs of terms which share letters in the sequence B if $1 - \rho > \theta(a; \nu, \mu)$, and not if $1 - \rho < \theta(a; \nu, \mu)$. Notice the interchange in the order of μ and ν in the arguments to θ. Thus, the key condition for there to be a successful Poisson approximation using W, for *some* choice of the relative growth rates of m and n, is that

$$\theta(a; \mu, \nu) + \theta(a; \nu, \mu) > 1, \tag{33}$$

so that we can find a value $\rho \in (1 - \theta(a; \nu, \mu), \theta(a; \mu, \nu))$.

For the case $\mu = \nu$ and $a = 1$, we have $G(1; \mu, \mu) = \log(1/c)$, so that $\theta(1; \mu, \mu) = \log c / \log p$, with $\theta(1; \mu, \mu) > \frac{1}{2}$ by Jensen's inequality. By continuity, for each nontrivial μ (so that $0 < p < 1$), (33) is satisfied for all a sufficiently close to 1.

Example 1. Condition (33) is not always satisfied. With $\mu = \nu = (0.75, 0.09, 0.09, 0.07)$ we compute $p = 0.5836$ and $c = 0.72597$. Using (31) yields $\theta = 0.498$ when $a = 0.7$. Thus, the variance of W explodes for all choices $m, n \to \infty$.

Example 2. With $\mu = (0.5, 0.5)$, $\nu = (x, 1 - x)$, $a \in (0.5, 1]$ we have $\theta(a; \nu, \mu) = 1$ and, if $x = 0$, then $\theta(a; \mu, \nu) = 0$. Thus, by picking x positive but close to zero, we have an example with $\frac{1}{2} \notin (1 - \theta(a; \nu, \mu), \theta(a; \mu, \nu)) \neq \varnothing$. Thus, a Poisson approximation is proved by our method for some $m, n \to \infty$ but not with $m = n$.

In particular, with $a = 1$ and $m = n$, the condition for a Poisson approximation to be proved by our method is that $x \in (x_0, 1 - x_0)$, where $x_0 = 0.178\ldots$ is the smaller solution of $\frac{1}{2} = \theta(1; \mu, \nu) = -\log_2(1 - 2x(1 - x))$. This same example is treated in Arratia and Waterman [(1985), page 1237], where it is seen that $\lim(M^1_{m,n}/R^1_{mn}) = 1$ a.s. if and only if $x \in [x_1, 1 - x_1]$, where $x_1 = 0.11002\ldots$ is the smaller solution of $x \log(x) + (1 - x)\log(1 - x) = -\log(2)/2$.

Example 3. The vector of probabilities

$$\mu = \nu = \left(\frac{42896}{121024}, \frac{17309}{121024}, \frac{17556}{121024}, \frac{43263}{121024} \right)$$

figures prominently in the experiments reported in Section 7. For these alphabet probabilities, $p = 0.2949$ and $c = 0.3262$. Corresponding to $a = 0.6, 0.7, 0.8, 0.9$, we obtain from (31) and (32) that $\theta = 0.9062, 0.9056, 0.9062, 0.9088$. Note that θ is not monotone in a.

6.3. *Putting it all together.* To use the Chen–Stein method, for $\alpha = (i, j) \in I$, we take $B_\alpha \equiv \{\beta = (i', j') \in I\colon |i - i'| < 2t \text{ or } |j - j'| < 2t\}$. With this choice, $b_3 = 0$ and $b_1 < 4t(m + n)\lambda EX_\alpha$.

There is a small contribution to b_2 from pairs α, β which are not adjacent, but for which $\beta \in B_\alpha$, i.e., with $t \le |i - i'| < 2t$ and $t \le |j - j'|$, or with $t \le |i - i'|$ and $t \le |j - j'| < 2t$. For these pairs, we have that X_α and X_β are dependent, but either X_α and Y_β are independent, so that $E(X_\alpha X_\beta) \le E(X_\alpha Y_\beta) = (EX_\alpha)(EY_\beta)$, or Y_α and X_β are independent, yielding the same upper bound. Thus, the total contribution here is at most $4t^2 \lambda EY_\alpha$.

There are fewer than $4t^2 mn$ doubly adjacent pairs (α, β). The total contribution to b_2 from these, by Lemma 3, is at most $4t^2 mn \exp(-2tI(a))$. In all situations of interest this contribution decays exponentially fast in t, because $mn \approx \exp(tH(a, p))$ and $2I(a) > H(a, p)$.

The main contribution to b_2 is from singly adjacent pairs (α, β), i.e., crabgrass. Consider those pairs which share letters in the A sequence, i.e., with $|i - i'| < t$ and $|j - j'| \geq t$. There are fewer than $2tmn^2$ such pairs, so by Lemma 4, the total contribution is at most $2tmn^2 \exp(-t(1 + \theta(a; \mu, \nu))H(a, p))$. The contribution from pairs which share letters in the B sequence has a similar bound, interchanging m with n and μ with ν. In situations of interest, our upper bound has the same exponential growth rate as the contribution itself, and may decay or grow exponentially with t.

The net result for using the Chen–Stein method is that

$$b_3 = 0, \quad b_1 < 4t(m + n)\lambda EX_\alpha$$

and

$$b_2 < 4t^2\lambda EY_\alpha + 4t^2mne^{-2tI(a)} + 2tmn^2 \exp[-t(1 + \theta(a; \mu, \nu))H(a, p)] \\ + 2tnm^2 \exp[-t(1 + \theta(a; \nu, \mu))H(a, p)].$$

Theorem 3. *Let $N(m, n; t)$ be the maximum number of matches between a word of length t taken from $A_1 \dots A_m$ and a word of length t from $B_1 \dots B_n$, with independent letters. Let $S(n, t)$ be the maximum number of heads in a run of length t starting within the first n tosses of p-coins Z_i, as described in Theorem 1, with $p = P(A_i = B_j)$. If $\rho \equiv \log n/\log(mn)$ satisfies $\theta(a; \mu, \nu) > \rho > 1 - \theta(a; \nu, \mu)$, where θ is defined by (32), and $s = at$, then $P(N(m, n; t) < s)$ can be approximated by $P(S_{mn;t} < s)$. More precisely,*

$$|P(N(m, n; t) < s) - P(S((m - t)(n - t), t) < s)| \\ < (4t^2 + 5t(m + n) + 7t + 1)P(Z_1 + \cdots + Z_t \geq s) + 4t^2mne^{-2tI(a)} \\ + 2tmn^2 \exp[-t(1 + \theta(a; \mu, \nu))H(a, p)] \\ + 2tnm^2 \exp[-t(1 + \theta(a; \nu, \mu))H(a, p)].$$

This upper bound converges to zero faster than some negative power of (mn).

Proof. To control boundary effects in using $W \equiv W(m, n, s, t)$ to approximate the distribution of $N \equiv N(m, n; t)$, we observe that $\{W \neq 0\} \subset \{N \geq s\}$ and

$$\{N \geq s, W = 0\} \subset \bigcup_\alpha \{V_\alpha = 1\},$$

where the union is over $\alpha = (i, j) \in I$, with $i \leq t$ or $j \leq t$, so that

$$\begin{aligned} |P(W = 0) - P(N < s)| &\leq t(m + n - 2t + 2)EV_\alpha \\ &\leq t(m + n)\exp(-tH(a, p)). \end{aligned} \tag{34}$$

A similar bound is given as (9) in approximating a probability for head rich regions in coin tossing, $P(S_{n;t} < s)$ by $P(W(n, s, t) = 0)$. We recopy (9) below, changing the dummy variable "n" to "l":

$$|P(S_{l;t} < s) - P(W(l, s, t) = 0)| \leq tP(Z_1 + \cdots + Z_t = s) \\ + P(Z_1 + \cdots + Z_t > s).$$

There is a potentially confusing reuse of notation, since in both Section 4 and here, we have a random variable W which is a sum of random variables called X_α. However, if the values of p, s, t agree, then EX_α is the same for the two different versions of X_α. With $l = (m - t)(n - t)$, the random variable $W(l, s, t)$ of Section 4 and the random variable $W \equiv W(m, n, s, t)$ of this section have the same expectation, namely, $EW = lEX_\alpha$. As in the proof of Theorem 1, the Chen–Stein method gives a bound of the form $|P(W = 0) - e^{-EW}| < (1 \wedge (EW)^{-1})(b_1 + b_2 + b_3)$. Thus our theorem follows by combining four upper bounds, namely for b_1, b_2, boundary effects and (11) of Theorem 1. □

THEOREM 4. *Let $M^a_{m,n}$ be the length of the longest quality a matching pair of words, one taken from $A_1 \dots A_m$ and the other from $B_1 \dots B_n$, with independent letters. Let R^a_n be the length of the longest quality a head run starting within the first n tosses of p-coins Z_i, as described in Theorem 2, with $p = P(A_i = B_j)$. If $\log n/\log(mn) \to \rho$ as $m, n \to \infty$, and ρ satisfies $\theta(a; \mu, \nu) > \rho > 1 - \theta(a; \nu, \mu)$, where θ is defined by (32), then the total variation distance between $M^a_{m,n}$ and $R^a_{(mn)}$ converges to zero, faster than some negative power of (mn).*

PROOF. The proof of this theorem bears exactly the same relation to the proof of Theorem 3 that the proof of Theorem 2 bears to the proof of Theorem 1. It is actually $R^a_{((m-t)(n-t))}$ that is involved, but this in turn is close to $R^a_{(mn)}$. Thus, for $a \in (p, 1]$, a positive integer t, and positional indices $\alpha = (i, j) \in [1, m - t + 1] \times [1, n - t + 1]$, we define indicators $X'_\alpha \le Y'_\alpha \le V'_\alpha$ to parallel the construction of Section 5. In particular, X'_α will have the same expectation as it had in Section 5. We have $V'_{ij} \equiv \max_{t \le k \le 2t}\{V(i, j, \lceil ak \rceil, k\}$, so the upper bound of Lemma 3 applies to V' after throwing in a factor of $1/(1 - e^{-I(a)})$ corresponding to summing a geometric series. The bound of Lemma 4 applies unchanged to V', by using Cramér's argument, as in the proof of the upper bound (20). The combination of upper bounds to control boundary effects and two Poisson approximations with the same $EW' = lEX'_\alpha$, with $l = (m - t)(n - t)$, is an upper bound on $|P(M^a_{m,n} < t) - P(R^a_l < t)|$. This upper bound is less than some constant, independent of m, n, t, times the upper bound from Theorem 3. □

7. Two experiments. In this section, we test the applicability of our results. We first present a simulation experiment in which the four letter alphabet $\{a, c, g, t\}$ is generated according to probabilities $\mu = \nu = (0.3544, 0.1430, 0.1451, 0.3575)$. These probabilities are the subject of Example 3 in Section 6.2. The probability of a single match of two independently chosen letters is $p = 0.2949$. The probabilities are determined by the proportions of base pairs in the subject of our second experiment, the complete chloroplast genome of the liverwort *Marchantia polymorpha*, taken from the GenBank

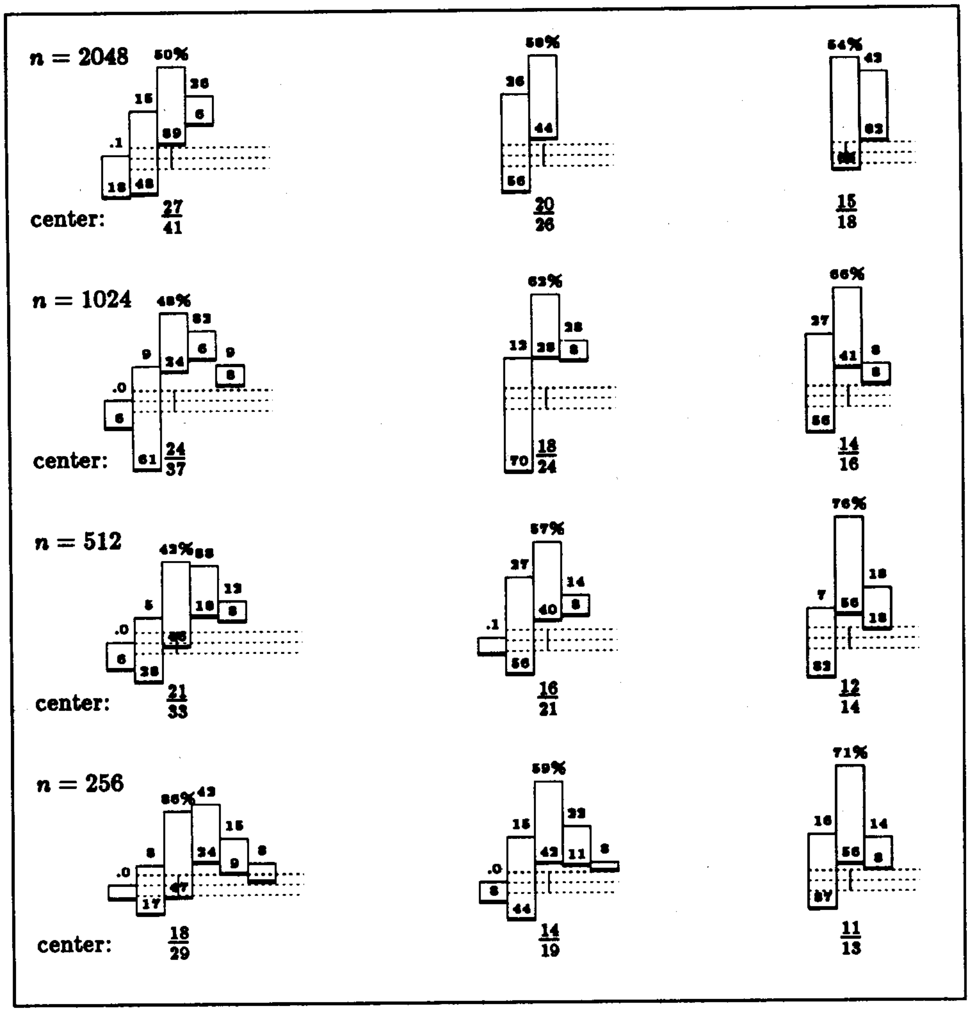

FIG. 1. *Simulation results—Empirical histogram hanging from predicted histogram after arcsine transformation.*

database. See Fickett and Burks (1989) for a description of the GenBank database.

The results of our first simulation experiment are presented in Figure 1. For each choice of $n \in \{256, 512, 1024, 2048\}$ and three choices of t, we present histograms summarizing 200 realizations of $N(n, n; t)$, the maximum number of matches belonging to length t windows in two independently generated sequences, each of length n, consisting of i.i.d. letters from a four letter alphabet.

Corresponding to n, we select window lengths $t = \lfloor \ln((a-p)n^2)/H(a,p) \rfloor$, for $a \in \{0.7, 0.8, 0.9\}$. Histograms are independent for differing values of n, but the same random sequences are used to find the richest matching windows for fixed n and varying t. The random number generator used was that provided by the MATLAB package, in which all programming was done on a Sun 3/260 computer. See Moler, Ullman, Little and Bangert (1987).

We use Theorem 3 to center the histograms of simulation results. Specifically, for given values of n and t, we call $\lfloor s_c \rfloor$ the center of the histogram where s_c is the solution to the equation

$$1 = n^2\left(\frac{s}{t} - p\right)\frac{\Gamma(1+t)}{\Gamma(1+s)\Gamma(1+t-s)}p^s(1-p)^{t-s}. \tag{35}$$

The centering constant s_c used here is asymptotically equivalent to the center $b(n,t)$ defined in (15). For example, for $n = 2048$ and $t = 41$, the integer part of the solution to (35) is $\lfloor s_c \rfloor = 27$. These values appear underneath corresponding histograms labeled as "center," in the form $\lfloor s_c \rfloor / t$.

Tolerance intervals based on Theorem 3 are given by the triple horizontal dotted lines. These cover horizontally at least 99% of the predicted distribution, from the lower 0.005 fractile to the upper 0.995 fractile. We report on $3 \times 4 \times 200 = 2400$ realizations. Were the approximate distribution exact, one would expect fewer than two simulated values to exceed the tolerance limits in each combination of a and n.

Values of the approximating distribution using Theorem 3 appear above the corresponding histogram bars. For example, we approximate the probability that for $n = 2048$ the maximally rich 41-regions of two independent sequences has approximate probability

$$\begin{aligned} P\{N(2048,2048;41) = 27\} &\\ = \exp[&-2007^2(28/41 - p)P\{\text{binomial}(41, 0.2949) = 28\}] \\ &- \exp[-2007^2(27/41 - p)P\{\text{binomial}(41, 0.2949) = 27\}] \\ = 0.5046&. \end{aligned}$$

The predicted value at the centered value is indicated by appending a percent sign. The heights of the histogram bars are determined for the arcsine transformation of the predicted approximating probability. Specifically, if Theorem (3) yields h, for the predicted probability of observing exactly s matches in the richest matching t-interval, we plot the top of the bar at $2\arcsin(\sqrt{h_s})$.

The results of the simulation experiment are reported numerically above the heavy black lines in Figure 1. Consider, for example, the upper left glyph. For the 200 simulations with $n = 2048$ and $t = 41$, 39% were observed to have had exactly 27 matches in common for maximally comparable 41-segments, while 6% were observed to have had exactly 28 matches in common. The range of maximally rich comparisons is $28 - 25 = 3$.

The thick lower bar is plotted at the difference of predicted minus observed on the arcsine transformation scale. Specifically, let $\hat{h}$, denote the observed frequency of observing $\{N_{n,n;t} = s\}$ in $m = 200$ simulations. The bottom of bar s appears at height $2\arcsin(\sqrt{h_s}) - 2\arcsin(\sqrt{\hat{h}_s})$.

Were our predicted frequencies to fit the observed simulation histograms perfectly, we would see the usual histogram shape, sitting on a horizontal fat black line. Deviations of the bottoms of the bars from a common base gives a graphical depiction of deviations of observed from expected.

The vertical distance between the highest and lowest of the three horizontal dotted lines correspond to ±2.57 standard errors of $\hat{h}_s$, meaningful across a broad range of values because arcsine transformation stabilizes binomial variances. Bottoms of histogram bars lying outside the dotted range suggest a lack of fit in the Poisson approximation to the distribution of $N(n, n; t)$.

A number of features of the simulation experiment are anticipated by mathematical results in the preceding sections. First, observe that the conditions of Theorem 3 are satisfied. From Example 3, $\theta > 0.9$, so that use of (13) is warranted by Theorem 3.

On a gross basis, the distribution of $N(n, n; t)$ is well approximated by the Poisson. The predicted center $\lfloor s_c \rfloor$ is a satisfactory location parameter. The predicted 95% tolerance region contains most of the empirical distribution. The distributions for similar values of a appear tight. The predicted shape is similar to extreme value, with a very rapidly decaying left tail and a much thicker right tail.

The fine structure of the empirical distribution is not so well predicted. The simulated empirical distributions appear to be greater than predicted by (13) for $s < s_0$, and smaller than predicted for $s > s_c$. Perhaps this is not so surprising, given that the tails of the binomial distribution are heavier than extreme value on the right, and lighter than extreme value on the left.

We present in Figure 2 the results of our second experiment in a format identical to that of Figure 1. The complete genome of *Marchantia polymorpha* is given as a sequence of 121,024 letters from the alphabet $\{a, c, g, t\}$. For each of four choices of n, the sequence was cut into blocks of exactly 256, 512, 1024 or 2048 letters, with the remaining letters ignored. A sample of 200 pairs from the population of all pairs having identical lengths was taken, and the richest common matching t-segment was found for the same choices of t as in the simulation experiment.

It is remarkable, given the known dependence of adjacent nucleotides, that predictions based on assumptions of i.i.d. generation of sequences should fit as well as they do. For an analysis of dependence among adjacent nucleotides, see Tavare and Giddings (1989).

The predicted center continues to be a good center for the empirical distribution. In all cases the empirical mode is at $\lfloor s_c \rfloor$ or at $\lfloor s_c \rfloor + 1$. The empirical distribution is less concentrated than the simulated distribution, no doubt attributable to departures from distributional assumptions. There are happily few outliers, suggesting that the approximate distribution of richest matching segments could be a suitable tool for screening for interesting regions of similarity.

Ohyama et al. (1986) study the chloroplast gene organization using the complete DNA sequence. We cannot relate the outlier found in the $n = 1024$ comparisons to their work, but the outlying comparison for $n = 512$ has

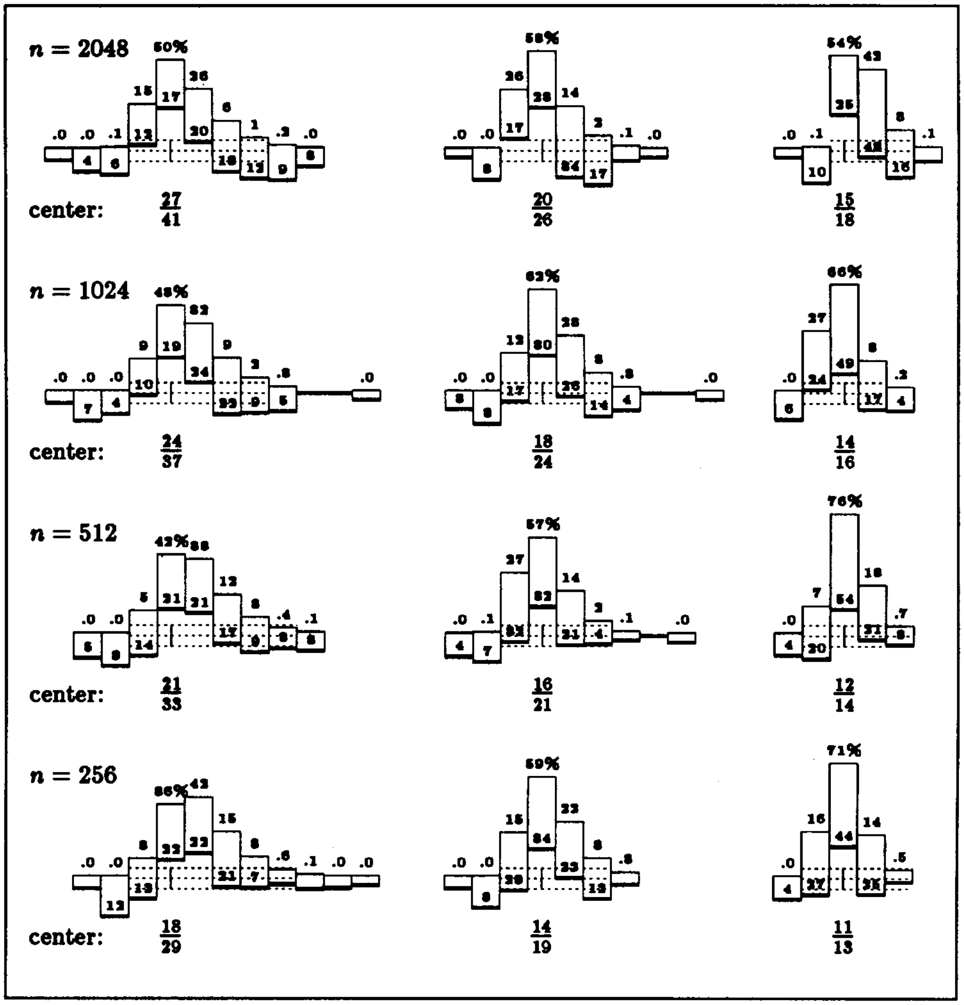

FIG. 2. *Chloroplast results—Empirical histogram hanging from predicted histogram after arcsine transformation.*

intriguing biological implications. Higher organisms have gene (protein encoding) DNA sequences interrupted by so-called intervening sequences which are removed from transcribed RNA by a mechanism known as splicing. There is as yet no consensus regarding the biological role of intervening sequences.

The $n = 512$ outlier corresponds to a match of length $t = 21$ beginning at nucleotide 26,665 and at nucleotide 67,475. The first segment is located in an intervening sequence within the gene coding a tRNA for lysine, while the second segment is located in an intervening sequence within open reading frame ORF203. (Open reading frames indicate regions of DNA that could encode proteins, although it has not yet been determined whether the region's

DNA is actually translated into proteins.) We are continuing to study both the $n = 512$ and $n = 1024$ outliers.

Although our results apply to segment comparison with mismatches, insertions and deletions are common in the evolutionary process. We hope ultimately to include the case of insertions and deletions in a distributional Erdös-Rényi law. We expect, because the dependence introduced by comparisons with insertions and deletions remains local, that the methods developed here—in particular, use of the Chen-Stein Poisson approximation—will be useful in the more complicated and scientifically more interesting problem.

REFERENCES

ALDOUS, D. (1989). Probability approximations via the Poisson clumping heuristic. *Applied Mathematical Sciences* **77** Springer, Berlin.

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

ARRATIA, R. and WATERMAN, M. S. (1989). The Erdös-Rényi strong law for pattern matching with a given proportion of mismatches. *Ann. Probab.* **17** 1152–1169.

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. (1986). An extreme value theory for sequence matching. *Ann. Statist.* **14** 971–993.

BARBOUR, A. D. (1982). Poisson convergence and random graphs. *Math. Proc. Cambridge Philos. Soc.* **92** 349–359.

BARBOUR, A. D. and HOLST, L. (1989). Some applications of the Stein-Chen method for proving Poisson convergence. *Adv. Appl. Probab.* **21** 74–90.

CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545.

DEHEUVELS, P. and DEVROYE, L. (1987). Limit laws of Erdös-Rényi-Shepp type. *Ann. Probab.* **15** 1363–1386.

DEHEUVELS, P., DEVROYE, L. and LYNCH, J. (1986). Exact convergence rate in the limit theorems of Erdös-Rényi and Shepp. *Ann. Probab.* **14** 209–223.

ELLIS, R. S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*. Springer, Berlin.

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

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

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

FICKETT, J. W. and BURKS, C. (1989). Development of a database for nucleotide sequences. In *Mathematical Methods for DNA Sequences* (M. S. Waterman, ed.) 1–44. CRC Press, Boca Raton, Fla.

GONCHAROV, V. L. (1942). On the field of combinatory analysis. *Doklady Akad. Nauk SSR* **35** 9. [Translated in *Amer. Math. Soc. Transl.* (*2*) **19** 1–46.]

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

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

HALL, P. (1980). Estimating probabilities for normal extremes. *Adv. Appl. Probab.* **12** 491–500.

KARLIN, S. and OST, F. (1987). Counts of long aligned word matches among random letter sequences. *Adv. Appl. Probab.* **19** 293–351.

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

KÓMLOS, J. and TUSNÁDY, G. (1975). On sequences of "pure heads." *Ann. Probab.* **3** 608–617.

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

MOLER, C., ULLMAN, M., LITTLE, J. and BANGERT, S. (1987). Pro-MATLAB User's Manual. The Math Works, Inc. Sherborn, Mass.

NAUS, J. I. (1979). An indexed bibliography of clusters, clumps, and coincidences. *Internat. Statist. Rev.* **47** 47–78.

NAUS, J. I. (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77** 177–183.

OHYAMA, K. ET AL. (1986). Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* **322** 572–574.

RÉVÉSZ, P. (1978). Strong theorems on coin tossing. In *Proc. International Congress of Mathematicians, Helsinki, 1978.*

STEIN, C. M. (1986). *Approximate Computation of Expectations*. IMS, Hayward, Calif.

TAVARE, S. and GIDDINGS, B. W. (1989). Some statistical aspects of the primary structure of nucleotide sequences. In *Mathematical Methods for DNA Sequences* (M. S. Waterman, ed.) 117–132. CRC Press, Boca Raton, Fla.

VARADHAN, S. R. S. (1984). Large deviations and applications. *SIAM Regional Conference Series in Applied Mathematics* **46**. SIAM, Philadelphia.

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

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