The Annals of Applied Probability
2000, Vol. 10, No. 3, 709–725

r-SCAN STATISTICS OF A MARKER ARRAY IN MULTIPLE SEQUENCES DERIVED FROM A COMMON PROGENITOR[1]

BY SAMUEL KARLIN AND CHINGFER CHEN

Stanford University

This study is motivated by problems of molecular sequence comparisons for biological traits conserved or lost over evolution time. A marker of interest is distributed in the genome of the ancestor and inherited among l offspring species which descend from this common ancestor. Each marker will be retained or lost during the evolution of the descendent species. The objective of the analysis here is to ascertain probabilities of clustering or overdispersion of the marker array among the sequences of the descendent species. Limiting distributions for the extremal r-scan statistics (defined in text) of the trait distributed among the l dependent offspring processes are derived by adapting the Chen–Stein Poisson approximation method. Results that accommodate new occurrences of the trait (gene) arising from duplications and transposition occurrences are also described. The r-scan statistical analysis is further applied to a multi sequence combined Poisson model where $\{B_1, \ldots, B_l\}$ are generated from m independent Poisson processes $\{A_1, \ldots, A_m\}$ such that $B_k = \cup_{i \in Z_k} A_i$, where $\{Z_k\}_{1 \le k \le l}$ are subsets of $\{1, 2, \ldots, m\}$.

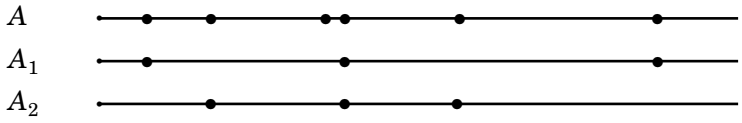
1. Introduction. The r-scan statistics (see below) of a single sequence were introduced in Karlin and Macken (1991) for purposes of characterizing nonrandomness in the distribution of a marker array in DNA or amino acid sequence data. r-scan statistics can also be used to identify significant peaks in analysis of counts in sliding windows. The new models and statistics presented in this paper aim to handle *correlated* sequence data and *multiple* arrays of markers. A marker of interest is distributed in the genome of an ancestor and species descendent from this common ancestor where their DNA sequences maintain some characteristics of the ancestor. Each marker will be retained or lost and some will be newly acquired during the evolution of the descendent species. The objective of our analysis is to evaluate probabilities of clustering or overdispersion of the marker array across the ensemble sequences of the descendent species. In separate publications, we will present models that extend the r-scan statistics to deal with processes of gene and/or marker deletions, duplication and displacement.

Studies of inhomogeneities in long DNA sequences can be insightful to the organization of the human genome. Particular markers (e.g., specific DNA restriction sites, nucleosome placements, gene locations) are distributed over the genome along chromosomes. Let X_i be the gap (in DNA units) between

Received July 1997; revised October 1999.
[1]Supported in part by NIH Grant 2R01HG00335-11 and 5R01GM10452-35.
AMS 1991 *subject classifications*. Primary 60E05; secondary 60G50.
Key words and phrases. r-scan statistics, Chen-Stein Poisson approximation, Poisson processes, total variation distance, asymptotic distributions.

the ith and the $i+1$th markers. Questions about spacings of a marker array and general issues of sequence heterogeneity lead to statistical considerations of the r-scan process $\{R_i = \sum_{j=i}^{i+r-1} X_j\}$, the array of distances between the ith and the $i+r$th markers, $i = 1, 2, 3, \ldots$, where r is an integer parameter. By varying r, organization on different scales can be detected, for example, $r = 2$ and 3 can aptly detect near neighbor interactions while $r = 10$ can discern marker concentrations over a greater range. Moreover, the r-scan ($r \geq 2$) statistics are better able to tolerate measurement errors and reduce effects of statistical fluctuations compared with $r = 1$ lengths. For previous literature and applications of r-scan statistics in molecular sequence analysis, see Karlin and Macken (1991), Dembo and Karlin (1992), Karlin and Brendel (1992), Masse, Karlin, Schachtel and Mocarski (1992), Karlin and Cardon (1994), Karlin, Mrázek and Campbell (1996), Gerstein (1997) and Reinert and Schbath (1998). For studies of clustering in other domains with an extensive bibliography, see Naus (1979, 1982).

Consider a marker distributed in the genome of an ancestor and l species descendent from this common ancestor. The DNA sequences of the l offspring species maintain some characteristics of the forebear according to the following assumptions. (1) The occurrences of the marker in the ancestor sequence are distributed randomly as the points of a renewal process A following a distribution function $F(x)$ governing the interval lengths $\{X_i\}_{i\geq 1}$ between successive marker points. (2) With each marker point in the ancestor sequence, there is a corresponding random indicator $\underline{D}$ of l components that describes the retention of the marker in the l offspring sequences. Thus if the kth component of the indicator is 1, the marker is retained in the kth species sequence and deleted otherwise. (3) All the indicator variables $\{\underline{D}_i\}_{i\geq 1}$ are independently distributed following the distribution $P_{\underline{D}}$ on $\{0,1\}^l$. Then the construction of the l offspring processes from the ancestor array A depends on the realization of $\{\underline{D}_i\}_{i\geq 1}$ such that if the kth component of $\underline{D}_i$ equals 1, the ith marker point of A is maintained in the kth sequence; otherwise, there is no point in the kth sequence at the given location. The following display illustrates the construction:



where $D_1 = \binom{1}{0}$, $D_2 = \binom{0}{1}$, $D_3 = \binom{0}{0}$, $D_4 = \binom{1}{1}$, $D_5 = \binom{0}{1}$ and $D_6 = \binom{1}{0}, \ldots$.

There is another model, which we refer to as the combined Poisson model, which is also of interest from an evolutionary perspective. Explicitly, given m independent Poisson processes $A_1, A_2, \ldots, A_m$, there are l dependent point processes $B_1, B_2, \ldots, B_l$, constructed from these independent arrays according to l subsets of $\{1, 2, \ldots, m\}$, labeled $Z_1, Z_2, \ldots, Z_l$. The point process B_k, $1 \leq k \leq l$, is generated by aggregating all the events (points) from $\{A_i\}_{i\in Z_k}$, producing $B_k = \cup_{i\in Z_k} A_i$. Let $\{a_i\}_{i=1}^m$ denote the parameters for the Poisson processes $\{A_i\}_{i=1}^m$. The parameter for the Poisson process B_k is $b_k = \sum_{i\in Z_k} a_i$.

We assume, for convenience, a smallest value among $\{b_k\}_{k=1}^{l}$, $b_1 < \min(b_2, b_3, \ldots, b_l)$. In Section 5, the asymptotic distributions of the smallest and the largest r-scan lengths among the l dependent arrays $\{B_k\}_{k=1}^{l}$ is set forth relying on Theorems 1 and 2 of this paper.

In the course of evolution, the number of markers in each offspring species sometimes increases (e.g., by duplications and transpositions caused by species specific selection). We can accommodate this effect in the following formulation. Assume the ancestral marker is distributed as a homogeneous Poisson (α) array. Consider l descendent sequences of retained or deleted trait positions. Assume, in addition, new trait positions are added to each offspring sequence distributed as independent Poisson processes with parameters $\gamma_1, \gamma_2, \ldots, \gamma_l$, respectively. Then the trait distribution in sequence k, $k = 1, 2, \ldots, l$, is again Poisson with parameter $\alpha p_k + \gamma_k$, where p_k is a probability that a trait position of the ancestral sequence is retained in the kth offspring process.

A concrete example for the combined Poisson model related to evolutionary phylogeny can be instructive. Consider three independent Poisson processes on $(0, \infty)$ with parameters α, β and γ, labeled Processes A, B and C, respectively. We construct two dependent Poisson processes A_1 and A_2 from A in such a way that each A point is retained in Processes A_1 and A_2 with probability p_1 and p_2 separately. This applies independently for each point of A. Join points from Processes A_1 and B to form the point process B_1. Also join points from Processes A_2 and C to form a point process C_1. The asymptotic distributions of the smallest and largest r-scan lengths among the processes of B_1 and C_1 can be deduced from a two-array model as follows. First consider a Poisson process $\tilde{A}$ with parameter $\alpha + \beta + \gamma$. Let $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. Then processes B_1 and C_1 can be extracted from $\tilde{A}$ specifying the probability density $P_{\underline{D}}$ on $\{0, 1\}^2$ as

$$P_{\underline{D}}\binom{1}{1} = \frac{\alpha}{\alpha+\beta+\gamma} p_1 p_2, \qquad P_{\underline{D}}\binom{1}{0} = \frac{\beta}{\alpha+\beta+\gamma} + \frac{\alpha}{\alpha+\beta+\gamma} p_1 q_2,$$

$$P_{\underline{D}}\binom{0}{1} = \frac{\gamma}{\alpha+\beta+\gamma} + \frac{\alpha}{\alpha+\beta+\gamma} q_1 p_2 \quad \text{and} \quad P_{\underline{D}}\binom{0}{0} = \frac{\alpha}{\alpha+\beta+\gamma} q_1 q_2.$$

We return to the general model of an ancestor sequence and l descendent sequences. To analyze the smallest r-scan length among the descendent sequences, we will first construct an associated rth order Markov chain realizing the evolutionary process. Let $\{S_i\}$ denote a sequence of $l \times (r+1)$ random matrices, which are generated from $\{\underline{D}_i\}$ as follows. Let $\underline{1}$ designate the l-length column vector with all components 1. For each realization $\{\underline{d}_i\}$, define $S_1 = (\underline{1}, \underline{d}_1, \ldots, \underline{d}_r)$, and $S_i = (\underline{d}_{i-1}, \ldots, \underline{d}_{i+r-1})$ for $i \geq 2$. Clearly $\{S_i\}$ is a Markov chain of order r, whose stationary distribution is

$$\Pi((\underline{d}_1, \ldots, \underline{d}_{1+r})) = \prod_{i=1}^{r+1} P_{\underline{D}}(\underline{d}_i).$$

Let Q be the collection of possible realizations of S_i, $i \geq 2$. Also let Q_r be a subset of Q, consisting of all states containing at least one row with all

components 1. In the ongoing discussion, let F_m denote the m-fold convolution of F, for $m \geq 1$. The following theorem describes the asymptotic distribution of the minimum r-scan length for the l-array model.

THEOREM 1. *Let m_t be the minimum r-scan length in $(0, t)$ from all the l subprocesses, and let n_t be the renewal count of the A process in $(0, t)$. For a prescribed positive constant λ, let a_t be determined to satisfy the equation*

$$\lambda = (N_t - r + 1)\Bigg(\sum_{(\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_{r+1}) \in Q_r} \Pi\Big((\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_{r+1})\Big)\Bigg) F_r(a_t), \tag{1}$$

where $N_t = E[n_t]$. Then m_t has the asymptotic distribution

$$\lim_{t \to \infty} \Pr\{m_t > a_t\} = \exp\{-\lambda\}. \tag{2}$$

An error estimate for the asymptotic distribution (2) *is given by*

$$O\left(\sqrt{\frac{\ln t}{t}}\right) + O(M_F(a_t)) + O\left(\max_{\underline{s} \in Q_r} E_{\Xi_t}\left|1 - \frac{n(\underline{s}|\Xi_t)}{N_{\underline{s}, t}}\right|\right),$$

for $M_F(a) = \sum_{k=1}^{\infty} F_k(a)$ (the renewal function of F),

$$\begin{aligned} \Xi_t &= \text{the sample path of } (S_1, S_2, \ldots, S_{N_t - r + 1}), \\ n(\underline{s}|\Xi_t) &= \text{the number of occurrences of the state } \underline{s} \text{ given the realization } \Xi_t, \\ N_{\underline{s}, t} &= \text{the unconditional mean number of occurrences of state } \underline{s} \text{ in} \\ &\qquad S_1, S_2, \ldots, S_{N_t - r + 1}. \end{aligned}$$

The proof is elaborated in Section 3.

For the discussion of the limiting distribution of the maximal r-scan length M_t, let H denote the set of all the l-length column vectors $\{\underline{d} = (d_1, d_2, \ldots, d_l), d_1, d_2, \ldots, d_l \in \{0, 1\}\}$. The marginal probability p_k for an A point to be retained in the kth subprocess is $p_k = \sum_{\underline{d} \in H,\, d_k = 1} P_{\underline{D}}(\underline{d})$, $1 \leq k \leq l$. The distribution F of the A-renewal process is assumed to satisfy the property

$$\lim_{x \to \infty} \frac{1 - F_1(x - a)}{1 - F_2(x)} = 0 \quad \text{for each positive constant } a, \tag{3}$$

and consequently the property

$$\lim_{x \to \infty} \frac{1 - F_m(x - a)}{1 - F_{m+1}(x)} = 0 \quad \text{for all } m \geq 1$$

[see Lemma 4.2, Dembo and Karlin (1992)]. The general condition

$$0 < F(x) < 1 \quad \text{for } 0 < x < \infty \quad \text{and } F(x) \text{ is continuous at } 0 \tag{4}$$

is assumed for F.

Let $G^{(k)}$, $1 \leq k \leq l$, denote the distribution of interval lengths between events in the kth subprocess. An easy calculation gives

$$1 - G^{(k)}(x) = (1 - F(x)) + \sum_{\nu=1}^{\infty} \{F_\nu(x) - F_{\nu+1}(x)\}(1 - p_k)^\nu. \tag{5}$$

Let $G_m^{(k)}$ denote the m-fold convolution of $G^{(k)}$, for $m \geq 1$. If $p_1 < p_k$, for $k \geq 2$, (5) yields easily

$$1 - G^{(k)}(x) < 1 - G^{(1)}(x) \quad \text{for all } x > 0.$$

and consequently,

$$1 - G_m^{(k)}(x) < 1 - G_m^{(1)}(x) \quad \text{for } m = 1, 2, \ldots . \tag{6}$$

Again, let $n_t^{(k)}$ denote the renewal count of the kth subprocess in $(0, t)$ and $N_t^{(k)} = |E[n_t^{(k)}]|$, for $1 \leq k \leq l$. The asymptotic limit law for the maximal r-scan length is as follows.

THEOREM 2. *Let M_t be the maximum r-scan length in $(0, t)$ from all the l subprocesses. Assume the distribution function F possesses properties* (3) *and* (4)*, and among the marginal probabilities $p_1, p_2, \ldots, p_l$, say for definiteness, $p_1 < \min(p_2, p_3, \ldots, p_l)$. For a given positive constant μ, let b_t be determined to satisfy the condition*

$$\mu = (N_t^{(1)} - r + 1)(1 - G_r^{(1)}(b_t)), \tag{7}$$

We have

$$\lim_{t \to \infty} \Pr\{M_t < b_t\} = \exp\{-\mu\}. \tag{8}$$

An error estimate for the Poisson distribution (8) *is*

$$O\left(\sqrt{\frac{\ln t}{t}}\right) + (2r-1)O\left(1 - G_r^{(1)}(b_t)\right) + O\left(\sum_{j=1}^{r-1} \Pr\{\tilde{R}_{j+1} \geq b_t | \tilde{R}_1 \geq b_t\}\right)$$
$$+ O\left(\max_{2 \leq k \leq l}\left\{\frac{1 - G_r^{(k)}(b_t)}{1 - G_r^{(1)}(b_t)}\right\}\right),$$

where $\tilde{R}_j = \sum_{i=j}^{j+r-1} \tilde{X}_i$, with $\{\tilde{X}_i\} \sim G^{(1)}$ i.i.d.

See Section 4 for the proof of Theorem 2.

2. Preliminary lemmas. Proofs are based on application of the Chen–Stein Poisson approximation method. Error estimates of the approximation typically involve only the first two moments of the sum in question. Within the context of this paper, let Z_λ denote the Poisson(λ) random variable and let the total variation distance of two random variables U, V be defined as usual and written as

$$d(U, V) = \sup_{\mathscr{A}} |\Pr\{U \in \mathscr{A}\} - \Pr\{V \in \mathscr{A}\}|.$$

Several elementary properties of the total variation distance which are referenced in the sequel are reviewed in Dembo and Karlin [(1992), page 336].

The Chen–Stein method following the Arratia, Goldstein and Gordon (1989) formulation is stated in the following lemma [see also Chen (1975), Barbour, Holst and Janson (1992)].

LEMMA 1 [Arratia, Goldstein and Gordon (1989)]. *Let* $\{U_i\}$ *be Bernoulli random variables with corresponding parameters* $\{\theta_i\}$ *and* $C = \sum_{i\in I} U_i$, *where* I *is a finite or countable index set. Then*

$$d(C, Z_\lambda) \le (c_1 + c_2)\frac{1 - e^{-\lambda}}{\lambda} + c_3 \ \min\left(1, \frac{\sqrt{2}}{\sqrt{\lambda}}\right), \tag{9}$$

where

$$\lambda = \sum_{i\in I} \theta_i, \qquad c_1 = \sum_{i\in I}\sum_{j\in \mathscr{W}_i} \theta_i\,\theta_j,$$

$$c_2 = \sum_{i\in I}\sum_{\substack{j\in \mathscr{W}_i \\ j\neq i}} E[U_i\,U_j], \qquad c_3 = \sum_{i\in I} E[|E[U_i|\{U_j\}_{j\notin \mathscr{W}_i}] - \theta_i|],$$

and $\{\mathscr{W}_i\}$ *is an appropriate family of subsets indexed by* I.

The r-scan process $\{R_i = \sum_{j=i}^{i-r+1} X_j\}$ generated by $\{X_j\}$ of i.i.d. variables is discussed in the paper of Dembo and Karlin (1992). In that context, the Chen–Stein method is applied to the Bernoulli sums $C^-(a) = \sum_{i=1}^{n-r+1} U_i^-(a)$ and $C^+(b) = \sum_{i=1}^{n-r+1} V_i^+(b)$, for $U_i^-(a) = \mathbf{I}\{R_i \le a\}$ and $V_i^+(b) = \mathbf{I}\{R_i \ge b\}$, where $\mathbf{I}\{\cdot\}$ denotes the indicator function of $\{\cdot\}$. The Poisson approximations of $C^-(a)$ and $C^+(b)$ by Z_λ and Z_μ, for $\lambda = E[C^-(a)]$ and $\mu = E[C^+(b)]$, is described in the following lemma.

LEMMA 2. *Let* $X_1, X_2, \ldots, X_n$ *be i.i.d. positive random variables with distribution function* $F(x)$ *and denote by* $F_m(x)$ *the* m*-fold convolution of* $F(x)$. *Define*

$$\lambda = (n - r + 1)F_r(a), \qquad \mu = (n - r + 1)[1 - F_r(b)].$$

Then

$$d(C^-(a), Z_\lambda) \le (1 - e^{-\lambda})\left[(2r - 1)F_r(a) + 2\sum_{m=1}^{r-1} F_m(a)\right],$$

$$d(C^+(b), Z_\mu) \le (1 - e^{-\mu})\left[(2r - 1)(1 - F_r(b)) + 2\sum_{m=1}^{r-1} \Pr\{R_{m+1} > b | R_1 > b\}\right].$$

The extremal statistics for the r-scan lengths from a stationary process are also developed in Dembo and Karlin (*1992*). We state next the corresponding result.

LEMMA 3. *Consider a finite state stationary Markov chain $\mathscr{M}$ taking values from a finite set $S = \{1, 2, \ldots, s\}$. Conditioned on the realization $\Lambda_1, \Lambda_2, \ldots, \Lambda_n$ from $\mathscr{M}$, let $Y_1, \ldots, Y_n$ be independent positive random variables and identically distributed for those Λ_i in the same state, say γ, with the distribution function $F^{(\gamma)}$. We abbreviate the event $\Lambda_1 = l_1, \Lambda_2 = l_2, \ldots, \Lambda_n = l_n$ by $\mathbf{\Lambda_n} = \underline{l}$ [$\mathbf{n} = (1, \ldots, n)$, $\underline{l} = (l_1, \ldots, l_n)$]. Then*

$$d(C^-(a), Z_\lambda) \le 4r \max_\gamma F^{(\gamma)}(a) + E_{\mathbf{\Lambda_n}}|\lambda(\mathbf{\Lambda_n}) - \lambda|,$$

$$d(C^+(b), Z_\mu) \le 4r \max_{\gamma_1, \ldots, \gamma_{r+1}} \Pr\left(\sum_{i=2}^{r+1} Y_i(\gamma_i) > b \,\middle|\, \sum_{i=1}^{r} Y_i(\gamma_i) > b\right) + E_{\mathbf{\Lambda_n}}|\mu(\mathbf{\Lambda_n}) - \mu|.$$

for

$$\lambda = E[C^-(a)], \qquad \lambda(\mathbf{\Lambda_n}) = E[C^-(a)|\mathbf{\Lambda_n}],$$
$$\mu = E[C^+(b)], \qquad \mu(\mathbf{\Lambda_n}) = E[C^+(b)|\mathbf{\Lambda_n}].$$

3. The asymptotic theorem for the minimal r-scan length.

PROOF OF THEOREM 1. Let $R_i = \sum_{j=i}^{i+r-1} X_j$ denote the ith r-scan length of the A process and let $R_{(r+1),i} = \sum_{j=i}^{i+r} X_j$ denote the ith $(r+1)$-scan length of A. We then define the Bernoulli random variables $U_i^-(a_t) = \mathbf{I}\{S_i \in Q_r, R_i \le a_t\}$, for $i \ge 1$. Two sums of the Bernoulli variables, $\tilde{C}_t^-(a_t) = \sum_{i=1}^{n_t - r+1} U_i^-(a_t)$ and $C^-_{N_t - r+1}(a_t) = \sum_{i=1}^{N_t - r+1} U_i^-(a_t)$, are constructed. Paralleling the discussion of Dembo and Karlin (1992), we bound the total variation distance between these two Bernoulli sums by conditioning its value on the following two events:

$$\mathscr{E}_t = \left\{\left|\frac{n_t}{N_t} - 1\right| \le \sqrt{\frac{\ln t}{t}}\right\},$$
$$\mathscr{E}_t^c = \text{the complement of } \mathscr{E}_t.$$

The Berry–Esseen estimate applied to n_t gives

$$d(\tilde{C}_t^-(a_t), C^-_{N_t - r+1}(a_t)) \le O\left(\sqrt{\frac{\ln t}{t}}\right) + O\left(\sqrt{\frac{1}{t}}\right) = O\left(\sqrt{\frac{\ln t}{t}}\right). \tag{10}$$

For $\mathbf{\Xi_t}$ a sample path of $\{S_i\}_{i=1}^{N_t - r+1}$, let $(\lambda|\mathbf{\Xi_t}) = E[C^-_{N_t - r+1}(a_t)|\mathbf{\Xi_t}]$ and $(Z_\lambda|\mathbf{\Xi_t}) = Z_{(\lambda|\mathbf{\Xi_t})}$. Then by specifying the neighborhood sets $\{\mathscr{W}_i\}$ as $\mathscr{W}_i = \{j: |j - i| \le r\}$, the Poisson approximation for $C^-_{N_t - r+1}(a_t)$ can be deduced by

the Chen–Stein method (Lemma 1) as follows:

$$
\begin{aligned}
&d\big(C^-_{N_t-r+1}(a_t), Z_\lambda\big)\\
&\quad\le E_{\Xi_t}\big[d\big((C^-_{N_t-r+1}(a_t)|\boldsymbol{\Xi}_\mathbf{t}\big), \big(Z_\lambda|\boldsymbol{\Xi}_\mathbf{t}\big)\big)\big] + E_{\Xi_t}\big[d\big((Z_\lambda|\boldsymbol{\Xi}_\mathbf{t}\big), Z_\lambda\big)\big]\\
&\quad\le E_{\Xi_t}\Bigg\{\frac{1-\exp(-(\lambda|\boldsymbol{\Xi}_\mathbf{t}))}{(\lambda|\boldsymbol{\Xi}_\mathbf{t})}\Bigg(\sum_{i=1}^{N_t-r+1}\sum_{j\in\mathscr{W}_i} E[U_i^-(a_t)|\boldsymbol{\Xi}_\mathbf{t}]E[U_j^-(a_t)|\boldsymbol{\Xi}_\mathbf{t}]\\
&\qquad\qquad + \sum_{i=1}^{N_t-r+1}\sum_{j\in\mathscr{W}_i,\, j\ne i} E[U_i^-(a_t)U_j^-(a_t)|\boldsymbol{\Xi}_\mathbf{t}]\Bigg)\Bigg\} + E_{\Xi_t}|(\lambda|\boldsymbol{\Xi}_\mathbf{t}) - \lambda| \qquad (11)\\
&\quad\le O(F_r(a_t)) + O(M_F(a_t)) + E_{\Xi_t}|(\lambda|\boldsymbol{\Xi}_\mathbf{t}) - \lambda|.
\end{aligned}
$$

The first two terms of (11) will converge to 0 as a_t tends to 0, stipulating that F is continuous at 0. And following the argument of (7.14) of Dembo and Karlin (1992), we evaluate the last term of (11),

$$
\begin{aligned}
E_{\Xi_t}|(\lambda|\boldsymbol{\Xi}_\mathbf{t}) - \lambda| &\le \sum_{\underline{s}\in Q_r} N_{\underline{s},\,t} E[R_i \le a_t | S_i = \underline{s}] E_{\Xi_t}\left|1 - \frac{n(\underline{s}|\boldsymbol{\Xi}_\mathbf{t})}{N_{\underline{s},\,t}}\right|\\
&\le \lambda \max_{\underline{s}\in Q_r} E_{\Xi_t}\left|1 - \frac{n(\underline{s}|\boldsymbol{\Xi}_\mathbf{t})}{N_{\underline{s},\,t}}\right|.
\end{aligned}
$$

The quantity above approaches 0 by the ergodic theorem.

Finally, since

$$
\begin{aligned}
\{\tilde{C}_t^-(a_t) = 0\} &= \{m_t > a_t;\ \tilde{C}_t^-(a_t) = 0\} \cup \{m_t \le a_t;\ \tilde{C}_t^-(a_t) = 0\}\\
&= \{m_t > a_t\} \cup \{m_t \le a_t;\ \tilde{C}_t^-(a_t) = 0\}
\end{aligned}
$$

and

$$
\{m_t \le a_t;\ \tilde{C}_t^-(a_t) = 0\} \subseteq \{\exists\, i,\ i \le n_t - r,\ S_i \notin Q_r,\ R_{(r+1),\,i} \le a_t\},
$$

we have

$$
\begin{aligned}
&|\Pr\{m_t > a_t\} - \Pr\{\tilde{C}_t^-(a_t) = 0\}|\\
&\quad\le \Pr\{\exists\, i,\ i \le n_t - r,\ S_i \notin Q_r,\ R_{(r+1),\,i} \le a_t\}\\
&\quad\le \Pr\{\exists\, i,\ i \le n_t - r,\ R_{(r+1),\,i} \le a_t\}\\
&\quad\le \Pr\left\{\left|\frac{n_t}{N_t} - 1\right| \le \sqrt{\frac{\ln t}{t}};\ \exists i,\ i \le n_t - r,\ R_{(r+1),\,i} \le a_t\right\} \qquad (12)\\
&\qquad + \Pr\left\{\left|\frac{n_t}{N_t} - 1\right| > \sqrt{\frac{\ln t}{t}}\right\}\\
&\quad\le \lambda\left(1 + \sqrt{\frac{\ln t}{t}}\right) O(F(a_t)) + O\left(\sqrt{\frac{1}{t}}\right).
\end{aligned}
$$

The conjunction of (10), (11) and (12) produces the estimate

$$|\Pr\{m_t > a_t\} - \exp\{-\lambda\}| \le O\left(\sqrt{\frac{\ln t}{t}}\right) + O(M_F(a_t)) + O\left(\max_{\underline{s}\in Q_r} E_{\Xi_t}\left|1 - \frac{n(\underline{s}|\Xi_t)}{N_{\underline{s},t}}\right|\right).$$

This completes the proof of Theorem 1. □

4. The asymptotic theorem for the largest r-scan length. Under the formulation of the multiple-array model, the l subprocesses are also renewal processes with the sojourn distributions $\{G^{(k)}\}_{k=1}^{l}$ defined in (5). To validate Theorem 2, the following lemmas are germane.

LEMMA 4. *If F has the property of* (3)*, then for any fixed constant $a \ge 0$, $G^{(1)}$ possesses the same property,*

$$\lim_{x\to\infty} \frac{1 - G^{(1)}(x-a)}{1 - G_2^{(1)}(x)} = 0. \tag{13}$$

PROOF. The formulation of (5) yields

$$\begin{aligned}
&\frac{1 - G^{(1)}(x)}{1 - G_2^{(1)}(x)} \\
&\quad = \frac{(1-F(x)) + \sum_{\nu=1}^{\infty}(F_\nu(x) - F_{\nu+1}(x))(1-p_1)^\nu}{(1-F_2(x)) + \sum_{\nu=2}^{\infty}(F_\nu(x) - F_{\nu+1}(x))\{(1-p_1)^\nu + \nu p_1(1-p_1)^{\nu-1}\}} \\
&\quad \le \frac{1-F(x)}{1-F_2(x)} + \frac{\sum_{\nu=1}^{L-1}(F_\nu(x) - F_{\nu+1}(x))(1-p_1)^\nu}{(F_L(x) - F_{L+1}(x))(1-p_1)^L} \\
&\qquad + \frac{\{\sum_{\nu=L}^{\infty}(F_\nu(x) - F_{\nu+1}(x))(1-p_1)^\nu\}}{L(p_1/(1-p_1))\{\sum_{\nu=L}^{\infty}(F_\nu(x) - F_{\nu+1}(x))(1-p_1)^\nu\}} \qquad (14) \\
&\quad \le \frac{1-F(x)}{1-F_2(x)} + \frac{\sum_{\nu=1}^{L-1}(F_\nu(x) - F_{\nu+1}(x))(1-p_1)^\nu}{(F_L(x) - F_{L+1}(x))(1-p_1)^L} + \frac{1-p_1}{Lp_1}.
\end{aligned}$$

First we choose L large to make the last term of (14) small. Then, for L fixed, we can choose x large to make the first two terms small under the force of (3). This assures $\lim_{x\to\infty}[(1 - G^{(1)}(x))/(1 - G_2^{(1)}(x))] = 0$. For any fixed constant $a > 0$, we exploit the renewal structure of the A process and write $G^{(1)}(x)$ $(x \ge 2a)$ as

$$\begin{aligned}
1 - G^{(1)}(x) &= \int_0^x \{1 - G^{(1)}(x-y)\}(1-p_1)\, dF(y) + (1 - F(x)) \\
&\ge \int_a^{2a} \{1 - G^{(1)}(x-y)\}(1-p_1)\, dF(y) \\
&\ge (1 - G^{(1)}(x-a))(1-p_1)(F(2a) - F(a)).
\end{aligned}$$

Therefore,

$$\frac{1-G^{(1)}(x-a)}{1-G^{(1)}(x)} \le \frac{1}{(1-p_1)(F(2a)-F(a))}.$$

It follows that

$$\begin{aligned}\lim_{x\to\infty}\frac{1-G^{(1)}(x-a)}{1-G_2^{(1)}(x)} &= \lim_{x\to\infty}\frac{1-G^{(1)}(x-a)}{1-G^{(1)}(x)}\,\frac{1-G^{(1)}(x)}{1-G_2^{(1)}(x)}\\ &\le \frac{1}{(1-p_1)(F(2a)-F(a))}\lim_{x\to\infty}\frac{1-G^{(1)}(x)}{1-G_2^{(1)}(x)}\\ &= 0.\end{aligned}$$

The proof of Lemma 4 is now complete. □

Also from (13) and Dembo and Karlin [(1992), Lemma 4.2], we have that for any integer $m \ge 1$,

$$\lim_{x\to\infty}\frac{1-G_m^{(1)}(x-a)}{1-G_{m+1}^{(1)}(x)} = 0. \tag{15}$$

The following lemma can be validated from the above lemma.

LEMMA 5. *Suppose* $p_1 < \min(p_2,\ldots,p_l)$ *and* F *satisfies the conditions* (3) *and* (4). *Then for* $2 \le k \le l$,

$$\lim_{t\to\infty}(N_t^{(k)}-r+1)(1-G_r^{(k)}(b_t)) = 0.$$

PROOF. Since $\lim_{t\to\infty}(N_t^{(k)}/N_t^{(1)}) = p_k/p_1$, we only need to prove

$$\lim_{t\to\infty}\frac{1-G_r^{(k)}(b_t)}{1-G_r^{(1)}(b_t)} = 0. \tag{16}$$

To this end, first we claim

$$\lim_{x\to\infty}\frac{1-G^{(k)}(x)}{1-G^{(1)}(x)} = 0.$$

A direct calculation from (5) yields

$$\begin{aligned}
&\frac{1-G^{(k)}(x)}{1-G^{(1)}(x)}\\
&\quad=\frac{(1-F(x))+\sum_{\nu=1}^{\infty}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_k)^\nu}{(1-F(x))+\sum_{\nu=1}^{\infty}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_1)^\nu}\\
&\quad\le\frac{(1-F(x))+\sum_{\nu=1}^{L-1}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_k)^\nu}{\{F_L(x)-F_{L+1}(x)\}(1-p_1)^L}\\
&\qquad+\frac{\sum_{\nu=L}^{\infty}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_k)^\nu}{\sum_{\nu=L}^{\infty}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_1)^\nu}\\
&\quad\le\frac{(1-F(x))+\sum_{\nu=1}^{L-1}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_k)^\nu}{\{F_L(x)-F_{L+1}(x)\}(1-p_1)^L}\\
&\qquad+\sum_{\nu=L}^{\infty}\frac{\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_k)^\nu}{\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_1)^\nu}\\
&\quad=\frac{(1-F(x))+\sum_{\nu=1}^{L-1}\{F_\nu(x)-F_{\nu+1}(x)\}(1-p_k)^\nu}{\{F_L(x)-F_{L+1}(x)\}(1-p_1)^L}+\left(\frac{1-p_k}{1-p_1}\right)^L\frac{1-p_1}{p_k-p_1}.
\end{aligned}$$

First we choose L large, and then x appropriately large to confirm the desired convergence.

Next we advance the induction from r to $r+1$ for (16). Consider

$$\begin{aligned}
&\frac{1-G_{r+1}^{(k)}(x)}{1-G_{r+1}^{(1)}(x)}\\
&\quad=\frac{1}{1-G_{r+1}^{(1)}(x)}\cdot\Bigg[1-G_r^{(k)}(x)+\int_0^{x-L}\left(1-G^{(k)}(x-\eta)\right)dG_r^{(k)}(\eta)\\
&\qquad\qquad+\int_{x-L}^{x}\left(1-G^{(k)}(x-\eta)\right)dG_r^{(k)}(\eta)\Bigg] \qquad (17)\\
&\quad\le\frac{1-G_r^{(k)}(x)}{1-G_r^{(1)}(x)}+\frac{\int_0^{x-L}\left(1-G^{(k)}(x-\eta)\right)dG_r^{(k)}(\eta)}{1-G_{r+1}^{(1)}(x)}\\
&\qquad+\frac{1-G_r^{(k)}(x-L)}{1-G_{r+1}^{(1)}(x)}.
\end{aligned}$$

The first term of (17) goes to 0 owing to the induction hypothesis.

The third term of (17) is estimated above by

$$\begin{aligned}\frac{1-G_r^{(k)}(x-L)}{1-G_{r+1}^{(1)}(x)} &= \frac{1-G_r^{(k)}(x-L)}{1-G_r^{(1)}(x-L)}\ \frac{1-G_r^{(1)}(x-L)}{1-G_{r+1}^{(1)}(x)} \\ &< \frac{1-G_r^{(1)}(x-L)}{1-G_{r+1}^{(1)}(x)}.\end{aligned} \tag{18}$$

The second term of (17) is estimated above by

$$\begin{aligned}&\frac{\int_0^{x-L}(1-G^{(k)}(x-\eta))dG_r^{(k)}(\eta)}{1-G_{r+1}^{(1)}(x)} \\ &\quad=\frac{\int_0^{x-L}\big[\big(1-G^{(k)}(x-\eta)\big)/\big(1-G^{(1)}(x-\eta)\big)\big][1-G^{(1)}(x-\eta)]dG_r^{(k)}(\eta)}{1-G_{r+1}^{(1)}(x)} \\ &\quad\le\sup_{\tilde\eta\ge L}\left[\frac{1-G^{(k)}(\tilde\eta)}{1-G^{(1)}(\tilde\eta)}\right]\frac{1-G^{(1)}*G_r^{(k)}(x)}{1-G_{r+1}^{(1)}(x)} \\ &\qquad\text{(here } * \text{ denotes the convolution operation)} \\ &\quad<\sup_{\tilde\eta\ge L}\left[\frac{1-G^{(k)}(\tilde\eta)}{1-G^{(1)}(\tilde\eta)}\right].\end{aligned} \tag{19}$$

Therefore, choose L large and then x sufficiently large after L is fixed to make all three terms of (17) small. The induction is established.

Since $b_t \to \infty$ as $t \to \infty$, the proof of Lemma 5 is now complete. □

Let $\tilde C_k^+(b_t)$ denote the number of the r-scan intervals in $(0,t)$ generated from the kth subprocess and with lengths $\ge b_t$. Lemma 5 also assures that for $k \ge 2$,

$$\begin{aligned}&\Pr\{\tilde C_k^+(b_t)\ne 0\} \\ &\le \Pr\left\{\tilde C_k^+(b_t)\ne 0;\left|\frac{n_t^{(k)}}{N_t^{(k)}}-1\right|\le\sqrt{\frac{\ln t}{t}}\right\}+\Pr\left\{\left|\frac{n_t^{(k)}}{N_t^{(k)}}-1\right|>\sqrt{\frac{\ln t}{t}}\right\} \\ &\le O\left(\frac{1-G_r^{(k)}(b_t)}{1-G_r^{(1)}(b_t)}\right)+O\left(\sqrt{\frac{1}{t}}\right) \\ &\to 0.\end{aligned}$$

From the above result we can prove the convergence result that

$$\lim_{t\to\infty} d\Big(\tilde C_1^+(b_t),(\tilde C_1^+(b_t)\,|\,\tilde C_k^+(b_t)=0,\ \forall\, k, 2\le k\le l)\Big)=0.$$

To this end, let

$$\begin{aligned}\tilde{\mathscr{E}}_t &= \{\tilde C_k^+(b_t)=0,\ \forall\, k, 2\le k\le l\}, \\ \tilde{\mathscr{E}}_t^c &= \text{complement of } \tilde{\mathscr{E}}_t.\end{aligned}$$

Then for $\mathscr{A}$ a set of nonnegative integers,

$$\begin{aligned}&\Pr\{\tilde{C}_1^+(b_t) \in \mathscr{A}\}\\ &\quad = \Pr\{\tilde{C}_1^+(b_t) \in \mathscr{A} \mid \tilde{\mathscr{E}_t}\}\Pr\{\tilde{\mathscr{E}_t}\} + \Pr\{\tilde{C}_1^+(b_t) \in \mathscr{A} \mid \tilde{\mathscr{E}_t^c}\}\Pr\{\tilde{\mathscr{E}_t^c}\},\end{aligned}$$

and thus

$$\begin{aligned}&\left|\Pr\{\tilde{C}_1^+(b_t) \in \mathscr{A}\} - \Pr\{\tilde{C}_1^+(b_t) \in \mathscr{A} | \tilde{\mathscr{E}_t}\}\right|\\ &\quad \le \Pr\{\tilde{C}_1^+(b_t) \in \mathscr{A} \mid \tilde{\mathscr{E}_t}\}\left|\Pr\{\tilde{\mathscr{E}_t}\} - 1\right| + \Pr\{\tilde{\mathscr{E}_t^c}\}\\ &\quad \le 2\ \Pr\{\tilde{\mathscr{E}_t^c}\}.\end{aligned}$$

Since the above bound does not depend on the choice of $\mathscr{A}$, it is true that

$$\begin{aligned}&d(\tilde{C}_1^+(b_t), (\tilde{C}_1^+(b_t) \mid \tilde{C}_k^+(b_t) = 0,\ \forall\, k, 2 \le k \le l))\\ &\quad \le 2\ \Pr\{\tilde{\mathscr{E}_t^c}\}\\ &\quad \le O\left(\max_{2\le k\le l}\left\{\frac{1 - G_r^{(k)}(b_t)}{1 - G_r^{(1)}(b_t)}\right\}\right) + O\left(\sqrt{\frac{1}{t}}\right)\\ &\quad \to 0 \quad \text{as } t \to \infty.\end{aligned}$$

Now we are ready to prove Theorem 2.

PROOF OF THEOREM 2. The Poisson approximation for $\tilde{C}_1^+(b_t)$ is assured by the Chen–Stein method as discussed in Dembo and Karlin (1992) when $G^{(1)}$ possesses the property of (13). Thus for $\tilde{C}^+(b_t) = \sum_{k=1}^{l} \tilde{C}_k^+(b_t)$, we can discuss the total variation distance $d(\tilde{C}^+(b_t), Z_\mu)$ by conditioning on the two events: $\tilde{\mathscr{E}_t} = \{\tilde{C}_k^+(b_t) = 0,\ \forall\, k,\ 2 \le k \le l\}$ and $\tilde{\mathscr{E}_t^c}$ = *the complement of* $\tilde{\mathscr{E}_t}$. The following calculation produces the desired result:

$$\begin{aligned}&d(\tilde{C}^+(b_t), Z_\mu)\\ &\quad \le d((\tilde{C}^+(b_t) \mid \tilde{\mathscr{E}_t}), Z_\mu)\Pr\{\tilde{\mathscr{E}_t}\} + d((\tilde{C}^+(b_t) | \tilde{\mathscr{E}_t^c}), Z_\mu)\Pr\{\tilde{\mathscr{E}_t^c}\}\\ &\quad \le d((\tilde{C}^+(b_t) \mid \tilde{C}_k^+(b_t) = 0,\ \forall\, k, 2 \le k \le l), Z_\mu) + \Pr\{\tilde{\mathscr{E}_t^c}\}\\ &\quad = d((\tilde{C}_1^+(b_t) \mid \tilde{C}_k^+(b_t) = 0, \forall\, k, 2 \le k \le l), Z_\mu) + \Pr\{\tilde{\mathscr{E}_t^c}\}\\ &\quad \le d(\tilde{C}_1^+(b_t), Z_\mu) + d(\tilde{C}_1^+(b_t), (\tilde{C}_1^+(b_t) \mid \tilde{C}_k^+(b_t) = 0,\ \forall\, k, 2 \le k \le l)) + \Pr\{\tilde{\mathscr{E}_t^c}\}\\ &\quad \le O\left(\sqrt{\frac{\ln t}{t}}\right) + (2r-1)O(1 - G_r^{(1)}(b_t)) + O\left(\sum_{j=1}^{r-1}\Pr\{\tilde{R}_{j+1} \ge b_t \mid \tilde{R}_1 \ge b_t\}\right)\\ &\qquad + O\left(\max_{2\le k\le l}\left\{\frac{1 - G_r^{(k)}(b_t)}{1 - G_r^{(1)}(b_t)}\right\}\right)\\ &\quad \to 0,\end{aligned}$$

where $\tilde{R}_i = \sum_{j=i}^{i+r-1} \tilde{X}_j$, for $\{\tilde{X}_j\} \sim G^{(1)}$ i.i.d. The proof of Theorem 2 is now complete. □

5. The asymptotic theorem for the combined Poisson model. Following the construction and notations of the combined Poisson model as formulated in Section 1, we first discuss the embedding into a multiple-array model described in the previous part of this paper. Let $\tilde{A}$ denote a Poisson process with parameter $\tilde{a} = \sum_{i=1}^{m} a_i$. Given a family of sets $\{Z_k\}_{k=1}^{l}$, there are m l-length column vectors $\underline{D}_1, \underline{D}_2, \ldots, \underline{D}_m$ which can be determined as follows. For $1 \le \eta \le m$, $\underline{D}_\eta = (d_{\eta,1}, d_{\eta,2}, \ldots, d_{\eta,l})$, where $d_{\eta,k}$ is 1 if $\eta \in Z_k$ and 0 otherwise. Generally, there could exist $\eta \ne \zeta$, such that $\underline{D}_\eta = \underline{D}_\zeta$. By rearranging the order of $\{\underline{D}_\eta\}_{\eta=1}^{m}$, let $\{\underline{D}_\eta\}_{\eta=1}^{w}$, $w \le m$, denote the largest subset of $\{\underline{D}_\eta\}_{\eta=1}^{m}$ which has different representations for each column vector $\underline{D}_\eta$. We then define a discrete probability density $P_{\underline{D}}$ on $\{0,1\}^l$ such that for $1 \le \eta \le w$,

$$P_{\underline{D}}(\underline{D}_\eta) = \sum_{\substack{1 \le \zeta \le m \\ \underline{D}_\zeta = \underline{D}_\eta}} \frac{a_\zeta}{\tilde{a}}. \tag{20}$$

From $\tilde{A}$, $\{\underline{D}_\eta\}_{\eta=1}^{w}$ and $P_{\underline{D}}$, we can construct l dependent arrays following the rule of the l-array model. To clarify the construction, we introduce an example first. Suppose there are three independent Poisson processes, A_1, A_2 and A_3, with parameters a_1, a_2 and a_3, respectively. We combine A_1 and A_2 yielding the B_1 process, and combine A_1 and A_3 yielding the B_2 process. The $\tilde{A}$ process for this problem is a Poisson process with parameter $a_1 + a_2 + a_3$. The sets Z_1, Z_2 are

$$Z_1 = \{1,2\}, \qquad Z_2 = \{1,3\},$$

and the vectors $\{\underline{D}_\eta\}$ as well as the probability density of $\underline{D}$ are

$$\underline{D}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \underline{D}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \underline{D}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

such that

$$P_{\underline{D}}(\underline{D}_1) = \frac{a_1}{\tilde{a}}, \qquad P_{\underline{D}}(\underline{D}_2) = \frac{a_2}{\tilde{a}}, \qquad P_{\underline{D}}(\underline{D}_3) = \frac{a_3}{\tilde{a}}, \qquad \tilde{a} = a_1 + a_2 + a_3.$$

The equivalence between the newly constructed l arrays and the original combined Poisson arrays can be assured by the classical coloring theorem of Poisson processes given in the following.

LEMMA 6 (Coloring theorem of the Poisson process). *Assume the existence of a homogeneous Poisson(θ) process Π in $(0,\infty)$. Given $(p_1, p_2, \ldots, p_m)$, $p_i > 0$ and $p_1 + p_2 + \cdots + p_m = 1$, we generate m point processes, $\Pi_1, \Pi_2, \ldots, \Pi_m$, from Π in such a way that p_i is the probability for each Π point to occur in the Π_i process but be absent from the other point processes $\{\Pi_j\}_{j \ne i}$. The determinations are independent with respect to each Π point. Then, the point*

processes $\{\Pi_i\}_{i=1}^m$ *are independent Poisson processes with the parameters* $\theta p_1, \theta p_2, \ldots, \theta p_m$, *respectively.*

Thus the independent Poisson processes $A_1, A_2, \ldots, A_m$ are induced from the Poisson process $\tilde{A}$ by fixing $p_i = a_i/\tilde{a}$, $1 \le i \le m$, and $\{\underline{D}_i\}_{i=1}^m$ and $P_{\underline{D}}$ are assured as before. Then the asymptotic distributions for the largest and the smallest r-scan lengths across the l induced Poisson arrays can be ascertained by the results from the multiple-array model, Theorems 1 and 2, leading to the following result.

THEOREM 3. *Let* $B_k = \cup_{i \in Z_k} A_i$, $1 \le i \le l$, *be dependent Poisson processes as described before and let* $P_{\underline{D}}$ *be the probability law on* $\{0,1\}^l$ *defined as in* (20). *For* $b_k = \sum_{i \in Z_k} a_i$, $1 \le k \le l$, *assume* $b_1 < \min_{2 \le k \le l}\{b_k\}$. *Then the asymptotic distributions for the smallest and the largest r-scan lengths,* m_t *and* M_t, *across the l dependent Poisson arrays in* $(0,t)$ *are as follows:*

$$\lim_{t\to\infty} \Pr\left\{m_t > \sqrt[r]{\frac{x}{t}}\right\} = \exp\left\{-\frac{(\sum_{i=1}^m a_i)^{r+1}x}{r!} \sum_{(\underline{d}_1, \underline{d}_2, \ldots, \underline{d}_{r+1}) \in Q_r} \left(\prod_{i=1}^{r+1} P_{\underline{D}}(\underline{d}_i)\right)\right\}, \tag{21}$$

$$\lim_{t\to\infty} \Pr\left\{M_t < \frac{\ln t + (r-1)\ln\ln t + x}{\sum_{i \in Z_1} a_i}\right\} = \exp\left\{-e^{-x}\frac{(\sum_{i \in Z_1} a_i)}{(r-1)!}\right\}, \tag{22}$$

where

$\underline{d}_i$ = *l-length vector with values in* $\{0,1\}$,

Q_r = *set of* $l \times (r+1)$ *matrices with elements taking values in* $\{0,1\}$ *such that each matrix of* Q_r *contains at least one row with all components 1.*

The r-scan analysis of the concrete example related to evolutionary phylogeny mentioned in Section 1 ensues from the above theorem. In this case, the corresponding probability density $P_{\underline{D}}$ on $\{0,1\}^2$ is

$$\begin{aligned}
P_{\underline{D}}\binom{1}{1} &= \frac{\alpha}{\alpha+\beta+\gamma} p_1 p_2 = z_1,\\
P_{\underline{D}}\binom{1}{0} &= \frac{\beta}{\alpha+\beta+\gamma} + \frac{\alpha}{\alpha+\beta+\gamma} p_1 q_2 = z_2,\\
P_{\underline{D}}\binom{0}{1} &= \frac{\gamma}{\alpha+\beta+\gamma} + \frac{\alpha}{\alpha+\beta+\gamma} q_1 p_2 = z_3,\\
P_{\underline{D}}\binom{0}{0} &= \frac{\alpha}{\alpha+\beta+\gamma} q_1 q_2 = z_4
\end{aligned}$$

as stated in Section 1.

Let m_t be the minimal r-scan length across the r-scans of Processes B_1 and C_1 in $(0, t)$. Then according to (21), we have

$$\begin{aligned}\lim_{t\to\infty} \Pr\left\{m_t > \sqrt[r]{\frac{x}{t}}\right\} &\\ = \exp\left\{-\left[(z_1+z_2+z_3)^{r+1} - \sum_{\substack{i+j+k=r+1\\ j>0,\,k>0}} \frac{(r+1)!}{i!\,j!k!} z_1^i z_2^j z_3^k\right]\right.&\\ \left.\times \frac{x(\alpha+\beta+\gamma)^{r+1}}{r!}\right\}.&\end{aligned} \tag{23}$$

Also the asymptotic distribution for the smallest r-scan length of the ancestor Poisson process A can be determined as satisfies

$$\lim_{t\to\infty} \Pr\left\{s_t > \sqrt[r]{\frac{x}{t}}\right\} = \exp\left\{-\frac{x\alpha^{r+1}}{r!}\right\}. \tag{24}$$

Thus from (23) and (24), a simple comparison shows that if

$$(z_1+z_2+z_3)^{r+1} - \sum_{\substack{i+j+k=r+1\\ j>0,\,k>0}} \frac{(r+1)!}{i!\,j!k!} z_1^i z_2^j z_3^k > \frac{\alpha^{r+1}}{(\alpha+\beta+\gamma)^{r+1}}, \tag{25}$$

then asymptotically ($t \to \infty$)

$$\Pr\left\{m_t > \sqrt[r]{\frac{x}{t}}\right\} < \Pr\left\{s_t > \sqrt[r]{\frac{x}{t}}\right\} \quad \text{for all positive } x.$$

This means that m_t is stochastically smaller than s_t as $t \to \infty$.

For $r = 1$, a sufficient condition for (25) is

$$(\alpha p_1 p_2 + \beta + \gamma)^2 > \alpha^2 + 2\beta\gamma.$$

For the maximal r-scan length, we need only consider the Poisson process B_1 or C_1, whichever has the smaller parameter as discussed in Section 4.

REFERENCES

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

BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson Approximation*. Oxford Scientific Publications.

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

DEMBO, A. and KARLIN, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Probab.* **2** 329–357.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

GERSTEIN, M. (1997). A structure census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Molecular Biology* **274** 562–576.

KARLIN, S. and BRENDEL, V. (1992). Chance and statistical significance in protein and DNA sequence analysis. *Science* **257** 39–49.

KARLIN, S. and CARDON, L. R. (1994). Computational DNA sequence analysis. *Ann. Rev. Microbiology* **48** 619–654.

KARLIN, S. and MACKEN, C. (1991). Some statistical problems in the assessment of inhomogeneities of DNA sequence data. *J. Amer. Statist. Assoc.* **86** 26–33.

KARLIN, S., MRÁZEK, J. and CAMPBELL, A. (1996). Frequent oligonucleotides and peptides of the Haemophilus influenzae genome. *Nucleic Acids Research* **24** 4263–4272.

KARLIN, S. and TAYLOR, H. M. (1981). *A Second Course in Stochastic Processes*, 2nd ed. Academic Press, New York.

MASSE, M. J. O., KARLIN, S., SCHACHTEL, A. and MOCARSKI, E. S. (1992). Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. *Proc. Nat. Acad. Sci. U.S.A.* **89** 5246–5250.

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

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

REINERT, G. and SCHBATH, S. (1998). Compound Poisson and Poisson approximations for occurrences of multiple words in Markov chains. *J. Comput. Biology* **5** 223–253.

STEIN, C. (1986). *Approximation Computation of Expectations*. IMS, Hayward, CA.

DEPARTMENT OF MATHEMATICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-2125
E-MAIL: fd.zgg@forsythe.stanford.edu