A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 287–290

r-scan extremal statistics of inhomogeneous Poisson processes

Samuel Karlin[1] **and Chingfer Chen**[1]

Stanford University

Abstract: Studies of inhomogeneities in long DNA sequences can be insightful to the organization of the human genome (or any genome). Questions about the spacings of a marker array and general issues of sequence heterogeneity in our studies of DNA and protein sequences led us to statistical considerations of r-scan lengths, the distances between marker i and marker $i+r$, $i = 1, 2, 3, \ldots$. It is interesting to characterize the r-scan lengths harboring clusters or indicating regions of over-dispersion of the markers along the sequence. Applications are reviewed for certain words in the *Haemophilus* genome and the *Cyanobacter* genome.

1. Introduction

We are happy to contribute this paper to the festscript volume in honor of Dr. H. Rubin. The paper is of pratical and theoretical application. I also had the pleasure to develop with Herman an extended analysis concerning a family of distributions in possession of a monotone likelihood ratio ([1, 2]).

Question about spacings of a marker array and general issues of sequence heterogeneity in our studies of DNA and protein sequences led us to statistical considerations of r-scan lengths, the distances between marker i and marker $i + r$, $i = 1, 2, 3, \ldots$. It is interesting to charaterize the r-scan lengths harboring clusters or indicating regions of over-dispersion of the markers along the sequence. Concretely, a typical objective is to determine the probability of successive $\{r + 1\}$ markers falling within a DNA sequence stretch under an appropriate stochastic model of the marker array. There are similar issues pertaining to sparseness of markers. Particular markers (in the language of DNA, e.g., specific restriction sites, nucleosome placements, locations of genes) are distributed over the genome along chromosomes. The r-scan analysis has been largely applied to the homogeneous Poisson processes for a marker array distributed over a long contig. It is known that the organization of mammalian genomes shows substantial inhomogeneities, including "isochores", regions dominated by either C + G or A + T DNA base content.

Here we consider an inhomogeneous Poisson process Π on the real axis $(0, \infty)$ with an intensity $\lambda(s)$, $0 \leq s < \infty$. The intensity function $\lambda(s)$ can be of different types, for example, periodic or constant in successive intervals, depending on different applications. In this context, we would like to determine the asymptotic distribution of the kth minimum among the r-scan lengths over the interval horizon $(0, t)$, as $t \to \infty$.

2. Preliminaries. Minimal r-scan lengths from a general distribution

In the paper [3], the asymptotic distribution of the kth minimum r-scan length from a general distribution function has been studied by applying the Chen-Stein

[1]Department of Mathematics, Stanford University, Stanford, CA 94305-2125, USA. e-mail: karlin@math.stanford.edu

Keywords and phrases: r-scan statistics, inhomogeneous Poisson marker array, asymptotic distributions.

AMS 2000 subject classifications: 92B05, 92D20.

method [4]. In that context, an r-scan process is generated following a piecewise constant function or continuous general density $f(x)$ with bounded support $(0, T]$. Thus let $V_1, V_2, \ldots, V_{n-1}$ be $n-1$ *i.i.d.* samples drawn from the density $f(x)$, and let $V_1^* \le V_2^* \le \cdots \le V_{n-1}^*$ be the order statistics corresponding to $\{V_i\}$. For convenience, let $V_0^* = 0$ and $V_n^* = T$. Then the associated r-scan fragments $R_i = V_{i+r-1}^* - V_{i-1}^*$, $i = 1, \ldots, n-r+1$, and their order statistics R_i^* are defined in the usual way such that $R_1^* \le R_2^* \le \cdots \le R_{n-r+1}^*$. For an extensive review of r-scan statistics, see the book [5].

From $\{R_i\}$, we define the Bernoulli random variables

$$\begin{aligned} U_i^-(a) &= 1, \quad \text{if } R_i \le a \\ &= 0, \quad \text{if } R_i > a \end{aligned}$$

and their sum

$$N_{n-r+1}^-(a) = \sum_{i=1}^{n-r+1} U_i^-(a).$$

Denote by $m_{n,k} = R_k^*$. The asymptotic distribution (as $n \to \infty$) for $m_{n,k}$ is as follows.

Lemma 1. *For a given positive constant μ, let a_n be determined to satisfy*

$$\mu = \frac{(na_n)^r}{r!} n \int [f(x)]^{r+1}\, dx. \tag{1}$$

Then we have the Poisson approximation

$$\lim_{n\to\infty} d\big(N_{n-r+1}^-(a_n), Po(\mu)\big) = 0,$$

for $Po(\mu)$ the Poisson distribution with parameter μ. Here $d(\ ,\ )$ is the total variational distance between two random variables defined by

$$d(U, V) = \sup_A \big[Pr\{U \in A\} - Pr\{V \in A\}\big].$$

Moreover, the kth minimal r-scan length, $m_{n,k}$, possesses the asymptotic distribution

$$\lim_{n\to\infty} \Pr\{m_{n,k} > a_n\} = \sum_{i=0}^{k-1} e^{-\mu} \frac{\mu^i}{i!}. \tag{2}$$

Proof of the above lemma is given in [3], Section 8. Here, by adapting the foregoing result, we will determine the asymptotic distribution of the kth minimal r-scan length corresponding to an inhomogeneous Poisson process in $(0, t)$, as $t \to \infty$.

3. Minimal r-scan lengths for an inhomogeneous Poisson process

The asymptotic theorem for the minimal r-scan length will be derived from the distributional property of $\widetilde{N}_t^-(a)$, where $\widetilde{N}_t^-(a)$ is the number of r-scan segments of lengths $\le a$ over the interval horizon $(0, t)$. It is clear that if $\widetilde{N}_t^-(a) < k$, the kth minimal r-scan length $m_{t,k}$ exceeds the level a. Thus if the Poisson approximation holds for $\widetilde{N}_t^-(a)$, we can access the asymptotic law for $m_{t,k}$. Here the r-scan process of interest is generated from an inhomogeneous Poisson process Π with an intensity function $\lambda(s)$, $0 \le s < \infty$. The main theorem is as follows.

Theorem 1. *Assume $\lambda(s)$ defined for $s \geq 0$ satisfies*

$$\int_0^t \lambda(s)\,ds \to \infty, \quad \text{as } t \to \infty. \tag{3}$$

For a given positive constant μ, let a_t be determined to satisfy the equation

$$\mu = \frac{a_t^r}{r!} \int_0^t \lambda^{r+1}(s)\,ds. \tag{4}$$

Then we have the Poisson approximation

$$\lim_{t\to\infty} d\big(\widetilde{N}_t^-(a_t), Po(\mu)\big) = 0.$$

Moreover, the kth minimal r-scan length, $m_{t,k}$, possesses the asymptotic distribution

$$\lim_{t\to\infty} \Pr\{m_{t,k} > a_t\} = \sum_{i=0}^{k-1} e^{-\mu} \frac{\mu^i}{i!}.$$

Proof of Theorem 1. If n_t denotes the point count of the Poisson Process Π in (0, t), then

$$E[n_t] = \int_0^t \lambda(s)ds, \qquad Var(n_t) = \int_0^t \lambda(s)\,ds.$$

For convenience, let $\bar{n}_t = \lfloor E[n_t] \rfloor$. Thus the Berry–Esseen estimate assures

$$\Pr\left\{ \left| \frac{n_t}{\bar{n}_t} - 1 \right| > \sqrt{\frac{\ln \bar{n}_t}{\bar{n}_t}} \right\} = O\left(\frac{1}{\sqrt{\bar{n}_t}} \right).$$

Therefore

$$d\big(\widetilde{N}_t^-(a_t), Po(\mu)\big) \leq d\big(N_{\bar{n}_t - r+1}^-(a_t), Po(\mu)\big) + O\left(\sqrt{\frac{\ln \bar{n}_t}{\bar{n}_t}} \right).$$

If $n_t = \bar{n}_t$, the $\bar{n}_t$ points in $(0, t)$ are distributed independently according to $g(x)$, with

$$g(x) = \frac{\lambda(x)}{\int_0^t \lambda(x)dx}, \quad 0 \leq x \leq t. \tag{5}$$

Following the result of Lemma 1, we have

$$\lim_{n\to\infty} d\big(N_{n-r+1}^-(a_n), Po(\mu)\big) = 0$$

for

$$a_n = \sqrt[r]{\frac{r!\mu}{n^{r+1} \int_0^t g^{r+1}(x)\,dx}}. \tag{6}$$

Since

$$\bar{n}_t = \int_0^t \lambda(s)\,ds \to \infty,$$

we replace n with $\bar{n}_t$ in formula (6) and $g(x)$ with $\frac{\lambda(x)}{\int_0^t \lambda(x)dx}$ to verify equation (4). On this bases, we obtain

$$\lim_{t\to\infty} d\big(\widetilde{N}_t^-(a_t), Po(\mu)\big) = 0,$$

with

$$a_t = \sqrt[r]{\frac{r!\mu}{\int_0^t \lambda^{r+1}(s)\,ds}}.$$

This completes the proof of Theorem 1.

4. Examples

Haemophilus influenza is a bacterium which engenders an infection in the lungs of humans [6]. The study of the USSs (uptake signal sequences) AAGTGCGGT (USS+) and its inverted complement (USS-) in the *H. influenza* genome (length of 1.83 Mb, Rd strain) provides opportunities for characterizing global genomic inhomogeneities. The result of homogeneous r-scan tests for $r = 1, 2, \ldots, 6$ shows a significant even spacings between the markers such that the USS+ and USS- are remarkably evenly spaced around the genome such that both USS+ positions and USS- positions have respective minimum spacings higher than expected by chance with probability 0.001. This rare possibility may suggest that the homogeniety assumption doesn't fit the real distributions of the markers and an inhomogeneous r-san test should be applied for the marker array.

Another example is the distribution of the palindrome GGCGATCGCC labeled HIP1 (highly iterated palindrome) in the genome of the organism *Synechocystis* (3.6 Mb). *Synechocystis* is thought to be the evolutionary precursor of vascular plant plastids [7]. The photosynthetic endosymbiont became dependent on host genetic information for maintenance and evolved into an organelle specialized for CO_2 fixation. The r-scan analysis of the genome shows in this case a significantly even distribution. The observed minimal *1*-scan spacing is 52 bp (base pair) which is much larger than the threshold of 9 bp with the probability of 0.001. Similar conclusions apply to the r-scan tests of $r = 2, ..., 6$. The even spacing of HIP1 in *Synechocystis* is more dramatic than the situation of USSs in *H. influenza*.

References

[1] Karlin, S. and Rubin, H. (1956). The theory of decision procedures for distributions with monotone likelihood ratio. *Annals of Mathematical Statistics.* **27** 272–299. MR81593

[2] Karlin, S. and Rubin, H. (1956). Distributions possessing a monotone likelihood ratio. *Journal of American Statistical Association.* **51** 637–643. MR104303

[3] Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes. *Ann. Appl. Probab.* **2** 329–357. MR1161058

[4] Arratia, R., Goldstein, L., and Gordon, L. (1990). Poisson approximation and the Chen–Stein method. *Statist. Sci.* **5** 403–434. MR1092983

[5] Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics.* Springer Verlag. MR1869112

[6] Karlin, S., Mrázek, J. and Campbell, A. (1996). Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. *Nucleic Acids Research.* **24** No. 21, 4263–4272.

[7] Mrázek, J., Bhaya, D., Grossman, A. R. and Karlin, S. (2001). Highly expressed and alien genes of the *Synechocystis* genome. *Nucleic Acids Research.* **29**, No. 7, 1590–1601.