

Scan Statistics With Weighted Observations

Hock Peng CHAN and Nancy Ruonan ZHANG

We examine scan statistics for one-dimensional marked Poisson processes. Such statistics tabulate the maximum weighted count of event occurrences within a window of predetermined width over all windows within an observed interval. We derive analytical formulas and also give an importance sampling method for approximating the tail probabilities of scan statistics. Because high-throughput genomic sequencing has led to the availability of massive amounts of biomolecular sequence data, it is often of interest to search long DNA or protein sequences for local regions that are enriched for a certain characteristic. Thus scan statistics have become a useful tool in modern computational biology. We illustrate the application of our p value approximations with such examples.

KEY WORDS: Change of measure; DNA sequence; Importance sampling; Large deviation; Marked Poisson process; Scan statistics.

1. INTRODUCTION

Scan statistics are used for detecting the presence of an unusually large cluster of events ordered by either time or location. A scanning window is moved along the time (or location) interval, and the maximum number of events captured by this window at some point in time is recorded. Due to the multiple comparisons effect from maximizing over all possible windows, most of which are nonoverlapping, scan statistics are often larger than expected from mere intuition when in fact no clustering is present. Glaz, Naus, and Wallenstein (2001, pp. 29–31) illustrated this point with interesting examples and motivated the need for precise p value computations for scan statistics. They documented the development of scan statistic tail approximations and bounds over the past 40 years and provided detailed derivations for these estimates. (For more detailed discussions, see Naus 1965, 1982; Cressie 1980; Glaz 1989; Glaz and Naus 1991; Loader 1991.)

One particular area in which scan statistics have been useful in recent years is the computational analysis of DNA and protein sequences. For example, Lifanov, Makeev, Nazina, and Papatsenko (2003) scanned DNA sequences for clusters of transcription factor-binding sites to locate genes that relate to specific biological processes. They used position weight matrices to score words for similarity to a given transcription factor pattern and used a cutoff value for the word score to determine locations of pattern occurrence. Rajewsky, Vergassola, Gaul, and Siggia (2002) considered a similar problem except that they used the total score of all words in a window exceeding the cutoff instead of the number of words exceeding the cutoff to compute the scan statistics. Currently available p value approximations for scan statistics of point processes treat only the case of 0–1 processes that are applicable to the approach of Lifanov et al. (2003) but cannot directly handle a scoring or weighting scheme used proposed by Rajewsky et al. (2002).

In Section 2 we provide p value approximations for scan statistics of marked Poisson processes. Our approximations are applicable to general scoring schemes used in computational biology. A novel feature of our formula is an overshoot correction term that disappears in the special case of 0–1 processes. In Section 3 we apply our p value approximations to various problems in computational biology. In Section 4 we use Monte

Carlo methods to check the analytical p value approximations. Because some of the p values considered are very small, which renders direct Monte Carlo inefficient, we propose an importance sampling scheme to provide more accurate estimates. We end the article in Section 5 with a discussion on computational issues and choice of scoring functions.

2. THEORETICAL RESULTS

Let N be a Poisson process on $(0, n]$ with constant rate $\lambda > 0$. Let X_1, X_2, \dots be independent and identically distributed (iid) random variables with cumulative distribution function F and independent of N such that $\mu := EX_1 > 0$. We say that F is arithmetic if the support of F lies on $\pm\eta, \pm 2\eta, \dots$ for some $\eta > 0$. The largest η with this property is known as the span of F (see Feller 1971, sec. 5.2). Let $S_k = X_1 + \dots + X_k$. Thus for any window $(t, t + u]$, $S_{N(t+u)} - S_{N(t)}$ is the weighted count of the point process in that window. We define

$$M_{n,u} = \sup_{0 \leq t \leq n-u} [S_{N(t+u)} - S_{N(t)}], \quad (1)$$

where $u \in (0, n)$ is a predetermined width of the sliding window. The widely studied scan statistic

$$\sup_{0 \leq t \leq n-u} [N(t+u) - N(t)] \quad (2)$$

is a special case of (1) when F is degenerate at 1. In (1), the occurrence of the i th jump in N is weighted by a score X_i , and hence we call $M_{n,u}$ a weighted scan statistic. Similarly, we call (2) an unweighted scan statistic.

In Theorem 1, we give an asymptotic approximation for the tail probability of the weighted scan statistic. Before stating the theorem, we need to define some constants. Assume that the moment-generating function of F , say $K(\theta) = E(e^{\theta X_1})$, is finite for some $\theta > 0$. Let $K'(\theta) = \frac{d}{d\theta} K(\theta) = E(X_1 e^{\theta X_1})$ and $K''(\theta) = \frac{d^2}{d\theta^2} K(\theta) = E(X_1^2 e^{\theta X_1})$. Given $x > \lambda\mu$, let $\theta_x > 0$ and $\alpha_x > \lambda$ be the unique constants that satisfy

$$K'(\theta_x) = (x/\lambda) \quad \text{and} \quad \alpha_x = \lambda K(\theta_x). \quad (3)$$

We also define the large deviation rate function,

$$I(x) = -(\alpha_x - \lambda) + \theta_x x. \quad (4)$$

To make the notation simple, we assume here that F is either continuous with density f or discrete with probability mass function f . Embed F in an exponential family of distribution

Hock Peng Chan is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, 119260 (E-mail: stachp@nus.edu.sg). Nancy Ruonan Zhang is Assistant Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: nzhang@stanford.edu). The authors thank an associate editor and two referees for their useful comments, suggestions, and references, which have led to a substantial improvement in the presentation of this article.

functions $\{F_\theta\}$, where F_θ has a density or probability mass function f_θ satisfying

$$f_\theta(y) = e^{\theta y} f(y) / K(\theta) \quad \text{whenever } K(\theta) < \infty. \quad (5)$$

Let Y_1, Y_2, \dots be iid random variables with the mixture density or probability mass function

$$g(y) = \left(\frac{\alpha_x}{\lambda + \alpha_x} \right) f_{\theta_x}(y) + \left(\frac{\lambda}{\lambda + \alpha_x} \right) f(-y), \quad (6)$$

and let $R_k = Y_1 + \dots + Y_k$. We define the overshoot constant

$$v_x = \lim_{b \rightarrow \infty} E[e^{-\theta_x(R_{\tau_b} - b)}], \quad \text{where } \tau_b = \inf\{k \geq 1 : R_k \geq b\}, \quad (7)$$

with the convention that b in (7) is a multiple of η if F is arithmetic with span η . (For more details on the overshoot constant, such as its existence and closed-form expressions, see Siegmund 1985, chap. 8; Tu and Siegmund 1999; Storey and Siegmund 2001.)

For any $a \in \mathbf{R}$, we let $\lfloor a \rfloor$ denote the greatest integer not exceeding a . For two sequences a_n and b_n , by $a_n \sim b_n$, we mean that $\lim_{n \rightarrow \infty} (a_n/b_n) = 1$.

Theorem 1. Let $v > \lambda\mu$ be fixed. If F is arithmetic with span $\eta > 0$, then let $x (= x_u) = \eta u^{-1} \lfloor uv/\eta \rfloor$. Thus ux is the largest multiple of η not exceeding uv . For F that is nonarithmetic, let $x = v$. Let $u \rightarrow \infty$ as $n \rightarrow \infty$ such that $(n - u) \rightarrow \infty$. Then

$$P\{M_{n,u} \geq ux\} \sim 1 - \exp\left\{-\frac{(n-u)v_x e^{-u(x-\lambda\mu)}}{\sqrt{2\pi u \lambda K''(\theta_x)}}\right\}. \quad (8)$$

It follows from (3), (6), and (7) that for the unweighted scan statistic in which F is degenerate at 1,

$$K(\theta_x) = K'(\theta_x) = K''(\theta_x) = e^{\theta_x}, \quad \alpha_x = x, \quad \theta_x = \log(x/\lambda), \quad \text{and } v_x = 1. \quad (9)$$

Thus Theorem 1 reduces to the following for this special case.

Corollary 1. Let $v > \lambda$ be fixed and define $x = u^{-1} \lfloor uv \rfloor$. Let $u \rightarrow \infty$ as $n \rightarrow \infty$ such that $(n - u) \rightarrow \infty$. Then

$$P\left\{\sup_{0 \leq t \leq n-u} [N(t+u) - N(t)] \geq ux\right\} \sim 1 - \exp\left\{-\frac{(n-u)e^{u(x-\lambda)}(\lambda/x)^{ux}(x-\lambda)}{\sqrt{2\pi ux}}\right\}. \quad (10)$$

Remark 1. If $u \rightarrow \infty$ and $n/u \rightarrow \infty$ as $n \rightarrow \infty$, then the relations (8) and (10) still hold if λ is replaced by the consistent estimator $\hat{\lambda} = N(n)/n$. For problems in computational biology, n often represents the length of a genome. Because genomes are very long, and the purpose of using scan statistics is to target relatively short segments of the genome of length u with “unusual” characteristics for further biological analysis, the required assumption is satisfied in practice. This allows us to estimate the Poisson rate parameter λ from the data.

Remark 2. Frolov (2005) considered the class of stochastically continuous processes with independent increments for which the marked Poisson process that we consider here is a special case. Bounds on the tail probabilities of $M_{n,u}$ were obtained and applied, using Borel–Cantelli lemmas, to derive almost-sure limits of $M_{n,u}$ as $u \rightarrow \infty$. The tail approximations in (8) are much sharper than those of Frolov (2005), however.

3. EXAMPLES

Before getting into the examples, we first describe the computation of the weighted scan statistic. Let t_i denote the time of occurrence of the i th event of interest and let X_i denote the score associated with this event. We construct the counting process $N(t) = \sum_i \mathbb{1}_{\{t_i \leq t\}}$, where $\mathbb{1}$ denotes the indicator function. Then, for a given window-size u , the weighted scan statistic is

$$M_{n,u} = \sup_{0 \leq t \leq n-u} \left(\sum_{i:t < t_i \leq t+u} X_i \right). \quad (11)$$

For computational purposes, it suffices to evaluate the sum on the right side of (11) at $t = t_i$ and $t = t_i - u$ for all i .

We next describe computation of the p value approximations in (8). If λ is unknown, then we replace it by $\hat{\lambda} = N(n)/n$. Unknown parameters of F (e.g., μ) also can be estimated from the data. We elaborate on this in Example 2. Because K' is an increasing function, θ_x can be computed through bijection methods using (3) if it cannot be obtained analytically. The constant $K''(\theta_x)$ and large deviation rate $I(x)$ can then be computed using (3) and (4).

It remains for us to compute the overshoot constant v_x . Let $\tau_+ = \inf\{k \geq 1 : R_k > 0\}$. By renewal theory (see, e.g., Siegmund 1985, cor. 8.33) it follows that if F is nonarithmetic, then

$$\lim_{b \rightarrow \infty} P\{R_{\tau_b} - b > y\} = (ER_{\tau_+})^{-1} \int_y^\infty P\{R_{\tau_+} > z\} dz \quad \text{for } y \geq 0, \quad (12)$$

and if F is arithmetic with span η , then

$$\lim_{b \rightarrow \infty, b \in \eta\mathbf{Z}} P\{R_{\tau_b} - b = j\eta\} = \eta(ER_{\tau_+})^{-1} P\{R_{\tau_+} \geq (j+1)\eta\} \quad \text{for } j = 0, 1, \dots \quad (13)$$

In Example 2 we consider F geometric with mean μ . By (5), F_{θ_x} is also geometric. Thus, by (6) and the memoryless property of the geometric distribution, R_{τ_+} is geometric with distribution F_{θ_x} . We can then use (7) and (13) to show that

$$v_x = \mu[1 - (1 - \mu^{-1})e^{\theta_x}]. \quad (14)$$

In Examples 1 and 3, we check Corollary 1 against p value approximations given by Naus (1982) for the unweighted scan statistics. For the weighted scan statistics in Example 2, we rely on importance sampling to demonstrate the accuracy of (8).

Example 1. Biologists are interested in finding segments of the genome that contain a high concentration of palindromic patterns (PLPs), because these segments are likely to be near an origin of replication of the virus (cf. Masse, Karlin, Schachtel, and Mocarski 1992). The DNA alphabet has four letters, A, T, C, and G with A–T and C–G complementary pairs on opposite strands of the DNA helix. Thus the complementary DNA sequence of GGATCC would be CCTAGG. The DNA sequence GGATCC is a PLP because its complement reads the same as itself backward. We define the length of a palindrome to be the number of complementary pairs that it contains; for example, the length of GGATCC is 3.

Denote by PLP* a PLP with length of at least 5 bp that is not nested inside another PLP. Leung, Schachtel, and Yu (1994),

modeled the occurrence of PLP* in the human cytomegalovirus (HCMV) genome as a Poisson process. The genome contains $n = 229,354$ base pairs, and a total of $N(n) = 296$ PLP* were observed. Thus rate of the Poisson process was estimated by $\hat{\lambda} = N(n)/n = 296/229,354 = .00129$. The unweighted scan statistic for window size $u = 1,000$ bp was computed and found to be equal to 10. The corresponding p value of .00195 computed by Corollary 1 agrees quite well with the estimate of .00193 obtained using the method Naus (1982) and the Monte Carlo estimate of .0021 obtained in the next section (see Table 3).

Example 2. We continue with the problem of finding clusters of PLPs in viral genomes. Instead of giving equal weigh to each PLP*, we now assign a score of $X_i = \ell_i - 4$, where ℓ_i is the length of the i th PLP*. By definition, palindromes always have even lengths. We define the location t_i of the i th PLP* to be the location of its left center. We then compute the weighted scan statistics $M_{n,u}$ with window length u equal to .5% of the genome length, rounded off to the nearest 100 bases. We apply (8) with estimated Poisson rate $\hat{\lambda} = N(n)/n$ and geometric F with estimated mean $\hat{\mu} = (1 - 2\hat{a}_A\hat{a}_T - 2\hat{a}_C\hat{a}_G)^{-1}$, where $(\hat{a}_A, \hat{a}_C, \hat{a}_G, \hat{a}_T)$ are the empirical probabilities of the four bases in the genome.

Chew, Choi, and Leung (2005) also studied clustering of PLP* but used the score $X_i = \ell_i$ (or, equivalently, $X_i = \ell_i/5$), corresponding to a shifted geometric distribution for X_i . They

did not provide any p value approximations and instead of (1), they considered a scan statistic in which consecutive windows differ by half the window length or, more specifically, $\sup_{k \in \mathbf{Z}, 0 \leq k \leq (2n/u)-1} [S_{N(uk/2+u)} - S_{N(uk/2)}]$.

Figure 1 plots the unweighted and weighted scan statistics against genome location for three viruses: cercopithecine herpesvirus 1 (CeHV1), bovine herpesvirus 1 (BoHV1), and bovine herpesvirus 5 (BoHV5). The figure also shows origins of replication for these viruses that have been validated experimentally. To avoid an excessive number of false positives when working with a large number of genomes, we choose a conservative p value cutoff of .001 and use (8) to determine the threshold levels corresponding to this cutoff. Table 1 provides the computed scan statistics and their p value approximations. We see from Figure 1 that using a length-based weighting scheme improves the power for both CeHV1 and BoHV1. For BoHV5, significant clusters of palindromes are found in the vicinity of the replication origins; however, there are also many false-positives for this genome.

Example 3. The location of the GATC pattern in a DNA sequence is known as a DAM site and is related to the repair and replication of DNA. An *Escherichia coli* genome sequence has approximately $n = 4.7$ million bp, with an approximate rate of $\lambda = 1.1/250$ occurrences of DAM sites per base pair. We are interested here in finding an unusually large number of DAM sites

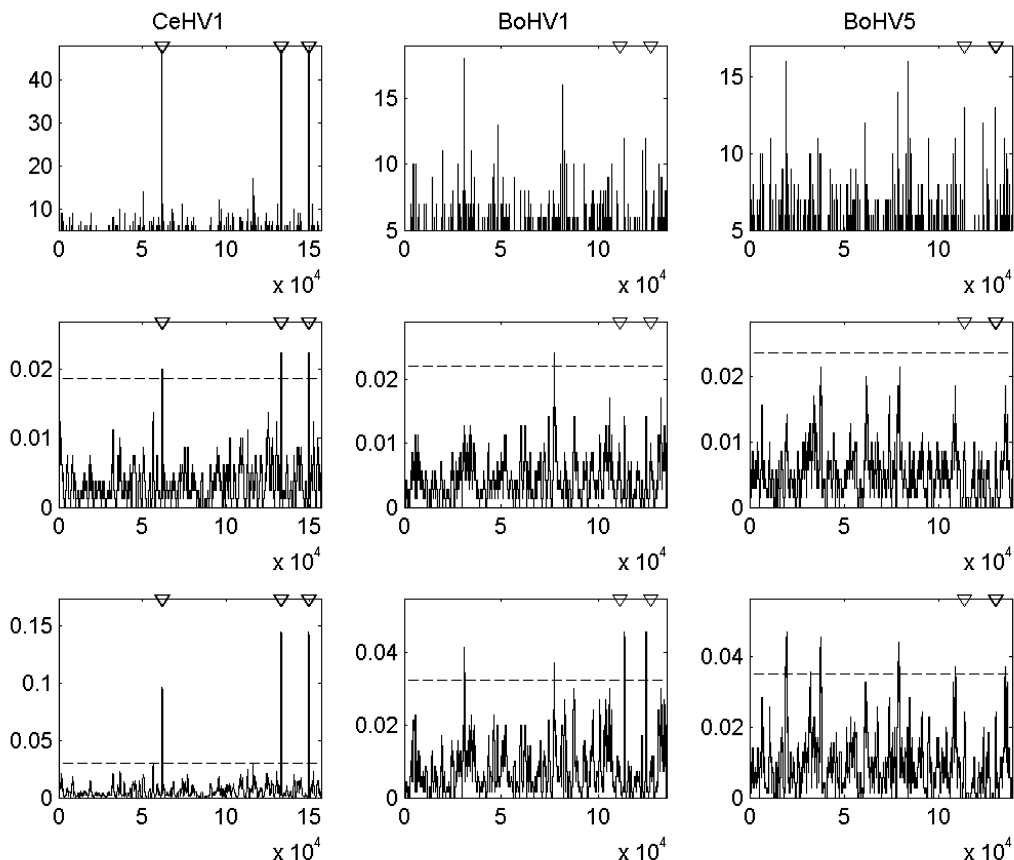


Figure 1. Comparison of Weighted and Unweighted Scan Statistics for Three Viral Genomes. For all plots, the horizontal axis is the location in the genome. The top plots show the locations and lengths of palindromes longer than 4. The middle plots show $u^{-1}[N(t + u/2) - N(t - u/2)]$ against t in the unweighted case. The bottom plots show $u^{-1}(S_{N(t+u/2)} - S_{N(t-u/2)})$ against t in the weighted case. Experimentally validated replication origins are indicated by triangles at the top of the plots. Dashed horizontal lines show thresholds for $p = .001$.

Table 1. Summary of Information for the Scan Statistics of Three Viral Genomes

	$(\hat{a}_A, \hat{a}_C, \hat{a}_G, \hat{a}_T)$	n	$N(n)$	u	Unweighted		F geometric	
					$M_{n,u}$	p	$M_{n,u}$	p
CeHV1	(.13, .37, .38, .13)	156,789	580	800	18	7.23×10^{-6}	116	0
BoHV1	(.14, .36, .37, .14)	135,301	615	700	17	1.09×10^{-4}	32	6.08×10^{-5}
BoHV5	(.12, .37, .38, .13)	138,390	714	700	15	1.07×10^{-2}	33	1.74×10^{-4}

in a sliding window of length $u = 245$ bp. A formula equivalent to the Newell–Ikeda approximation was used by Karlin and Brendel (1992) to compute the p value of the scan statistics. Glaz et al. (2001), however, suggested using the approach of Naus (1982). Table 2 compares these estimates with (10) and a Monte Carlo–derived estimate involving 2,000 simulation runs.

4. IMPORTANCE SAMPLING FOR SMALL p VALUES

Let $\mathbf{X}^{(i)} = \{N^{(i)}, X_1^{(i)}, \dots, X_{N^{(i)}(n)}^{(i)}\}$, $1 \leq i \leq B$, be B independent realizations of the underlying marked Poisson process with rate λ and weights following a distribution F . Then

$$\hat{p}_D = B^{-1} \sum_{i=1}^B \mathbb{1}_{\{M_{n,u}^{(i)} \geq ux\}} \quad \text{and} \quad (15)$$

$$\widehat{s.e.}_D = [\hat{p}_D(1 - \hat{p}_D)/B]^{1/2}$$

are the unbiased direct Monte Carlo estimate of $p := P\{M_{n,u} \geq ux\}$ and its consistent standard error estimate. We provide an alternative way of simulating the probability of $\{M_{n,u} \geq ux\}$ that attains substantial variance reduction when the probabilities are small. Let $\theta_x > 0$ and $\alpha_x > \lambda$ satisfy (3). For each $1 \leq i \leq B$, we do the following:

1. Generate $t^{(i)}$ uniformly from $[0, n - u]$.
2. Generate $N^{(i)}$ from a nonuniform Poisson process with rate α_x on the interval $(t^{(i)}, t^{(i)} + u]$ and rate λ elsewhere on $(0, n]$.
3. Generate independent random variables $X_1^{(i)}, \dots, X_{N^{(i)}(n)}^{(i)}$ with $X_j^{(i)}$ having distribution F_{θ_x} for $N^{(i)}(t^{(i)}) < j \leq N^{(i)}(t^{(i)} + u)$ and distribution F otherwise.

Let $S_k^{(i)} = X_1^{(i)} + \dots + X_k^{(i)}$. Then, by (3) and (4), the likelihood ratio of $\mathbf{X}^{(i)}$ between the underlying marked point process and the process generated through steps 1–3 is

$$L_i = \left[(n - u)^{-1} \int_0^{n-u} \left\{ e^{\theta_x(S_{N^{(i)}(t+u)}^{(i)} - S_{N^{(i)}(t)}^{(i)})} / [K(\theta_x)]^{N^{(i)}(t+u) - N^{(i)}(t)} \right\} \times e^{-(\alpha_x - \lambda)u} (\alpha_x / \lambda)^{N^{(i)}(t+u) - N^{(i)}(t)} dt \right]^{-1}$$

Table 2. Estimation of $P\{M_{n,u} \geq ux\}$ (\pm standard error for direct Monte Carlo)

ux	Direct Monte Carlo	Analytical estimate (10)	Naus (1982)	Newell–Ikeda
8	.876 \pm .007	.873	.870	.999
9	.25 \pm .01	.246	.244	.0987
10	.034 \pm .004	.0337	.0334	.011

$$= e^{-uI^{(x)}(n - u)} / \left(\int_0^{n-u} e^{\theta_x(S_{N^{(i)}(t+u)}^{(i)} - S_{N^{(i)}(t)}^{(i)} - ux)} dt \right). \quad (16)$$

The corresponding unbiased estimate of p and its respective standard error estimate are then

$$\hat{p}_I = B^{-1} \sum_{i=1}^B L_i \mathbb{1}_{\{M_{n,u}^{(i)} \geq ux\}} \quad \text{and} \quad (17)$$

$$\widehat{s.e.}_I = B^{-1} \left(\sum_{i=1}^B L_i^2 \mathbb{1}_{\{M_{n,u}^{(i)} \geq ux\}} - B \hat{p}_I^2 \right)^{1/2}.$$

For the unweighted scan statistics in which $F = F_{\theta_x}$ are degenerate at 1, step 3 can be omitted. By (9) and (16),

$$L_i = (n - u) e^{(x - \lambda)u} / \left(\int_0^{n-u} (x/\lambda)^{N^{(i)}(t+u) - N^{(i)}(t)} dt \right). \quad (18)$$

For the degenerate case, a similar importance sampling scheme was introduced by Naiman and Priebe (2001). Change of measure importance sampling schemes have also been used in sequential analysis (Siegmund 1976; Chan and Lai 2000), change-point detection (Lai and Shan 1999), and sequence alignments (Chan 2003). The change of measure associated with the foregoing importance sampling scheme also plays an important role in deriving the asymptotic expression of p in Theorem 1.

Example 4. We check some of the analytical estimates of $p = P\{M_{n,u} \geq ux\}$ that were applied in Examples 1 and 2 through direct Monte Carlo and importance sampling with $B = 2,000$ runs for each entry. In Table 3 we check the approximations for the HCMV genome with $n = 229,354$, $u = 1,000$ and $\lambda = 296/n = .00129$. Here we consider the unweighted scan statistics. In Table 4 we check the approximations for the BoHV1 genome with $n = 135,301$, $u = 700$, $\lambda = 615/n = .00455$, and F geometric with mean $\mu = (.700)^{-1}$. In Table 5 we check the approximations with parameters taken from the BoHV5 genomes, with $n = 138,390$, $u = 700$, $\lambda = 714/n = .00516$, and $P\{X_1 = k\} = (.3124)^{k-5} (.6876)$ for $k \geq 5$. We note that for F degenerate at 1, $v_x = 1$; for the geometric distribution, we compute $v_x = \mu[1 - (1 - \mu^{-1})e^{\theta_x}]$ [see (14)]. We calculate the overshoot of the shifted geometric by recursively computing the distribution of $R_{\min(t, \tau_+)}$ for $t = 1, 2, \dots$ and then applying (7) and (13).

Tables 3–5 reveal substantial variance reduction for probabilities of order 10^{-2} or 10^{-3} when importance sampling is used in place of direct Monte Carlo. For probabilities of order 10^{-4} or 10^{-5} , direct Monte Carlo breaks down whereas the importance sampling algorithm still provides reliable estimates. The

Table 3. Estimation of $p \pm$ s.e. With F Degenerate at 1

ux	Direct Monte Carlo	Importance sampling	Analytical estimate (10)	Naus (1982)
9	$(1.5 \pm .3) \times 10^{-2}$	$(1.3 \pm .07) \times 10^{-2}$	1.32×10^{-2}	1.32×10^{-2}
10	$(1 \pm 1) \times 10^{-3}$	$(2.1 \pm .1) \times 10^{-3}$	1.95×10^{-3}	1.93×10^{-3}
11	0	$(2.2 \pm .2) \times 10^{-4}$	2.53×10^{-4}	2.53×10^{-4}

analytical estimates (8) and (10) agree very well with the simulation results and are within 2 standard errors of the importance sampling estimates.

5. DISCUSSION

Here we have derived asymptotic approximations for the large-exceedance p values of scan statistics for marked Poisson processes. The proof is based on a change-of-measure approach, through which we also designed an importance sampling method for fast Monte Carlo evaluation of p values. The numerical studies presented in Section 4 show that the asymptotic approximation is close to the importance sampling estimates.

We applied our approximations to several classical examples in DNA sequence analysis and found them to be in agreement with previous approximations for the p values for the unweighted scan statistics. We also calculated the p values for high scoring windows for a weighted scoring scheme in the search for clusters of palindromes in viral genomes, with each incidence of a palindrome weighted by its length. We assumed a geometric distribution for the palindrome lengths and provided a simple formula for the overshoot constant ν_x . For more general scoring schemes, we could compute the overshoot constant using (12) or (13) and recursive numerical evaluation. This was illustrated in Example 4 for the shifted geometric distribution.

Our theoretical results depend on the assumption of independent X_i 's. For the problems in genome analysis that we study, this is roughly true because the locations of the events (palindromes in Examples 1 and 2 and DAM sites in Example 3) are spaced on the order of hundreds of bases apart, far enough to escape the documented local dependence in DNA sequences. In support of this assumption, for each genome analyzed, we tested the hypothesis that $S_{N(t)}$ has independent increments using the procedure described in Section A.2. Table 6 shows the test results for the three virus genomes analyzed in Examples 1 and 2. The p values for all tests are $>.05$.

A more involved scoring scheme for locating replication origin in viruses could take into account sequence composition, as well as allow (and compensate) for mismatches and gaps in the palindrome. For example, we could count the occurrences of inverted repeats, which are patterns constructed by inserting

Table 4. Estimation of $p \pm$ s.e. With F Geometric

ux	Direct Monte Carlo	Importance sampling	Analytical estimate (8)
24	$(1.0 \pm .2) \times 10^{-2}$	$(1.42 \pm .07) \times 10^{-2}$	1.5×10^{-2}
26	$(4 \pm 1) \times 10^{-3}$	$(4.3 \pm .3) \times 10^{-3}$	4.04×10^{-3}
28	0	$(1.00 \pm .07) \times 10^{-3}$	1.04×10^{-3}
30	0	$(2.5 \pm .2) \times 10^{-4}$	2.55×10^{-4}
32	0	$(5.3 \pm .4) \times 10^{-5}$	6.08×10^{-5}

Table 5. Estimation of $p \pm$ s.e. With F Shifted Geometric

ux	Direct Monte Carlo	Importance sampling	Analytical estimate (8)
80	$(1.5 \pm .3) \times 10^{-2}$	$(1.46 \pm .09) \times 10^{-2}$	1.43×10^{-2}
85	$(3 \pm 1) \times 10^{-3}$	$(4.7 \pm .3) \times 10^{-3}$	4.21×10^{-3}
90	0	$(1.16 \pm .08) \times 10^{-3}$	1.17×10^{-3}
95	0	$(3.2 \pm .5) \times 10^{-4}$	3.11×10^{-4}
100	0	$(7.5 \pm .5) \times 10^{-5}$	7.87×10^{-5}

arbitrary base pairs in the middle of a PLP. The GGACCTCC pattern is an example of an inverted repeat, constructed by inserting CC into the middle of GGATCC. We then assign scores that reward longer underlying PLPs as well as penalize for the number of inserted base pairs.

Recently there has been increasing interest of scanning entire genomes for regions that are enriched with a certain type of signal, such as transcription factor binding sites (Lifanov et al. 2003; Rajewsky et al. 2002), segments of high evolutionary conservation (Siepel et al. 1985, or a combination of both (Blanchette et al. 2006). The scoring functions used for such scans are often quite complex. For example, Blanchette et al. (2006) used a multiple alignment of the human, mouse, and rat genomes to locate the ‘‘hits,’’ and combined the pattern scores of the sequence in all three genomes when computing the score of a hit. Such complex scoring schemes can be analyzed using our methods by choosing the appropriate distribution function F .

APPENDIX: PROOFS AND JUSTIFICATIONS

A.1 Proof of Theorem 1

Let $0 < m < u$ and define

$$B_{w,m} = \left\{ \sup_{w < t \leq w+m} [S_{N(t+u)} - S_{N(t)}] \geq ux, \right. \\ \left. [S_{N(w+u)} - S_{N(w)}] < ux \right\}. \quad (A.1)$$

We show that if $m \rightarrow \infty$ as $u \rightarrow \infty$ with $m = o(u)$, then for all $0 \leq w \leq n - m$,

$$P(B_{w,m}) = P(B_{0,m}) \\ \sim m\nu_x e^{-u\lambda} (x - \lambda\mu) / [2\pi u\lambda K''(\theta_x)]^{1/2}. \quad (A.2)$$

Let $C > (x - \lambda\mu)$ and $\ell > 0$. We also show that

$$J_{1,r} := P\{S_{N(rm+u)} - S_{N(rm)} < ux - Cm, \\ \sup_{rm < t \leq (r+1)m} [S_{N(t+u)} - S_{N(t)}] \geq ux\} \\ = o(me^{-u\lambda} u^{-1/2}), \\ J_{2,r} := P(B_{rm,m} \cap B_{(r+1)m, (\ell-1)m}) \\ = o(me^{-u\lambda} u^{-1/2}). \quad (A.3)$$

Moreover, there exists $\ell > 0$ (depending on C) such that for all $\kappa > 0$,

$$J_{3,r} := \sum_{q=\ell}^{\lfloor \kappa u/m \rfloor - r} P\{S_{N(rm+u)} - S_{N(rm)} \geq ux - Cm, \\ S_{N(rm+qm+u)} - S_{N(rm+qm)} \geq ux - Cm\} \\ = o(me^{-u\lambda} u^{-1/2}), \quad (A.4)$$

uniformly over all $0 \leq r < \kappa u/m$. We can conclude from the inequalities

$$\sum_{r=0}^{\lfloor \kappa u/m \rfloor - 1} [P(B_{rm,m}) - J_{1,r} - J_{2,r} - J_{3,r}] \leq P(B_{0,\kappa u})$$

$$\leq \sum_{r=0}^{\lfloor \kappa u/m \rfloor} P(B_{rm,m}),$$

and (A.2)–(A.4), that for any fixed $\kappa > 0$,

$$P(B_{0,\kappa u}) \sim \kappa u v_x e^{-uI(x)} (x - \lambda u) / [2\pi u \lambda K''(\theta_x)]^{1/2}. \tag{A.5}$$

Theorem 1 then follows from (A.5) with κ large, the bound

$$P\{S_{N(t+u)} - S_{N(t)} \geq ux\} = P\{S_{N(u)} \geq ux\}$$

$$= O(u^{-1/2} e^{-uI(x)}), \tag{A.6}$$

and the independence of $B_{t,\kappa u}$ and $B_{w,\kappa u}$ for $(w - t) > (\kappa + 1)u$.

Proof of (A.2). Let us first assume that F is arithmetic with span 1. Let $\theta_x > 0$ and $\alpha_x > \lambda$ satisfy (3). We introduce here the probability measure Q , under which N is nonuniform Poisson with rate α_x on $(0, u]$ and rate λ on $(u, n]$, and $X_1, \dots, X_{N(n)}$ are independent random variables with X_j having distribution F_{θ_x} for $1 \leq j \leq N(u)$ and distribution F for $N(u) < j \leq N(n)$. Let $\mathbf{X} = \{N, X_1, \dots, X_{N(n)}\}$. By (3)–(5),

$$\frac{dQ}{dP}(\mathbf{X}) = \{e^{\theta_x S_{N(u)}} / [K(\theta_x)]^{N(u)}\}$$

$$\times e^{-\alpha_x u} (\alpha_x u)^{N(u)} / [e^{-\lambda u} (\lambda u)^{N(u)}]$$

$$= e^{\theta_x S_{N(u)} - (\alpha_x - \lambda)u}$$

$$= e^{uI(x) + \theta_x(S_{N(u)} - ux)}. \tag{A.7}$$

By stationarity,

$$P(B_{w,m}) = P(B_{0,m})$$

$$= E_Q[e^{-uI(x) - \theta_x(S_{N(u)} - ux)} \mathbb{1}_{B_{0,m}}]$$

$$= e^{-uI(x)} \sum_{b=1}^{\infty} e^{\theta_x b}$$

$$\times Q\left\{\sup_{0 < t \leq m} [S_{N(t+u)} - S_{N(t)}] \geq ux \mid S_{N(u)} = ux - b\right\}$$

$$\times Q\{S_{N(u)} = ux - b\}, \tag{A.8}$$

where the notation E_Q denotes expectation under probability measure Q and $Q(A)$ denotes the probability that the event A occurs under Q . It follows from (3) that under Q , the sum $S_{N(u)}$ is asymptotically normal with mean $u\alpha_x K'(\theta_x) / K(\theta_x) = ux$ and variance $u\alpha_x K''(\theta_x) / K(\theta_x) = u\lambda K''(\theta_x)$. Hence for any b fixed,

$$Q\{S_{N(u)} = ux - b\} \sim [2\pi u \lambda K''(\theta_x)]^{-1/2} \text{ as } u \rightarrow \infty. \tag{A.9}$$

We now evaluate the conditional term in (A.8). Observe that $[S_{N(t+u)} - S_{N(t)}]$ is the sum of all X_j lying inside the window $(t, t + u]$ [in other words, for all $N(t) < j \leq N(t + u)$]. We slide the window $(t, t + u]$ from $t = 0$ to $t = m$. At each point t whereby there is an inclusion or exclusion of an X_j from the interval $(t, t + u]$, there is a jump in the sum $[S_{N(t+u)} - S_{N(t)}]$. We represent the amount of this jump (possibly negative) by some Y_k . Conditioned on $S_{N(u)} = ux - b$, the crossing $[S_{N(t+u)} - S_{N(t)}] \geq ux$ occurs if there is an accumulation of jumps exceeding b . Because $m = o(u)$ and the Poisson process on $(u, m + u]$ has rate λ with observations X_j lying inside $(u, m + u]$ following distribution F , whereas the Poisson process on $(0, m]$ is of

rate α_x with observations X_j lying inside $(0, m]$ of distribution F_{θ_x} , we have

$$Q\left\{\sup_{0 < t \leq m} [S_{N(t+u)} - S_{N(t)}] \geq ux \mid S_{N(u)} = ux - b\right\}$$

$$\sim P\left\{\sup_{1 \leq k \leq T} R_k \geq b\right\}, \tag{A.10}$$

where T is Poisson with mean $m(\lambda + \alpha_x)$ and $R_k = Y_1 + \dots + Y_k$ with Y_1, Y_2, \dots iid random variables such that

$$P\{Y_1 = y\} = \left(\frac{\lambda}{\lambda + \alpha_x}\right) f(y) + \left(\frac{\alpha_x}{\lambda + \alpha_x}\right) f_{\theta_x}(-y). \tag{A.11}$$

Let P_* be another probability measure under which Y_1, Y_2, \dots are iid with probability mass function g [see (6)]. Then, by (3), (5), and (A.11),

$$P_*\{Y_1 = y\}$$

$$= g(y)$$

$$= \left(\frac{\alpha_x}{\lambda + \alpha_x}\right) f_{\theta_x}(y) + \left(\frac{\lambda}{\lambda + \alpha_x}\right) f(-y)$$

$$= \left(\frac{\alpha_x}{\lambda + \alpha_x}\right) e^{\theta_x y} (\lambda / \alpha_x) f(y) + \left(\frac{\lambda}{\lambda + \alpha_x}\right) e^{\theta_x y} (\alpha_x / \lambda) f_{\theta_x}(-y)$$

$$= e^{\theta_x y} P\{Y_1 = y\}. \tag{A.12}$$

Let E_* denote expectation with respect to P_* and $\tau_b = \inf\{k : R_k \geq b\}$. Then, by (7) and (A.12),

$$P\left\{\sup_{1 \leq k \leq T} R_k \geq b\right\} = E_*(e^{-\theta_x R_{\tau_b}} \mathbb{1}_{\{\sup_{1 \leq k \leq T} R_k \geq b\}})$$

$$\sim v_x e^{-\theta_x b} P_*\left\{\sup_{1 \leq k \leq T} R_k \geq b\right\}. \tag{A.13}$$

By substituting (A.13) into (A.10) and then substituting (A.9) and (A.10) into (A.8), we obtain

$$e^{uI(x)} [2\pi u \lambda K''(\theta_x)]^{1/2} P(B_{w,m})$$

$$\sim v_x \sum_{b=1}^{\infty} P_*\left\{\sup_{1 \leq k \leq T} R_k \geq b\right\} = v_x E_*\left(\sup_{1 \leq k \leq T} R_k\right)$$

$$\sim v_x (E_* R T) = v_x m (\alpha_x E_{\theta_x} X_1 - \lambda \mu), \tag{A.14}$$

and (A.2) follows from (3) because $E_{\theta_x} X_1 = K'(\theta_x) / K(\theta_x) = (x / \alpha_x)$.

The foregoing proof also can be modified to show (A.2) when F is arithmetic with span η by replacing the sum $\sum_{b=1}^{\infty}$ in (A.8) by $\sum_{b \in \eta \mathbf{Z}^+}$. Similarly, when F has a density function, (A.2) can be shown by the foregoing steps by changing the sum in (A.8) by a corresponding integral.

Proof of (A.3). Let $x^+ = \max(x, 0)$. By substituting (A.13) into (A.10) and then substituting (A.9) and (A.10) into (A.8), we obtain

$$e^{uI(x)} [2\pi u \lambda K''(\theta_x)]^{1/2}$$

$$\times P\left\{S_{N(m)} < ux - Cm, \sup_{rm < t \leq (r+1)m} S_{N(t)} \geq ux\right\}$$

$$\sim v_x \sum_{b > Cm} P_*\left\{\sup_{1 \leq k \leq T} R_k \geq b\right\}$$

$$\sim m v_x E_*\left[\left(m^{-1} \sup_{1 \leq k \leq T} R_k - C\right)^+\right] = o(m),$$

because $m^{-1} \sup_{1 \leq k \leq T} R_k$ has asymptotic mean $(x - \lambda \mu) / C$ and variance converging to 0 as $m \rightarrow \infty$. Thus the first part of (A.3) holds.

The second part of (A.3) follows from (A.2), (A.6), and the relation

$$J_{2,r} = P(B_{rm,m} \cup B_{(r+1)m,(\ell-1)m}) - P(B_{rm,m}) - P(B_{(r+1)m,(\ell-1)m}). \quad (\text{A.15})$$

Proof of (A.4). By stationarity, $J_{3,r}$ is monotone decreasing with respect to r , and hence we need consider only $r = 0$. By the independence of $S_{N(u)}$ and $(S_{N(u+qm)} - S_{N(qm)})$ for $q \geq u/m$,

$$\sum_{q=\lfloor u/m \rfloor + 1}^{\lfloor \kappa u/m \rfloor} P\{S_{N(u)} \geq ux - Cm, S_{N(u+qm)} - S_{N(qm)} \geq ux - Cm\} = (\lfloor \kappa u/m \rfloor - \lfloor u/m \rfloor) P^2\{S_{N(u)} \geq ux - Cm\}. \quad (\text{A.16})$$

We next consider a fixed $q < u/m$. Let \tilde{Q} be the probability measure under which $\mathbf{X} = \{N, X_1, \dots, X_{N(n)}\}$ is a marked point process with rate $\alpha_x = \lambda K(\theta_x)$ on $(qm, u]$, rate $\tilde{\alpha}_x := \lambda K(\theta_x/2)$ on $[0, qm] \cup (u, qm + u]$, and rate λ on $(qm + u, n]$. Moreover, we require that under \tilde{Q} ,

$$X_j \sim \begin{cases} F_{\theta_x} & \text{for } N(qm) < j \leq N(u) \\ F_{\theta_x/2} & \text{for } j \leq N(qm) \text{ and } N(u) < j \leq N(qm + u) \\ F & \text{otherwise.} \end{cases}$$

By (5),

$$\begin{aligned} \frac{d\tilde{Q}}{dP}(\mathbf{X}|N) &= \{e^{\theta_x[S_{N(u)} - S_{N(qm)}]}/[K(\theta_x)]^{N(u) - N(qm)}\} \\ &\quad \times \{e^{(\theta_x/2)(S_{N(qm)} + S_{N(u+qm)} - S_{N(u)})} \\ &\quad / [K(\theta_x/2)]^{N(qm) + N(u+qm) - N(u)}\} \end{aligned}$$

and

$$\begin{aligned} \frac{d\tilde{Q}}{dP}(N) &= \{e^{-2qm(\tilde{\alpha}_x - \lambda)} (\tilde{\alpha}_x/\lambda)^{N(qm) + N(u+qm) - N(u)}\} \\ &\quad \times \{e^{-(u - qm)(\alpha_x - \lambda)} (\alpha_x/\lambda)^{N(u) - N(qm)}\}. \end{aligned}$$

Because $(d\tilde{Q}/dP)(\mathbf{X}) = (d\tilde{Q}/dP)(\mathbf{X}|N) \times (d\tilde{Q}/dP)(N)$, it follows from (3) that

$$\frac{d\tilde{Q}}{dP}(\mathbf{X}) = e^{\theta_x(S_{N(u)} + S_{N(qm+u)} - S_{N(qm)})/2 - 2qm(\tilde{\alpha}_x - \lambda) - (u - qm)(\alpha_x - \lambda)}. \quad (\text{A.17})$$

Because $K(\theta_x) \geq [K(\theta_x/2)]^2$, it follows that $(\alpha_x/\lambda) \geq (\tilde{\alpha}_x/\lambda)^2$, and, noting that $\tilde{\alpha}_x^2/\lambda - (2\tilde{\alpha}_x - \lambda) = \lambda^{-1}(\tilde{\alpha}_x - \lambda)^2 > 0$, we conclude that

$$(\alpha_x - \lambda) \geq (\tilde{\alpha}_x^2/\lambda - \lambda) > 2(\tilde{\alpha}_x - \lambda). \quad (\text{A.18})$$

Let $\zeta = (\alpha_x - \lambda) - 2(\tilde{\alpha}_x - \lambda) (> 0)$. By (4), (A.17), and (A.18),

$$\begin{aligned} \frac{d\tilde{Q}}{dP}(\mathbf{X}) &\geq e^{\theta_x(S_{N(u)} + S_{N(qm+u)} - S_{N(qm)})/2 - u(\alpha_x - \lambda) + \zeta qm} \\ &= e^{u\ell(x) + \zeta qm + \theta_x[(S_{N(u)} + S_{N(qm+u)} - S_{N(qm)})/2 - ux]}, \end{aligned}$$

and thus by an analog of (A.9),

$$\begin{aligned} P\{S_{N(u)} \geq ux - Cm, S_{N(qm+u)} - S_{N(qm)} \geq ux - Cm\} &\leq P\{S_{N(u)} + S_{N(qm+u)} - S_{N(qm)} \geq 2(ux - Cm)\} \\ &= E_{\tilde{Q}} \left[\frac{dP}{d\tilde{Q}}(\mathbf{X}) \mathbb{1}_{\{S_{N(u)} + S_{N(qm+u)} - S_{N(qm)} \geq 2(ux - Cm)\}} \right] \\ &= O(u^{-1/2} e^{-u\ell(x) + \theta_x Cm - \zeta qm}), \end{aligned} \quad (\text{A.19})$$

and (A.4) follows from (A.6), (A.16), and (A.19) by choosing $\ell > \theta_x C/\zeta$.

Table A.1. Test of Independent Increments for Three Virus Genomes

Virus	n	u	l	r	χ^2	p
CeHV1	156,789	800	160	15	198.3317	.08
BoHV1	135,301	700	140	7	48.2131	.80
BoHV5	138,390	700	140	5	11.1809	.44

A.2 A Test for Independence

To test whether $S_{N(t)}$ has independent increments, we used the following procedure:

1. Divide $(0, n]$ into l segments, each of length n/l . Set $Z_k = S_{N(kn/l)} - S_{N((k-1)n/l)}$ for $k = 1, \dots, l$.
2. Divide the state space of Z_k evenly into r blocks $\{R_1, \dots, R_r\}$. Perform the nonparametric chi-squared goodness-of-fit test for independence of the pairs (Z_k, Z_{k+1}) using a $r \times r$ contingency table with entries

$$N_{i,j} = \#\{k : Z_k \in R_i, Z_{k+1} \in R_j\}.$$

The parameter l in this procedure should be chosen to be smaller than the scanning window size u , because dependence within this range would cause a ‘‘clumping’’ effect that renders our approximations inaccurate. We set l to be $u/5$. The number of blocks r is set to be $\text{Range}(Z_k : 1 \leq k \leq l - 1)/4$. Table A.1 shows the parameters, chi-squared statistics, and p values for the three genomes in Example 2.

[Received May 2006. Revised September 2006.]

REFERENCES

- Blanchette, M., Bataille, A., Chen, X., Poitras, C., Laganière, J., Lefévre, C., Deblois, G., Giguère, V., Ferretti, V., Bergeron, D., Coulombe, P., and Robert, F. (2006), ‘‘Genome-Wide Computational Prediction of Transcriptional Regulatory Modules Reveal New Insights Into Human Gene Expression,’’ *Genome Research*, 16, 656–668.
- Chan, H. (2003), ‘‘Upper Bounds and Importance Sampling of p -Values for DNA and Protein Sequence Alignments,’’ *Bernoulli*, 9, 183–199.
- Chan, H., and Lai, T. (2000), ‘‘Asymptotic Approximations for Error Probabilities of Sequential or Fixed Sample Size Tests in Exponential Families,’’ *The Annals of Statistics*, 28, 1638–1669.
- Chew, D., Choi, K., and Leung, M. (2005), ‘‘Scoring Schemes of Palindrome Clusters for More Sensitive Prediction of Replication Origins in Herpesviruses,’’ *Nucleic Acids Research*, 33, e134.
- Cressie, N. (1980), ‘‘The Asymptotic Distribution of the Scan Statistic Under Uniformity,’’ *The Annals of Probability*, 8, 828–840.
- Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. II (2nd ed.), New York: Wiley.
- Frolov, A. (2005), ‘‘Unified Limit Theorems for Increments of Processes With Independent Increments,’’ *Theory of Probability and Its Applications*, 49, 531–540.
- Glaz, J. (1989), ‘‘Approximations and Bounds for the Distribution of the Scan Statistic,’’ *Journal of the American Statistical Association*, 84, 560–566.
- Glaz, J., and Naus, J. (1991), ‘‘Tight Bounds and Approximations for Scan Statistic Probabilities for Discrete Data,’’ *The Annals of Applied Probability*, 1, 306–318.
- Glaz, J., Naus, J., and Wallenstein, S. (2001), *Scan Statistics*, New York: Springer-Verlag.
- Karlin, S., and Brendel, V. (1992), ‘‘Chance and Statistical Significance in Protein and DNA Sequence Analysis,’’ *Science*, 257, 39–49.
- Lai, T., and Shan, Z. (1999), ‘‘Efficient Recursive Algorithm for Detection of Abrupt Changes in Signal and Control Systems,’’ *IEEE Transactions on Automatic Control*, 44, 952–966.
- Leung, M., Schachtel, G., and Yu, H. (1994), ‘‘Scan Statistics and DNA Sequence Analysis: The Search for an Origin of Replication in a Virus,’’ *Nonlinear World*, 1, 445–471.
- Lifanov, A., Makeev, V., Nazina, A., and Papatsenko, D. (2003), ‘‘Homotypic Regulatory Clusters in *Drosophila*,’’ *Genome Research*, 13, 579–588.
- Loader, C. (1991), ‘‘Large Deviation Approximations to the Distribution of Scan Statistics,’’ *Advances in Applied Probability*, 23, 751–771.
- Masse, M. J. O., Karlin, S., Schachtel, G. A., and Mocarski, E. S. (1992), ‘‘Human Cytomegalovirus Origin of DNA Replication (oriLyt) Residues Within a Highly Complex Repetitive Region,’’ *Proceedings of the National Academy of Sciences*, 89, 5246–5250.

- Naiman, D., and Preibe, C. (2001), "Computing Scan Statistics p -Values Using Importance Sampling, With Applications to Genetics and Medical Image Analysis," [*Journal of Computational and Graphical Statistics*, 10, 296–328.](#)
- Naus, J. (1965), "The Distribution of the Size of the Maximum Cluster of Points on a Line," *Journal of the American Statistical Association*, 60, 532–538.
- (1982), "Approximations for Distributions of Scan Statistics," *Journal of the American Statistical Association*, 77, 177–183.
- Rajewsky, N., Vergassola, M., Gaul, U., and Siggia, E. (2002), "Computational Detection of Genomic *cis*-Regulatory Modules Applied to Body Patterning in the Early *Drosophila* Embryo," *BMC Bioinformatics*, 3, e30.
- Siegmund, D. (1976), "Importance Sampling in the Monte Carlo Study of Sequential Tests," *The Annals of Statistics*, 4, 673–684.
- (1985), *Sequential Analysis*, New York: Springer-Verlag.
- Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S., Weinstock, G., Wilson, R., Gibbs, R., Kent, J., Miller, W., and Haussler, D. (1985), "Evolutionarily Conserved Elements in Vertebrate, Insect, Worm, and Yeast Genomes," [*Genome Research*, 15, 1034–1050.](#)
- Storey, J. D., and Siegmund, D. (2001), "Approximate p -Values for Local Sequence Alignments: Numerical Studies," [*Journal of Computational Biology*, 8, 549–556.](#)
- Tu, I., and Siegmund, D. (1999), "The Maximum of a Function of a Markov Chain and Application to Linkage Analysis," [*The Advances in Applied Probability*, 31, 510–531.](#)