

Estimating the Occurrence Rate of DNA Palindromes

I-Ping Tu^{a*}, Shao-Hsuan Wang^b, and Yuan-Fu Huang^a

^aInstitute of Statistical Science, Academia Sinica

^bDepartment of Mathematics, National Taiwan University

December 20, 2012

Abstract

A DNA palindrome is a segment of letters along a DNA sequence with inversion symmetry that one strand is identical to its complementary one running in the opposite direction. Searching non-random clusters of DNA palindromes, an interesting bioinformatic problem, relies on the estimation of the null palindrome occurrence rate. The most commonly used approach for estimating this number is the average rate method. However, we observed that the average rate could exceed the actual rate by 50% when inserting 5,000 bp hot-spot regions with 15-fold rate in a simulated 150,000 bp genome sequence. Here, we propose a Markov based estimator to avoid counting the number of palindromes directly, and thus to reduce the impact from the hot-spots. Our simulation shows that this method is more robust against the hot-spot effect than the average rate method. Furthermore, this method can be generalized to either a higher order Markov model or a segmented Markov model, and extended to calculate the occurrence rate for palindromes with gaps. We also provide a p -value approximation for various scan statistics to test non-random palindrome clusters under a Markov model.

Keywords and phrases: DNA palindrome, genome sequence, hairpin structure, higher order Markov model, hot-spot, Markov model, occurrence rate, Poisson process, power, p -value, segmented Markov model.

*Corresponding author. *Email address:* iping@stat.sinica.edu.tw

1 Introduction

A chromosome is a long double-stranded helix of DNA that consists of adenine-thymine (A-T) pairs or cytosine-guanine (C-G) pairs. Thus, one DNA strand decides the sequence of its complementary strand. A DNA palindrome with minimum half length L is defined as a segment of DNA letters with half length greater than or equal to L that one strand is identical to its complementary one running in the opposite direction. This inversion symmetry increases the probability to form secondary structures conferring significant biological functions ranging from RNA transcription to DNA replication (Leach (1994)).

It has been observed that DNA palindromes are common candidates for searching genetic motifs involved in different cellular processes, including gene transcriptions, gene replications, and gene deletions. For example, among nine octameres (segments with 8 bp) suggested to be transcription factor binding sites, three are palindromes (FitzGerald et al. (2004)). Many studies have focused on investigating the occurrence rates of palindromes in suspicious regions against random sequences. For example, Lisnic and Svetec (2005) investigated the frequencies of palindromes in the yeast *Saccharmyces cerevisiae* genome. Chew et al. (2005) proposed various score schemes to quantify palindromes and found an association between high score regions and the replication origins. Lu et al. (2007) compared the scores of the suspicious regions, including introns, exons, and upstream of transcription start sites, against simulated random sequences, and reported that meaningful sites tend to have higher palindrome scores.

The analysis of these comparisons strongly depends on the null occurrence rate. This rate usually is estimated by the genome-wide average or the iid model based method using the DNA letter frequencies (Chew et al. (2005)). We tested these two methods on a herpes virous sequence bohv1 (sequence ID 'BHV1CGEN'). Its average rate is 0.00178 and the iid model based estimate is 0.00073. The large discrepancy between these two alerted us, and further studies indicated that the average rate might be biased due to hot-spots, and the iid model might be too naive to describe the DNA sequence. Therefore, we propose a Markov based estimator using the DNA pairs' frequencies in addition to the letters' and get the estimate 0.00109. Compared to the iid model, the Markov model is more close to real sequences but yet not too complicated to estimate its parameters. The simulation shows that our method performs better than the average rate in estimating the null occurrence rate against hot-spots under a variety of model

settings. We also show that this method can be generalized for either a higher order Markov model or a segmented Markov model. Furthermore, we demonstrate that this method can be extended to calculate the occurrence rate when the DNA palindrome contains a gap.

Many related p -value approximations have been developed based on the assumption that the events can be modeled as a Poisson process. This assumption has been justified in many cases including DNA palindromes (Reinert and Schbath (1998); Leung et al. (2005); Hansen (2009)). Chan and Zhang (2007) developed a method to approximate the p -value for a scan of score statistics over a Poisson process, when the score can be modeled through an exponential family. To apply their method, the analytic formula for the moment generating function (MGF) of the score is required. However, the distribution of the palindrome scores has not been well studied except the length score under the iid assumption. Thus, we develop a method to derive the analytic formulae for the MGF of various scores under a Markov model. Another challenge in calculating the p -value approximations of scan statistics is to calculate an overshoot function (Woodroffe (1978); Siegmund (1985); Tu (2009)). This function relates to the characteristic function of a random variable defined by the difference between the statistic and the threshold given the condition that the statistic exceeds the threshold. We further extend the p -value approximations developed by Chan and Zhang (2007) to provide a more general formula to calculate the overshoot functions on various scores under a Markov model.

In this paper, we first show that three scores for detecting DNA palindrome clusters proposed by Chew et al. (2005) can be formulated as likelihood ratio statistics. Second, we show that the occurrence rates can be calculated accurately under a Markov model. Third, we derive the moment generating function for various scores under a Markov model. Fourth, we present a p -value approximation method for those statistics to detect DNA palindrome clusters. In Section 3, we show the results on both real data and simulated data. This paper ends with a brief discussion. The related derivations are collected in online supplementary materials.

2 Method

2.1 Notations and the Log Likelihood Ratio Statistics

Let $N(t)$ be the counting process for the palindrome events and let $N_w(t) = N(t+w) - N(t)$ denote the number of palindromes whose starting positions fall in the interval $(t, t+w]$. Leung et al. (2005) proved that $N(t)$ can be approximated by a Poisson process under a Markov Model. We let X_i be the score for the i th palindrome (event) along the genome sequence, and $S_{N_w(t)}$ is the summation of the palindrome scores:

$$S_{N_w(t)} = \sum_{i=N(t)+1}^{N(t+w)} X_i. \quad (1)$$

To search the clusters of palindromes, Chew et al. (2005) proposed three schemes to quantify palindromes including the palindrome count score (PCS), the palindrome length score (PLS), and the base-pair weighted score of order m (BWS_m). PCS gives the same score for each DNA palindrome; PLS gives the score as the palindrome length divided by its minimum required length; and BWS_m gives the score as the minus log-likelihood under Markov order m assumption.

We would like to show that both $N_w(t)$ and $S_{N_w(t)}$ are equivalent to the log-likelihood ratio statistics when the alternative hypotheses are properly constructed. This equivalence is useful for developing the p -value approximations. Under the Poisson process model, we also assume that X_i 's can be treated as iid with a density function $f_\theta(x) = f_0(x) \exp(\theta x - \phi(\theta))$, where $f_0(x)$ is an unknown distribution and $\phi(\theta) = \log \int e^{\theta x} f_0(x) dx$ is its log of the MGF. For events that occur outside of the interval $(t_a, t_a + w]$, the parameters are (λ_0, θ_0) . For events that occur in the interval $(t_a, t_a + w]$, the parameters for $N(t)$ and X_i , are (λ_a, θ_a) . The null hypothesis is that $\lambda_a = \lambda_0$ and $\theta_a = \theta_0$. When t_a is known, the likelihood ratio is $f_{\lambda_a, \theta_a}(N_w(t_a), S_{N_w(t_a)}) / f_{\lambda_0, \theta_0}(N_w(t_a), S_{N_w(t_a)})$, and the likelihood is as follows:

$$\begin{aligned} & f_{\lambda, \theta}(N_w(t), S_{N_w(t)}) \\ = & f_\lambda(N_w(t)) f_\theta(S_{N_w(t)} | N_w(t)) \\ = & \frac{(\lambda w)^{N_w(t)} e^{-\lambda w}}{N_w(t)!} \left(\prod_{i=N(t)+1}^{N(t+w)} f_0(x_i) \right) \exp(\theta S_{N_w(t)} - N_w(t) \phi(\theta)). \end{aligned}$$

Because t_a is usually unknown, we search for the maximum of the statistic over all possible t .

Case 1. If the alternative hypothesis is constructed as $H_a : \lambda_a = \lambda_1 > \lambda_0$ and $\theta_a = \theta_0$, then the log-likelihood ratio statistic is equivalent to PCS in Chew et al. (2005):

$$\begin{aligned} \max_t l_t(\lambda_1, \theta_0) &= \max_t \log \left(\frac{f_{\lambda_1, \theta_0}(N_w(t), S_{N_w(t)})}{f_{\lambda_0, \theta_0}(N_w(t), S_{N_w(t)})} \right) \\ &= \max_t N_w(t) \log \left(\frac{\lambda_1}{\lambda_0} \right) - (\lambda_1 - \lambda_0)w. \end{aligned} \quad (2)$$

Case 2. If the alternative hypothesis is constructed as $H_a : \lambda_a = \lambda_1 > \lambda_0$ and $\theta_a = \theta_1 > \theta_0$, where λ_1 and θ_1 are constrained to satisfy

$$\log \left(\frac{\lambda_1}{\lambda_0} \right) - (\phi(\theta_1) - \phi(\theta_0)) = 0, \quad (3)$$

the log-likelihood ratio statistic in formula (4) can be equivalent to PLS or BWS_m proposed by Chew et al. (2005), depending on the definition of X_i 's.

$$\begin{aligned} \max_t l_t(\lambda_1, \theta_1) &= \max_t \log \left(\frac{f_{\lambda_1, \theta_1}(N_w(t), S_{N_w(t)})}{f_{\lambda_0, \theta_0}(N_w(t), S_{N_w(t)})} \right) \\ &= \max_t \{ -(\lambda_1 - \lambda_0)w + (\theta_1 - \theta_0)S_{N_w(t)} \} \end{aligned} \quad (4)$$

It can be observed that (2) is equivalent to $\max_t N_w(t)$ and (4) is equivalent to $\max_t S_{N_w(t)}$. While (2) tests only the Poisson parameter λ , (4) tests both the Poisson parameter λ and the score parameter θ with the constraint (3). $N_w(t)$ can be treated as a special case of $S_{N_w(t)}$ with $X_i = 1$ for each i .

We applied the method developed by Chan and Zhang (2007) to derive the threshold value of $\max_t S_{N_w(t)}$. Let $N(t)$ be a Poisson process with mean λ_0 and the log of the MGF of X_i is $\phi(\theta)$. X_i 's are iid with mean μ_0 , then

$$\begin{aligned} &P_0 \left(\max_{0 < t < W} S_{N_w(t)} \geq b \right) \\ &\sim 1 - \exp \left\{ -(W - w) \nu_{\lambda_1, \theta_1} (b - \lambda_0 \mu_0) e^{-[b\theta_1 - w(\lambda_1 - \lambda_0)]} \left(2\pi w \lambda_1 [\ddot{\phi}(\theta_1) + \dot{\phi}^2(\theta_1)] \right)^{-1/2} \right\}, \end{aligned} \quad (5)$$

where W is the total length of the sequence, $\dot{\phi}(\theta_1)$ and $\ddot{\phi}(\theta_1)$ are the first and the second derivative of $\phi(\theta_1)$ representing the mean and the variance of X_i with density $f_{\theta_1}(x)$, and $\nu_{\lambda_1, \theta_1}$ is an overshoot function indexed with (θ_1, λ_1) satisfying the equations:

$$\begin{aligned} w \lambda_1 \phi'(\theta_1) &= b, \\ \log(\lambda_1 / \lambda_0) &= \phi(\theta_1) - \phi(\theta_0). \end{aligned}$$

It is obvious from (5) that λ_0 always plays a crucial role in deciding the critical value for the tests. The most challenging part in applying (5) to calculate the p -value is to calculate the overshoot function $\nu_{\lambda_1, \theta_1}$. As a pioneer, Woodroffe (1978) derived a computable formula to calculate the overshoot function for iid non-arithmetic random variables given the characteristic functions. Tu (2009) generalized this formula for iid arithmetic random variables. Incorporating their results, we develop Theorem 4 for more general cases.

2.2 Occurrence rate of DNA palindromes under Markov model

Let T be a 4×4 matrix with $T_{ij} = P_{b_i b_j} P_{\tilde{b}_j \tilde{b}_i}$ which groups together the transition probabilities of symmetric complementary pairs, where $(b_1, b_2, b_3, b_4) = (A, C, G, T)$ and \tilde{b}_j refers to the complementary letter of b_j . T is considered as a quasi transition matrix because the sum of its rows do not equal one.

Theorem 1 Assume that DNA letters along the genome sequence follow a Markov model with transition probability $\{P_{a,b} \mid a, b \in \{A, C, G, T\}\}$ and letter frequency $P'_0 = (\pi_A \ \pi_C \ \pi_G \ \pi_T)$. The occurrence probability of a palindrome I_i (given a starting position i) with minimum half length L is

$$\lambda_M \equiv P(\|I_i\| \geq L) = P'_0 T^{L-1} P_1 \quad (6)$$

where $\|I_i\|$ denotes the corresponding maximum length, and

$$P'_1 = (P_{AT} \ P_{CG} \ P_{GC} \ P_{TA}),$$

and

$$T = \begin{pmatrix} P_{AA}P_{TT} & P_{AC}P_{GT} & P_{AG}P_{CT} & P_{AT}P_{AT} \\ P_{CA}P_{TG} & P_{CC}P_{GG} & P_{CG}P_{CG} & P_{CT}P_{AG} \\ P_{GA}P_{TC} & P_{GC}P_{GC} & P_{GG}P_{CC} & P_{GT}P_{AC} \\ P_{TA}P_{TA} & P_{TC}P_{GA} & P_{TG}P_{CA} & P_{TT}P_{AA} \end{pmatrix}.$$

Remark 1.1

Theorem 1 gives an exact formula to calculate the palindrome occurrence rate under a Markov model. To apply this formula, one needs to estimate the Markov parameters

including the stationary probabilities $\{\pi_a \mid a \in \{A, C, G, T\}\}$ and the transition matrix $\{P_{a,b} \mid a, b \in \{A, C, G, T\}\}$ which usually are estimated by the letter frequencies of W letters and the pair frequencies of $W - 1$ letter pairs, where W is the total sequence length. Because the size of the hot-spot regions is much smaller than W that these Markov parameter estimations are not heavily influenced by hot-spots and neither is λ_M to estimate the null occurrence rate. On the other side, for the rare events like the DNA palindromes, the average rate counts the total number of palindromes of which a non-negligible portion is potentially contributed from the hot-spots in real sequences, and hence the average rate is easily inflated when estimating the null occurrence rate.

Remark 1.2 iid Model

When the Markov model is reduced to the iid model, P'_1 becomes

$$P'_2 = (\pi_T \ \pi_G \ \pi_C \ \pi_A),$$

and T becomes $P_2 P'_0$. Thus,

$$\lambda_{\text{iid}} \equiv P(\|I_i\| \geq L) = P'_0 (P_2 P'_0)^{L-1} P_2 = (P'_0 P_2)^L = \gamma^L, \quad (7)$$

where $\gamma = 2(\pi_A \pi_T + \pi_C \pi_G)$. (7) has been shown in Leung et al. (2005).

Remark 1.3 Higher Order Markov Models

When the DNA sequence does not follow a first-order Markov model, a higher order Markov model may be considered. Theorem 1 can also be applied to higher order Markov models. For example, a four-state second-order Markov model can be described as a first-order Markov model with sixteen states, in which each state represents one adjacent letter pair, like $a_1 a_2$ where $a_i \in \{A, T, G, C\}$ and the probability model becomes $P(a_1 a_2 a_3) = P(a_1 a_2) P(a_2 a_3 \mid a_1 a_2)$. Under this setting, only four elements in each row or each column of the 16×16 transition matrix will be non-zero, because $P_{a_1 a_2, a_3 a_4} = 0$ if $a_2 \neq a_3$. The elements in the corresponding quasi transition matrix will be like $P_{a_1 a_2, a_3 a_4} P_{\tilde{a}_4 \tilde{a}_3, \tilde{a}_2 \tilde{a}_1}$. The corresponding P_1 in (6) is a column vector with length sixteen, in which the elements are like $P_{a_1 a_2, a_2 \tilde{a}_2} P_{a_2 \tilde{a}_2, \tilde{a}_2 \tilde{a}_1}$.

Remark 1.4 Segmented Markov Models

Another alternative model is the segmented Markov model which relaxes the stationary

condition (Chen and Zhou (2010)). Considering a three-state segmented Markov model $\{\xi_1, \xi_2, \xi_3\}$ that each state contributes total length W_1, W_2 and W_3 and has its own p -values p_1, p_2 and p_3 . Two constraints on these p -values are $1 - (1 - p_1)(1 - p_2)(1 - p_3) = 0.05$ by the Bonferroni approximation on the overall 0.05 significance level and $\frac{p_1}{W_1} = \frac{p_2}{W_2} = \frac{p_3}{W_3}$ that the p -values are proportional to their contributed length. Given change points that separate different hidden states ξ 's, each parameter set can be estimated respectively. As such, Theorem 1 and Theorem 4 can be applied to calculate the palindrome occurrence rate for each set and their corresponding thresholds.

Remark 1.5 Hairpin Structures

It is of interest to consider DNA palindromes with gaps. When a gap exists at the center position of a DNA palindrome, the single strand segment may form a hairpin secondary structure (Leach (1994)). Theorem 1 can be extended to calculate the occurrence rate for such patterns. Consider a palindrome with the half-length $\geq L$ and a gap with length β at the center position of the palindrome, then the probability to see such a pattern given a start position is $P'_0 T^{L-1} \text{diag}((P^\beta - P^{\beta-2})\tilde{P})$ when $\beta \geq 2$ and $P'_0 T^{L-1} \text{diag}(P\tilde{P})$ when $\beta = 1$, where \tilde{P} is the transition matrix with the index order (A C G T) for row and (T G C A) for column. The technical derivation is in the Appendix A.2.

Theorem 2 Under the same assumption discussed in Theorem 1, the PLS score for the i th palindrome is defined as $X_i = \|I_i\|/L$ conditional on $\|I_i\| \geq L$. As such, the MGF for X_i is

$$K_{\text{PLS}}(t) \equiv E(e^{X_i t} \mid \|I_i\| \geq L) = \frac{e^t}{\lambda_M} P'_0 T^{L-1} (I - e^{t/L} T)^{-1} (I - T) P_1. \quad (8)$$

Remark 2.1 iid model

When the Markov model is reduced to the iid model,

$$K_{\text{PLS}}(t) = \sum_{k=L}^{\infty} e^{kt/L} (\gamma^k - \gamma^{k+1}) / \gamma^L = \frac{e^t(1 - \gamma)}{1 - e^{t/L}\gamma}.$$

Theorem 3 Under the same assumption discussed in Theorem 1, the BWS score is defined as $X_i = -\log(P(I_i))$ conditional on $\|I_i\| \geq L$. Then, the MGF for X_i is

$$K_{\text{BWS}}(t) \equiv E[e^{X_i t} \mid \|I_i\| \geq L] = \frac{1}{\lambda_M} \mathbf{v}'(t)(I - Q(t))^{-1}(Q(t))^{L-1}\mathbf{u}(t), \quad (9)$$

where $\mathbf{v}(t) = (v_1(t) \ v_2(t) \ v_3(t) \ v_4(t))'$ is defined as $v_i(t) = ([(I-T)P_0]_i)^{1-t}$; $Q(t)$ is defined as $Q_{ij}(t) = (T_{ij})^{(1-t)}$; and $\mathbf{u}(t) = (u_1(t) \ u_2(t) \ u_3(t) \ u_4(t))'$ is defined as $u_i(t) = ([P_1]_i)^{1-t}$ with $i = 1, \dots, 4$.

Remark 3.1 iid model

When the Markov model is reduced to the iid model,

$$K_{\text{BWS}}(t) = \frac{(1-\gamma)^{1-t}}{1-\gamma_t} \left(\frac{\gamma_t}{\gamma} \right)^L, \quad (10)$$

where $\gamma_t = P'_0(t)P_2(t) = 2[(\pi_A\pi_T)^{1-t} + (\pi_C\pi_G)^{1-t}]$. To provide a more general approximation for the p -value of maximum of (1), we develop Theorem 4.

Theorem 4 Let N be a Poisson process with constant rate $\lambda_0 > 0$ and let random variables $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f_{\theta_0}(\cdot)$ with the MGF $\exp(\phi(\theta))$. Let λ_1 and θ_1 satisfy two conditions : (a) $w\lambda_1\phi'(\theta_1) = b$. (b) $\log(\lambda_1/\lambda_0) - (\phi(\theta_1) - \phi(\theta_0)) = 0$. Let $W \rightarrow \infty$ as $w \rightarrow \infty$ such that $W - w \rightarrow \infty$. Then

$$P_0(\max_{0 < s < W} S_{N_w(s)} \geq b) \approx 1 - \exp \left\{ -(W - w)\nu_{\lambda_1, \theta_1}(b - \lambda_0\mu_0)e^{-I(b)w} \left(2\pi w\lambda_1[\ddot{\phi}(\theta_1) + \dot{\phi}^2(\theta_1)] \right)^{-1/2} \right\},$$

where $\nu_{\lambda_1, \theta_1} = \frac{1 - E_0 e^{-S_{\tau_+}(\theta_1 - \theta_0)}}{(1 - e^{-(\theta_1 - \theta_0)})E_0 S_{\tau_+}}$ and $I(b) = b(\theta_1 - \theta_0)/w - (\lambda_1 - \lambda_0)$. The definition of S_{τ_+} and the proof of Theorem 4 are put online.

Remark 4.1

Theorem 4 follows Theorem 1 of Chan and Zhang (2007) but extends their result to allow more varieties of scores on the events. The assumptions of Theorem 4 include that the event can be modeled as a Poisson process and the scores for those events are iid. It has been reported that the approximation error percentage (compared to the Monte Carol simulation) is less than 5% when $W = 20w$ and $w = 9$ unit length in Tu (2012). However, in such an application to search non-random palindrome clusters, the window

size 9 bp can not cover even one palindrome. Thus, the issue of choosing w is addressed by efficiently searching the clusters. The criterion should be that w needs to be large enough to cover the clusters but not too large to dilute its density. In our experience, 500 to 1,000 bp is a reasonable size of w to search the palindrome clusters of the herpes virus genomes. Another example appeared in Chan and Zhang (2007) that they used the window size 245 to search GATC clusters as DAM sites in an *E. coli* genome sequence.

3 Real Data Analysis and Simulations

We studied 27 herpes virus genome sequences among which bohv1 with total length 135,301 bp is chosen as our model sequence. Two replication origins of bohv1 have been reported in the literature (Leung et al. (2005)) and there exist distinct deviations among its average rate (0.00178), Markov model estimate (0.00109) and the iid estimate (0.00073). Its transition matrix and stationary probabilities are estimated as¹

$$P_{\text{bohv1}} = \begin{pmatrix} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & 0.1854 & 0.3288 & 0.3556 & 0.1303 \\ \text{C} & 0.1258 & 0.2932 & 0.4347 & 0.1463 \\ \text{G} & 0.1343 & 0.4512 & 0.2994 & 0.1151 \\ \text{T} & 0.1141 & 0.3151 & 0.3695 & 0.2012 \end{pmatrix},$$

$$\pi_{\text{bohv1}} = (\quad 0.1354 \quad 0.3588 \quad 0.3654 \quad 0.1405). \quad (11)$$

We also employed a second-order Markov model on bohv1, and estimated the occurrence rate as 0.00113. The closeness of this value to that from a first-order Markov model (0.00109) suggests the appropriateness of a first-order Markov model for this sequence. We followed the criterion that minimum half length times the square of the occurrence rate is most close to 0.16 proposed in Leung et al. (2005). Thus, among the 27 herpes virus sequences, we use $L \geq 6$ as the palindrome criterion for 5 sequences including bohv1, cev1, hsv2, muhv4 and thy, and $L \geq 5$ for the remaining 22 sequences.

3.1 Real Data Analysis

We downloaded from the EBI Nucleotide Sequences database 27 herpes virus genome sequences. For each sequence, we estimated its own transition matrix and stationary

¹Standard errors of the estimated transition probabilities are roughly $\sqrt{0.4 \cdot 0.6/135301} \approx 0.0013$.

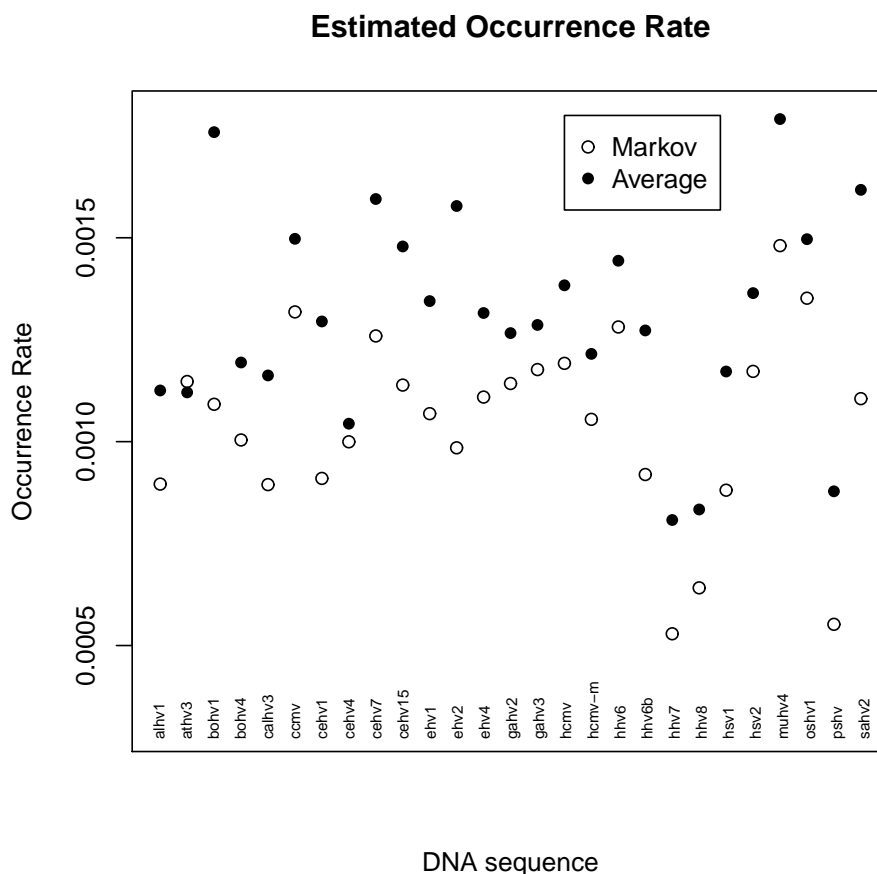


Figure 1: 27 herpes virus genomic sequences were downloaded from the EBI Nucleotide Sequences database. Two methods for estimating the null palindrome rates are presented, including the average rate and the Markov model based estimator. We used the abbreviation for naming the genome sequences that was used in Leung et al. (2005).

probabilities. Theorem 1 was then applied to estimate the null occurrence rate for each sequence. These results are compared with the average rate estimates in Figure 1. As shown, the average rate estimates have higher values except for the sequence athv3. Based on these two occurrence rate estimates and given the total length for each sequence, Theorem 4 is applied to derive the 0.05 significance thresholds for the scan statistics of PLS and BWS, shown in Figure 2. Both the PLS and the BWS scan scores of bohv1, hhv6 and hhv8 become significant at the 0.05 thresholds when the average rate estimator is replaced by the Markov rate estimate. These results suggest that the Markov rate estimate is potentially more powerful in detecting non-random clusters.

Figure 3 presents a more detailed analysis for bohv1. For both PLS and BWS plots, two peaks occurs at position 113,488 and 124,582, and they are close to two replication

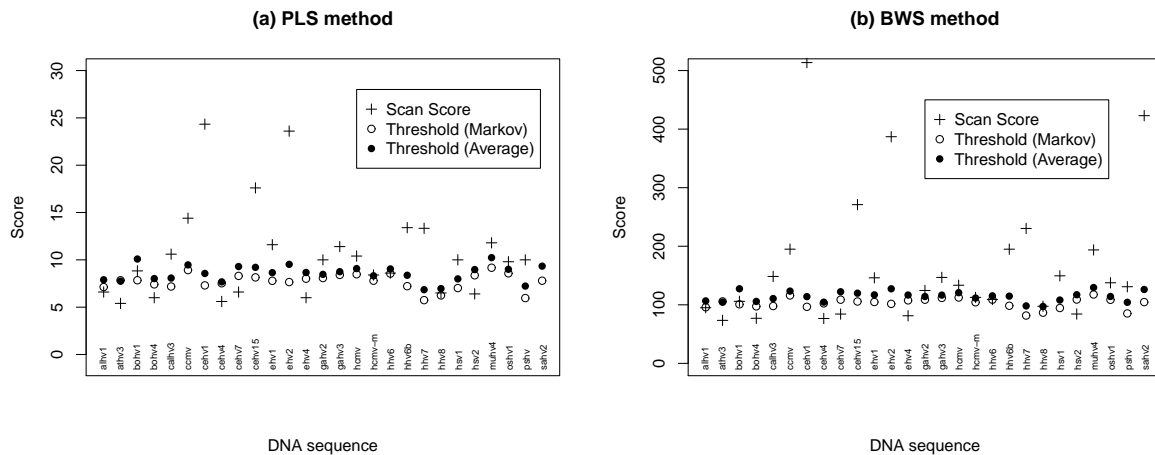


Figure 2: The thresholds based on the two occurrence rate estimates for the 27 herpes virus genomic sequences are compared. The solid circles label the thresholds derived by the average rates and the empty circles by the Markov model estimates. The crosses are the maximum scores for each sequence.

bohv1 Sequence	Scores of Two Peaks	Thresholds by Markov Model	Thresholds by Average Method
PLS	8.67	7.84	10.08
BWS	105.3	100.9	127.4

Table 1: Two peaks which are close to two replication origins of bohv1 happen to share the same scores for both PLS and BWS. Markov model estimate provides a more robust occurrence rate estimate against non-random clusters which leads to more appropriate threshold values and gains power.

origins respectively at positions 111,080–111,300 (OriS) and 126,918–127,138(OriS) (Lung et al. (2005)). The scores and the threshold values by both the Markov rate estimate and the average rate are summarized in Table 1. The scores are above the thresholds by the Markov rate and below that by the average rate. The analysis based on the Markov rate estimate is consistent with the hypothesis that non-random palindrome clusters may play a role to search the replication origins, proposed by Chew et al. (2007). The discrepancies between these two methods could be due to the non-random palindrome clusters of bohv1 which cause the inflation of the average rate. The Markov rate estimates are based on the letter frequencies and the pair frequencies of the whole genome in which the effect from those clusters becomes negligible.

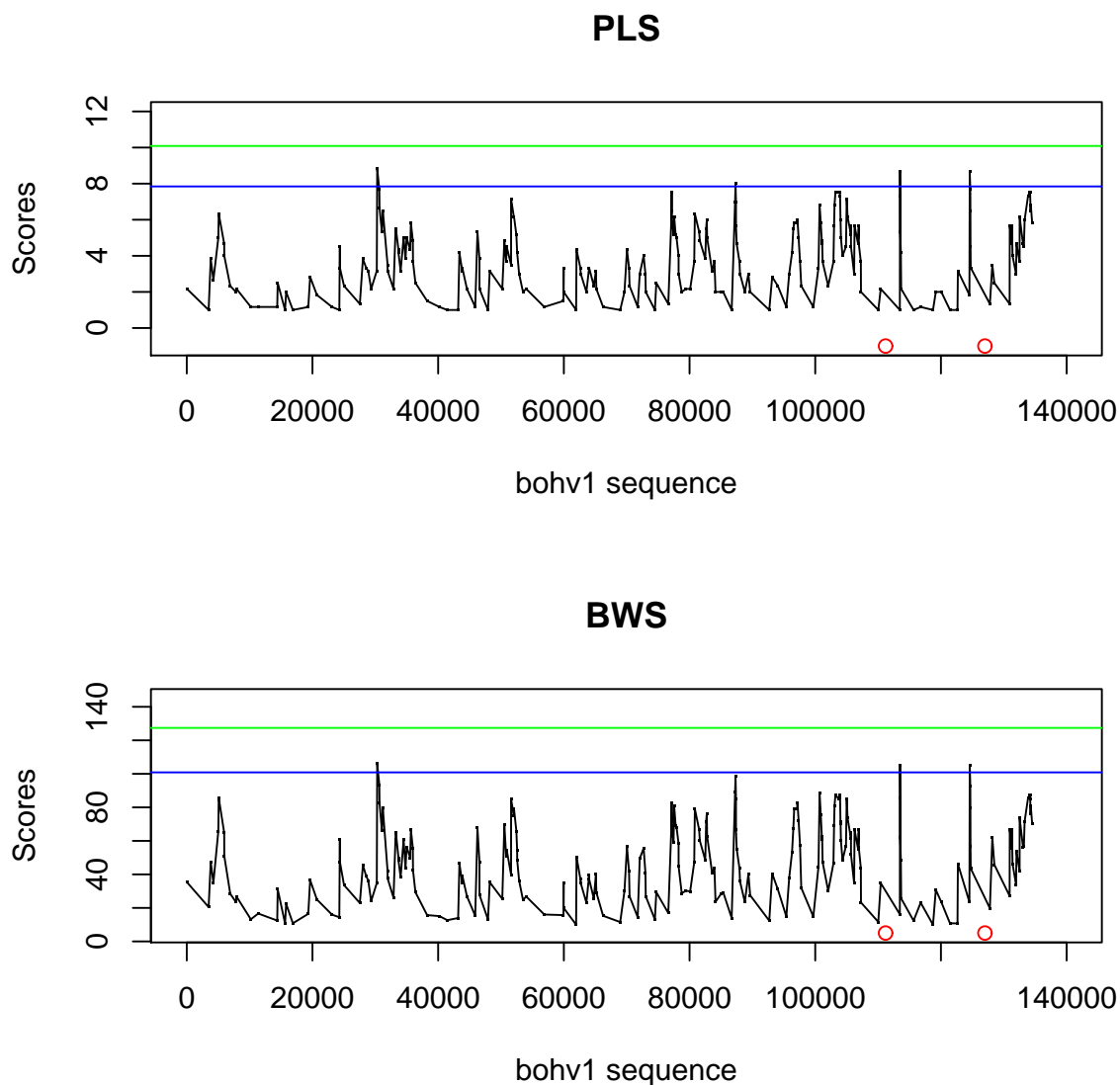


Figure 3: *The PLS and BWS scan scores are shown verse their genomic positions of bohv1. The blue horizontal lines are the thresholds by our occurrence rate estimate and the green lines are by the average rates. The red circles label the position of the replication origins of bohv1. Two peaks of PLS and BWS occurs at position 113,488 and 124,582, and they are close to two replication origins respectively at positions 111,080–111,300 (OriS) and 126,918–127,138 (OriS) (Leung et al. (2005)).*

No hot-spots	$\hat{\lambda}_i$	$\hat{\lambda}_M$	$\bar{\lambda}$
iid random sequences	0.00073	0.00073	0.00071
Markov random sequences	0.00073	0.00109	0.00101

Table 2: Three methods for estimating the palindrome occurrence rate are compared on both iid and Markov random sequences. If the model is correctly specified, both model calculation estimators are consistent with the mean rate. If the model is not correct, $\hat{\lambda}_M$ still performs well but $\hat{\lambda}_i$ does not.

3.2 Simulation Study

When there exists no non-random clusters, the palindrome events along these random sequences could be well approximated by a homogeneous Poisson process, for either a Markovian sequence or an iid sequence (Leung et al. (2005)). In this case, the average rate is the maximum likelihood estimator (MLE) for the occurrence rate and can be treated as a target reference. Table 2 shows that when iid random sequences are generated, all the three methods perform equally well. However, when Markov random sequences are generated, the iid model-based estimator 0.00073 falls below 27.7% of the target 0.00101. The reason is that the iid model is a sub-model of the Markov model while the reverse is not true. Thus, the Markov model is a better choice than the iid model.

We designed a simulation experiment to investigate the power performance of the estimates when hot-spot regions exist for a first-order Markov model. We used P_{bohV1} and π_{bohV1} to generate stationary random sequence of length 150,000 and then simulated the hot-spots by inserting various intensity of palindromes that are resampled from the bohV1 palindrome bank. The length distribution of this bank is shown in Table 3. We set five 1,000 bp regions with palindrome occurrence rates $(r_1, r_1, r_1, r_2, r_2) \times \hat{\lambda}_M$, where $\hat{\lambda}_M = 0.00109$ is the Markov rate estimate of the bohV1 sequence.

We chose $r_1 = 20$ to make the average rate close to that of bohV1 and let $r_2 \in \{6, 7, 8, 9, 10\}$ to check their power performance. The palindrome insertion for each hot-spot includes three steps: (a) to generate a Poisson random number \hat{M} with mean $1000 \hat{\lambda}_M r_i$; (b) to resample \hat{M} palindromes from the bohV1 palindrome bank; (c) to start at \hat{M} uniformly random positions inside the 1,000 bp segments, the DNA letters are replaced with the resampled palindromes. For each generated sequence, both Markov

half length	6	7	8	9	10	11	12
counts	132	54	22	10	12	4	4
half length	13	14	15	16	17	18	> 19
counts	1	0	0	1	0	1	0

Table 3: *The length distribution of the palindromes collected from bohv1. The genome length is 135,301 bp and the total number of palindromes (half length ≥ 6) is 241.*

rate and average rate are estimated and their corresponding threshold values are derived by Theorem 4. The comparisons between the two estimate methods on occurrence rates, threshold values and powers based on 250 replicas are presented in Table 4. The powers for those hot-spots with intensity $r_1 = 20$ -fold intensity reach one for both two methods of all cases and thus are omitted from the table.

When there exists no hot-spot ($r_1 = r_2 = 0$), both methods match very well on estimating the occurrence rate and thus share similar thresholds. When the three hot-spots are generated ($r_1 = 20$ and $r_2 = 0$), the average rate becomes 0.00151 (39% increase) while the Markov rate is 0.00111 (2% increase), resulting in threshold values 9.21 and 7.97 for PLS, and 122.40 and 102.34 for BWS. When $r_1 = 20$ and r_2 increases up to 10, the difference made by the Markov estimate is no more than 3%, while that by the average rate goes up to 54%. The PLS and BWS threshold values for our Markov method are virtually the same and that for the average method goes more than 23%. Evidently, our method gains more power.

We further applied a hidden Markov model (HMM) to generate segmented Markov DNA sequences. We first generated a three-state-Markov chain $S_t \in \{\xi_1, \xi_2, \xi_3\}$ whose stationary probabilities are (0.8, 0.1, 0.1) for $1 \leq t \leq 30$. Given each state $S_t = \xi_i$, we used its own transition matrix to generate the DNA letters of length 5,000. Thus, we generated a segmented Markovian DNA sequence of length 150,000 (Chen and Zhou (2010)). We used P_{bohv1} as the transition matrix for ξ_1 and added 0.05 on the second and the third column and deducted 0.05 from the first and the fourth columns of P_{bohv1} to generate the transition matrix for ξ_2 and then we exchanged the addition and deduction to generate that for ξ_3 . By doing so, ξ_2 has higher CG ratio and ξ_3 has lower CG ratio compared to that of bohv1 and the average transition matrix keeps the same as P_{bohv1} . The stationary probabilities and the occurrence rates for the three states are summarized

PLS									
r_1	r_2	Markov Model				Average Method			
		$\hat{\lambda}_M$	Threshold	Powers		$\bar{\lambda}$	Threshold	Powers	
0	0	0.00109	7.90	0.000	0.000	0.00109	7.88	0.000	0.000
20	0	0.00111	7.97	0.000	0.000	0.00151	9.21	0.000	0.000
20	6	0.00111	7.98	0.644	0.616	0.00161	9.53	0.480	0.464
20	7	0.00111	7.97	0.756	0.776	0.00164	9.63	0.636	0.628
20	8	0.00112	7.99	0.844	0.836	0.00165	9.67	0.708	0.716
20	9	0.00111	7.99	0.920	0.900	0.00166	9.74	0.840	0.800
20	10	0.00111	7.99	0.972	0.964	0.00168	9.79	0.912	0.916
BWS									
r_1	r_2	Markov Model				Average Method			
		$\hat{\lambda}_M$	Threshold	Powers		$\bar{\lambda}$	Threshold	Powers	
0	0	0.00109	101.65	0.000	0.000	0.00109	101.45	0.000	0.000
20	0	0.00111	102.22	0.000	0.000	0.00151	118.62	0.000	0.000
20	6	0.00111	102.34	0.644	0.608	0.00161	122.40	0.452	0.460
20	7	0.00111	102.28	0.736	0.768	0.00164	123.41	0.640	0.572
20	8	0.00112	102.41	0.856	0.828	0.00165	123.78	0.696	0.708
20	9	0.00111	102.39	0.912	0.884	0.00166	124.41	0.812	0.788
20	10	0.00111	102.39	0.972	0.956	0.00168	124.91	0.892	0.908

Table 4: Powers are compared for using $\hat{\lambda}_M$ and $\bar{\lambda}$ to estimate the null occurrence rates of DNA palindromes when five 1,000 bp hot-spots are inserted with relative intensity $(r_1, r_1, r_1, r_2, r_2)$ into a 150,000 bp DNA sequence generated by a Markov model. $r_1 = r_2 = 0$ is presented as the control group that no palindromes are inserted. Powers are defined as the frequencies of detecting the hot-spot based on 250 replicates. For the hot-spot region with $r_1 = 20$, the power reaches 1 and hence is not shown in the table. Only the powers for the two sites with r_2 intensity are shown. $\bar{\lambda}$ tends to overestimate the occurrence rates and constructs an overly conservative test, leading to power loss.

State	A	C	G	T	Occurrence Rate
ξ_1	0.1354	0.3588	0.3654	0.1405	0.00109
ξ_2	0.0833	0.4138	0.4162	0.0867	0.00310
ξ_3	0.1874	0.3038	0.3146	0.1941	0.00049

Table 5: *Three hidden states (ξ_1, ξ_2, ξ_3) are constructed to generate the segmented Markov model. The letter frequencies and the palindrome occurrence rates for the three states ξ_1, ξ_2 and ξ_3 are presented. ξ_2 has a higher CG ratio and ξ_3 has a lower one. The stationary probability distribution for (ξ_1, ξ_2, ξ_3) is (0.8, 0.1, 0.1).*

in Table 5.

In Table 6, we compare the power performance of the average method and the Markov method when the underlying sequence follows a segmented model. Table 6 has similar results as that in Table 4. The two methods, by the Markov model and the average rate, match very well when no hot-spot exists. When hot-spots constitute a significant portion of the total counts, the average rate is inflated to result in power loss and, by contrast, our Markov method is robust against the hot-spot effect and gains more power in detecting the non-random clusters.

In summary, both simulations of the first-order Markov model and the segmented model show that the average rate can overestimate the occurrence rate seriously due to the hot-spots effect and lead to power loss eventually. On the other hand, the Markov rate estimate is robust against the hot-spots and can maintain the threshold values appropriately and thus gains more power than the average rate method.

4 Discussion

In scan statistics, the average rate method is popular for estimating the null occurrence rate. In this paper, however, we report that the average rate method does not always work. The average rate can overestimate the null occurrence rate up to 50% above the actual number, because the hot-spots have the potential of contributing to a large portion of the number of events, especially when the null occurrence rate is very low. Thus, we propose an estimator based on a Markov model and define it as a function of the Markov parameters so we can estimate the Markov parameters by the letter frequencies and the adjacent pair frequencies without using the number of events. Therefore, as

PLS									
r_1	r_2	Markov Model				Average Method			
		$\hat{\lambda}_M$	Threshold	Powers		$\bar{\lambda}$	Threshold	Powers	
0	0	0.00109	7.91	0.000	0.000	0.00109	7.89	0.004	0.000
20	0	0.00111	7.96	0.000	0.000	0.0015	9.18	0.000	0.000
20	6	0.00111	7.97	0.632	0.656	0.00161	9.54	0.488	0.512
20	7	0.00111	7.98	0.768	0.764	0.00164	9.62	0.628	0.644
20	8	0.00111	7.98	0.896	0.868	0.00164	9.64	0.780	0.712
20	9	0.00111	7.99	0.940	0.936	0.00166	9.70	0.872	0.836
20	10	0.00112	7.99	0.948	0.924	0.00167	9.78	0.884	0.880

BWS									
r_1	r_2	Markov Model				Average Method			
		$\hat{\lambda}_M$	Threshold	Powers		$\bar{\lambda}$	Threshold	Powers	
0	0	0.00109	101.74	0.000	0.000	0.00109	101.54	0.000	0.000
20	0	0.00111	102.17	0.000	0.000	0.00150	118.34	0.000	0.000
20	6	0.00111	102.27	0.636	0.640	0.00161	122.57	0.464	0.488
20	7	0.00111	102.32	0.744	0.764	0.00164	123.36	0.584	0.600
20	8	0.00111	102.33	0.876	0.848	0.00164	123.53	0.776	0.728
20	9	0.00111	102.39	0.944	0.924	0.00166	124.18	0.860	0.804
20	10	0.00112	102.4	0.932	0.916	0.00167	124.72	0.888	0.872

Table 6: Powers are compared for using $\hat{\lambda}_M$ and $\bar{\lambda}$ to estimate the null occurrence rates of DNA palindromes when five 1,000 bp hot-spots are inserted with relative intensity $(r_1, r_1, r_1, r_2, r_2)$ into a 150,000 bp DNA sequence generated by a three-state segmented Markov model. The parameters of the three states are in Table 5. $r_1 = r_2 = 0$ is presented as the control group that no palindromes are inserted. Powers are defined as the frequencies of detecting the hot-spot based on 250 replicates. For the hot-spot region with $r_1 = 20$, the power reaches 1 and hence is not shown in the table. Only the powers for the two sites with r_2 intensity are shown. $\bar{\lambda}$ tends to overestimate the occurrence rates and constructs an overly conservative test, leading to power loss.

long as the size of the hot-spot regions is much smaller than the total length of the genomes, the estimated Markov parameters would have little influence on the presence of the hot-spots, rendering our method insensitive to the hot-spot effect. Our study suggests that a model based estimator might be more appropriate than the average rate for null occurrence rate estimation, especially when the Poisson process involves rare events with hot-spot regions, which are quite common in epidemiology studies involving rare diseases.

References

- [1] Chan, H.P. and Zhang, N.R. (2007). Scan statistics with weighted observations. *Journal of the American Statistical Association* **102**, 595–602.
- [2] Chen, G. and Zhou, Q. (2010). Heterogeneity in DNA multiple alignments: Modeling, inference, and applications in motif finding. *Biometrics* **66**, 694–704.
- [3] Chew, D., Cho, K., and Leung, M. (2005). Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. *Nucleic Acids Research* **33**, 134.
- [4] FitzGerald, P., Shlyakhtenko, A., Mir, A., and Vinson, C. (2004). Clustering of DNA sequences in human promoters. *Genome Research* **14**, 1562–1574
- [5] Hansen, N.R. (2009). Statistical models for local occurrences of RNA structures. *J. Computational Biology* **16**, 845–858.
- [6] Leach, D.(1994). Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays* **16**, 893–900.
- [7] Leung, M.Y., Choi, K.P., Xia, A., and Chen, L.H.Y. (2005). Nonrandom clusters of palindromes in Herpesvirus genomes. *J. Computational Biology* **12**, 331–354.
- [8] Lisnic B., Svetec I.K., Saric H., Nikolic I., and Zgaga Z. (2005). Palindrome content of the yeast *Saccharomyces cerevisiae* genome. *Current Genetics* **47**, 289–97
- [9] Lu, L., Jia, H., Droge, P., and Li, J. (2007). The human genome-wide distribution of DNA palindromes. *Functional Integrative Genomics* **7**, 221–227.

- [10] Reinert, G. and Schbath, S.(1998). Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Computational Biology* **5**, 223–253.
- [11] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.
- [12] Tu, I-P. and Siegmund, D. (1999). The maximum of a function of a Markov chain and application to linkage analysis. *Advances in Applied Probability* **31**, 510–531.
- [13] Tu, I-P. (2009). Asymptotic overshoots for arithmetic i.i.d. random variables. *Statistica Sinica* **19**, 315–323.
- [14] Tu, I-P. (2012). The maximum of a ratchet scanning process over a Poisson random field. *to appear in Statistica Sinica*.
- [15] Woodroffe, M. (1979). Repeated likelihood ratio tests. *Biometrika* **66**, 454–463.