Nonrandom Clusters of Palindromes in Herpesvirus Genomes

MING-YING LEUNG,[1] KWOK PUI CHOI,[2] AIHUA XIA,[3] and LOUIS H.Y. CHEN[4]

ABSTRACT

Palindromes are symmetrical words of DNA in the sense that they read exactly the same as their reverse complementary sequences. Representing the occurrences of palindromes in a DNA molecule as points on the unit interval, the scan statistics can be used to identify regions of unusually high concentration of palindromes. These regions have been associated with the replication origins on a few herpesviruses in previous studies. However, the use of scan statistics requires the assumption that the points representing the palindromes are independently and uniformly distributed on the unit interval. In this paper, we provide a mathematical basis for this assumption by showing that in randomly generated DNA sequences, the occurrences of palindromes can be approximated by a Poisson process. An easily computable upper bound on the Wasserstein distance between the palindrome process and the Poisson process is obtained. This bound is then used as a guide to choose an optimal palindrome length in the analysis of a collection of 16 herpesvirus genomes. Regions harboring significant palindrome clusters are identified and compared to known locations of replication origins. This analysis brings out a few interesting extensions of the scan statistics that can help formulate an algorithm for more accurate prediction of replication origins.

Key words: DNA palindromes, Wasserstein distance, Poisson process approximation, scan statistics, replication origins.

1. INTRODUCTION

DNA PALINDROMES ARE WORDS FROM THE NUCLEOTIDE BASE ALPHABET $\mathcal{A} = \{A, C, G, T\}$ that are symmetrical in the sense that they read exactly the same as their complementary sequences in the reverse direction (see Fig. 1(a)). A DNA palindrome is necessarily even in length because the middle base in any odd-length nucleotide string cannot be identical to its complement. Palindromes are involved in a variety of biological processes. For example, the recognition sites for bacterial restriction enzymes to cut foreign DNA are mostly palindromic (Waterman, 1995, Chapter 2). Palindromes also play important roles

---

[1]Department of Mathematical Sciences, University of Texas at El Paso, El Paso, TX 79968-0514.
[2]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117543.
[3]Department of Mathematics and Statistics, University of Melbourne, VIC 3010, Australia.
[4]Institute for Mathematical Sciences, National University of Singapore, Singapore 118402.
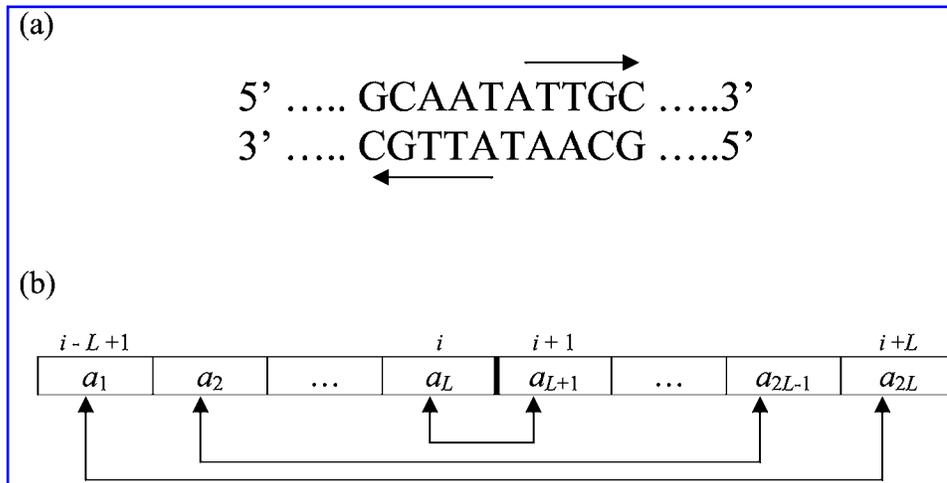
(a)

$$5' \ldots.. \text{GCAATATTGC} \ldots..3'$$
$$3' \ldots.. \text{CGTTATAACG} \ldots..5'$$

(b)

$i-L+1$			i	$i+1$			$i+L$
a_1	a_2	...	a_L	a_{L+1}	...	a_{2L-1}	a_{2L}

FIG. 1. DNA palindrome. (**a**) This is a palindromic nucleotide sequence on the two complementary strands of DNA which are always read in opposite directions from the $5'$ end to the $3'$ end as shown by the arrows. The displayed segment reads exactly the same on both strands. (**b**) On each strand, the first base of the palindrome is complementary to the last, the second to the second last, and so on. This is a schematic representation of such complementary pairing between the bases in a $2L$-palindrome centered at base i.

in gene regulation and DNA replication processes (Wagner, 1991, Chapters 6, 12, 18; Kornberg and Baker, 1992, Chapter 1). It appears that palindromes have to do with DNA–protein binding. The local two-fold symmetry created by the palindrome provides a binding site for DNA-binding proteins which are often dimeric in structure. Such double binding markedly increases the strength and specificity of the binding interaction (Creighton, 1993, Chapter 8).

The herpesvirus family includes some of the well-known pathogenic viruses such as herpes simplex, varicella-zoster, Epstein-Barr, and cytomegalovirus. Some of these viruses are believed to pose major risks in immunosuppressive posttransplantation therapies, while others have been associated with life-threatening diseases such as AIDS and various cancers (Bennett *et al.*, 2001; Biswas *et al.*, 2001; Labrecque *et al.*, 1995; Vital *et al.*, 1995). A number of the animal herpesviruses are of agricultural concern. For example, the alcelaphine herpesvirus 1, indigenous to the wildebeest, is a causative agent of the fatal lymphoproliferative disease malignant catarrhal fever in cattle and deer (Bridgen, 1991).

Replication origins are places on the DNA molecules where replication processes are initiated. As DNA replication is the central step in the reproduction of many viruses, understanding the molecular mechanisms involved in DNA replication is of great importance in developing strategies to control the growth and spread of viruses (Delecluse and Hammerschmidt, 2000). For Epstein-Barr virus, one of these replication origins has been shown to associate with cellular proteins that regulate the initiation of DNA synthesis in human cells (Sugden, 2002). This suggests that these replication origins are also important locations for studying possible mechanisms of infecting human host cells. Knowledge of the locations of these replication origins will enhance the development of antiviral agents by blocking viral DNA replication or by interfering with the infection process.

As replication origins in DNA are considered major sites for regulating genome replication in general, labor-intensive laboratory procedures have been used to search for replication origins in various organisms (e.g., see Hamzeh, 1990; Zhu, 1998; Newlon and Theis, 2002). With the increasing availability of genomic DNA sequence data, the value of using computational methods to predict likely locations of replication origins before the experimental search has already been recognized although no prediction scheme that works for all DNA in general is available to date. The success of the computational prediction depends critically on the observation of the characterizing patterns in the nucleotide sequence around the replication origins of the particular kind of organisms under study. For example, the algorithm of Salzberg *et al.* (1998) predicted the replication origins for a number of bacterial and archaeal genomes based on the finding of seven-base and eight-base oligomers whose orientation is preferentially skewed around the replication origins. However, as pointed out by the authors, this algorithm is not suited for DNA molecules, like those

in many viruses and their eukaryotic hosts, where multiple replication origins exist. In those cases, one would need to rely on other relevant sequence patterns to locate the replication origins.

The existence of high concentrations of palindromes in proximity of the replication origins of herpesviruses has been reported in some early studies (Weller *et al.*, 1985; Reisman *et al.*, 1985; Masse *et al.*, 1992). This phenomenon is generally attributed to the fact that initiation of DNA replication typically requires an assembly of enzymes such as the helicases to locally unwind the helical structure of DNA and pull apart the two complementary strands. Furthermore, Masse *et al.* (1992) have demonstrated that by looking for palindrome clusters, among other features such as clusters of close repeats and close inversions on the nucleotide sequence, likely regions containing replication origins can be predicted.

Leung *et al.* (1994) describe how a statistical criterion, based on the scan statistics (Glaz, 1989; Dembo and Karlin, 1992), is developed for identifying nonrandom palindrome clusters by modeling the occurrences of palindromes in the genome as points randomly sampled from the unit interval according to the uniform distribution. Despite the fact that the criterion worked well for the cytomegalovirus genome sequence used for illustration in that article, the authors point out that the assumption of uniform distribution of palindromes has not yet been mathematically justified. In this paper, we shall justify this claim under the model that the nucleotide sequence is generated as a sequence of independent and identically distributed (i.i.d.) random variables.

The second, and more important, aim of this paper is to analyze a collection of herpesvirus genome sequences for nonrandom palindrome clusters and examine their connections with replication origins. Table 1 presents the collection of herpesvirus genomes to be analyzed. The dataset comprises all the complete genome sequences of the herpesvirus family downloaded from GenBank at the NCBI website in June 2001. Listed along with each virus name in the table are an abbreviation that will be used throughout this paper, its accession number in the GenBank database, its genome sequence length in number of bases, and the relative frequencies of the four nucleotide bases in the genome. The experimentally confirmed replication origins among the viruses in this dataset will help us assess the palindrome-based algorithm and suggest directions for improving the rate of successful prediction.

It would be of interest to ask whether our algorithm, developed for the herpesviruses, can be applied to identify replication origins in other organisms, or even to identify other functionally important regions, such as regulatory sites. While one would not anticipate that palindrome clusters will be the universal characterizing sequence pattern for all kinds of organisms and all types of functional sites, existence of palindrome clusters may serve as one possible criterion for these general purposes. For example, we have already noted that this approach adds to the skewed-oligomers method described by Salzberg *et al.* (1998) because it is not limited to DNA molecules with single replication origins. It is our hope that it will contribute to a general prediction tool that can be broadly applied to various domains of the tree of life.

TABLE 1.   THE LIST OF HERPESVIRUS GENOMES TO BE ANALYZED

Name	Abbrev.	Accession	Length	Base composition
Alcelaphine herpesvirus 1	AHV1	NC_002531	130608	(.27, .24, .22, .26)
Ateline herpesvirus 3	AtHV3	NC_001987	108409	(.32, .19, .17, .31)
Bovine herpesvirus 1.1	BHV1	NC_001847	135301	(.14, .36, .37, .14)
Equine herpesvirus 1	EHV1	NC_001491	150223	(.22, .29, .28, .22)
Equine herpesvirus 4	EHV4	NC_001844	145597	(.25, .25, .25, .25)
Gallid herpesvirus 1	MDV2	NC_002530	110637	(.24, .25, .25, .25)
Gallid herpesvirus 2	MDV	NC_002229	138675	(.28, .22, .21, .29)
Human herpesvirus 1	HSV1	NC_001806	152261	(.16, .34, .34, .16)
Human herpesvirus 2	HSV2	NC_001798	154746	(.15, .35, .35, .15)
Human herpesvirus 3	VZV	NC_001348	124884	(.27, .23, .23, .27)
Human herpesvirus 4	EBV	NC_001345	172281	(.20, .30, .29, .20)
Human herpesvirus 5	HCMV	NC_001347	229354	(.22, .28, .29, .21)
Human herpesvirus 6	HHV6	NC_001664	159321	(.29, .22, .21, .29)
Human herpesvirus 7	HHV7	NC_001716	144861	(.32, .18, .17, .32)
Ictalurid herpesvirus	CCV1	NC_001493	134226	(.21, .28, .28, .22)
Saimiriine herpesvirus 2	HVS2	NC_001350	112930	(.33, .18, .16, .32)

The organization of the paper is as follows: Section 2 formulates the random process representing the palindrome occurrences on a nucleotide sequence and introduces the Wasserstein distance for measuring the difference between the palindrome process and the Poisson process. Using a general mathematical Poisson process approximation theorem, we derive an explicitly computable upper bound for this distance which approaches zero under suitable conditions. Section 3 briefly reviews how the scan statistics are used to identify nonrandom clusters of palindromes, treating them as points randomly sampled from a uniform distribution. The significant palindrome clusters obtained from the herpesviruses are presented in Section 4 where their association with replication origins is also discussed. We conclude with a few remarks about future works towards a more accurate replication origin prediction scheme in Section 5.

2. DISTRIBUTION OF PALINDROMES ON RANDOM DNA SEQUENCES

In this section, we shall see that if a DNA genome is assumed to be a sequence of nucleotide bases generated as i.i.d. random variables taking values A, C, G, T with probabilities p_A, p_C, p_G, p_T, respectively, the occurrences of palindromes above a certain minimal length can be approximated by a Poisson process. This is achieved by deriving an upper bound for the Wasserstein distance, also called the d_2 metric (Barbour *et al.*, 1992, Chapter 10), between the palindrome process and the Poisson process. Under suitable conditions, this Wasserstein distance dwindles to 0 as the sequence length increases indefinitely. Before these results can be stated, we need to first make precise the notion of the palindrome process and explain the concept of the Wasserstein distance.

2.1. The palindrome process and Wasserstein distance

Due to the complementary base pairing, it is often sufficient to represent DNA as a single nucleotide sequence. Any segment of the nucleotide sequence consisting of $2L$ bases will be a palindrome if its first base and its $2L$th base form a complementary pair, and so do its second and $(2L - 1)$st bases, the third and $(2L - 2)$nd, ..., up to the Lth and $(L + 1)$st bases in the center of the segment (Fig. 1(b)). Using the center of the palindrome to indicate its position, we say that a palindrome of length $2L$ occurs at position i of the sequence if the bases $i - j + 1$ and $i + j$ are complementary to each other for $j = 1, \ldots, L$. This characterization does not preclude the possibility that a palindrome of length $2L$ may be extended to a longer length if the complementary pairing continues on for $j > L$. In this paper, it will be understood that the term "palindrome of length $2L$" actually means "palindrome of length $2L$ or more." For short, we shall just refer to it as a $2L$-palindrome. If there are multiple $2L$-palindromes centered at the same position of the sequence, only the longest one will be counted.

Because it is impossible for a nucleotide sequence of length M to have any $2L$-palindrome centered at positions $1, \ldots, L - 1$ and $M - L + 1, \ldots, M$, we shall represent the occurrences of palindromes as a random process on $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$, where $n = M - 2L + 1$. The palindrome process is defined as

$$\Xi = \sum_{i=1}^{n} I_i \delta_{i/n}.$$

Here I_i is the indicator random variable for the occurrence of a $2L$-palindrome centered at base $i + L - 1$ of the DNA sequence and $\delta_{i/n}$ denotes the unit point mass at i/n. In this definition, the DNA segment H of length $2L$ spanning bases $i, \ldots, i + L - 1, i + L, \ldots, i + 2L - 1$ is associated with the indicator I_i and the unit point mass at $\frac{i}{n}$. For brevity, we shall say that the DNA segment H is positioned at i for the rest of the paper.

In an i.i.d. random nucleotide sequence, the success probability for the random variable I_i is

$$p_i = P(I_i = 1) = 1 - P(I_i = 0) = \theta^L, \tag{1}$$

with $\theta = 2(p_A p_T + p_C p_G)$, and the expected number of palindromes is

$$\lambda = \sum_{i=1}^{n} p_i = n\theta^L. \tag{2}$$

Because palindromes occurring close to each other overlap, the I_i's are locally dependent. The neighborhood of dependence of I_i is

$$A_i = \{1 \le j \le n : |j - i| < 2L\}. \tag{3}$$

For all j outside of A_i, I_i and I_j are independent of each other.

We want to approximate the palindrome process Ξ defined on the n equally spaced discrete points $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}$ in $[0, 1]$ by the Poisson process Z_λ with intensity λ on the continuous interval $[0, 1]$. In a number of DNA related studies (e.g., Arratia $et\ al.$, 1990, 1996; Reinert $et\ al.$, 2000; Reinert and Schbath, 1998; Schbath, 1995), the differences between two random processes are quantified by the total variation distance d_{TV} between them. However, as explained in Chapter 10 of Barbour $et\ al.$ (1992), the total variation distance is too strong to be useful for point processes like Ξ and Z_λ. Indeed,

$$d_{TV}(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) = \sup_A |P(\Xi \in A) - P(Z_\lambda \in A)|$$

where the supremum is taken over all measurable subsets[1] of $\mathcal{Y}$ where

$$\mathcal{Y} = \left\{ \sum_{i=1}^n \delta_{x_i} : n \ge 0, x_i \in [0, 1] \right\}$$

is often called the configuration space of $[0, 1]$. From this, it can be shown that

$$d_{TV}(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) = 1/2 \sup_{-1 \le f \le 1} |Ef(\Xi) - Ef(Z_\lambda)|$$

which always equals 1 because Ξ has support only in $\{i/n : 1 \le i \le n\}$ whereas Z_λ has no points in $\{i/n : 1 \le i \le n\}$ with probability 1.

Unlike d_{TV}, the Wasserstein distance, also called the d_2 metric, is less sensitive to small changes in the positions of points. It is more suitable for our purpose of measuring the discrepancy between the palindrome process and the Poisson process. Essentially, the Wasserstein distance between two point processes X and Y is the supremum of all expected differences between X and Y under test functions f that do not fluctuate too vigorously on the configuration space of $[0, 1]$ with respect to the d_1 metric explained below. These test functions are called Lipschitz functions. We give a concise explanation of the Wasserstein distance below. A more detailed description can be found in Barbour $et\ al.$ (1992, Chapter 10, Section 2).

The Wasserstein distance between two point processes X and Y is defined as

$$d_2(\mathcal{L}(X), \mathcal{L}(Y)) = \sup\{|Ef(X) - Ef(Y)| : ||f||_{Lip(\mathcal{Y})} \le 1\}$$

where $f : \mathcal{Y} \longrightarrow \mathbf{R}$ and

$$||f||_{Lip(\mathcal{Y})} = \sup \left\{ \frac{|f(\xi_1) - f(\xi_2)|}{d_1(\xi_1, \xi_2)} : \xi_1 \ne \xi_2 \in \mathcal{Y} \right\}.$$

Here, $d_1(\xi_1, \xi_2)$ denotes the distance between two configurations $\xi_1 = (y_{11}, \ldots, y_{1m_1})$ and $\xi_2 = (y_{21}, \ldots, y_{2m_2})$. It is defined to be 1 if ξ_1 and ξ_2 have different numbers of points in $[0, 1]$ and is defined to be the average distance between ξ_1 and ξ_2 under the closest matching if they have the same number of points. More precisely,

$$d_1(\xi_1, \xi_2) = \begin{cases} 1 & \text{if } m_1 \ne m_2 \\ \min\{1/m \sum_{i=1}^m |y_{1i} - y_{2\pi(i)}|\} & \text{if } m_1 = m_2 \end{cases}$$

---

[1] A subset A of $\mathcal{Y}$ is measurable if $A \in \mathcal{B}$ where $\mathcal{B}$ is the smallest sigma algebra making the mappings from $\mathcal{Y}$ to $\mathbf{R}: \xi \mapsto \xi(C)$ measurable for all Borel sets $C \subset [0, 1]$.

where the minimum is taken over all permutations π of $(1, 2, \ldots, m)$ with m being the common value of m_1 and m_2. Since our point processes are defined on $[0, 1]$, without loss of generality, we can assume that $y_{11} \leq y_{12} \leq \ldots \leq y_{1m_1}$ and $y_{21} \leq y_{22} \leq \ldots \leq y_{2m_2}$. Then, when $m_1 = m_2 = m$,

$$d_1(\xi_1, \xi_2) = \frac{1}{m} \sum_{i=1}^{m} |y_{1i} - y_{2i}|.$$

The above simplified expression for $d_1(\xi_1, \xi_2)$ will be formally proved as Proposition 3 in the appendix.

With this background, we present the following general theorem giving an upper bound for the Wasserstein distance between a point process Ξ and the Poisson process Z_λ. We will use the following notations. Let $I_1, \ldots, I_n$ be indicator random variables with $P(I_i = 1) = 1 - P(I_i = 0) = p_i$, and $p_{ij} = P(I_i = I_j = 1)$ for $1 \leq i, j \leq n$. For each i, I_i has a neighborhood of dependence A_i which is a collection of those indices j such that I_j may be dependent on I_i. Define a random point process

$$\Xi = \sum_{i=1}^{n} I_i \delta_{i/n}$$

where $\delta_{i/n}$ denotes the unit point mass at i/n. We also define the point processes

$$V_i = \sum_{j \notin A_i} I_j \delta_{j/n} \quad \text{and} \quad V_{ij} = \sum_{k \notin A_i \cup A_j} I_k \delta_{k/n}.$$

We let

$$|V_i| = \sum_{j \notin A_i} I_j \quad \text{and} \quad |V_{ij}| = \sum_{k \notin A_i \cup A_j} I_k$$

denote the number of points in these processes.

Theorem 1. *Let Z_λ denote the Poisson process on $[0, 1]$ with intensity $\lambda = \sum_{i=1}^{n} p_i$. Assume $(I_j, j \notin A_i)$ is independent of I_i and $(I_k, k \notin A_i \cup A_j)$ is independent of (I_i, I_j) for all i, j. Then we have*

$$d_2(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) \leq \left[\frac{5}{\lambda} + 3 \sup_{\substack{1 \leq i \leq n \\ j \in A_i \setminus \{i\}}} E\left(\frac{1}{|V_{ij}| + 1} \right) \right] (b_1 + b_2) + \frac{1}{2n} \tag{4}$$

where

$$b_1 = \sum_{i=1}^{n} \sum_{j \in A_i} p_i p_j$$

is the sum of products of the probabilities of observing $2L$-palindromes at positions i and j within the neighborhood of dependence of each other and

$$b_2 = \sum_{i=1}^{n} \sum_{j \in A_i \setminus \{i\}} p_{ij}$$

is the sum of probabilities of observing overlapping $2L$-palindromes at both positions i and $j (\neq i)$ within the neighborhood of dependence of each other.

The proof of this theorem, given in the appendix, is based on an adaptation of Stein's method (1972) for the Poisson process setting. We shall now examine how to use Theorem 1 to obtain a Poisson limit for the palindrome process. With long sequences, n is large, making the last term in the above bound negligible. We have noted in Equations (1) and (2) that for any i, $p_i = \theta^L$, giving $\lambda = n\theta^L$. It is also easy to see

from (3) that in the palindrome process Ξ, the neighborhood of dependence of I_i can stretch in either the forward or backward direction from base i for at most only $2L - 1$ bases. This implies

$$b_1 \leq n(4L - 1)\theta^{2L}. \tag{5}$$

To obtain an explicit upper bound for the Wasserstein distance between Ξ and Z_λ, two things remain to be done. First, we need to examine the b_2 term which involves the probabilities of overlapping palindromes. Second, we need to work out a uniform upper bound for $E\left(\frac{1}{|V_{ij}|+1}\right)$ which is independent of i, j.

The b_2 term comprises the probabilities p_{ij} of observing overlapping $2L$-palindromes at positions i and j that are no more than $2L - 1$ bases apart. For example, the sequence ATCGATCG contains a 6-palindrome centered at position $i = 3$ overlapping with another 6-palindrome centered at position $j = 5$. The following proposition expresses the overlapping probability p_{ij} in terms of the base probabilities p_A, p_C, p_G, and p_T, the length parameter L, and the distance $h = |i - j|$ between the centers of the two palindromes.

Proposition 1 (probability of overlapping palindromes). *Let $h = |i - j| > 0$. If $h < 2L$, the probability p_{ij} of having overlapping palindromes at both positions i and j are given by the following.*
Case (a): $L \leq h \leq 2L - 1$. We have

$$p_{ij} = \theta^{2(h-L)}[p_A p_T(p_A + p_T) + p_C p_G(p_C + p_G)]^{2L-h}.$$

Case (b): $0 < h < L$. Here we let $L = qh + r$ and consider two subcases according to how big the remainder r is in relation to h.
Subcase (b1): $0 \leq r < (h + 1)/2$.

$$p_{ij} = [2((p_A p_T)^{q+1} + (p_C p_G)^{q+1})]^{2r}$$
$$\cdot [(p_A p_T)^q(p_A + p_T) + (p_C p_G)^q(p_C + p_G)]^{h-2r}.$$

Subcase (b2): $(h + 1)/2 \leq r < h$.

$$p_{ij} = [2((p_A p_T)^{q+1} + (p_C p_G)^{q+1})]^{2(h-r)}$$
$$\cdot [(p_A p_T)^{q+1}(p_A + p_T) + (p_C p_G)^{q+1}(p_C + p_G)]^{2r-h}.$$

Proof. Without loss of generality, we can assume $i < j < 2L + i$. Let H_i and H_j denote the DNA segments of length $2L$ positioned at i and j, respectively. Because of their overlap, H_i and H_j must share a common subsegment of length $2L - h$ at the right end of H_i and the left end of H_j. Throughout the proof, we shall use a' to denote the complement of base a (e.g., $A' = T$), $\mathbf{w}'$ to denote the inverse complement of word $\mathbf{w}$, and $P(\mathbf{w})$ to denote the probability for $\mathbf{w}$ (e.g., if $\mathbf{w} = (ATC)$, then $P(\mathbf{w}) = p_A p_T p_C$). We discuss the cases separately.

Case (a): Let $\mathbf{w} = (a_1, \ldots, a_{2L-h})$ denote the common subsegment. For both H_i and H_j to be palindromes, we must have the arrangement as shown in Fig. 2(a). At the left end of H_i and the right end of H_j, the sequence must be $\mathbf{w}'$. The center portions of H_i and H_j must be $2(h - L)$-palindromes denoted, respectively, by $(\mathbf{u}', \mathbf{u})$ and $(\mathbf{v}, \mathbf{v}')$. The probability of this arrangement is

$$p_{ij} = \theta^{2(h-L)} \sum_{\mathbf{w}} P(\mathbf{w})P(\mathbf{w}')^2$$

where the sum is taken over all possible DNA words $\mathbf{w}$ of length $2L - h$. Writing out $P(\mathbf{w})$ in terms of the base probabilities, we have

$$p_{ij} = \theta^{2(h-L)} \sum_{a_1,\ldots,a_{2L-h}\in\mathcal{A}} (p_{a_1} p_{a_1'}^2)\cdots(p_{a_{2L-h}} p_{a_{2L-h}'}^2)$$

$$= \theta^{2(h-L)}\left[\sum_{a\in\mathcal{A}} p_a p_{a'}^2\right]^{2L-h}$$

$$= \theta^{2(h-L)}[p_A p_T(p_A + p_T) + p_C p_G(p_C + p_G)]^{2L-h}$$

where a is summed over the four bases in the alphabet $\mathcal{A}$.
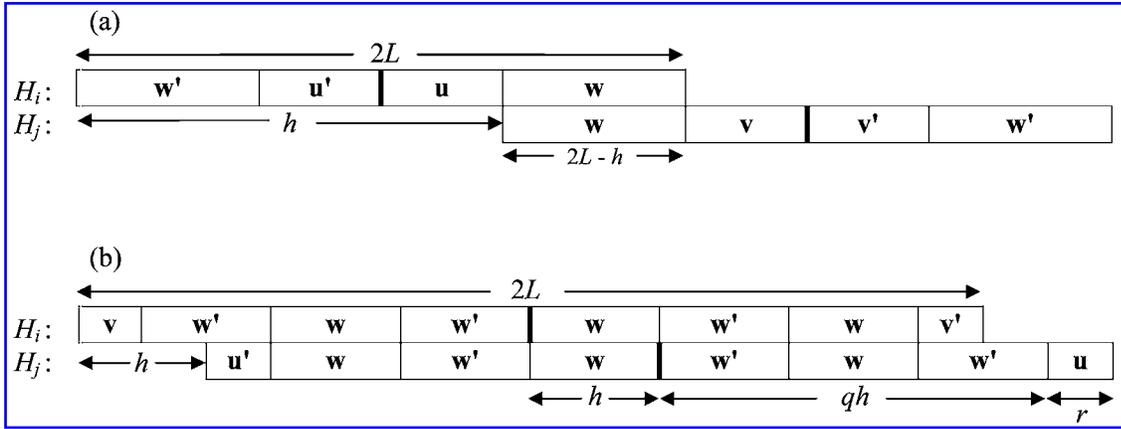
FIG. 2. Overlapping palindromes. (**a**) When $L \leq h \leq 2L - 1$, the DNA segments H_i and H_j must have this structure for both of them to be $2L$-palindromes centered at positions i and j, respectively. The center position of each segment is represented by a bold vertical bar. (**b**) Similarly, this displays the structures of H_i and H_j for both of them to be palindromes when $0 < h < L$. In this illustration, $q = 3$.

Case (b): This time, let $\mathbf{w} = (a_1, \ldots, a_h)$ denote the first h bases to the right of the center of H_i and to the left of the center of H_j. Let $\mathbf{u} = (a_1, \ldots, a_r)$ and $\mathbf{v} = (a_{h-r+1}, \ldots, a_h)$, respectively, stand for the first and last r bases of $\mathbf{w}$. Figure 2(b) displays the necessary structure in H_i and H_j for both of them to be palindromes when $q = 3$. In general, the probability of such arrangements can be computed as

$$
p_{ij} = \begin{cases} \sum_{\mathbf{w}} [P(\mathbf{w})]^q [P(\mathbf{w}')]^{q+1} P(\mathbf{u}) P(\mathbf{v}) & \text{when } q \text{ is odd,} \\[2ex] \sum_{\mathbf{w}} [P(\mathbf{w}')]^q [P(\mathbf{w})]^{q+1} P(\mathbf{u}') P(\mathbf{v}') & \text{when } q \text{ is even.} \end{cases}
$$

Subcase (b1): $0 \leq r < \frac{h+1}{2}$. If q is odd,

$$
p_{ij} = \left[\sum_{a \in \mathcal{A}} (p_a p_{a'})^{q+1} \right]^{2r} \left[\sum_{a \in \mathcal{A}} p_{a'} (p_a p_{a'})^q \right]^{h-2r}
$$

$$
= [2((p_A p_T)^{q+1} + (p_C p_G)^{q+1})]^{2r} [(p_A p_T)^q (p_A + p_T) + (p_C p_G)^q (p_C + p_G)]^{h-2r}.
$$

If q is even, the calculation is exactly the same except a and a' need to be swapped, reducing to the same expression in terms of the base probabilities.

Subcase (b2): $\frac{h+1}{2} \leq r < h$.

$$
p_{ij} = \left[\sum_{a \in \mathcal{A}} (p_a p_{a'})^{q+1} \right]^{2(h-r)} \left[\sum_{a \in \mathcal{A}} p_a (p_a p_{a'})^{q+1} \right]^{2r-h}
$$

when q is odd. Just like subcase (b1), a and a' need to be swapped when q is even. Either way, the expression reduces to

$$
[2((p_A p_T)^{q+1} + (p_C p_G)^{q+1})]^{2(h-r)} [(p_A p_T)^{q+1} (p_A + p_T) + (p_C p_G)^{q+1} (p_C + p_G)]^{2r-h}. \qquad \blacksquare
$$

The following two lemmas will facilitate proving that the palindrome process approaches a Poisson limit under the equal complementary probability (ECP) assumption of $p_A = p_T$ and $p_C = p_G$.

Lemma 1. *Under the ECP assumption, $p_{ij} \leq \theta^{3L/2}$ for $1 \leq i \neq j \leq n$.*

Proof. When $j \notin A_i$, the inequality is trivially true because I_i and I_j are independent and $p_{ij} = p_i p_j = \theta^{2L}$. We therefore only need to look at those p_{ij}'s with $|j - i| < 2L$. The calculations needed involve expressions of the form $(0.5 - x)^l + x^l$ where $x = p_G = p_C$ is a number between 0 and 0.5 and l is a positive integer. We note the following elementary algebraic properties:

A. $1/8 \leq (0.5 - x)^2 + x^2 \leq 1/4$ (i.e., $1/4 \leq \theta \leq 1/2$).
B. If $\alpha > \beta \geq 0$, then

$$(0.5 - x)^\alpha + x^\alpha \leq [(0.5 - x)^\beta + x^\beta][(0.5 - x)^{\alpha - \beta} + x^{\alpha - \beta}].$$

It follows that for any positive integer q

$$(0.5 - x)^{2q} + x^{2q} \leq [(0.5 - x)^2 + x^2]^q$$

and

$$(0.5 - x)^{2q+1} + x^{2q+1} \leq 0.5[(0.5 - x)^2 + x^2]^q.$$

In Case (a), $L \leq h \leq 2L - 1$. Here

$$
\begin{aligned}
p_{ij} &= \{2[(0.5 - x)^2 + x^2]\}^{2(h-L)}\{2[(0.5 - x)^3 + x^3]\}^{2L-h} \\
&\leq 2^h[(0.5 - x)^2 + x^2]^{2(h-L)}\{0.5[(0.5 - x)^2 + x^2]\}^{2L-h} \qquad \text{(property B)} \\
&= 2^{2(h-L)}[(0.5 - x)^2 + x^2]^h.
\end{aligned}
$$

With $\theta^L = 2^L[(0.5 - x)^2 + x^2]^L$, we have

$$\frac{p_{ij}}{\theta^{3L/2}} \leq \frac{2^{2(h-L)}[(0.5 - x)^2 + x^2]^{h-L}}{2^{3L/2}[(0.5 - x)^2 + x^2]^{L/2}}.$$

This ratio is ≤ 1 because the numerator is ≤ 1 while the denominator is ≥ 1 by property A noted above. Similar calculations lead to

$$\frac{p_{ij}}{\theta^{3L/2}} \leq \frac{2^{2r}[(0.5 - x)^2 + x^2]^r}{2^{3L/2}[(0.5 - x)^2 + x^2]^{L/2}}$$

for Subcase (b1) and

$$\frac{p_{ij}}{\theta^{3L/2}} \leq \frac{2^{2(h-r)}[(0.5 - x)^2 + x^2]^{(h-r)}}{2^{3L/2}[(0.5 - x)^2 + x^2]^{L/2}}$$

for Subcase (b2). Both of these ratios are ≤ 1 again because of property A.   ∎

Remark on Lemma 1. Although not directly relevant to the analysis in this paper, it is of interest to point out that calculations similar to those in Lemma 1 will also show that if $p_A = p_T$ and $p_C = p_G$, $p_{ij} \geq p_i p_j = \theta^{2L}$. Equality is obtained when one of the following special situations occur: Either all four probabilities are equal to 1/4, or one pair of probabilities is equal to 0 while the other pair is equal to 0.5. These equalities have been obtained by Ghosh and Godbole (1996). Other than these special cases, we now see that under the ECP assumption, I_i and I_j are nonnegatively correlated indicator random variables. Furthermore, we have carried out quite extensive computations, and the results indicate that the positive correlation between I_i and I_j holds even without the ECP assumption. We have yet to prove it analytically though. This positive correlation essentially says that when a palindrome occurs, it enhances the probability of having another palindrome occurring nearby that overlaps with it.

Lemma 2. *Assuming ECP, then for $1 \leq i \leq n$ and $j \in A_i$, we have*

$$E\left(\frac{1}{1 + |V_{ij}|}\right) \leq \frac{7}{\lambda} \tag{6}$$

provided that $4 \leq L \leq n/500$.

Proof. Notice that $|V_{ij}| = \sum_{k \notin A_i \cup A_j} I_k$, which is of the form $\sum_{k \in \Gamma_{ij}} I_k$. For simplicity, we suppress the notational dependence of Γ_{ij} on i, j and write

$$W = |V_{ij}| = \sum_{k \in \Gamma} I_k, \quad \mu = E(W), \text{ and } \sigma^2 = Var(W).$$

Lemma 3.1 from Brown *et al.* (2000) states that for any random variable $X \geq 1$,

$$E\left(\frac{1}{X}\right) \leq \frac{1 + \kappa/2 + \sqrt{\kappa(1 + \kappa/4)}}{E(X)} \tag{7}$$

where $\kappa = Var(X)/E(X)$. Here, we let $X = 1 + W$. Then

$$\kappa = \frac{Var(X)}{E(X)} = \frac{\sigma^2}{1 + \mu} \leq \frac{\sigma^2}{\mu}.$$

For $k, l \in \Gamma$ where $|k - l| \geq 2L$, I_k and I_l are independent. So we have

$$\sigma^2 = \sum_{k \in \Gamma} \theta^L (1 - \theta^L) + \sum_{k \in \Gamma} \sum_{l \in A_k, l \neq k} Cov(I_k, I_l)$$

$$\leq |\Gamma| \theta^L (1 - \theta^L) + 2(2L - 1)|\Gamma| \theta^{3L/2} \quad \text{from Lemma 1}$$

$$\leq |\Gamma| \theta^L (1 - \theta^L + 4L\theta^{L/2})$$

$$\leq 5\mu \quad \text{as } L\theta^{L/2} \leq 1 \text{ for } L \geq 4.$$

Consequently, $\kappa \leq 5$. We observe that the upper bound on $E(1/X)$ in (7) is an increasing function in κ. Replacing κ by 5 yields

$$E\left(\frac{1}{1 + |V_{ij}|}\right) \leq \frac{1 + 5/2 + \sqrt{5(1 + 5/4)}}{E(|V_{ij}|)} \leq \frac{6.86}{[1 - 4(2L - 1)/n]\lambda} \leq \frac{7}{\lambda}.$$

The middle inequality comes from the fact that

$$E(|V_{ij}|) = \sum_{k \notin A_i \cup A_j} E(I_k) \geq [n - 4(2L - 1)]\theta^L = [1 - 4(2L - 1)/n]\lambda.$$

The last inequality follows from $L \leq n/500$.   ∎

2.2. Poisson limit for the palindrome process

We shall make use of Lemmas 1 and 2 in conjunction with Theorem 1 to prove the following proposition stating that the palindrome process approaches a Poisson process in the limit under suitable conditions. This is done by showing that the Wasserstein distance between Ξ and Z_λ can be made arbitrarily small provided that $n \to \infty$ and L grows at a suitable rate proportional to $\log n$.

Proposition 2. *Assume ECP and suppose that n, $L \to \infty$ in such a way that $n\theta^L = \lambda$, where $\lambda \geq 1/32$ is a fixed positive constant, then*

$$d_2(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) \leq cL\theta^{L/2} \to 0$$

where c is an absolute constant no greater than 131.

Proof. First note that the condition $4 \leq L \leq n/500$ in Lemma 2 is easily satisfied when n and L become large with $n\theta^L = \lambda$. This condition will continue to be assumed true. It has been pointed out

in (5) that $b_1 \leq n(4L-1)\theta^{2L} \leq 4L\lambda\theta^L$. From Lemma 1, it follows that $b_2 \leq n(4L-2)\theta^{3L/2} \leq 4L\theta^{L/2}\lambda$. Combining the two inequalities gives

$$b_1 + b_2 \leq 4L\lambda\theta^{L/2}(1 + \theta^{L/2}) \leq 5L\lambda\theta^{L/2} \tag{8}$$

because $\theta^{L/2} \leq (1/2)^{L/2} \leq 0.25$ for $L \geq 4$. From Lemma 2, it follows that

$$\frac{5}{\lambda} + 3 \sup E\left(\frac{1}{|V_{ij}| + 1}\right) \leq \frac{26}{\lambda} \tag{9}$$

where the supremum is taken over $1 \leq i \leq n$, $j \in A_i \setminus \{i\}$. Combining Inequalities (8) and (9), we see that the first term in the upper bound for the Wasserstein distance in Theorem 1 is less than $130L\theta^{L/2}$. It can be verified that the second term $1/(2n) \leq L\theta^{L/2}$ when $\lambda \geq 1/32$. Putting the two terms together gives

$$d_2(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) \leq 131L\theta^{L/2}.$$

Furthermore, as $n\theta^L = \lambda$, L must be growing at the rate $\log n$, and hence

$$L\theta^{L/2} \leq c' \sqrt{\lambda} \log n/\sqrt{n} \to 0. \qquad \blacksquare$$

2.3. Remarks on various assumptions in Proposition 2

Various assumptions and constraints have been made in order to prove the result in Proposition 2. How restrictive are these assumptions? Are they satisfied, at least to a reasonable extent, by the actual viral genome DNA data? We shall discuss these questions in the following remarks.

Remark 1. The ECP assumption. It should be noted that the ECP assumption is sufficient but not necessary for obtaining the Poisson process limit. The proofs of Lemma 2 and Proposition 2 require only the inequality $p_{ij} \leq \theta^{3L/2}$, which can be computationally verified to hold for many base frequencies that do not satisfy the ECP assumption (e.g., $p_A = 0.1$, $p_C = 0.2$, $p_G = 0.3$, $p_T = 0.4$). In those cases, the palindrome process still approaches a limiting Poisson process.

Intuitively, the ECP assumption is quite believable, and it has been used in a number of genomic or chromosomal DNA sequence analysis studies (e.g., Burge *et al.*, 1992; Karlin *et al.*, 1993). Looking at the base composition of the herpesviruses in Table 1, we see that the relative frequencies of A and T are quite close to each other, and so are those of C and G. It is tempting to believe in the ECP assumption.

To examine this more objectively, one can turn to the Bayesian information criterion (BIC). For each genome, we compare the saturated multinomial model where the bases are generated with probabilities (p_A, p_C, p_G, p_T) summing to 1 against the ECP model with the extra condition that $p_A = p_T$ and $p_C = p_G$. As summarized in Tavaré and Giddings (1989), the BIC of a model D can be computed by

$$BIC(D) = -2l + k \log M$$

where M is the sequence length, k is the number of free parameters in model D ($k = 3$ in the saturated model and $k = 1$ in the ECP model), and

$$l = \sum_{a \in \mathcal{A}} n(a) \log \hat{p}_a$$

is the log likelihood for the data, with $n(a)$ representing the count of base a in the sequence and $\hat{p}_a$ the maximum likelihood estimate of the base probability p_a. In the saturated model, $\hat{p}_a$ is the relative frequency of a. In the ECP model, $\hat{p}_a$ is the averaged relative frequency of a and a'. When comparing two statistical models, the model with smaller BIC is considered superior.

It turns out from the BIC that in 8 out of the 16 herpesviruses in our dataset, the ECP model is preferred hence justifying the ECP assumption for their DNA sequences. For the rest of the viruses, we do not have such a nice statistical justification. Fortunately, one can computationally verify, by working out the values of p_{ij} and θ, that the inequality in Lemma 1 remains true for those base probabilities estimated from each

of the 16 viruses, ascertaining the Poisson process limit for the palindrome process in each genome as explained above.

Remark 2. Restrictions on n, L, and λ. We have also posed the restriction of $4 \leq L \leq n/500$ and $\lambda \geq 1/32$ in Lemma 2 and Proposition 2. These restrictions offer no difficulty at all to our application. In each of the herpesviruses to be analyzed, n is of the order 10^5 and θ is close to $1/4$. The values of L of interest are in the range of 4–8. One can easily verify that in each case, these requirements on n, L, and λ are satisfied.

Remark 3. The constant c. The constant c in Proposition 2 can be made much smaller than 131 in most cases of practical interest in DNA sequence analysis. The θ value calculated from real DNA base frequencies are usually close to $1/4$ rather than $1/2$, making the upper bound in (8) quite close to $4.25L\lambda\theta^{L/2}$ and the upper bound in (9) close to $17/\lambda$. This will reduce the constant c to about 73. There are places in Theorem 1 and Lemma 1 where the bounds can be made tighter also. However, since the main goal of Proposition 2 is to give a Poisson limit for the palindromes process, the actual value of c is not of great concern.

Remark 4. Can the Poisson limit approximate the palindrome process in real DNA? Proposition 2 gives only an asymptotic result stating that the palindrome process will approach a Poisson process in the limit. In the proof of Proposition 2, it was further shown that the rate of convergence is of the order $\log n/\sqrt{n}$. As this result is to serve as the basis of approximating the palindrome process in a real DNA sequence which is finite in length, one would naturally have to ask how large n and L have to be in order to make the Poisson process a good approximation of the palindrome process. Since the d_2 distance has not been previously applied to any practical setting, at this point there is no firm basis for a reliable assessment of how good the approximation is. However, we refer to a previous work of Leung *et al.* (1994) where the Poisson approximation was demonstrated to be reasonable by Q-Q plots for the cytomegalovirus genome with $n = 229,354$ and $L = 5$. The d_2 distance calculated for this case is 0.1644 (see Table 2). In our application to the herpesvirus genome coming up in Section 4, we shall use this value of the d_2 distance as a benchmark to pick values of the parameter L for the given genome lengths and compositions.

Remark 5. The i.i.d. sequence model. Proposition 2 is proved only for a random nucleotide sequence generated as i.i.d. random variables. While in many studies it has been pointed out that this i.i.d. sequence model does not fit well with real DNA sequences, the model is still frequently used for deriving statistical criteria to evaluate whether certain observed sequence patterns are unlikely to be observed simply by

TABLE 2.   VALUES OF $L\theta^{L/2}$ AT $L = 4, 5, 6, 7, 8$

Virus	$L = 4$	$L = 5$	$L = 6$	$L = 7$	$L = 8$
AHV1	0.2523	**0.1580**	0.0950	0.0556	0.0318
AtHV3	0.2867	**0.1854**	0.1151	0.0695	0.0411
BHV1	0.3605	0.2469	**0.1624**	0.1038	0.0650
EHV1	0.2586	**0.1630**	0.0986	0.0580	0.0334
EHV4	0.2500	**0.1563**	0.0938	0.0547	0.0313
MDV2	0.2500	**0.1562**	0.0937	0.0547	0.0312
MDV	0.2600	**0.1641**	0.0994	0.0586	0.0338
HSV1	0.3213	0.2138	**0.1366**	0.0848	0.0516
HSV2	0.3400	0.2295	**0.1487**	0.0937	0.0578
VZV	0.2531	**0.1587**	0.0955	0.0559	0.0320
EBV	0.2700	**0.1720**	0.1052	0.0626	0.0365
HCMV	0.2603	**0.1644**	0.0996	0.0587	0.0339
HHV6	0.2615	**0.1653**	0.1003	0.0592	0.0342
HHV7	0.2948	**0.1920**	0.1200	0.0730	0.0434
CCV1	0.2577	**0.1623**	0.0981	0.0577	0.0332
HVS2	0.2997	**0.1960**	0.1231	0.0751	0.0449

chance, mostly because of the difficulty in deriving analytical results with more elaborate models. It is also because of this limitation that the i.i.d. model is used to derive Proposition 2. The assumption of independent bases is rather hard to relax. However, one may consider allowing the base probabilities p_A, p_C, p_G, p_T to vary from one region of the sequence to another so that the local base frequencies are better reflected in the model. Indeed, there is evidence indicating that base frequencies change from one region of the genome to another. For example, Mrazrek and Karlin (1998) report a change of base preference in some herpesvirus replication origins from having more C's than C's on one side of the origin to having more C's than G's on the other. Because of the possibility of fluctuation of palindrome probabilities due to these changes in base composition, we have checked the regions around all the known replication origins in our dataset to see whether there are base compositional differences substantial enough to cause the θ value to become significantly different from that estimated from other segments of similar lengths sampled from the rest of the genome. The result in every case shows no significant difference. We have therefore chosen to simply use a value of θ estimated by the overall base frequencies in the entire genome for our analysis.

3. USE OF THE SCAN STATISTICS TO LOCATE PALINDROME CLUSTERS

In light of the mathematical result of the previous section, we can justifiably make the assumption that the midpoints of the $2L$-palindromes are distributed like the events of a Poisson process on the unit interval. It then follows from the properties of a Poisson process that if the total number of $2L$-palindromes is known, say, m, then these m points are distributed in the same way as m i.i.d. uniform random variables on $(0, 1)$ (Karlin and Taylor, 1981, Chapter 4).

For a set of points $X_1, \ldots, X_m$ distributed independently and uniformly over the unit interval $(0, 1)$, the traditional scan statistic $N_w = \max_{1 \leq i \leq m} N_w(i)$, where $0 < w < 1$ is a prescribed window length and $N_w(i)$ is the number of points in the interval $[X_{(i)}, X_{(i)} + w)$, is a generalized likelihood ratio test statistic that has been shown to be most powerful among a class of statistics to test against the clustering alternative (Naus, 1965). More recently, Dembo and Karlin (1992) define the r-scan statistic A_r to be the minimal cumulative lengths of r consecutive distances between the ordered statistics $X_{(1)}, \ldots, X_{(m)}$. Formally, let S_i denote the distance between the ordered ith and $(i + 1)$st points, i.e., $S_i = X_{(i+1)} - X_{(i)}$, $i = 1, \ldots, m - 1$. For any fixed integer r between 1 and $m - 1$, the r-scan is $A_r = \min\{A_r(i), i = 1, \ldots, m - r\}$ where $A_r(i) = \sum_{j=i}^{i+r-1} S_j, i = 1, \ldots, m - r$. Both the traditional and the r-scan statistics are used quite extensively in DNA sequence analysis (see Glaz *et al.*, 2001, Chapter 6).

These two scan statistics are essentially equivalent. Consider the event $\{N_w(i) \geq r + 1\}$ for $i = 1, \ldots, m - r$, which says that there are at least $r + 1$ points contained in the window $[X_{(i)}, X_{(i)} + w)$. This is equivalent to the event $\{A_r(i) < w\}$ which says that there are at least r adjoining spacings, starting at $X_{(i)}$, whose cumulative length is less than w. See Fig. 3 for illustration with $r = 3$. We therefore have a simple duality relationship between N_w and A_r:

$$P(N_w \geq r + 1) = P(A_r < w). \tag{10}$$

If either of the above probabilities is too small (say, < 0.05), then a cluster of at least r points in a window of length w is statistically significant.
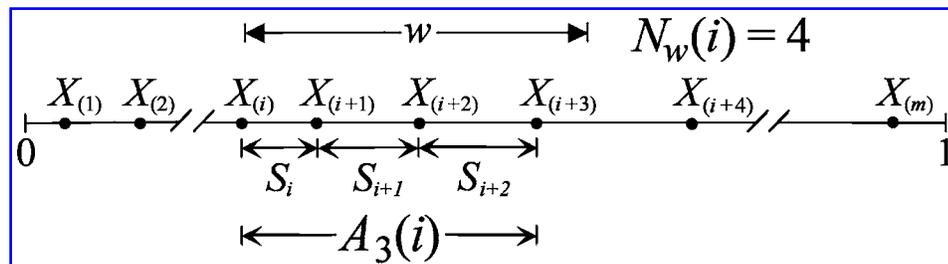

FIG. 3. The w- and r-scan statistics. The equivalence of the events $\{N_w \geq r + 1\}$ and $\{A_r < w\}$ is illustrated with $r = 3$.

There is a very simple asymptotic approximation for the distribution of A_r (Cressie, 1977; Dembo and Karlin, 1992): For any $x > 0$,

$$\lim_{m \to \infty} P(A_r \le \frac{x}{m^{1+1/r}}) = e^{-x^r/r!}.$$

When m is large, it yields the following approximation:

$$P(A_r \le w) \approx 1 - \exp\left(\frac{-m^{r+1}w^r}{r!}\right). \tag{11}$$

Leung *et al.* (1994) make use of Equation (11) to identify nonrandom palindrome clusters for the HCMV but they observe that in some cases the approximate probabilities so computed have rather large discrepancies when compared with simulated probabilities such as those obtained by Glaz (1989). It is therefore not advisable to routinely apply this approximation to evaluate the statistical significance of palindrome clusters without first considering whether the desired accuracy can be achieved.

Leung and Yamashita (1999) review a few other Poisson type approximations to the distribution of the scan statistics with special interest in their accuracy when used on data from the palindrome occurrences in seven herpes genomes which is part of our dataset in the present paper. Their simulation results show that the compound Poisson distribution put forth by Glaz *et al.* (1994) based on a result of Roos (1993) produces the best approximation to A_r. We have, therefore, adopted their result to evaluate the statistical significance of palindrome clusters for the genomes in our dataset. Explicitly, this approximation is

$$P(A_r \le w) \approx 1 - \exp\{-(m-r)\pi(1 - p + p^r(r + p - rp))\}, \tag{12}$$

where $\pi = Q_1$, $p = 1 - Q_2/Q_1$, with

$$Q_1 = \sum_{j=r}^{m} B(j; m, w), \tag{13}$$

$$Q_2 = \sum_{j=r}^{m} (-1)^{r+j} B(j; m, w), \tag{14}$$

and $B(j; m, w) = \binom{m}{j} w^j (1 - w)^{m-j}$. The approximation in (12) is used in the next section to assess the statistical significance of the r-clusters of palindromes observed in the herpesvirus genomes.

4. PALINDROME CLUSTERS IN HERPESVIRUS GENOMES

4.1. Choosing L

Our study of palindromes is motivated by their association with replication origins. In HCMV, the replication origin "oriLyt" is successfully located by a significant cluster of palindromes with $L = 5$ using the r-scan statistic The name "oriLyt" is designated to the origin of replication where DNA replication is initiated during the lytic phase of the virus life cycle. A detailed explanation of the terminology as well as the methodology is given in the article of Leung *et al.* (1994). The choice of $L = 5$ turns out to be crucial for the successful detection of the oriLyt. When the analysis is done with $L = 4$, too many statistically significant clusters were detected over very diverse regions of the genome. With $L = 6$, no significant palindrome cluster was detected at all. This indicates that the choice of L can be quite influential on the success rate of predicting a replication origin. Too small a value of L degrades the specificity while too large a value reduces the sensitivity of this method of replication origin prediction. However, apart from the knowledge that L should increase like the logarithm of the sequence length for the palindrome process to approach the Poisson limit, we do not have any guideline on what is the best length of palindromes one should examine.

We have therefore set up our algorithm in such a way that L is a parameter in the program so that its value can be easily reset by the user. The prediction accuracy is tested on our dataset with $L = 4, 5, 6, 7,$

and 8. $L = 5$ is found to work best for many of the genomes, but there are several exceptions. For example, choosing $L = 5$ for BHV1 produced 17 significant palindrome clusters, only two of which are close to replication origins. Increasing L to 6 reduces the number of significant clusters to five, which is a more acceptable number. This is at first somewhat surprising because the BHV1 genome is much shorter than that of HCMV. On closer examination, we notice that BHV1 has a much higher C/G content than all the other genomes, yielding a higher θ, which affects the palindrome probability.

In using the scan statistics to detect palindrome clusters, the choice of L should be made so that the overall palindrome process can be approximated reasonably by a Poisson process on one hand, but it should still allow the unusual palindrome clusters to be detected on the other hand. Proposition 2 tells us that the Wasserstein distance between the palindrome process and the Poisson process is bounded above by a quantity proportional to $L\theta^{L/2}$, which depends on the base frequencies as well as L. Table 2 contains the values of $L\theta^{L/2}$ computed at $L = 4, 5, 6, 7, 8$ for the base frequencies of the 16 herpesvirus genomes. If we use the value of $L\theta^{L/2}$ with $L = 5$ for HCMV as a benchmark (that is, for each virus, we choose L such that $L\theta^{L/2}$ is closest to 0.1644), L is chosen to be 5 for most of the viruses with the exceptions of BHV1, HSV1, and HSV2 whose genomes are more C/G rich than the others. For these three viruses, $L = 6$ is the closest choice. The entries corresponding to the chosen L are bold-printed in Table 2.

Eventually, we hope to be able to formulate a criterion of choosing L for each genome that has a biological basis on the type of genome structure under analysis. However, this can be achieved only if we have a larger database of experimentally confirmed replication origins. For now, our hope is that by reporting these regions of significant palindrome clusters, we will be able to facilitate the experimentation to expand this database, which will in turn help improve the prediction scheme.

4.2. The palindrome clusters and replication origins

Having chosen L, we have a computer program, which is a simple adaptation of an algorithm of Leung *et al.* (1991), to examine each of the herpes genome nucleotide sequences and find all $2L$-palindromes. It should be noted that we choose to use this particular program mainly because we have easy access to it and there are other programs for finding palindromes available in standard sequence analysis packages such as EMBOSS (Rice *et al.*, 2000). Only the nonredundant $2L$-palindromes are used for the analysis. That is, if one palindrome is completely contained in a longer one, the shorter palindrome will be discarded. The S-Plus functions developed by Leung and Yamashita (1999), based on the compound Poisson approximation to the r-scans described in Section 3, are then applied to examine the palindrome locations and identify all the nonrandom $(r + 1)$-clusters with r ranging from 1 to 15. Table 3 shows the spans of all the significant clusters found in HCMV, where the span of a cluster is the range of bases starting at the beginning position of the first palindrome in the cluster to the ending position of the last palindrome in the

TABLE 3. SEGMENTS OF THE HCMV GENOME SPANNED BY SIGNIFICANT $r + 1$-CLUSTERS

r	Clusters
1	None
2	None
3	92701–92792
4	92526–92718, 92569–92756, 92643–92792, 92701–92868, 195029–195227, 195109–195268
5	92526–92756, 92569–92792, 92643–92868, 195029–195268
6	92526–92792, 92569–92868, 92643–93119
7	91953–92792, 92526–92868, 92569–93119, 92643–93260, 92701–93520, 92709–93610
8	91635–92792, 91953–92868, 92526–93119, 92569–93260, 92643–93520, 92701–93610
9	91490–92792, 91635–92868, 91953–93119, 92526–93260, 92569–93520, 92643–93610, 92701–94183
10	91490–92868, 91635–93119, 91953–93260, 92526–93520, 92569–93610, 92643–94183
11	90759–92868, 91490–93119, 91635–93260, 91953–93520, 92526–93610, 92569–94183
12	90251–92868, 90759–93119, 91490–93260, 91635–93520, 91953–93610, 92526–94183
13	90251–93119, 90759–93260, 91490–93520, 91635–93610, 91953–94183
14	89585–93119, 90251–93260, 90759–93520, 91490–93610, 91635–94183
15	89585–93260, 90251–93520, 90759–93610, 91490–94183

TABLE 4.   REGIONS OF THE HERPESVIRUS GENOMES CONTAINING
SIGNIFICANT CLUSTERS

Genome	Region	Palindrome	Feature
BHV1	77155–77168	3	
	102895–106948	22	
	113462–113636	5	1.75 mu[a] from Ori
	124582–124756	5	1.61 mu from Ori
	131268–135221	21	
EHV1	115125–119094	17	Overlaps transcriptional regulator
	144064–148033	17	Overlaps transcriptional regulator
EHV4	None		
HSV1	None		
HSV2	None		
VZV	None		
EBV	6772–11675	19	Contains OriP
	49460–54858	25	Contains OriLyt
HCMV	89585–94183	19	Contains OriLyt
	195029–195268	8	Enhancer element
HHV6	None		
HHV7	120758–124422	16	
AHV1	113456–113759	5	
ATHV3	95350–100098	17	
MDV2	93143–93243	4	
	109331–110590	8	
MDV	None		
CCV1	None		
HVS2	None		

[a]Here, mu stands for a map unit, which is 1% of the genome length. The distance is calculated from the midpoint of the cluster region to the midpoint of the closest replication origin.

cluster. Not surprisingly, the spans of many of these clusters overlap one another. To reduce redundancy, we go through the list of clusters and join them to become one if their spans overlap. After this joining process, typically only a few nonoverlapping regions of a genome emerge. Each region contains one or more significant clusters. Table 4 lists all the regions found from the herpesvirus genomes.

Although our ultimate goal is to eventually make use of palindrome clusters to help predict the likely locations of replication origins, it must be recognized that at this stage, it is not yet possible to achieve much prediction accuracy. There are two main problems. First, the prediction procedure must also include information about clusters of close repeats and inversions which are also known to be characteristics of replication origins. A close repeat (respectively, inversion) is a segment of DNA with an exact (respectively, inverted complementary) copy of itself present in close vicinity, say, within 150 bases. A palindrome is actually a special case of close inversion because it is a segment of DNA followed immediately by its inverted complement. The statistical assessments of clusters for close repeats and inversions still need to be developed. Second, reports on confirmed location of replication origins is relatively scarce. We hope that the findings of the palindrome clusters in this paper will be helpful towards the experimental determination of more replication origins so that more information is available for prediction accuracy testing in the future.

Nevertheless, even with limited information, it is still of interest to examine the correspondence between these significant palindrome clusters and the actual confirmed locations of the replication origins. From various sources like the annotations in the GenBank file of these sequences and the references therein, plus published genetic maps and other biomedical articles (Farrel, 1993; Masse et al., 1992; McGeoch and Schaffer, 1993; Baumann et al., 1989), we are able to compile a list of replication origins in 10 of the 16 herpesviruses. These include one herpesvirus hosted in the cow, two in the horse, and seven in humans. It is not surprising that these viruses have been studied more than the others because of their agricultural and medical importance. The location of those origins are displayed in Table 5, and we also indicate them

TABLE 5. LOCATION OF REPLICATION ORIGINS IN TEN HERPESVIRUSES

Virus	Replication origins
BHV1	111080–111300(OriS), 126918–127138(OriS)
EHV1	126187–126338
EHV4	73900–73919(OriL), 119462–119481(OriS), 138568–138587(OriS)
HSV1	62475(OriL), 131999(OriS), 146235(OriS)
HSV2	62930(OriL), 132760(OriS), 148981(OriS)
VZV	110087–110350, 119547–119810
EBV	7315–9312(OriP), 52589–53581(OriLyt)
HCMV	92270–93715(OriLyt)
HHV6	67617–67993(OriLyt)
HHV7	66685–67298

in the last column of Table 4 whenever they are close (within 2% of the genome length) to one of the regions found to contain significant clusters.

While we see some agreements between palindrome cluster regions and replication origins in BHV1, EBV, and HCMV, many clusters regions have not been found to contain replication origins. There are two possibilities. First, there may be a replication origin which has not yet been experimentally located so that it is not documented in the biomedical literature. Second, a cluster region may correspond to a regulatory site rather than a replication origin. For example, the region 195029–195268 in HCMV is actually an enhancer element (Weston, 1988).

We also note that among the 19 replication origins compiled in Table 5, only five of them have palindrome clusters in their proximity. Palindrome clusters by themselves, therefore, will not be sufficient in terms of replication origin prediction. This is not unexpected because the sequence features of close repeat and inversion still need to be incorporated into the prediction scheme. Indeed, for the 14 replication origin sequences which do not contain any palindrome clusters, we further analyze the sequence structure around them using the program developed by Leung *et al.* (1991). Highly noticeable close repeats or inversions are found in all of these sequences, with only one exception.

The one exception is the oriL in the herpes Simplex I (HSV1) virus. This replication origin, located at position 62475 of the genome, is not identified by any significant palindrome clusters, nor does it have any unusual clusters of close repeats and inversions in its proximity. Instead, it has a perfect palindrome of length 144 stretching from base 62404 to base 62547. This means that it is possible to have a replication origin sequence with a highly unusual palindrome, yet our method has failed to identify it because it contains only a single long palindrome instead of a cluster of shorter ones. Upon further examination for long palindromes in the other genomes in our dataset, we find that the two replication origins of the chicken pox virus (VZV) located at bases 110087–110350 and 119547–119810 also contain two palindromes of length 36 from base 110194 to 110229 and 119668 to 119703. Again, despite the presence of a long palindrome, no significant palindrome cluster is detected.

These observations suggest the necessity of a generalization of our method of identifying palindrome clusters to take the length of the palindrome into consideration. Presently, we represent the occurrence of a palindrome by a point placed at its center. These points are considered equally as events in a Poisson process. Hence, the 144 bp palindrome carries the same weight as a 10 bp palindrome and counts as only one point. Suppose that we weigh the palindromes proportional to their lengths; then the 144 bp palindrome would count as 14 points and be deemed a significant cluster by itself. The idea of letting weights be given to events in a point process is encompassed in the theory of marked point process where each point is attached with a real-valued random variable called a mark. The Wasserstein distance between the new palindrome process and the appropriate marked Poisson process, as well as the distribution of the scan statistics will need to be developed. Such extension of the present model is under way. Together with the future incorporation of close repeats and inversions, we anticipate that a more accurate and efficient prediction scheme for replication origins can be established.

For the time being, the nonrandom palindrome clusters located by the method in this paper can be used for prioritizing which segments of DNA should be experimentally tested for replication origins first.

Among the 10 herpesvirus genomes in Table 5, 12 nonrandom palindrome clusters have been identified of which 5 (over 40%) are in proximity of replication origins. One would, therefore, expect that for those herpesviruses where no replication origins have yet been identified, the chance of finding a replication origin in a segment of about 5% the length of the genome around a palindrome cluster would be much better than that of a random genome segment of similar length. We believe that, like the work of Masse *et al.* (1992) on HCMV, a good initial choice of genome segments for testing can expedite the experimental search for replication origins.

5. CONCLUDING REMARKS

In this paper, we have focused our analysis on herpesvirus genomes. We shall be conducting a broader analysis on other families of double-stranded DNA viruses such as the adenoviruses and papillomaviruses to gain more insight into the relationship between palindrome clusters and replication origins in other viral genomes. Moreover, we would like to point out that palindrome clusters may also be associated with the other biologically relevant functional sites (e.g., enhancer elements, transcriptional regulators as indicated in Table 4). The general statistical criterion for significant palindrome clusters will allow exploratory studies of such possible associations to be undertaken. The computer programs, implemented in C and S-Plus, for locating statistically significant clusters of palindromes using scan statistics will be accessible from the web page *www.bioinformatics.utep.edu/mleung* so that interested readers can adapt them for other applications.

Recently, predictions for various biological features, such as CpG islands and coding regions on the genome sequence, have been accomplished using hidden Markov models. We expect that the hidden Markov model approach can also be useful for prediction of replication origins. The difficulty, at least for now, is insufficient known replication origins that can be used for estimation of the model parameters. In cases where data is scanty, it has been suggested (Durbin *et al.*, 1998) that Bayesian estimation should be used instead of maximum likelihood estimation. In this regard, the knowledge gained from understanding the connection between clusters of palindromes and repeats can be useful for choosing reasonable prior distributions for the model parameters.

It is a well known fact that nucleotide sequences in real DNA molecules do not fit well with the model of i.i.d. random variables (Philips *et al.*, 1987; Prum *et al.*, 1995; Leung *et al.*, 1996; just to name a few). Preferably a Markov model of order at least 3 should be used. This motivates the need to generalize Theorem 1 to a Markov chain context. Even from a purely mathematical point of view, such a generalization is an interesting problem worthy of in depth investigation. Once more, this illustrates that the interaction between mathematical and biological research will prove fruitful for the advancement of both sciences.

APPENDIX: TECHNICAL PROOFS

Proposition 3. *Let $\Gamma = [0, 1]$ with metric $d(x, y) = |x - y|$. If $\xi_1 = \sum_{i=1}^{m} \delta_{t_i}$ with $0 \leq t_1 \leq \cdots \leq t_m \leq 1$ and $\xi_2 = \sum_{i=1}^{m} \delta_{s_i}$ with $0 \leq s_1 \leq \cdots \leq s_m \leq 1$, then*

$$\sum_{i=1}^{m} |t_i - s_i| \leq \sum_{i=1}^{m} |t_i - s_{\pi(i)}| \tag{15}$$

for all permutations π of $(1, \cdots, m)$, and hence

$$d_1(\xi_1, \xi_2) = \frac{1}{m} \sum_{i=1}^{m} |t_i - s_i|.$$

Proof. We use mathematical induction to prove the claim. Suppose first that $m = 2$. Without loss of generality, we may assume $t_1 = \min\{t_1, t_2, s_1, s_2\}$. Then it suffices to consider the following three cases.

(i) $t_2 \geq s_2$, then

$$|t_1 - s_1| + |t_2 - s_2| = s_1 - t_1 + t_2 - s_2 \leq |t_1 - s_2| + |t_2 - s_1|$$

(ii) $s_1 \leq t_2 < s_2$, then

$$|t_1 - s_1| + |t_2 - s_2| = s_1 - t_1 + s_2 - t_2 \leq |t_1 - s_2| + |t_2 - s_1|$$

(iii) $t_2 < s_1$, then

$$|t_1 - s_1| + |t_2 - s_2| = s_1 - t_1 + s_2 - t_2 = |t_1 - s_2| + |t_2 - s_1|$$

Now, suppose (15) holds for $m \leq k$ with $k \geq 2$; we shall prove it holds for $m = k + 1$ and all permutations π of $(1, \cdots, k + 1)$. As a matter of fact, the claim is obvious if $\pi(k + 1) = k + 1$. Assume $\pi(k + 1) \neq k + 1$, then it follows that

$$\sum_{i=1}^{k+1} |t_i - s_{\pi(i)}| = \sum_{i \neq k+1, i \neq \pi^{-1}(k+1)} |t_i - s_{\pi(i)}| + [|t_{k+1} - s_{\pi(k+1)}| + |t_{\pi^{-1}(k+1)} - s_{k+1}|]$$

$$\geq \sum_{i \neq k+1, i \neq \pi^{-1}(k+1)} |t_i - s_{\pi(i)}| + |t_{\pi^{-1}(k+1)} - s_{\pi(k+1)}| + |t_{k+1} - s_{k+1}|$$

$$\geq \sum_{i=1}^{k} |t_i - s_i| + |t_{k+1} - s_{k+1}|. \qquad \blacksquare$$

Proof of Theorem 1. Define a Poisson process, $\tilde{Z}$, on $J := \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}\}$ as follows:

$$\tilde{Z} = \sum_{i=1}^{N} \delta_{\zeta_i}$$

where $\zeta_1, \zeta_2, \ldots$ are independent random variables uniformly distributed on J and N is a Poisson random variable with mean λ and is independent of the ζ_i's. For coupling argument below, we need to represent the usual Poisson process on $[0, 1]$ as

$$Z_\lambda = \sum_{i=1}^{N} \delta_{\zeta_i - U_i/n}$$

where U_i's are independent and uniformly distributed on $[0, 1]$. This can be seen as follows: first we pick a point from J at random and then move it to the left in a uniform manner within a distance $1/n$. We shall prove Theorem 1 in two steps: first, we apply a coupling argument to show that

$$d_2(\mathcal{L}(\tilde{Z}), \mathcal{L}(Z_\lambda)) \leq \frac{1 - e^{-\lambda}}{2n}.$$

And in the second step, we apply Stein's method to bound $d_2(\mathcal{L}(\Xi), \mathcal{L}(\tilde{Z}))$, giving the first two terms in Theorem 1. Then Theorem 1 will follow immediately by the triangle inequality:

$$d_2(\mathcal{L}(\Xi), \mathcal{L}(Z_\lambda)) \leq d_2(\mathcal{L}(\Xi), \mathcal{L}(\tilde{Z})) + d_2(\mathcal{L}(\tilde{Z}), \mathcal{L}(Z_\lambda)).$$

Step 1: We see that,

$$d_2(\mathcal{L}(\tilde{Z}), \mathcal{L}(Z_\lambda)) = \sup\{|Eh(\tilde{Z}) - Eh(Z_\lambda)| : ||h||_{Lip(\mathcal{Y})} \le 1\}$$

$$\le \sup\{E|h(\tilde{Z}) - h(Z_\lambda)| : ||h||_{Lip(\mathcal{Y})} \le 1\}$$

$$\le Ed_1(\tilde{Z}, Z_\lambda)$$

$$= Ed_1\left(\sum_{i=1}^N \delta_{\zeta_i}, \sum_{i=1}^N \delta_{\zeta_i - U_i/n}\right) I(N \ge 1)$$

$$= E\left[\frac{1}{N}\sum_{i=1}^N |\zeta_i - (\zeta_i - U_i/n)|\right] I(N \ge 1) \quad \text{(by Proposition 3)}$$

$$= E\frac{I(N \ge 1)}{nN} \sum_{i=1}^N U_i$$

$$= \frac{P(N \ge 1)}{2n} \quad \text{as } EU_i = 1/2 \text{ and } U_i's \text{ independent of } N$$

$$= \frac{1 - e^{-\lambda}}{2n}.$$

Step 2: Let $\{I_i', 1 \le i \le n\}$ be an independent copy of $\{I_i, 1 \le i \le n\}$. We first derive Stein's equation (16) as follows:

$$E\int [h(\Xi) - h(\Xi - \delta_x)]\Xi(dx)$$

$$= \sum_{i=1}^n EI_i[h(\Xi) - h(\Xi - \delta_{i/n}) - h(V_i + \delta_{i/n}) + h(V_i)] + \sum_{i=1}^n EI_i[h(V_i + \delta_{i/n}) - h(V_i)]$$

$$= \sum_{i=1}^n EI_i[h(\Xi) - h(\Xi - \delta_{i/n}) - h(V_i + \delta_{i/n}) + h(V_i)] + \sum_{i=1}^n EI_i'[h(V_i + \delta_{i/n}) - h(V_i)]$$

$$= \sum_{i=1}^n EI_i[h(\Xi) - h(\Xi - \delta_{i/n}) - h(V_i + \delta_{i/n}) + h(V_i)]$$

$$- \sum_{i=1}^n EI_i'[h(\Xi + \delta_{i/n}) - h(\Xi) - h(V_i + \delta_{i/n}) + h(V_i)]$$

$$+ \sum_{i=1}^n EI_i'[h(\Xi + \delta_{i/n}) - h(\Xi)]$$

$$= R_1(h) - R_2(h) + E\int [h(\Xi + \delta_x) - h(\Xi)]\tilde{\lambda}(dx)$$

where

$$\tilde{\lambda}(dx) = \sum_{i=1}^n p_i \delta_{i/n}(dx),$$

$$R_1(h) = \sum_{i=1}^n EI_i[h(\Xi) - h(\Xi - \delta_{i/n}) - h(V_i + \delta_{i/n}) + h(V_i)],$$

and

$$R_2(h) = \sum_{i=1}^{n} E I_i'[h(\Xi + \delta_{i/n}) - h(\Xi) - h(V_i + \delta_{i/n}) + h(V_i)].$$

Hence, we have derived Stein's identity,

$$E \int [h(\Xi + \delta_x) - h(\Xi)]\tilde{\lambda}(dx) + E \int [h(\Xi - \delta_x) - h(\Xi)]\Xi(dx) = -R_1(h) + R_2(h).$$

Consider Stein's equation:

$$\int [h(\xi + \delta_x) - h(\xi)]\tilde{\lambda}(dx) + \int [h(\xi - \delta_x) - h(\xi)]\xi(dx) = f(\xi) - \int f d\pi \qquad (16)$$

where $||f||_{Lip(\mathcal{Y})} \leq 1$ and $\pi = \mathcal{L}(\tilde{Z})$.

We need a result from Brown and Xia (2001, Theorem 5.1) which gives a nonuniform bound on the solution of Stein's equation: Suppose f satisfies $||f||_{Lip(\mathcal{Y})} \leq 1$ and h_f is the solution to Stein's equation; then we have

$$\left| h_f(\xi + \delta_x + \delta_y) - h_f(\xi + \delta_x) - h_f(\xi + \delta_y) + h_f(\xi) \right| \leq \frac{5}{\lambda} + \frac{3}{|\xi| + 1} \qquad (17)$$

where $|\xi|$ denotes the number of points in the configuration ξ.

Back to the proof of Theorem 1. We have

$$Ef(\Xi) - Ef(\tilde{Z}) = -R_1(h_f) + R_2(h_f)$$

where

$$|R_1(h_f)| \leq \sum_{i=1}^{n} E I_i \left| h_f(\Xi) - h_f(\Xi - \delta_{i/n}) - h_f(V_i + \delta_{i/n}) + h_f(V_i) \right|$$

$$\leq \sum_{i=1}^{n} E I_i \sum_{j \in A_i, j \neq i} I_j \left[\frac{5}{\lambda} + \frac{3}{1 + |V_i|} \right]$$

$$\leq \sum_{i=1}^{n} \sum_{j \in A_i, j \neq i} E I_i I_j \left[\frac{5}{\lambda} + \frac{3}{1 + |V_{ij}|} \right]$$

$$= \sum_{i=1}^{n} \sum_{j \in A_i, j \neq i} p_{ij} E \left[\frac{5}{\lambda} + \frac{3}{1 + |V_{ij}|} \right].$$

In the second inequality, we write $h_f(\Xi) - h_f(\Xi - \delta_{i/n}) - h_f(V_i + \delta_{i/n}) + h_f(V_i)$ as a telescoping sum and apply (17). Similarly,

$$|R_2(h_f)| \leq \sum_{i=1}^{n} p_i E \left| h_f(\Xi + \delta_{i/n}) - h_f(\Xi) - h_f(V_i + \delta_{i/n}) + h_f(V_i) \right|$$

$$\leq \sum_{i=1}^{n} p_i \sum_{j \in A_i, j \neq i} E I_j \left[\frac{5}{\lambda} + \frac{3}{1 + |V_i|} \right]$$

$$\leq \sum_{i=1}^{n} \sum_{j \in A_i, j \neq i} p_i p_j E \left[\frac{5}{\lambda} + \frac{3}{1 + |V_{ij}|} \right].$$

Combining the bounds on the error terms $R_1(h_f)$ and $R_2(h_f)$ which are independent of f, we take the supremum over all f with $||f||_{Lip(\mathcal{Y})} \leq 1$ to get the d_2 distance. This completes the proof of Theorem 1.   ■

Remark. Theorem 1 has been extended by Chen and Xia (2004) using Palm theory to a more general setting with wider applicability.

ACKNOWLEDGMENTS

REFERENCES

Arratia, R., Goldstein, L., and Gordon, L. 1990. Poisson approximation and the Chen–Stein method. *Statist. Sci.* 5, 403–434.

Arratia, R., Martin, D., Reinert, G., and Waterman, M.S. 1996. Poisson process approximation for sequence repeats, and sequencing by hybridization. *J. Comp. Biol.* 3, 425–463.

Barbour, A.D., Holst, L., and Janson, S. 1992. *Poisson Approximation*, Clarendon Press, Oxford.

Baumann, R.F., Yalamanchili, V.R.R., and O'Callaghan, D.J. 1988. Functional mapping a DNA sequence of an equine herpesvirus 1 origin of replication. *J. Virol.* 63(3), 1275–1283.

Bennett, J.J., Tjuvajev, J., Johnson, P., Doubrovin, M., Akhurst, T., Malholtra, S., Hackman, T., Balatoni, J., Finn, R., Larson, S.M., Federoff, H., Blasberg, R., and Fong, Y. 2001. Positron emission tomography imaging for herpes virus infection: Implications for oncolytic viral treatments of cancer. *Nat. Med.* 7(7), 859–863.

Biswas, J., Deka, S., Padmaja, S., Madhavan, H.N., Kumarasamy, N., and Solomon, S. 2001. Central retinal vein occlusion due to herpes zoster as the initial presenting sign in a patient with acquired immunodeficiency syndrome (AIDS). *Occl. Immunol. Inflamm.* 9(2), 103–109.

Bridgen, A. 1991. A restriction endonuclease map for Alcelaphine herpesvirus 1 DNA, *in* S.J. O'Brien, ed., *Genetic Maps*, 6th ed., *Book 1, Viruses*, Cold Spring Harbor Laboratory Press.

Brown, T.C., Weinberg, G.V., and Xia, A. 2000. Removing logarithms from Poisson process error bounds. *Stochastic Process. Appl.* 87, 149–165.

Brown, T.C., and Xia, A. 2001. Stein's method and birth–death processes. *Ann. Probab.* 29, 1373–1403.

Burge, C., Campbell, A.M., and Karlin, S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA.* 89, 1358–1362.

Chen, L.H.Y., and Xia, A. 2004. Stein's method, Palm theory and Poisson process approximation. To appear in *Ann. Probab.* 32, 2545–2569.

Creighton, T.E. 1993. *Proteins*, W.H. Freeman, New York.

Cressie, N. 1977. The minimum of higher order gaps. *Austral. J. Stat.* 19, 132–143.

Delecluse, H.J., and Hammerschmidt, W. 2000. The genetic approach to the Epstein-Barr virus: From basic virology to gene therapy. *J. Clin. Pathol. Mol. Pathol.* 53(5), 270–279.

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

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. 1998. *Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, London.

Farrell, P.J. 1993. Epstein-Barr virus, *in* O'Brien, S.J., ed., *Genetic Maps*, 6th ed., *Book 1, Viruses*, Cold Spring Harbor Laboratory Press.

Ghosh, D., and Godbole, A.P. 1996. Palindromes in random letter generation: Poisson approximations, rates of growth and Erdös-Rényi laws. *Lecture Notes in Statistics* 114, 99–115.

Glaz, J. 1989. Approximations and bounds for the distribution of the scan statistics. *J. Am. Statist. Assoc.* 84(406), 560–566.

Glaz, J., Naus, J., Roos, M., and Wallenstein, S. 1994. Poisson approximations for the distribution and moments of ordered m-spacings. *J. Appl. Prob.* 31A, 271–281.

Glaz, J., Naus, J., and Wallenstein, S. 2001. *Scan Statistics*, Springer-Verlag, New York.

Hamzeh, F.M., Lietman, P.S., Gibson, W., and Hayward, G.S. 1990. Identification of the lytic origin of DNA replication in human cytomegalovirus by a novel approach utilizing ganciclovir-induced chain termination. *J. Virol.* 64, 6184–6195.

Karlin, S., Blaisdell, B.E., Sapolsky, R.J., Cardon, L., and Burge, C. 1993. Assessments of DNA inhomogeneities in yeast chromosome III. *Nucl. Acids Res.* 21(3), 703–711.

Karlin, S., and Taylor, H.M. 1981. *A Second Course in Stochastic Processes*, 2nd ed., Academic Press, New York.

Kornberg, A., and Baker, T.A. 1992. *DNA Replication*, 2nd ed., W. Freeman, New York.

Labrecque, L.G., Barnes, D.M., Fentiman, I.S., and Griffin, B.E. 1995. Epstein-Barr virus in epithelial cell tumors: A breast cancer study. *Cancer Res.* 55(1), 39–45.

Leung, M.Y., Burge, C., Blaisdell, B.E., and Karlin, S. 1991. An efficient algorithm for identifying matches with errors in multiple long molecular sequences. *J. Mol. Biol.* 221, 1367–1378.

Leung, M.Y., Marsh, G.M., and Speed, T.P. 1996. Over- and underrepresentation of short DNA words in herpesvirus genomes. *J. Comp. Biol.* 3(3), 345–360.

Leung, M.Y., Schachtel, G.A., and Yu, H.S. 1994. Scan statistics and DNA sequence analysis: The search for an origin of replication in a virus. *Nonlinear World* 1, 445–471.

Leung, M.Y., and Yamashita, T.E. 1999. Applications of the scan statistic in DNA sequence analysis, *in* Glaz, J., and Balakrishnan, N., eds., *Scan Statistics and Applications*, 269–286, Birkhauser Publishers.

Masse, M.J., Karlin, S., Schachtel, G.A., and Mocarski, E.S. 1992. Human cytomegalo-virus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. *Proc. Natl. Acad. Sci. USA.* 89, 5246–5250.

McGeoch, D.J., and Schaffer, P.A. 1993. Herpes simplex virus, *in* O'Brien, S.J., ed., *Genetic Maps*, 6th ed., *Book 1, Viruses*, Cold Spring Harbor Laboratory Press.

Mrazek, J., and Karlin, S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA* 95, 3720–3725.

Naus, J.I. 1965. The distribution of the size of the maximum cluster of points on a line. *J. Am. Statist. Assoc.* 60, 532–538.

Newlon, C.S., and Theis, J.F. 2002. DNA replication joins the revolution: Whole-genome views of DNA replication in budding yeast. *BioEssays* 24, 300–304.

Phillips, G., Arnold, J., and Ivarie, R. 1987. The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. *Nucl. Acids Res.* 15, 2627–2638.

Prum, B., Rodolphe, F., and De Turckheim, E. 1995. Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B* 57(1), 205–220.

Reinert, G., and Schbath, S. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* 5, 223–253.

Reinert, G., Schbath, S., and Waterman, M.S. 2000. Probabilistic and statistical properties of words: An overview. *J. Comp. Biol.* 7, 1–46.

Reisman, D., Yates, J., and Sugden, B. 1985. A putative origin of replication of plasmids derived from Epstein-Barr virus is composed of two cis-acting components. *Mol. Cell. Biol.* 5, 1822–1832.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* 16(6), 276–277.

Roos, M. 1993. Compound Poisson approximations for the numbers of extreme spacings. *Adv. Appl. Probab.* 25, 847–874.

Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., and Tomb, J.-F. 1998. Skewed oligomers and origins of replication. *Gene* 217, 57–67.

Schbath, S. 1995. Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Prob. and Stat.* 1, 1–16.

Stein, C. 1972. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. 6th Berkeley Symp. Math Statist. Probab.* 2, 583–602.

Sugden, B. 2002. In the beginning: A viral origin exploits the cell. *Trends Biochem. Sci.* 27(1), 1–3.

Tavaré, S., and Giddings, W. 1989. Some statistical aspects of the primary structure of nucleotide sequences, *in* Waterman, M.S., ed., *Mathematical Methods for DNA Sequences*, CRC Press, Boca Raton, FL.

Vital, C., Monlun, E., Vital, A., Martin-Negrier, M.L., Cales, V., Leger, F., Longy-Boursier, M., Le Bras, M., and Bloch, B. 1995. Concurrent herpes simplex type 1 necrotizing encephalitis, cytomegalovirus ventriculoencephalitis and cerebral lymphoma in an AIDS patient. *Acta pathologica* 89(1), 105–108.

Wagner, E.K., ed. 1991. *Herpesvirus Transcription and its Regulation*, CRC Press, Boca Raton, FL.

Waterman, M.S. 1995. *Introduction to Computational Biology*, Chapman and Hall, London.

Weller, S.K., Spadaro, A., Schaffer, J.E., Murray, A.W., Maxam, A.M., and Schaffer, P.A. 1985. Cloning, sequencing, and functional analysis of ori_L, a herpes simplex virus type 1 origin of DNA synthesis. *Mol. Cell. Biol.* 5, 930–942.

Weston, K. 1988. An enhancer element in the short unique region of human cytomegalo virus regulates the production of a group of abundant immediate early transcripts. *Virology* 162, 406–416.

Zhu, Y., Huang, L., and Anders, D.G. 1998. Human cytomegalovirus oriLyt sequence requirements. *J. Virol.* 72, 4989–4996.

Address correspondence to:
Ming-Ying Leung
Department of Mathematical Sciences
University of Texas
El Paso, TX 79968-0514

E-mail: mleung@utep.edu

This article has been cited by:

1. Hock Peng Chan , Nancy Ruonan Zhang , Louis H.Y. Chen . 2010. Importance Sampling of Word Patterns in DNA and Protein SequencesImportance Sampling of Word Patterns in DNA and Protein Sequences. *Journal of Computational Biology* **17**:12, 1697-1709. [Abstract] [Full Text] [PDF] [PDF Plus]

2. Eva M. Strawbridge, Gary Benson, Yevgeniy Gelfand, Craig J. Benham. 2010. The distribution of inverted repeat sequences in the Saccharomyces cerevisiae genome. *Current Genetics* **56**:4, 321-340. [CrossRef]

3. Minge Xie, Qiankun Sun, Joseph Naus. 2009. A Latent Model to Detect Multiple Clusters of Varying Sizes. *Biometrics* **65**:4, 1011-1020. [CrossRef]

4. M. Spanò, F. Lillo, S. Miccichè, R. N. Mantegna. 2008. Statistical properties of thermodynamically predicted RNA secondary structures in viral genomes. *The European Physical Journal B* **65**:3, 323-331. [CrossRef]

5. Fabrizio Lillo, Marco Spanò. 2007. Inverted and mirror repeats in model nucleotide sequences. *Physical Review E* **76**:4. . [CrossRef]