AGGREGATE AND SPATIAL DISTRIBUTIONS OF DNA PALINDROMES AND THEIR APPLICATIONS TO REPLICATION ORIGINS PREDICTION IN SOME VIRAL GENOMES

CHEW SOON HUAT DAVID
(M.Sc, B.Sc.(Hons.), NUS)

A THESIS SUBMITTED
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
DEPARTMENT OF MATHEMATICS
NATIONAL UNIVERSITY OF SINGAPORE
2006

To Carolyn . . .

ACKNOWLEDGEMENTS

I would like to thank my advisor and friend, Professor Choi Kwok Pui, for investing a great deal of his time and energy during the past few years in me. Thanks for helping me go through this "enduring" process. I am very grateful for all you have done for me, in particular, the last few months while applying for jobs. The conversations we had in your office, especially the encouragement you gave, advice for my career; I will bear them in my mind for a long time to come. I feel blessed and fortunate to have you as my advisor.

My gratitude also goes to Professor Leung Ming-Ying, for your guidance all this while. I can still remember the day I first heard about the palindrome problem in a seminar you gave, which started my journey in this field. I have learnt a great deal from you even though we work long distance most of the time. Therefore, I greatly cherish the few times we were able to work together in person. I especially remember the encouragement you gave on the last day of my visit to El Paso in December 2005.

I would also like to thank the Department of Mathematics, especially Professor Tan Eng Chye, for employing me as a TA with the department throughout my candidature. It has enabled me to pursue my PhD degree and at the same time help support my brothers through university, which I otherwise would not have been able to do. Many thanks.

I am indebted to my family, who have supported me in their own quiet ways all these years.

Most of all, I want to thank my fiancée Carolyn, for standing by, encouraging, cheering me on and taking very good care of me, evermore so during the last stage of this journey. You are God's gift to me.

DAVID CHEW

July 2006

Table Of Contents

SUMMARY

One of the problems we will look at in this thesis concerns the over-representation (or under-representation) of palindromic words in genomic sequences, particularly in the SARS and other coronavirus genomes. Based on a Markov-chain model for the genome sequence, the mean and standard deviation of the number of palindromes at or above a certain length are derived. Using these results and extensive simulation, palindromes of a certain length are assessed whether they are statistically over-represented (or under-represented). Chapt. 2

Many empirical studies show that there are unusual clusters of palindromes, closely spaced repeats and inverted repeats around the replication origins of herpesviruses. As the search for replication origins involves labor-intensive laboratory procedures, the long-term goal of my project is to develop sound computational and statistical methods to predict the likely locations of replication origins in the herpesvirus families. This results in huge savings of time and resources. This long-term project consists of two stages.

Stage 1 is to devise new scoring schemes to measure the spatial abundance of palindromes, which generalize and refine the scan-statistics approach of Leung et al. (Leung et al., 2005, 1994; Leung and Yamashita, 1999). The new prediction methods, based on these new scoring schemes, when applied to 39 known or annotated replication origins in 19 herpesviruses have close to 80% sensitivity in the prediction accuracy (compared to about 15% by the scan statistics approach). Chapt. 3

Stage 2 is to develop the mathematics needed to compute or approximate the distribution of the scores so as to determine which scores obtained are statistically significant. We approximate the scores in one of the new schemes, the *Palindrome Length Score* by a compound Poisson distribution with parameters entirely determined by the base pair composition of the genome. Chapt. 4

As an alternative approach to predict the locations of replication origins in the double stranded herpesviruses, we propose looking at a simple, yet natural, sequence feature - the AT content. We adopt Karlin's score based approach (Karlin, 1994, 2005; Karlin and Altschul, 1990, 1993; Karlin et al., 1992) to quantitate local AT abundance reflecting the genome's base pairs composition. We then develop a computational method, called the AT excursion method, to complement the prediction methods we have developed in the first part of the thesis. Chapt. 5

Finally, we conclude this thesis by reporting some preliminary results on our attempt in adopting Karlin's excursion approach to palindromic word patterns. A summary of the approaches we have tried in this thesis in predicting locations of replication origins is presented. Some possible extensions to works in this thesis are also proposed. Chapt. 6

List of Tables

List of Figures

He giveth power to the faint; and to them that have no might he increaseth strength . . .

Isaiah

This grace gives me fear, and this grace draws me near,
And all that it asks it provides . . .

Sandra McCracken

CHAPTER 1

INTRODUCTION

Advances in biochemical techniques have led to an exponential increase in the amount of genomic sequence data available to us. Mathematical and computational methods play an increasingly important role in managing, organizing, analyzing and interpreting the rapidly accumulating DNA data. Computer algorithms can be used to compare and extract sequence features of interest while probability models and statistical techniques tell us if these features are random or not.

This thesis deals with measuring spatial abundance of some word patterns in genomic sequences. There are three main themes that we will be looking at:

(i) Over-representation (or under-representation) of RNA-palindromes in the SARS and other coronaviruses;

(ii) Novel scoring schemes to quantify the spatial abundance of DNA-palindromes; and

(iii) `AT` excursions to quantitate local `AT` abundance in genomic sequences.

In particular, we are interested to look at (ii) and (iii) and make use of them to predict the locations of replication origins in some families of double stranded viruses which includes the herpesviruses, amongst others.

1.1 A Little Biology for the Mathematician

Before we go on, let us review some relevant biological concepts and background.

Deoxyribonucleic acid (DNA) is a nucleic acid – usually in the form of a double helix – that contains the genetic instructions specifying the biological development of all cellular forms of life, and many viruses. Contrary to a common misconception, DNA is not a single molecule, but rather a pair of molecules joined by hydrogen bonds: it is organized as two complementary strands, head-to-toe, with the hydrogen bonds between them. Each strand of DNA is a chain of chemical "building blocks", called nucleotides, of which there are four types: adenine (abbreviated `A`), cytosine (`C`), guanine (`G`) and thymine (`T`).

Between the two strands, each base can only "pair up" with one single predetermined other base: `A+T`, `T+A`, `C+G` and `G+C` are the only possible combinations; that is, an "`A`" on one strand of double stranded DNA will "mate" properly only with a "`T`" on the other, complementary strand; therefore, naming the bases on the conventionally chosen side of the strand is enough to describe the entire double strand sequence. We call `A` the complement of `T` (*vice versa*), and `C` the complement of `G`. Two nucleotides paired together are called a base pair.

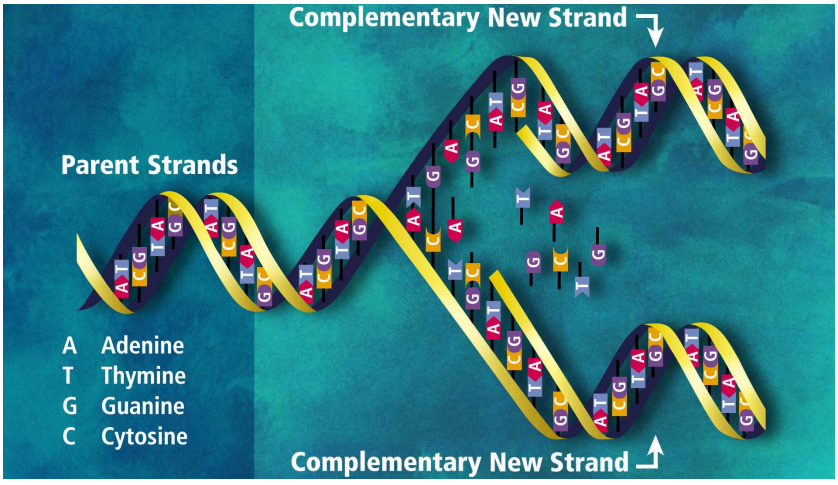


Figure 1.1 *– DNA replication*

The double stranded structure of DNA provides a simple mechanism for DNA replication: the DNA double strand is first "unzipped" down the middle, and the

"other half" of each new single strand is recreated by exposing each half to a mixture of the four bases. An enzyme makes a new strand by finding the correct base in the mixture and pairing it with the original strand. In this way, the base on the old strand dictates which base will be on the new strand, and the cell ends up with an extra copy of its DNA.

DNA palindromes are DNA words which are symmetrical in the sense that they read exactly the same as their complementary sequences in the reverse direction (see Figure 1.2 for example). A DNA palindrome is necessarily even in length because the middle base in any odd-length nucleotide string cannot be identical to its complement. More precisely, we can define a palindrome to be a word pattern of the form $b_1 \dots b_L b'_L \dots b'_1$, where b' is the complement of base b and L is called the stem length (or half-length) of the palindrome. We call the letter b_L the left-center and b'_L the right-center of the palindrome. The length of the palindrome in Figure 1.2 is 10 and $L = 5$.

$$5' \dots \overleftarrow{\text{GCAATA}}\text{TTGC} \dots 3'$$
$$3' \dots \text{CGTTA}\underrightarrow{\text{TAACG}} \dots 5'$$

Figure 1.2 *– A palindrome of length 10.*

Palindromes play important roles as protein binding sites in DNA replication processes (Kornberg and Baker, 1992, Chapter 1). 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). High concentration of palindromes around replication origins is generally attributed to the reason that the initiation of DNA replication typically requires the binding of an assembly of enzymes to these DNA sequences. Helicase is an example of these enzymes known to bind with the initiation site, locally unwind the DNA helical structure, and pull apart the two complementary strands. This explanation is consistent with the observation

of AT-rich regions, believed to facilitate the unwinding, in replication origin domains of the genome (Lin et al., 2003).

Ribonucleic acid (RNA) is primarily made up of four different bases: adenine, guanine, cytosine, and uracil (abbreviated `U`). The first three are the same as those found in DNA, but uracil replaces thymine as the base complementary to adenine. RNA serves as the template for translation of genes into proteins, transferring amino acids to the ribosome to form proteins, and also translating the transcript into proteins. The definition of a RNA palindrome is similar to that of a DNA palindrome, with uracil (`U`) taking on the role of thymine (`T`).

1.2 Organization of the Thesis

We are firstly interested to measure the abundance of *palindromic* word pattern at a global and local level. The assessment of whether DNA/RNA palindromes are over-represented or under-represented can be broadly classified into (i) global count – total count of palindromes in a biological sequence; and (ii) local count – spatial distributions of palindromes in a biological sequence.

One of the problems we will look at in this thesis concerns the over-representation (or under-representation) of palindromic words in genomic sequences, particularly in the SARS and other coronavirus genomes. Based on a Markov-chain model for the genome sequence, the mean and standard deviation of the number of palindromes at or above a certain length are derived. Using these results and extensive simulation, palindromes of a certain length are assessed whether they are statistically over-represented (or under-represented). Our conclusions are (i) length 4 palindromes are statistically significantly under-represented in all coronaviruses; and (ii) most interestingly, length 6 palindromes are significantly under-represented only in the SARS sequence and not in any other coronaviruses. These findings lead to the hypothesis that this avoidance of length-six palindromes in the SARS genome perhaps offers a protective effect on the virus, making it comparatively more difficult to be de-

Chapt. 2

stroyed. This is a joint work with Kwok Pui Choi (NUS), Hans Heidner (University of Texas, San Antonio) and Ming-Ying Leung (University of Texas, El Paso) and has been published in a special issue on computational molecular biology/bioinformatics of *INFORMS Journal on Computing*, 16(4):331-340 (Chew et al., 2004).

Many empirical studies show that there are unusual clusters of palindromes, closely spaced repeats and inverted repeats around the replication origins of herpesviruses. As the central step in the reproduction of herpesviruses, viral DNA replication has been the target for a number of anti-herpesvirus drugs. Understanding the molecular mechanisms involved in DNA replication is of great importance in further developing strategies to control the growth and spread of viruses. As the search for replication origins involves labor-intensive laboratory procedures, the long-term goal of my project is to develop sound computational and statistical methods to predict the likely locations of replication origins in the herpesvirus families. This results in huge savings of time and resources. This long-term project consists of two stages.

Stage 1 is to devise new scoring schemes to measure the spatial abundance of palindromes, which generalize and refine the scan-statistics approach of Leung et al. (Leung et al., 2005, 1994; Leung and Yamashita, 1999). The new prediction methods, based on these new scoring schemes, when applied to 39 known or annotated replication origins in 19 herpesviruses have close to 80% sensitivity in the prediction accuracy (compared to about 15% by the scan statistics approach). [1] This joint work with Kwok Pui Choi and Ming-Ying Leung has been published in *Nucleic Acids Research*, 33(15):e134 (Chew et al., 2005). Chapt. 3

Stage 2 is to develop the mathematics needed to compute or approximate the distribution of the scores so as to determine which scores obtained are statistically significant. We approximate the scores in one of the new schemes, the *Palindrome Length Score* by a compound Poisson distribution with parameters entirely deter- Chapt. 4

[1] For this thesis, we work with a slightly larger data set and so the above sentence would read "... 43 known or annotated replication origins in 20 herpesviruses...".

mined by the base pair composition of the genome. Based on this approximation, we are able to identify windows with statistically high scores which are then proposed as possible locations of replication origins of herpesviruses. Work is in progress for the other scheme.

Chapt. 5

As an alternative approach to predict the locations of replication origins in the double stranded herpesviruses, we propose looking at a simple, yet natural, sequence feature - the AT content. It has been observed that regions around the replication origins are rich in AT. One possible explanation is that segments of DNA with high AT content, i.e., lower GC content, are less stable and hence more likely candidates for replication origins. We adopt Karlin's score based approach (Karlin, 1994, 2005; Karlin and Altschul, 1990, 1993; Karlin et al., 1992) to quantitate local AT abundance reflecting the genome's base pairs composition. We then develop a computational method, called the AT excursion method, to complement the prediction methods we have developed in the first part of the thesis. The idea is to assign positive scores to AT bases and negative ones to CG bases and look for regions in the genomic sequence with high positive additive scores. Our method is statistical-based. Building on the work of Karlin and his collaborators, we have statistical tools to determine statistically high scoring segments. When this is used to predict replication origins of viruses from the herpesvirus family, we obtained results that complement the approach mentioned earlier.

Chapt. 6

Finally, we conclude this thesis by reporting some preliminary results on our attempt in adopting Karlin's excursion approach to palindromic word patterns. A summary of the approaches we have tried in this thesis in predicting locations of replication origins is presented. Some possible extensions to works in this thesis are also proposed.

CHAPTER 2

PALINDROMES IN SARS AND OTHER CORONAVIRUSES

2.1 Introduction

In March 2003, a novel coronavirus associated with the *severe acute respiratory syndrome* (SARS) was identified. The outbreak of SARS in different parts of the world, causing hundreds of deaths, has initiated much international effort that includes clinical, epidemiologic, and laboratory investigations with the aim of controlling the spread of the virus (Bloom, 2003; Marra et al., 2003; Rota et al., 2003; Ruan et al., 2003). Although the world was cleared of new SARS cases by July 2003, the pursuit for a thorough understanding of the origin, evolution, and pathogenicity of this deadly virus continues.

With the availability of the complete genome sequence of the SARS and several other coronaviruses in public databases (e.g., GenBank), it is possible to do a computational analysis of the viral genome, looking for unusual genome sequence features either unique to the SARS virus or common to the coronavirus family. Such information can give clues to the origin, natural reservoir, and evolution of the virus. It may

contribute to the studies of the immune response to this virus and the pathogenesis of SARS-related disease (Rota et al., 2003).

Statistical and experimental studies of palindromes in the other classes of viral genomes, such as the double stranded DNA viruses, bacteriophages, retroviruses, etc., have been performed (Cain et al., 2001; Dirac et al., 2002; Hill et al., 2003; Karlin et al., 1992; Leung et al., 2005; Rocha et al., 2001, among others). These studies have suggested that palindromes might be involved in the viral packaging, replication, and defense mechanisms. Unlike these well-studied viruses involved in fatal diseases such as AIDS and various cancers, the coronaviruses have not received as much attention until the recent outbreak of SARS.

In the present study, we focus our attention on palindromes in the positive stranded RNA genomes of coronaviruses. In accordance with GenBank convention, we represent an RNA sequence as a string of letters from the alphabet $\mathscr{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. The four letters respectively stand for the RNA bases adenine, cytosine, guanine, and uracil. The letters `A` and `T` are complementary to each other because adenine and uracil form hydrogen bonds with each other. The same applies to `C` and `G`. A palindrome is a symmetrical word such that when it is read in the reverse direction, it is exactly the complement of itself. For example, `ACGT` is a palindrome of length four. A palindrome is necessarily even in length because the middle base in any odd-length nucleotide string cannot be identical to its complement.

Several points are worth noting from this initial exploratory analysis of palindromes in the coronavirus genome sequences:

(1) The palindrome counts in the coronavirus genomes seem lower than what would be expected from random sequences.

(2) The SARS virus contains an exceptionally long palindrome with 22 nucleotide bases. This is the longest among all palindromes observed in the coronaviruses.

(3) There are two copies of a length-12 palindrome situated within 100 bases of each other in the SARS genome. This is not observed in the other coronaviruses.

Whether or not these palindrome-related features have any biological relevance will, of course, have to rely on careful laboratory investigations by the virologists. At this stage, however,it would be only reasonable to assess whether these features can indeed be considered statistically unusual when compared to random-sequence models. Our observations call for investigations into the probability distributions of palindrome counts, lengths, and locations in a random sequence. For this chapter, we will focus only on the palindrome counts, leaving the others for future studies.

In the next section, the mathematical formulas for the theoretical mean and variance for the number of palindromes at or above a prescribed length are derived based on a Markov-chain random-sequence model. Section 2.3 summarizes the computational results in comparing palindrome counts of the coronavirus genomes to the random-sequence models. In Section 2.4, we propose some biological questions that may be investigated in relation to these observed nonrandom features. A few concluding remarks are given in Section 2.5.

2.2 Palindrome Counts in Markov-Chain Models

The main objective of this chapter is to assess whether the palindrome counts in the coronavirus genomes are observed more (or less) frequently than expected, under some specified probability models. We model the genome sequence as a realization of a sequence of random variables $\xi_1, \xi_2, \ldots, \xi_n$ taking values in $\mathscr{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ where n is the genome length.

Throughout, we will assume that either

(i) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ are independent and identically distributed (M0); or

(ii) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ form a stationary Markov chain of order one (M1).

For studying DNA words of length k, one can choose to use Markov chains of order up to the maximum order of $k-2$ as the sequence model. A higher-order Markov chain will better fit the data sequence, but at the same time the number of parameters in the model increases exponentially. In this study, we carried out some sim-

ulations using the second-order Markov-chain model (M2). The computation takes much longer but the z scores obtained gave the same interpretation as that of the M1 model. We therefore content ourselves with the M0 and M1 models for our analysis of palindromes of length four and above.

We are interested in deriving the mean and standard deviation of the random variable X_L, total number of palindromes of length at least $2L$ under the M0 and M1 sequence models. This will help quantify the extent of deviation of the observed palindrome counts in the coronavirus genome from the expected counts under the specified probability model.

For $L \le k \le n-L$, define

$$I_k = \begin{cases} 1 & \text{if the } k\text{th base is the left center of a palindrome of length } \ge 2L \\ 0 & \text{otherwise} \end{cases}.$$

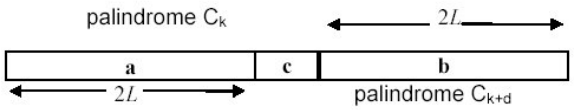
We say that a palindrome *occurs* at k when $I_k = 1$. Therefore, $X_L = \sum_{k=L}^{n-L} I_k$. Note that the distribution of I_k depends only on the joint distribution of $(\xi_{k-L+1}, \ldots, \xi_{k+L})$. Under the M0 or M1 model, the joint distribution of $(\xi_{k-L+1}, \ldots, \xi_{k+L})$ is independent of k. Hence $\mathbb{P}[I_k = 1]$ is a constant in k. Similarly $\mathbb{P}[I_j = 1,\ I_k = 1]$ depends only on $|j-k|$. Therefore, for $L \le k \le n-L$ and $1 \le d \le n-L-k$, we define

$$\gamma(0) := \mathbb{P}[I_k = 1] \quad \text{and} \quad \gamma(d) := \mathbb{P}[I_k = 1,\ I_{k+d} = 1].$$

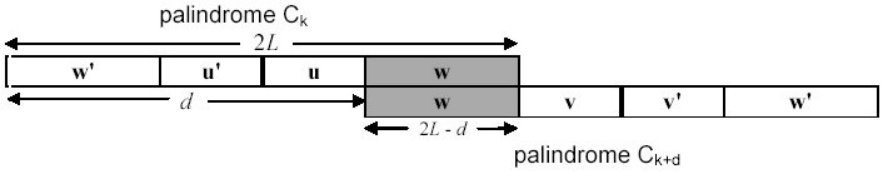
The expressions of $\gamma(0)$ and $\gamma(d)$ are crucial to calculating the mean and variance of X_L (see Proposition 2.3 below). Lemma 2.1 (respectively, Lemma 2.2) deals with the computation of $\gamma(0)$ and $\gamma(d)$ under the M1 (respectively, M0) sequence model. Indeed, we will deduce Lemma 2.2 from Lemma 2.1.

Throughout, we use b' to denote the complementary base of b, and $\mathbf{w}'$ the inversion (i.e., the complementary word read in reverse) of the word $\mathbf{w}$. There are quite a few details to work out all the possible overlap cases since the overlap structures depend on the relative sizes of d (the extent of overlap) and $2L$ (the cut-off length of

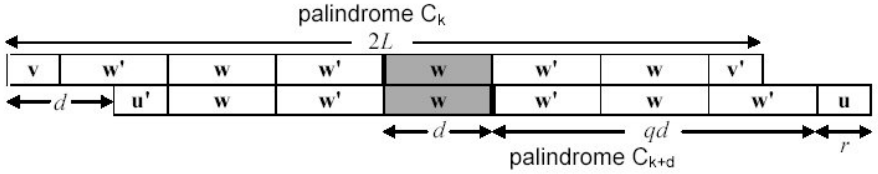
a palindrome). However, there are only two basic patterns in the overlap. In the first pattern (as illustrated by Figure 2.1(b)), the shaded segment, due to the complimentary requirement of a palindrome, will uniquely determine the left and right ends of C_k and C_{k+d}. And in the other pattern (as illustrated by Figure 2.1(c)), the shaded segment will determine the rest of both palindromes. In Figure 2.1(a), even though palindromes C_k and C_{k+d} do not actually overlap (i.e., $d \geq 2L$), the occurrence of a palindrome at k will still have an effect on the probability that a palindrome will occur at $k+d$ under the M1 sequence model. Lemma 2.1 provides expressions of $\gamma(d)$ under all possible situations.



(a) $d \geq 2L$. Here the palindromes C_k and C_{k+d} do not overlap and **c** denotes the segment between them.



(b) $L \leq d < 2L$. Here **w** denotes the common segment of palindromes C_k and C_{k+d}. And **w** determines the left end and right end of C_k and C_{k+d}.



(c) $1 \leq d < L$ with q as quotient when L is divided by d and r the remainder. The shaded segment determines the rest of both palindromes

Figure 2.1 *– Overlapping Structures of Palindromes C_k and C_{k+d} for Different Values of d. Note that (a), (b), and (c) are Drawn with Different Scales.*

Lemma 2.1. *Suppose the genome sequence is modeled as a stationary Markov chain of order one with stationary distribution* $\pi := (\pi(A), \pi(C), \pi(G), \pi(T))$. *For* $a, b \in \mathscr{A}$ *and* $m \geq 1$, *let* $P(a,b)$ *and* $P^{(m)}(a,b)$ *respectively denote the transition probability and the* m*-step transition probability from base* a *to base* b.

(a) We have

$$\gamma(0) = \sum_{b_1,\ldots,b_L \in \mathscr{A}} \pi(b_1) P(b_L, b_L^{'}) \prod_{j=1}^{L-1} \left[P(b_j, b_{j+1}) P(b_{j+1}^{'}, b_j^{'}) \right]. \tag{2.1}$$

(b) For $d \geq 1$, *we have the following three cases:*

(i) $d \geq 2L$*:*

$$\begin{aligned} \gamma(d) &= \sum_{\substack{1 \leq i \leq L \\ a_i,\, b_i \in \mathscr{A}}} \pi(a_1) P(a_L, a_L^{'}) P(b_L, b_L^{'}) P^{(d-2L+1)}(a_1^{'}, b_1) \\ &\quad \times \prod_{j=1}^{L-1} \left[P(a_j, a_{j+1}) P(a_{j+1}^{'}, a_j^{'}) P(b_j, b_{j+1}) P(b_{j+1}^{'}, b_j^{'}) \right]. \end{aligned}$$

(ii) $L \leq d < 2L$*:*

$$\begin{aligned} \gamma(d) &= \sum_{b_1,\cdots,b_d \in \mathscr{A}} \pi(b_L^{'}) P(b_1^{'}, b_1) P(b_d, b_d^{'}) \prod_{j=1}^{d-1} P(b_j, b_{j+1}) \\ &\quad \times \prod_{l=1}^{L-1} \left[P(b_{l+1}^{'}, b_l^{'}) P(b_{d-L+l+1}^{'}, b_{d-L+l}^{'}) \right]. \end{aligned}$$

(iii) $1 \leq d < L$*: we let* $L = qd + r$.

$$\begin{aligned} \gamma(d) &= \sum_{b_1,\cdots,b_d \in \mathscr{A}} K_{r,d}(b_1, \ldots, b_d) \left[P(b_d, b_d^{'}) \prod_{j=1}^{d-1} P(b_{j+1}^{'}, b_j^{'}) \right]^{q+1} \\ &\quad \times \left[P(b_1^{'}, b_1) \prod_{j=1}^{d-1} P(b_j, b_{j+1}) \right]^{q} \end{aligned}$$

where

$$K_{r,d}(b_1,\ldots,b_d) = \begin{cases} \pi(b_{d-r+1})P(b_1',b_1)\prod_{j=1}^{r-1}P(b_j,b_{j+1})\prod_{j=d-r+1}^{d-1}P(b_j,b_{j+1}) & r\geq 2 \\ \pi(b_{d-r+1})P(b_1',b_1) & r=1 \\ \frac{\pi(b_d')}{P(b_d,b_d')} & r=0 \end{cases}.$$

Proof. (a) Note that a palindrome of length at least $2L$ is of the form $b_1\cdots b_L b_L'\cdots b_1'$ where $b_1,\ldots,b_L\in\mathscr{A}$. Therefore

$$\gamma(0)=\sum_{b_1,\ldots,b_L\in\mathscr{A}}\mathbb{P}[b_1\cdots b_L b_L'\cdots b_1'].$$

Since

$$\mathbb{P}[b_1\cdots b_L b_L'\cdots b_1'] = \pi(b_1)\left[\prod_{j=1}^{L-1}P(b_j,b_{j+1})\right]P(b_L,b_L')\left[\prod_{j=1}^{L-1}P(b_{j+1}',b_j')\right],$$

(2.1) follows immediately after rearranging terms.

(b) To compute the overlap probability $\gamma(d)$, i.e., the probability that there are palindromes at k and $k+d$, we call the stretch of bases $\xi_{k-L+1}\cdots\xi_{k+d+L}$ the *span* of palindromes C_k and C_{k+d}.

For (i) $d\geq 2L$: the span $\mathbf{s}$ of the two palindromes C_k and C_{k+d} is of the form $\mathbf{acb}$ where $\mathbf{a}=a_1\cdots a_L a_L'\cdots a_1'$, $\mathbf{c}=c_1\cdots c_{d-2L}$, and $\mathbf{b}=b_1\cdots b_L b_L'\cdots b_1'$. Hence,

$$\begin{aligned}\gamma(d) &= \sum_{\mathbf{a},\mathbf{c},\mathbf{b}}\mathbb{P}[\mathbf{s}]=\sum_{\mathbf{a},\mathbf{b}}\sum_{\mathbf{c}}\mathbb{P}[\mathbf{a}]\,\mathbb{P}[\mathbf{c}b_1|a_1']\,\mathbb{P}[\mathbf{b}|b_1] \\ &= \sum_{\mathbf{a},\mathbf{b}}\mathbb{P}[\mathbf{a}]\,P^{(d-2L+1)}(a_1',b_1)\,\mathbb{P}[\mathbf{b}|b_1].\end{aligned}$$

Hence (i) follows immediately from

$$\mathbb{P}[\mathbf{a}]=\pi(a_1)\left[\prod_{j=1}^{L-1}P(a_j,a_{j+1})\right]P(a_L,a_L')\left[\prod_{j=1}^{L-1}P(a_{j+1}',a_j')\right];$$

and

$$\mathbb{P}[\mathbf{b}|b_1] = \left[\prod_{j=1}^{L-1} P(b_j, b_{j+1})\right] P(b_L, b_L^{'}) \left[\prod_{j=1}^{L-1} P(b_{j+1}^{'}, b_j^{'})\right].$$

For (ii) $L \le d < 2L$: refer to Figure 2.1(b), let $\mathbf{w} = b_{d-L+1} \cdots b_L$ denote the common segment of palindromes C_k and C_{k+d}. Assuming $d > L$, let $\mathbf{u} = b_1 \cdots b_{d-L}$ and $\mathbf{v} = b_{L+1} \cdots b_d$; we can represent $C_k = \mathbf{w}^{'}\mathbf{u}^{'}\mathbf{u}\mathbf{w}$ and $C_{k+d} = \mathbf{w}\mathbf{v}\mathbf{v}^{'}\mathbf{w}^{'}$ where $b_1, \ldots, b_d \in \mathscr{A}$. Therefore

$$\gamma(d) = \sum_{b_1,\cdots,b_d \in \mathscr{A}} \mathbb{P}[\mathbf{w}^{'}\mathbf{u}^{'}\mathbf{u}\mathbf{w}\mathbf{v}\mathbf{v}^{'}\mathbf{w}^{'}] = \sum_{b_1,\cdots,b_d \in \mathscr{A}} \mathbb{P}[b_L^{'} \cdots b_1^{'} b_1 \cdots b_d\, b_d^{'} \cdots b_{d-L+1}^{'}].$$

Writing it out in terms of the initial distribution and transition probabilities, we have proved (ii) for $d > L$. The case for $d = L$ is similar: take $\mathbf{u}$ and $\mathbf{v}$ as null words and proceed as in the case $d > L$.

To prove (iii), we consider the case $r \ge 1$ first. This time, let $\mathbf{w} = b_1 \cdots b_d$ denote the first d bases to the right of the center of C_k and to the left of the center of C_{k+d}. Let $\mathbf{u} = b_1 \cdots b_r$ and $\mathbf{v} = b_{d-r+1} \cdots b_d$ respectively denote the first and last r bases of $\mathbf{w}$. Figure 2.1(c) displays the necessary structure in C_k and C_{k+d} for both of them to be palindromes when $q = 3$.

If q is odd, then the span of C_k and C_{k+d} is of the form $\mathbf{v}\underbrace{\mathbf{w}^{'}\mathbf{w}}_{1}\cdots\underbrace{\mathbf{w}^{'}\mathbf{w}}_{q}\mathbf{w}^{'}\mathbf{u}$. Therefore,

$$\gamma(d) = \sum_{b_1,\cdots,b_d \in \mathscr{A}} \mathbb{P}[b_{d-r+1} \cdots b_d \underbrace{b_d^{'} \cdots b_1^{'} b_1 \cdots b_d}_{1} \cdots \underbrace{b_d^{'} \cdots b_1^{'} b_1 \cdots b_d}_{q} b_d^{'} \cdots b_1^{'} b_1 \cdots b_r]. \tag{2.2}$$

If q is even, then the span of C_k and C_{k+d} is changed accordingly to the form $\mathbf{u}^{'}\underbrace{\mathbf{w}\mathbf{w}^{'}}_{1}\cdots\underbrace{\mathbf{w}\mathbf{w}^{'}}_{q}\mathbf{w}\mathbf{v}^{'}$ and

$$\gamma(d) = \sum_{b_1,\cdots,b_d \in \mathscr{A}} \mathbb{P}[b_r^{'} \cdots b_1^{'} \underbrace{b_1 \cdots b_d\, b_d^{'} \cdots b_1^{'}}_{1} \cdots \underbrace{b_1 \cdots b_d\, b_d^{'} \cdots b_1^{'}}_{q} b_1 \cdots b_d\, b_d^{'} \cdots b_{d-r+1}^{'}]. \tag{2.3}$$

By making the one-to-one transformation in the summation, $b_1 \to b_d^{'}, \ldots, b_d \to$

$b_1^{'}$, and we can see that both sums on the RHS of (2.2) and (2.3) are the same. So without loss of generality, we compute $\gamma(d)$ under the assumption that q is odd. The crucial step is then to calculate the probability of the span of C_k and C_{k+d}, and part (iii) will follow immediately from summing over all possible $b_1,\ldots,b_d$. We first consider $r \geq 2$, then

$$\begin{aligned}
&\mathbb{P}[b_{d-r+1}\cdots b_d \underbrace{b_d^{'}\cdots b_1^{'}b_1\cdots b_d}_{1}\cdots\underbrace{b_d^{'}\cdots b_1^{'}b_1\cdots b_d}_{q} b_d^{'}\cdots b_1^{'}b_1\cdots b_r] \qquad (2.4)\\
&= \pi(b_{d-r+1})P(b_1^{'},b_1)\left[\prod_{j=1}^{r-1}P(b_j,b_{j+1})\right]\left[\prod_{j=d-r+1}^{d-1}P(b_j,b_{j+1})\right]\\
&\quad\times\left[P(b_d,b_d^{'})\prod_{j=1}^{d-1}P(b_{j+1}^{'},b_j^{'})\right]^{q+1}\left[P(b_1^{'},b_1)\prod_{j=1}^{d-1}P(b_j,b_{j+1})\right]^{q}.
\end{aligned}$$

For $r=1$, (2.4) becomes

$$\begin{aligned}
&\mathbb{P}[b_d\underbrace{b_d^{'}\cdots b_1^{'}b_1\cdots b_d}_{1}\cdots\underbrace{b_d^{'}\cdots b_1^{'}b_1\cdots b_d}_{q} b_d^{'}\cdots b_1^{'}b_1]\\
=\;&\pi(b_d)P(b_1^{'},b_1)\left[P(b_d,b_d^{'})\prod_{j=1}^{d-1}P(b_{j+1}^{'},b_j^{'})\right]^{q+1}\left[P(b_1^{'},b_1)\prod_{j=1}^{d-1}P(b_j,b_{j+1})\right]^{q}.
\end{aligned}$$

If $r=0$, reasoning similar to the above leads us to consider just the case q is odd. However, the span of C_k and C_{k+d} becomes (one can take $\mathbf{u}$ and $\mathbf{v}$ as empty words) $\underbrace{\mathbf{w}^{'}\mathbf{w}}_{1}\cdots\underbrace{\mathbf{w}^{'}\mathbf{w}}_{q}\mathbf{w}^{'}$. And hence

$$\begin{aligned}
&\mathbb{P}[\underbrace{b_d^{'}\cdots b_1^{'}b_1\cdots b_d}_{1}\cdots\underbrace{b_d^{'}\cdots b_1^{'}b_1\cdots b_d}_{q} b_d^{'}\cdots b_1^{'}]\\
=\;&\frac{\pi(b_d^{'})}{P(b_d,b_d^{'})}\left[P(b_d,b_d^{'})\prod_{j=1}^{d-1}P(b_{j+1}^{'},b_j^{'})\right]^{q+1}\left[P(b_1^{'},b_1)\prod_{j=1}^{d-1}P(b_j,b_{j+1})\right]^{q}.
\end{aligned}$$

□

Under the M0 model, the stationary distribution $\pi=(p_A,p_C,p_G,p_T)$, and the transition probabilities $P(a,b)=p_b$ and $P^{(m)}(a,b)=p_b$ for any $a,b\in\mathscr{A}, m\geq 1$. Substituting these into Lemma 2.1(a) and (i) and (ii) of Lemma 2.1(b) immediately gives

us the corresponding parts in Lemma 2.2 below. Part (iii) of Lemma 2.1(b) can be simplified further according to how big the remainder r is in relation to d. We shall omit the details. In this way, we have deduced the following Lemma 2.2, which was first proved in Leung et al. (2005).

Lemma 2.2. *Suppose the genome sequence is modeled as M0 and let*

$$\theta := 2(p_A p_T + p_C p_G).$$

(a) We have

$$\gamma(0) = \theta^L.$$

(b) For $d \geq 1$, we have the following four cases:

(i) $d \geq 2L$*:*

$$\gamma(d) = \theta^{2L};$$

(ii) $L \leq d < 2L$*:*

$$\gamma(d) = \theta^{2(d-L)} \left[p_A p_T (p_A + p_T) + p_C p_G (p_C + p_G) \right]^{2L-d};$$

when $1 \leq d < L$ we let $L = qd + r$ where $0 \leq r < d$, and consider two subcases according to how big the remainder r is in relation to d.

(iii) $1 \leq d < L$ *and* $0 \leq r < (d+1)/2$*:*

$$\gamma(d) = \left[2 \left((p_A p_T)^{q+1} + (p_C p_G)^{q+1} \right) \right]^{2r} \times \left[(p_A p_T)^q (p_A + p_T) + (p_C p_G)^q (p_C + p_G) \right]^{d-2r}.$$

(iv) $1 \leq d < L$ *and* $(d+1)/2 \leq r < d$*:*

$$\gamma(d) = \left[2 \left((p_A p_T)^{q+1} + (p_C p_G)^{q+1} \right) \right]^{2(d-r)} \times \left[(p_A p_T)^{q+1} (p_A + p_T) + (p_C p_G)^{q+1} (p_C + p_G) \right]^{2r-d}.$$

Proposition 2.3. *With the I_k's as defined at the beginning of Section 2.2, the total number of palindromes of length at least $2L$ is given by $X_L := \sum_{k=L}^{n-L} I_k$. And hence,*

$$\lambda_L := E(X_L) = (n-2L+1)\gamma(0)$$

and

$$\sigma_L^2 := Var(X_L) = (n-2L+1)\gamma(0)(1-\gamma(0)) + 2\sum_{d=1}^{n-2L}(n-2L+1-d)\left[\gamma(d)-\gamma(0)^2\right]$$

where $\gamma(0)$ and $\gamma(d)$ are given as in Lemma 2.2 under the M0 sequence model, and Lemma 2.1 under M1 sequence model.

Proof. The first equation follows immediately from taking expectations on both sides of $X_L := \sum_{k=L}^{n-L} I_k$. And

$$\begin{aligned}
\sigma_L^2 &= \sum_{j=L}^{n-L}\mathrm{Var}(I_j) + 2\sum_{j=L}^{n-L-1}\sum_{k=j+1}^{n-L}\mathrm{Cov}(I_j, I_k) \\
&= (n-2L+1)\gamma(0)(1-\gamma(0)) + 2\sum_{j=L}^{n-L-1}\sum_{d=1}^{n-L-j}\left[\mathbb{P}[I_j=1,\, I_{j+d}=1]-\gamma(0)^2\right] \\
&= (n-2L+1)\gamma(0)(1-\gamma(0)) + 2\sum_{d=1}^{n-2L}(n-2L+1-d)\left[\gamma(d)-\gamma(0)^2\right].
\end{aligned}$$

□

2.3 Palindrome Counts in Coronaviruses

The derived means and variances under the M0 and M1 sequence models enable us to assess whether the observed palindrome count in a genome is too abundant or rare. The z score defined in (2.5) below is a modification of a generally accepted measure of over- (or under-) representation of a DNA word. For $L \geq 2$, a standardized frequency under the assumption of the M1 sequence model is defined as

$$z_{\mathrm{M1}} = \frac{X_L - \mu_{\mathrm{M1}}}{\sigma_{\mathrm{M1}}} \tag{2.5}$$

where X_L is the observed number of palindromes of length at least $2L$, while μ_{M1} and σ_{M1} denote its expected value and standard deviation, respectively. (For simplicity, we do not indicate the dependence of μ and σ on L.) The corresponding z score is defined similarly for the M0 sequence model. When L is small compared with the genome length n, X_L is a sum of weakly dependent random indicators I_k and it is therefore well approximated by a normal distribution. Indeed, if we let $X_L^{(j)}$ denote the number of occurrences of the jth palindrome in the genome, then the count vector $(X_L^{(1)}, X_L^{(2)}, \ldots, X_L^{(4^L)})$ will converge to a multivariate normal distribution as $n \to \infty$ (see Theorem 12.5 in Waterman (1995)). And hence $X_L = \sum_{1 \le j \le 4^L} X_L^{(j)}$ will converge to a normal distribution as $n \to \infty$. For $L = 2$ or 3, and n in the range 30000, we expect that the distribution of the z scores will be approximately standard normal. The near-straight lines in the Q-Q plots in Figure 2.2 confirmed that this is the case. This motivates our definition: the count is said to be *over-* (or *under-*)*represented*, if the z score is greater than 1.645 or less than -1.645, respectively (i.e., in the upper or lower 5% of a standard normal distribution, as commonly used in one-tailed hypothesis tests in biological experiments). However, it should be emphasized that these cutoff z score values can only be considered as a convenient statistical guideline to help bring out interesting observations rather than a strict criterion to lead to a definitive conclusion.

We compute the z scores of the genomes in a data set that comprises seven coronaviruses with complete genome sequences and four other RNA viruses. For some coronaviruses, the genome sequences of multiple strains of the same virus are available. Only one strain is included in our data set because their genomes are very similar. Four other RNA viruses outside the coronavirus family are included in the data set. Two of these (the rubella virus and the equine arteritis virus) have positive-stranded RNA genomes like the coronoviruses, one (rabies virus) has a negative stranded RNA genome, and the remaining one (HIV) is a retrovirus. Table 2.1 lists the names of the viruses, abbreviations, GenBank accession numbers, genome lengths, and base compositions of the seven coronaviruses and the other four RNA viruses. Table 2.2

displays the z scores for counts of palindromes of length four and above under the M0 and M1 models.

Table 2.1 *– List of Seven Coronaviruses and Four Other RNA Viruses to be Analyzed*

Name	Abbrev.	Accession	Length	Base Composition
SARS coronavirus Urbani	SARS	AY278741	29727	(0.28, 0.20, 0.21, 0.31)
Avian infectious bronchitis virus	AIBV	NC_001451.1	27608	(0.29, 0.16, 0.22, 0.33)
Bovine coronavirus	BCoV	NC_003045.1	31028	(0.27, 0.15, 0.22, 0.36)
Human coronavirus 229E	HCoV	NC_002645.1	27317	(0.27, 0.17, 0.22, 0.35)
Murine hepatitis virus	MHV	NC_001846	31357	(0.26, 0.18, 0.24, 0.32)
Porcine epidemic diarrhea virus	PEDV	NC_003436.1	28033	(0.25, 0.19, 0.23, 0.33)
Transmissible gastroenteritis virus	TGV	NC_002306.2	28586	(0.29, 0.17, 0.21, 0.33)
Rubella virus	RUV	NC_001545.1	9755	(0.15, 0.39, 0.31, 0.15)
Equine arteritis virus	EAV	NC_002532.2	12704	(0.21, 0.26, 0.26, 0.27)
Rabies virus	RV	NC_001542.1	11932	(0.29, 0.22, 0.23, 0.26)
Human immunodeficiency virus 1	HIV-1	NC_001802.1	9181	(0.36, 0.18, 0.24, 0.22)

Table 2.2 *– z Scores for Counts of Palindromes of Length Four and Above*

Virus	Counts	$\mu_{M0}(\sigma_{M0})$	$\mu_{M1}(\sigma_{M1})$	z_{M0}	z_{M1}
SARS	1554	1981.0 (43.4)	1687.6 (40.3)	-9.83	-3.32
AIBV	1578	1896.6 (42.8)	1675.3 (38.2)	-7.45	-2.54
BCoV	1886	2115.6 (45.4)	2007.5 (45.5)	-5.06	-2.67
HCoV	1451	1843.6 (42.2)	1567.6 (37.0)	-9.30	-3.15
MHV	1793	2006.6 (43.8)	1911.3 (41.4)	-4.88	-2.86
PEDV	1457	1781.6 (41.2)	1578.8 (38.3)	-7.87	-3.18
TGV	1610	1993.9 (43.8)	1695.6 (38.9)	-8.76	-2.20
RUV	868	793.2 (28.0)	845.6 (28.3)	2.67	0.79
EAV	672	784.3 (27.2)	710.4 (25.8)	-4.13	-1.49
RV	559	758.0 (26.7)	564.3 (23.0)	-7.45	-0.23
HIV-1	475	551.9 (23.1)	480.2 (21.9)	-3.33	-0.24

Table 2.2 indicates that there is a general avoidance of palindromes of length four and above in the coronavirus genomes. A natural question that follows is whether palindromes of a given exact length are also under-represented in these viruses.

To answer this question, one would need the mean ν and standard deviation τ for the count Y_L of palindromes of exact length $2L$. It is easy to obtain the mean because $\nu = E(Y_L) = E(X_L) - E(X_{L+1})$. The standard deviation of Y_L can be derived with suitable modification of the method of proofs in Lemmas 2.1 and 2.2, but the expression

obtained is rather lengthy due to an increase in the overlapping structures. Instead, we adopt an alternative approach to estimate the standard deviation by simulation, which at the same time serves to validate our derived means and standard deviations. This approach has a further advantage of giving us the empirical distributions, and Figure 2.2 shows that for small values of L, the distributions are well approximated by normal distributions.

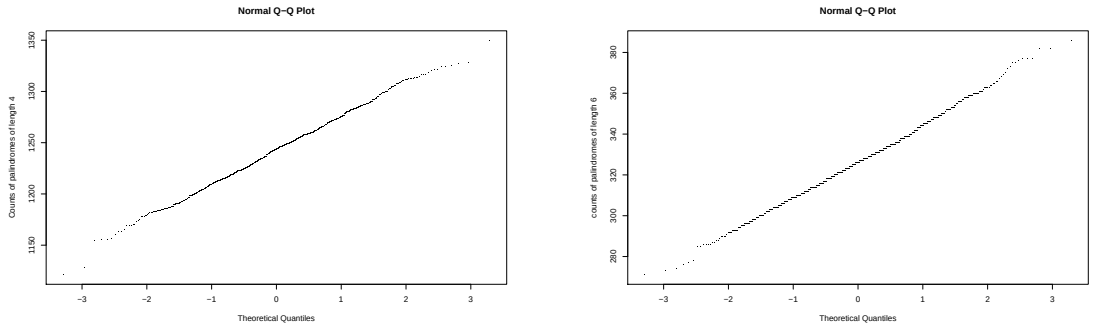


Figure 2.2 – *Normal Q-Q Plots of Counts of Palindromes of Length Four (Left) and Six (Right) in the 1000 Random Sequences Under the M1 Model for the SARS Genome*

For each virus in Table 2.1, 1000 random sequences were generated for both the M0 and M1 models using scripts written in the R language (http://www.r-project.org/). The sequences are run through the *palindrome* program which is part of EMBOSS (European Molecular Biology Open Software Suite, Rice et al. (2000)) to extract the palindrome positions and length. Each output is then read by R again and the counts of palindromes of various length are tabulated.

Tables 2.3 and 2.4 present the counts of palindromes of exact length four, six, and eight, along with their expected values ν, estimated standard deviations $\hat{\tau}$, and z scores.

Based on the z scores, Tables 2.3 and 2.4 indicate that length-four palindromes are significantly under-represented across the coronavirus family under both the M0 and M1 sequence models. However, for length-six palindromes, SARS is the only member of the coronavirus family that shows under-representation under the M1 sequence model. For length eight or above, no distinct patterns are observed.

Table 2.3 *– z Scores for Palindromes of Various Lengths Under the M0 Model*

	Length-Four Palindromes			Length-Six Palindromes			Length-Eight Palindromes		
	Count	$\nu_{M0}(\hat{\tau}_{M0})$	z_{M0}	Count	$\nu_{M0}(\hat{\tau}_{M0})$	z_{M0}	Count	$\nu_{M0}(\hat{\tau}_{M0})$	z_{M0}
SARS	1144	1469.6 (36.9)	−8.82	284	379.4 (19.4)	−4.92	90	97.9 (9.7)	−0.82
AIBV	1142	1399.5 (37.5)	−6.87	320	366.8 (18.6)	−2.52	91	96.1 (9.9)	−0.52
BCoV	1360	1563.2 (40.4)	−5.03	389	408.2 (20.4)	−0.94	98	106.6 (10.7)	−0.80
HCoV	1054	1364.7 (36.9)	−8.42	287	354.5 (18.9)	−3.57	82	92.1 (9.8)	−1.03
MHV	1328	1499.0 (38.0)	−4.50	340	379.2 (19.5)	−2.01	82	95.9 (9.9)	−1.41
PEDV	1079	1332.5 (36.5)	−6.94	274	335.9 (18.5)	−3.35	79	84.7 (9.2)	−0.62
TGV	1180	1467.3 (38.4)	−7.48	306	387.5 (19.7)	−4.14	85	102.3 (9.8)	−1.77
RUV	610	567.0 (22.8)	+1.89	167	161.7 (12.6)	+0.42	68	46.1 (6.9)	+3.17
EAV	479	589.4 (23.8)	−4.64	145	146.4 (12.3)	−0.12	36	36.4 (6.1)	−0.06
RV	407	567.0 (23.7)	−6.75	102	142.9 (12.4)	−3.30	38	36.0 (5.9)	+0.34
HIV-1	347	416.6 (20.1)	−3.46	89	102.1 (10.2)	−1.29	34	25.0 (4.8)	+1.87

Table 2.4 *– z Scores for Palindromes of Various Lengths Under the M1 Model*

	Length-Four Palindromes			Length-Six Palindromes			Length-Eight Palindromes		
	Count	$\nu_{M1}(\hat{\tau}_{M1})$	z_{M1}	Count	$\nu_{M1}(\hat{\tau}_{M1})$	z_{M1}	Count	$\nu_{M1}(\hat{\tau}_{M1})$	z_{M1}
SARS	1144	1242.7 (33.4)	−2.96	284	327.3 (18.0)	−2.41	90	86.5 (9.4)	+0.37
AIBV	1142	1229.8 (35.4)	−2.48	320	326.9 (17.8)	−0.39	91	87.0 (9.4)	+0.42
BCoV	1360	1476.5 (37.2)	−3.13	389	390.4 (19.5)	−0.07	98	103.4 (9.8)	−0.55
HCoV	1054	1146.9 (34.5)	−2.69	287	307.6 (17.4)	−1.18	82	82.7 (8.9)	−0.08
MHV	1328	1421.3 (37.8)	−2.47	340	364.3 (18.8)	−1.29	82	93.5 (9.8)	−1.17
PEDV	1079	1169.8 (34.5)	−2.63	274	302.9 (17.5)	−1.65	79	78.6 (9.1)	+0.05
TGV	1180	1239.5 (34.0)	−1.75	306	333.2 (18.4)	−1.48	85	89.8 (9.7)	−0.49
RUV	610	604.3 (24.5)	+0.23	167	172.5 (13.8)	−0.40	68	49.2 (6.9)	+2.72
EAV	479	529.6 (22.5)	−2.25	145	134.8 (11.3)	+0.91	36	34.3 (5.7)	+0.30
RV	407	415.2 (19.1)	−0.43	102	109.8 (10.4)	−0.75	38	28.9 (5.3)	+1.71
HIV-1	347	358.3 (18.7)	−0.60	89	91.0 (9.6)	−0.21	34	23.1 (4.5)	+2.42

For palindromes of length four and above, it is possible to fit higher-order Markov models to the genome sequence. For example, the second-order Markov-chain model that takes the base, dinucleotide, as well as trinucleotide composition into account, can be used to calculate the z scores. We simulated 1000 random sequences with the M2 model, but the results did not differ much from the M1 model.

As the EMBOSS *palindrome* program provides us with a detailed listing of all occurrences of palindromes of length four and above, we are able to notice two unique features in SARS. First, the SARS sequence contains a long palindrome of length 22, the longest among all palindromes observed in the coronaviruses. Second, there are two identical, length-12 palindromes situated within 100 bases of each other in the SARS genome. These are not observed in the other coronaviruses. Although con-

tributing little to the total palindrome counts, these three palindromes appear unusual enough to warrant further study of their possible biological roles, as discussed in the next section.

2.4 Discussion

Various statistical assessments of unusual abundance and rarity of individual words, including individual palindromes, in nucleotide sequences have been done using random-sequence models in a number of previous studies (Karlin et al., 1992; Merkl and Fritz, 1996; Rocha et al., 2001, 1998; Schbath et al., 1995, to name just a few). The present study, however, aims at investigating the unusual abundance and rarity of palindromes collectively rather than individually. The mathematical results in Section 2.2 provide a directly computable formula to give a single z score for all palindromes with a given minimal length. We hope the exploratory results in this chapter will serve as a basis for more detailed investigations to see how palindromes might be involved in important biological mechanisms of the coronaviruses.

There are two random sequence models M0 and M1 used in this chapter. Since M1 can take the genome dinucleotide compositions into consideration while M0 cannot, M1 is preferred over M0. Comparatively, the z scores under M1 are less extreme than those of M0. M1 is therefore more conservative in declaring the palindrome counts in a genome to be significantly different from those in random sequences. We shall base our discussion of the results on M1 whenever possible.

The counts of palindromes of length at least four in each coronavirus analyzed are significantly lower than expected (see Table 2.2). As the palindrome length increases to six and above, the under-representation of palindromes no longer holds across the family (theoretical z scores under M1 range from -1.66 to 0.46.) This suggests that there is a family-wide avoidance of palindromes of exact length four in the coronaviruses, which is confirmed by the empirical z scores for exact-length palindromes in Tables 2.3 and 2.4. With this knowledge, a thorough examination of the

relative abundance of individual length-four palindromes, conditional on the total length-four palindrome count is called for. We are in the process of setting up such a study.

Although the under-representation of length-four palindromes is observed for all of the coronaviruses in our data set that include members from all three antigenic groups (Marra et al., 2003), this under-representation is not universally true in all RNA viruses, as demonstrated by the other RNA viruses outside the coronavirus family. While it is conceivable that palindrome under-representation is just a characteristic of the common ancestor of the coronaviruses, it is worth noting that the characteristic is preserved in the family despite the reputation for RNA viruses to be nature's swiftest evolvers (Worobey and Holmes, 1999). So far, we cannot find any previous report of under-representation of short palindromes in RNA viruses with eukaryotic hosts. However, avoidance of short palindromes in some bacterial and phage DNA genomes has been reported in several studies (Karlin et al., 1992; Merkl and Fritz, 1996; Rocha et al., 2001, 1998, among others). The phenomenon is generally explained in relation to the defense mechanisms of the bacterial and phage genomes, protecting themselves against being destroyed by restriction enzymes capable of cutting up DNA molecules at certain palindromic sites. It will be interesting to investigate whether there is any possible interaction of the short palindromes in the coronavirus genomes with the immune system of the host cells that might have detrimental effects on the survival of the virus.

Length-six palindromes are found significantly under-represented only in SARS but not in the other six coronaviruses (see Table 2.4). Would this avoidance of length-six palindromes in the SARS genome offer a protective effect on the virus, making it comparatively more difficult to be destroyed and contributing to the rapid spread and the severity of the disease? This will be an interesting point to observe as we seek to learn more about the SARS virus.

Among all palindromes found in the seven coronaviruses genomes we analyzed, the longest one resides in SARS, composed of the 22 bases `TCTTTAACAAGCTTGTTAAAGA`

spanning positions 25962–25983. Since the probability distribution of palindrome lengths has not been rigorously obtained, we can only attempt a rough estimation, based on the simple M0 sequence model, of observing a length-22 palindrome in a genome with base composition like that of SARS. It has been demonstrated in Leung et al. (2005) that for larger values of L (say ≥ 5), we may approximate the counts of palindromes at or above length $2L$ by a Poisson random variable with parameter λ equal to the expected count. We therefore have $\mathbb{P}[\text{maximal palindrome length} \geq 22] = \mathbb{P}[X_{11} \geq 1]$, which can be approximated by the corresponding Poisson probability with $\lambda_{11} = E(X_{11}) = 0.01008$ by Proposition 2.3. This Poisson probability is equal to $1 - e^{-\lambda_{11}}$, about 1%.

Knowing that this long palindrome is quite unlikely to occur by chance, one would logically ask the question of whether it plays any particular functional role. According to the classification of open reading frames (ORFs) encoding potential nonstructural proteins of the SARS virus (Rota et al., 2003, Table 1), this palindrome occurs in the overlapping region of the two ORFs designated X1 and X2. Due to the location of this palindrome, it is tempting to speculate that it might be involved in some secondary structures serving similar purposes like those of a pseudoknot, which is typically found at frame-shift locations in overlapping coding sequences (Giedroc et al., 2000). One would have to perform a detailed secondary structure prediction on this part of the SARS and other coronavirus genomes before further suggestions can be made. The methods and tools used by Qin et al. (2003) to predict the secondary structure in another part of the SARS virus genome (around the packaging-signal sequence) are likely to be applicable here as well.

Another feature unique to SARS is the occurrence of two repeating length-12 palindromes `TTATAATTATAA` spanning positions 22712–22723 and 22796–22807, all within 100 bases of the genome in the coding sequence of the surface-spike glycoprotein, which is important for virus entry and virus-receptor interactions (Yu et al., 2003). Both copies begin on the third position of a codon. Three amino acids Tyr-Asn-Tyr are coded by the second through tenth bases of the palindrome. No such

repeating palindromes are observed in the corresponding glycoprotein-coding sequences for any of the other six coronaviruses. Probabilistic assessment of close repeating palindromes occurring in random sequences has yet to be formulated mathematically or estimated by simulation. (The method of Robin and Daudin (1999) can be used to assess the probability that a given palindrome repeats itself in close proximity.) If such an observation is found to be unlikely to occur by chance, then these repeating palindromes might be tested for potential regulatory functions. Large palindromes present in single-stranded RNA have the inherent ability to form double stranded stem structures through the formation of intramolecular base pairs; thus, it is possible that these sequences form secondary RNA structures in the genomic RNA and in one or more subgenomic RNAs of the SARS virus. In many of the single-stranded RNA viruses, stem structures play important regulatory roles in genome replication or gene expression. It should be possible to investigate potential regulatory roles of these repeated length-12 palindromes by engineering silent mutations within these sequences such that the encoded protein is not altered but the palindromes and putative secondary structures are lost.

2.5 Concluding Remarks

While we hope that there will never be another outbreak of SARS, we believe that detailed analysis of the SARS genome sequence can help generate useful information for understanding the biology of the coronaviruses and perhaps other RNA viruses in general. This first exploration about palindromes in the coronavirus family generates many questions to be investigated in greater detail mathematically, computationally, as well as biologically.

Closely related to palindromes is the sequence feature of close inversion, which is a palindrome with its two halves separated by a short stretch of intervening nucleotides. These close inversions are well known to form stem-loop and other secondary structures involved in the viral recombination and packaging process (Qin

et al., 2003; Rowe et al., 1997). We anticipate that a set of interesting and challenging questions in random-sequence models will again emerge from the analysis of close inversions.

CHAPTER 3

SCORING SCHEMES OF PALINDROME CLUSTERS FOR PREDICTION OF REPLICATION ORIGINS IN HERPESVIRUSES

3.1 Introduction

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 post-transplantation therapies, while others have been associated with life-threatening disease 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 animal herpesviruses are also of agricultural concern.

Example of 80 or more herpesviruses that infect a variety of animal species are the herpes simplex virus (HSV1 and HSV2), which causes cold sores and genital tract

infections in humans; Epstein-Barr virus (EBV), associated with infectious mononucleosis and with two-human cancers, Burkitt's lymphoma and nasopharyngeal carcinoma; cytomegalovirus (HCMV), causing human and animal diseases, particularly in immunodeficient individuals; varicella-zoster virus (VZV), producing chickenpox in children and shingles in adults; and Marek's herpesvirus (GaHV2), which causes malignant avian lymphoma (Kornberg and Baker, 1992).

Early studies (Reisman et al., 1985; Weller et al., 1985) have reported that the nucleotide sequences around replication origins of certain herpesviruses have complex repetitive structures of closely spaced direct and inverted repeats. A high concentration of palindromes around replication origins have been found in these herpesviruses.

Herpesviruses utilize two different types of replication origins during lytic and latent infections. For each type of origins, the count and locations in the genome vary from one kind of herpesvirus to another. Most herpesviruses have one to two copies of latent and lytic origins. Presence of palindromes around replication origins is prevalent in both latent and lytic types (Leung et al., 2005; Lin et al., 2003; Masse et al., 1992; Reisman et al., 1985; Weller et al., 1985).

As the central step in the reproduction of herpesviruses, viral DNA replication has been the target for a number of anti-herpesvirus drugs (e.g., acyclovir). Understanding the molecular mechanisms involved in DNA replication is of great importance in further developing strategies to control the growth and spread of viruses (Delecluse and Hammerschmidt, 2000; Hartline et al., 2005; Villarreal, 2003). Since replication origins are regarded as major sites for regulating genome replication, labor-intensive laboratory procedures have been used to search for replication origins (See, for example, Deng et al., 2004; Newlon and Theis, 2002; Zhu et al., 1998).

With the increasing availability of genomic DNA sequence data, one way that may save time and resources would be to scan the viral genome sequence for the expected sequence features by a computer program before an experimental search for replication origins is launched. Masse et al. (1992) first used this computational ap-

proach to predict the replication origin oriLyt on the human cytomegalovirus (HCMV) and then confirmed it by experimentation. In that computational analysis, one of the sequence features being scanned for in the genome sequence is the presence of a high concentration of palindromes of length 10 or above clustering within a window of 1000 bases.

Leung et al. (1994) describe how an evaluation criterion, based on the scan statistics (Dembo and Karlin, 1992; Glaz, 1989), is developed for assessing palindrome clusters by modeling the occurrences of palindromes in the genome as points randomly sampled from the unit interval according to the uniform distribution. By identifying windows on the genome sequence containing statistically significant clusters of palindromes, the scan statistics, in principle, provide a method to predict likely locations of replication origins. This criterion, however, essentially assesses a window of the genome by only the counts of palindrome contained in it, regardless of the actual extent of the palindrome lengths. This drawback has led to missing some replication origins which contain one extremely long palindrome rather than a cluster of moderately long ones. In the present chapter, we propose two new schemes for evaluating palindrome clusters and use the rankings of these evaluation criteria to predict the replication origins in the herpesviruses. By checking with known replication origins reported either in published literature or GenBank annotations, we assess the accuracy of the new prediction schemes. These assessments demonstrate that there is a substantial improvement over the original scan statistics criterion.

In section 2, we describe the main steps of the prediction method and three scoring schemes. The first scoring scheme, called the *palindrome count scheme* (PCS), is essentially the scan statistics method first described by Leung et al. (1994), and further discussed in the articles of Leung and Yamashita (1999), and Leung et al. (2005). Two new scoring schemes, namely, the *palindrome length scheme* (PLS) and the *base-pair weighted scheme* (BWS) are introduced as measures of palindrome clusters. In section 3, we report the results of applying these scoring schemes to predict the locations of replication origins for 42 fully sequenced herpesviruses, and compare the

prediction accuracies in terms of sensitivity and positive predictive value. A few concluding remarks are given in section 4.

3.2 Methods

We propose a computational method to identify regions of a genome which harbor unusual clusters of palindromes. This, in turn, becomes the basis of our method to predict replication origins for the herpesviruses. Table 3.1 on the following page presents the viruses to be analyzed. The data set comprises all complete genome sequences of the herpesvirus family downloaded from GenBank at the NCBI web site in March 2006. For each virus, we list its abbreviation, accession number, sequence length, and the relative frequencies of the four nucleotide bases in the genome.

Our method for predicting replication origins consists of 4 basic steps: (1) locate palindromes at or above a prescribed length; (2) choose a scoring scheme for palindromes; (3) compute a score for each window of the genome according to the chosen scoring scheme; and (4) select regions with high scores.

Step (1): Locating palindromes at or above a prescribed length:

As very short palindromes occur frequently by chance, a parameter, L, needs to be chosen where palindromes of length below $2L$ will not be considered in the analysis. Leung et al. (2005) propose a procedure, which is based on bench-marking with the well-studied HCMV virus, for the choice of L. This choice takes into account the length of the sequence, as well as the base frequencies in the genome. Using this criterion, L is chosen to be 6 for the BOHV1, BOHV5, CEHV1, CEHV2, CEHV16, HSV1, HSV2, SHV1 and THV sequences and 5 for the other sequences. Once the minimal palindrome length has been chosen, the sequences are run through the *palindrome* program, which is part of EMBOSS (European Molecular Biology Open Software Suite, Rice et al., 2000), to extract the palindrome positions and lengths. Each of these palindromes will be assigned a score

Table 3.1 *– The list of herpesviruses to be analyzed.*

Virus	Abbreviation	Accession	Length	Base Composition
Alcelaphine herpesvirus 1	alhv1	NC_002531	130608	(0.27, 0.24, 0.22, 0.26)
Ateline herpesvirus 3	athv3	NC_001987	108409	(0.32, 0.19, 0.17, 0.31)
Bovine herpesvirus 1	bohv1	NC_001847	135301	(0.14, 0.36, 0.37, 0.14)
Bovine herpesvirus 4	bohv4	NC_002665	108873	(0.30, 0.21, 0.20, 0.29)
Bovine herpesvirus 5	bohv5	NC_005261	138390	(0.12, 0.37, 0.38, 0.13)
Callitrichine herpesvirus 3	calhv3	NC_004367	149696	(0.26, 0.25, 0.25, 0.25)
Cercopithecine herpesvirus 1	cehv1	NC_004812	156789	(0.13, 0.37, 0.38, 0.13)
Cercopithecine herpesvirus 2	cehv2	NC_006560	150715	(0.12, 0.38, 0.38, 0.12)
Cercopithecine herpesvirus 8	cehv8	NC_006150	221454	(0.26, 0.25, 0.24, 0.25)
Cercopithecine herpesvirus 9	cehv7	NC_002686	124138	(0.29, 0.21, 0.20, 0.30)
Cercopithecine herpesvirus 15	cehv15	NC_006146	171096	(0.18, 0.31, 0.31, 0.20)
Cercopithecine herpesvirus 16	cehv16	NC_007653	156487	(0.12, 0.38, 0.38, 0.12)
Cercopithecine herpesvirus 17	mmrv	NC_003401	133719	(0.24, 0.27, 0.26, 0.23)
Equid herpesvirus 1	ehv1	NC_001491	150224	(0.22, 0.29, 0.28, 0.22)
Equid herpesvirus 2	ehv2	NC_001650	184427	(0.22, 0.29, 0.28, 0.21)
Equid herpesvirus 4	ehv4	NC_001844	145597	(0.25, 0.25, 0.25, 0.25)
Gallid herpesvirus 1	gahv1	NC_006623	148687	(0.26, 0.24, 0.24, 0.26)
Gallid herpesvirus 2	gahv2	NC_002229	174077	(0.28, 0.22, 0.22, 0.28)
Gallid herpesvirus 3	gahv3	NC_002577	164270	(0.23, 0.27, 0.27, 0.23)
Human herpesvirus 1	hsv1	NC_001806	152261	(0.16, 0.34, 0.34, 0.16)
Human herpesvirus 2	hsv2	NC_001798	154746	(0.15, 0.35, 0.35, 0.15)
Human herpesvirus 3	vzv	NC_001348	124884	(0.27, 0.23, 0.23, 0.27)
Human herpesvirus 4	ebv	NC_007605	171823	(0.20, 0.30, 0.30, 0.21)
Human herpesvirus 5 strain AD169	hcmv	NC_001347	230287	(0.22, 0.28, 0.29, 0.21)
Human herpesvirus 5 strain Merlin	hcmv-m	NC_006273	235645	(0.21, 0.29, 0.29, 0.21)
Human herpesvirus 6	hhv6	NC_001664	159321	(0.29, 0.22, 0.21, 0.29)
Human herpesvirus 6B	hhv6b	NC_000898	162114	(0.29, 0.22, 0.21, 0.29)
Human herpesvirus 7	hhv7	NC_001716	153080	(0.32, 0.20, 0.17, 0.31)
Human herpesvirus 8	hhv8	NC_003409	137508	(0.24, 0.27, 0.26, 0.23)
Ictalurid herpesvirus 1	ichv1	NC_001493	134226	(0.21, 0.28, 0.28, 0.22)
Meleagrid herpesvirus 1	mehv1	NC_002641	159160	(0.26, 0.24, 0.24, 0.26)
Murid herpesvirus 1	mcmv	NC_004065	230278	(0.20, 0.29, 0.30, 0.21)
Murid herpesvirus 2	rcmv	NC_002512	230138	(0.19, 0.30, 0.31, 0.20)
Murid herpesvirus 4	muhv4	NC_001826	119450	(0.27, 0.24, 0.23, 0.26)
Macaca fuscata rhadinovirus	mfrv	NC_007016	131217	(0.25, 0.27, 0.25, 0.23)
Ostreid herpesvirus 1	oshv1	NC_005881	207439	(0.31, 0.19, 0.19, 0.30)
Ovine herpesvirus 2	ohv2	NC_007646	135135	(0.23, 0.29, 0.24, 0.24)
Pongine herpesvirus 4	ccmv	NC_003521	241087	(0.19, 0.31, 0.31, 0.19)
Psittacid herpesvirus 1	pshv1	NC_005264	163025	(0.19, 0.31, 0.30, 0.20)
Saimiriine herpesvirus 2	sahv2	NC_001350	112930	(0.33, 0.18, 0.16, 0.32)
Suid herpesvirus 1	shv1	NC_006151	143461	(0.13, 0.37, 0.37, 0.13)
Tupaiid herpesvirus 1	thv	NC_002794	195859	(0.17, 0.33, 0.34, 0.17)

according to a scoring scheme chosen in the next step. Note that although it is possible for one palindrome to contain a shorter one in it (e.g. the length 12 palindrome `ACCGTGCACGGT` contains the length 10 palindrome `CCGTGCACGG`), EMBOSS automatically discards the shorter redundant palindrome and report only the longest one.

Step (2): Choosing a scoring scheme for palindromes:

Three schemes for scoring palindromes are described. In all of them, any palindrome of length less than $2L$ will always get a score 0.

(i) Palindrome count score (PCS):

In this scoring scheme, a palindrome is given a score 1 when its length is at or above $2L$.

(ii) Palindrome length score (PLS):

A palindrome of length $2s \geq 2L$ is given a score s/L. For example, if we let $L = 5$, a palindrome of length 10 will get a score of 1, while one of length 24 will get a score of 2.4.

(iii) Base-pair weighted score of order m (BWS_m):

The idea behind BWS is that a higher score should be given to rarer palindromes, namely those which have lower probabilities to occur by chance. We assess the probability of occurrence of a particular palindrome based on Markov type sequence models (Durbin et al., 2000, Chapter 3). Here m denotes the order of the Markov chain. Then, we take the negative logarithm of the probability of a palindrome to give it a positive score which is higher when the probability is lower.

We give a simple example of calculating the BWS_0 score. In the Markov model with order $m = 0$, the letters in the sequence are independent of each other. A palindrome containing respectively n_A, n_C, n_G, n_T of `A`, `C`, `G`, and `T` occurs with probability $p_A^{n_A} p_C^{n_C} p_G^{n_G} p_T^{n_T}$ where p_A, p_C, p_G, p_T are the

relative base frequencies in the sequence. The BWS_0 score of such a palindrome will be the negative logarithm of this probability, which is equal to $-(n_A \log p_A + n_C \log p_C + n_G \log p_G + n_T \log p_T)$. Consider two palindromes: `CACGTACGTG` and `TTTTTAAAAA` in a very `CG`-rich genome, say, with relative base frequencies $p_A = p_T = 0.1$, and $p_C = p_G = 0.4$. The latter palindrome is much less likely to occur than the former, and accordingly should receive a higher score to reflect its rarity compared with the former. Indeed, the calculated scores of the two palindromes turn out to be 14.7 for the former and 23.0 for the latter.

Step (3): Computing the Window score:

The score of a window in the genome is simply the total of the scores of all the palindromes occurring in this window. A palindrome is considered in the window if its left-center is. By trying out a variety of window lengths with the method, we have found that it is best to choose the window length w at 0.5% of the genome length, rounded down to the nearest hundred bases for convenience. Also, we let consecutive windows overlap by half their lengths. That is, the first window spans the first through the w-th bases, the second the $\left(\frac{w}{2}+1\right)$st to $\left(\frac{3w}{2}\right)$th bases, and so on. Because of the way the sliding windows are constructed, the length of the last window is usually shorter than w.

Step (4): Selecting regions with significant palindrome clusters:

For the PCS, regions that harbor statistical significant clusters of palindromes are identified using the scan statistics criterion as described in Leung et al. (1994). For this chapter, we use a nonparametric approach where a fixed number of top scoring windows are chosen as the predicted locations of replication origins. It is well known that herpesviruses have multiple replication origins. However, there does not appear to be any obvious rule to determine the number of top scoring windows that one should take. Based

on sensitivity and positive predictive value consideration (defined below), we find that using the top 3 to 5 ranked windows for prediction works well for the herpesviruses.

3.3 Results And Discussion

3.3.1 Scan Statistics method versus the new scoring schemes

To compare and contrast the two new scoring schemes with the scan statistics method, now called PCS, the sliding window plots for HCMV and HSV1 using PCS, PLS and BWS_0 score schemes are displayed in Figure 3.1. In each plot, the scores of the windows are plotted against the position of the window. For HCMV, the highest scoring window is the same for all three schemes. This window corresponds to the oriLyt of the HCMV identified by Masse et al. (1992). For HSV1, however, the plot of the PCS look rather different from those of the PLS and BWS. The highest scoring window in each of PLS and BWS corresponds to the oriL, and the two next highest peaks are close to the two oriS's. In contrast, the PCS fails to locate any significant clusters of palindromes.

Tables 3.2 and 3.3 shows the top 10 scoring windows for each of the 42 viruses under both the PLS and BWS schemes. The numbers in the table indicate the middle positions of the windows. In cases where two or more high scoring windows are close to one another, only one of them is picked to represent the region that gave the high scores. We adopt the practice that when a certain high scoring window is chosen, the neighboring 8 windows both to the left and to the right of it will not be considered subsequently. Rows that are shaded indicate that the particular viruses have known replication origins either from literature or from annotation. Underlined entries denote the middle positions of the windows which are within 2 map units [1] of known replication origins. Shaded rows without any underlined entries show that the computational method fails to predict the known origins of replication. Finally, rows that

[1] One map unit is one percent of the genome length, and will be abbreviated as 'mu' from now on.

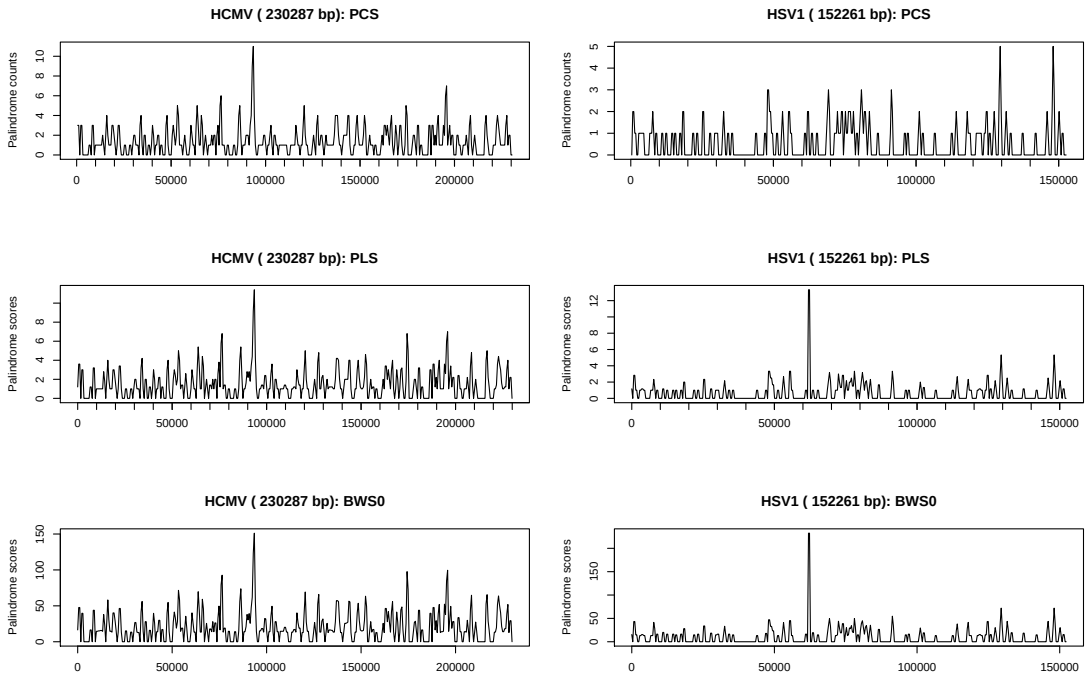


Figure 3.1 – *Sliding window plots of HCMV and HSV1 using PCS, PLS and BWS_0. The first window spans the first through the w-th bases, the second the $\left(\frac{w}{2}+1\right)$st to $\left(\frac{3w}{2}\right)$th bases, and so on. The score of a window is the total of the scores of all the palindromes occurring in this window according to PCS, PLS or BWS_0.*

are not shaded denote those viruses whose origins of replication are not known, as far as we know. Table 3.4 lists the regions with significant clusters of palindromes as found by the PCS scheme.

3.3.2 Prediction accuracy

We next examine the correspondence between the locations of these high scoring windows and those of the known replication origins. From Genbank sequence entries, annotations and literature, we are able to compile a list of 43 known replication origins for some of the viruses in our data set. Table 3.5 shows the distance between each known origin from the nearest significant palindrome cluster for PCS, or the nearest high scoring window for PLS and BWS_1 if the center of the cluster or window is within 2 mu of the origin. Otherwise a "-" is entered. The distance is calculated from the mid-point of the window to the mid-point of the closest replication origin.

Table 3.2 – *High Scoring Windows of PLS. The numbers in the table indicate the middle positions of the windows. Rows that are shaded indicate that the particular viruses have known replication origins either from literature or from annotation. Underlined entries denote the middle positions of the windows which are within* 2 *map units (i.e.* 2% *of the genome length) of known replication origins.*

	PLS Rankings									
Virus	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
alhv1	113701	32701	123301	27301	127501	110701	95101	1501	64201	120301
athv3	99001	54751	97001	1001	25501	36751	107751	86751	49501	43501
bohv1	113401	124501	103801	134401	87301	107101	131101	82801	30901	101101
bohv4	30251	54751	72251	26501	11501	48501	19751	46251	52251	2501
bohv5	19201	78001	107401	135601	31501	36901	6601	90901	67501	84301
calhv3	116201	133351	23101	56351	14001	18901	30101	100101	143851	148751
ccmv	91201	207001	177001	130201	24001	142201	63601	154201	49201	149401
cehv1	133001	149451	61601	113051	117601	109901	8051	36401	44451	140701
cehv2	129501	144201	61601	75951	123551	150501	107101	19951	92051	79101
cehv7	18601	93601	15601	24601	110701	117601	51301	101701	106201	121801
cehv8	161151	147401	198001	166651	44551	122651	88551	136401	207901	76451
cehv15	8001	34801	138801	109201	152001	68801	114001	57201	126401	100401
cehv16	21001	137201	8751	118301	154001	143151	127751	63351	38151	76301
ebv	7601	141201	41201	73201	115201	66401	12401	121201	155601	63201
ehv1	116201	146651	47601	123201	140001	94151	50751	9801	24851	56001
ehv2	6301	54001	173251	140401	46351	131851	164701	17551	160651	24751
ehv4	105351	142801	3851	109901	53551	64751	115151	27651	21001	42351
gahv1	41651	68601	99751	31851	111651	57401	126351	26951	36401	71751
gahv2	160801	801	137601	42401	46401	75201	108801	144801	168001	5601
gahv3	158801	138401	11201	122401	105201	154801	1201	132401	142401	52401
hcmv	94051	196351	77001	174901	64351	86901	53901	121001	217251	128151
hcmv-m	175451	94051	153451	77001	86901	167751	201301	190301	229351	551
hhv6	30101	8051	110601	67901	89251	125651	98701	132651	20651	24501
hhv6b	90401	69201	132801	8801	12001	60801	44001	57201	111601	31601
hhv7	133351	9451	127401	152251	29751	140701	43751	49001	62651	78401
hhv8	23401	119401	15001	136501	19201	29101	130801	102001	108601	38701
hsv1	62301	129851	148401	48301	55651	78401	91701	69651	81201	72801
hsv2	74551	7351	119701	28001	45151	12951	48651	81201	77351	1051
ichv1	55501	9301	89701	124801	19201	15001	130501	32401	108301	2101
mcmv	92951	142451	200201	130351	210651	67101	108351	101201	191401	182601
mehv1	5601	117951	11551	40951	97651	134751	72801	65451	86101	51101
mfrv	130501	115501	54601	13201	23401	75301	127201	10801	32401	101401
mmrv	132601	3301	117601	35101	87001	60001	22801	55801	32701	76801
muhv4	99251	26251	62001	50751	106251	751	30251	42251	66751	19501
ohv2	117601	134401	81001	103801	90001	99901	42001	49201	87001	16801
oshv1	21001	144001	185001	197501	204501	2501	180001	49501	67501	92501
pshv1	130401	151601	26801	60801	18801	43201	106801	11201	103201	114801
rcmv	75901	110551	83601	101751	127601	118251	8801	37401	155101	95151
sahv2	103751	112501	27751	81501	3251	6751	76501	51001	109251	90501
shv1	38151	93101	11551	46201	58451	1401	25901	85051	122851	53551
thv	134101	10801	50401	144901	85051	107551	58501	163801	54451	157951
vzv	119401	110101	100501	49201	1501	60001	13501	57301	66901	6601

Table 3.3 – *High Scoring Windows of BWS_1.*

	BWS_1 Rankings									
Virus	1	2	3	4	5	6	7	8	9	10
alhv1	113701	123301	32701	27301	127501	110701	95101	1501	64201	4801
athv3	99251	97001	54751	107751	36751	1001	25501	87501	67001	49501
bohv1	113401	124501	87301	104101	134101	82801	107101	131101	30901	101101
bohv4	54751	30251	72251	26501	11501	48501	52251	46251	19751	7751
bohv5	18901	113401	129601	78001	107401	135601	84301	31501	90901	36901
calhv3	116201	133351	23101	56351	100101	18901	14001	30101	143851	148751
ccmv	91201	207001	177001	24001	130201	142201	63601	49201	154201	115801
cehv1	133001	149451	61601	113051	117601	36401	109901	44451	8051	68601
cehv2	129501	144201	61601	75951	79101	92051	123551	150501	32201	107101
cehv7	18601	106201	121801	24601	15601	93601	110701	117601	51301	68401
cehv8	161151	147401	198001	166651	44551	122651	136401	88551	207901	76451
cehv15	8001	34801	138801	152001	109201	68801	114001	57201	126401	98801
cehv16	21001	137201	154001	8751	118301	143151	63351	38151	127751	76301
ebv	7601	41201	144001	73201	115201	66401	121201	155601	63201	78801
ehv1	116201	147001	47601	123201	140001	51101	94151	76651	73501	9801
ehv2	54001	6301	173251	140401	46351	131851	164701	160651	17551	72901
ehv4	105351	143151	109901	3851	53551	64751	115151	27651	21001	42351
gahv1	68601	41651	99751	31851	111651	57401	26951	122501	126351	36401
gahv2	160801	801	137601	46401	145201	75201	168001	20401	42401	5601
gahv3	158801	138401	11201	122401	105201	154801	142401	52401	1201	132401
hcmv	94051	174901	196351	77001	86901	53901	121001	64351	217251	209001
hcmv-m	175451	94051	153451	77001	86901	201301	167751	190301	229351	23101
hhv6	8051	30101	110601	67901	89251	132651	98701	125651	24501	93101
hhv6b	90801	132801	8801	69201	12001	60801	57201	44001	111601	2001
hhv7	9451	152251	133351	127401	29751	140701	43751	62651	49001	78401
hhv8	23401	119701	136501	15001	19201	29101	102001	130801	108601	38701
hsv1	62301	129851	148401	91701	78401	69651	48301	81201	55651	1051
hsv2	74551	28001	12951	45151	7351	119701	81201	48651	89251	77351
ichv1	55501	89701	9301	124801	19201	15001	130501	32401	108301	117901
mcmv	92951	142451	200201	130351	210651	182601	101201	108351	67101	191401
mehv1	5601	117951	11551	97651	40951	134751	72801	86101	65451	51101
mfrv	130501	115501	54601	23401	13501	75301	33601	127201	101401	10801
mmrv	132601	117601	3301	35101	87001	60001	22801	55801	123901	32701
muhv4	99251	26251	62001	50751	106251	66751	30251	42251	751	87501
ohv2	117601	134701	81001	103801	49201	42001	90001	99901	87001	16801
oshv1	21001	144001	187501	204501	197501	2501	180001	93001	103001	44001
pshv1	130401	151601	18801	26801	60801	43201	103201	106801	114801	11201
rcmv	75901	110551	83601	127601	101751	118251	207351	8801	155101	147951
sahv2	103751	112501	81501	29751	6751	3251	76501	51001	90501	11501
shv1	38151	11551	93101	46201	115151	130201	58451	1401	64051	122851
thv	134101	10801	144901	107551	49951	85501	163801	58501	54451	38251
vzv	119401	110101	100501	1501	49201	60001	66901	57301	13501	63901

Table 3.4 – *Regions with significant clusters of palindromes as found by the PCS. For example, for the virus EBV, the region 6771-10590 bp is deemed to contain a high concentration of palindromes. BOHV4, BOHV5, CEHV2, CEHV7, EHV4, GAHV1, GAHV2, HHV6, HSV1, HSV2, ICHV1, OSHV1, SAHV2 and VZV have no significant clusters of palindromes.*

Virus	Region
alhv1	113456-113759
athv3	95350-100098
bohv1	77155-77168, 102895-106948, 113462-113636, 124582-124756, 131268-135221
calhv3	21899-23918, 115406-117660, 133180-133587
ccmv	88376-93659, 206555-207582
cehv1	112833-113219
cehv8	147015-147280, 158953-164225
cehv15	5182-10840, 32483-36810, 137852-139781, 150277-152289
cehv16	20343-21242
ebv	6771-10590, 37173-42573, 138248-145848
ehv1	115125-119096, 144064-148035
ehv2	4911-9106, 147228-147250, 171785-175980
gahv3	10409-11952, 104965-105067, 121153-123174, 138321-138935, 158536-159150
hcmv	90515-95115, 195962-196203
hcmv-m	90881-96835, 175177-176003, 201246-201487
hhv6b	88469-94716
hhv7	124985-128653
hhv8	21913-23705
mcmv	92621-93412, 142118-142186
mehv1	116644-116667
mmrv	3464-3517, 130148-132723
muhv4	96755-105094
mfrv	114579-118884, 127211-130650
ohv2	113104-121989, 130697-134852
pshv1	128677-131155, 151017-153495
rcmv	74134-76485, 118126-118854
shv1	36683-41606
thv	10089-11213

Clearly, Table 3.5 shows that both PLS and BWS present a substantial improvement in the prediction accuracy of replication origins. For the PLS and BWS, we have used the top 3 scoring windows for each virus to construct this table.

Prediction accuracy of the different schemes can be quantified by two commonly accepted measures: sensitivity and positive predictive value (PPV). In our context, sensitivity is the percentage of known origins that are close to the regions suggested by the prediction; and positive predictive value is the percentage of identified regions that are close to the known origins.

Table 3.5 – *Prediction performance of various scoring schemes, PLS and BWS, based on top 3 scoring windows. The table shows the distance between each known origin from the nearest significant palindrome cluster for PCS, or the nearest high scoring window for PLS and BWS_1 if the center of the cluster or window is within 2 mu of the origin. For example, one of the top 3 scoring windows under the PLS (and BWS) for RCMV is 0.62 map unit away from the RCMV oriLyt.*

Virus	Known ORIs/ Names		PCS	PLS	BWS_1
bohv1	111080-111300	(OriS)	1.96mu	1.63mu	1.63mu
	126918-127138	(OriS)	1.52mu	1.87mu	1.87mu
bohv4	97143-98850	(OriLyt)	-	-	-
bohv5	113206-113418	(OriLyt)	-	-	0.064mu
	129595-129807	(OriLyt)	-	-	0.072mu
cehv1	61592-61789	(OriL1)	-	0.057mu	0.057mu
	61795-61992	(OriL2)	-	0.18mu	0.18mu
	132795-132796	(OriS1)	-	0.13mu	0.13mu
	132998-132999	(OriS2)	-	0.0016mu	0.0016mu
	149425-149426	(OriS2)	-	0.016mu	0.016mu
	149628-149629	(OriS1)	-	0.11mu	0.11mu
cehv2	61445-61542	(OriL)	-	0.071mu	0.071mu
	129452-129623	(OriS)	-	0.024mu	0.024mu
	144386-144557	(OriS)	-	0.18mu	0.18mu
cehv7	109627-109646		-	-	-
	118613-118632		-	-	-
cehv16	62892-63070	(OriL)	-	-	-
	133380-133578	(OriS)	-	-	-
	149725-149923	(OriS)	-	-	-
ebv	7315-9312	(OriP)	contains ori	0.41mu	0.41mu
	40301-41293	(OriLyt)	contains ori	0.23mu	0.23mu
	143207-144444	(OriLyt)	contains ori	1.52mu	0.10mu
ehv1	126187-126338		-	-	-
ehv4	73900-73919	(OriL)	-	-	-
	119462-119481	(OriS)	-	-	-
	138568-138587	(OriS)	-	-	-
gahv1	24738-25005	(OriL)	-	-	-
hcmv	93201-94646	(OriLyt)	contains ori	0.055mu	0.055mu
hhv6	67617-67993	(OriLyt)	-	-	-
hhv6b	68740-69581	(OriLyt)	-	0.024mu	-
hhv7	66685-67298		-	-	-
hsv1	62475	(OriL)	-	0.11mu	0.11mu
	131999	(OriS)	-	1.41mu	1.41mu
	146235	(OriS)	-	1.42mu	1.42mu
hsv2	62930	(OriL)	-	-	-
	132760	(OriS)	-	-	-
	148981	(OriS)	-	-	-
rcmv	75666-78970	(OriLyt)	overlaps ori	0.62mu	0.62mu
shv1	63848-63908	(OriL)	-	-	-
	114393-115009	(OriS)	-	-	-
	129593-130209	(OriS)	-	-	-
vzv	110087-110350		-	0.094mu	0.094mu
	119547-119810		-	0.22mu	0.22mu

Figure 3.2 shows the performance of the various schemes. For the PLS and BWS_1, the sensitivity and positive predictive value using one to ten top scoring windows are given in percentages. Results from BWS_0 and BWS_2 are also obtained (not shown). Their prediction accuracies are close to but slightly less than that of BWS_1. Note that as the number of windows increases, we gain in sensitivity but at the same time lose in positive predictive value. The highest sensitivities attained by PLS and BWS_1 are 67% and 79% respectively. The highest positive predictive values for both schemes are 47%.

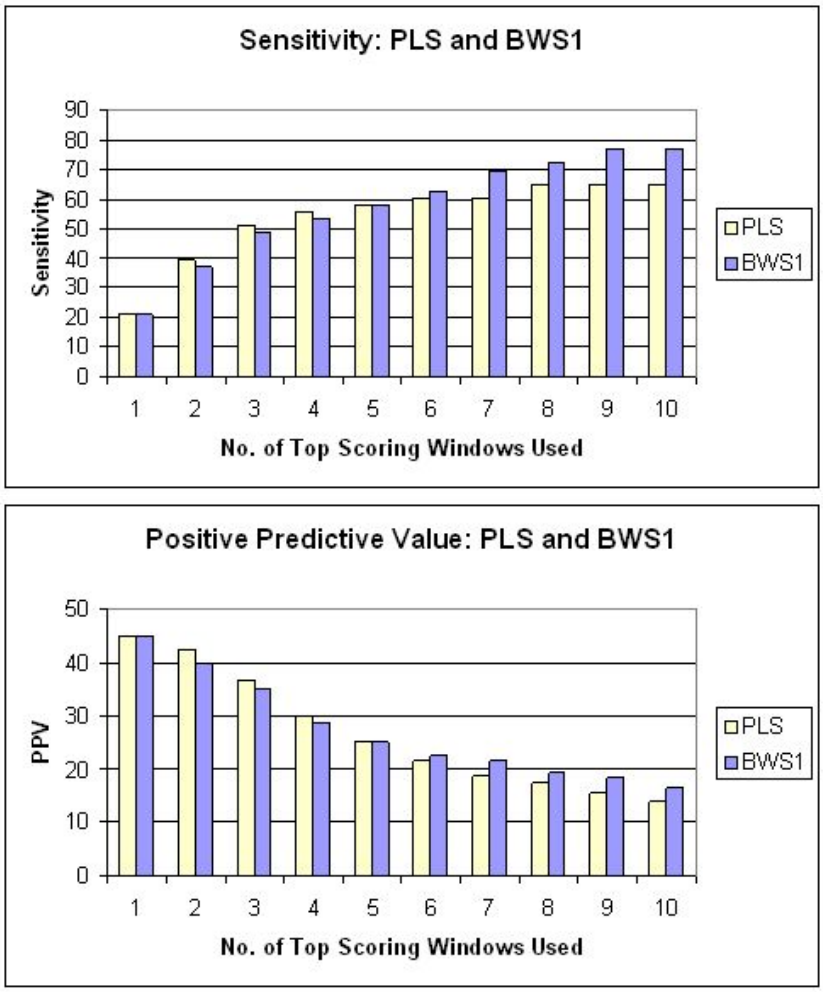


Figure 3.2 – *Sensitivity and positive predictive values of the PLS and BWS. In our context, sensitivity is the percentage of known origins that are close to the regions suggested by the prediction; and positive predictive value is the percentage of identified regions that are close to the known origins. The sensitivity and positive predictive values of the PCS are* 16 *and* 37 *respectively.*

3.3.3 Difference between PLS and BWS

Note that both PLS and BWS take the length of the palindromes into account, as longer palindromes have lower probability of occurrence than shorter ones. Moreover, the BWS takes into account the base and word frequencies which affect the probability of occurrence of the palindrome. Consider, for example, the BWS_0 score can be viewed as a weighted sum, with weights according to the negative logarithms of the base frequencies. If the base probabilities are all equal, the BWS_0 will reduce to $\log 4 \times (n_A + n_C + n_G + n_T)$ which is equal to $\log 4 \times$ Length of palindrome and hence is equivalent to the PLS.

In essence, the BWS includes more information about the sequence in its prediction and so we expect it to give better prediction accuracy. Our results show that this is indeed true. When we choose to use 6 or more top ranking windows, the BWS performs better than the PLS in terms of (higher) sensitivity and positive predictive value.

Suspecting that the probability of occurrence of palindromes might not be well estimated on the basis of a global base and word frequencies, we also try calculating palindrome probabilities using the base and word frequencies of those at the local window rather than those of the entire genome.

Figure 3.3 shows the sensitivity and positive predictive values of the local BWS of order $0, 1, 2$. We use BWS_m(Local) to represent the local version of BWS of order m. According to these results, the local version still does not perform any better than BWS_1.

3.3.4 Further improvement of the algorithm

While our results show that using PLS and BWS with the ranking approach clearly outperforms the PCS, we have to note that the PCS is the only scheme where a rigorous statistical significance criterion, based on the probability distribution of the scan statistics, is currently available. The probability distributions of the maximal window scores with PLS and BWS have yet to be established. We have some prelim-

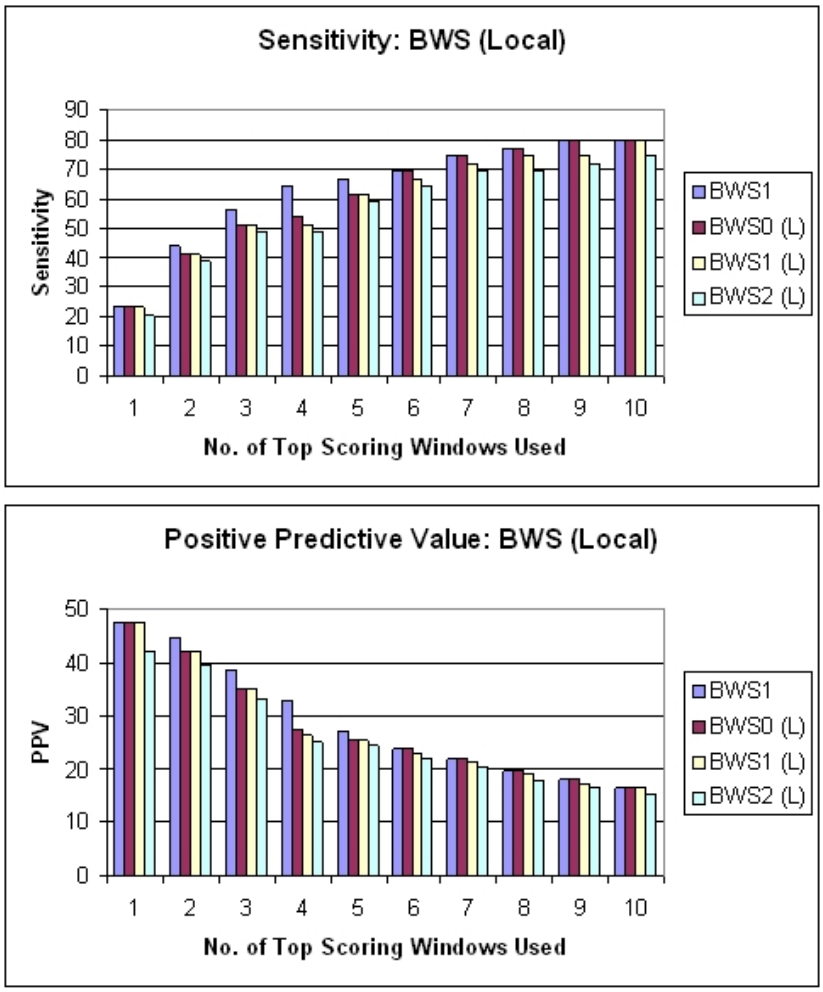


Figure 3.3 – *Sensitivity and positive predictive values of Local BWS.*

inary results on approximating the distributions of the window score under PLS by compound Poisson distribution. The compound Poisson distribution is motivated from a marked Poisson process point of view. The occurrence of a palindrome of length $2L$ and above is modeled by a Poisson process (Leung et al., 2005), and the actual length of this palindrome is modeled by a geometric distribution.

On closer examination of the known replication origins in this set of genome sequences, we notice that some of the origins missed by this prediction algorithm are actually rather long approximate palindromes. They are missed because we choose to consider only the perfect palindromes. For example, in HSV2, allowing just one error would have let us pick up a 136 base long approximate palindrome centered at 62930, which is where the reported replication origin is located. If we include

these approximate palindromes in our consideration, the sensitivity can be further increased.

3.4 Concluding Remarks

It is mentioned in the introduction that palindromes are merely one type of sequence features known to be associated with replication origins. Other frequently observed characteristics around replication origins include clustering of closely spaced direct and inverted repeats, as well as high AT content. We have actually examined each of these other types of sequence features and found that none of them, when used alone on our data set, reaches the same level of prediction accuracy offered by the BWS. However, it is likely that the prediction accuracy can be further improved by appropriately incorporating them in the prediction scheme. In fact, several replication origins in BoHV4, EHV4 and HSV2 which are not identified by any of PCS, PLS, or BWS can be easily detected by the high local AT content around them. Exactly in what way all the different sequence features should be combined to produce the optimal prediction results is the subject of an ongoing investigation.

While it is encouraging to see that close to 80% of replication origins can be predicted using a palindrome based scoring scheme like BWS, we have also noted that the positive predictive value is rather low whenever the corresponding sensitivity exceeds 50%. This means that a substantial percentage of the high-scoring windows do not correspond to confirmed replication origins. On closer examination of these high scoring windows which are not replication origins, some of them turn out to be regulatory sequences such as transcription factor binding sites. So far, we have not made use of palindromes to predict regulatory sites, but this would be an important area to explore.

Our prediction scheme is geared towards herpesviruses and still needs to be tested on other DNA viruses. There are a few other methods proposed for prediction of replication origins for bacterial, archaeal, and yeast genomes (Breier et al., 2004;

Mackiewicz et al., 2004; Salzberg et al., 1998; Zhang and Zhang, 2005). These methods, which are based on DNA asymmetry, flanking sequence similarity, z-curves, might be adapted to work on viral DNA as well.

CHAPTER 4

COMPOUND POISSON APPROXIMATION OF PALINDROME LENGTH SCORE

4.1 Introduction

In the previous chapter, we introduced several scoring schemes to measure spatial concentration of palindromic patterns in genomic sequences. The aim is to locate regions in herpesvirus genomes that has a high concentration of palindromic patterns and ultimately suggest them as potential replication origin sites. While our prediction methods are rather successful in terms of sensitivity measure, they lack a rigorous statistical significance criterion.

In this chapter, we will approximate the distributions of the window scores under the *Palindrome Length Score* (PLS) by a Compound Poisson distribution. The occurrence of a palindrome of length $2L$ and above is modeled by a Poisson process (Leung et al., 2005), and the actual length of this palindrome is modeled by a geometric distribution.

We will discuss very briefly some properties of the Compound Poisson Distribution, before going on to describe our approximation of the PLS. Based on this ap-

proximation, we will then locate windows with scores in the herpesvirus genomes that are statistically significant at the 5% and 1% level.

4.2 Implementing The Palindrome Length Score

Recall that the Palindrome Length Score assigns to a palindromic pattern appearing in a genomic sequence a score that is proportionate to its length. For each of the viruses listed in Table 3.1 on page 31, we do the following:

(1) Locate palindromes at or above a prescribed length;

(2) Score the palindromes according to the PLS scheme;

(3) Compute the score for each window of the genome according to the PLS scheme; and

(4) Select regions with high scores.

The reader may refer to Section 3.2 on page 30 for details.

Note that in the previous chapter, (4) was done by selecting a pre-determined number of top scoring windows. In this chapter, we approximate the PLS score of a sliding window using a *Compound Poisson* random variable. Based on this, we are able to locate windows that have statistically significant high scores.

4.3 Properties of the Compound Poisson Distribution

Before we proceed with the modeling of the PLS window score, it is perhaps timely to have a quick and brief review of the *Compound Poisson Distribution*.

Definition 4.1. *Let $X_1, X_2, \ldots$ be positive, integer valued, independent and identically distributed random variables with a common distribution F with finite moments up to a certain order. Let N be a Poisson random variable, independent of $X_1, X_2, \ldots$, with parameter λ.*

We define the Compound Poisson random variable S_N *to be*

$$S_N = \begin{cases} X_1 + \ldots + X_N, & N \geq 1 \\ 0, & N = 0 \end{cases}.$$

We now establish the Stein identity for the Compound Poisson distribution.

Proposition 4.2. *Let X be a random variable having the same distribution as in the definition of S_N above, and independent of $X_1, X_2, \ldots$ and N. Then, for $f : \mathbb{Z}^+ \to \mathbb{R}$ bounded, we have*

$$\lambda \mathbb{E}X f(X + S_N) = \mathbb{E}S_N f(S_N).$$

Proof.

$$\begin{aligned} \mathbb{E}S_N f(S_N) &= \sum_{n=0}^{\infty} \mathbb{E}\left[S_N f(S_N) \mid N = n\right] \mathbb{P}(N = n) \\ &= \sum_{n=1}^{\infty} \mathbb{E}\left[S_n f(S_n)\right] \mathbb{P}(N = n) \\ &= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \mathbb{E}\left[X_k f(S_n)\right] \mathbb{P}(N = n) \end{aligned}$$

By symmetry, $\mathbb{E}X_k f(S_n) = \mathbb{E}X_n f(S_n)$, for all $1 \leq k \leq n$ so the above expression becomes

$$\begin{aligned} \mathbb{E}S_N f(S_N) &= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \mathbb{E}\left[X_n f(S_n)\right] \mathbb{P}(N = n) \\ &= \sum_{n=1}^{\infty} n\mathbb{E}\left[X_n f(S_{n-1} + X_n)\right] \mathbb{P}(N = n) \\ &= \lambda \sum_{n=1}^{\infty} \mathbb{E}\left[X_n f(S_{n-1} + X_n)\right] \mathbb{P}(N = n-1) \\ &= \lambda \sum_{n=1}^{\infty} \mathbb{E}X f(S_{n-1} + X) \mathbb{P}(N = n-1) \\ &= \lambda \mathbb{E}X f(S_n + X). \end{aligned}$$

□

We will also have the following corollary:

Corollary 4.3. *Let $\alpha_k = \mathbb{P}(X = k)$, for $k \geq 1$. Then $\mathbb{P}(S_N = 0) = e^{-\lambda}$, and*

$$\mathbb{P}(S_N = n) = \frac{\lambda}{n} \sum_{k=1}^{n} k\alpha_k \mathbb{P}(S_N = n-k), \text{ for } n \geq 1$$

.

Proof. Firstly, $\mathbb{P}(S_N = 0) = \mathbb{P}(N = 0) = e^{-\lambda}$.

Using Proposition 4.2 with the indicator function I, we get

$$\begin{aligned} n\mathbb{P}(S_N = n) &= \mathbb{E}S_N I_{\{n\}}(S_N) = \lambda \mathbb{E}X I_{\{n\}}(S_N + X) \\ &= \lambda \sum_{j=1}^{\infty} j\alpha_j \mathbb{E} I_{\{n\}}(S_N + j) \\ &= \lambda \sum_{j=1}^{n} j\alpha_j \mathbb{P}(S_N = n-j). \end{aligned}$$

□

4.4 Modeling the Palindrome Length Score

The modeling of the Palindrome Length Score consists of the following stages:

1. **Probability model on DNA genome:**

   We model the genome sequence as a realization of a sequence of random variables $\xi_1, \xi_2, \ldots, \xi_n$ taking values in $\mathscr{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ where n is the genome length.

   We will assume that either

   a) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ are independent and identically distributed (M0); or

   b) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ form a stationary Markov chain of order one or two (M1 or M2).

2. **Occurrences of palindromes:**

   Let $L \geq 1$ be fixed, L is our lower cutoff, i.e., the minimum length of palindrome that we consider. For $L \leq k \leq n-L$, define the indicator random variable

$$I_k = \begin{cases} 1, & \text{if the } k\text{th base is the left center of a palindrome of length } \geq 2L \\ 0, & \text{otherwise} \end{cases}.$$

3. **Scoring the Palindrome:**

   Suppose there exist a palindrome of length at least $2L$ with left center at k. That is, assume that $I_k = 1$. Let $M > L$, where M denotes our upper cutoff for the palindrome length. For our application, we have $M = 3L$, which we will justify later in the chapter. We consider the following cases:

   a) $L \leq k < M$ or $n - M + 1 < k \leq n - L + 1$.

      For $L \leq s \leq k$, we define $X_k = s$ if there is a maximally extended palindrome of length exactly $2s$ with left center at k.

   b) $M \leq k \leq n - M + 1$.

      For $L \leq s < M$, we define $X_k = s$ if there is a maximally extended palindrome of length exactly $2s$ with left center at k.

      For $s = M$, we define $X_k = M$ if there is a maximally extended palindrome of length at least $2M$ with left center at k.

4. **Window score:**

   Recall that we construct overlapping windows along the span of the genome. Say a total of T of them. For a typical window (ignoring the edge effect), the window score W_i is given by

$$W_i = \sum_{(i-1)w/2 < k \leq (i+1)w/2} X_k I_k$$

   where w is the window width.

The random variable

$$V_i = \sum_{(i-1)w/2<k\le(i+1)w/2} I_k$$

counts the number of palindromes of length at least $2L$ in the window i.

4.5 Compound Poisson Approximation

For $1 \le i \le T$, we construct a compound Poisson random variable Z_i to approximate the window score as below. Let N_i denote a Poisson random variable of parameter

$$\lambda := w\mathbb{P}[\xi_1 \cdots \xi_{2L} \text{ forms a palindrome}].$$

Here N_i models the number of palindromes that occur in window i. To model the length of the palindromes, let $Y, Y_{i,1}, Y_{i,2}, \ldots$ be independent random variables taking values $L, L+1, \ldots, M$ with a common probability mass function p_Y to be specified later, i.e,

$$\mathbb{P}[Y = j] = p_Y(j), \qquad \text{for } L \le j \le M. \tag{4.1}$$

We remark that probability mass function of Z_i can be computed once the probability mass function as given by (4.1) is computed, and is given by the following recursive formula (See Corollary 4.3 on page 47)

$$\mathbb{P}[Z_i = k] = \frac{\lambda}{k} \sum_{j=1}^{k} j\, p_Y(j) \mathbb{P}[Z_i = k-j], \qquad \text{for } k \ge 1. \tag{4.2}$$

with initial value $\mathbb{P}[Z_i = 0] = e^{-\lambda}$.

4.6 Probability Mass Function of Y

The probability mass function of Y depends on the sequence model of the genome. We will show how to compute p_Y under the assumption of IID (M_0), or stationary Markov chain of order r with $r = 1$ or 2. Computation of p_Y when the underlying

probability model are higher order Markov chains can be done in a similar fashion.

1. **DNA sequence model is M_0:**

   Let $\theta = 2(p_A p_T + p_C p_G)$ and $\lambda = |w|\theta^L$, where θ is the probability that a pair of bases being complementary to each other. We will define p_Y to be

   $$p_Y(j) = \begin{cases} (1-\theta)\theta^{j-L}, & \text{if } L \le j < M \\ \theta^{M-L}, & \text{if } s = M \end{cases}.$$

2. **DNA sequence model is M_1:**

   Let $P(a,b)$ denote the transition probability of a to b for $a, b \in \mathscr{A}$. Let $(\pi(a))_{a\in\mathscr{A}}$ be the stationary distribution of this Markov chain. We shall illustrate how $\mathbb{P}[Y \ge j]$ can be computed, and hence $\mathbb{P}[Y = j]$.

   A brute force way to compute $\mathbb{P}[Y \ge j]$ is by exhaustive enumeration:

   $$\mathbb{P}[Y \ge j] = \sum_{\mathbf{w}\in\mathscr{A}^j} \mathbb{P}[\mathbf{w}\mathbf{w}']$$

   for $L \le j \le M$.

   For the herpesviruses, recall that we pick $L = \{5, 6\}$ and hence $M = 3L = \{15, 18\}$. Note that computation by exhaustive enumeration soon becomes impractical for when $j = M$, it will be summing about 68 billion ($4^{18} = 68,719,476,736$) terms. Fortunately, we have a dynamic programming algorithm to do the computation effectively.

Proposition 4.4 (The Outside-In Algorithm for M_1)**.**

a) Define, for $a \in \mathscr{A}$, $\alpha_1(a) = P(a, a')$.

b) For $m \geq 2$ and $a \in \mathscr{A}$, define

$$\alpha_m(a) := \sum_{b \in \mathscr{A}} P(a,b)\, \alpha_{m-1}(b)\, P(b', a').$$

c) Then

$$p_Y(m) = \sum_{a \in \mathscr{A}} \pi(a)[\alpha_m(a) - \alpha_{m+1}(a)]. \tag{4.3}$$

Proof. Let $\xi_1 \cdots \xi_{2m} \in \mathscr{P}$ to denote that $\xi_1 \cdots \xi_{2m}$ forms a palindrome. We claim that

$$\mathbb{P}[\xi_1 \cdots \xi_{2m} \in \mathscr{P} | \xi_1 = a] = \alpha_m(a), \qquad a \in \mathscr{A},\ m \geq 1. \tag{4.4}$$

Assuming that (4.4) holds, then

$$\begin{aligned} \mathbb{P}[Y \geq m] &= \mathbb{P}[\xi_1 \cdots \xi_{2m} \in \mathscr{P}] \\ &= \sum_{a \in \mathscr{A}} \pi(a) \mathbb{P}[\xi_1 \cdots \xi_{2m} \in \mathscr{P} | \xi_1 = a] \\ &= \sum_{a \in \mathscr{A}} \pi(a) \alpha_m(a) \end{aligned}$$

proving (4.3). We shall prove (4.4) by induction.

For $m = 1$,

$$\mathbb{P}[\xi_1 \xi_2 \in \mathscr{P} | \xi_1 = a] = \mathbb{P}[aa'] = P(a, a') = \alpha_1(a)$$

showing that (4.4) holds for $m = 1$.

Assuming (4.4) holds for m, then

$$\begin{aligned}
&\mathbb{P}[\xi_1 \cdots \xi_{2m+2} \in \mathscr{P} | \xi_1 = a] \\
&= \mathbb{P}[a\xi_2 \cdots \xi_{2m+1} a' \in \mathscr{P} | \xi_1 = a] \\
&= \sum_{b \in \mathscr{A}} \mathbb{P}[ab\xi_3 \cdots \xi_{2m} b'a' \in \mathscr{P} | \xi_1 = a] \\
&= \sum_{b \in \mathscr{A}} \mathbb{P}[ab\xi_3 \cdots \xi_{2m} b'a' \in \mathscr{P} | \xi_1 = a, \xi_2 = b] \mathbb{P}[\xi_2 = b | \xi_1 = a] \\
&= \sum_{b \in \mathscr{A}} P(a,b) \mathbb{P}[b\xi_3 \cdots \xi_{2m} b' \in \mathscr{P} | \xi_2 = b] P(b', a') \\
&= \sum_{b \in \mathscr{A}} P(a,b) \alpha_m(b) P(b', a') \\
&= \alpha_{m+1}(a)
\end{aligned}$$

showing that (4.4) holds for $m+1$. Induction finishes the proof. □

3. **DNA sequence model is M_2:**

Let $P(ab,c)$ denote the transition probability of the dinucleotide ab to c for $a,b,c \in \mathscr{A}$. Let $(\pi(ab))_{ab \in \mathscr{A}^2}$ be the stationary distribution of this Markov chain of order 2. The Outside-In algorithm can be extended to the M_2 case.

Proposition 4.5 (The Outside-In Algorithm for M_2)**.**

a) Define, for $a,b \in \mathscr{A}$,

$$\beta_1(a,b) = \begin{cases} 1, & \text{if } b = a' \\ 0, & \text{otherwise} \end{cases}.$$

b) For $m \geq 2$ and $a,b \in \mathscr{A}$, define

$$\beta_m(a,b) := \sum_{c \in \mathscr{A}} P(ab,c)\, \beta_{m-1}(b,c)\, P(c'b', a').$$

c) Then

$$p_Y(m) = \sum_{a \in \mathscr{A}} \pi(ab)[\beta_m(a,b) - \beta_{m+1}(a,b)]. \tag{4.5}$$

Proof. The proof is similar to the above. We observe that

$$\beta_2(a,b) = \sum_{c\in\mathscr{A}} P(ab,c)\beta_1(b,c)P(c'b',a') = P(ab,b')P(bb',a')$$
$$= \mathbb{P}[\xi_1\xi_2\xi_3\xi_4 \in \mathscr{A} | \xi_1\xi_2 = ab].$$

For $m \geq 2$,

$$\begin{aligned}
&\mathbb{P}[\xi_1\cdots\xi_{2m+2} \in \mathscr{A} | \xi_1\xi_2 = ab] \\
&= \sum_{c\in\mathscr{A}} \mathbb{P}[abc\xi_4\cdots\xi_{2m-3}c'b'a' \in \mathscr{A} | \xi_1\xi_2 = ab] \\
&= \sum_{c\in\mathscr{A}} \mathbb{P}[abc\xi_4\cdots\xi_{2m-3}c'b'a' \in \mathscr{A} | \xi_1\xi_2\xi_3 = abc]\mathbb{P}[\xi_3 = c | \xi_1\xi_2 = ab] \\
&= \sum_{c\in\mathscr{A}} P(ab,c)\mathbb{P}[bc\xi_4\cdots\xi_{2m-3}c'b' \in \mathscr{A} | \xi_2\xi_3 = bc]P(c'b',a') \\
&= \sum_{c\in\mathscr{A}} P(ab,c)\beta_m(b,c)P(c'b',a') \\
&= \beta_{m+1}(a,b).
\end{aligned}$$

□

4.7 Goodness of Approximation

We now proceed to demonstrate that the Compound Poisson random variable we constructed approximates the window score under the Palindrome Length Score well.

Recall that for our model we needed to apply a cut-off to the (stem) length of the palindromes we consider (see Section 4.4 on page 48). We have chosen the lower cut-off for the stem length of the palindromes to be L (which is 5 for most of the viruses in our herpesvirus data set and 6 for a handful of others) and M to be the upper cut-off.

A few questions arise? What would be a good upper cut-off for M? How good is our compound Poisson approximation? To answer these questions, we resort to simulation.

This is what we did (for each of the viral genomes in our data set):

1. Under the assumed Markov chain order ($M0$ or $M1$), estimate the transition probabilities from the real DNA sequence.

2. Simulate 10,000 DNA sequences (of length w each) using the estimated transition probabilities and stationary distribution.

3. Form the empirical distribution of window score under the PLS from the window scores of these 10,000 simulated windows.

4. Measure the discrepancy of the Compound Poisson distribution and the empirical distribution by the *Total Variational distance* and *Kolmogorov distance*.

   The *Total Variational distance* d_{TV} between two discrete random variables X and Y is given by

   $$d_{TV} = \sum_k (\mathbb{P}(X = k) - \mathbb{P}(Y = k))_-$$

   and the *Kolmogorov distance* d_K is given by

   $$d_K = \sup_k |P(X \leq k) - P(Y \leq k)|.$$

We have to decide upon a good value of M to be used. To answer this question, we picked several representative members of the herpesvirus family, namely the hcmv, vzv, ebv, hsv1 and cehv1 to form our training set. By letting the value of M to be $2L$, $3L$ and $4L$, we were able to measure the total variational and Kolmogorov distances between our compound Poisson and empirical distributions under those assumed values of M. In fact, we did a total of two runs of simulations under the $M0$ model. Table 4.1 on the next page gives the results of these simulation studies.

As you will discover from the table, there is an substantial improvement in the values of d_{TV} and d_K (i.e, the distance between the theoretical compound poisson and the empirical distribution gets smaller) when we change the value of M from $M = 2L$ to $M = 3L$, but no substantial change when we change M from $M = 3L$ to

Table 4.1 *– Total Variational Distance (d_{TV}) and Kolmogorov Distance (d_K) between the Compound Poisson and Empirical Distributions for the training set.*

			1st Run		2nd Run			
	L	M	d_{TV}	d_K	d_{TV}	d_K	min(d_{TV})	min(d_K)
		M=2L						
hcmv	5	10	0.034657	0.026505	0.045732	0.033084	0.034657	0.026505
vzv	5	10	0.013849	0.010177	0.018386	0.010799	0.013849	0.010177
ebv	5	10	0.036998	0.022327	0.028022	0.020017	0.028022	0.020017
hsv1	6	12	0.012643	0.008720	0.015082	0.010020	0.012643	0.008720
cehv1	6	12	0.025797	0.016437	0.026785	0.013875	0.025797	0.013875
		M=3L						
hcmv	5	15	0.013494	0.004203	0.017642	0.004435	0.013494	0.004203
vzv	5	15	0.005284	0.001312	0.009679	0.002863	0.005284	0.001312
ebv	5	15	0.016800	0.013452	0.009851	0.007279	0.009851	0.007279
hsv1	6	18	0.010954	0.008720	0.013199	0.010020	0.010954	0.008720
cehv1	6	18	0.011316	0.004575	0.019632	0.013875	0.011316	0.004575
		M=4L						
hcmv	5	20	0.013485	0.004203	0.017626	0.004435	0.013485	0.004203
vzv	5	20	0.005280	0.001312	0.009675	0.002863	0.005280	0.001312
ebv	5	20	0.016788	0.013452	0.009840	0.007279	0.009840	0.007279
hsv1	6	24	0.010953	0.008720	0.013199	0.010020	0.010953	0.008720
cehv1	6	24	0.011307	0.004575	0.019630	0.013875	0.011307	0.004575

$M = 4L$. Based on these observations, we decide to select $M = 3L$ across the board for all the viral genomes in our data set.

Using $M = 3L$ we proceed to compute d_{TV} and d_K for all the viruses in our data set. Table 4.4 on page 59 gives the details of the d_{TV}'s and d_K's for all the viruses in our data set. Table 4.2 gives some statistics of these distances.

Table 4.2 *– Summary for Total Variational Distance (d_{TV}) and Kolmogorov Distance (d_K) between the Compound Poisson and Empirical Distributions.*

	M0		M1	
	d_{TV} in 10^{-2}	d_K in 10^{-2}	d_{TV} in 10^{-2}	d_K in 10^{-2}
minimum	0.503	0.246	0.409	0.251
maximum	2.166	1.868	3.052	2.683
mean	1.310	0.799	1.432	0.939
standard deviation	0.383	0.362	0.490	0.457

These results demonstrates that our compound Poisson approximation of the window score under PLS is good. We will then be able to use this compound Poisson approximation to come out with a cut-off for the window scores under the PLS

scheme say at 1%, and use that to determine windows that have statistically significant high scores.

4.8 Identifying High Scoring Windows

In the previous section, we have established that the compound Poisson random variable is a good approximation of the PLS window score. We now have the means to identify statistically high scoring windows.

Recall that the genome is "covered" by T windows. We want to approximate the distribution of $W^* := \max_{1\le i\le T} W_i$, the maximum of all the window scores W_i. Let c be the 95 (or 99) percentile of W^*. To determine c, we apply the usual Poisson approximation argument.

$$\begin{aligned} 0.05 = \mathbb{P}[W^* \ge c] &= \mathbb{P}\left[\sum_{i=1}^{T} I(W_i \ge c)\right] \\ &\approx 1 - \exp\left\{-\sum_{i=1}^{T} \mathbb{P}[W_i \ge c]\right\} \\ &= 1 - \exp\{-T\mathbb{P}[W_1 \ge c]\} \\ &\approx 1 - \exp\{-T\mathbb{P}[Z_1 \ge c]\}. \end{aligned}$$

Based on Equation (4.2) on page 50, c can be chosen so that

$$\mathbb{P}[Z_1 \ge c] = -\frac{\log 0.95}{T}.$$

We present in Tables 4.5 on page 60 and 4.6 on page 61 the high-scoring windows under the M_0 and M_1 models respectively, at both 5% and 1%.

As in the previous chapter, to assess the prediction performance of these schemes, we look at the sensitivity and positive predictive power of them. The summary of the performance is given in Table 4.3.

Notice that the sensitivity of our prediction decreases as compared to the results

Table 4.3 *– Prediction performance of PLS with compound Poisson approximation.*

	M0		M1		
	1%	5%	1%	5%	PLS (10 windows)
Sensitivity	47	51	47	49	65
PPV	25	28	25	27	14

for PLS (using 10 windows) in the previous chapter. However, this is not surprising to us. We had expected that when a cut-off is applied, we would have lesser windows to be put forth as potential replication origin sites, and hence lesser replication origins will be predicted. We are however glad to see that the positive predictive power nearly doubles for both approximation schemes (M0 and M1) at 1% and 5%. This means that we will have lesser false prediction when we use such an approximation scheme. Finally, we could not stress enough that the advantage that this approach has over the non-parametric approach in the previous chapter is that we have able to know the statistical significance of the scores of the windows that we use as potential replication origin sites.

Last but not least, we do want to remark that the compound Poisson approximation for the BWS scheme as described in the previous chapter has not yet been worked out. Recall that the prediction performance of the BWS is better than that of the PLS. We would then expect that under the corresponding compound Poisson approximation we will get results that will be better that those we have seen in this chapter.

Table 4.4 – *Total Variational Distance (d_{TV}) and Kolmogorov Distance (d_K) between the Compound Poisson and Empirical Distributions under M0 and M1 model.*

			M0		M1	
Virus	L	M=3L	d_{TV}	d_K	d_{TV}	d_K
alhv1	5	15	0.00990927	0.00490914	0.00409128	0.00250852
athv3	5	15	0.01546954	0.01377820	0.01880386	0.01496876
bohv1	6	18	0.01887245	0.01475483	0.03052115	0.02683041
bohv4	5	15	0.02077680	0.01868002	0.01645188	0.01413331
bohv5	6	18	0.01112975	0.00531951	0.02040017	0.01520402
calhv3	5	15	0.01091473	0.00821730	0.00777838	0.00501029
ccmv	5	15	0.01688846	0.00631984	0.01575265	0.00817880
cehv1	6	18	0.01564512	0.01175157	0.01692463	0.01381026
cehv15	5	15	0.01588958	0.00949796	0.01520027	0.01158785
cehv16	6	18	0.01655194	0.00796564	0.02133380	0.01455390
cehv2	6	18	0.01819161	0.01035406	0.01918128	0.01026972
cehv7	5	15	0.01408748	0.01005454	0.01134546	0.00617778
cehv8	5	15	0.01217465	0.00334885	0.01550555	0.00582608
ebv	5	15	0.02166096	0.01079030	0.01331154	0.00891508
ehv1	5	15	0.00967108	0.00277337	0.01215583	0.00544480
ehv2	5	15	0.01592747	0.00915296	0.00995649	0.00604436
ehv4	5	15	0.00976372	0.00430894	0.01227365	0.00932433
gahv1	5	15	0.00968592	0.00655585	0.01347043	0.01023047
gahv2	5	15	0.01235920	0.00584615	0.00942289	0.00701855
gahv3	5	15	0.00918203	0.00501825	0.01047886	0.00394285
hcmv	5	15	0.01395996	0.00661273	0.02144730	0.00924072
hcmv-m	5	15	0.01416455	0.00898589	0.02041383	0.01258359
hhv6	5	15	0.00961172	0.00647553	0.01597751	0.00757296
hhv6b	5	15	0.01241257	0.00723543	0.01577835	0.01131182
hhv7	5	15	0.01405826	0.00571736	0.01034384	0.00352412
hhv8	5	15	0.01053258	0.00821349	0.01094572	0.00718649
hsv1	6	18	0.00742916	0.00622034	0.01016645	0.00784827
hsv2	6	18	0.01501359	0.01334275	0.01526228	0.01254315
ichv1	5	15	0.01191611	0.00910209	0.00980290	0.00660671
mcmv	5	15	0.01365551	0.00595827	0.01509266	0.00597760
mehv1	5	15	0.01068748	0.00537938	0.00976498	0.00486943
mfrv	5	15	0.00800154	0.00398064	0.00876273	0.00631488
mmrv	5	15	0.00863562	0.00399538	0.01848031	0.01552133
muhv4	5	15	0.01130150	0.00794064	0.01027370	0.00930311
ohv2	5	15	0.00751198	0.00526105	0.00959869	0.00641516
oshv1	5	15	0.01586530	0.00867926	0.01300445	0.00336977
pshv1	5	15	0.01409341	0.00596304	0.02154577	0.01596890
rcmv	5	15	0.01924567	0.01268640	0.01592607	0.00829536
sahv2	5	15	0.01510956	0.01216842	0.01458957	0.00971274
shv1	6	18	0.01699541	0.01396658	0.01723847	0.01264602
thv	6	18	0.00502617	0.00246431	0.01222325	0.00970855
vzv	5	15	0.01029122	0.00566254	0.01044034	0.00799625

Table 4.5 – *Windows with scores exceeding the critical score at 5% for M0 Model. Rows on upper half list viruses with known replication origins, those on lower half without. Entries in bold indicate that window score is also significantly high at 1%. Underlined entries indicate that window is within 2mu of some known ORI.*

Virus	Mid point of Window
bohv1	**105901,113401,124501,132301,77401,87601,82801,30901,35701,4801,51901,** 74401, 70201,96901
bohv4	
bohv5	**78001,108301,134701,19201,36901,6601,33901,61801,84301,31501,67501,94201, 53101,118501,** 10501,90901,101401,4201
cehv1	**133001,149451,61601,**113051,125301,701 ,102201,109901,32551,36401,50751,117601, 144551,154701
cehv2	**129501,144201,61601,**123551,22051,75951,107101,92401, 111301,150501,35351,8051, 32201,115151
cehv7	
cehv16	**118301,8751,21001,37801,137201,33251,154001,** 51451,102901,77701, 47251,132301
ebv	7601,141201,41201
ehv1	116201,146651
ehv4	
gahv1	
hcmv	94051
hhv6	
hhv6b	90401
hhv7	
hsv1	**62301,129851,148401,72801,** 1051,124951
hsv2	**74551,7351,119701,28001,128801,152951,** 22401,45151
rcmv	**75901,** 110551,83601,101751,127601
shv1	**37801,58451,93101,30451,85051,78751,124601,75251,11551,** 20301,6651
vzv	**119401,** 110101,100501
alhv1	
athv3	
calhv3	116201,133351,23101
ccmv	91201,207001,177001
cehv15	8001,34801,138801
cehv8	161151,147401
ehv2	**6301,54001,173251,** 140401
gahv2	
gahv3	**158801,138401,** 11201,122401
hcmv-m	**175451,94051,** 153451
hhv8	23401
ichv1	
mcmv	**92951,142451,** 200201
mehv1	
mfrv	130501
mmrv	132601
muhv4	
ohv2	**117601,** 134401
oshv1	
pshv1	130401,151601
sahv2	**103751,** 112501
thv	**134101,10801,50401,144901,85051,107551,58501,163801,54451,** 157951,38251,181801

Table 4.6 *– Windows with scores exceeding the critical score at 5% for M1 Model.*

Virus	Mid point of Window
bohv1	**105901,113401,124501,132301,77401,87601,82801,30901,35701,**4801
bohv4	
bohv5	**78001,108301,134701,19201,36901,6601,33901,61801,84301,31501,67501,**94201,53101
cehv1	**133001,149451,61601,113051,125301,701,102201,109901,32551,36401,50751,117601,** 144551,154701
cehv2	**129501,144201,61601,123551,22051,75951,107101,92401,**111301,150501,35351,8051, 32201,115151
cehv7	
cehv16	**118301,8751,21001,37801,137201,33251,154001,**51451,102901,77701
ebv	7601,141201,41201
ehv1	116201,146651
ehv4	
gahv1	
hcmv	94051
hhv6	
hhv6b	90401
hhv7	
hsv1	**62301,129851,148401,72801,**1051,124951
hsv2	**74551,7351,119701,28001,**128801,152951,22401,45151
rcmv	**75901,**110551
shv1	**37801,58451,93101,30451,85051,78751,124601,75251,11551,**20301,6651
vzv	**119401,**110101
alhv1	113701
athv3	
calhv3	116201,133351,23101
ccmv	91201,207001,177001
cehv15	**8001,34801,138801,**109201,152001
cehv8	161151,147401
ehv2	**6301,54001,173251,**140401
gahv2	
gahv3	**158801,138401,**11201,122401
hcmv-m	175451,94051
hhv8	23401
ichv1	
mcmv	92951,142451
mehv1	
mfrv	130501
mmrv	132601
muhv4	99251
ohv2	**117601,**134401
oshv1	
pshv1	**130401,**151601
sahv2	103751,112501
thv	**134101,10801,50401,144901,85051,107551,58501,163801,54451,**157951

4.9 Binomial Approximation to the AT Sliding Window Score

We reported in the last section of Chapter 3 using sliding windows of AT percentages as a prediction tool. Based on windows with top AT percentages we were able to predict 28 replication origins out of 43 known origins in the herpesviruses.

In this section we want to describe our attempt to approximate the AT content sliding window scores using a Binomial distribution. Think of the AT content in a typical sliding window of length w as a realization of a Binomial distribution with parameters (w, b), where b is the probability of success, which in our case corresponds to the event the base A or T is chosen.

Suppose again that there are T sliding windows constructed for a given viral genome. For each $1 \leq i \leq T$, we construct a Binomial random variable W_i with parameters (w, b) where b will be estimated using the global AT frequency of that particular genome. Note when $|i - j| > 1$, windows i and j do not overlap and hence W_i will be independent of W_j. This fact will be useful in the derivation of Equation (4.6).

We are interested to estimate the probability that the maximum of A plus T base count amongst all sliding windows exceeds a certain number, x, given by

$$\mathbb{P}\left(\max_{1\leq i\leq T} W_i \geq x\right).$$

The aim then is to preset this probability to some significance level, say 5% and find the critical x value. Windows with AT counts exceeding this critical x would then be deemed to be abundant in AT content at the significance level chosen.

Note that by Boole's Inequality

$$\sum_{1\leq i\leq T} \mathbb{P}(W_i \geq x) - \sum_{1\leq i<j\leq T} \mathbb{P}(W_i \geq x, W_j \geq x) \leq \mathbb{P}\left(\max_{1\leq i\leq T} W_i \geq x\right) \leq \sum_{1\leq i\leq T} \mathbb{P}(W_i \geq x).$$

Using the fact that W_i and W_j are independent when $|i - j| > 1$ and some simpli-

fication, we arrive at

$$T\alpha - \frac{(T-1)(T-2)}{2}\alpha^2 + (T-1)\mathbb{P}(W_1 \geq x, W_2 \geq x) \leq \mathbb{P}\left(\max_{1\leq i\leq T} W_i \geq x\right) \leq T\alpha, \quad (4.6)$$

where $\alpha = \mathbb{P}(W_1 \geq x)$.

We had in fact estimated the lower and upper bounds of the above inequality numerically and found them to be tight. This means that we can use the following approximation

$$\mathbb{P}\left(\max_{1\leq i\leq T} W_i \geq x\right) \approx T\mathbb{P}(W_1 \geq x), \quad (4.7)$$

which the right hand side term can be easily computed.

Setting the right hand side term of Equation (4.7) to some predetermined value, say p%, we are able to locate windows with statistically significant scores at p% for each of the viral genomes in our data set. In fact, when we set the significance level to 5%, the set of windows with statistically significant AT count (content) was able to predict correctly 31 out of 43 known replication origins. Compared to the 51% sensitivity (22 origins correctly predicted) that the PLS compound Poisson approximation yielded, this is a huge improvement.

CHAPTER 5

AT EXCURSIONS FOR PREDICTION OF REPLICATION ORIGINS

5.1 Background

Besides the methods described in the previous chapters, many computational methods to predict likely locations of replication origins have also been developed for the prediction of replication origins in bacterial, archaeal and yeast genomes. All these methods exploit certain sequence features often found around the replication origins for their prediction. For example, Lobry (1996) employed the GC skew plot to predict replication origins and terminals in bacterial genomes. The skew, calculated as (G-C)/(G+C) for a window sliding along the sequence, was shown to switch polarity in the vicinity of the terminus and replication origin, with the leading strand manifesting a positive skew. This method is commonly used to identify the putative oriC region within chromosomes, particularly before experimental analysis. However, when we applied the GC skew plot to the herpesviruses, no clear cut switches of polarity could be observed.

Salzberg et al. (1998) predicted the replication origins for a number of bacterial and archael genomes based on the identifying some 7-mers and/or 8-mers whose orientation is preferentially skewed around the replication origins. However, as pointed out by the authors, this method may not suited for many viral DNA genomes with multiple replication origins. Breier et al. (2004) developed the Oriscan algorithm to predict the exact location of replication origins in S. cerevisiae genome. The algorithm searched for sequences similar to a training set of 26 known yeast origins that were pinpointed by site-directed mutagenesis. Oriscan uses both the origin recognition complex binding site and its flanking regions to identify candidates, and it then ranks potential origins by their likelihood of activity. Zhang and Zhang (2005) applied the Z-curve method successfully to identify several replication origins in bacterial and archaeal genomes. The Z-curve is a three-dimensional curve that constitutes a unique representation of a DNA sequence. This means that for any DNA sequence and its associated Z-curve, each can be uniquely reconstructed from the other. One of the advantages of the Z-curve is its intuitiveness; the entire Z-curve of a genome can be viewed on a computer screen or on paper, regardless of genome length, thus allowing both global and local compositional features of genomes to be easily grasped.

These methods do not seem to work well in predicting the likely locations of replication origins in viral genomes with multiple replication origins.

A simple, yet natural, sequence feature that can possibly be exploited to predict the locations of replication origins in the doubly stranded herpesviruses is the AT content. Segments of DNA with high GC content, i.e., lower AT content, are more stable and hence less likely candidates for replication origins. Segurado et al. (2003) used a sliding window approach to find "islands" within the *Schizosaccharomyces pombe* genome that have high AT content. They measured base composition using sliding windows of different sizes and found that the highest A+T content for each window was significantly higher for ORI-containing regions than for regions that replicated passively.

It has also been observed that regions around the replication origins are rich in AT (see Chapter 1 in Kornberg and Baker (1992), Bramhill and Kornberg (1988)). Chew et al. (2005) reported using sliding windows of AT percentages. Based on windows with top AT percentages they were able to predict 28 replication origins out of 43 known origins in the herpesviruses. Moreover, 4 origins, which were predicted by AT percentages, failed to be detected by their based weighted score method. We are thus led to adopt a more refined score based approach as in Karlin (1994) to quantitate the AT content and hence a computational method to predict the replication origins in the herpesviruses. This score based approach has a further advantage as Karlin and his collaborators have worked out the limiting statistical distribution which enables us to identify statistically significant high scoring segments.

There are 3 main objectives in this chapter. Our first objective is to adopt Karlin's score based approach to quantitate local AT abundance reflecting the genome's base pairs composition. Moreover, this approach does away the choice of window size. We then develop a computational method, called AT excursion method, to complement the existing prediction methods. The second objective is to apply the AT excursion method to predict the replication origins in herpesviruses. And from known locations of the replication origins, we can then assess the performance of this method. Our result demonstrates that the AT excursion method compares very well with the other methods, and this method is also shown to complement these methods. Having established that AT excursion method is a credible prediction tool, our third objective is to apply the AT excursion method to predict the locations of replication origins in two other classes of viruses, the *Iridoviruses* and the *Poxviruses*. These two families are chosen because, like the herpesviruses, they are double stranded viruses with no RNA stage and their lengths are similar in magnitude to that of the herpesviruses. Moreover, the replication origins of these two classes of viruses are either unknown or not available in the public domain. Indeed, amongst these two classes of viruses, we could only find one virus with 6 known replication origins, when we checked the literature.

5.2 Methods

We propose a computational method to identify segments of a genome that have high AT concentration. This, in turn, forms the basis of our method to predict replication origins for the herpesviruses. As with the previous chapters, Table 3.1 on page 31 presents the viruses to be analyzed. The data set comprises all complete genome sequences of the herpesvirus family downloaded from GenBank at the NCBI web site in March 2006. For each virus, we list its abbreviation, accession number, sequence length, and the relative frequencies of the four nucleotide bases in the genome.

The approach that we adopt here will be score-based sequence analysis.

5.2.1 Score-based sequence analysis

The aim of score-based sequence analysis is to identify segments of DNA sequences with high additive scores by assigning appropriate scores to individual residues in those sequences.

Karlin and his collaborators were among the first to use this approach to identify interesting biological features using various score schemes. For details, see, for example, Karlin (1994, 2005); Karlin and Altschul (1990, 1993); Karlin et al. (1992).

5.2.2 Scoring the bases.

In this chapter, we are interested to find segments of DNA sequences with high AT concentration. We classify the four nucleotide bases {A, C, G, T} as "strongly bonding" or "weakly bonding" bases, denoted by S and W respectively. Under this formulation, S bases (C or G) are given a score of s_s and W bases (A or T), a score of s_w.

The probabilities $p_s := P(\text{base chosen is S})$ and $p_w := P(\text{base chosen is W})$ are estimated using the relative frequencies of the four nucleotide bases in the genome we are considering.

5.2.3 Probability Model.

The sequence model we will work with is adapted from the work of Karlin and his collaborators. For the general mathematical theory, interested readers may refer to Karlin et al. (1990) and Dembo and Karlin (1991a) for details.

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed letters drawn from the alphabet set {S, W} with associated scores $\{s_s, s_w\}$ such that $P(X = \mathtt{S}) = p_s$, $P(X = \mathtt{W}) = p_w$, where $p_s = 1 - p_w > 0$. The interpretation is that if you sample the letter W, say, the score associated with that draw is $X = s_w$. In order to have meaningful conclusions, we further require that the expected score per base $\mu = p_s s_s + p_w s_w$ should be negative, with at least one of the scores, s_w in our case, taking positive value.

Following a hint from Karlin (1994), we let $s_w = 1$ and chose s_s to be a (negative) integer so that the expected score per base, $\mu = p_s s_s + p_w s_w$ is close to the value of -0.5. In fact, as we prefer to deal with integer-valued scores, s_s is chosen to be

$$\lfloor \tfrac{\mu - p_w s_w}{p_s} \rfloor,$$

where $\mu = -0.5$ and $\lfloor \cdot \rfloor$ denotes the *integer floor* function.

5.2.4 Excursions and their value.

We next compute the cumulative scores and seek to identify segments of the genome that have significantly high scores. As we are only interested in segments with positive additive scores, we reset our cumulative scores to zero whenever it becomes non-positive.

The *excursion scores* E_i are defined recursively as

$$E_0 = 0, \qquad E_i = \max\{E_{i-1} + X_i, 0\}, \qquad \text{for } 1 \le i \le n.$$

Using this recursive definition, we are able to construct "excursions" for each of the genomes. An *excursion* starts at a point i where E_i is zero and ends at $j > i$ where

E_j first becomes zero. The score then stays at zero until it first becomes positive again for the start of the next excursion. The *value* of an excursion is defined to be the peak score during the course of that particular excursion.

5.2.5 Distribution of the Maximal Aggregate Score.

For each value of x, the maximal aggregate score

$$M_n = \max_{1 \le k \le n} E_k$$

satisfies

$$P\left(M_n > \frac{\ln n}{\lambda^*} + x\right) \approx 1 - \exp\{-K^* e^{-\lambda^* x}\}, \tag{5.1}$$

where λ^* is the unique positive solution to the equation

$$E\left(e^{\lambda X}\right) = p_s e^{\lambda s_s} + p_w e^{\lambda s_w} = 1$$

and K^* is a parameter given by an explicit series expansion (See Karlin and Altschul (1990)).

When X is a lattice variable of span δ, we have a simpler expression for K^* (See Karlin et al. (1990)):

$$\begin{aligned} \exp\{-K_+ e^{-\lambda^* x}\} &\le \liminf_{n\to\infty} P\left(M_n - \frac{\ln n}{\lambda^*} < x\right) \\ &\le \limsup_{n\to\infty} P\left(M_n - \frac{\ln n}{\lambda^*} < x\right) \\ &\le \exp\{-K_- e^{-\lambda^* x}\}, \end{aligned}$$

where

$$K_- = \frac{\lambda^* \delta}{e^{\lambda^* \delta} - 1} K^*, \qquad K_+ = \frac{\lambda^* \delta}{1 - e^{\lambda^* \delta}} K^*. \tag{5.2}$$

For the simple score scheme with values $\{-m, \ldots, -1, 0, 1\}$ occurring with proba-

bilities $\{p_{-m},\ldots,p_{-1},p_0,p_1\}$ we have,

$$K_- = (e^{-\lambda^*} - e^{-2\lambda^*})E\left(Xe^{\lambda^* X}\right).$$

We can set the left hand side of Equation (5.1) to some predetermined significance level, say $P = 0.01$, and solve for x. A segment with score exceeding $M_P = \frac{\ln n}{\lambda^*} + x$ is then said to be significant at the P level.

For our approach, we use K_- in place of K^* in Equation (5.1) for a "conservative" estimate of the probability and K_+ for a "generous" one.

5.2.6 High-scoring Segments.

We use Equation (5.1) with $P = 0.05$ and $P = 0.01$ to get $M_{0.05}$ and $M_{0.01}$ respectively. If the value of an excursion exceeds the critical values $M_{0.05}$ or $M_{0.01}$, then the segment from the beginning of the excursion up to the base where the peak value is realized is known as a high-scoring segment (HSS), significant at the 5% or 1% level.

We show the excursion plot of the Human Herpesvirus 3 (the VZV virus) in Figure 5.1.

For each of the viral genomes list in Table 3.1, we obtain a set of high-scoring segments, significant at the 0.05 (or 0.01) level. In each set of high-scoring segments, it is common to find that several of them are actually very close to one another. We thus apply a filtering procedure so that, if this happens, we will only take one out of several "neighboring" excursions as a "representative" for that part of the genome.

Table 5.4 on page 79 lists the high-scoring segments for each virus in the *Herpesviridae* family.

5.2.7 Prediction Performance.

The high-scoring segments are then checked against known replication origins in herpesviruses to evaluate their performance as a prediction tool for replication origins.

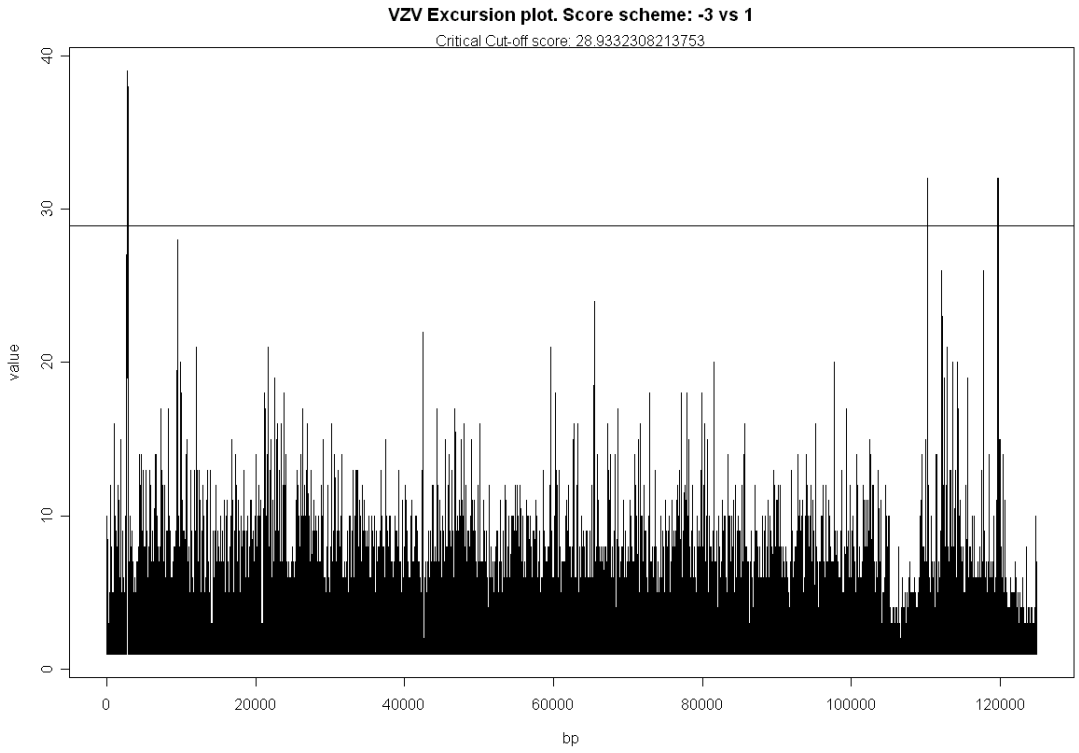


Figure 5.1 *– The Excursion Plot of the VZV virus.*

We list in Table 5.1 on the following page all the known replication origins for the viruses in the Herpesviridae family. These replication origins are reported either in published literature or GenBank annotations. For each replication origin, we list the high-scoring segment (at 5% level) closest to it. For this table we had used the "conservative" estimate for the value of K^* (See Equations (5.1) and (5.2)). When the peak of a high-scoring segment is less than 2 map units away from the center of a replication origin, we say that our method has correctly predicted that particular replication origin.

From Table 5.1 on the next page, we see that of the 43 replication origins known to us, 32 of them are close to the high-scoring segments that we have identified. This suggests that regions with high `AT` concentration are potential replication origin sites.

We had also tried using the "generous" estimate for K^* at the 5% and 1% level of significance. Table 5.2 on page 73 gives a summary of the performance of our prediction scheme when those bounds were used. The first two columns of the table gives the *sensitivity* level and *positive prediction power* of our scheme. *APD (average*

Table 5.1 *– Prediction results at 5% level using the conservative bound.*

		Nearest HSS			
Virus	**Ori Center**	**Start**	**Peak**	**Value**	**Prediction**
bohv1	111190	109702	109730	25	Yes
bohv1	127028	128487	128515	25	Yes
bohv4	97996.5	60687	60826	35	No
bohv5	113312	113549	113583	28	Yes
bohv5	129701	129429	129463	28	Yes
cehv1	61690.5	61680	61700	20	Yes
cehv1	61893.5	61680	61700	20	Yes
cehv1	132795.5	132785	132805	20	Yes
cehv1	132998.5	132785	132805	20	Yes
cehv1	149425.5	149415	149435	20	Yes
cehv1	149628.5	149415	149435	20	Yes
cehv16	62981	62970	62991	21	Yes
cehv16	133479	133468	133489	21	Yes
cehv16	149824	149813	149834	21	Yes
cehv2	61493.5	61483	61503	20	Yes
cehv2	129537.5	129527	129547	20	Yes
cehv2	144471.5	144461	144481	20	Yes
cehv7	109636.5	86167	86296	37	No
cehv7	118622.5	86167	86296	37	No
ebv	8313.5	11854	11950	45	No
ebv	40797	43158	43235	23	Yes
ebv	143825.5	77111	77150	24	No
ehv1	126262.5	128924	128992	23	Yes
ehv4	73909.5	73340	73509	37	Yes
ehv4	119471.5	112929	112967	29	No
ehv4	138577.5	132383	132462	49	No
gahv1	24871.5	24852	24890	30	Yes
hcmv	93923.5	96685	96824	34	Yes
hhv6	67805	130410	130501	59	No
hhv6b	69160.5	132997	133163	62	No
hhv7	66991.5	128589	128984	70	No
hsv1	62475	62465	62485	20	Yes
hsv1	131999	131990	132008	18	Yes
hsv1	146235	144115	144142	18	Yes
hsv2	62930	62919	62939	17	Yes
hsv2	132760	132691	132711	17	Yes
hsv2	148981	146600	146631	19	Yes
rcmv	77318	24072	24108	21	No
shv1	63878	63862	63892	24	Yes
shv1	114701	114686	114715	20	Yes
shv1	129901	129607	129636	20	Yes
vzv	110218.5	110195	110227	32	Yes
vzv	119678.5	119669	119701	32	Yes

predictive distance) shows the average of the distances between the center of each replication origin and a HSS that predicts it in map units. We also did some simple analysis of the location of the center of each replication origin with respect to the HSS closest to it. We count the number of times the center of replication origin falls within the left, right or center of the HSS. $\%L$, $\%R$ and $\%C$ gives these proportions.

Table 5.2 – *Prediction Performance: Summary. (C) indicates that the "Conservative" bound is used while (G) indicates that the "Generous" bound is used.*

Significance	Sensitivity	PPV	APD	%L	%R	%C
5% (C)	74%	22%	0.34 ± 0.57	16%	31%	53%
5% (G)	86%	17%	0.35 ± 0.53	24%	30%	46%
1% (C)	67%	25%	0.31 ± 0.52	14%	34%	52%
1% (G)	74%	18%	0.34 ± 0.57	16%	31%	53%

We see from the table that the prediction performance is rather good, with a sensitivity value of up to 86% when we use the "generous" bound at 5%. Another thing to note is that, on average, when we have a replication origin correctly predicted, the high-scoring segment closest to it is only 0.34 map units away from the true origin.

5.3 Discussion/Conclusion

We have also done some comparison studies between the methods described in this chapter and that of Chapter 3. Investigations revealed that amongst the methods mentioned in Chapter 3, BWS_1 performs the best when used to predict replication origins of viruses from the Herpesviridae family.

As mentioned in Chapter 3, we tried using a `AT`-content sliding window approach (say we call it `AT` sliding window) on the herpesviruses. 28 out of 43 replication origins have been predicted using this approach and some of these were not predicted by the BWS_1 or PLS method. Curious, we had in fact made some attempts to investigate the association of the `AT` sliding window approach with that of the BWS_1. Scatter plots and the Spearman's rank correlation coefficient were examined and we found that there is no association between the two schemes.

Further, we also tried used a "voting" scheme, to combine the two features in the hope of better prediction performance, in the following way:

1. For each sliding window constructed, we compute the BWS_1 score and AT content of it.

2. We rank the windows two times, the first ranking them according to its BWS_1 score and the other their AT content. Thus each window will have two ranks, one due to its BWS_1 score and the another its AT content.

3. For each window, we compute a "combined" rank, which is the average of the two ranks. Say if for the 100th window, its BWS_1 rank is 12 and AT sliding window rank is 30, then the combined rank will be 21.

4. We then sort all windows according to this combined rank and list out the top 10 windows, applying the filtering process mentioned in Chapter 3. That is, if the rank ith window is already chosen, then 8 windows to the left and right of it will not be considered for further ranking.

5. Using these top 10 windows, we access its prediction performance. Our results shows that this "voting" scheme does not add much value even though it managed to predict 32 out of 43 replication origins. There are only 3 replication origins that were predicted by this "voting " scheme but not previously predicted by either one of the two methods, the BWS_1 and AT sliding window approach. However, 9 origins that were predicted by the BWS_1 or AT sliding window approach were not picked up by this new approach.

This investigation justifies that the AT excursion approach we introduced in this chapter is a more refined method to measure AT content.

We now do a comparison of the AT excursion method and the BWS_1 scheme. The number of predictions suggested by both the AT excursion method and BWS_1 scheme are presented in Figure 5.2.

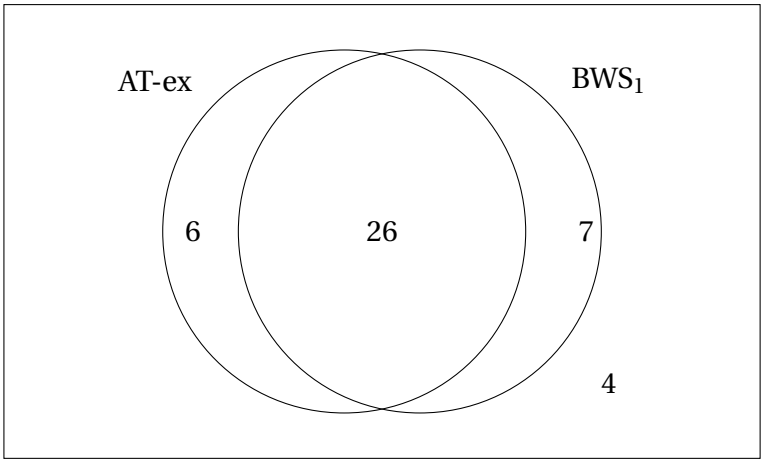


Figure 5.2 – *Predictions of AT excursion and BWS$_1$. In this figure, the set A consists of origin replications predicted by the AT excursion method and B consists of those predicted by the BWS$_1$ method. $A \cap B^C$ = {cehv7$_1$, cehv7$_2$, ehv4$_1$, hsv2$_1$, hsv2$_2$, hsv2$_3$}, $A^C \cap B$ = {cehv16$_2$, cehv16$_3$, ebv$_1$, ebv$_3$, hhv6, hhv6b, rcmv}, $(A \cup B)^C$ = {bohv4, ehv4$_2$, ehv4$_3$, hhv7}. The rest of the replication origins (26 of them) are predicted by both methods. (Note: For viruses with several known replication origins, such as hsv2, we denote the replication origins as hsv2$_1$,hsv2$_2$,hsv2$_3$, etc.)*

From the diagram, we see that the two methods complement one another. Majority of replication origins are predicted by both methods and most of the remaining ones are predicted by either methods. Of the 43 known replication origins, only 4 of them failed to be predicted by either one of the methods. This suggests that when searching for potential replication origin sites, AT concentration and palindromic concentration are two features that could be worth a look at.

We would like to point out several advantages of this approach.

1. It is "window size free". Unlike the approach mentioned in Chapter 3, the methods described in this chapter does not require the use of any sliding window to measure AT concentration.

2. The palindromes considered in Chapter 3 are of length at least 10 in most cases, and in some cases 12. These lengths were chosen after bench-marking with the well studied HCMV. The AT excursion method does not need to impose this kind of parameter.

3. Our method is more elaborate than merely measuring A/T percentage. Hopefully, this will more correctly capture the essence of A/T abundance.

   This is indeed the case for the herpesvirus data set. Out of 43 known replication origins, 23 are predicted by AT sliding window plot (AT-SWP) and AT Excursion method (AT-Ex); 9 are predicted by AT-Ex but not AT-SWP; whereas 5 by AT-SWP but not AT-EX.

4. Our method is statistical-based. Building on the work of Karlin and his collaborators (Karlin, 1994, 2005; Karlin and Altschul, 1990, 1993; Karlin et al., 1992), we have statistical tools to determine statistically high scoring segments.

5. It picks up some origins not detected by BWS1 as shown in Figure 2. This shows that the AT-excursion method complements the BWS1 method.

We have also tried locating high-scoring segments by running the excursions from the $3'$ end to $5'$ end of the genome. The results we obtained is not significantly different from the "vanilla" version (i.e., from 5' to 3').

5.3.1 Other Families of Viruses

Iridoviruses are a family of viruses that contain DNA as their genetic material and have an icosahedral (20-sided) capsid. Iridoviruses have been found in a wide variety of fish, including both freshwater and saltwater species. Some iridoviruses have been associated with serious diseases (e.g., viral erythrocytic necrosis of salmonids) while others have not and have only been found in apparently healthy animals (e.g., goldfish iridovirus). One iridovirus causes a disease called lymphocystis which causes unsightly skin lesions on infected fish. Iridoviruses associated with disease and mortality of tropical fish have been reported in Ramirez dwarf cichlids, angelfish, and, most recently, gouramis from the genus Trichogaster .

Poxviruses are the largest and most complex viruses. They are linear double stranded DNA viruses of 130 – 300 kilobase pair. The major human disease caused by a poxvirus (variola virus) is smallpox. Smallpox is caused by the Variola virus. Many animal

species have their own specific poxvirus infections, usually in the form of skin lesions. There are many poxviruses in nature, affecting species that gather in swarms and herds. Insects are also tortured with poxviruses. There are three groups of insect poxviruses: beetlepox, butterflypox (which includes mothpoxes), and flypox (including those of mosquitoes).

We will repeat our methods described in the previous sections on these two classes of viruses and identify high-scoring segments of these viral genomes. Viruses from the two families that are completely sequenced as of April 2006 are listed in Table 5.3. These two families of viruses are chosen because, like the herpesviruses, they are double stranded viruses with no RNA stage and their lengths are similar in magnitude to that of the herpesviruses. Amongst these two classes of viruses, we could only find one virus with 6 known replication origins, when we checked the literature. Our methods, however, could only correctly predict the location of one of the replication origins. We list out the high-scoring segments for each of the viruses in Table 5.5 on page 82.

Also, Table 5.6 on page 83 list the high-scoring windows as per the Base Weighted Scheme described in Chew et al. (2005) for the Irido and Pox viruses.

We do hope that the high-scoring segments and high-scoring windows will prove to be useful in identifying replication origins in these two families of viruses too.

Table 5.3 *– The list of Irido and Pox viruses to be analyzed.*

Accession	Virus	Length	Base Composition
	— Irido Viruses —		
NC_001824	Lymphocystis disease virus 1	102653	(0.35, 0.15, 0.14, 0.36)
NC_003038	Invertebrate iridescent virus 6	212482	(0.35, 0.15, 0.14, 0.36)
NC_003494	Infectious spleen and kidney necrosis virus	111362	(0.23, 0.28, 0.27, 0.23)
NC_005832	Ambystoma tigrinum virus	106332	(0.23, 0.27, 0.27, 0.23)
NC_005902	Lymphocystis disease virus - isolate China	186250	(0.36, 0.13, 0.14, 0.36)
NC_005946	Frog virus 3	105903	(0.23, 0.27, 0.28, 0.22)
NC_006549	Singapore grouper iridovirus	140131	(0.25, 0.24, 0.24, 0.26)
	— Pox Viruses —		
NC_001132	Myxoma virus	161773	(0.29, 0.22, 0.22, 0.28)
NC_001266	Rabbit fibroma virus	159857	(0.31, 0.20, 0.20, 0.30)
NC_001611	Variola virus	185578	(0.34, 0.16, 0.16, 0.33)
NC_001731	Molluscum contagiosum virus	190289	(0.18, 0.32, 0.32, 0.18)
NC_001993	Melanoplus sanguinipes entomopoxvirus	236120	(0.41, 0.09, 0.09, 0.41)
NC_002188	Fowlpox virus	288539	(0.35, 0.15, 0.15, 0.34)
NC_002520	Amsacta moorei entomopoxvirus	232392	(0.41, 0.09, 0.09, 0.42)
NC_002642	Yaba-like disease virus	144575	(0.37, 0.13, 0.14, 0.36)
NC_003027	Lumpy skin disease virus NI-2490	150773	(0.38, 0.13, 0.13, 0.36)
NC_003310	Monkeypox virus	196858	(0.34, 0.17, 0.17, 0.33)
NC_003389	Swinepox virus	146454	(0.37, 0.14, 0.14, 0.36)
NC_003391	Camelpox virus	205719	(0.34, 0.17, 0.17, 0.33)
NC_003663	Cowpox virus	224499	(0.33, 0.17, 0.17, 0.33)
NC_004002	Sheeppox virus 17077-99	149955	(0.38, 0.12, 0.13, 0.37)
NC_004003	Goatpox virus Pellor	149599	(0.38, 0.12, 0.13, 0.37)
NC_004105	Ectromelia virus	209771	(0.33, 0.17, 0.17, 0.33)
NC_005179	Yaba monkey tumor virus	134721	(0.35, 0.15, 0.15, 0.35)
NC_005309	Canarypox virus	359853	(0.35, 0.15, 0.15, 0.34)
NC_005336	Orf virus	139962	(0.18, 0.32, 0.32, 0.18)
NC_005337	Bovine papular stomatitis virus	134431	(0.18, 0.32, 0.32, 0.18)
NC_006966	Mule deer poxvirus	166259	(0.37, 0.13, 0.13, 0.37)
NC_006998	Vaccinia virus	194711	(0.33, 0.17, 0.17, 0.33)

Table 5.4 *– Herpesviruses : HSS at 5% level using the conservative bound.*

Accession	HSS Start	HSS Peak	HSS Value
alhv1	1204	1370	54
	32478	32850	48
	113630	113684	46
	85923	85992	45
	72999	73115	44
	125691	125726	31
athv3	8827	8892	40
bohv1	100410	100484	26
	109702	109730	25
	128487	128515	25
	16593	16626	21
	113720	113738	18
	124479	124497	18
	29	45	16
	58542	58569	15
bohv4	60687	60826	35
bohv5	68440	68507	49
	113549	113583	28
	129429	129463	28
	592	616	21
	86191	86215	21
	102074	102106	17
	92511	92535	15
	120935	120959	15
	59921	59938	14
	17408	17433	13
	41883	41899	13
calhv3	70131	70198	31
ccmv	50872	50973	50
	158344	158701	45
	95375	95603	39
	3519	3602	35
	24084	24156	33
	182982	183136	31
	14314	14370	23
	177170	177247	23
	189041	189075	22
	147310	147384	20
cehv1	116723	116836	53
	92092	92118	26
	61680	61700	20
	132785	132805	20
	149415	149435	20
	52055	52075	17
	42984	43006	16
	11389	11407	15
	24415	24441	14
cehv15	11965	12011	28
	114927	114988	19
cehv16	92913	92940	23
	62970	62991	21
	133468	133489	21
	149813	149834	21
	8303	8331	20
	118685	118713	20
	53056	53100	18
	25423	25473	16
	1717	1736	15
	114861	114890	15
	125280	125299	15
	30975	30991	14

Accession	HSS Start	HSS Peak	HSS Value
cehv2	7681	7738	33
	115791	115848	33
	61483	61503	20
	129527	129547	20
	144461	144481	20
	90857	90884	19
	51884	51910	14
	93873	93887	14
	112292	112320	14
cehv7	86167	86296	37
cehv8	149643	149720	33
	15671	15733	30
	29233	29278	29
	163766	163806	28
	177904	178092	28
	89538	89589	27
ebv	11854	11950	45
	77111	77150	24
	43158	43235	23
ehv1	20348	20431	47
	134195	134276	36
	65055	65126	35
	99301	99374	34
	11034	11141	32
	105796	105862	30
	73653	73746	27
	113818	113849	25
	149310	149341	25
	110314	110352	23
	128924	128992	23
ehv2	160281	160518	102
	86522	86622	76
	53843	54012	61
	140661	140826	57
	4580	4655	51
	171454	171529	51
	95342	95440	50
	10772	10820	48
	39893	39977	48
	177646	177694	48
	113310	113399	47
	134709	134772	45
	166114	166207	42
	45831	45965	41
	15443	15482	39
	19722	19845	39
	182317	182356	39
	153977	154145	36
	123321	123362	35
	147222	147341	35
	34816	34884	29
	76380	76454	29
	103167	103223	29
	64344	64402	25
	786	831	24

Table 5.4 *– Herpesviridae : HSS at 5% level using the conservative bound. (Cont'd)*

Accession	Start	HSS Peak	Value
ehv4	109852	110086	60
	19878	19943	50
	132383	132462	49
	105284	105365	48
	23895	24016	43
	3984	4110	42
	73340	73509	37
	98849	98930	33
	46612	46674	32
	10630	10697	31
	58833	58906	31
	82616	82701	31
	127230	127351	31
	112929	112967	29
	145082	145120	29
gahv1	24852	24890	30
gahv2	106724	106811	35
gahv3	11168	11198	27
	122384	122414	27
	134414	134461	26
	162999	163046	26
	58953	58999	25
hcmv	3402	3542	41
	186855	186995	41
	16757	16915	35
	96685	96824	34
	11713	11808	32
	198116	198171	31
	173560	173599	30
	210724	210781	30
	26361	26475	27
	108222	108303	24
	159296	159380	24
	71011	71055	23
	226192	226230	23
hcmv-m	3798	3939	42
	181238	181334	33
	97069	97206	32
	173950	173994	32
	216020	216077	30
	203400	203456	29
	17082	17297	26
	12060	12145	25
	157590	157726	25
hhv6	130410	130501	59
	3605	3712	51
	154838	154945	51
	137079	137210	43
hhv6b	132997	133163	62
	139482	139569	51
	3911	3988	37
	157232	157309	37
hhv7	134169	134376	117
	128589	128984	70
hhv8	136287	136704	93
	982	1125	44
	58833	58906	28
	23547	23598	27
	30712	30775	27
	119416	119467	27
	106412	106452	25

Accession	Start	HSS Peak	Value
hsv1	62465	62485	20
	35000	35034	19
	115242	115303	19
	131990	132008	18
	144115	144142	18
	11705	11734	17
	52753	52818	17
	96047	96069	16
	136146	136162	16
hsv2	5584	5628	35
	121621	121665	35
	52978	53003	19
	91716	91747	19
	146600	146631	19
	95238	95256	18
	48761	48778	17
	62919	62939	17
	132691	132711	17
	81195	81220	16
	99337	99370	15
ichv1	6068	6290	81
	121738	121960	81
	104134	104399	70
	17065	17333	58
	132735	133003	58
	451	726	50
	116121	116396	50
	60752	60845	30
	42919	43007	28
	20109	20187	24
	10016	10063	23
	125686	125733	23
mcmv	155163	156341	125
	161228	161391	40
	115543	115640	37
	102865	102960	35
	79497	79573	34
	15628	15724	33
	144170	144290	33
	73525	73579	27
	39209	39248	24
	92997	93036	24
	219239	219282	22
mehv1		NIL	
mfrv	128046	128640	114
	23139	23374	109
	2488	3068	106
	32573	33752	84
	64296	64454	62
	111496	111624	44
	72739	72809	43
	53766	53825	32
	69912	70061	32
	114828	114860	32
mmrv	2388	2967	111
	23902	24187	108
	33761	35136	103
	130346	131085	97
	65611	65853	56
	74140	74204	37
	71311	71462	31
	117507	117551	29
	112930	113033	28

Table 5.4 *– Herpesviridae : HSS at 5% level using the conservative bound. (Cont'd)*

Accession	HSS Start	HSS Peak	HSS Value
muhv4	6000	6037	29
ohv2	115365	115545	72
	126823	127116	68
	118943	118988	42
	72630	72699	36
	1269	1370	29
	27589	27633	29
	76335	76370	26
	79158	79265	26
oshv1	73292	73460	64
	35416	35493	61
	146021	146164	55
	190174	190312	54
	195928	196026	54
	201648	201786	54
	23065	23135	50
	161395	161505	50
	2682	2735	49
	180276	180329	49
	108068	108173	45
	171433	171549	44
	67872	67975	43
	114689	114763	42
pshv1	18751	18791	31
	121452	121486	31
	160685	160719	31
	130332	130365	27
	151806	151839	27
	23896	23942	22
	134013	134049	21
	78233	78256	20

Accession	HSS Start	HSS Peak	HSS Value
rcmv	150923	151612	92
	207600	207980	80
	143617	144150	74
	178241	178326	37
	214638	214702	37
	219069	219153	33
	201767	201885	28
	161797	161929	27
	171828	171870	27
	24072	24108	21
sahv2	28533	28613	45
shv1	63862	63892	24
	96251	96275	21
	114686	114715	20
	129607	129636	20
	50382	50407	19
	75955	75984	17
	16151	16172	15
	33045	33063	15
	109083	109098	15
	135503	135518	15
	8432	8455	14
thv	168842	168927	25
	24153	24200	23
	28257	28286	17
vzv	2574	2785	39
	110195	110227	32
	119669	119701	32

Table 5.5 *– Irido and Pox viruses: HSS at 5% level using the conservative bound.*

Accession	HSS Start	HSS Peak	HSS Value
— Irido Viruses —			
NC_001824		NIL	
NC_003038	79081	79249	66
	183476	183613	65
	105552	105759	63
	153314	153430	62
	174299	174361	62
	58432	58537	57
NC_003494	14511	14546	26
	8403	8449	25
NC_005832	57645	58021	103
	28404	28881	102
	34433	34957	86
	93135	93532	82
	60770	61386	76
	8445	8824	67
	23967	24236	62
	49790	50135	60
	14551	14919	59
	73173	74011	58
	81666	81980	56
	101304	101712	54
	98977	99415	48
	55341	55467	45
	21468	21707	41
	3416	3623	36
	88908	89282	35
	104642	104781	31
	37618	37803	26
	71482	71548	24
	84341	84389	24
NC_005902	25319	25426	83
NC_005946	91036	91300	60
	65486	65899	56
	84444	84664	43
	11456	11662	41
	100217	100399	41
	87467	87807	37
	70406	70535	33
	60587	60658	32
	42478	42647	31
	75837	76051	31
	1169	1483	29
	30559	30801	29
	16604	16868	27
	50384	50630	27
	80707	80812	27
	104871	105033	27
	6300	6550	25
	37735	37917	23
	46489	46554	23
NC_006549	81306	81382	32
	110066	110102	32
	37472	37506	30
	135884	135929	29
— Pox Viruses —			
NC_001132	70592	70699	47
	115736	115827	47
NC_001266	69787	69895	52
	114903	114989	50
	143583	143744	41
NC_001611		NIL	
NC_001731	142080	142122	36
	3323	3353	24
	186936	186966	24
	158487	158508	18
NC_001993	99125	99560	147
	32990	33296	99
NC_002188	73269	73413	74
	15448	15664	66
	232716	232804	58
NC_002520	200361	200872	214
	207549	208262	155
	165286	165627	152
	16532	17256	130
	139207	139516	111
	133061	133305	109
	89235	89522	107
	110336	110659	107
	58328	58497	106
	99032	99612	103
NC_002642	23141	23269	68
	69383	69515	66
NC_003027		NIL	
NC_003310	178751	178921	80
	148715	148766	51
NC_003389	67171	67316	73
NC_003391	144645	144737	92
	191172	191354	57
	24592	24653	51
NC_003663	3	114	76
	159398	159460	62
NC_004002	17624	18041	87
	117891	118153	76
NC_004003		NIL	
NC_004105	177738	177834	91
	16773	17140	57
NC_005179		NIL	
NC_005309	151555	151798	78
NC_005336	4745	4880	51
	135222	135298	43
	129134	129329	42
	113376	113456	32
	119112	119205	30
	108363	108520	25
	318	350	23
	60440	60478	23
	139546	139578	23
	46479	46515	21
	54314	54343	20
	124682	124703	18
NC_005337	126750	126962	35
	4662	4781	26
	13133	13183	23
	54399	54440	23
	108478	108513	23
	46663	46700	22
	60477	60503	20
	9108	9145	19
	113089	113137	18
	62922	62972	17
	69451	69477	17
	119341	119370	17
NC_006966	120853	121259	88
NC_006998	1	114	74

Table 5.6 *– Irido and Pox viruses: Top 10 high-scoring windows under BWS_1.*

Virus	Accession	Rank of Window 1	2	3	4	5	6	7	8	9	10
— Irido Viruses —											
Lymphocystis disease virus 1	NC_001824	35251	11251	39501	60751	100251	71001	47501	50751	2001	9001
Invertebrate iridescent virus 6	NC_003038	90501	86501	110001	143501	34501	155501	170001	49501	150501	208501
Infectious spleen and kidney necrosis virus	NC_003494	103001	16751	47251	58751	96751	31001	99751	94251	78001	39751
Ambystoma tigrinum virus	NC_005832	51251	28251	57501	60751	73501	34251	14501	81501	45001	71251
Lymphocystis disease virus - isolate China	NC_005902	24751	19801	130501	49501	88201	451	98101	174151	162451	5851
Frog virus 3	NC_005946	65501	50251	87251	60501	37501	70501	6251	73001	103001	40251
Singapore grouper iridovirus	NC_006549	50401	41301	24151	4901	139301	36751	129151	29051	12601	80851
— Pox Viruses —											
Myxoma virus	NC_001132	77201	9201	151601	103201	16401	56401	145601	60001	118401	131201
Rabbit fibroma virus	NC_001266	69301	33251	143851	106051	26251	109551	57751	89601	82951	18551
Variola virus	NC_001611	26101	70201	77851	159751	13951	61651	169651	88651	139501	20251
Molluscum contagiosum virus	NC_001731	143101	35551	60301	54901	135901	67501	22951	105301	109801	7651
Melanoplus sanguinipes entomopoxvirus	NC_001993	223301	191401	198551	6601	185351	65451	118801	33001	18151	73701
Fowlpox virus	NC_002188	193901	155401	242901	261101	268101	232401	72101	49701	30801	109201
Amsacta moorei entomopoxvirus	NC_002520	7151	32451	223851	132001	93501	140801	184801	37951	202401	104501
Yaba-like disease virus	NC_002642	1401	142451	35351	134051	60901	68951	138601	90301	82601	15401
Lumpy skin disease virus NI-2490	NC_003027	8051	40251	131951	97651	122151	44451	80851	29751	138951	119351
Monkeypox virus	NC_003110	8101	141751	194401	166951	20701	97651	171001	901	87301	59401
Swinepox virus	NC_003389	85401	105701	128801	77351	93101	135801	51451	66151	48301	10151
Camelpox virus	NC_003391	144501	181501	186501	100501	59501	82001	172001	32001	90001	17001
Cowpox virus	NC_003663	551	222751	158401	187551	203501	49501	27501	114401	139701	72601
Sheeppox virus 17077-99	NC_004002	122151	119001	39901	7701	45501	17501	97651	80151	29051	138601
Goatpox virus Pellor	NC_004003	7351	121801	65451	17851	39551	44101	137901	29051	96951	2801
Ectromelia virus	NC_004105	501	208501	177501	194001	184501	159001	42001	18501	28501	106501
Yaba monkey tumor virus	NC_005179	63001	73801	105001	10501	32401	85801	67201	39301	129901	22801
Canarypox virus	NC_005309	257551	195501	203151	269451	1	54401	71401	309401	281351	160651
Orf virus	NC_005336	99301	39901	3001	21001	126601	88801	85201	79201	136501	27901
Bovine papular stomatitis virus	NC_005337	45901	68101	6901	117301	112501	33901	100801	60601	2101	133201
Mule deer poxvirus	NC_006966	39201	26801	140401	159201	14801	34401	92001	126401	20801	149201
Vaccinia virus	NC_006998	2701	191251	141751	169201	21601	97651	13951	79201	177751	122851

CHAPTER 6

PALINDROME EXCURSIONS AND SUMMARY

Encouraged by our success with the AT excursion, we want to try to extend the approach to work with palindromes. In this chapter we will give some preliminary results of our investigation. We will also conclude this thesis by giving a summary of our efforts in the prediction of replication origins and suggest some possible extensions of the problems we have considered in this thesis.

6.1 Palindrome Excursions

We will describe in this section our attempts to adapt once again Karlin's score based approach to the setting of palindromes. Recall that the idea is to assign scores to different bases in the genomic sequences and look for regions with statistically high scores. So to make the approach work with the palindromes, we score a base according to if it is part of a palindrome, that is, bases that form part of a palindrome will be given a score say s_p and those that do not form part of a palindrome will be given a score say s_q.

Further, we need to compute the probability that a base is part of a palindrome, so let us define

$$\psi := p(k\text{-th base pair is part of a palindrome of length at least } 2L).$$

Note that we would once again consider palindromes above a certain length, consistent with the approach of this thesis.

Let A_j denote the event that there is a palindrome of length at least $2L$ starting from base j. Then

$$\begin{aligned}
\psi &= P(\xi_j\xi_{j+1}\cdots\xi_{j+2L-1} \text{ forms a palindrome, for some } k-2L+1 \le j \le k.) \\
&= P(\cup_{j=1}^{2L} A_j) \\
&= P(A_1) + P(A_1^C A_2) + \cdots + P(A_1^C \cdots A_{2L-1}^C A_{2L}) \\
&= \sum_{i=1}^{2L} P(A_i) - \sum_{i=2}^{2L} P\left(\left[\cup_{j=1}^{i-1} A_j\right] A_i\right) \\
&\le 2LP(A_1) - \sum_{i=2}^{2L} P(A_{i-1}A_i) \\
&= 2LP(A_1) - (2L-1)P(A_1A_2) \\
&:= \psi_U.
\end{aligned}$$

So we have an upper bound (which we define as ψ_U) for the probability ψ. Note that the term $P(A_1A_2)$ is actually the term $\gamma(1)$ as defined in Lemma 2.1 on page 12.

Following a hint from Galambos and Simonelli (1996)(Inequality I.7, p.22), we will also have an lower bound for ψ, given by

$$\psi_L := \max_{2\le k\le 2L}\left\{\frac{2S_1}{k} - \frac{2S_2}{k(k-1)}\right\},$$

where

$$S_1 := \sum_{i=1}^{2L} P(A_i)$$

$$S_2 := \sum_{1 \le i < j \le 2L} P(A_i A_j) = \sum_{i=1}^{2L-1} \sum_{j=i+1}^{2L} P(A_i A_j) = \sum_{i=1}^{2L-1} \sum_{r=1}^{2L-i} P(A_1 A_{r+1})$$
$$= \sum_{r=1}^{2L-1} \sum_{i=1}^{2L-r} P(A_1 A_{r+1}) = \sum_{r=1}^{2L-1} (2L-r) P(A_1 A_{r+1}).$$

Similarly, the term $P(A_1 A_{r+1})$ is $\gamma(r)$ as defined in Lemma 2.1.

The values of ψ_L and ψ_U for the herpesviruses are listed in Table 6.1 on the following page. From the last column of the table, we see that the upper and lower bounds of ψ are rather close, which means that our bounds are tight.

Even though we do not have the exact form of the probability expression ψ, it does seem reasonable to use an approximation of it and apply the excursion approach to it. However, Karlin's results require an i.i.d. or Markov chain assumption (See, for example Dembo and Karlin, 1991a,b), whereas for our case here, there is some local dependence in the way the bases are related. For if a base is part of a palindrome, then bases near it is likely to be part of a palindrome too. Hence we cannot directly apply Karlin's results to this problem.

Nonetheless, we decide to try an non-parametric approach like we did for the scoring schemes in Chapter 3. We will run the excursions on the palindromes over the family of herpesviruses and list out the top high scoring segments and use them as our prediction regions. The procedure will be similar to what we have described in the previous chapter on `AT` excursion. We will not be able to apply Karlin's results to come up with any statistically high scoring windows though.

However, we will still use ψ_U as an conservative approximation for ψ. The rational is that we want to control the "drift" of the excursion process, which is dependent on the expected value per base $\mu = s_p\psi + s_q(1-\psi)$. Note that as in the previous chapter, we will set μ to some negative value, and let s_p be 1. The value s_q of will then be determined according to the definition of μ.

Table 6.1 – *Herpesviruses: ψ values.*

Virus	ψ_U	ψ_U	ψ_L/ψ_U
alhv1	0.00993734	0.00997550	0.99617406
athv3	0.01358172	0.01369403	0.99179838
bohv1	0.00866357	0.00874687	0.99047745
bohv4	0.01117899	0.01123763	0.99478183
bohv5	0.01081602	0.01095066	0.98770574
calhv3	0.00970650	0.00974086	0.99647229
ccmv	0.01261702	0.01270470	0.99309886
cehv1	0.01042433	0.01054889	0.98819164
cehv15	0.01273456	0.01282521	0.99293252
cehv16	0.01216570	0.01233720	0.98609926
cehv2	0.01202644	0.01219396	0.98626209
cehv7	0.01156561	0.01163149	0.99433630
cehv8	0.00972483	0.00975931	0.99646721
ebv	0.01155367	0.01161911	0.99436842
ehv1	0.01059220	0.01064006	0.99550220
ehv2	0.01082274	0.01087468	0.99522403
ehv4	0.00972554	0.00975992	0.99647710
gahv1	0.00978630	0.00982155	0.99641095
gahv2	0.01045407	0.01049959	0.99566507
gahv3	0.00997182	0.01000979	0.99620665
hcmv	0.01074478	0.01079521	0.99532850
hcmv-m	0.01083288	0.01088484	0.99522684
hhv6	0.01085252	0.01090489	0.99519764
hhv6b	0.01075493	0.01080555	0.99531574
hhv7	0.01377465	0.01389278	0.99149707
hhv8	0.00992798	0.00996557	0.99622790
hsv1	0.00616250	0.00619926	0.99407027
hsv2	0.00728651	0.00734222	0.99241291
ichv1	0.01047484	0.01052080	0.99563175
mcmv	0.01126013	0.01131982	0.99472622
mehv1	0.00983418	0.00987013	0.99635727
mfrv	0.00976015	0.00979539	0.99640227
mmrv	0.00981245	0.00984835	0.99635514
muhv4	0.00985069	0.00988705	0.99632187
ohv2	0.00969729	0.00973306	0.99632489
oshv1	0.01235602	0.01243828	0.99338643
pshv1	0.01221558	0.01229454	0.99357746
rcmv	0.01224677	0.01232638	0.99354180
sahv2	0.01513100	0.01528533	0.98990333
shv1	0.00960496	0.00970946	0.98923778
thv	0.00543967	0.00546611	0.99516254
vzv	0.01002025	0.01005902	0.99614604

Table 6.2 – *Prediction Performance of Palindrome Excursion.*

	1	2	3	4	5	6	7	8	9	10
Sensitivity	16	28	37	44	51	51	53	56	60	63
PPV	35	30	27	24	22	18	16	15	14	14

We tried setting $\mu = -5, -10, -15, -20$ and found that for our purpose, $\mu = -10$ works the best. Table 6.2 shows the performance of this "Palindrome Excursion" scheme when a certain number of top scoring windows are chosen. Comparing with the non-parametric approach we adapted for Chapter 3, we see that the performance of this approach is just slightly inferior to the PLS scheme.

6.2 Summary

In this section we do a summary of the various approaches we have looked at in this thesis in the problem of predicting replication origins in the herpesviruses. Table 6.3 on the following page lists all the known replication origins of the herpesviruses, together with the prediction outcomes of the various schemes of prediction, namely the PLS, BWS1, PLS with compound Poisson approximation (PLS-CPA) at 5% under the M0 model, the AT sliding window with Binomial approximation (AT-swp-Binomial) at 5%, the AT excursion (AT-ex) at 5% and the palindrome excursion (Pal-ex). Entries under the columns "PLS", "BWS1" and "Pal-ex" indicate the rank of the window/segment that predicts the replication origin listed on that row. For the other columns, a " Y" indicates that the high-scoring window/segment is successful in predicting that particular replication origin, and a "N" indicates otherwise. A "-" indicates that there are no statistically significant high scoring windows/segments.

We note that most of the replication origins are predicted by either one of the prediction schemes except a few, namely one of the replication origins of ehv4, and that of hhv7. We suspect that other features such as approximate palindromes (imperfect palindromes with one or more mismatch), inverted repeats might be useful in the prediction of these replication origins. Indeed, Qin (2005) reported in her thesis her attempts to use approximate palindromes in the prediction of replication origins in

Table 6.3 – *Summary of All Prediction Schemes.*

Virus	ORI Center	Non-Parametric PLS	Non-Parametric BWS1	PLS-CPA 5% M0	AT-swp-Binomial 5%	AT-ex 5%	Pal-ex
bohv1	111190	1	1	Y	Y	Y	3
bohv1	127028	2	2	Y	Y	Y	4
bohv4	97996.5	0	0	-	Y	N	0
bohv5	113312	0	2	N	Y	Y	9
bohv5	129701	0	3	N	Y	Y	10
cehv1	61690.5	3	3	Y	Y	Y	1
cehv1	61893.5	3	3	Y	Y	Y	1
cehv1	132795.5	1	1	Y	Y	Y	2
cehv1	132998.5	1	1	Y	Y	Y	2
cehv1	149425.5	2	2	Y	Y	Y	3
cehv1	149628.5	2	2	Y	Y	Y	3
cehv2	61493.5	3	3	Y	Y	Y	3
cehv2	129537.5	1	1	Y	N	Y	1
cehv2	144471.5	2	2	Y	Y	Y	2
cehv7	109636.5	5	7	-	N	N	0
cehv7	118622.5	6	8	-	N	N	0
cehv16	62981	8	7	N	Y	Y	1
cehv16	133479	0	0	Y	N	Y	20
cehv16	149824	0	0	N	Y	Y	21
ebv	8313.5	1	1	Y	Y	N	5
ebv	40797	3	2	Y	Y	Y	1
ebv	143825.5	2	3	Y	Y	N	2
ehv1	126262.5	4	4	N	Y	Y	5
ehv4	73909.5	0	0	-	Y	Y	0
ehv4	119471.5	0	0	-	N	N	0
ehv4	138577.5	0	0	-	Y	N	0
gahv1	24871.5	8	7	-	N	Y	0
hcmv	93923.5	1	1	Y	Y	Y	4
hhv6	67805	4	4	-	Y	N	5
hhv6b	69160.5	2	4	N	Y	N	4
hhv7	66991.5	0	0	-	N	N	0
hsv1	62475	1	1	Y	Y	Y	1
hsv1	131999	2	2	Y	N	Y	15
hsv1	146235	3	3	Y	Y	Y	16
hsv2	62930	0	0	N	Y	Y	11
hsv2	132760	0	0	N	N	Y	0
hsv2	148981	0	0	N	Y	Y	0
rcmv	77318	1	1	Y	N	N	8
shv1	63878	0	9	N	Y	Y	16
shv1	114701	0	5	N	Y	Y	7
shv1	129901	0	6	N	Y	Y	9
vzv	110218.5	2	2	Y	N	Y	1
vzv	119678.5	1	1	Y	N	Y	2

the herpesviruses. She extended the palindrome length scheme to work with the approximate palindromes and reported that the prediction performance of her scheme shows an improvement over that of the PLS in terms of sensitivity and positive predictive power.

6.3 Future Work

In this thesis, we had devoted a great deal of effort in the problem of predicting replication origins in the herpesviruses (primarily).

There are still a few problems that we can work on. One of it is the problem of approximating the window score under the Base-pair Weighted Scheme by possibly a compound Poisson distribution.

The excursion approach of Karlin could also be adapted to work with palindromes. Because of the local dependence structure embedded in the problem, we suspect the Chen-Stein method of Poisson approximation might be relevant to this problem.

Finally, we note that these endeavors to accurately predict replication origins had motivated several interesting and challenging mathematical problems and will continue to do so.

BIBLIOGRAPHY

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. Available from: http://dx.doi.org/10.1038/89991.

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). *Ocul Immunol Inflamm*, 9(2):125–130.

Bloom, B. R. (2003). Lessons from SARS. *Science*, 300(5620):701. Available from: http://dx.doi.org/10.1126/science.300.5620.701.

Bramhill, D. and Kornberg, A. (1988). A model for initiation at origins of DNA replication. *Cell*, 54(7):915–918.

Breier, A. M., Chatterji, S., and Cozzarelli, N. R. (2004). Prediction of Saccharomyces cerevisiae replication origins. *Genome Biol*, 5(4):R22. Available from: http://dx.doi.org/10.1186/gb-2004-5-4-r22.

Cain, D., Erlwein, O., Grigg, A., Russell, R. A., and McClure, M. O. (2001). Palindromic sequence plays a critical role in human foamy virus dimerization. *J Virol*, 75(8):3731–3739. Available from: http://dx.doi.org/10.1128/JVI.75.8.3731-3739.2001.

Chew, D. S. H., Choi, K. P., Heidner, H., and Leung, M.-Y. (2004). Palindromes in SARS and Other Coronaviruses. *INFORMS Journal on Computing*, 16(4):331–340.

Chew, D. S. H., Choi, K. P., and Leung, M.-Y. (2005). Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses. *Nucleic Acids Res*, 33(15):e134. Available from: http://dx.doi.org/10.1093/nar/gni135.

Creighton, T. E. (1993). *Proteins*. WH Freeman and Company, New York, New York.

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

Dembo, A. and Karlin, S. (1991a). Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables. *Ann. Probab.*, 19(4):1756–1767.

Dembo, A. and Karlin, S. (1991b). Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Probab.*, 19(4):1737–1755.

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

Deng, H., Chu, J. T., Park, N.-H., and Sun, R. (2004). Identification of cis sequences required for lytic DNA replication and packaging of murine gammaherpesvirus 68. *J Virol*, 78(17):9123–9131. Available from: http://dx.doi.org/10.1128/JVI.78.17.9123-9131.2004.

Dirac, A. M. G., Huthoff, H., Kjems, J., and Berkhout, B. (2002). Requirements for RNA heterodimerization of the human immunodeficiency virus type 1 (HIV-1) and HIV-2 genomes. *J Gen Virol*, 83(Pt 10):2533–2542.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (2000). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge Univ. Press, Cambridge, UK.

Galambos, J. and Simonelli, I. (1996). *Bonferroni-type inequalities with applications*. Probability and its Applications (New York). Springer-Verlag, New York.

Giedroc, D. P., Theimer, C. A., and Nixon, P. L. (2000). Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. *J Mol Biol*, 298(2):167–185. Available from: http://dx.doi.org/10.1006/jmbi.2000.3668.

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

Hartline, C. B., Harden, E. A., Williams-Aziz, S. L., Kushner, N. L., Brideau, R. J., and Kern, E. R. (2005). Inhibition of herpesvirus replication by a series of 4-oxo-dihydroquinolines with viral polymerase activity. *Antiviral Res*, 65(2):97–105. Available from: http://dx.doi.org/10.1016/j.antiviral.2004.10.003.

Hill, M. K., Shehu-Xhilaga, M., Campbell, S. M., Poumbourios, P., Crowe, S. M., and Mak, J. (2003). The dimer initiation sequence stem-loop of human immunodeficiency virus type 1 is dispensable for viral replication in peripheral blood mononuclear cells. *J Virol*, 77(15):8329–8335.

Karlin, S. (1994). Statistical studies of biomolecular sequences: score-based methods. *Philos Trans R Soc Lond B Biol Sci*, 344(1310):391–402.

Karlin, S. (2005). Statistical signals in bioinformatics. *Proc Natl Acad Sci U S A*, 102(38):13355–13362. Available from: http://dx.doi.org/10.1073/pnas.0501804102.

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, 87(6):2264–2268.

Karlin, S. and Altschul, S. F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc Natl Acad Sci U S A*, 90(12):5873–5877.

Karlin, S., Burge, C., and Campbell, A. M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucleic Acids Res*, 20(6):1363–1370.

Karlin, S., Dembo, A., and Kawabata, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Statist.*, 18(2):571–581.

Kornberg, A. and Baker, T. A. (1992). *DNA Replication*. WH Freeman and Company, New York, New York, 2nd edition.

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., Choi, K. P., Xia, A., and Chen, L. H. Y. (2005). Nonrandom clusters of palindromes in herpesvirus genomes. *J Comput Biol*, 12(3):331–354. Available from: http://dx.doi.org/10.1089/cmb.2005.12.331.

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(4):445–471.

Leung, M.-Y. and Yamashita, T. E. (1999). Applications of the scan statistic in DNA sequence analysis. In *Scan statistics and applications*, Stat. Ind. Technol., pages 269–286. Birkhäuser Boston, Boston, MA.

Lin, C. L., Li, H., Wang, Y., Zhu, F. X., Kudchodkar, S., and Yuan, Y. (2003). Kaposi's sarcoma-associated herpesvirus lytic origin (ori-Lyt)-dependent DNA replication: identification of the ori-Lyt and association of K8 bZip protein with the origin. *J Virol*, 77(10):5578–5588.

Lobry, J. R. (1996). Asymmetric substitution patterns in the two dna strands of bacteria. *Mol Biol Evol*, 13(5):660–665.

Mackiewicz, P., Zakrzewska-Czerwinska, J., Zawilak, A., Dudek, M. R., and Cebrat, S. (2004). Where does bacterial replication start? Rules for predicting the oriC

region. *Nucleic Acids Res*, 32(13):3781–3791. Available from: http://dx.doi.org/10.1093/nar/gkh699.

Marra, M. A., Jones, S. J. M., Astell, C. R., Holt, R. A., Brooks-Wilson, A., Butterfield, Y. S. N., Khattra, J., Asano, J. K., Barber, S. A., Chan, S. Y., Cloutier, A., Coughlin, S. M., Freeman, D., Girn, N., Griffith, O. L., Leach, S. R., Mayo, M., McDonald, H., Montgomery, S. B., Pandoh, P. K., Petrescu, A. S., Robertson, A. G., Schein, J. E., Siddiqui, A., Smailus, D. E., Stott, J. M., Yang, G. S., Plummer, F., Andonov, A., Artsob, H., Bastien, N., Bernard, K., Booth, T. F., Bowness, D., Czub, M., Drebot, M., Fernando, L., Flick, R., Garbutt, M., Gray, M., Grolla, A., Jones, S., Feldmann, H., Meyers, A., Kabani, A., Li, Y., Normand, S., Stroher, U., Tipples, G. A., Tyler, S., Vogrig, R., Ward, D., Watson, B., Brunham, R. C., Krajden, M., Petric, M., Skowronski, D. M., Upton, C., and Roper, R. L. (2003). The genome sequence of the SARS-associated coronavirus. *Science*, 300(5624):1399–1404. Available from: http://dx.doi.org/10.1126/science.1085953.

Masse, M. J., Karlin, S., Schachtel, G. A., and Mocarski, E. S. (1992). Human cytomegalovirus origin of DNA replication (oriLyt) resides within a highly complex repetitive region. *Proc Natl Acad Sci U S A*, 89(12):5246–5250.

Merkl, R. and Fritz, H. J. (1996). Statistical evidence for a biochemical pathway of natural, sequence-targeted G/C to C/G transversion mutagenesis in Haemophilus influenzae Rd. *Nucleic Acids Res*, 24(21):4146–4151.

Newlon, C. S. and Theis, J. F. (2002). DNA replication joins the revolution: whole-genome views of DNA replication in budding yeast. *Bioessays*, 24(4):300–304. Available from: http://dx.doi.org/10.1002/bies.10075.

Qin, L., Xiong, B., Luo, C., Guo, Z.-M., Hao, P., Su, J., Nan, P., Feng, Y., Shi, Y.-X., Yu, X.-J., Luo, X.-M., Chen, K.-X., Shen, X., Shen, J.-H., Zou, J.-P., Zhao, G.-P., Shi, T.-L., He, W.-Z., Zhong, Y., Jiang, H.-L., and Li, Y.-X. (2003). Identification of probable genomic packaging signal sequence from SARS-CoV genome by bioinformatics analysis. *Acta Pharmacol Sin*, 24(6):489–496.

Qin, X. (2005). Palindrome distributions and their applications. Master's thesis, National University Of Singapore.

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(8):1822–1832.

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

Robin, S. and Daudin, J. J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.*, 36(1):179–193.

Rocha, E. P., Danchin, A., and Viari, A. (2001). Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res*, 11(6):946–958. Available from: http://dx.doi.org/10.1101/gr.153101.

Rocha, E. P., Viari, A., and Danchin, A. (1998). Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. *Nucleic Acids Res*, 26(12):2971–2980.

Rota, P. A., Oberste, M. S., Monroe, S. S., Nix, W. A., Campagnoli, R., Icenogle, J. P., Peñaranda, S., Bankamp, B., Maher, K., Chen, M.-H., Tong, S., Tamin, A., Lowe, L., Frace, M., DeRisi, J. L., Chen, Q., Wang, D., Erdman, D. D., Peret, T. C. T., Burns, C., Ksiazek, T. G., Rollin, P. E., Sanchez, A., Liffick, S., Holloway, B., Limor, J., McCaustland, K., Olsen-Rasmussen, M., Fouchier, R., Günther, S., Osterhaus, A. D. M. E., Drosten, C., Pallansch, M. A., Anderson, L. J., and Bellini, W. J. (2003). Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science*, 300(5624):1394–1399. Available from: http://dx.doi.org/10.1126/science.1085952.

Rowe, C. L., Fleming, J. O., Nathan, M. J., Sgro, J. Y., Palmenberg, A. C., and Baker, S. C. (1997). Generation of coronavirus spike deletion variants by high-frequency re-

combination at regions of predicted RNA secondary structure. *J Virol*, 71(8):6183–6190.

Ruan, Y. J., Wei, C. L., Ee, A. L., Vega, V. B., Thoreau, H., Su, S. T. Y., Chia, J.-M., Ng, P., Chiu, K. P., Lim, L., Zhang, T., Peng, C. K., Lin, E. O. L., Lee, N. M., Yee, S. L., Ng, L. F. P., Chee, R. E., Stanton, L. W., Long, P. M., and Liu, E. T. (2003). Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet*, 361(9371):1779–1785.

Salzberg, S. L., Salzberg, A. J., Kerlavage, A. R., and Tomb, J. F. (1998). Skewed oligomers and origins of replication. *Gene*, 217(1-2):57–67.

Schbath, S., Prum, B., and de Turckheim, E. (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J Comput Biol*, 2(3):417–437.

Segurado, M., de Luis, A., and Antequera, F. (2003). Genome-wide distribution of DNA replication origins at A+T-rich islands in Schizosaccharomyces pombe. *EMBO Rep*, 4(11):1048–1053. Available from: http://dx.doi.org/10.1038/sj.embor.embor7400008.

Villarreal, E. C. (2003). Current and potential therapies for the treatment of herpesvirus infections. *Prog Drug Res*, 60:263–307.

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

Waterman, M. S. (1995). *Introduction to Computational Biology*. Chapman and Hall, New York.

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 oriL, a herpes simplex virus type 1 origin of DNA synthesis. *Mol Cell Biol*, 5(5):930–942.

Worobey, M. and Holmes, E. C. (1999). Evolutionary aspects of recombination in RNA viruses. *J Gen Virol*, 80 (Pt 10):2535–2543.

Yu, X.-J., Luo, C., Lin, J.-C., Hao, P., He, Y.-Y., Guo, Z.-M., Qin, L., Su, J., Liu, B.-S., Huang, Y., Nan, P., Li, C.-S., Xiong, B., Luo, X.-M., Zhao, G.-P., Pei, G., Chen, K.-X., Shen, X., Shen, J.-H., Zou, J.-P., He, W.-Z., Shi, T.-L., Zhong, Y., Jiang, H.-L., and Li, Y.-X. (2003). Putative hAPN receptor binding sites in SARS_CoV spike protein. *Acta Pharmacol Sin*, 24(6):481–488.

Zhang, R. and Zhang, C.-T. (2005). Identification of replication origins in archaeal genomes based on the Z-curve method. *Archaea*, 1(5):335–346.

Zhu, Y., Huang, L., and Anders, D. G. (1998). Human cytomegalovirus oriLyt sequence requirements. *J Virol*, 72(6):4989–4996.