PALINDROME DISTRIBUTIONS AND THEIR APPLICATIONS

QIN XUAN

NATIONAL UNIVERSITY OF SINGAPORE
2005

PALINDROME DISTRIBUTIONS AND THEIR APPLICATIONS

QIN XUAN
(*BSc, Beijing University of Technology, China*)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
2005

Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my supervisor Associate Professor Choi Kwok Pui. During my research, he has given me invaluable advice and guidance with endless patience, kindness and encouragement. I truly appreciate all the time and effort he has spent in helping me to solve the problems encountered even when he is in the midst of his work. Thank you very much.

I also wish to express my sincere gratitude and appreciation to all the staff in Department of Statistics and Applied Probability for providing me a pleasant environment for study and work.

Special thanks to all my friends who helped me in one way or another for their encouragement. Thank you for all the precious advice and help in my study.

Contents

Summary

We analyze DNA palindromes in the Coronavirus and Herpesvirus families. Specifically we study two problems. Problem 1 deals with the overall count of palindromes of a certain length in a genome where we compare the observed number of palindromes of a certain length against its expected number under Markov chain sequence models of the genome. We derive expressions for the mean and standard deviation of the number of palindromes. The resulting z-score enables us to explore whether the observed number of palindromes of a certain length is over- (or under-)represented.

Problem 2 deals with a measure of local clusters of nearly palindromes at or above a certain length. This measure leads to a statistical procedure to predict the replication origins of these viruses.

Key words: DNA sequences, palindrome distributions, Markov chain, under- and over-representation, z-scores, replication origins

List of Tables

List of Figures

Chapter 1

Introduction

This thesis focuses on a special biological word pattern–palindromes. Palindromes (explained below) are involved in a variety of biological processes. For example, the recognition sites for bacterial restriction enzymes to cut foreign DNA are mostly palindromic (Waterman 1995, Chapter 2). Palindromes also play important roles in gene regulation and DNA replication processes (Wagner 1991, Chapters 6, 12, 18; Kornberg and Baker 1992, Chapter 1). It appears that palindromes have to do with DNA-protein binding. The local two-fold symmetry created by the palindrome provides a binding site for DNA-binding proteins which are often dimeric in structure. Such double binding markedly increases the strength and specificity of the binding interaction (Creighton 1993, Chapter 8).

In this thesis, we apply our results to two virus families, namely, the Coronaviruses and the Herpesvirus family.

Unlike these well-studied viruses involved in fatal diseases such as AIDS and various cancers, the coronaviruses have not received much attention until the recent outbreak of SARS. So in this thesis we pay special attention to this SARS virus.

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

We first introduce some relevant DNA concepts and background.

1.1 Examples and Notation

GenBank

GenBank is a free public database where we can access the original sequence of many kinds of genome. The raw data in this paper are all downloaded from the GenBank in 2005.

DNA and RNA

The DNA molecule is in the form of a twisted ladder shape scientists call a "double helix". The rungs of this ladder make up the four-letter DNA alphabet: A, C, G, T. These alphabet pieces bond together according to special rules. A always pairs with T and C always with G. RNA is a single-stranded molecule composed of nucleotide sequences that is similar to the double-stranded DNA. The following is a double strand DNA. It reads exactly the same from the $5'$ to $3'$ on both strands.

$$5' \cdots\cdots GCAATATTGC \cdots\cdots 3'$$

$$3' \cdots\cdots CGTTATAACG \cdots\cdots 5'$$

Herpesvirus and Coronavirus

The Herpesvirus is a double-stranded DNA sequence over the alphabet $\mathscr{A} = \{A, C, G, T\}$. The Coronavirus is a single stranded RNA. In accordance with GenBank convention, we also represent an RNA sequence as a string of letters from $\mathscr{A} = \{A, C, G, T\}$ (although RNA is actually a sequence from $\mathscr{A} = \{A, C, G, U\}$).

DNA word

A DNA word is a segment of DNA. We use $\boldsymbol{w}$ to denote such a word and $w_1, w_2, \cdots, w_m$ to denote the bases of this word. Here m stands for the length of the word $\boldsymbol{w}$. For example, a word ATCG can be expressed as $\boldsymbol{w} = w_1 w_2 w_3 w_4$ where $w_1 = A, w_2 = T, w_3 = C$

and $w_4 = G$. We use w_1' to denote the complementary base of w_1, and $\boldsymbol{w'}$ to denote the inversion of the word $\boldsymbol{w}$. For example, if $w_1 = A$, then $w_1' = T$. If $\boldsymbol{w} = ATC$, then the inversion of the word $\boldsymbol{w}$ is $\boldsymbol{w'} = GAT$.

Palindrome

DNA palindromes (we will abbreviate it to palindromes) are DNA words which are symmetrical in the sense that they read exactly the same as their complementary sequences in the reverse direction. A DNA palindrome is necessarily even in length because the middle base in any odd-length nucleotide string cannot be identical to its complement. For example, ACGT is a palindrome of length four; AATGCATT is a palindrome of length eight. We denote the half length of palindrome by L. So $L = 2$ for palindrome ACGT, and $L = 4$ for palindrome AATGCATT.

For convenience, we define the"left center" of a palindrome. For example, for the palindrome

$$A\underline{C}|GT$$

the base C is the left center; for the palindrome

$$AAT\underline{G}|CATT$$

the base G is the left center.

EMBOSS

EMBOSS (European Molecular Biology Open Software Suite) is a suite of free software tools for nucleotide and protein sequence analysis. It consists of more than 140 programs, ranging from sequence alignment to restriction enzyme mapping. We used the "palindrome" and "comseq" programs.

M0, M1 and M2 model

We analyze the sequences by Markov-Chain models. M0 denotes the i.i.d. Model and M1, M2 denote the Markov chain of order one and order two respectively.

1.2 Main Results

In Chapter 2, we derive the mathematical formulas for the theoretical mean and variance for the number of palindromes at a prescribed length based on a Markov-Chain random-sequence model. We give the specific expressions of their variances under two Markov-Chain models (M0 and M1). For M2 model, because the expressions are complicated and lengthy, we provide an algorithm to calculate them, which can be programmed for numerical calculation.

In Chapter 3, we design a new scoring scheme using approximate palindromes (to be explained in Chapter 3) to provide a measure of abundance of palindromes to predict the locations of replication origins. Then we compare with the current scoring scheme

based on perfect palindromes. The new scoring scheme improves the current work of Chew *et al*. (2005).

1.3 Organization of the Thesis

The organization of the thesis is as follows: In Chapter 2, we will focus on the distribution of the aggregate palindrome counts in a DNA sequence based on Markov-Chain models. In Chapter 3, we will focus on the spatial distribution of the approximate palindrome length and its application for predicting the replication origins. We provide the necessary introduction and literature review in each chapter .

Chapter 2

Exact Length Palindrome Distribution

2.1 Introduction

In this chapter we focus on the aggregate palindrome counts in a DNA sequence. We are interested in whether palindrome counts in a genome is more or less than what would be expected based on some random sequences. We model the genome as a sequence of random variables from some Markov-Chain models. The distribution of the aggregate palindrome counts will be used to assess whether the observed aggregate palindrome count is over-(or under-)represented.

Chew *et al.* (2004) have analyzed the number of palindromes at or above prescribed length. They have derived the theoretical mean and variance for the number of palindromes at or above a prescribed length under the Markov-Chain models. They did not give the theoretical mean and variance for the number of *exact* length palindromes but

rather estimated it by simulation method. This is because the standard deviation of counts of exact length palindromes has not been derived. However, their approach becomes impractical as the Herpesviruses are much longer. Moreover, there are 37 of these herpesviruses now and the increase of the viruses takes even longer time for simulation.

In this chapter we will derive the expressions of theoretical mean and variance for the number of palindromes at a prescribed length under the Markov-Chain sequence model for the genome. Chapter 2 is as follows: In Section 2.2, we will model the genome by Markov-Chain models. In Section 2.3, the mathematical formulas for the theoretical mean and variance for the number of palindromes at a prescribed length are derived based on M0 and M1 models. Then in Section 2.4 we will compare the observed palindrome counts with the expected palindrome counts derived from our models. We apply these models to Coronavirus and Herpesvirus families in this section. Some suggestions on future investigations are provided in Section 2.5.

2.2 Modeling the DNA Sequences

We model the DNA genome as a realization of a sequence of random variables $\xi_1, \xi_2, \ldots, \xi_n$ taking values in $\mathscr{A} = \{A, C, G, T\}$, where n denotes the genome length. Throughout this Chapter, we will assume one of the following:

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

(ii) $\{\xi_1, \xi_2, \ldots, \xi_n\}$form a stationary Markov chain of order 1 (M1);

(iii) $\{\xi_1, \xi_2, \ldots, \xi_n\}$ form a stationary Markov chain of order 2 (M2).

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

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

We say that a palindrome of length *at least* $2L$ occurs at k when $I_{k,L}$ =1. Let random variable X_L denote the total number of palindromes of length *at least* $2L$, that is, $X_L = \sum_{k=L}^{n-L} I_{k,L}$. We are interested in deriving the mean and standard deviation of the random variable Y_L, which is the total number of palindromes of *exact* length $2L$ under the above three Markov-Chain Models. By definitions of Y_L and X_L, it easy to see that $Y_L = X_L - X_{L+1}$. So

$$EY_L = E(X_L - X_{L+1})$$

$$\text{Var}(Y_L) = \text{Var}(X_L) + \text{Var}(X_{L+1}) - 2\text{Cov}(X_L, X_{L+1}).$$

The expectation and variance of X_L have been derived by Chew *et al.* (2004). They have derived the expressions for the expectation and variance of X_L in terms of $\gamma_L(0)$ and $\gamma_L(d)$, where

$$\gamma_L(0) := P[I_{k,L} = 1] \quad \textit{and} \quad \gamma_L(d) := P[I_{k,L} = 1, I_{k+d,L} = 1], \quad d \geq 1.$$

According to Chew *et al.* (2004),

$$E(X_L) = (n-2L+1)\gamma_L(0)$$

$$\text{Var}(X_L) = (n-2L+1)\gamma_L(0)(1-\gamma_L(0)) + 2\sum_{d=1}^{n-2L}(n-2L+1-d)[\gamma_L(d) - \gamma_L(0)^2].$$

What we are interested in is the expectation and variance of Y_L. Hence it follows that

$$EY_L = E(X_L - X_{L+1}) = (n-2L+1)\gamma_L(0) - (n-2L-1)\gamma_{L+1}(0), \tag{2.1}$$

and

$$\begin{aligned}
\mathrm{Var}(Y_L) &= \mathrm{Var}(X_L) + \mathrm{Var}(X_{L+1}) - 2\mathrm{Cov}(X_L, X_{L+1}) \\
&= (n-2L+1)\gamma_L(0)(1-\gamma_L(0)) + 2\sum_{d=1}^{n-2L}(n-2L+1-d)[\gamma_L(d) - \gamma_L(0)^2] \\
&\quad + (n-2L-1)\gamma_{L+1}(0)(1-\gamma_{L+1}(0)) \\
&\quad + 2\sum_{d=1}^{n-2(L+1)}(n-2L-1-d)[\gamma_{L+1}(d) - \gamma_{L+1}(0)^2] \\
&\quad - 2\mathrm{Cov}(X_L, X_{L+1}).
\end{aligned} \tag{2.2}$$

Therefore $E(Y_L)$ can be computed. In order to compute $\mathrm{Var}(Y_L)$, we only need to calculate $\mathrm{Cov}(X_L, X_{L+1})$. For $j \neq i$, we denote $\mathrm{Cov}(\mathrm{I}_{i,L}, \mathrm{I}_{j,L+1})$ by $c_L(j-i)$, and $d = j - i$.

$$\begin{aligned}
&\mathrm{Cov}(X_L, X_{L+1}) \\
&= \mathrm{Cov}\left(\sum_{i=L}^{n-L}\mathrm{I}_{i,L}, \sum_{j=L+1}^{n-L-1}\mathrm{I}_{j,L+1}\right) \\
&= \sum_{i=L}^{n-L}\sum_{j=L+1}^{n-L-1}\mathrm{Cov}(\mathrm{I}_{i,L}, \mathrm{I}_{j,L+1})
\end{aligned}$$

For convenience, we use $c_L(d)$ in the following the calculation. Thus

$$\begin{aligned}
&\mathrm{Cov}(X_L, X_{L+1}) \\
&= \sum_{i=L}^{n-L} \sum_{j=L+1}^{n-L-1} c_L(j-i) \\
&= \sum_{j=L+1}^{n-L-1} c_L(j-L) + \sum_{j=L+1}^{n-L-1} c_L(j-n+L) + \sum_{i=L+1}^{n-L-1} \sum_{j=L+1}^{n-L-1} c_L(j-i) \\
&= (n-2L-1)c_L(0) + \sum_{d=1}^{n-2L-1} [c_L(d) + c_L(-d)] \\
&\quad + \sum_{i=L+1}^{n-L-2} \sum_{j=i+1}^{n-L-1} [c_L(j-i) + c_L(i-j)] \\
&= (n-2L-1)c_L(0) + \sum_{d=1}^{n-2L-1} [c_L(d) + c_L(-d)] \\
&\quad + \sum_{d=1}^{n-2L-2} (n-2L-1-d)[c_L(d) + c_L(-d)] \\
&= mc_L(0) + \sum_{d=1}^{m} (m-d+1)[c_L(d) + c_L(-d)]. \qquad (2.3)
\end{aligned}$$

where $m = n-2L-1$.

The $c_L(d)$ can be further simplified from below:

If $d = 0$, then

$$\begin{aligned}
c_L(0) &= \mathrm{Cov}(\mathrm{I}_{j,L}, \mathrm{I}_{j,L+1}) \\
&= P(\mathrm{I}_{j,L} = 1, \mathrm{I}_{j,L+1} = 1) - \gamma_L(0) \cdot \gamma_{L+1}(0) \\
&= P(\mathrm{I}_{j,L+1} = 1) - \gamma_L(0) \cdot \gamma_{L+1}(0) \\
&= \gamma_{L+1}(0) - \gamma_L(0) \cdot \gamma_{L+1}(0) \\
&= \gamma_{L+1}(0)[1 - \gamma_L(0)].
\end{aligned}$$

If $d \neq 0$, then

$$\begin{aligned} c_L(d) &= \mathrm{Cov}(\mathrm{I}_{i,L}, \mathrm{I}_{i+d,L+1}) \\ &= P(\mathrm{I}_{i,L} \cdot \mathrm{I}_{i+d,L+1} = 1) - P(\mathrm{I}_{i,L} = 1) \cdot P(\mathrm{I}_{i+d,L+1} = 1) \\ &= P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{j,L+1} = 1) - \gamma_L(0) \cdot \gamma_{L+1}(0). \end{aligned}$$

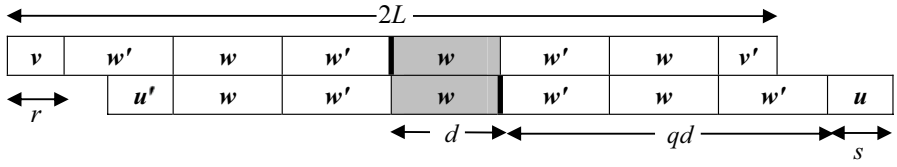
In order to deduce $\mathrm{Var}(Y_L)$, it suffices to calculate the overlapping probabilities $P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1)$ for $d \neq 0$.

2.3 Calculating the Overlapping Probability

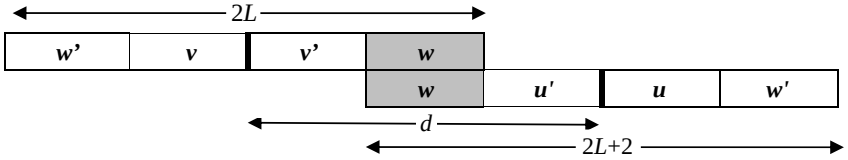
The Markov-Chain model we choose and the value of d determine the overlapping probability. We will first present the general structure of two overlapping palindromes in Section 2.3.1. Then we will derive the overlapping probability under M0 and M1 models separately.

2.3.1 Structure of Two Overlapping Palindromes

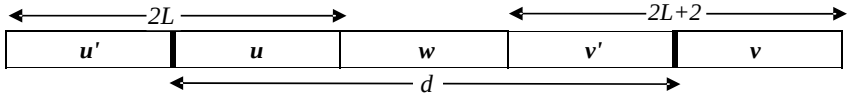
In order to calculate the $P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1)$, we need to find out the general structure of two overlapping palindromes. One palindrome is of length at least $2L$, the other is of length at least $2(L+1)$. Note that d is in fact the distance between the left centers of these two palindromes. We use w' to denote the complementary base of w, and $\boldsymbol{w'}$ to denote the inversion of the word $\boldsymbol{w}$. For example, the inversion of the word $\boldsymbol{w} = w_1 w_2 w_3$ is $\boldsymbol{w'} = w'_3 w'_2 w'_1$. Recall that d is the distance between the left centers of



(a) $1 \leq d \leq L+1$. Here q is quotient when L is divided by d and r is the remainder. The shaded segment w determines the rest of both palindromes



(b) $L+1 < d \leq 2L$. The shaded segment w determines the rest of both palindromes



(c) $d \geq 2L+1$. The two palindromes do not overlap and w denotes the segment between them.

Figure 2.1: Overlapping structures of the two palindromes for different d

the two palindromes and it represents the extent of their overlap. There are three basic patterns in the overlap according to d. We first describe these three patterns when $d > 0$ followed by the description of these structures when $d < 0$.

Lemma 2.3.1 *Suppose a palindrome of length at least* $2L$ *occurs at* i *and another palindrome of length at least* $2L+2$ *occurs at* $i+d$ $(d > 0)$. *We write*

$$L = qd + r, \quad 0 \leq r < d,$$

where q *is the quotient and* r *is the remainder when* L *is divided by* d. *It follows that*

$$L+1 = qd + s, \quad s = r+1.$$

(1) When $1 \leq d \leq L+1$, the span of the two palindromes can be expressed as

$$w_{d-r+1} \dots w_d \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{1} \cdots \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{q} \boldsymbol{w}' w_1 \dots w_s.$$

where $\boldsymbol{w} \in \mathscr{A}^d$.

Note when $r = 0$, the span takes the form of

$$\underbrace{\boldsymbol{w}'\boldsymbol{w}}_{1} \cdots \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{q} \boldsymbol{w}' w_1.$$

(2) When $L+1 < d \leq 2L$, the span of the two palindromes can be expressed as

$$\boldsymbol{w}'\boldsymbol{v}\boldsymbol{v}'\boldsymbol{w}\boldsymbol{u}'\boldsymbol{u}\boldsymbol{w}'$$

where $\boldsymbol{w} \in \mathscr{A}^{2L+1-d}, \boldsymbol{v} \in \mathscr{A}^{d-L-1}$, and $\boldsymbol{u} \in \mathscr{A}^{d-L}$.

(3) When $d \geq 2L+1$, the span of the two palindromes can be expressed as

$$\boldsymbol{u}'\boldsymbol{u}\boldsymbol{w}\boldsymbol{v}'\boldsymbol{v}$$

where $\boldsymbol{u} \in \mathscr{A}^L, \boldsymbol{w} \in \mathscr{A}^{d-2L-1}$, and $\boldsymbol{v} \in \mathscr{A}^{L+1}$.

Proof. We shall prove case 1 first. If $r \neq 0$ and q is odd, the span of the two palindromes is of the form $\boldsymbol{v}\underbrace{\boldsymbol{w}'\boldsymbol{w}}_{1} \cdots \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{q}\boldsymbol{w}'\boldsymbol{u}$. As illustrated by Figure 2.1(a), the overlapping structure of the two palindromes is uniquely determined by the shaded segment $\boldsymbol{w}$. A close examination of v and u show that $\boldsymbol{v} = w_{d-r+1} \dots w_d$ and $\boldsymbol{u} = w_1 \dots w_s$, therefore, the span is

$$w_{d-r+1} \dots w_d \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{1} \cdots \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{q} \boldsymbol{w}' w_1 \dots w_s.$$

If $r \neq 0$ and q is even, however, the span will be the form of

$$w'_r \dots w'_1 \underbrace{\boldsymbol{w}\boldsymbol{w}'}_{1} \cdots \underbrace{\boldsymbol{w}\boldsymbol{w}'}_{q} \boldsymbol{w} w'_d \dots w'_{d-s+1}.$$

We can see that the above two expressions are essentially the same. In fact, we can make the one-to-one transformation: $w_1 \to w'_d, \cdots, w_d \to w'_1$ which reduces to the case when q is odd. So the form of the span does not depend on whether q is even or odd.

In case 1 when $r = 0$, it can be easily checked that the span is the form of

$$\underbrace{\boldsymbol{w}'\boldsymbol{w}}_{1} \cdots \underbrace{\boldsymbol{w}'\boldsymbol{w}}_{q} \boldsymbol{w}' w_1.$$

And similar to the case $r \neq 0$, it does not matter whether q is even or odd.

In case 2 when $L+1 < d \leq 2L$, the span of the two palindromes can be illustrated by Figure 2.1(b). We can see that $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ altogether will determine the whole span. Obviously the lengths of $\boldsymbol{w}, \boldsymbol{v}$, and $\boldsymbol{u}$ are $2L+1-d, d-L-1$ and $d-L$ respectively.

Similarly, from Figure 2.1(c), when $d \geq 2L+1$, the span is of the form of $\boldsymbol{u}\boldsymbol{u}'\boldsymbol{v}\boldsymbol{w}\boldsymbol{w}'$ where the lengths of $\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{v}$ are $L, d-2L-1$ and $L+1$ respectively. □

Now we consider the structure when $d < 0$, that is, the left center of the longer palindrome is on the left of the left center of the shorter palindrome. In fact, when $d < 0$ the overlapping structure is just the reverse of the three basic patterns in Figure 1: if we read the Figure 1 from right to left.

2.3.2 Overlapping Probability for M0 Model

We will abbreviate $c_L(d)$ to $c(d)$. Under M0 Model, the expression of $\mathrm{Cov}(X_L, X_{L+1})$ can be simplified for two reasons:

(i) $c(d) = c(-d)$ when $d \geq 1$;

(ii) $c(d) = 0$ when $d \geq 2L+1$

To see (i),we know from Section 2.2.1 that when $j - i < 0$, the overlapping structure is just the reverse of the stricture when $j - i > 0$. Since under M0 model, that is, the i.i.d. model, the overlapping probability is just the sum over all possible $\boldsymbol{w}$ in case 1, $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$ in cases 2 and 3. Probabilities of this word and its reverse coincide under M0 and hence the sum. When $d \geq 2L+1$, the two palindromes do not physically overlap. By i.i.d. Model, $\mathrm{I}_{i,L}$ and $\mathrm{I}_{i+d,L+1}$ are independent and therefore $\mathrm{Cov}(\mathrm{I}_{i,L}, \mathrm{I}_{i+d,L+1}) = 0$. That is, $c(d)$=0.

These two simplifications lead to

$$\mathrm{Cov}(X_L, X_{L+1}) = \sum_{d=1}^{2L} 2(m-d+1)c(d) + mc(0)$$

where $m = n - 2L - 1$.

In the following lemma we will deduce the $c(d)$ when $d \geq 0$.

Lemma 2.3.2 *Under the assumption of i.i.d. sequence model where (p_A, p_T, p_C, p_G) is the nucleotide distribution, define*

$$\theta := 2(p_A p_T + p_C p_G). \tag{2.4}$$

(1)

$$c(0) = \theta^{L+1}(1-\theta^L). \tag{2.5}$$

(2)

$$c(1) = 2(p_A p_T)^{L+1} + 2(p_C p_G)^{L+1} - \theta^{2L+1}. \tag{2.6}$$

(3) For $2 \le d \le L$, we have the following 2 cases:

(a) $r+s \le d$:

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

(b) $r+s > d$:

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

(4) For $L+1 \le d \le 2L$:

$$c(d) = \left[p_A p_T (p_A + p_T) + p_C p_G (p_C + p_G)\right]^{2L+1-d} \cdot \theta^{2d-2L-1} - \theta^{2L+1}.$$

Proof. To show (1), it has been previously observed that $c(0) = \gamma_{L+1}(0)[1-\gamma_L(0)]$. In Chew *et al*. (2004), it has been proved that $\gamma_L(0) = \theta^L$, so case (1) follows immediately:

$$c(0) = \gamma_{L+1}(0)[1-\gamma_L(0)] = \theta^{L+1}(1-\theta^L).$$

To show (2) when $d = 1$,

$$P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+1,L+1} = 1) = \sum_{w_1 \in \mathscr{A}} P(w_1')^{L+1} \cdot P(w_1)^{L+1}.$$

Thus

$$c(1) = 2(p_A p_T)^{L+1} + 2(p_C p_G)^{L+1} - \theta^{2L+1}.$$

So equation (2.6) follows.

Since $c(d) = P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1) - \gamma_L(0) \cdot \gamma_{L+1}(0)$, we only consider $P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1)$. By Lemma 2.3.1, the span is the of form

$$w_{d-r+1} \dots w_d \underbrace{\boldsymbol{w'w}}_{1} \cdots \underbrace{\boldsymbol{w'w}}_{q} \boldsymbol{w'} w_1 \dots w_s.$$

Let $\boldsymbol{w}^q$ denote the concatenation of $\boldsymbol{w}$ by itself q times. Then the span can be expressed as $w_{d-r+1} \dots w_d \, (\boldsymbol{w'w})^q \boldsymbol{w'} w_1 \dots w_s$.

Therefore,

$$P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{j,L+1} = 1)$$
$$= \sum_{w_1, \cdots, w_d \in \mathscr{A}} P\Bigg(w_{d-r+1} \cdots w_d \underbrace{w'_d \cdots w'_1 w_1 \cdots w_d}_{1} \cdots \underbrace{w'_d \cdots w'_1 w_1 \cdots w_d}_{q} w'_d \cdots w'_1 w_1 \cdots w_s \Bigg).$$

If $r + s \leq d$, we split the $\boldsymbol{w}$ into three parts, α, β and γ where $\alpha = w_1 \cdots w_s, \beta = w_{s+1} \cdots w_{d-r}$ and $\gamma = w_{d-r+1} \cdots w_d$ as illustrated below:

$$\overbrace{w_1 \cdots w_s}^{\alpha} \underbrace{w_{s+1} \cdots w_{d-r}}_{\beta} \overbrace{w_{d-r+1} \cdots w_d}^{\gamma}$$

Hence the overlapping probability

$$
\begin{aligned}
P(\mathrm{I}_{i,L}=1,\mathrm{I}_{j,L+1}=1) &= \sum_{\boldsymbol{\alpha}\in\mathscr{A}^s} P[(\boldsymbol{\alpha}\boldsymbol{\alpha}')^{q+1}] \sum_{\boldsymbol{\gamma}\in\mathscr{A}^r} P[(\boldsymbol{\gamma}\boldsymbol{\gamma}')^{q+1}] \sum_{\boldsymbol{\beta}\in\mathscr{A}^{d-r-s}} P[(\boldsymbol{\beta}\boldsymbol{\beta}')^{q}\boldsymbol{\beta}'] \\
&= \left[\sum_{\alpha_1\in\mathscr{A}} P(\alpha_1)^{q+1}P(\alpha_1')^{q+1}\right]^s \left[\sum_{\gamma_1\in\mathscr{A}} P(\gamma_1)^{q+1}P(\gamma_1')^{q+1}\right]^r \\
&\quad \times \left[\sum_{\beta_1\in\mathscr{A}} P(\beta_1)^{q}P(\beta_1')^{q+1}\right]^{d-r-s} \\
&= \left(p_A^{q+1}p_T^{q+1}+p_T^{q+1}p_A^{q+1}+p_C^{q+1}p_G^{q+1}+p_G^{q+1}p_C^{q+1}\right)^{r+s} \\
&\quad \times \left(p_A^{q}p_T^{q+1}+p_T^{q}p_A^{q+1}+p_C^{q}p_G^{q+1}+p_G^{q}p_C^{q+1}\right)^{d-r-s} \\
&= \left[2(p_Ap_T)^{q+1}+2(p_Cp_G)^{q+1}\right]^{r+s} \\
&\quad \times \left[(p_Ap_T)^q(p_A+p_T)+(p_Cp_G)^q(p_C+p_G)\right]^{d-r-s}.
\end{aligned}
$$

If $r+s>d$, similarly, we split the $\boldsymbol{w}$ into three parts, α,β and γ where $\alpha=w_1\cdots w_{d-r}, \beta = w_{d-r+1}\cdots w_s$ and $\gamma = w_{s+1}\cdots w_d$ as illustrated below:

$$
\overbrace{w_1\cdots w_{d-r}}^{\alpha}\underbrace{w_{d-r+1}\cdots w_s}_{\beta}\overbrace{w_{s+1}\cdots w_d}^{\gamma}
$$

Hence the overlapping probability

$$\begin{aligned}
P(\mathrm{I}_{i,L}=1,\mathrm{I}_{j,L+1}=1) &= \sum_{\boldsymbol{\alpha}\in\mathscr{A}^{d-r}} P[(\boldsymbol{\alpha}\boldsymbol{\alpha}')^{q+1}] \sum_{\boldsymbol{\gamma}\in\mathscr{A}^{d-s}} P[(\boldsymbol{\gamma}\boldsymbol{\gamma}')^{q+1}] \sum_{\boldsymbol{\beta}\in\mathscr{A}^{r+s-d}} P[(\boldsymbol{\beta}\boldsymbol{\beta}')^{q+1}\boldsymbol{\beta}] \\
&= \left[\sum_{\alpha_1\in\mathscr{A}} P(\alpha_1)^{q+1}P(\alpha_1')^{q+1}\right]^{d-r} \left[\sum_{\gamma_1\in\mathscr{A}} P(\gamma_1)^{q+1}P(\gamma_1')^{q+1}\right]^{d-s} \\
&\quad \times \left[\sum_{\beta_1\in\mathscr{A}} P(\beta_1)^{q}P(\beta_1')^{q+1}\right]^{r+s-d} \\
&= \left(p_A^{q+1}p_T^{q+1}+p_T^{q+1}p_A^{q+1}+p_C^{q+1}p_G^{q+1}+p_G^{q+1}p_C^{q+1}\right)^{2d-(r+s)} \\
&\quad \times \left(p_A^{q+2}p_T^{q+1}+p_T^{q+2}p_A^{q+1}+p_C^{q+2}p_G^{q+1}+p_G^{q+2}p_C^{q+1}\right)^{r+s-d} \\
&= \left[2(p_Ap_T)^{q+1}+2(p_Cp_G)^{q+1}\right]^{2d-r-s} \\
&\quad \times \left[(p_Ap_T)^{q+1}(p_A+p_T)+(p_Cp_G)^{q+1}(p_C+p_G)\right]^{r+s-d}.
\end{aligned}$$

For $L+1 \le d \le 2L$:

$$\begin{aligned}
P(\mathrm{I}_{i,L}=1,\mathrm{I}_{j,L+1}=1) &= \sum_{\boldsymbol{u}\in\mathscr{A}^{2L+1-d}} \sum_{\boldsymbol{v}\in\mathscr{A}^{d-L-1}} \sum_{\boldsymbol{w}\in\mathscr{A}^{d-L}} P(\boldsymbol{u}'\boldsymbol{v}\boldsymbol{v}'\boldsymbol{u}\boldsymbol{w}\boldsymbol{w}'\boldsymbol{u}') \\
&= \sum_{\boldsymbol{u}\in\mathscr{A}^{2L+1-d}} P(\boldsymbol{u}'\boldsymbol{u}\boldsymbol{u}') \sum_{\boldsymbol{v}\in\mathscr{A}^{d-L-1}} P(\boldsymbol{v}\boldsymbol{v}') \sum_{\boldsymbol{w}\in\mathscr{A}^{d-L}} P(\boldsymbol{w}\boldsymbol{w}') \\
&= (p_Ap_T^2+p_Tp_A^2+p_Cp_G^2+p_Gp_C^2)^{2L+1-d}\theta^{d-L-1}\theta^{d-L} \\
&= [p_Ap_T(p_A+p_T)+p_Cp_G(p_C+p_G)]^{2L+1-d}\theta^{2d-2L-1}.
\end{aligned}$$

□

2.3.3 Overlapping Probability for M1 Model

For M1 model we observe numerically that $c(d) \neq c(-d)$ for some $d \geq 1$ so we must calculate them separately. First we will calculate $P(\mathrm{I}_{i,L}=1,\mathrm{I}_{i+d,L+1}=1)$ as shown in

the following Lemma 2.3.3. We will explain in Appendix how to deduce $c(-d)$, that is, how to deduce $P(\mathrm{I}_{i,L}=1,\mathrm{I}_{i-d,L+1}=1)$.

Lemma 2.3.3 *Under the assumption of M1 sequences model where $P(w_1,w_2)$ denotes the transition probability from base w_1 to base w_2 and stationary distribution $\pi := (\pi_A,\pi_T,\pi_C,\pi_G)$,*

(1) When $1 \le d \le L$,

$$\begin{aligned} &P(\mathrm{I}_{i,L}=1,\mathrm{I}_{i+d,L+1}=1) \\ &\quad= \sum_{w_1,\cdots,w_d\in\mathscr{A}} K_{r,s,d}P(w_1',w_1)\left[P(w_d,w_d')\prod_{j=1}^{d-1}P(w_{j+1}',w_j')\right]^{q+1} \\ &\qquad\times\left[P(w_1',w_1)\prod_{j=1}^{d-1}P(w_j,w_{j+1})\right]^{q} \end{aligned}$$

where

$$K_{r,s,d}=\begin{cases}\pi(w_{d-r+1})\prod_{j=1}^{s-1}P(w_j,w_{j+1})\prod_{j=d-r+1}^{d-1}P(w_j,w_{j+1}) & r\ge 2\\ \pi(w_d)P(w_1,w_2) & r=1\\ \frac{\pi(w_d')}{P(w_d,w_d')} & r=0\end{cases}.$$

(2) When $d \ge L+1$,

$$\begin{aligned} &P(\mathrm{I}_{i,L}=1,\mathrm{I}_{i+d,L+1}=1) \\ &\quad= \sum_{w_1,\cdots,w_d\in\mathscr{A}} \pi(w_L')P(w_1',w_1)P(w_d,w_d')\prod_{j=1}^{L-1}P(w_{j+1}',w_j') \\ &\qquad\times\prod_{j=1}^{d-1}P(w_j,w_{j+1})\prod_{j=d-L}^{d-1}P(w_{j+1},w_j) \end{aligned}$$

Proof. From Lemma 2.3.1 we can see that when $0 \leq d \leq L$ the span is the form of

$$w_{d-r+1} \dots w_d \underbrace{\boldsymbol{w'w}}_{1} \cdots \underbrace{\boldsymbol{w'w}}_{q} \boldsymbol{w'} w_1 \dots w_s.$$

For $r \geq 2$,

$$\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1) \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} P\left[w_{d-r+1} \cdots w_d \underbrace{w_d^{'} \cdots w_1^{'} w_1 \cdots w_d}_{1} \cdots \underbrace{w_d^{'} \cdots w_1^{'} w_1 \cdots w_d}_{q} w_d^{'} \cdots w_1^{'} w_1 \cdots w_s \right] \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w_{d-r+1}) \prod_{j=1}^{s-1} P(w_j, w_{j+1}) \prod_{j=d-r+1}^{d-1} P(w_j, w_{j+1}) P(w_1', w_1) \\
&\quad \times \left[P(w_d, w_d^{'}) \prod_{j=1}^{d-1} P(w_{j+1}^{'}, w_j^{'}) \right]^{q+1} \left[P(w_1^{'}, w_1) \prod_{j=1}^{d-1} P(w_j, w_{j+1}) \right]^{q}.
\end{aligned}$$

For $r = 1$,

$$\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1) \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} P\left[w_1 \underbrace{w_d^{'} \cdots w_1^{'} w_1 \cdots w_d}_{1} \cdots \underbrace{w_d^{'} \cdots w_1^{'} w_1 \cdots w_d}_{q} w_d^{'} \cdots w_1^{'} w_1 w_2 \right] \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w_d) P(w_1, w_2) P(w_1', w_1) \left[P(w_d, w_d^{'}) \prod_{j=1}^{d-1} P(w_{j+1}^{'}, w_j^{'}) \right]^{q+1} \\
&\quad \times \left[P(w_1^{'}, w_1) \prod_{j=1}^{d-1} P(w_j, w_{j+1}) \right]^{q}.
\end{aligned}$$

For $r = 0$,

$$\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1) \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} P\left[\underbrace{w_d^{'} \cdots w_1^{'} w_1 \cdots w_d}_{1} \cdots \underbrace{w_d^{'} \cdots w_1^{'} w_1 \cdots w_d}_{q} w_d^{'} \cdots w_1^{'} w_1 \right] \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} \frac{\pi(w_d^{'})}{P(w_d, w_d^{'})} P(w_1', w_1) \left[P(w_d, w_d^{'}) \prod_{j=1}^{d-1} P(w_{j+1}^{'}, w_j^{'}) \right]^{q+1} \\
&\quad \times \left[P(w_1^{'}, w_1) \prod_{j=1}^{d-1} P(w_j, w_{j+1}) \right]^{q}
\end{aligned}$$

This complete the proof of the case $1 \leq d \leq L$. Now consider the case $d \geq L+1$.

When $d \geq L+1$, recall from Lemma 2.3.1 (also see Figure 2.3.1 (b) and (c)) that when $L+1 < d \leq 2L$, the span of the two palindromes can be expressed as

$$\boldsymbol{w'vv'wu'uw'}$$

where $\boldsymbol{w} \in \mathscr{A}^{2L+1-d}, \boldsymbol{v} \in \mathscr{A}^{d-L-1}, \boldsymbol{u} \in \mathscr{A}^{d-L}$;

When $d \geq 2L+1$, the span of the two palindromes can be expressed as

$$\boldsymbol{u'uwv'v}$$

where $\boldsymbol{u} \in \mathscr{A}^{L}, \boldsymbol{w} \in \mathscr{A}^{d-2L-1}, \boldsymbol{v} \in \mathscr{A}^{L+1}$.

For convenience, we can combine the above two expressions into a simpler one as a more general form of span when $d \geq L+1$. If we take the bases between the left centers of the palindromes as our $\boldsymbol{w}$. Obviously $\boldsymbol{w} = w_1 \cdots w_d$. We can get the span of form

$$w'_L \cdots w'_1 w_1 \cdots w_d w'_d \cdots w'_{d-L}.$$

Thus we can deduce the overlapping probability from the above form as:

$$\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i+d,L+1} = 1) \\
&\quad = \sum_{w_1,\cdots,w_d \in \mathscr{A}} P[w'_L \cdots w'_1 w_1 \cdots w_d w'_d \cdots w'_{d-L}] \\
&\quad = \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w'_L) P(w'_1, w_1) P(w_d, w'_d) \prod_{j=1}^{L-1} P(w'_{j+1}, w'_j) \\
&\qquad \times \prod_{j=1}^{d-1} P(w_j, w_{j+1}) \prod_{j=d-L}^{d-1} P(w_{j+1}, w_j)
\end{aligned}$$

□

The method of computation is similar for $c(-d)$. Furthermore, the method can be easily adapted to the M2 sequence model.

2.4 Palindrome Counts in Coronaviruses and Herpesviruses

Now that we have derived the theoretical mean and variance of Y_L under the M0, M1 and M2 models, this will enable us to assess whether the observed palindrome count in a genome is too abundant or too rare.

2.4.1 z Scores

Our objective is to assess whether the observed palindrome count of a given exact length in the Coronaviruses and Herpesviruses is more (or less) than the expected, under some specified probability models. We need a statistic to measure the extent of over (or under) representation of a DNA word. The z score is such a statistics. For $L \geq 2$, a standardized frequency under a Markov-Chain Model (M0, M1 or M2) is defined as

$$Z = \frac{Y_L - \mu}{\sigma}$$

where Y_L is the observed number of palindromes of exact length $2L$, and μ and σ denote its expected value and standard deviation respectively. When L is small compared with the genome length n, the distribution of z score will be approximately standard normal. In fact, when L is small compared with the genome length n, X_L is a sum of weakly

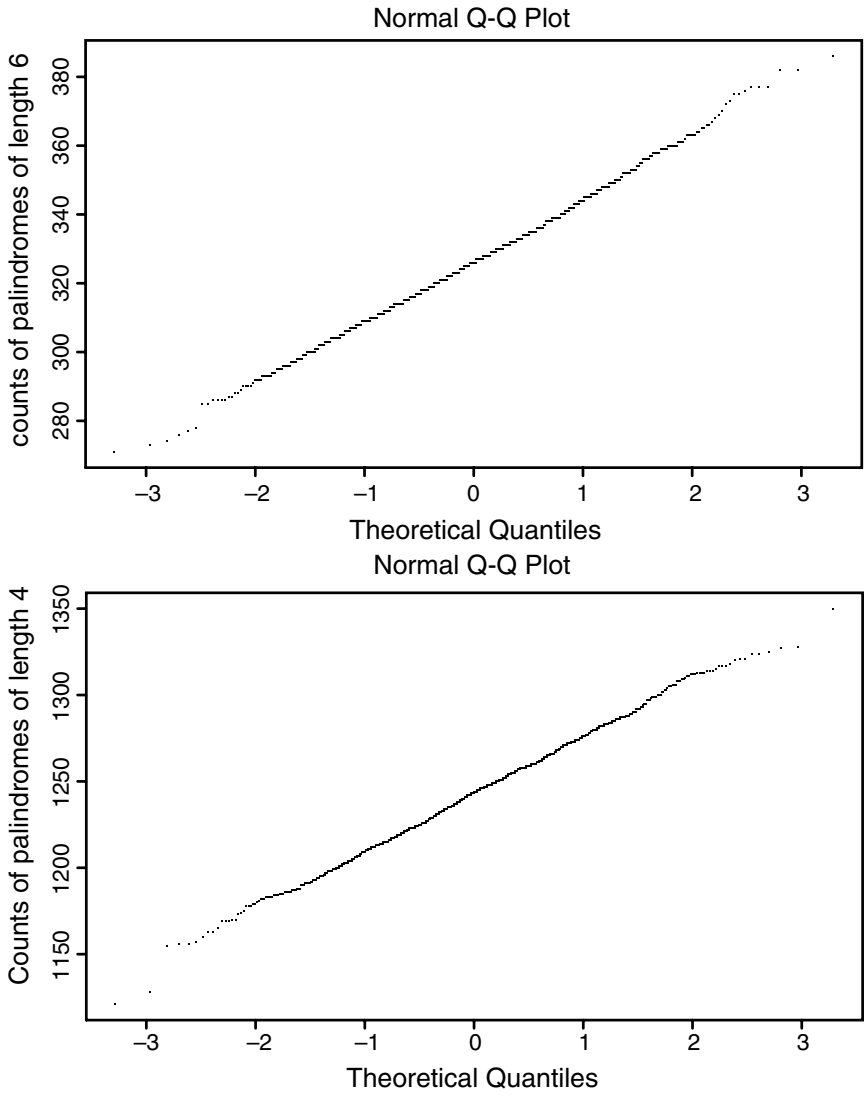


Figure 2.2: Normal Q-Q Plots of Counts of Palindromes of Length Four (Top) and Six (Bottom) in the 1,000 Random Sequences Under the M1 Model for the SARS Genome (Chew et al. 2004)

dependent random indicators $I_{k,L}$ and it is therefore well approximated by a normal distribution (Chew *et al.* 2004). If we let $X_L^{(j)}$ denote the number of occurrences of the j^{th} palindrome in the genome, then the count vector $(X_L^{(1)}, X_L^{(2)}, \cdots X_L^{(4^L)})$ will converge to a multivariate normal distribution as $n \to \infty$ (see Theorem 12.5 in Waterman 1995). Hence X_L will converge to a normal distribution as $n \to \infty$. So for $L = 2$ or 3, and n

in the range 30,000 for Coronaviruses and 100,000 for Herpesviruses, we expect that the distribution of the z scores will be approximately standard normal. This has been justified graphically by $Q-Q$ plots in Chew *et al*. (2004), which are reproduced in Figure 2.2.

Since the z score is approximately standard normal, we can say the count is said to be *over-(or under-) represented*, if the z score is greater than 1.645 (or less than -1.645), that is, in the upper (or lower) 5% of a standard normal distribution, as commonly used in one-tailed hypothesis tests in biological experiments. It should be noted that these cutoff z score values are only a guideline to help us find out interesting observations rather than a strict criterion to make a conclusion.

We compute the z scores of each of the genomes in these two families of viruses: Coronavirus and Herpesvirus.

2.4.2 Palindrome Counts in Coronaviruses

We compute the z scores of the Coronaviruses family in the following data set. It is composed of seven coronaviruses with complete genome sequences. Table 2.1 lists the names of the viruses, their abbreviations, GenBank accession numbers, genome lengths, and base composition of the seven coronaviruses. Tables 2.2–2.4 present the counts of palindromes of exact length four, six, and eight, along with their expected values μ, estimated standard deviations σ, and z scores under M0, M1, and M2 models respectively. From Tables 2.2–2.4 we can see that the exact length four palindrome

Table 2.1: List of coronaviruses to be analyzed.

Name	Abbrev.	Accession	Length	Base composition
SARS coronavirus Urbani	SARS	AY278741	29,727	(0.28,0.20,0.21,0.31)
Avian infectious bronchitis virus	AIBV	NC_0014511	27,608	(0.29,0.16,0.22,0.33)
Bovine coronavirus	BCoV	NC_0030451	31,028	(0.27,0.15,0.22,0.36)
Human coronavirus 229E	Hcov	NC_0026451	27,317	(0.27,0.17,0.22,0.35)
Murine hepatitis virus	MHV	NC_001846	31,357	(0.26,0,18,0.24,0.32)
Porcine epidemic diarrhea virus	PEDV	NC_0034361	28,033	(0.25,0.19,0.23,0.33)
Transmissible gastroenteritis virus	TGV	NC_0023062	28,586	(0.29,0.17,0.21,0.33)

count in each coronavirus analyzed is significantly lower than expected under M0 or M1 model. As for exact length six palindrome count, under-representation of palindromes no longer holds across the whole family, only SARS shows underrepresentation under M1 model. No other obvious patterns exist for length eight palindrome.

M1 model is preferred because variables under M1 model are dependent so the genome dinucleotide compositions can be used. Besides, z scores under M1 are less extreme than those under M0, and thus M1 is more conservative in declaring the palindrome counts in a genome to be significantly different from those in random sequences. M2 model does not show much difference with M1 in this context. So we will use M1 model in the following discussions.

The wide avoidance of palindromes of exact length four in the coronaviruses may have some biological implications. Although there is no previous report of underrepresentation of short palindromes in RNA viruses with eukaryotic hosts, there are some reports about other genomes. The avoidance of short palindromes in some bacterial and phage DNA genomes has been reported in several studies (Karlin *et al*. 1992; Merkl and Fritz

Table 2.2: z scores for coronaviruses palindromes under M0 model

Abbrev.	2				3				4			
	Counts	μ	σ	z	Counts	μ	σ	z	Counts	μ	σ	z
SARS	1144	1464.4	37.54	**-8.53**	284	377.19	19.43	**-4.80**	90	97.2	9.88	-0.72
AIBV	1142	1396.7	36.83	**-6.92**	320	365.64	19.23	**-2.37**	91	95.7	9.85	-0.48
BCoV	1360	1556.0	39.06	**-5.02**	389	405.15	20.31	-0.79	98	105.5	10.36	-0.72
HCoV	1054	1399.4	36.81	**-9.38**	287	369.13	19.28	**-4.26**	82	97.4	9.92	-1.55
MHV	1328	1497.2	38.04	**-4.45**	340	378.48	19.47	**-1.98**	82	95.7	9.81	-1.39
PEDV	1079	1335.0	35.93	**-7.12**	274	336.93	18.37	**-3.43**	79	85.0	9.25	-0.65
TGV	1180	1455.3	37.56	**-7.33**	306	382.43	19.65	**-3.89**	85	100.5	10.08	-1.54

Table 2.3: z scores for coronaviruses palindromes under M1 model

Abbrev.	2				3				4			
	Counts	μ	σ	z	Counts	μ	σ	z	Counts	μ	σ	z
SARS	1144	1242.6	34.82	**-2.83**	284	327.30	17.98	**-2.41**	90	86.5	9.29	0.38
AIBV	1142	1229.7	34.17	**-2.57**	320	326.84	17.89	-0.38	91	87.0	9.30	0.43
BCoV	1360	1476.4	36.60	**-3.18**	389	390.31	19.39	-0.07	98	103.3	10.11	-0.53
HCoV	1054	1146.8	33.44	**-2.78**	287	307.52	17.40	-1.18	82	82.7	9.08	-0.07
MHV	1328	1421.2	36.57	**-2.55**	340	364.28	18.88	-1.29	82	93.4	9.64	-1.19
PEDV	1079	1169.7	33.73	**-2.69**	274	302.85	17.29	**-1.67**	79	78.6	8.85	0.05
TGV	1180	1239.4	34.57	**-1.72**	306	333.16	18.10	-1.50	85	89.8	9.46	-0.50

Table 2.4: z scores for coronaviruses palindromes under M2 model

Abbrev.	2				3				4			
	Counts	μ	σ	z	Counts	μ	σ	z	Counts	μ	σ	z
SARS	1144	1214.1	34.53	**-2.03**	284	320.52	17.80	**-2.05**	90	84.3	9.17	0.62
AIBV	1142	1216.5	33.92	**-2.19**	320	322.59	17.77	-0.15	91	85.6	9.23	0.58
BCoV	1360	1459.8	36.56	**-2.73**	389	384.35	19.27	0.24	98	101.0	9.99	-0.30
HCoV	1054	1127.2	33.22	**-2.20**	287	301.57	17.25	-0.84	82	80.6	8.97	0.16
MHV	1328	1406.5	36.33	**-2.16**	340	359.78	18.75	-1.06	82	91.7	9.55	-1.02
PEDV	1079	1152.7	33.60	**-2.19**	274	299.39	17.20	-1.48	79	77.5	8.79	0.17
TGV	1180	1233.5	34.63	-1.54	306	330.10	18.03	-1.34	85	88.7	9.40	-0.39

1996; Rocha *et al*. 1998, 2001). This is generally explained as defense mechanisms of the bacterial and phage genomes. This could help genomes to protect themselves against being destroyed by restriction enzymes capable of cutting up DNA molecules

at certain palindromic sites. From our observation of avoidance of short palindromes in coronavirus genomes, we are interested in investigating whether there is any possible interaction of the short palindromes in the coronavirus genomes with the immune system of the host cells that might do harm to virus.

For length-six palindromes under M1 model only SARS is found to be significantly underrepresented while the other six coronaviruses are not. This avoidance of length-six palindromes might offer a more effective protection for SARS virus, making it more difficult to be destroyed. Would this contribute 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.

2.4.3 Palindrome Counts in Herpesviruses

We compute the z scores of the Herpesviruses family in the following data set. It consists of 37 Herpesviruses with complete genome sequences. Table 2.5 lists the names of the viruses, abbreviations, GenBank accession numbers, genome lengths, and base composition of the 37 Herpesviruses. Tables 2.6–2.8 present the counts of palindromes of exact length four, six and eight along with their expected values μ, estimated standard deviations σ and z scores under M0, M1, and M2 models respectively.

We find that the Herpesviruses family is quite different from the Coronavirus family. For $L = 2$, for example, there are 17 viruses which are underrepresented under M0 model, while under M1 model only 9 viruses are underpresented. It indicates that the

model selection heavily influences the z scores. Recall that for Coronaviruses (See Tables 2.2–2.4) the number of underrepresented viruses are almost the same among three models (M0, M1 and M2). We also find that for viruses AIHV-1, CeHV-15, EHV-2, HHV-4, IcHV-1 and MuHV-4, the z scores change dramtically from underrepresented to overrepresented. It means that the model selection has more influences on these viruses. We may look into the reasons in future research.

The underrepresentation of different length palindromes in Herpesviruses are different. The shorter palindromes tend to have more underrepresentation under each model(M0, M1 or M2) which are similar to the Coronaviruses. For example, we observe that 17 viruses are underrepresented under M0 model and 13 and 4 viruses are underrepresented under M1 and M2 models respectively. One difference from the Coronavirus is that the avoidance of shorter palindromes is not across the whole family. Since the Herpesviruses are divided into several subfamilies, the relationship between the classification and the underrepresentation should be an area to explore.

2.5 Future Investigation of Coronavirus

We have analyzed the total length-four palindrome count in Coronavirus. However, we have not looked into the individual length-four palindromes. For example, the length-four palindrome ACGT and another length-four palindrome TTAA are both counted as total length-four palindrome. But they may have quite different influence on the underrepresentation of the total palindrome count. Consequently, a thorough examination

Table 2.5: List of herpesviruses to be analyzed.

Name	Abbrev.	Accession	Length	Base composition
Alcelaphine herpesvirus 1	AlHV-1	NC_002531.1	130,608	(0.27, 0.24, 0.22, 0.26)
Ateline herpesvirus 3	AtHV-3	NC_001987.1	108,409	(0.32, 0.19, 0.17, 0.31)
Bovine herpesvirus 1	BoHV-1	NC_001847.1	135,301	(0.14, 0.36, 0.37, 0.14)
Bovine herpesvirus 4	BoHV-4	NC_002665.1	108,873	(0.30, 0.21, 0.20, 0.29)
Bovine herpesvirus 5	BoHV-5	NC_005261.1	138,390	(0.12, 0.37, 0.38, 0.13)
Callitrichine herpesvirus 3	CalHV-3	NC_004367.1	149,696	(0.26, 0.25, 0.25, 0.25)
Cercopithecine herpesvirus 1	CeHV-1	NC_004812.1	156,789	(0.13, 0.37, 0.38, 0.13)
Cercopithecine herpesvirus 7	CeHV-7	NC_002686.1	124,138	(0.29, 0.21, 0.20, 0.30)
Cercopithecine herpesvirus 8	CeHV-8	NC_006150.1	221,454	(0.26, 0.25, 0.24, 0.25)
Cercopithecine herpesvirus 15	CeHV-15	NC_006146.1	171,096	(0.18, 0.31, 0.31, 0.20)
Cercopithecine herpesvirus 17	CeHV-17	NC_003401.1	133,719	(0.24, 0.27, 0.26, 0.23)
Equine herpesvirus 1	EHV-1	NC_001491.2	150,224	(0.22, 0.29, 0.28, 0.22)
Equine herpesvirus 2	EHV-2	NC_001650.1	184,427	(0.22, 0.29, 0.28, 0.21)
Equine herpesvirus 4	EHV-4	NC_001844.1	145,597	(0.25, 0.25, 0.25, 0.25)
Gallid herpesvirus 2	GaHV-2	NC_002229.2	174,077	(0.28, 0.22, 0.22, 0.28)
Gallid herpesvirus 3	GaHV-3	NC_002577.1	164,270	(0.23, 0.27, 0.27, 0.23)
Human herpesvirus 1	HHV-1	NC_001806.1	152,261	(0.16, 0.34, 0.34, 0.16)
Human herpesvirus 2	HHV-2	NC_001798.1	154,746	(0.15, 0.35, 0.35, 0.15)
Human herpesvirus 3	HHV-3	NC_001348.1	124,884	(0.27, 0.23, 0.23, 0.27)
Human herpesvirus 4	HHV-4	NC_001345.1	172,281	(0.20, 0.30, 0.29, 0.20)
Human herpesvirus 5 strain AD169	HHV-5A	NC_001347.2	230,287	(0.22, 0.28, 0.29, 0.21)
Human herpesvirus 5 strain Merlin	HHV-5M	NC_006273.1	235,645	(0.21, 0.29, 0.29, 0.21)
Human herpesvirus 6	HHV-6	NC_001664.1	159,321	(0.29, 0.22, 0.21, 0.29)
Human herpesvirus 6B	HHV-6B	NC_000898.1	162,114	(0.29, 0.22, 0.21, 0.29)
Human herpesvirus 7	HHV-7	NC_001716.2	153,080	(0.32, 0.20, 0.17, 0.31)
Human herpesvirus 8	HHV-8	NC_003409.1	137,508	(0.24, 0.27, 0.26, 0.23)
Ictalurid herpesv 1	IcHV-1	NC_001493.1	134,226	(0.21, 0.28, 0.28, 0.22)
Meleagrid herpesvirus 1	MeHV-1	NC_002641.1	159,160	(0.26, 0.24, 0.24, 0.26)
Murid herpesvirus 1	MuHV-1	NC_004065.1	230,278	(0.20, 0.29, 0.30, 0.21)
Murid herpesvirus 2	MuHV-2	NC_002512.2	230,138	(0.19, 0.30, 0.31, 0.20)
Murid herpesvirus 4	MuHV-4	NC_001826.1	119,450	(0.27, 0.24, 0.23, 0.26)
Ostreid herpesvirus 1	OsHV-1	NC_005881.1	207,439	(0.31, 0.19, 0.19, 0.30)
Pongine herpesvirus 4	PoHV-4	NC_003521.1	241,087	(0.19, 0.31, 0.31, 0.19)
Psittacid herpesvirus 1	PsHV-1	NC_005264.1	163,025	(0.19, 0.31, 0.30, 0.20)
Saimiriine herpesvirus 2	SaHV-2	NC_001350.1	112,930	(0.33, 0.18, 0.16, 0.32)
Suid herpesvirus 1	SuHV-1	NC_006151.1	143,461	(0.13, 0.37, 0.37, 0.13)
Tupaiid herpesvirus 1	TuHV-1	NC_002794.1	195,859	(0.17, 0.33, 0.34, 0.17)

Table 2.6: z scores for herpesvirus palindrome of length four ($L = 2$) under M0, M1 and M2 model

Abbrev.	Counts	M0			M1			M2		
		μ	σ	z	μ	σ	z	μ	σ	z
AIHV-1	5,046	6168.0	88.2	**-12.72**	4934.5	70.0	1.59	4880.3	69.6	2.38
AtHV-3	4,575	5689.7	88.2	**-12.64**	4765.5	67.5	**-2.82**	4723.4	65.9	**-2.25**
BoHV-1	10,548	8533.8	114.3	17.63	10356.4	50.7	3.78	10366.5	50.7	3.58
BoHV-4	4,121	5350.8	83.4	**-14.74**	4188.3	63.7	-1.06	4203.5	61.3	-1.35
BoHV-5	11,183	9256.1	120.8	15.96	10864.2	51.2	6.22	10971.3	46.7	4.53
CalHV-3	5,834	7013.0	93.7	**-12.59**	6061.0	77.4	**-2.93**	6058.5	77.2	**-2.91**
CeHV-1	11,027	10385.5	127.6	5.03	10559.7	80.0	5.84	10716.0	74.0	4.20
CeHV-7	6,412	6171.6	90.0	2.67	6261.7	76.1	1.98	6312.8	74.6	1.33
CeHV-8	9,336	10381.4	113.9	**-9.17**	9924.0	97.9	**-6.01**	9888.4	97.4	**-5.67**
CeHV-15	7,738	8787.0	108.7	**-9.65**	7170.6	83.8	6.77	7287.2	84.1	5.36
CeHV-17	6,435	6287.7	88.8	1.66	6205.2	76.7	3.00	6213.4	76.5	2.90
EHV-1	7,169	7249.4	96.3	-0.83	7181.1	81.6	-0.15	7215.3	80.3	-0.58
EHV-2	7,745	8965.2	107.5	**-11.35**	7261.3	85.0	5.69	7190.3	83.9	6.61
EHV-4	6,654	6825.4	92.4	**-1.86**	6731.6	80.3	-0.97	6727.2	79.1	-0.93
GaHV-2	8,659	8363.2	103.3	2.86	8565.7	89.1	1.05	8615.1	87.8	0.50
GaHV-3	8,367	7766.5	98.9	6.07	8481.3	87.5	-1.31	8459.0	87.1	-1.06
HHV-1	8,743	8763.5	112.8	-0.18	8465.3	83.5	3.33	8481.4	80.8	3.24
HHV-2	9,692	9318.7	117.9	3.17	9315.2	83.0	4.54	9376.3	79.4	3.98
HHV-3	6,304	5914.1	86.4	4.51	6074.4	75.7	3.03	6115.8	74.8	2.52
HHV-4	7,016	8608.0	106.4	**-14.96**	6814.0	82.4	2.45	6916.8	82.7	1.20
HHV-5A	11,462	11167.1	119.8	2.46	11642.7	100.6	**-1.80**	11684.7	100.3	**-2.22**
HHV-5M	11,645	11458.5	121.5	1.53	11989.4	101.7	**-3.39**	12020.0	101.6	**-3.69**
HHV-6	6,882	7751.9	100.0	**-8.70**	7248.1	83.7	**-4.37**	7141.7	82.7	**-3.14**
H6B	6,922	7863.7	100.6	**-9.37**	7293.0	84.1	**-4.41**	7185.3	83.2	**-3.16**
HHV-7	6,772	8072.7	105.3	**-12.35**	6872.3	81.1	-1.24	6739.3	80.6	0.41
HHV-8	5,664	6491.6	90.4	**-9.16**	5793.9	75.3	**-1.72**	5763.7	74.6	-1.34
IcHV-1	6,267	6453.0	90.8	**-2.05**	6041.5	76.1	2.96	6110.1	75.9	2.07
MeHV-1	7,928	7489.5	96.9	4.52	8012.0	86.3	-0.97	8037.7	85.5	-1.28
MuHV-1	11,467	11345.1	121.5	1.00	11578.8	101.4	-1.10	11682.3	101.5	**-2.12**
MuHV-2	12,561	11664.6	124.6	7.20	12055.1	102.1	4.95	12087.2	101.7	4.66
MuHV-4	4,489	5624.1	84.0	**-13.51**	4428.0	66.4	0.92	4363.0	65.5	1.92
OsHV-1	8,767	10545.7	118.6	**-15.00**	9290.7	93.6	**-5.59**	9236.6	91.7	**-5.12**
PoHV-4	12,496	12342.8	128.6	1.19	12566.3	102.7	-0.69	12617.4	103.0	-1.18
PsHV-1	9,465	8255.8	104.8	11.54	9312.8	83.8	1.81	9442.7	81.7	0.27
SaHV-2	5,175	6145.5	92.8	**-10.45**	5196.4	70.2	-0.30	5135.0	68.9	0.58
SuHV-1	10,375	9299.6	120.1	8.95	9779.0	72.8	8.19	10008.9	64.1	5.71
TuHV-1	12,031	10896.9	124.4	9.12	11926.4	89.1	1.17	11750.8	93.7	2.99

The underlined values are over-presented and the bold ones are under-presented.

Table 2.7: z scores for herpesvirus palindrome of length six ($L = 3$) under M0, M1 and M2 model

Abbrev.	Counts	M0			M1			M2		
		μ	σ	z	μ	σ	z	μ	σ	z
AIHV-1	1,353	1548.9	45.2	**-4.33**	1283.0	35.7	1.96	1288.0	35.8	1.82
AtHV-3	1,321	1523.2	46.5	**-4.35**	1297.7	35.6	0.65	1302.6	35.5	0.52
BoHV-1	3,356	2562.0	63.9	12.42	3257.0	39.8	2.49	3274.5	39.9	2.04
BoHV-4	1,186	1376.4	43.2	**-4.41**	1121.3	33.3	1.94	1153.4	33.2	0.98
BoHV-5	3,661	2885.4	69.0	11.24	3491.7	41.0	4.13	3546.7	39.4	2.90
CalHV-3	1,537	1752.7	48.0	**-4.50**	1542.2	39.1	-0.13	1552.8	39.3	-0.40
CeHV-1	3,267	3217.3	72.7	0.68	3280.0	50.5	-0.26	3349.9	48.6	-1.71
CeHV-7	1,693	1598.6	46.7	2.02	1652.1	40.0	1.02	1689.3	40.1	0.09
CeHV-8	2,467	2595.5	58.4	**-2.20**	2502.9	49.8	-0.72	2508.3	49.8	-0.83
CeHV-15	2,094	2321.4	57.0	**-3.99**	1940.1	43.7	3.52	1976.7	44.0	2.67
CeHV-17	1,765	1574.9	45.5	4.18	1577.0	39.4	4.77	1596.2	39.6	4.26
EHV-1	1,825	1844.3	49.7	-0.39	1833.3	42.3	-0.20	1864.3	42.3	-0.93
EHV-2	2,367	2290.9	55.5	1.37	1916.9	43.6	10.32	1923.2	43.7	10.16
EHV-4	1,738	1706.4	47.3	0.67	1701.7	41.0	0.89	1722.6	41.1	0.38
GaHV-2	2,280	2121.9	53.2	2.97	2192.3	46.3	1.89	2223.6	46.3	1.22
GaHV-3	2,245	1951.7	50.8	5.78	2159.2	45.8	1.88	2163.1	45.8	1.79
HHV-1	2,538	2483.6	61.1	0.89	2429.2	47.1	2.31	2455.7	46.6	1.77
HHV-2	2,886	2716.8	64.9	2.61	2743.0	48.8	2.93	2786.6	47.9	2.08
HHV-3	1,606	1487.6	44.3	2.67	1558.4	39.1	1.22	1587.7	39.3	0.47
HHV-4	1,973	2236.5	55.4	**-4.76**	1834.7	42.6	3.24	1865.4	42.9	2.51
HHV-5A	2,972	2849.3	61.9	1.98	3009.2	53.5	-0.70	3049.1	53.8	-1.43
HHV-5M	3,032	2928.6	62.8	1.65	3104.5	54.3	-1.34	3141.9	54.6	**-2.01**
HHV-6	1,904	1982.0	51.7	-1.51	1878.8	43.1	0.58	1870.4	42.9	0.78
H6B	1,856	2006.8	51.9	**-2.90**	1888.4	43.2	-0.75	1877.9	43.1	-0.51
HHV-7	1,869	2167.5	55.6	**-5.37**	1887.2	42.9	-0.42	1855.9	42.6	0.31
HHV-8	1,527	1629.8	46.4	**-2.22**	1468.7	38.2	1.53	1475.8	38.2	1.34
IcHV-1	1,769	1637.9	46.8	2.80	1572.6	39.3	4.99	1608.5	39.7	4.04
MeHV-1	2,047	1876.7	49.7	3.43	2026.5	44.6	0.46	2047.2	44.6	0.00
MuHV-1	3,009	2922.6	63.0	1.37	3037.1	54.0	-0.52	3075.2	54.3	-1.22
MuHV-2	3,429	3057.1	65.1	5.72	3231.2	55.3	3.57	3271.1	55.5	2.84
MuHV-4	1,235	1409.8	43.1	**-4.06**	1152.5	33.9	2.44	1154.4	33.9	2.38
OsHV-1	2,224	2768.9	62.0	**-8.79**	2485.3	49.3	**-5.30**	2495.5	49.1	**-5.54**
PoHV-4	3,249	3254.6	67.3	-0.08	3329.0	55.8	-1.43	3364.2	56.1	**-2.05**
PsHV-1	2,366	2162.5	54.7	3.72	2485.5	47.0	**-2.55**	2544.8	46.6	**-3.84**
SaHV-2	1,533	1682.2	49.5	**-3.02**	1435.8	37.4	2.60	1430.4	37.2	2.75
SuHV-1	3,279	2841.2	67.8	6.45	2994.3	47.0	6.06	3100.0	43.9	4.07
TuHV-1	3,449	3023.9	66.7	6.37	3385.2	53.0	1.20	3375.5	54.7	1.34

The underlined values are over-presented and the bold ones are under-presented.

Table 2.8: z scores for herpesvirus palindrome of length eight ($L = 4$) under M0, M1 and M2 model

Abbrev.	Counts	M0			M1			M2		
		μ	σ	z	μ	σ	z	μ	σ	z
AIHV-1	335	389.0	22.8	**-2.37**	333.3	18.2	0.09	340.1	18.4	-0.28
AtHV-3	405	407.8	23.9	-0.12	353.5	18.7	2.75	360.1	18.9	2.38
BoHV-1	1,193	769.2	34.5	12.27	1024.1	26.0	6.51	1034.9	26.1	6.05
BoHV-4	358	354.0	21.9	0.18	299.7	17.3	3.37	313.0	17.5	2.57
BoHV-5	1,268	899.5	38.0	9.71	1122.2	27.2	5.37	1146.1	26.7	4.57
CalHV-3	427	438.0	24.1	-0.46	392.4	19.8	1.75	397.3	19.9	1.49
CeHV-1	1,144	996.7	39.9	3.70	1018.9	30.0	4.17	1045.4	29.5	3.35
CeHV-7	442	414.1	23.8	1.17	436.0	20.8	0.29	452.2	21.1	-0.49
CeHV-8	607	648.9	29.4	-1.43	631.3	25.1	-0.97	636.3	25.2	-1.16
CeHV-15	510	613.3	29.2	**-3.54**	525.0	22.9	-0.66	540.1	23.2	-1.30
CeHV-17	448	394.5	22.9	2.34	400.8	20.0	2.36	410.6	20.2	1.85
EHV-1	462	469.2	25.1	-0.29	468.0	21.6	-0.28	481.7	21.8	-0.90
EHV-2	564	585.4	28.1	-0.76	506.1	22.5	2.57	516.3	22.7	2.10
EHV-4	425	426.6	23.8	-0.07	430.2	20.7	-0.25	441.8	20.9	-0.80
GaHV-2	558	538.4	26.9	0.73	561.2	23.6	-0.13	573.8	23.8	-0.66
GaHV-3	523	490.4	25.6	1.27	549.6	23.3	-1.14	553.7	23.4	-1.31
HHV-1	699	703.9	32.2	-0.15	697.1	25.9	0.08	710.3	25.9	-0.44
HHV-2	815	792.1	34.6	0.66	807.7	27.5	0.27	827.0	27.4	-0.44
HHV-3	415	374.2	22.3	1.83	399.9	19.9	0.76	412.3	20.2	0.13
HHV-4	485	581.1	28.2	**-3.41**	494.1	22.2	-0.41	506.5	22.5	-0.96
HHV-5A	791	727.0	31.3	2.04	777.8	27.6	0.48	795.1	27.9	-0.15
HHV-5M	820	748.5	31.8	2.25	803.9	28.1	0.57	820.5	28.3	-0.02
HHV-6	511	506.7	26.2	0.16	487.1	22.0	1.09	493.0	22.2	0.81
H6B	472	512.1	26.3	-1.53	489.1	22.1	-0.77	493.9	22.2	-0.99
HHV-7	567	582.0	28.6	-0.52	518.1	22.7	2.16	514.8	22.6	2.31
HHV-8	413	409.2	23.4	0.16	372.3	19.3	2.11	378.3	19.4	1.79
IcHV-1	471	415.8	23.6	2.34	409.4	20.2	3.05	424.6	20.5	2.26
MeHV-1	526	470.3	25.0	2.23	512.6	22.6	0.59	521.5	22.7	0.20
MuHV-1	851	752.9	32.0	3.07	796.6	28.0	1.94	811.4	28.3	1.40
MuHV-2	976	801.2	33.2	5.26	866.2	29.1	3.77	886.1	29.4	3.05
MuHV-4	322	353.4	21.7	-1.45	299.6	17.3	1.29	303.9	17.4	1.04
OsHV-1	616	727.0	31.7	**-3.50**	665.2	25.7	**-1.91**	673.5	25.8	**-2.23**
PoHV-4	916	858.2	34.5	1.68	881.8	29.3	1.17	897.4	29.6	0.63
PsHV-1	734	566.5	27.9	6.00	663.3	25.0	2.82	684.7	25.1	1.96
SaHV-2	445	460.5	25.7	-0.60	398.0	19.9	2.36	401.2	19.9	2.20
SuHV-1	1,027	868.0	37.0	4.30	916.8	28.0	3.94	956.9	27.1	2.58
TuHV-1	1,080	839.1	34.9	6.91	961.0	29.6	4.02	964.9	30.2	3.82

The underlined values are over-presented and the bold ones are under-presented.

of the relative abundance of individual length-four palindromes, conditional on the total length-four palindrome count may shed further light of the biological importance of palindromes in these genomes.

Chapter 3

Scoring Approximate Palindrome Clusters in the Prediction of Replication Origins

3.1 Introduction

Recall that a palindrome is a special word in which a short segment of nucleotide bases is immediately followed by its reverse complement. Previous studies show that around the replication origins of some viruses there is a high concentration of palindromes. Therefore describing the spatial abundance of palindromes in a genome may provide a good computational tool to predict where the replication origins are. Replication origins are places on the DNA molecules where replication processes are initiated. As DNA

replication is the central step in the reproduction of many viruses, understanding the molecular mechanisms involved in DNA replication is of great importance in developing strategies to control the growth and spread of viruses (Delecluse and Hammerschmidt, 2000). As the experimental determination of replication origins in DNA involves labor-intensive laboratory procedures (Hamzeh 1990; Zhu 1998; Newton and Theis 2002), one way that may save time and resources would be to scan the genome sequence for the expected palindromes by a computer program before an experimental search for replication origins is launched. This computational approach has successfully located the replication origin oriLyt on the human cytomegalovirus (HHV-5A) by Masse *et al*. (1992) and then been confirmed by experimentation. Masse *et al*. (1992) analyzed these data by the high concentration of palindromes of length 10 or above clustering within a window of 1000 bases.

Leung *et al*. (1994) first provided an evaluation criterion for assessing palindrome clusters by modeling the occurrences of palindromes using the scan statistics (Glaz 1989, Dembo and Karlin 1992).We call the scoring scheme Palindrome Count Scheme (PCS). This scoring scheme is further developed in the articles of Leung and Yamashita (1999), and Leung *et al.* (2005). This scheme, however, essentially assesses a window of the genome by only the counts of palindrome contained in it. It ignores the actual extent of the palindrome lengths. This drawback has caused it to miss some replication origins which contain one extremely long palindrome rather than a cluster of moderately ones. Chew *et al* (2004) recognize this drawback and present another two new schemes for evaluating palindrome clusters and use the new schemes to predict the

origins of replication in the herpesvirus. Their new schemes have showed substantial improvement over the original scan statistics criterion. The two new schemes are called Palindrome Length Scheme (PLS) and Base-pair Weighted Scheme (BWS) respectively. However, we observe that the new scheme (PLS) can be further improved. As Chew *et al.* (2005) mentioned that some of the origins missed by their new algorithms are actually rather long approximate palindromes. They are missed in the PLS because only the exact palindromes are considered. Approximate palindromes are similar to the perfect palindromes except that approximate palindromes allows up to one error in the reverse complement.

In the following we will present another new scoring scheme using approximate palindromes allowing up to ONE error, namely, the Approximate Palindromes Length Score Scheme (APLS). The known (experimentally confirmed or) replication origins among the herpesviruses will help us assess the approximate palindrome-based algorithm.

The organization of Chapter 3 is as follows: In Section 3.2, we will first introduce our new Approximate Palindromes Length Scheme (APLS). In Section 3.3, the significant approximate palindrome clusters obtained from the herpesviruses are presented and their association with replication origins is also discussed. Comparison between the new scoring scheme and the PLS is also discussed here. Finally in Section 3.4 we conclude with a few remarks about future works towards a more accurate replication origin prediction scheme.

3.2 Approximate Palindrome Length Scheme (APLS)

Unusual clusters of palindromes can be exploited to predict replication origins for the herpesvirus family. We propose a computational method to identify unusual clusters of palindromes. Table 2.5 (in Chapter 2) 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 April 2005. 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 approximate palindromes at or above a prescribed length; (2) use Approximate Palindromes Length Score (APLS) to score the palindromes; (3) compute a score for each window of the genome according to the chosen scoring scheme; and (4) select regions with high scores.

3.2.1 Locating Palindromes at or Above a Prescribed Length

Choosing L

We need to consider palindromes at or above a prescribed length because very short palindromes occur frequently by chance. So 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 HHV-5A virus, for the choice of L. This choice incorporates the length of the sequence,

as well as the base frequencies in the choice of L. Using this criterion, L is chosen to be 6 for the BoHV-1, BoHV-5,BoHV-1, CeHV-1, HHV-1 and , HHV-2 and SuHV-1 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 2000), to extract the palindrome positions and lengths. Each of these palindromes will be assigned a score according to a our APLS scoring scheme chosen in the Section 3.2.2.

Filtering Redundant Approximate Palindromes

Only the nonredundant palindromes are kept for the analysis. That is, if one approximate palindrome is completely contained in a longer one, the shorter approximate palindrome will be discarded. There are two types of redundant palindromes: One type is that a shorter palindrome is contained in a longer palindrome with the same left center. For example, the length 12 palindrome ACCGTGCACGGT contains the length 10 palindrome CCGTGCACGG (G is their common left center). EMBOSS automatically discards all the shorter palindromes and report only the longest one. Another type is that shorter palindrome is contained in a longer palindrome WITHOUT using the same left center. For example,the length 12 palindrome GATATGCATATC contains the two length 4 palindromes ATAT. They have some common pieces but do not have same left center. EMBOSS will report GATAT, TGATATGCATATC, and ATAT, however, we propose that we should only count once and write a short program to filter out the two ATAT's which lie inside this length 12 palindromes.

3.2.2 Using Approximate Palindromes Length Scheme (APLS) to Score the Palindromes

We propose a refinement of Palindromes Length Scheme in Chew *et al.*(2005). Our new score scheme -Approximate Palindrome Length Scheme (APLS) must have three characteristics: First, recall that we only analyze palindromes of length at least $2L$ so any palindrome of length less than $2L$ will always get a score 0. Second, the longer length palindromes will receive higher scores; Third, since we allow up to one error in palindrome, the position of the occurrence of the error will affect the final score. So some adjustments according to the error position need to be done.

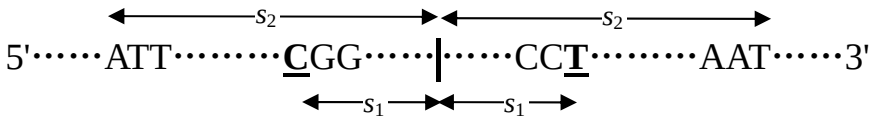


Figure 3.1: Approximate palindrome of length $2s_2$

For convenience, we define two lengths: s_1 and s_2 where $2s_2$ is the length of the approximate palindrome and $2s_1$ is the length of the exact palindrome contained in the approximate palindrome (see one example in Figure 3.1). The underlined C and T are not complementary. So C is the error base. Obviously the $(s_1 + 1)$ is the distance between the position of the error base and the left center of the palindrome.

(1) If the approximate palindrome does not contain the error base, it is in fact a real perfect palindrome. In this case, an approximate palindrome of length $2s_2$ ($2s_2 \geqslant 2L$) is given a score s_2/L. For example, if we let $L = 6$, a palindrome of length 12

(s_2=6) will get a score of 1(6/6=1), while another one of length 14 will get a score of 1.17 (7/6=1.17).

(2) If the approximate palindrome contains one error base such that $s_1 < L$, the approximate palindrome of length $2s_2$ (note $2s_2 \geqslant 2L$) is given a score $s_2/L - 1$.

(3) If the approximate palindrome contains the error base but $s_1 \geqslant L$, the approximate palindrome of length $2s_2$ ($2s_2 \geqslant 2L$) is given a score $(s_2 + s_1)/2L$.

The theoretical justification of this scoring scheme is like this: The score in case 1 is adopted to agree with that in Chew *et al.* (2005). For case 2, since $s_1 < L$, this length $2s_2$ approximate palindrome would not have been extracted if we only consider perfect palindromes of length at or above $2L$. So we assign score $s_2/L - 1$. If $s_1 \geqslant L$, the score should be between s_2/L and s_1/L. So we use the average score $(s_2 + s_1)/2L$.

3.2.3 Computing the Window Score

After every approximate palindrome has been assigned a score, a series of window scores need to be calculated. The score of a window in the genome is simply the total of the scores of all the approximate palindromes occurring in this window. An approximate palindrome is considered to be in the window if its left center is. Following Chew *et al.* (2005), we choose the window length m at 0.5% of the genome length, rounded down to the nearest hundred bases. Also, we let consecutive windows overlap by half their lengths. That is, the first window spans the first through the m-th bases, the second the

(m/2+1)st to (3m/2)th bases, and so on. Every window score is recorded for ranking in the next step.

3.2.4 Selecting Regions With Significant Approximate Palindrome Clusters

We rank top scoring windows for predicting locations of replication origins. There does not appear to be any obvious rule to determine the number of top scoring windows that one should take. In accordance with Chew *et al.* (2005), we first select top 7 windows. We find that using the top 3 to 5 ranked windows for prediction works well for the herpesviruses. The middle position of each selected top window is the specific predicted location we are looking for.

3.3 Result and Discussion

Our interest is to examine the correspondence between these significant approximate palindrome clusters and the actual confirmed locations of the replication origins. From various sources like the annotations in the GenBank file of these sequences and the references therein, plus published genetic maps and other biomedical articles (Farrel, 1993; Masse *et al*., 1992; McGeoch and Schaffer, 1993; Baumann *et al*., 1988), Chew *et al.* (2005) compile a list of replication origins in 17 herpesviruses. Table 3.1 presents the name of virus and also the location range of the replication origins. It is well known

Table 3.1: Known replication origins of Herpesvirus

Virus	Known ORIs/Names		Virus	Known ORIs/Names	
BoHV–1	111080–111300	(OriS)	HHV–1	62475	(OriL)
	126918–127138	(OriS)		131999	(OriS)
BoHV–4	97143–98850	(OriLyt)		146235	(OriS)
BoHV–5	113206–113418	(OriLyt)	HHV–2	62930	(OriL)
	129595–129807	(OriLyt)		132760	(OriS)
CeHV–1	61592–61789	(OriL1)		148981	(OriS)
	61795–61992	(OriL2)	HHV–3	110087–110350	
	132795–132796	(OriS1)		119547–119810	
	132998–132999	(OriS2)	HHV–4	7315–9312	(OriP)
	149425–149426	(OriS2)		52589–53581	(OriLyt)
	149628–149629	(OriS1)	HHV–5	93201–94646	(OriLyt)
CeHV–7	109627–109646		HHV–6	67617–67993	(OriLyt)
	118613–118632		HHV–6B	68740–69581	(OriLyt)
EHV–1	126187–126338		HHV–7	66685–67298	
EHV–4	73900–73919	(OriL)	MuHV–2	75666–78970	(OriLyt)
	119462–119481	(OriS)	SHV1	63848–63908	(OriL)
	138568–138587	(OriS)		114393–115009	(OriS)
				129593–130209	(OriS)

that herpesviruses have multiple replication origins. So we altogether have 35 known replication origins in 17 viruses. Note we take the middle points of the replication origins range as the the real exact replication origins.

Table 3.2 lists the regions with significant clusters of palindromes as found by the PCS and APLS. It shows the top 7 scoring windows for each of the 37 herpesviruses under both the PLS and APLS 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. In practice, 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. Bold entries denote the middle positions of the windows which are within 2 map units of known replication origins where a map unit stands for 1% of the genome length. If the distance from the mid-point of the window to the mid-point of the closest replication origin is within 2 map units, we say this middle position of window correctly predicted the replication origin. Shaded rows without any bold entries show that the computational method fails to predict the known origins of replication. Finally, rows that are not shaded denote those viruses whose origins of replication are not known, as far as we know. The number underlined under APLS scheme is the new correctly predicted origin compared with the PLS scheme.

Table 3.2: High scoring windows of PLS and APLS

Virus	PLS Scoring							APLS Scoring						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
AIHV-1	113701	32701	123301	27301	127501	110701	95101	32701	127501	60901	123301	1501	42001	39901
AtHV-3	99001	54751	97001	1001	25501	36751	107751	67251	98501	63751	57501	38001	25501	47501
BoHV-1	**113401**	**124501**	103801	134401	87301	107101	131101	105901	132301	**113701**	**124501**	48001	5701	31201
BoHV-4	30251	54751	72251	26501	11501	48501	19751	**97501**	39251	54751	11501	48001	35001	78251
BoHV-5	78001	108301	134701	19201	6601	36901	33901	35101	108301	134701	6601	32101	42601	20101
CalHV-3	116201	133351	23101	56351	14001	18901	30101	78751	116201	133351	52851	23101	14701	16801
CeHV-1	**133001**	**149451**	**61601**	113051	56351	117601	109901	**133001**	**149451**	**61601**	152951	129151	95551	116551
CeHV-7	18601	93601	15601	24601	**110701**	**117601**	51301	78301	24601	34201	**108601**	68101	**118501**	48001
CeHV-8	161151	147401	198001	170501	166651	44551	122651	161151	147401	88551	170501	184801	108901	5501
CeHV-15	8001	34801	138801	109201	152001	68801	114001	8001	138801	35201	168401	109601	78801	170401
CeHV-17	132601	117601	3301	35101	87001	60001	22801	132601	5701	117601	3301	34501	60901	105601
EHV-1	146651	116201	47601	123201	140001	94151	50751	116551	146651	**125301**	137901	108851	44801	82951
EHV-2	6301	54001	173251	140401	46351	131851	164701	54001	140401	46351	17551	6301	173251	145801
EHV-4	105351	142801	3851	109901	53551	64751	115151	142451	115501	105001	108501	**138951**	**119001**	6301
GaHV-2	160801	801	137601	42401	46401	75201	108801	137601	801	11601	46401	34001	126801	114401
GaHV-3	158801	138401	11201	122401	105201	154801	1201	158801	138401	41201	134401	122001	8001	125601
HHV-1	**62301**	**129851**	**148401**	48301	55651	78401	91701	**62301**	1051	125301	149451	128801	151901	78401
HHV-2	74551	7351	119701	28001	45151	5251	12951	**63001**	125651	1401	79801	**129851**	**151901**	72451
HHV-3	**119401**	**110101**	100501	49201	1501	60001	13501	**110101**	**119701**	100501	30901	12301	21901	34501

Table 3.2 (continued): High scoring windows of PLS and APLS

Virus	PLS Scoring							APLS Scoring						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
HHV-4	**7601**	**53201**	**51201**	127601	79201	85601	81201	**7601**	**51201**	**53201**	31201	40401	12801	16001
HHV-5A	**94051**	196351	77001	174901	64351	86901	53901	**94051**	174901	19251	163351	229351	167201	190851
HHV-5M	175451	94051	153451	77001	86901	167751	201301	144651	175451	154001	94051	163901	99001	132551
HHV-6	30101	8051	110601	**67901**	89251	125651	98701	**67901**	151201	29051	8051	97651	136151	21001
HHV-6B	8801	90401	**69201**	132801	162001	12001	60801	**69201**	132801	8801	90401	139601	98801	30001
HHV-7	133351	9451	127401	152251	29751	140701	43751	9451	152601	128451	133351	78401	107101	81551
HHV-8	23401	9451	15001	136501	19201	29101	130801	119401	23401	102001	29101	125701	18901	27001
IcHV-1	55501	9451	89701	124801	19201	15001	130501	6001	121501	110701	63901	55501	34201	58801
MeHV-1	5601	9451	11551	40951	97651	134751	72801	79451	134751	117951	5601	83651	67901	87501
MuHV-1	92951	9451	200201	130351	210651	67101	108351	92951	132001	142451	128701	201301	45101	68751
MuHV-2	**75901**	9451	83601	101751	127601	118251	**79201**	**75901**	83601	45101	103401	8251	155101	95701
MuHV-4	99251	9451	62001	50751	106251	751	30251	99251	26251	119001	49501	101251	37001	90501
OsHV-1	21001	9451	185001	187501	197501	204501	207001	72501	146001	17001	22501	126501	144001	174501
PoHV-4	91201	9451	177001	130201	24001	142201	63601	101401	149401	137401	65401	90601	130201	142201
PsHV-1	130401	9451	26801	60801	18801	43201	154801	18801	151601	130401	126001	156001	26401	64801
SaHV-2	103751	9451	27751	29751	81501	3251	6751	103751	10751	2751	33251	75001	66501	57001
SuHV-1	37801	9451	93101	30451	85051	78751	43051	58801	39201	86101	93101	12951	43051	78751
TuHV-1	134101	9451	144901	50401	85051	107551	58501	134101	10801	7651	50401	128701	85051	107551

Table 3.3: Sensitivity and PPV measures of the two scoring schemes

	PLS							APLS						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Sensitivity	23	40	54	57	60	63	66	31	43	57	63	69	77	77
PPV	47	41	37	29	25	22	19	65	44	39	32	28	26	23

Prediction Accuracy

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

$$\text{Sensitivity} = \frac{\text{No. of ORIs that are significant clusters}}{\text{No. of ORIs}}$$

$$\text{PPV} = \frac{\text{No. of significant clusters that are ORIs}}{\text{No. of significant clusters}}$$

The sensitivity and PPV using one to 7 top scoring windows are given in percentages. Note that as the number of windows increases, we gain in sensitivity but at the same time loses in PPV.

Table 3.3 shows the performance of the PLS and APLS schemes. We can see that the sensitivity and PPV are both improved by APLS. More importantly, from Table 3.2 we can see that APLS predicted 7 more new origins of four viruses than PLS. This is a big improvement since we only have 17 viruses under analysis with known origins. Note from Table 3.2 that APLS missed two origins **129851** and **148401** compared with PLS under the virus HHV-1. This is because we only consider middle positions of the

windows which are within 2 map units of known replication origins. These two locations **129851, 148401** happened to be 2.1 map units away. So these two positions are missed. However, the distance 2 map units is just an approximate criterion so if we relax a little this criterion value we would get an even much more improved result from APLS.

3.4 Concluding Remark

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

Our APLS scheme is tested on herpesviruses and still needs to be tested on other DNA viruses. We have allowed one error base in approximate palindromes under APLS. So

far, we have not made use of approximate palindromes that allow several more errors, but this would be an area to explore.

References

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

[2] 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.

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

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

[5] Chew, D.S.H., Choi, K.P., Heidner H., and Leung M.Y. 2004. Palindromes in SARS and Other Coronaviruses. *INFORMS Journal on Computing* Vol. 16, No. **4**, Fall 2004, 331–340

[6] Chew, D.S.H., Choi, K.P., Heidner H., and Leung M.Y. 2005. Scoring schemes of palindrome clusters for more sensitive prediction of replication origins in herpesviruses, manuscript.

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

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

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

[10] Farrell, P.J. 1993. Epstein-Barr virus, in OBrien, S.J., ed., *Genetic Maps*, 6th ed., *Book 1, Viruses*, Cold Spring Harbor Laboratory Press.

[11] Glaz , J. 1989. Approximations and bounds for the distribution of the scan statistics. *J. Am. Statist. Assoc.* **84**, 560-566.

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

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

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

[15] 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.

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

[17] Leung, M.Y. and Yamashita, T.E. 1999. Applications of the scan statistic in DNA sequence analysis. In *Scan Statistics and Applications* (eds. J. Glaz and N. Balakrishnan), 269-286. Birkhauser Publishers, Boston.

[18] Leung, M.Y., Choi, K.P., Xia, A. and Chen, L.H.Y. 2005. Nonrandom clusters of palindromes in herpesvirus genomes. *J. Computat. Biol.* **12**, 331-354.

[19] 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. USA.* **89**, 5246-5250.

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

[21] Merkl, R., H. J. Fritz. 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** 4146-4151.

[22] Newton, C.S. and Theis, J.F. 2002. DNA replication joins the revolution: whole genome views of DNA replication in budding yeast. *BioEssays* **24**, 300-304.

[23] Rice, P., Longden, I. and Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genetics.* **16**, 276-277.

[24] Rocha, E.P., Danchin, A. and Viari, A. 2001. Evolutionary role of restriction/ modification systems as revealed by comparative genome analysis. *Genome Res.* 11, 946-958.

[25] Rocha, E.P., Viari, A. and, Danchin, A. 1998. Oligonucleotide bias in *Bacillus subtilis:* General trends and taxonomic comparisons. *Nucleic Acids Res*. 26, 2971-2980.

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

[27] Waterman, M. S. 1995. *Introduction to Computational Biology*. Chapman & Hall, New York.

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

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

Appendix

Derivation of $c(-d)$

Overlapping Probability for M1 Model

For M1 model we observe numerically that $c(d) \neq c(-d)$ for some $d \geq 1$. In Chapter 2 we have deduced $c(d)$ when $d > 0$. In the following we will show how to deduce $c(-d)$, that is, how to deduce $P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1)$.

Lemma .0.1 *For M1 model,* $P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1)$ *is calculated as following*

(1) When $1 \leq d \leq L$,

$$
\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1) \\
&\quad = \sum_{w_1,\cdots,w_d \in \mathscr{A}} K_{r,s,d} \\
&\qquad \times \left[P(w_1, w_1') \prod_{j=1}^{d-1} P(w_j^{'}, w_{j+1}^{'}) \right]^{q+1} \left[P(w_d', w_d) \prod_{j=1}^{d-1} P(w_{j+1}, w_j) \right]^{q}.
\end{aligned}
$$

where

$$K_{r,s,d} = \begin{cases} \pi(w_s)P(w'_d, w_d)\prod_{j=1}^{s-1} P(w_{j+1}, w_j)\prod_{j=d-r+1}^{d-1} P(w_{j+1}, w_j) & r \geq 2 \\ \pi(w_2)P(w_2.w_1)P(w'_d, w_d) & r = 1 \cdot \\ \pi(w_1) & r = 0 \end{cases}$$

(2) When $d \geq L+1$,

$$\begin{aligned} &P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1) \\ &\quad = \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w'_L)P(w'_1, w_1)P(w_d, w'_d)\prod_{j=1}^{L-1} P(w'_{j+1}, w'_j) \\ &\qquad \times \prod_{j=1}^{d-1} P(w_j, w_{j+1}) \prod_{j=d-L}^{d-1} P(w_{j+1}, w_j) \end{aligned}$$

Proof. From Lemma 2.3.1 we can see that when $0 \leq d \leq L$ the span is the form of

$$w_s \cdots w_1 w'_1 \cdots w'_d \underbrace{w_d \cdots w_1 w'_1 \cdots w'_d}_{1} \cdots \underbrace{w_d \cdots w_1 w'_1 \cdots w'_d}_{q} w_d \cdots w_{d-r+1}$$

For $r \geq 2$,

$$\begin{aligned} &P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1) \\ &= \sum_{w_1,\cdots,w_d \in \mathscr{A}} P\left[w_s \cdots w_1 w'_1 \cdots w'_d \underbrace{w_d \cdots w_1 w'_1 \cdots w'_d}_{1} \cdots \underbrace{w_d \cdots w_1 w'_1 \cdots w'_d}_{q} w_d \cdots w_{d-r+1}\right] \\ &= \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w_s)P(w'_d, w_d)\prod_{j=1}^{s-1} P(w_{j+1}, w_j) \prod_{j=d-r+1}^{d-1} P(w_{j+1}, w_j) \\ &\qquad \times \left[P(w_1, w'_1)\prod_{j=1}^{d-1} P(w'_j, w'_{j+1})\right]^{q+1} \left[P(w'_d, w_d)\prod_{j=1}^{d-1} P(w_{j+1}, w_j)\right]^{q}. \end{aligned}$$

For $r = 1$,

$$
\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1) \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} P\left[w_2 w_1 w_1' \cdots w_d' \underbrace{w_d \cdots w_1 w_1' \cdots w_d'}_{1} \cdots \underbrace{w_d \cdots w_1 w_1' \cdots w_d'}_{q} w_d \right] \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w_2) P(w_2.w_1) P(w_d', w_d) \\
&\quad \times \left[P(w_1, w_1') \prod_{j=1}^{d-1} P(w_j^{'}, w_{j+1}^{'}) \right]^{q+1} \left[P(w_d', w_d) \prod_{j=1}^{d-1} P(w_{j+1}, w_j) \right]^{q}.
\end{aligned}
$$

For $r = 0$,

$$
\begin{aligned}
&P(\mathrm{I}_{i,L} = 1, \mathrm{I}_{i-d,L+1} = 1) \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} P\left[w_1 w_1' \cdots w_d' \underbrace{w_d \cdots w_1 w_1' \cdots w_d'}_{1} \cdots \underbrace{w_d \cdots w_1 w_1' \cdots w_d'}_{q} \right] \\
&= \sum_{w_1,\cdots,w_d \in \mathscr{A}} \pi(w_1) \\
&\quad \times \left[P(w_1, w_1') \prod_{j=1}^{d-1} P(w_j^{'}, w_{j+1}^{'}) \right]^{q+1} \left[P(w_d', w_d) \prod_{j=1}^{d-1} P(w_{j+1}, w_j) \right]^{q}.
\end{aligned}
$$

This complete the proof of the case $1 \leq d \leq L$. Now consider the case $d \geq L+1$.

From Lemma 2.3.3 we know when $d \geq 0$ the span of form is

$$w_L' \cdots w_1' w_1 \cdots w_d w_d' \cdots w_{d-L}'.$$

If we consider $d < 0$, the span form should be reversed as

$$w_{d-L}' \cdots w_d' w_d \cdots w_1 w_1' \cdots w_L'.$$

Thus we can deduce the overlapping probability from this reverse form as:

$$
\begin{aligned}
&P(\mathrm{I}_{i,L}=1,\mathrm{I}_{i-d,L+1}=1)\\
&\quad=\sum_{w_1,\cdots,w_d\in\mathscr{A}} P[w'_{d-L}\cdots w'_d w_d\cdots w_1 w'_1\cdots w'_L]\\
&\quad=\sum_{w_1,\cdots,w_d\in\mathscr{A}} \pi(w'_{d-L})P(w_1,w'_1)P(w'_d,w_d)\prod_{j=1}^{L-1}P(w'_j,w'_{j+1})\\
&\qquad\times\prod_{j=1}^{d-1}P(w_{j+1},w_j)\prod_{j=d-L}^{d-1}P(w'_j,w'_{j+1})
\end{aligned}
$$

□

Similar method can be easily adapted to the M2 sequence model.