INTEGRATING DNA SEQUENCE FEATURES FOR MORE ACCURATE PREDICTION OF REPLICATION ORIGINS IN SOME DOUBLE–STRANDED DNA VIRAL GENOMES

ZHAO WANTING

(Master of Science, Northeast Normal University, China)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2010

Acknowledgements

This thesis would not have been possible without the support and help of many people. It is my pleasure to express my gratitude to all of them.

I would like to thank my supervisors, Associate Professor Choi Kwok Pui and Dr. Li Jialiang, whose invaluable advice and guidance, endless patience and encouragement have been crucial to the completion of this thesis. During the past four years, I have been fortunate to receive their continuous support and to learn a lot from them, not only on the way to do research, but also the careful and precise manner to conduct scientific research. I truly appreciate all the time and effort they have spent in helping me to solve the problems encountered.

I would like to express my sincere gratitude and appreciation to Professor Bai Zhidong and Professor Chen Zehua for his continuous encouragement and support. My gratitude also goes to the National University of Singapore for awarding me a research scholarship, and the Department of Statistics and Applied Probability for providing an excellent research environment. During my Ph.D. programme

I received continuous help from staff in our department, especially our helpful IT support personnel Ms. Yvonne Chow and Mr. Zhang Rong for advice and assistance in computing.

I warmly thank Dr. Chew Soon Huat, David for his valuable advice and friendly help. His extensive discussions around my work have been very helpful for this study.

It is a great pleasure to thank my friendly colleagues Mr. Loke Chok Kang for much help learning computer software, and Dr. Wang Xiaoying and Dr. Zhao Jingyuan for useful discussion during my study. I also would like to thank my friends: Dr. Zhang Rongli, Mr. Wang Xiping, Ms. Li Hua, who have given me much help in my study and life. Sincere thanks to all my friends who helped me in one way or another.

Finally, I am greatly indebted to my parents, who have never failed to encourage me and to support me whenever they could. I feel a deep sense of gratitude for my husband Yu Dingyi, for his love, thoughtfulness and cheering me on.

Contents

Summary

The research of replication origins is critical to understanding the molecular mechanisms involved in DNA replication. Many computational methods based on on individual sequence feature have been developed for predicting locations of replication origins in viruses. However, a particular sequence feature known as close direct repeats has thus far not been used to predict replication origins in herpesviruses. In addition, no studies to date have predicted replication origins by integrating multiple, related sequence features. The aim of this study was to integrate DNA sequence features for more accurate prediction of replication origins in some double-stranded DNA viral genomes.

A computational method to predict the likely locations of replication origins was developed in this thesis. Empirical evidences showed that replication origins often located around regions with an unusually high concentration of palindromes, close direct repeats and AT content. Generalized additive models were then built up and fitted by quantifying these sequence features in herpesvirus genomes with known replication origins. The explanatory variables set of generalized additive

models contained window scores of palindromes, close direct repeats, AT content and their local maxima. The optimal model was chosen by the area under the ROC curve (AUC) criterion, and a standard leave-one-out cross-validation method was employed to assess the predictive performance of the model.

We further refined the GAM approach by integrating additional DNA sequence features, such as the subfamily of a virus family, standardized window numbers of virus genome sequences, and dinucleotide scores of each window of virus genome sequences. A stepwise model selection procedure (GAM31 (AUC)) was performed by the AUC criterion. The similar procedure was performed on caudoviruses, since they share some common properties with herpesviruses. The predictive accuracy of our GAM31 (AUC) approach surpassed existing methods of replication origins prediction in herpesviruses and caudoviruses. For herpesviruses, the GAM31 (AUC) approach outperforms Chew's palindrome-based approach by scoring schemes BWS_1 and PLS in terms of both the sensitivity and positive predictive values (PPV) using the top 1-10 windows. The highest sensitivity and PPV attained by our GAM31 (AUC) approach were 88% and 55% respectively, which were better than those of the best approach introduced by Chew *et al.* (2005), i.e., 79% and 47% respectively. For caudoviruses, the sensitivity and PPV achieved by the GAM31 (AUC) approach when we choose top 3 windows were 62% and 25% respectively, which were almost twice as the LSSVM23 approach introduced by Cruz-Cano *et al.* in 2010.

The key contribution of this study is that the generalized additive modeling approach extends previous work on integrating DNA sequence features for the more accurate prediction of replication origins in some double-stranded DNA viral genomes. Moreover, the AUC criterion, which is a good summary measure to evaluate the overall classification accuracy for identifying a dichotomous response, was applied to select the best model among several reasonable models to improve the predictive accuracy of replication origins in viruses. Our generalized additive modeling approach that integrates DNA sequence features appears effective in identifying replication origins in herpesviruses and caudoviruses.

List of Tables

List of Figures

Chapter 1

Introduction

Herpesviridae is a large, ancient family of DNA viruses that infect many vertebrates and even lower organisms (Davison *et al.*, 2005). Members of this family are also known as herpesviruses. Herpesviruses share a common structure–all herpesviruses are enveloped, double-stranded DNA viruses with relatively large complex genomes that range in size from 120 to over 230 k base-pairs (bp) (Roizman *et al.*, 1991). The base composition G+C content of herpesvirus DNA varies from 31% to 75% (Roizman *et al.*, 1991).

Herpesviruses inflict much harm to human beings and other animals. They have been associated with fatal diseases such as AIDS and cancers, while others pose risks in immunosuppressive post-transplantation therapies (Labrecque *et al.*, 1995; Vital *et al.*, 1995; Biswas *et al.*, 2001; Bennett *et al.*, 2001). Many animal herpesviruses are harmful to agriculture. For example, the alcelaphine herpesvirus

1 is a causative agent of the lethal lymphoproliferative disease malignant catarrhal fever in cattle and deer (Bridgen, 1991). Because herpesviruses endanger the health and lives of humans and animals, doing research on them in order to develop strategies to control their growth and spread is of great value.

As pointed out by Chew *et al.* in 2005, a detailed understanding of the molecular mechanisms involved in DNA replication is very crucial, because DNA replication plays a significant role in the reproduction of herpesviruses. An origin of replication (also known as replication origin) is a site on the genome at which DNA replication is initiated (Ghosh, 2005). Identification of these locations is crucial to understand DNA replication. However, identifying the location of replication origins in the genome is a labor-intensive task. With the increasing availability of genomic DNA sequence data, naturally, computational methodologies for predicting replication origins have been devised (Masse *et al.*, 1992). Thus far, a considerable number of herpesviruses have been completely sequenced, which can be obtained from the NCBI database (http://www.ncbi.nlm.nih.gov/). Based on the information of herpesvirus genome sequences, in the thesis, we build and explore appropriate statistical models that integrate genomic sequence features to improve the prediction of likely locations of replication origins in herpesviruses.

Sections 1.1 and 1.2 provide an overview of the motivation and background of our study. In Section 1.1, the basic biological background of DNA is introduced. In Section 1.2, we describe the genome characteristics and biological properties of

herpesviridae. In Section 1.3, we introduce the replication origins in herpesviruses in more detail. The overall organization of this thesis is given in Section 1.4.

1.1 Biological Background

We first introduce some relevant DNA concepts and background. DNA is short for deoxyribonucleic acid, the genetic material that determines the makeup of all living cells and many viruses. DNA is capable of self-replication and synthesis of RNA. The long-term storage of information is the main function of DNA molecules. The genome is the sequence of the individual bases of the nucleic acid that determines hereditary features of living organisms and some viruses. This sequence is used to make all the proteins of the organism in the appropriate time and place by way of a complex series of interactions (See Lewin, 2004. Chapter 1, section 1.1). The amounts of bases in DNAs vary among different species.

The DNA molecule consists of two long chains of nucleotides twisted into a shape called a "double helix". The DNA double helix is joined by hydrogen bonds between four kinds of bases: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). The DNA double helix exhibits a unique complementary base pairing structure, with each type of base on one strand forming a bond with only one type of base on the other strand; A only bonds to T, and C only bonds to G (see Figure 1.1). That is, purines form hydrogen bonds to pyrimidines (see

Watson *et al.*, 1953). The two strands in a double helix of DNA can be pulled apart like a zipper; either high temperatures or a mechanical force can separate two strands of DNA (Clausen-Schaumann *et al.*, 2000).

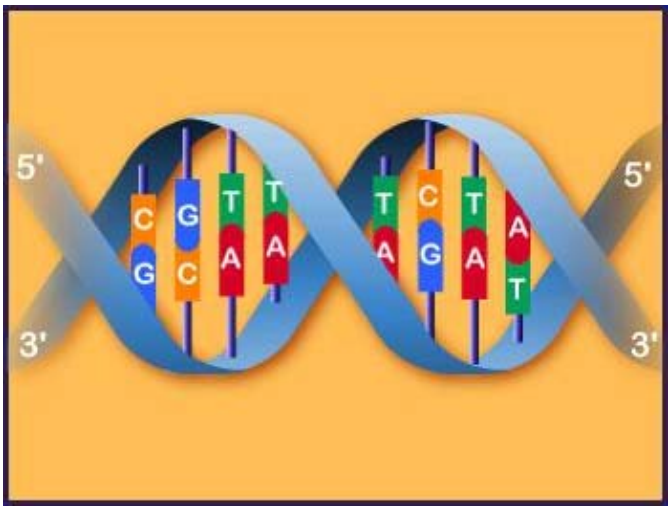


Figure 1.1: DNA base pairing helix.
A bonds to T, and C bonds to G.

(Retrieved 1 January 2010, from http://members.cox.net/amgough/Fanconi-genetics-genetics-primer.htm)

The two types of base pairs form distinct numbers of hydrogen bonds; G and C form three hydrogen bonds, while A and T form two hydrogen bonds (see Figure 1.2) (Roy *et al.*, 2008). DNA with low GC-content is less stable than DNA with high GC-content. Some people believe that this phenomenon is due to the extra hydrogen bond of a GC base pair (Nguyen *et al.*, 1998). However, contrary to popular belief, this is actually due to the contribution of stacking interactions, since hydrogen bonding does not provide stability, but rather specificity of the pairing (See Yakovchuk *et al.*, 2006). In the laboratory, the strength of the interaction of DNA double strands can be measured by determining the temperature

required to break the hydrogen bonds. The DNA double strands separate into two independent molecules when all the base pairs in the double strands melt. Both the length of a DNA double helix and the percentage of AT content determine the strength of the association between the two strands of DNA. Long DNA helices with a low AT content have stronger interacting strands, while short helices with a high percentage of AT base pairs have weaker interacting strands (Chalikian *et al.*, 1999). In biology, parts of the DNA double helix can be pulled apart easily due to high AT content (deHaseth *et al.*, 1995).

1.2 Herpesviruses

Herpesviridae is a large family of linear, double-stranded DNA viruses with relatively large complex genomes with lengths ranging from 120 to 230 kbp. Herpesviruses contain 60 to 120 genes and the content of bases A and T ranges from 25% to 69% in each herpesviruses sequence (Roizman *et al.*, 1991).

The members of the *herpesviridae* family have been classified into three subfamilies (*alphaherpesvirinae*, *betaherpesvirinae* and *gammaherpesvirinae*) by the Herpesvirus Study Group of the International Committee on the Taxonomy of Viruses (ICTV). The classification is based on virus host range, genome organization and homology, and other biological properties (Roizman *et al.*, 1981). The

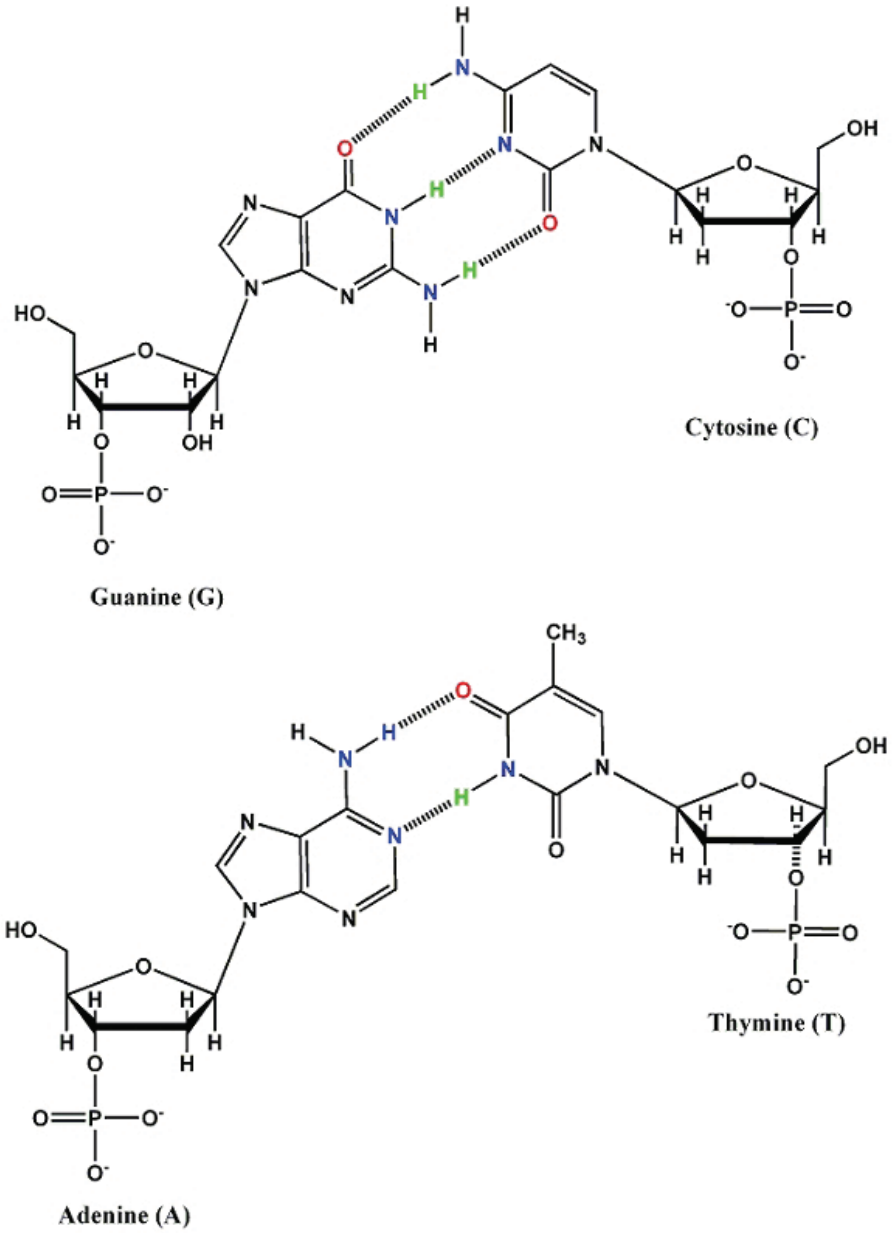


Figure 1.2: DNA base pairs.

Bottom, an AT base pair with two hydrogen bonds. Top, a GC base pair with three hydrogen bonds. The dashed lines denote non-covalent hydrogen bonds between the pairs.

α-herpesviruses grow rapidly in a wide range of tissues and efficiently destroy their host cell. The β-herpesviruses grow slowly and only in limited types of cells. Members of the γ-herpesviruses subfamily, grow slowly in, or immortalize, lymphoid cells of their natural host. Classifying viruses into subfamilies serves multiple purposes. The evolutionary relationship is often described by a classification scheme. Practically, it helps the laboratory worker predict the properties and identity of a new isolate (Roizman *et al.*, 1991).

Herpesviridae encompasses a large group of animal viruses with the distinguishing ability to establish latent, life-long infections. Members of this family have been observed in more than 80 different animal species (Frenkel *et al.*, 1990). Herpesvirus infections of human beings are a major public health issue, given their prevalence in the population. Examples of a variety of herpesviruses are the herpes simplex viruses (HSV-1 and HSV-2), which cause cold sores and genital tract infections in humans; Epstein-Barr virus (EBV) associated with infectious mononucleosis and with two-human cancer, Burkitt's lymphoma and nasopharyngeal carcinoma; human herpesvirus 8 (HHV8), linked to a variety of lymphomas which establishes latency in B lymphocytes and persists for the lifetime of the host; cytomegalovirus (CMV) which causes animal and human diseases, particularly in immunodeficient individuals; varicella-zostervirus (VZV), which induces chickenpox in children and shingles in adults; and Marek's herpesvirus, which causes malignant avian lymphoma (see p709 in Kornberg and Baker, 1992).

1.3 Replication Origins

DNA replication is a fundamental process in living cells that ensures transmission of genetic information between generations. The origin of replication is a particular sequence in a genome at which the replication process is initiated.

As Leung *et al.* (2005) indicated, the replication origin of Epstein-Barr Virus (EBV), which is a human herpesvirus, has been shown to associate with cellular proteins that regulate the initiation of DNA synthesis in human cells. EBV maintains its genome extra-chromosomally in infected cells (Sugden, 2002). Identifying the location of these replication origins is important in order to study the possible infection mechanisms of herpesviruses in human host cells. Knowledge of the precise locations of replication origins throughout herpesvirus genomes can provide a valuable resource to improve our understanding of DNA replication and lead to the development of antiviral agents by interfering with the infection process or by blocking viral DNA replication (Leung *et al.*, 2005).

1.4 Organization of the Thesis

The thesis is organized as follows:

In Chapter 2, we review the existing methods that are used to predict replication origins in bacterial, archaeal and eukaryotic genomes, especially in viruses.

We focus more on computational methods that use sequence features to predict replication origins in herpesviruses.

In Chapter 3, we focus on our approach based on the Generalized Additive Model (GAM) to predict replication origins. Before the models are built and fitted, we convert the sequence features into numerical data. We use the herpesvirus genomes with known replication origins to fit the model. We adopt the area under the Receiver Operating Curve (AUC) as the criterion for model selection. Then, further refinement of our GAM approach, which integrates multiple sequence features for more accurate prediction of replication origins in herpesviruses and other double-stranded DNA viral genomes, is discussed. Dominant sequence features are selected to build the Generalized Additive Models (GAMs). The stepwise model selection procedure is implemented in software R. We then apply the GAM approach to predict replication origins in Caudoviruses.

In Chapter 4, predictive results are presented and discussed. We select the best model from several reasonable models and employ a cross-validation method to assess the predictive performance of the model. We compare the predictive accuracies of different methods. Our approach exhibits respectable performance. In addition, we apply this GAM approach to other herpesviruses with unknown replication origins. The ultimately chosen and refined GAM approach performs much better than previous methods. It proves to be a valuable computational method of prediction for replication origins in Caudoviruses. We also applied

other approaches; however, our GAM approach outperformed them all.

In Chapter 5, we give the conclusions of this thesis and propose future steps including applying our approach to other organisms such as bacteria and yeasts, and exploring motifs around replication origins in order to predict the locations of the replication origins.

Chapter 2

Literature Review

2.1 Experimental Approaches to Identify Replication Origins

Because origins of replication in DNA of various organisms are considered important sites for regulating genome replication, much laboratory work has been done to search for replication origins (e.g., Stow, 1982; Brewer and Fangman, 1987; Zhu *et al.*, 1998; Hamzeh *et al.*, 1990; Wyrick *et al.*, 2001; Newlon and Theis, 2002).

As early as 1982, Stow developed an assay to locate an origin of DNA replication on the herpes simplex virus type 1 (HSV-1) genome, also known as human herpes virus 1 (HHV1). Stow transfected baby hamster kidney cells with circular plasmid molecules containing cloned copies of HSV-1 DNA fragments, and a su-

perinfection with wild-type HSV-1 provided helper functions. The presence of an HSV-1 origin of replication within a plasmid enabled amplification of the vector DNA sequences, which was detected by the incorporation of [^{32}P]orthophosphate. By screening various HSV-1 DNA fragments, Stow identified a 995-bp fragment containing all the cis-acting signals necessary to function as an origin of viral DNA replication. Brewer and Fangman (1987) developed an approach for physically mapping origins of replication by two-dimensional agarose gel electrophoresis, which was used to examine the replication of the native 2μm plasmid and a recombinant autonomous replication sequence (ARS) plasmid. The two-dimensional gel electrophoresis demonstrated that there was a single, specific origin of replication in each plasmid. In 2001, Wyrick *et al.* identified the positions of potential DNA replication origins across the *Saccharomyces cerevisiae* genome by determining the genome-wide locations of Origin Recognition Complex (ORC) and minichromosome maintenance (MCM) binding sites, because the binding of ORC and MCM proteins occurs at or very near the replication origin. Chromatin immunoprecipitation (ChIP) was used to identify the sites that ORC and MCM proteins bound. The ChIP-based method proposed 429 potential replication origins in the *S. cerevisiae* genome.

2.2 Computational Approaches to Predict Replication Origins

The increasing availability of sequence data of DNA data enables researchers to use computational approaches to predict likely locations of replication origins before applying experimentation. Many computational methods for predicting replication origins in bacterial, archaeal, eukaryotic and viral genomes were developed. They were reviewed in Chew *et al.* (2007). These algorithms are based on characteristic sequence features, rather than laboratory procedures, which can save significant money and time (Friedman *et al.*, 1995; Stow, 1982).

2.2.1 Prediction of Replication Origins in Bacterial, Archaeal and Eukaryotic Genomes

Mizraji and Ninio first introduced vectorial representations of sequences in 1985. The four bases, C, G, A and T, in a nucleic acid sequence were represented with vectors. The sequence was thus transformed into a trajectory in the plane. In 1996, Lobry adapted Mizraji and Ninio's vectorial representation (Mizraji and Ninio, 1985) of DNA sequences to locate replication origins in bacteria. Lobry (1996) replaced the four nucleic acid bases with vectors (see Figure 2.1). Then sequences could be represented in a planar trajectory. For example, the vectorial representation of the *Bacillus subtilis* sequence was given in Figure 2.2, where the

circle was used to indicate the location of a replication origin. Figure 2.2 showed that it was easy to detect a replication origin with this vectorial representation, since they were close to the reverse turn of the trajectory. With this graphical representation, the origins of replication in four bacterial species, *Escherichia coli, Bacillus subtilis, Haemophilus influenzae* and *Mycoplasma genitalium*, were well outlined.

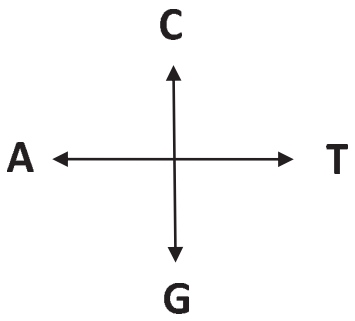


Figure 2.1: Each of the four nucleic acid bases is represented with a vector.(form Lobry, 1996)

Salzberg *et al.* (1998) employed the skewed oligomer method, a sequence-based method, to predict origins of replication in prokaryotic genomes, and in particular, in some bacterial and archaeal genomes. Short oligomers (seven-base and eight-base nucleic acid sequences), whose orientation is skewed around the origin, were found using this method. Here, "skewed orientation" means that short oligomers occur much more often on the leading strand in the direction of replication than it does on the lagging strand. They developed an algorithm for finding these skewed seven-base and eight-base sequences. They described

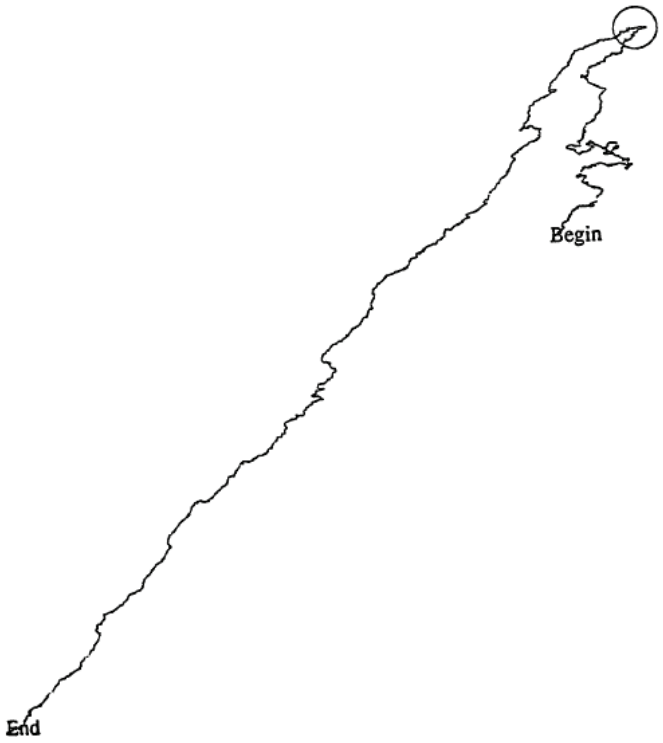


Figure 2.2: Vectorial representation of DNA sequences from Bacillus subtilis. The position of the origin of replication is outlined by a circle. (form Lobry, 1996)

a method for combining evidence from multiple skewed oligomers to locate the origins of replication accurately.

An approach based on base composition rather than specific sequences was used to predict replication origins in *Schizosaccharomyces pombe* by Segurado *et al.* in 2003. They used sliding windows of different sizes to determine base composition, and found that A+T content of windows close to replication origins were significantly higher.

Mackiewicz *et al.* (2004) applied three methods to identify the putative repli-

cation origins in 112 bacterial chromosomes, based on DNA asymmetry, DnaA box (a common motif) distribution and dnaA gene location. DNA asymmetry can be described in terms of the relationships between numbers of the four different nucleotides in DNA strands. They indicated that the most universal method of putative oriC identification in bacterial chromosomes is DNA asymmetry, although applying all three methods is necessary in some cases.

Breier *et al.* (2004) developed an algorithm called "Oriscan" to predict the exact location of replication origins in yeast genomes based on sequence information. Oriscan used 268 bp of sequence derived from a training set of 26 previously known replication origins. It was shown that accuracy was 94% in the top 100 predictions, but reliability decreased to 70% in the top 350 predictions.

For archaeal genomes, Zhang and Zhang (2005) applied the Z-curve method to identify several replication origins. The Z-curve is a three-dimensional curve that constitutes a unique representation of any given DNA sequence. Figure 2.3 shows an example of the three-dimensional Z-curve for the *Methanosarcina mazei* genome. The arrow indicates the position of the putative replication origin. Because the Z-curve contains all the information that the corresponding DNA sequence carries, we can study the DNA sequence by geometrical methods with the Z-curve. This method nicely complements widely used mathematical methods. In the same year, large-scale analysis of nucleotide compositional strand asymmetries were also developed (Brodie of Brodie *et al.*, 2005; Touchon *et al.*, 2005) for de-

tecting DNA replication origins in human chromosomes. More recently, Worning

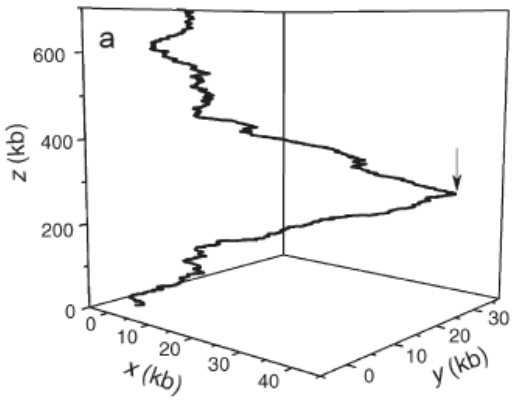


Figure 2.3: The three-dimensional Z-curve for the *Methanosarcina mazei* genome. (from Zhang and Zhang, 2005))

et al. (2006) developed a program that accurately located replication origins in prokaryotic chromosomes by measuring the differences between leading and lagging strands of all oligonucleotides up to 8 bp in length. This method was more sensitive than existing methods based on mononucleotide skews or the octamer skews.

Chew *et al.* (2005) pointed out that the method of predicting replication origins in one kind of genome may not necessarily work well on others, because sequence features around their replication origins in different organisms vary due to the differences in DNA replication mechanisms. Cells in the three major kingdoms, Bacteria, Archaea and Eukarya, use roughly similar strategies and mechanisms for genome replication; however, the mechanisms used are different from those of

viral genome replication (Stillman, 1996). Thus the computational methods for predicting the replication origins vary in viruses and other organisms. We will review the methods of predicting replication origins in viruses in the next section.

2.2.2 Prediction of Replication Origins in Viruses

Sequence Features to Predict Replication Origins

Many kinds of sequence features have been used to predict replication origins in herpesviruses. In this section, we first discuss the palindrome sequence feature (Chew *et al.*, 2005).

As defined by Chew *et al.* in 2005, a DNA palindrome is a segment of double-stranded DNA in which the nucleotide sequence of one strand reads exactly the same in reverse order with that of the complementary strand. A palindrome can also be defined as a word pattern of the form $a_1...a_L a'_L...a'_1$, where we denote a' to be the complement of base a, and the half-length of the palindrome is denoted as L. The letters a_L and a'_L are the left-center and the right-center of the palindrome, respectively. Figure 2.4 shows an example of a palindrome. The length of the palindrome in Figure 2.4 is 10 and its half-length L equals 5.

Early studies have reported that replication origins in herpesvirus genomes often lie around regions of the DNA sequence with an unusually high concentration of palindromes (Reisman *et al.*, 1985; Weller *et al.*, 1985; Masse *et al.*, 1992). The

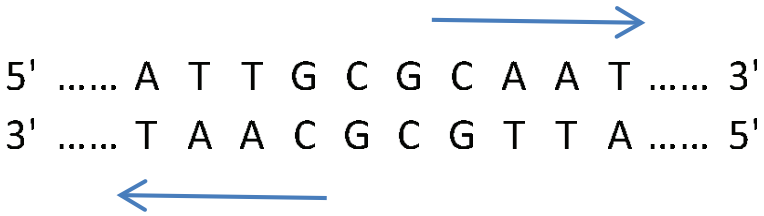


Figure 2.4: A palindrome of length 10.
The DNA sequence ATTGCGCAAT is a palindrome because its complement is TAACGCGTTA, which is equal to the original sequence in reverse complement.

general reason for this phenomenon is that initiation of DNA replication typically requires an assembly of enzymes to bind to the DNA, then locally unwind the helical structure and finally pull apart the two complementary strands (Chapter 1 in Kornberg and Baker, 1992; Bramhill, and Kornberg, 1998). The symmetry created by palindromes is advantageous for providing a suitable binding site for these DNA-binding proteins.

Another sequence feature that has been found in the vicinity of replication origins is the sequence of close direct repeats. Close direct repeats are short repeats separated by a spacer of several nucleotides (Rocha and Blanchard, 2002) (see Figure 2.5 for an illustration). The arrows under the DNA sequence indicate the sequence that is repeated. For instance, "bye-bye" is a Linguistical example of a direct repeat. The left part and right part of the close direct repeat are called the left stem and right stem, respectively. The starting positions of the left stem and right stem are called the left start and right start, respectively. We define the number of nucleotide bases in each stem as the stem length. For example, the

stem length of the close direct repeats in Figure 2.5 is 6.

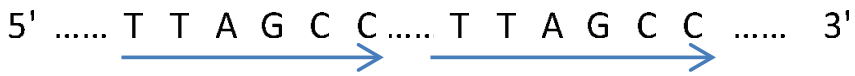


Figure 2.5: Close Direct Repeats.
The DNA sequence TTAGCC is repeated. The stem length is 6.

Empirical studies have suggested that close direct repeats are also found near replication origins in viral genomes (Hirsch *et al.*, 1977; Weller *et al.*, 1985; Reisman *et al.*, 1985; Dutch *et al.*, 1992; Masse *et al.*, 1992; Lehman and Boehmer, 1999). It was reported that in some herpesvirus genomes, the nucleotide sequences around replication origins are richer in A and T bases (Lin *et al.*, 2003). This is generally attributed to the fact that the two complementary DNA strands bond less strongly to each other due to the higher AT content around the origins (Segurado *et al.*, 2003; Sponer *et al.*, 1996). This facilitates the two complementary DNA strands to be pulled apart and initiate the replication process.

All these sequence features are relevant to replication origins in herpesviruses. Based on these observations, computational methods for replication origin prediction in herpesvirus genomes have been devised by using individual sequence feature palindromes and AT content (Chew *et al.*, 2005; Chew *et al.*, 2007). However, no one has yet predicted replication origins by the computational method using close direct repeats. We suggest that it is reasonable to introduce an approach based on close direct repeats to predict replication origins. Considering these sequence features jointly could also be compelling.

Existing Computational Methods to Predict Replication Origins in Viruses

So far, many computational methods to predict likely locations for replication origins in herpesviruses prior to experimentation have been developed. For example, Leung *et al.* (2005) suggested using scan statistics to locate statistically significant clusters of palindromes. Chew *et al.* (2005) further developed palindrome-based scoring schemes for quantifying palindrome concentrations to predict known replication origins in complete herpesvirus genomes and improve the sensitivity of the prediction. They introduced three scoring schemes for palindromes: *palindrome count score* (PCS), *palindrome length score* (PLS) and *base-pair weighted score* of order m (BWS_m). L was used to denote the benchmark of the minimum half length of a palindrome, where they only considered palindromes of at least $2L$ in length in their analysis. The *palindrome count score* (PCS) scheme, which was introduced by Leung *et al.* in 1994, gave a palindrome score of 1 when its length was at or above $2L$. A palindrome of length $2s \geq 2L$ was given a score s/L by the *palindrome length score* (PLS) scheme. Chew *et al.* (2005) highlighted the *base-pair weighted score* of order m (BWS_m) scheme, where m denotes the order of the Markov chain model of the DNA sequence. Under this scheme, the palindrome that had lower probabilities to occur by chance was given a higher score. Then, the score for a palindrome was the negative logarithm of the probability of a palindrome.

Using this scoring scheme, their method of predicting origins of replication

was to slide a window of fixed size over the sequence. The window scores for each window were calculated. A high window score reflected a high concentration of palindromes in the window, and vice versa. The windows with top scores were then selected as predicted locations of replication origins. However, the drawback to this method is that it does not make use of any information known about the replication origin locations in closely related members of the herpesvirus family. Since many members of the herpesvirus family were known to have a similar overall genome organization (Albrecht *et al.*, 1992), knowledge about the locations of replication origins in one herpesvirus should be relevant for predicting replication origins in other herpesviruses.

Another sequence feature known to be associated with replication origins is AT content. As reviewed by Chew *et al.* in 2007, Segurado *et al.* (2003) localized the positions of A+T rich "islands" in the Schizosaccharomyces pombe genome using sliding windows of different sizes. Genome-wide analysis enabled them to identify A+T rich "islands" regions, which predicted the localization of most origins of replication in the genome. Chew *et al.* (2005) also reported using the AT content feature on herpesviruses in order to identify replication origins. This method successfully identified several origins in some herpesviruses genomes (bohv4, ehv4 and hsv2) that were not predicted by any of the palindrome-based approaches using scoring schemes; namely, the *palindrome count score* (PCS), the *palindrome length score* (PLS), or the *base-pair weighted score* (BWS_m). This suggested that the sequence feature of AT content should be incorporated with

other predictive approaches to produce the optimal predictive results. Motivated by this, Chew *et al.* (2007) found a window free approach to better quantify the AT content variation in genome sequences. This score-based excursion approach was used to identify genome regions with high AT concentrations, called high-scoring segments. These segments were predicted as potential replication origin sites in herpesviruses. This AT excursion approach successfully identified several replication origins not previously predicted by the palindrome-based method. Therefore, the AT excursion approach was a valuable approach to predict replication origins in herpesviruses. However, it was observed that quite a number of regions predicted as potential replication origin sites by AT excursions were not close to replication origins. This meant that the positive predictive value of the AT excursion approach was low although the corresponding sensitivity was high. Thus, developing methods which can improve the positive predictive value could be very beneficial.

Besides palindromes and AT content, the sequence feature of close direct repeats has also been found to be concentrated around the replication origins in herpesviruses (Stow, 1982). However, this sequence feature has never been used to predict the locations of replication origins in herpesviruses. As such, an approach based on close direct repeats needs to be explored. All of the current methods have achieved success to some extent in predicting replication origins in herpesviruses by using an individual sequence feature. Therefore, it is reasonable to expect that the predictive accuracy can be improved by appropriately integrat-

ing sequence features, palindromes, close direct repeats and AT content.

Chapter 3

Methodology

From the above review, we can see that the replication origins in herpesviruses can be predicted with some degree of success by computational approaches that separately use sequence features, palindromes and AT content.

The aim of this research was to develop a statistical model that integrates multiple DNA sequence features for more accurate prediction of replication origins in herpesviruses, and also to extend this model to other similar viral families. We adopted the area under the Receiver Operating Curve (ROC) as the criterion for model selection (Pepe, 2003). The area under the ROC curve (AUC) is a numerical measure of a model's discrimination performance. We compared AUC scores of several models with different combinations of explanatory variables (i.e., sequence features) in order to select the best model.

We hope our model can improve the accuracy of predicting locations of replication origins. Our approach may be a promising computational tool for identifying replication origins in herpesvirus genomes. Also, the methodology we use may be applicable to other viral families. Furthermore, the identification of origins of replication was a labor-intensive task (Friedman *et al.*, 1995; Stow, 1982; Brewer and Fangman, 1987; Wyrick *et al.*, 2001). Therefore, our computational methods using relevant DNA sequence features to predict likely positions of replication origins before applying experimental methods should be highly valuable. The computational predictive approaches could help design finely-tuned experiments that efficiently locate positions of replication origins with fewer resources and less labor and in shorter time.

In this chapter, we propose a computational method to predict the locations of replication origins in herpesviruses. Here we give an overview of our method. After locating palindromes and close direct repeats in herpesvirus genome sequences, we convert the sequence features to numerical data. The AT content is also quantified. Then we model the data using Generalized Additive Models (GAMs), which are fitted by regressing the quantified sequence features and known replication origins in herpesvirus genomes. By using the AUC criterion, we select the best model which is used to predict replication origins in herpesviruses with unknown replication origin locations. Furthermore, we refine the GAM approach with more sequence features which may relate to replication origins. We select variables by a forward stepwise GAM approach. After finding the most effective way to

predict the likely locations of replication origins in herpesviruses, we apply the GAM approach to caudoviruses, which share several similar characteristics with herpesviruses.

3.1 Converting Sequence Features into Numerical Data

We consider the sequence features of palindromes, close direct repeats and AT content in herpesvirus genomes and local maxima of the scores of these sequence features. The following subsections discuss how to quantify these sequence features.

3.1.1 Data Set to Be Analyzed

The data set comprises all complete genome sequences of the herpesvirus family downloaded in June 2007 from GenBank at the NCBI web-site (http://www.ncbi.nlm.nih.gov/). The analysis encompasses 47 herpesviruses in all, which are presented in Table 3.1. Their sequence length and the percentage of nucleotide bases A and T are listed in Table 3.1 as well as their abbreviation and accession number. Chew *et al.* (2007) reported fourty-three replication origins in herpesviruses with known locations after extensive compilation from literature review and Genbank. This forms the basis of our data set.

Table 3.1: The list of herpesviruses to be analyzed.

Virus	Length	AT percentage	Abbreviation	Accession
Alcelaphine herpesvirus 1	130608	54	alhv1	NC_002531
Ateline herpesvirus 3	108409	64	athv3	NC_001987
Bovine herpesvirus 1	135301	28	bohv1	NC_001847
Bovine herpesvirus 4	108873	59	bohv4	NC_002665
Bovine herpesvirus 5	137821	26	bohv5	NC_005261
Callitrichine herpesvirus 3	149696	51	calhv3	NC_004367
Cercopithecine herpesvirus 1	156789	26	cehv1	NC_004812
Cercopithecine herpesvirus 2	150715	25	cehv2	NC_006560
Cercopithecine herpesvirus 8	221454	51	cehv8	NC_006150
Cercopithecine herpesvirus 9	124784	60	cehv9	NC_002686
Cercopithecine herpesvirus 15	171096	39	cehv15	NC_006146
Cercopithecine herpesvirus 16	156487	24	cehv16	NC_007653
Cercopithecine herpesvirus 17	133719	48	cehv17	NC_003401
Chimpanzee cytomegalovirus	241087	39	pohv4	NC_003521
Equid herpesvirus 1	150224	44	ehv1	NC_001491
Equid herpesvirus 2	184427	43	ehv2	NC_001650
Equid herpesvirus 4	145597	50	ehv4	NC_001844
Gallid herpesvirus 1	148687	52	gahv1	NC_006623
Gallid herpesvirus 2	177874	56	gahv2	NC_002229
Gallid herpesvirus 3	164270	47	gahv3	NC_002577
Human herpesvirus 1	152261	32	hhv1	NC_001806
Human herpesvirus 2	154746	30	hhv2	NC_001798
Human herpesvirus 3	124884	54	hhv3	NC_001348
Human herpesvirus 4 type 1	171823	41	ebv1	NC_007605

Virus	Length	AT percentage	Abbreviation	Accession
Human herpesvirus 4 type 2	172764	41	ebv2	NC_009334
Human herpesvirus 5 strain AD169	230287	43	hcmv	NC_001347
Human herpesvirus 5 strain Merlin	235645	43	hcmv-m	NC_006273
Human herpesvirus 6A	159321	58	hhv6a	NC_001664
Human herpesvirus 6B	162114	58	hhv6b	NC_000898
Human herpesvirus 7	153080	64	hhv7	NC_001716
Human herpesvirus 8 type p	137969	47	hhv8p	NC_009333
Human herpesvirus 8 type M	137508	47	hhv8m	NC_003409
Ictalurid herpesvirus 1	134226	44	ichv1	NC_001493
Koi herpesvirus	295146	41	khv	NC_009127
Macaca fuscata rhadinovirus	131217	49	mfrv	NC_007016
Meleagrid herpesvirus 1	159160	53	mehv1	NC_002641
Murid herpesvirus 1	230278	42	mcmv	NC_004065
Murid herpesvirus 2	230138	39	rcmv	NC_002512
Murid herpesvirus 4	119451	53	muhv4	NC_001826
Ostreid herpesvirus 1	207439	62	oshv1	NC_005881
Ovine herpesvirus 2	135135	48	ohv2	NC_007646
Psittacid herpesvirus 1	163025	40	pshv1	NC_005264
Ranid herpesvirus 1	220859	46	rahv1	NC_008211
Ranid herpesvirus 2	231801	48	rahv2	NC_008210
Saimiriine herpesvirus 2	112930	66	sahv2	NC_001350
Suid herpesvirus 1	143461	27	shv1	NC_006151
Tupaia herpesvirus	195859	34	thv	NC_002794

3.1.2 Converting Palindromes to Numerical Data

As argued by Chew *et al.* (2005), very short palindromes occur frequently by chance, so we need to fix a minimum half length of a palindrome L. Palindromes of length less than $2L$ will not be considered in the analysis. Based on the benchmark of the well-studied HCMV virus, Leung *et al.* (2005) proposed a procedure to choose the parameter L. Leung *et al.* (2005) chose $L = 5$ for most of the viruses, so we fix $L = 5$ here. We used the software EMBOSS [European Molecular Biology Open Software Suite] (Rice *et al.*, 2000) to locate palindromes in the genome. In order to extract the useful information of each palindrome, such as its length and position, the software was employed based on the minimal palindrome length $2L$. We will assign a score to each of these palindromes. Chew *et al.* (2005) found that the scoring scheme *base-pair weighted score* of order m (BWS_m) worked better than the *palindrome length score* (PLS) scheme in terms of predictive accuracy. But the BWS_m scoring scheme assigns scores to palindromes based on the dependence of the adjacent nucleotide bases, while our proposed method assumes two non-overlapping segments of genome sequences are independent. Thus we will adopt the PLS scoring scheme to assign scores for palindromes, then a palindrome of length $2h \geq 2L$ will be scored as h/L. For instance, if L is chosen to be 5, a palindrome of length 28 will be given a score of $\frac{28}{2}/5 = 2.8$.

3.1.3 Converting Close Direct Repeats to Numerical Data

We use the software "REPuter" (Kurtz *et al.*, 1999; Kurtz *et al.*, 2001) to locate close direct repeats. REPuter can find all repeats above a given level of significance in a complete genome. REPuter assesses the significance of each repeat by its E-value (i.e., "the number of repeats of the same length or longer and with the same number of errors or fewer that one would expect to find in a random DNA of the same length" (defined by Kurtz *et al.*, 2000)). The maximum computed repeats, the minimal repeat size and the error distance should be chosen before running the program. The range of maximum computed repeats is from 1 to 5000, and the minimal repeat size can be chosen from 8 to 200 using the software REPuter. Because we want to extract as many close direct repeats as possible from the genome sequences, we chose 5000 as the maximum computed repeats and 8 as the minimum repeat size. The maximum allowed error distance was chosen as 0, since only exact repeats were considered in this study. The REPuter results page gives an overview of the number, length and location of repeats in the uploaded sequence. The output is sorted by E-values.

In this study, the maximum allowable distance between the starting positions of the close direct repeats depends on the length of the genome sequence. The details will be described in a later section. Each pair of close direct repeats will be converted to a numerical score according to a scoring scheme. We introduce a scoring scheme, *repeats length score* (RLS), in which a pair of close direct repeats

of stem length S is given a score S. For example, a close direct repeats of stem length 18 will receive a score of 18.

3.1.4 Converting AT Content to Numerical Data

A replication origin often lies around an AT-rich region. We use the percentage of nucleotide bases A and T in the sequence segment as the score for AT content. For example, the score of the sequence segment AATGCTTATA is 80.

3.1.5 Computing the Window Scores

The entire genomic sequence is partitioned into non-overlapping windows of equal size. For each window, the palindrome score of the window is the total of each score of palindromes within this window. If the left-center of a palindrome is in this window, the palindrome will be considered in the window. Likewise, a close direct repeats score of a window is defined as follows. For a pair of close direct repeats, if the starting positions are in the same window, we will score this pair. Otherwise, we ignore it. The window score for AT content is the percentage of nucleotide bases A and T in this window.

Following the choice of window length by Chew *et al.* (2005), the window length w was chosen as 0.5% of the genome length, rounded down to the nearest

hundred bases for convenience. The length of the last window is usually shorter than w due to the way the windows are constructed.

3.1.6 Local Maxima

After determining the window scores of the three sequence features, we need to consider another variable, the local maxima of window scores. If a window score is higher than or equal to scores of its m neighboring windows both to the left and to the right, then we consider this window as a local maximum. Here, m is chosen to be 4. The reason we need this variable is that the windows with high window scores also identified as local maxima would be considered as potential locations of replication origins. If a window score is relatively high but is not a local maximum, the window is less likely to be around a replication origin compared to a window that is a local maximum with slightly less window score.

For example, in Figure 3.1, the AT scores of the windows in suhv1 sequence are plotted against the center position of the windows. Three circles in the graph indicate the locations of three known replication origins. Although the window marked by the cross has higher score than the window marked by the red circle, the latter one actually contains a replication origin. This is due to the fact that the window denoted by the red circle is a window with relatively higher score than most of the other windows, and is a local maximum. This location is more likely to be a replication origin. Therefore, we can identify the local maxima

among the window scores for such sequence features as palindromes, close direct repeats and AT content in herpesvirus genome sequences. LM is used to denote the variable local maximum. If a window is a local maximum, $LM = 1$; otherwise, $LM = 0$. LM_P, LM_R and LM_{AT} denote the local maxima of palindromes, close direct repeats and AT content–based window scores, respectively.

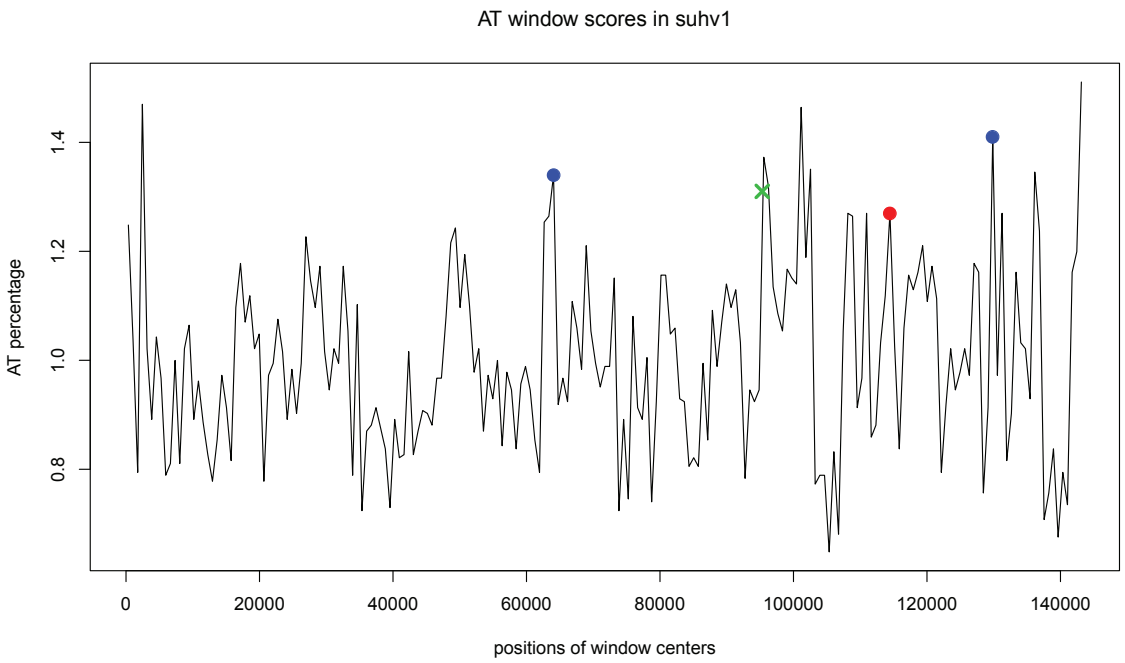


Figure 3.1: Local maximum of AT window scores in suhv1 genome sequence. The values on the x-axis of three circles (both red and blue ones) in the figure indicate the centers of windows that contain known replication origins. The green cross indicates a window that does not contain a replication origin.

3.2 Comparison of Approaches Based on Single Sequence Feature

To compare methods of replication origin prediction based on close direct repeats, palindromes and AT content individually, we examine the numbers of replication origins captured by top 10 ranked windows using these approaches, which are listed in Table 3.2. It can be seen that the close direct repeats–based method performs better to some extent. The method using close direct repeats identifies some replication origins which are not predicted by using palindromes or AT content.

We used a Venn diagram to display the numbers of replication origins in herpesviruses correctly predicted by any one or more methods out of close direct repeats, AT content and palindromes–based methods (see Figure 3.2).

Table 3.2: No. of replication origins captured by close direct repeats, palindromes, and AT content methods with top 10 windows.

Top	Repeats	Palindrome	AT content
1	7	5	6
2	14	15	15
3	19	23	17
4	23	23	22
5	24	23	22
6	25	23	24
7	27	23	25
8	31	24	27
9	32	24	29
10	34	26	29

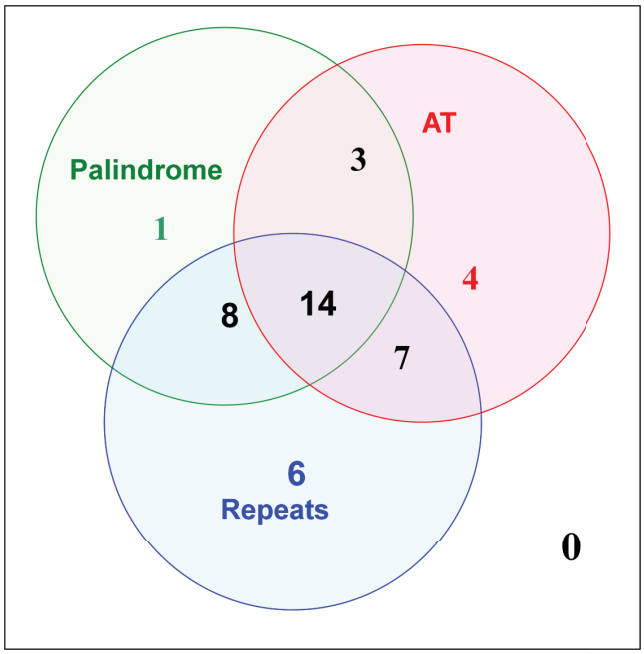


Figure 3.2: Numbers of replication origins correctly predicted based on palindromes, repeats and AT content approaches by top 10 ranked windows. Fourteen replication origins are predicted by all the three methods and all of the 43 known origins in the herpesviruses are predicted by at least one of these methods.

From this figure, we can see that 14 replication origins are predicted by all three methods and all of the 43 known origins in the herpesviruses are predicted by at least one of these methods. Some of the replication origins are captured by only one or two of these approaches. This suggests that individual close direct repeats, palindromes and AT content based methods complement each other in predicting replication origins very well. A natural question would be how to combine window scores of various sequence features to give more accurate predictions. An approach combining multiple sequence features will be developed later in this thesis.

3.3 Pre-processing of Data Set

Before setting up a model to integrate several sequence features, we combine all the window scores of the 20 herpesvirus genome sequences, whose locations of replication origins are known. Then we plot the histograms of window scores of repeats, AT content and palindromes. The plots are given in Figure 3.3.

The histogram for window scores of close direct repeats is extremely skewed, since the window scores are zero for most of the windows. In order to look at the distribution of positive window scores, in Figure 3.4, we plot the histogram for window scores of close direct repeats whose window scores are positive. Zooming in the histogram of window scores which are higher than 1000 (see the right plot in Figure 3.4), we find that few windows have extremely high window scores.

Likewise, we plot the histograms for window scores of palindromes, which are positive (see the left plot in Figure 3.5) and larger than 30 (see the right plot in Figure 3.5).

We therefore consider using the logarithmic transformation to transform the original window scores of repeats, because the logarithm function tends to squeeze together the larger values in the data set and stretches out the smaller values. This squeezing and stretching can correct our skewed data. Log transformation is valid only for positive numbers, so we transform each window score of repeats (R) to $\log(R + 1)$, and then analyze the resulting data. Figure 3.6 (right) shows the distribution of the transformed window scores. We can see that it spreads more than before.

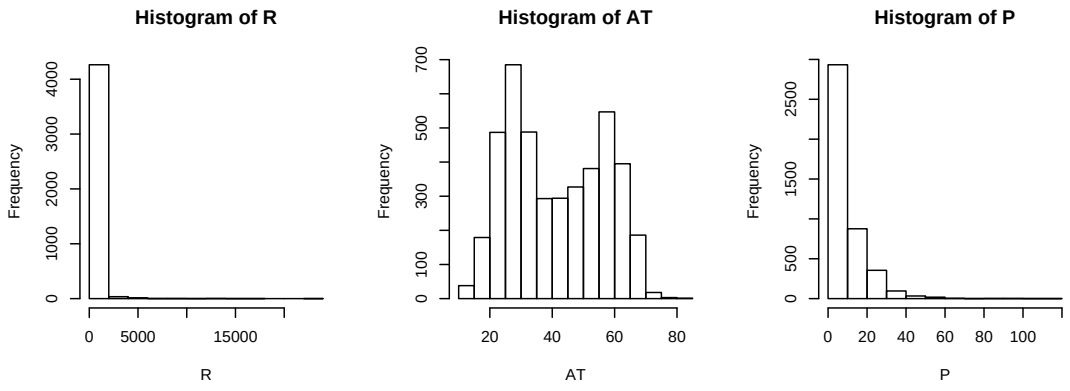


Figure 3.3: Histograms of window scores of repeats, AT content and palindromes.

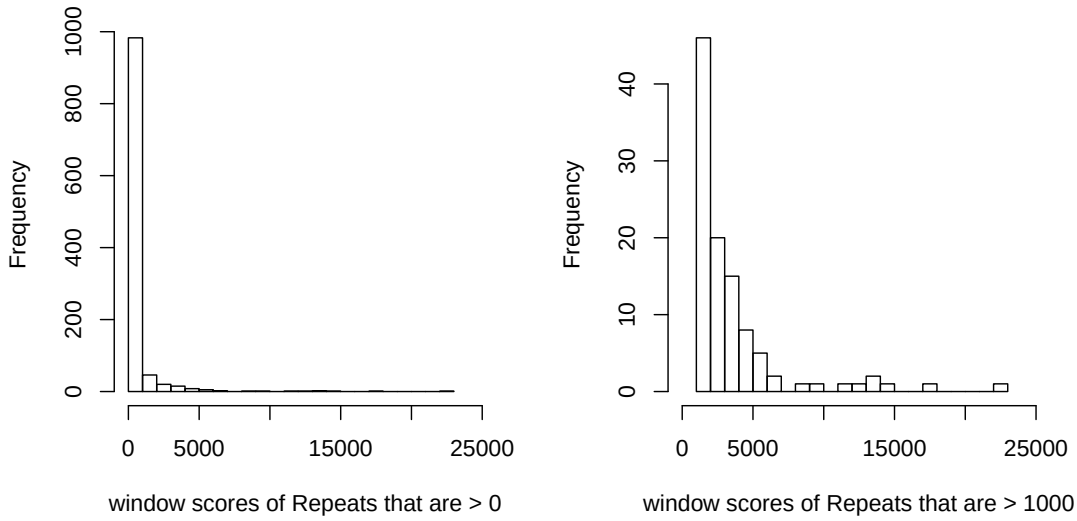


Figure 3.4: Histograms of window scores of close direct repeats whose window scores are positive and above 1000.

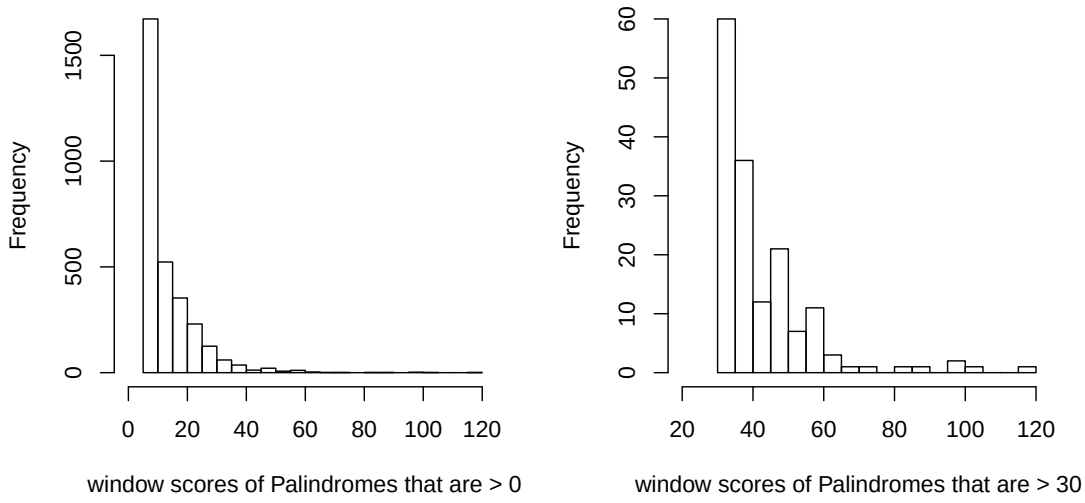


Figure 3.5: Histograms of window scores of Palindromes whose window scores are positive and above 30.

After the transformation, we pool the window scores in one data set and use them to build up and fit generalized additive models. Because the ranges and the distribution of window scores in various herpesviruses are different, simply pooling all window scores together is not reasonable. The summary statistics of window scores of repeats (log transformed), AT content in percentages and palindromes in herpesviruses with known replication origins are listed in Tables 3.3, 3.4 and 3.5. These tables show the minimum, the first quartile, the median, the mean, the third quartile and the maximum of window scores for each herpesvirus. Window scores vary among different members of the herpesvirus family. Take the window scores of palindromes for example. The window scores of palindromes range from 0 to 116 in cehv1, while in gahv1, the highest window score is only 17. Similarly,

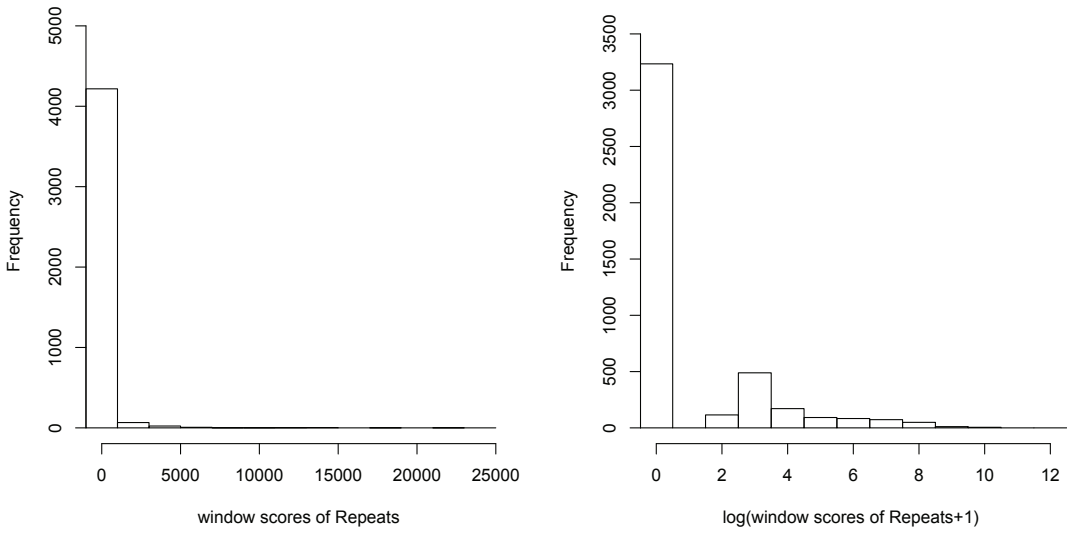


Figure 3.6: The log transform of scores of close direct repeats.

the range of AT content window scores varies from virus to virus. We notice that the maximum window score of cehv2 is 34, which is even less than the minimum AT window score of gahv1. If we combine the two sets of window scores to fit a statistical model without standardization, the model may mistakenly ignore the effects of AT content variation in cehv2 and put more weight on gahv1, since the largest window score of cehv2 is much lower than that of gahv1. So we should not simply combine them to fit the model. In order to show that the standardization is needed, we compare predictive results of the models using standardized and non-standardized data, shown in Chapter 4.

Hence we standardize the window scores for each herpesvirus before pooling them. The method of standardization is dividing the difference between an original window score and the mean of window scores in the same herpesvirus sequence by the standard deviation. For example, if the window score of repeats in the ith window of a particular herpesvirus sequence V is R_i, then the standardized window score will be

$$\frac{R_i - \text{the mean of window scores in the sequence } V}{\text{the standard deviation of window scores in the sequence } V}. \tag{3.1}$$

Finally, our data set is composed of all these standardized window scores.

Table 3.3: Summary of window scores of repeats in herpesviruses ($\log(R+1)$).

Name	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
bohv1	0	0	0	1.2	2.6	7.6
bohv4	0	0	0	0.5	0	8.0
bohv5	0	0	0	1.1	2.7	7.9
cehv1	0	0	0	1.4	2.8	9.6
cehv2	0	0	0	1.5	2.9	8.2
cehv9	0	0	0	0.7	0	6.8
cehv16	0	0	0	1.5	2.8	8.6
ebv	0	0	0	1.1	2.6	9.5
ehv1	0	0	0	1.0	2.4	9.0
ehv4	0	0	0	0.8	0	8.8
gahv1	0	0	0	0.6	0	5.5
hcmv	0	0	0	0.9	2.6	5.5
hhv1	0	0	0	1.2	2.6	8.0
hhv2	0	0	0	1.2	2.6	8.6
hhv3	0	0	0	0.5	0	7.3
hhv6a	0	0	0	0.6	0	10
hhv6b	0	0	0	0.7	0	9.8
hhv7	0	0	0	0.7	0	9.2
rcmv	0	0	0	1.2	2.6	7.0
suhv1	0	0	0	1.9	3.2	9.3

Table 3.4: Summary of window scores of AT content in percentages in herpesviruses.

Name	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
bohv1	11	25	28	28	31	44
bohv4	27	56	59	59	62	71
bohv5	12	22	25	25	28	43
cehv1	14	22	26	26	29	39
cehv2	12	21	24	24	27	34
cehv9	30	59	61	60	64	69
cehv16	13	21	24	24	27	37
ebv	15	37	42	40	46	60
ehv1	23	41	44	43	47	56
ehv4	27	48	51	50	54	66
gahv1	35	49	53	52	56	60
hcmv	24	39	42	43	46	64
hhv1	15	29	33	32	35	43
hhv2	14	26	30	30	34	40
hhv3	21	52	55	54	58	64
hhv6a	31	57	59	58	61	72
hhv6b	30	56	59	57	61	73
hhv7	39	63	66	64	68	80
rcmv	20	30	34	39	48	68
suhv1	17	23	26	26	29	40

Table 3.5: Summary of window scores of palindromes in herpesviruses.

Name	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
bohv1	0	5	12	15	21	57
bohv4	0	0	0	3	5	23
bohv5	0	6	16	17	23	72
cehv1	0	5	11	14	20	116
cehv2	0	6	15	16	22	97
cehv9	0	0	5	5	10	22
cehv16	0	5	12	14	21	51
ebv	0	0	5	7	10	87
ehv1	0	0	5	5	7	33
ehv4	0	0	3	4	6	23
gahv1	0	0	5	4	6	17
hcmv	0	0	5	8	10	57
hhv1	0	5	6	9	13	85
hhv2	0	5	7	9	12	48
hhv3	0	0	5	5	6	30
hhv6a	0	0	5	5	7	26
hhv6b	0	0	5	5	10	31
hhv7	0	0	5	6	10	32
rcmv	0	5	10	11	16	41
suhv1	0	6	15	16	22	62

3.4 Generalized Additive Models

The underlying principle of our method is that the regions near replication origins contain some characteristic features, such as palindromes, close direct repeats, and AT content, which make them distinct from regions far from replication origins. We want to find the dependence of the locations of replication origins on these types of sequence features. The predictive models, utilizing these features to discriminate regions that contain replication origins from others, are built and then used to predict the likely locations of replication origins in herpesvirus DNA sequences with unknown replication regions.

Generalized Additive Model (GAM), a non-parametric regression technique not restricted by linear relationships, is flexible regarding the statistical distribution of the data (Swartzman *et al.*, 1995). Because appropriate functional forms of covariates are unknown, GAMs are applied to our data set to investigate relationship between locations of replication origins in herpesviruses and several related sequence features. The GAMs (Hastie and Tibshirani, 1990) enable us to combine the information of spatial abundance of palindromes, close direct repeats, AT abundance, local maxima of these sequence features and their interactions in a meaningful way for better prediction. Moreover, data obtained in all windows of 20 herpesviruses are used to fit the models, enabling the information of the locations of replication origins in one virus to be available for predicting the likely locations of replication origins in the other viruses. The window which contains a

replication origin and the neighboring 4 windows both to the left and to the right of it will be considered as windows close to a replication origin. Let Y_i be the ith binary response variable defined as follows:

$$Y_i = \begin{cases} 1, & \text{if the } i\text{th window is close to a replication origin,} \\ 0, & \text{if the } i\text{th window is not close to a replication origin.} \end{cases} \tag{3.2}$$

Associated with this response are the possible explanatory variables R_i, AT_i, P_i, LM_{Ri}, LM_{ATi}, LM_{Pi}, where R_i, AT_i, P_i denote the ith window scores of close direct repeats, AT content and palindromes, and LM_{Ri}, LM_{ATi}, LM_{Pi} denote the local maxima of them respectively. p_i is defined by $p_i = P(Y_i = 1|\mathbf{X_i})$, where $\mathbf{X_i}$ is the collection of explanatory variables in the ith window. Our model will be in this form

$$\log(\frac{p_i}{1-p_i}) = m(\mathbf{X_i}) \tag{3.3}$$

where $m : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ is an unknown smooth function and $\mathcal{X}$ is bounded. The right hand side of the model equation is usually called a *risk score*. For a linear logistic regression model, $m(\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a finite dimensional unknown parameter. For a nonparametric additive logistic regression model, $m(\mathbf{X}) = m(X_1, ...X_d) = \sum_{j=1}^{d} m_j(X_j)$, where each m_j is a univariate unknown smooth function. To achieve a sensible interpretation, we usually require monotonicity of each m_j in the model. After fitting this model, we can find how the locations of replication origins are dependent on the windows scores of repeats, AT content, palindromes and their local maxima.

3.5 Software for Implementing Generalized Additive Models

We used the statistical software R (Ihaka and Gentleman, 1996) to implement the models. Functions are available in the R language, and so we used the `gam` package in R to fit the generalized additive models which we describe here.

The syntax of the `gam` is

```
gam(formula, family, ...)
```

The `formula` expression has the form `response` $\sim$ `predictors`. Irrespective of the error model, `response` describes what variable is to be used for the response, and `predictors` describe symbolically the composition of the additive model. For smoothing splines, `s` is used to indicate nonparametric smoothing terms. We describe it with an example.

```
y ~ s(repeats)+s(palindrome)+AT
```

The function `s` above indicates that smoothing splines are used to fit both `repeats` and `palindromes`, while `AT` is fit linearly.

The `family` argument is a description of the error distribution and link function to be used in the model. For example, the call

```
fit=gam(y ~ s(repeats)+AT, family=binomial)
```

assumes binomial data, uses the logit link and, by default, the binomial variance function. It will fit a smooth term in `repeats` and a linear term in `AT`. The result of the fit will be stored in the object `fit`.

Other arguments (...) to `gam` include `weight` for giving prior weights, `subset` for specifying a subset of the data for fitting the model, and `na.action`, a function for dealing with missing data.

The object returned by `gam` is an R list object, with elements such as `coefficients`, `deviance`, `fitted.values` and others that describe the fitted model. Other functions exist for summarizing and displaying the fit. Suppose we save the output of the `gam` procedure into an object called `fit`, then `summary(fit)` will give a detailed summary, and `plot(fit)` produces plots of the terms comprising the fit.

3.6 ROC and AUC

3.6.1 The Receiver Operating Characteristic (ROC) Curve

A useful statistical tool, the Receiver Operating Characteristic (ROC) curve, is often used to evaluate the accuracy of continuous diagnostic tests by (Pepe, 2003). We employ this tool to examine the predictive accuracy of our models, where the risk score $m(\mathbf{X})$ of each model can be viewed as a diagnostic test.

In order to comply with the notation introduced by Pepe, windows close to a replication origin are viewed as being in diseased status, while windows far from a replication origin are viewed as being in non-diseased status. We use the binary variable, D, to denote true diseased status:

$$D = \begin{cases} 1, & \text{for disease;} \\ 0, & \text{for non-disease.} \end{cases}$$

The risk score $m(\mathbf{X})$ is considered as the result of the test. By convention, larger values of $m(\mathbf{X})$ are more indicative of disease. Using a threshold c, we define a binary test from the continuous test result $m(\mathbf{X})$ as

$$\text{positive for disease if } \; m(\mathbf{X}) \geq c,$$

$$\text{negative for disease if } \; m(\mathbf{X}) < c.$$

Subscripts D and $\bar{D}$ are used to index quantities pertinent to diseased and non-diseased respectively. Thus, for example, m_D denotes the test result for a diseased subject.

The result of the test can be classified as a true positive, a true negative, a false positive or a false negative, as shown in Table 3.6. As the names suggest, a true positive occurs when a diseased subject is correctly tested positive, and a false negative when a diseased subject is incorrectly tested negative. Similarly, a true negative or a false positive occurs when a non-diseased subject has a negative or a positive result, respectively.

Table 3.6: Classification of test results by disease status.

	$D = 0$	$D = 1$
$m(\mathbf{X}) < c$	True negative	False negative
$m(\mathbf{X}) \geq c$	False positive	True positive

A test has two types of errors: false positive and false negative. An ideal test should have no false positives and no false negatives. We define the true and false positive fractions at threshold c, TPF(c) and FPF(c), as follows:

$$\text{TPF}(c) = P[m(\mathbf{X}) \geq c | D = 1], \tag{3.4}$$

$$\text{FPF}(c) = P[m(\mathbf{X}) \geq c | D = 0]. \tag{3.5}$$

The ROC curve is the entire set of possible true and false positive fractions attainable by dichotomizing $m(\mathbf{X})$ with different thresholds. That is, the ROC curve is

$$\text{ROC}(\cdot) = \{(\text{FPF}(c), \text{TPF}(c)), c \in (-\infty, \infty)\}. \tag{3.6}$$

Observe that, as the threshold c increases, both FPF(c) and TPF(c) decrease. On one extreme, assuming $c = \infty$, we have $\lim_{c\to\infty}\text{TPF}(c) = 0$ and $\lim_{c\to\infty}\text{FPF}(c) = 0$. On the other extreme, assuming $c = -\infty$, we have $\lim_{c\to-\infty}\text{TPF}(c) = 1$ and $\lim_{c\to-\infty}\text{FPF}(c) = 1$. Thus, the ROC curve is a monotonically increasing function in the positive quadrant. This is illustrated in Figure 3.7. We also write the ROC curve as

$$\text{ROC}(\cdot) = \{(t, \text{ROC}(t)), t \in (0, 1)\}. \tag{3.7}$$

where the ROC function maps t to TPF(c), and c is the threshold corresponding

to $\mathrm{FPF}(c) = t$.

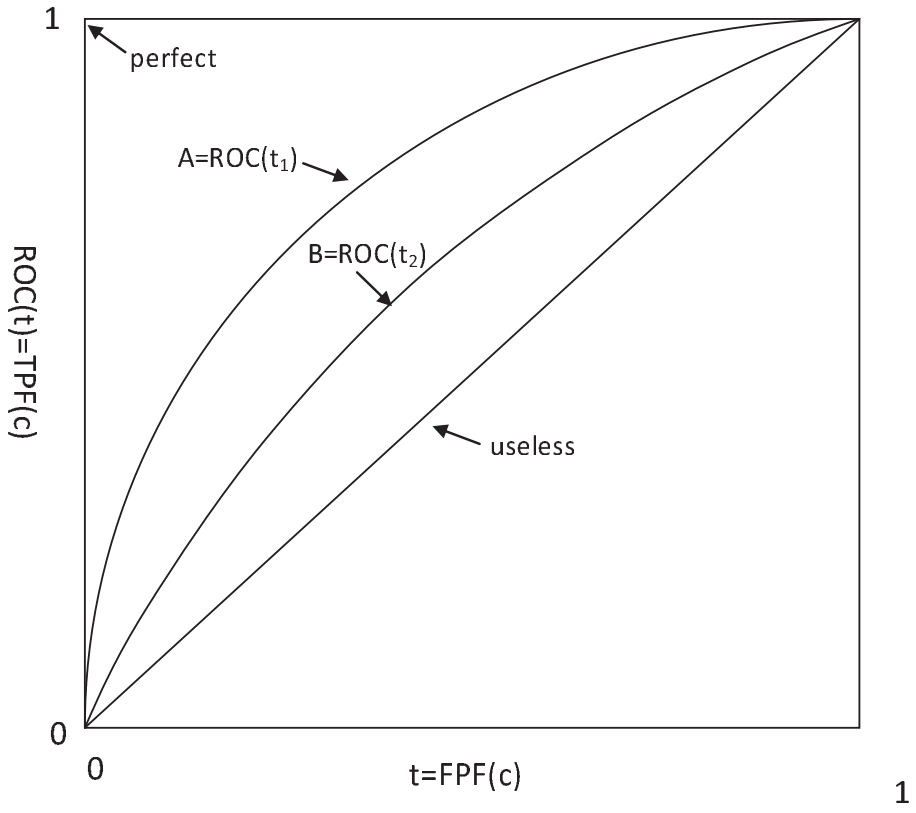


Figure 3.7: ROC curves.
Curves A and B are ROC curves for tests A and B, where the test A is uniformly better than the test B. Each point on an ROC curve is generated by a different decision threshold. ROC curves for the useless and perfect tests for comparison are also shown.

Generally speaking, the ROC is evaluated by means of a plot of a test's true positive fraction (plotted on the y-axis) versus its false positive fraction (plotted on the x-axis) using a continuously varying decision threshold. In practice, the plot is produced by classifying each window as positive or negative according to the outcome (i.e., whether the window is close to a replication origin or not).

An uninformative test is one such $m(\mathbf{X})$ that is unrelated to disease sta-

tus. That is, the probability distributions for $m(\mathbf{X})$ are the same in the diseased and non-diseased populations, and therefore for any threshold c, we have TPF(c) =FPF(c). The ROC curve for an uninformative test is therefore ROC(c) = t, which is a line with unit slope.

A perfect test on the other hand completely separates diseased and non-diseased subjects. That is, for some threshold c, we have TPF(c) = 1 and FPF(c) = 0. Its ROC curve is along the left and upper borders of the positive unit quadrant. Most tests have ROC curves that lie between those of the perfect and useless tests. Better tests have ROC curves closer to the upper left corner. See Figure 3.7, where test A, the better of the two tests, is such that at any false positive fraction its corresponding true positive fraction is higher than that of test B. Similarly, if we choose thresholds c_A and c_B for which $\text{TPF}_A(c_A) = \text{TPF}_B(c_B)$, the corresponding false positive fractions are ordered in favor of test A, that is $\text{FPF}_A(c_A) < \text{FPF}_B(c_B)$.

3.6.2 The Area Under the ROC Curve (AUC)

We adopt the area under the Receiver Operating Characteristic (ROC) Curve as the criterion for the model selection. The area under the ROC curve (AUC) is a numerical measure of a model's discrimination performance. AUC is defined as

$$\text{AUC} = \int_0^1 \text{ROC}(t)\text{d}t. \tag{3.8}$$

A high AUC value indicates favorable classification and prediction ability of the test (Pepe, 2003). An almost perfect test has the value AUC ≈ 1.0. Conversely, an uninformative test, with ROC(t) = t, has AUC = 0.5. Most tests result in values that fall in between (Pepe, 2003). Clearly, if two tests are ordered with test A uniformly better than test B in the sense that

$$\mathrm{ROC}_A(t) \geq \mathrm{ROC}_B(t) \quad \forall t \in (0,1) \tag{3.9}$$

(see Figure 4.2), then their AUC statistics are also ordered:

$$\mathrm{AUC}_A \geq \mathrm{AUC}_B. \tag{3.10}$$

The AUC is equal to the probability that test results from a randomly selected pair of diseased and non-diseased subjects are correctly ordered, namely $P[m(\mathbf{X}_D) > m(\mathbf{X}_{\bar{D}})]$ (Bamber, 1975; Hanley and McNeil, 1982).

The AUC for the joint accuracy of the d dimensional vector $\mathbf{X}$ can be expressed as the following probability

$$U_0 = P\{m(\mathbf{X}_D) > m(\mathbf{X}_{\bar{D}})\}. \tag{3.11}$$

This is interpreted as the probability that a randomly selected diseased subject has a risk score higher than a randomly selected non-diseased subject.

The AUC can be estimated nonparametrically by the following U-statistic,

$$\mathbb{U}(m) = n_D^{-1} n_{\bar{D}}^{-1} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} \mathbf{1}\{m(\mathbf{X}_{D,i}) > m(\mathbf{X}_{\bar{D},j})\}, \tag{3.12}$$

where $\mathbf{X}_{D,i}$ denotes the argument of the ith diseased subject, and $\mathbf{X}_{\bar{D},j}$ denotes the argument of the jth non-diseased subject. The estimator does not require any distributional assumption of the test results (Pepe, 2003). When m is known, it is easy to verify that $\mathbb{U}(m)$ is unbiased to U_0. We further establish the asymptotic normality in the following theorem. Denote the distribution functions for $m(\mathbf{X}_D)$ and $m(\mathbf{X}_{\bar{D}})$ to be F_D and $F_{\bar{D}}$.

Theorem 1. *Assume the d-dimensional density functions π_D and $\pi_{\bar{D}}$ for $\mathbf{X}_D$ and $\mathbf{X}_{\bar{D}}$, respectively to be continuous. Each of the functions m_i ($i = 1, \cdots, d$) is strictly monotone and has a bounded first order derivative m_i'. As sample size n_D and $n_{\bar{D}}$ tend to infinity, $n_D/n \to \lambda \in (0,1)$, we have*

$$\sqrt{n}\{\mathbb{U}(m) - U_0\} \to_d N(0, \sigma_u^2),$$

where $\sigma_u^2 = \lambda^{-1}\|F_D \cdot F_{\bar{D}}^{-1}\|^* + (1-\lambda)^{-1}\|F_{\bar{D}} \cdot F_D^{-1}\|^*$ and $\|h\|^* = \int_0^1 h^2(t)dt - (\int_0^1 h(t)dt)^2$.

Proof. Given the continuous density function π_D for $\mathbf{X}_D$, we can derive the density function of $m(\mathbf{X}_D)$. Let $y_1 = x_1$, $y_2 = x_2$, ..., $y_{d-1} = x_{d-1}$, $y_d = m(\mathbf{X}_D) = \sum_{i=1}^d m_i(x_i)$. Therefore, $x_1 = y_1$, $x_2 = y_2$, ..., $x_{d-1} = y_{d-1}$, $x_d = m_d^{-1}\left\{y_d - \sum_{i=1}^{d-1} m_i(y_i)\right\}$.

Define

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_{d-1}} & \frac{\partial x_1}{\partial y_d} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_{d-1}} & \frac{\partial x_2}{\partial y_d} \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{\partial x_d}{\partial y_1} & \frac{\partial x_d}{\partial y_2} & \cdots & \frac{\partial x_d}{\partial y_{d-1}} & \frac{\partial x_d}{\partial y_d} \end{vmatrix}$$

$$= \begin{vmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \frac{\partial x_d}{\partial y_1} & \frac{\partial x_d}{\partial y_2} & \cdots & \frac{\partial x_d}{\partial y_{d-1}} & \frac{\partial x_d}{\partial y_d} \end{vmatrix}$$

$$= \frac{\partial x_d}{\partial y_d} = \frac{1}{\frac{\partial y_d}{\partial x_d}} = \frac{1}{m_d'(x_d)} = \frac{1}{m_d' \left[m_d^{-1} \left\{ y_d - \sum_{i=1}^{d-1} m_i(y_i) \right\} \right]}.$$

The joint pdf of $(y_1, y_2, ..., y_d)$ is given by

$$\pi_D \left[y_1, \cdots, y_{d-1}, m_d^{-1} \left\{ y_d - \sum_{i=1}^{d-1} m_i(y_i) \right\} \right] |J|$$

$$= \pi_D \left[y_1, \cdots, y_{d-1}, m_d^{-1} \left\{ y_d - \sum_{i=1}^{d-1} m_i(y_i) \right\} \right] \left| m_d' \left[m_d^{-1} \left\{ y_d - \sum_{i=1}^{d-1} m_i(y_i) \right\} \right] \right|^{-1}$$

The density function of $m(\mathbf{X}_D)$ should be

$$f_D(y_d) = \int \pi_D \left[y_1, \cdots, y_{d-1}, m_d^{-1} \left\{ y_d - \sum_{i=1}^{d-1} m_i(y_i) \right\} \right] \left| m_d' \left[m_d^{-1} \left\{ y_d - \sum_{i=1}^{d-1} m_i(y_i) \right\} \right] \right|^{-1} dy_1 \cdots dy_{d-1}.$$

To simplify the notation, we use $f_D(y)$ to denote $f_D(y_d)$. It is easy to show that the density is uniformly continuous and bounded away from zero and infinity. Similarly we verify the same properties for $f_{\bar{D}}(y)$.

The estimator (3.12) is the sum of trapezoid areas under the empirical ROC curve which is constructed from the process $\hat{F}_D \cdot \hat{F}_{\bar{D}}^{-1}(y)$ by using the empirical versions of F_D and $F_{\bar{D}}$. Komlos *et al.* (1975) showed the strong approximation of the uniform empirical process by a sequence of Brownian bridge obtained from a single Kiefer process. Csorgo and Revesz (1978) established the strong approximation of the quantile process by a Kiefer process. Combining these two known results and assumed properties of the distribution functions, we obtain that the process $\sqrt{n}\{\hat{F}_D \cdot \hat{F}_{\bar{D}}^{-1}(y) - F_D \cdot F_{\bar{D}}^{-1}(y)\}$ can be approximated by a suitably defined Brownian bridge (Hsieh and Turnbull, 1996) with probability of one uniformly on $[a, b]$. □

In practice, we may estimate σ_u^2 by $\hat{\sigma}_u^2 = \lambda^{-1}\|\hat{F}_D \cdot \hat{F}_{\bar{D}}^{-1}\|^* + (1-\lambda)^{-1}\|\hat{F}_{\bar{D}} \cdot \hat{F}_D^{-1}\|^*$, where $\hat{F}_{\bar{D}}$ and $\hat{F}_{\bar{D}}$ are the empirical versions of the distribution functions for the two classes. We can easily verify that $\hat{\sigma}_u^2$ is strongly consistent to σ_u^2 (Komolos *et al.*, 1975).

When the function m is unknown, we need to first obtain an estimator $\hat{m}$. For a linear logistic regression, maximum likelihood estimation for $\boldsymbol{\beta}$ is a well-known approach. For a nonparametric additive logistic regression, smoothing methods such as the local polynomial regression or the smoothing spline are usually coupled with a back-fitting algorithm to yield the functional estimates. These estimation approaches have been implemented in many statistical packages, such as the function `gam` in `R`. The consistency and asymptotical normality have been

established for a variety of estimators $\hat{m}$. Therefore, in this thesis, we assume that we have chosen a consistent estimator $\hat{m}$ such that

$$\sqrt{n}\{\hat{m}(\mathbf{x}) - m(\mathbf{x})\} \to_d N(0, \sigma^2_{\mathbf{x}}),$$

for any $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$. For simplicity, we ignore the estimation bias when $\hat{m}$ is obtained via nonparametric regression methods.

The nonparametric estimation for the AUC is obtained by a plug-in method

$$\mathbb{U}(\hat{m}) = n_D^{-1} n_{\bar{D}}^{-1} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} \mathtt{1}\{\hat{m}(\mathbf{X}_{D,i}) > \hat{m}(\mathbf{X}_{\bar{D},j})\}. \tag{3.13}$$

We cannot establish the exact convergence rate of $\mathbb{U}(\hat{m})$ for U_0 and subsequently cannot evaluate the asymptotic variance analytically. However, we can show the following result which is relatively stronger than convergence in probability.

Theorem 2. *Assume the same conditions in Theorem 1 hold. As sample sizes n_D and $n_{\bar{D}}$ tend to infinity, we have*

$$E\{\mathbb{U}(\hat{m}) - \mathbb{U}(m)\}^2 \to 0.$$

Proof. We can write $E\{\mathbb{U}(\hat{m}) - \mathbb{U}(m)\}^2 = E[\mathbb{U}(\hat{m})]^2 - 2E[\mathbb{U}(\hat{m})\mathbb{U}(m)] + E[\mathbb{U}(m)]^2 = I_1 - 2I_2 + I_3$, where $I_k = n_D^{-2} n_{\bar{D}}^{-2} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} \sum_{i'=1}^{n_D} \sum_{j'=1}^{n_{\bar{D}}} I_{kiji'j'}$ and $I_{1iji'j'} = E\mathbf{1}\{\hat{m}(\mathbf{X}_{D,i}) > \hat{m}(\mathbf{X}_{\bar{D},j}), \hat{m}(\mathbf{X}_{D,i'}) > \hat{m}(\mathbf{X}_{\bar{D},j'})\}$, $I_{2iji'j'} = E\mathbf{1}\{\hat{m}(\mathbf{X}_{D,i}) > \hat{m}(\mathbf{X}_{\bar{D},j}), m(\mathbf{X}_{D,i'}) > m(\mathbf{X}_{\bar{D},j'})\}$, $I_{3iji'j'} = E\mathbf{1}\{m(\mathbf{X}_{D,i}) > m(\mathbf{X}_{\bar{D},j}), m(\mathbf{X}_{D,i'}) > m(\mathbf{X}_{\bar{D},j'})\}$. At any continuous point (say 0) of $m(\mathbf{X}_D) - m(\mathbf{X}_{\bar{D}})$ for any $\epsilon > 0$ one may find

a δ such that the difference between $P\{m(\mathbf{X}_D) - m(\mathbf{X}_{\bar{D}}) \leq 0\}$ and $P\{m(\mathbf{X}_D) - m(\mathbf{X}_{\bar{D}}) \leq -\delta\}$ is less than ϵ. We notice that $|I_{2iji'j'} - I_{3iji'j'}|$ is less than

$$\begin{aligned}
& P\{\hat{m}(\mathbf{X}_{D,i}) \leq \hat{m}(\mathbf{X}_{\bar{D},j}), m(\mathbf{X}_{D,i}) > m(\mathbf{X}_{\bar{D},j})\} \\
& +P\{\hat{m}(\mathbf{X}_{D,i}) > \hat{m}(\mathbf{X}_{\bar{D},j}), m(\mathbf{X}_{D,i}) \leq m(\mathbf{X}_{\bar{D},j})\} \\
\leq\ & P\{\hat{m}(\mathbf{X}_{D,i}) \leq \hat{m}(\mathbf{X}_{\bar{D},j}), m(\mathbf{X}_{D,i}) > m(\mathbf{X}_{\bar{D},j}) + \delta\} \\
& +P\{\hat{m}(\mathbf{X}_{D,i}) > \hat{m}(\mathbf{X}_{\bar{D},j}), m(\mathbf{X}_{D,i}) \\
\leq\ & m(\mathbf{X}_{\bar{D},j}) - \delta\} + \epsilon \\
\leq\ & 2P[|\hat{m}(\mathbf{X}_{D,i}) - \hat{m}(\mathbf{X}_{\bar{D},j}) - \{m(\mathbf{X}_{D,i}) - m(\mathbf{X}_{\bar{D},j})\}| \geq \delta] + \epsilon
\end{aligned}$$

Under the pre-assumed weak convergence of $\hat{m}$, this implies that each $I_{2iji'j'}$ converges to $I_{3iji'j'}$. Similarly we can show that $I_{1iji'j'}$ converges to $I_{3iji'j'}$. We further notice that the limit of I_3 is finite. That completes the proof. □

Since $\mathbb{U}(\hat{m})$ is convergent to U_0, we will use $\mathbb{U}(\hat{m})$ to estimate U_0. We compare AUC scores of several generalized additive models with different combinations of explanatory variables and their interaction in order to select the best model to do prediction for other herpesviruses.

3.7 Further Refinement of the GAM Approach

Natural DNA is replete with special sequence features and global or local heterogeneity in composition. We will explore more sequence features, which possibly

relate to the locations of replication origins in herpesviruses, for further refinement of the GAM approach.

3.7.1 Features to Be Selected

In addition to the sequence features used in previous chapters (i.e., palindromes, close direct repeats, AT content and local maxima), more features will be used for the construction of the GAM which are described below.

1. Subfamily: Members of the herpesvirus family are classified into the α, β, and γ subfamilies according to the virus host range and other biological properties. The relative locations of replication origins of herpesviruses in the same subfamily are more similar. Two variables, X_α and X_β, are used to indicate a window comes from:

$$X_\alpha = \begin{cases} 1, & \text{if this herpesvirus belongs to the subfamily } \alpha; \\ 0, & \text{otherwise.} \end{cases}$$

$$X_\beta = \begin{cases} 1, & \text{if this herpesvirus belongs to the subfamily } \beta; \\ 0, & \text{otherwise.} \end{cases}$$

If $X_\alpha = 0$ and $X_\beta = 0$, then the herpesvirus of the window under consideration belongs to the γ family.

2. Standardized window number: As described in Cruz-Cano et al. (2010), the standardized window number is defined as the window number divided by the

total number of windows in the genome sequence. Therefore, the window number will be a real number in the range from 0 to 1 after normalization. For example, if a genome sequence of a virus has 400 windows in all, then the corresponding standardized window number for the first window is 1/400=0.0025. The inclusion of this variable to build up our GAM was due to the empirical observation that the replication origins were located in relatively similar parts of the genome, especially for members in the same subfamily. Cruz-Cano et al. (2010) gave a schematic representation of the genomes as vertical bars in Figure 3.8, where the black colored regions are those windows close to known replication origins.

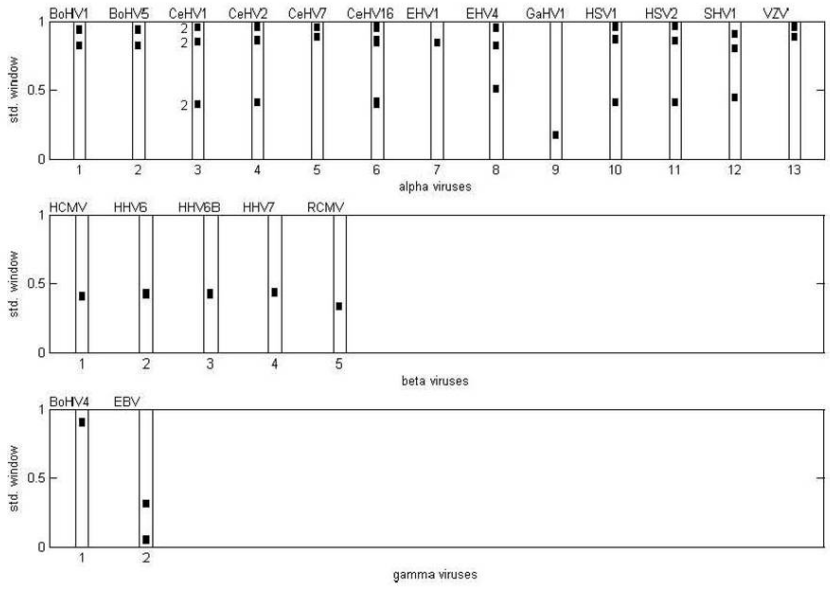


Figure 3.8: Replication origins of herpesviruses (from Cruz-Cano *et al.* (2010))

3. Dinucleotide scores: A dinucleotide is a single piece of DNA made up of any two contiguous nucleotide bases. There are 16 possible dinucleotides in DNA, which are `AA, AT, AG, AC, TA, TT,..., TC`. Genomic compositional inhomo-

geneities are widely recognized (Kozhukhin and Pevzner, 1991; Karlin *et al.* 1993). As early as the 1960s and 1970s, researchers extensively applied biochemical methods of DNA dinucleotide frequency analysis to estimate dinucleotide frequencies in samples of genomic DNA in many organisms (Josse *et al.*, 1961; Swartz *et al.*, 1962; Russell *et al.*, 1976; Russell *et al.*, 1977). The biochemical experiments established that the set of dinucleotide odds ratio values, or 'general design', is a remarkably stable property of the DNA of an organism. Early studies have demonstrated that the set of dinucleotide odds ratio values constitute a genomic signature which can discriminate between sequences from different organisms (Karlin and Burge, 1995). Dinucleotide odds ratio values reflect the species-specific property of DNA modification and replication. In our research, our extended dinucleotide score is a member of the set of standby variables, which will explore the relationship with the locations of replication origins in virus genomes.

Let f_X and f_Y denote the frequency of the nucleotides X and Y (A, T, G or C) in the sequence and f_{XY} the frequency of dinucleotide XY. A standard assessment of dinucleotide bias is through the odds-ratio calculation, $\rho_{XY} = f_{XY}/f_X f_Y$. If ρ_{XY} values much larger (smaller) than 1, then the dinucleotide XY will be considered of high (low) relative abundance compared with a random association of its component mononucleotides (Burge *et al.*, 1992).

The score for a dinucleotide XY in window w is

$$score(XY) = \log\left(\frac{f_{XY,w}}{f_{X,w}f_{Y,w}}\right),$$

where $f_{XY,w}$, $f_{X,w}$ and $f_{Y,w}$ denote the frequencies of dinucleotide XY, X and Y in window w respectively. We use $L(w)$ to denote the length of window w, and use $N(A_w)$ to denote times of the dinucleotide or nucleotide A that appears in window w. Then, the frequencies are as follows:

$$f_{XY,w} = \frac{N(XY_w)}{L(w)-1},\ f_{X,w} = \frac{N(X_w)}{L(w)},\ f_{Y,w} = \frac{N(Y_w)}{L(w)}.$$

Because the counts of some dinucleotides are 0 in some windows, a common device is to incorporate a pseudo count. We increase every dinucleotide count by 1 in each window. Then the pseudo frequencies of dinucleotide XY, single nucleotide base X and Y are converted to

$$f'_{XY,w} = \frac{N(XY_w)+1}{L(w)-1+16} = \frac{N(XY_w)+1}{L(w)+15},$$
$$f'_{X,w} = \frac{N(X_w)+1}{L(w)+4},$$
$$f'_{Y,w} = \frac{N(Y_w)+1}{L(w)+4}.$$

Now the score of dinucleotide XY is converted to the pseudo score,

$$score'(XY) = \log\left(\frac{f'_{XY,w}}{f'_{X,w}f'_{Y,w}}\right) \tag{3.14}$$

We combine the window pseudo dinucleotide scores of each herpesvirus genome sequence, and then build up and fit generalized additive models. Since the ranges and the distribution of these pseudo scores differ in various herpesviruses, we have to standardize these window pseudo dinucleotide scores for each herpesvirus

sequence before combining them:

$$\frac{score'(XY) - mean[score'(XY)]}{s.d.[score'(XY)]}.$$

The candidate features set of herpesviruses contains a total of 31 variables: 2 for subfamily classification (α, β subfamilies), 1 for palindromes(P), 1 for close direct repeats(R), 1 for AT content(ATcon), 3 for local maxima of P, R and ATcon, 3 for the interaction of two feature scores among P, R and ATcon, 3 for interaction of P, R and ATcon and their corresponding local maxima, 1 for standardized window number, and 16 for dinucleotide scores.

3.7.2 Model Selection

Among the 31 target variables, we ascertained the dominant features by a variable selection approach. We compared AUC values of Generalized Additive Models (GAMs) with single variable in order to rank the importance of these variables in predicting replication origins in herpesviruses. The nonparametric GAM model with a univariate smooth function is in the form

$$\log(\frac{p}{1-p}) = \alpha + s(X) \tag{3.15}$$

where X is any one variable in the 31-variable set, p is the probability that the window is close to a replication origin based on the variable X and $s(\cdot)$ is a univariate unknown smooth function. It is noted that if X is one of the local maxima,

subfamily classification or standardized window number (discrete variables), the model will be simplified to a generalized linear model

$$\log(\frac{p}{1-p}) = \alpha + \beta X. \tag{3.16}$$

Because the AUC value is a reasonable single measure to assess the predictive accuracy with a trade-off of sensitivity and specificity, the AUC was used as the criterion to select the best model. We applied a forward stepwise approach to select variables by the AUC criterion, where variables were added one by one to the model. We used $f_{AUC}(\mathbf{X})$ to denote the AUC value of a logistic generalized additive model with a vector of predictor variables $\mathbf{X}$. We selected variables from 31 possible predictor variables $\{X_1, X_2, ..., X_{31}\}$, which denote window scores of 31 candidate sequence features. The variable selection procedure is as follows:

1. The first selected variable was $X_{k_1} = \underset{1 \le k \le 31}{argmax} f_{AUC}(X_k)$, where *argmax* stands for the argument of the maximum, that is, the argument of AUC function $f_{AUC}(X_k)$ that achieved the maximum value when $1 \le k \le 31$. Then the model chosen at this step would be

$$\log(\frac{p}{1-p}) = \alpha + s(X_{k_1}). \tag{3.17}$$

2. The second step selected among the remaining 30 variables, exclusive of X_{k_1}. We chose the second important variable X_{k_2} in the sense that the value of $f_{AUC}(X_{k_1}, X_{k_2})$ is the largest among models with two variables: one is X_{k_1}, the

other one is any $X_{k_j}(1 \leq j \leq N, j \neq k_1)$, that is, $X_{k_2} = \underset{1 \leq j \leq N, j \neq k_1}{argmax} f_{AUC}(X_{k_1}, X_j)$. Then we got the best logistic generalized additive bivariate model containing X_{k_1}:

$$\log(\frac{p}{1-p}) = \alpha + s(X_{k_1}) + s(X_{k_2}). \tag{3.18}$$

3. Similarly, the next step selected among the remaining 29 variables, exclusive of X_{k_1} and X_{k_2}, and so on. The 31 variables were selected one by one in this forward stepwise way until all predictor variables were selected.

In each step, we include one more variable to our generalized additive model. The model with the highest AUC value was chosen to be the model to predict replication origins in herpesviruses. The stepwise model selection approach was denoted as GAM31 (AUC).

3.8 The Application of Generalized Additive Models to Prediction of Replication Origins in Caudoviruses

The *caudovirales* are an order of viruses that have double–stranded DNA genomes. The genome length ranges from 18 kbp to 500 kbp (Orlova, 2009). G+C contents in DNA are 27–72% and are usually similar to their host DNA (Fauquet, *et al.*, 2005).

As described in Fauquet *et al.* (2005), the order of *caudovirales* consists of the three families of bacterial viruses infecting Bacteria and Archaea: Myoviridae (long contractile tails), Siphoviridae (long non-contractile tails), and Podoviridae (short non-contractile tails). Tailed bacterial viruses are an extremely large group with highly diverse virion, genome, and replication properties. Over 4,500 descriptions have been published (as of November 2001). However, data pertaining to virion structure, genome organization and replication properties are available for only a small number of well-studied species (Fauquet *et al.*, 2005).

Early studies have suggested that *Herpesviridae* and caudoviruses share some common properties (Baker *et al.*, 2005). For example, the most fundamental common point is that they are all large linear double–stranded DNA viruses (Baker *et al.*, 2005). In addition, *Herpesviridae* and *Caudovirales* may share a common ancestry (Baker *et al.*, 2005). Ackermann (1998) found caudovirales and herpesviruses share elements of morphogenesis and life-style that are attributed to convergent evolution. In 2005, Baker *et al.* indicated that the *Herpesviridae* and *Caudovirales* are structurally and evolutionarily related based on analysis of their capsid structures. In the replication mechanism, *Herpesviridae* and *Caudovirales* have several direct evolutionary links (Baker *et al.*, 2005). With respect to DNA packaging, both the *Herpesviridae* and most of the *Caudovirales* possess a terminase-portal protein system of DNA packaging (Catalano, 2000; Newcomb *et al.*, 2001; Iyer, *et al.*, 2006). Since *Herpesviridae* and *Caudovirales* have a lot in common, we propose to apply the Generalized Additive Model, which has been

proved to be useful to predict replication origins in *Herpesviridae*, to the order of *Caudovirales*.

Twenty caudoviruses have known replication origins. Their genome sequences and the locations of known replication origins were obtained from Genbank at the NCBI web-site (http://www.ncbi.nlm.nih.gov/) in 2009. Table 3.7 displays the caudoviruses that were used in this study, their genome length, and family they belong to based on the documented annotations of the GenBank files.

Table 3.7: The list of *Caudovirales* to be analyzed.

Virus	Accession	Genome Length	Family
Enterobacteria phage T4	NC_000866	168903	Myoviridae
Streptococcus phage Sfi21	NC_000872	40739	Siphoviridae
Lactobacillus prophage phiadh	NC_000896	43785	Siphoviridae
Yersinia phage phiYeO3-12	NC_001271	39600	Podoviridae
Enterobacteria phage P4	NC_001609	11624	Myoviridae
Lactococcus phage c2	NC_001706	22172	Siphoviridae
Lactococcus phage sk1	NC_001835	28451	Siphoviridae
Enterobacteria phage P2	NC_001895	33593	Myoviridae
Streptococcus phage Sfi11	NC_002214	39807	Siphoviridae
Enterobacteria phage P22	NC_002371	41724	Podoviridae
Lactococcus phage Tuc2009	NC_002703	38347	Siphoviridae
Enterobacteria phage HK620	NC_002730	38297	Podoviridae
Enterobacteria phage T3	NC_003298	38208	Podoviridae
Lactobacillus phage A2	NC_004112	43411	Siphoviridae
Streptococcus prophage EJ-1	NC_005294	42935	Myoviridae
Enterobacteria phage Sf6	NC_005344	39043	Podoviridae
Enterobacteria phage P1	NC_005856	94800	Myoviridae
Enterobacteria phage ES18	NC_006949	46900	Siphoviridae
Staphylococcus phage 80alpha	NC_009526	43864	Siphoviridae
Salmonella phage epsilon34	NC_011976	43016	Podoviridae

The predictive procedure for caudoviruses is similar with that of herpesviruses. The genomic sequence is partitioned into non-overlapping windows. The window size w for each window is about 1% of the genome length, rounded down to the nearest hundred bases for convenience. We used 1% instead of 0.5% as we did for herpesviruses, because the average genome sequence length of caudoviruses is much shorter than that of herpesviruses.

Raw data were transformed and standardized as the procedure of analyzing herpesviruses. Then we pooled the standardized window scores of different caudoviruses together to build up and fit models. Similarly, the candidate features set of caudoviruses contained a total of 31 variables. The same forward stepwise variable selection approach as that for herpesviruses was used to choose the best model.

Chapter 4

Results and Discussion

We collect the results of our statistical analysis in this thesis in this chapter. Section 4.1 contained the predictive accuracies of generalized additive models (GAMs) and generalized linear models (GLMs) using repeats (R), AT content (AT), palindromes (P) and their local maxima (LM_R, LM_{AT} and LM_P) as covariates. The optimal model was selected by the AUC criterion. In Section 4.2, the predictive accuracy of the selected model for known replication origins in herpesviruses is given. In the next section, we predict potential locations of unknown replication origins in herpesviruses. In Section 4.4, we introduce the refined GAM approach and results, where a stepwise variable selection approach by the AUC criterion is applied. In Section 4.5, we compare our selected model with existing methods and find that our approach works better in predicting replication origins in herpesviruses than others in terms of sensitivity and positive predictive value. Our

optimal GAM method is applied to caudoviruses in Section 4.6. Finally, some discussion is given in Section 4.7.

4.1 Predictive Accuracies using Palindromes, AT content, Repeats and Their Local Maxima

Before the generalized additive model (GAM) was applied to predict replication origins in herpesviruses, the generalized linear model (GLM) was used firstly. In our problem, the response variable Y is dichotomous, whether the window is near a replication origin or not, and the data analysis is aimed at relating this outcome to the explanatory variables. We code the response variable Y as zero or one according to the outcome. The GLM approach is commonly used to predict a binary outcome from continuous and/or discrete explanatory variables, which models the *logit* of the response probability with a linear form

$$\text{logit}\{P(\boldsymbol{X})\} \equiv \log\left\{\frac{P(\boldsymbol{X})}{1-P(\boldsymbol{X})}\right\} = \boldsymbol{\beta X} \tag{4.1}$$

where $P(\boldsymbol{X}) = \text{pr}(Y = 1|\boldsymbol{X})$ and $\boldsymbol{X}$ is the vector of explanatory variables. The unknown parameter $\boldsymbol{\beta}$ is to be estimated. Our generalized linear models contain several combinations of variables among palindromes (P), repeats (R), AT content (AT) and their local maxima (LM). Because AUC is a widely accepted single measure to assess the predictive accuracy, we calculate AUC values of GLMs and GAMs to compare their predictive performance. Table 4.1 displays the comparison

of AUC values and their standard errors of GLMs and GAMs with the same explanatory variables.

Table 4.1: AUC values and their standard errors (s.e.) of GLMs and GAMs with the same explanatory variables.

Model No.	Covariate	AUC (GLM)	s.e. of AUC (GLM)	AUC (GAM)	s.e. of AUC (GAM)
1	R	0.628	0.018	0.627	0.017
2	AT	0.512	0.016	0.654	0.016
3	P	0.504	0.017	0.557	0.014
4	R, P, AT	0.637	0.018	0.713	0.015
5	R, P, AT, LM_{AT}	0.649	0.017	0.717	0.015
6	R, P, AT, LM_P	0.648	0.017	0.711	0.015
7	R, P, AT, LM_R	0.647	0.018	0.719	0.015
8	R, P, AT, LM_R, LM_{AT}	0.658	0.017	0.723	0.015
9	R, P, AT, LM_R, LM_P	0.658	0.017	0.717	0.015
10	R, P, AT, LM_{AT}, LM_P	0.661	0.017	0.716	0.015
11	R, P, AT,LM_R, LM_P, LM_{AT}	0.671	0.016	0.721	0.015

From Table 4.1, we notice that the GAM approach surpasses the GLM approach in terms of AUC values. The AUC value of each GAM is higher than that of the corresponding GLM except Model 1. Because GAMs achieve higher predictive accuracy, we employed the GAM approach rather than the GLM approach in this thesis.

We want to select the best model from GAMs with various covariates and forms of predictors, which are given in Table 4.2. AUC values for all kinds of

models are given in the fourth column in Table 4.2. We observe the following:

I. The AUC value of the model using close direct repeats (R) alone is approximately 0.63 with standard error of estimate 0.017, higher than AUC values of AT content and palindrome–based methods. This suggests that our close direct repeats-based method outperforms the AT content-based and palindrome-based methods.

II. Model 4 $\log[p/(1-p)] = f_1(R) + f_2(P) + f_3(AT)$ is then considered in our analysis. It can been seen from Table 4.2 that the AUC value of Model 4, which takes more sequence features into account simultaneously, surpasses univariate Models 1–3. This indicates that the model integrating more sequence features can achieve better prediction accuracy.

III. We include interactions of variables R, AT, P and their local maxima into the predictor of Model 4 sequentially. Each of Models 5–10 contains local maxima of any one or two variables among R, AT, P, and the interactions of the variables and their local maxima, respectively, which are in linear forms. The interactions of variables R, AT, P and their local maxima are included in Model 11 in a linear form. Lastly, we consider the most complicated model, Model 12. It differs from Model 11 in the functional form of interaction. Interactions of R, AT, P and their local maxima are in the form of a smooth function as in Model 12 instead of a linear form as in Model 11.

Table 4.2: The AUC values and their standard error (s.e)

for various Generalized Additive Models.

Model No.	Covariate	Predictor	AUC	s.e.
1	R	$s(R)$	0.627	0.017
2	AT	$s(AT)$	0.654	0.016
3	P	$s(P)$	0.557	0.014
4	R, P, AT	$s(R) + s(P) + s(AT)$	0.713	0.015
5	R, P, AT, LM_{AT}	$s(R) + s(P) + s(AT) + \beta_1 \cdot AT \cdot LM_{AT} + \beta_2 \cdot LM_{AT} + \beta_0$	0.717	0.015
6	R, P, AT, LM_P	$s(R) + s(P) + s(AT) + \beta_1 \cdot P \cdot LM_P + \beta_2 \cdot LM_P + \beta_0$	0.712	0.015
7	R, P, AT, LM_P	$s(R) + s(P) + s(AT) + \beta_1 \cdot R \cdot LM_R + \beta_2 \cdot LM_R + \beta_0$	0.719	0.015
8	R, P, AT, LM_R, LM_{AT}	$s(R) + s(P) + s(AT) + \beta_1 \cdot R \cdot LM_R + \beta_2 \cdot LM_R + \beta_3 \cdot AT \cdot LM_{AT} + \beta_4 \cdot LM_{AT} + \beta_0$	0.723	0.015
9	R, P, AT, LM_R, LM_P	$s(R) + s(P) + s(AT) + \beta_1 \cdot R \cdot LM_R + \beta_2 \cdot LM_R + \beta_3 \cdot P \cdot LM_P + \beta_4 \cdot LM_P + \beta_0$	0.717	0.015
10	R, P, AT, LM_{AT}, LM_P	$s(R) + s(P) + s(AT) + \beta_1 \cdot AT \cdot LM_{AT} + \beta_2 \cdot LM_{AT} + \beta_3 \cdot P \cdot LM_P + \beta_4 \cdot LM_P + \beta_0$	0.716	0.015
11	R, P, AT,LM_R, LM_P, LM_{AT}	$s(R) + s(P) + s(AT) + \beta_1 \cdot AT \cdot LM_{AT} + \beta_2 \cdot LM_{AT} + \beta_3 \cdot P \cdot LM_P + \beta_4 \cdot LM_P + \beta_5 \cdot P \cdot LM_R + \beta_6 \cdot LM_R + \beta_0$	0.721	0.015

12	R, P, AT,LM_R, LM_P, LM_{AT}	$s(R)+s(P)+s(AT)+s(R\cdot LM_R)+s(P\cdot LM_P)+s(AT\cdot LM_{AT})+\beta_1\cdot LM_P+\beta_2\cdot LM_R+\beta_3\cdot LM_{AT}+\beta_0$	0.724	0.015

R, AT and P denote the window scores of repeats, AT content, and palindromes, respectively. LM_R, LM_{AT}, and LM_P represent local maxima of repeats, AT content, and palindromes, respectively. Model 1-3 are univariate models. Model 4 contains three explanatory variables, R, AT and P. Model 5-12 include different combination of R, AT and P and their local maxima.

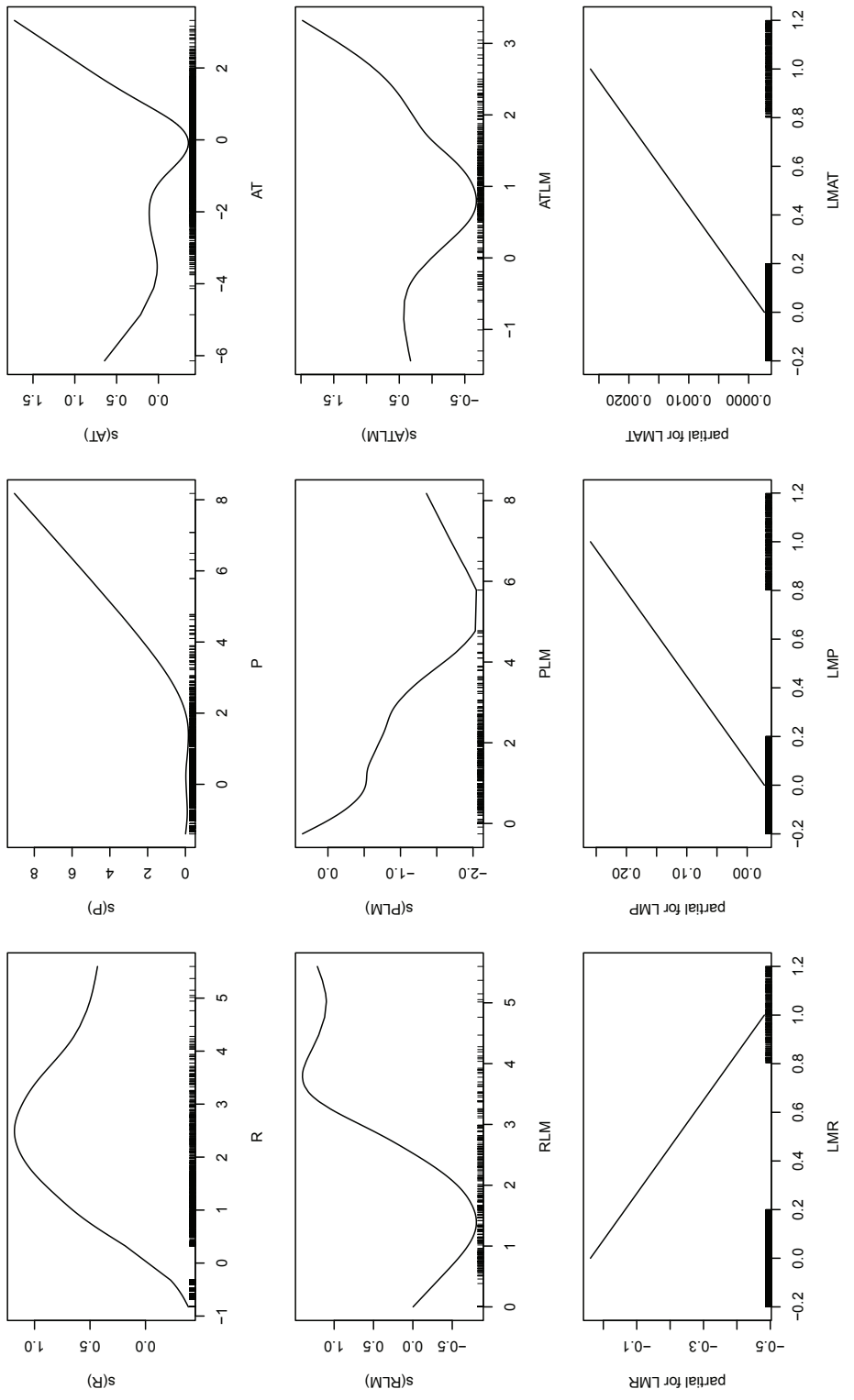


Figure 4.1: A graph showing the predictor effects of model 12.
RLM, PLM, and ATLM denote $R \cdot LM_R$, $P \cdot LM_P$, $AT \cdot LM_{AT}$ respectively. LMR, LMP and LMAT denote variables local maxima of repeats, palindromes and AT content respectively. The y-axis labels for three graphs in the last row "Partial for LMR, LMP, LMAT" indicate partial effects of parametric model components LMR, LMP, LMAT respectively. Black lines on x-axis of last three graphs indicate the covariate values of LMR, LMP, LMAT are 0 and 1.

Figure 4.1 shows how the predictors affect the response of Model 12. From the three graphs in the second row in Figure 4.1, we see that most of the data concentrate in the regions where the predictors affect response almost monotonically. In order to simplify the model, we consider the semi-parametric additive models 5-11 (refer to Table 4.2). In these models, the items with interaction $R \cdot LM_R$, $P \cdot LM_P$, $AT \cdot LM_{AT}$ are parametric, and different models contain different covariates.

After we fit the generalized additive model, predictor effects can be examined separately. We find that among Models 5-11, only predictors in Model 5 and Model 8 have approximately monotone increasing effects. This means that the higher the covariates score, the larger the predictor effect. This relationship is consistent with the empirical study results (Chew et al., 2005; Chew et al., 2007). Figure 4.2 and Figure 4.3 show the effects of key predictors in Model 5 and Model 8.

Thus we will choose one of the two models. The standard deviation of AUC for Model 5 and Model 8 are each around 0.015. So the difference of the AUC values between Model 5 and Model 8 is not significant. Therefore, Model 5, being a simpler model, is chosen as our final model. On the other hand, Model 8 differs from Model 5 in that Model 8 considers the local maxima of repeats and interaction of repeats and local maxima of repeats, while Model 5 does not. After investigating the plots of window centers versus window scores, we find that the local maxima of repeats are much fewer than those of AT content and palindromes. This can be

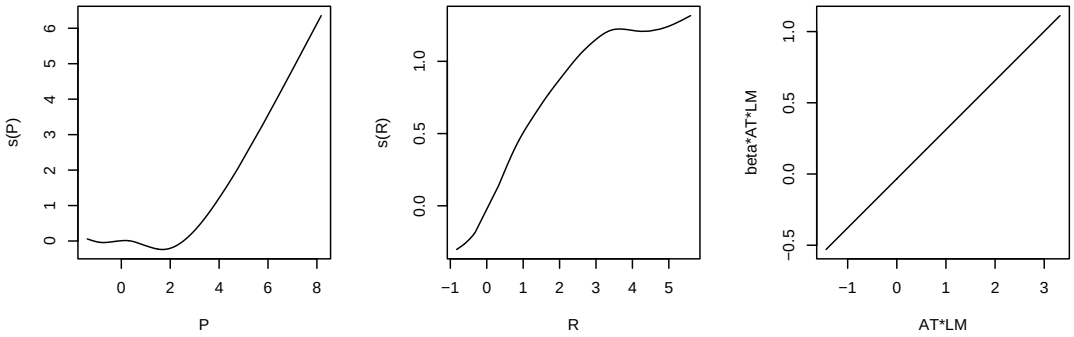


Figure 4.2: A graph showing the effects of the key predictors P, R, and $AT \cdot LM_{AT}$ in Model 5.

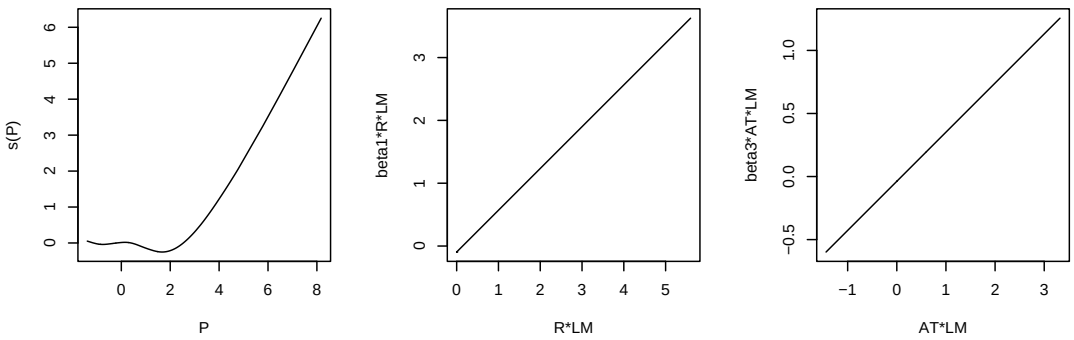


Figure 4.3: A graph showing the effects of the key predictors P, $R \cdot LM_R$, and $AT \cdot LM_{AT}$ in Model 8.

seen in Figure 4.4 and Figure 4.5. We give the examples of herpesviruses bohv4 and cehv2 in Figure 4.4 and Figure 4.5, where the comparisons of local maxima of repeats, AT content and palindromes are shown. The circles indicate the locations of replication origins.

These results suggest that the sequence feature local maxima of repeats may affect the model weakly; hence, it is reasonable to ignore the local maxima of repeats. For the reasons of the comparative AUC values and plots of local maxima, Model 5 is chosen rather than Model 8.

4.2 Predictive Accuracy for Known Replication Origins in Herpesviruses

In order to check the predictive accuracy, we examine the correspondence between the location predicted by our approach and those of the known replication origins. A cross-validation procedure is employed to evaluate the sensitivity and positive predictive value of the GAM approach. Our data set contains 20 viral genomes with known replication origins. We apply the commonly used leave-one-out cross-validation method to assess the predictive performance of the model (Ripley, 1996): using 19 viral genomes to fit the generalized additive model, and then predict the locations of replication origins of the remaining one. This procedure is repeated for each viral genome, in turn. Then, we compare the locations of replication origins

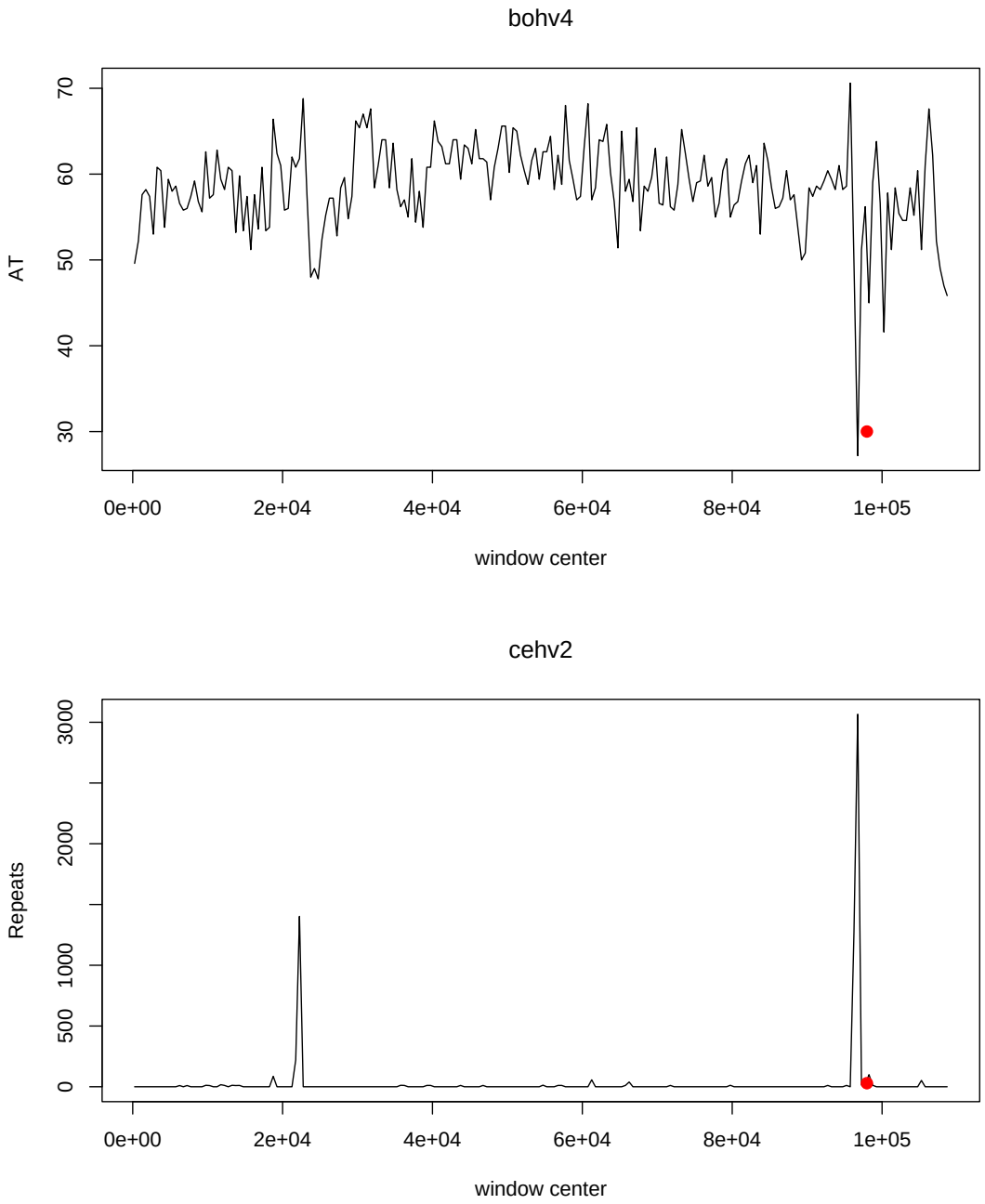


Figure 4.4: Window scores of AT content and Repeats in virus bohv4.

The x-axis indicates window centers along the genome sequence and the y-axis indicates corresponding window scores of AT content and repeats. The x-axis values represented by the circles indicate the true locations of replication origins.

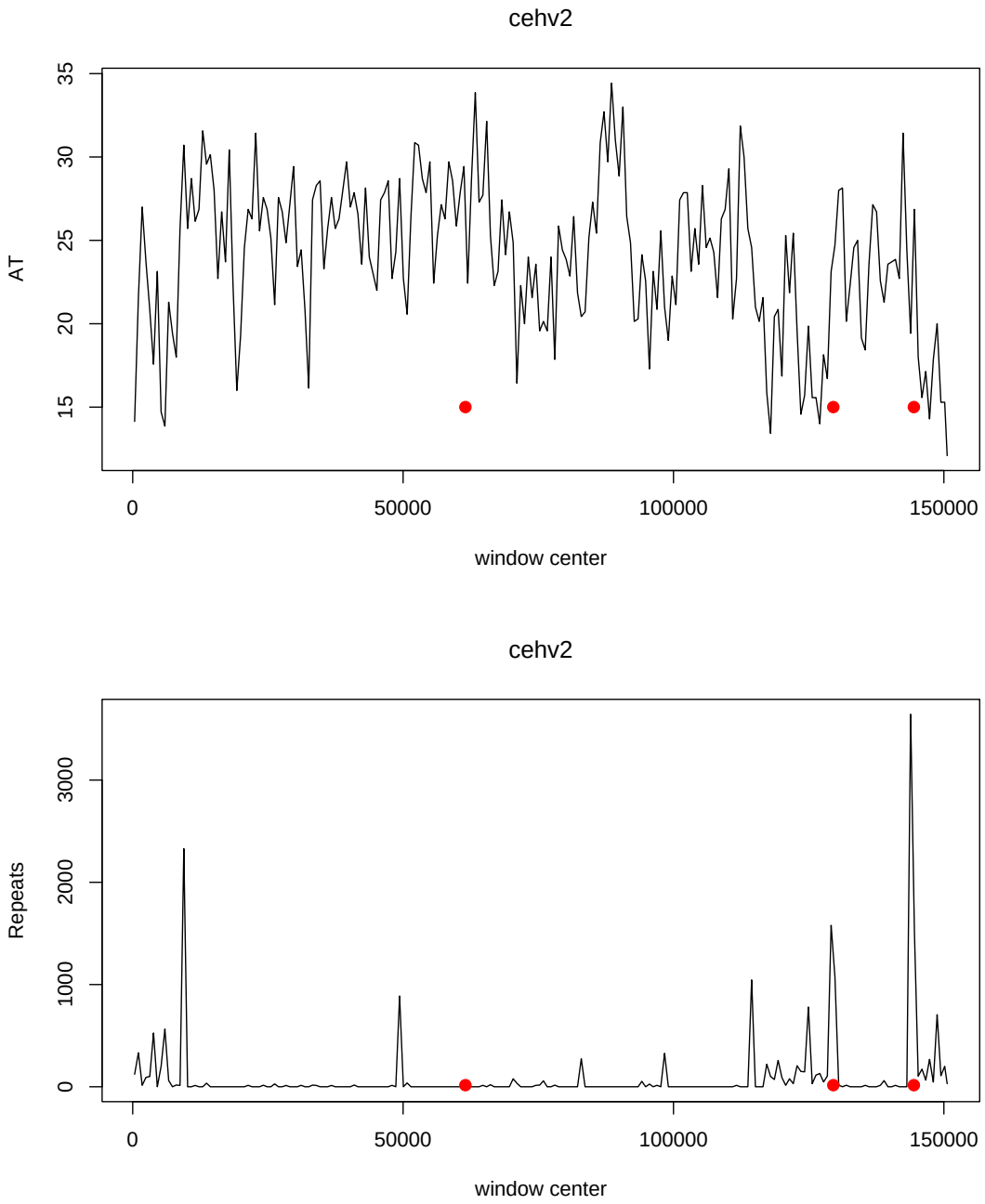
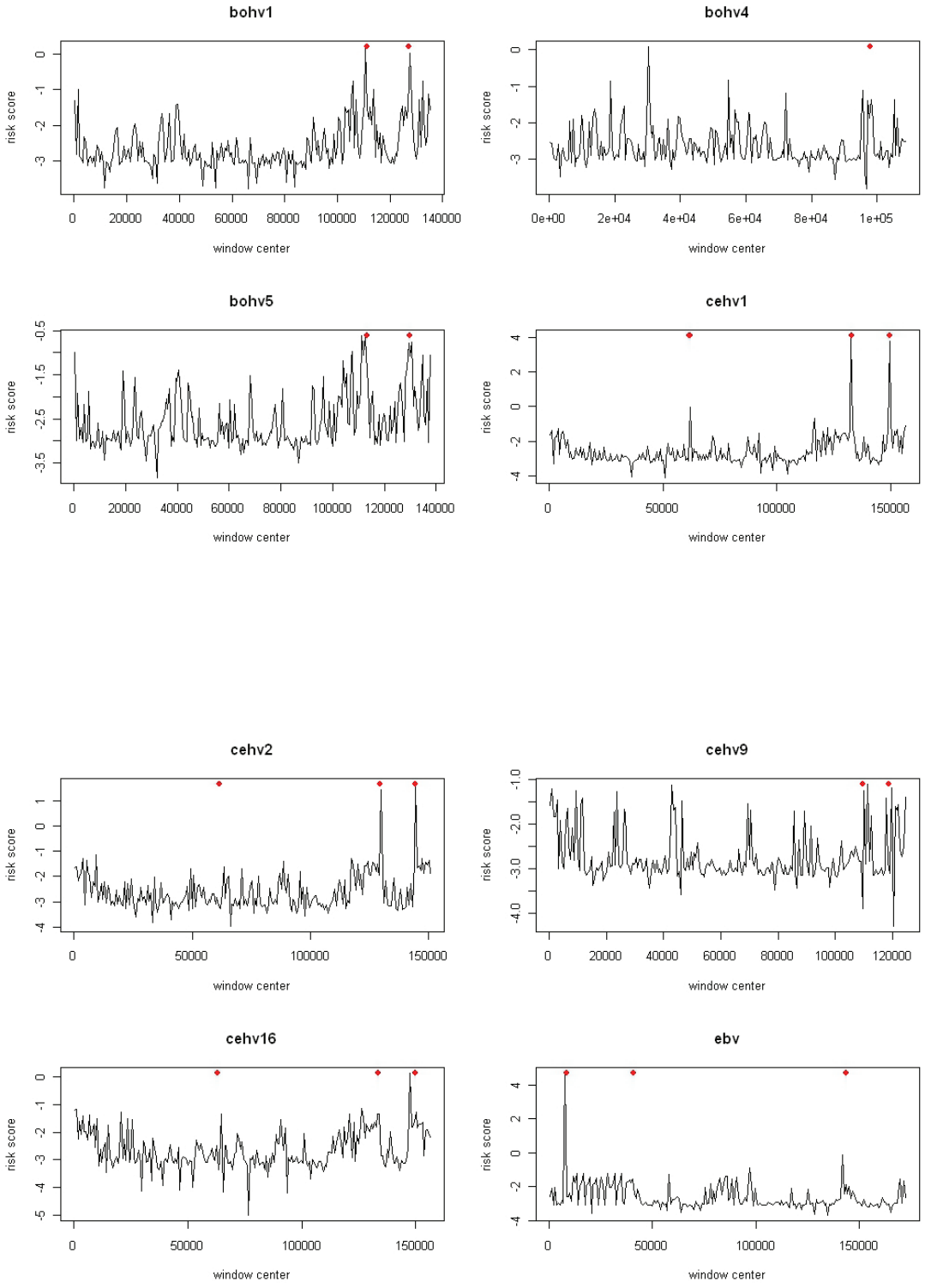


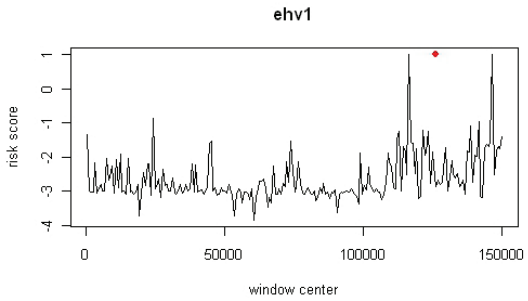
Figure 4.5: Window scores of AT content and Repeats in virus cehv2.

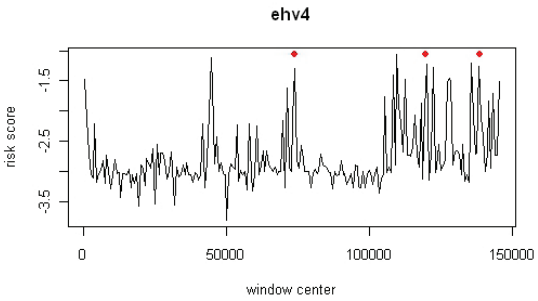
The x-axis indicates window centers along the genome sequence and the y-axis indicates corresponding window scores of AT content and repeats. The x-axis values of the circles indicate the true locations of replication origins.

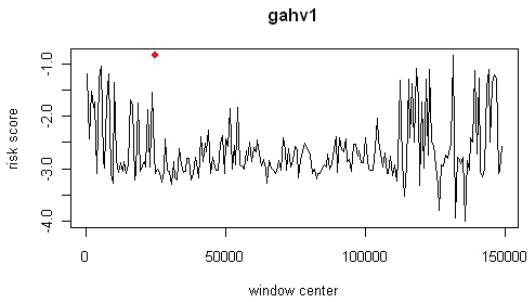
with the predicted locations of all 20 viral genomes. The positions of 43 replication origins of herpesviruses in our data set have been known so far. Locations of known replication origins were presented in Table 4.3. By the cross-validation approach, we can predict the risk score for each window of 20 herpesviruses genome sequences based on the chosen generalized additive model, Model 5. The fitted risk scores are plotted against the positions of window centers along the genome sequences for each of the herpesviruses with known replication origins in Figure 4.6. The x-axis values of red points indicate the known locations of replication origins. The replication origins are predicted at the peaks of the risk score curves. We can roughly see the predictive performance by comparing the peaks of the curves to the red point positions. For most of the herpesviruses, the positions of red points are consistent with the curve peaks, such as bohv1, cehv1, and hcmv, whose locations of replication origins are perfectly predicted by our generalized additive modeling approach. This means the highest risk scores correspond to the true locations of replication origins in herpesviruses. However, this approach fails to predict replication origins in hhv7. The general idea about how well the predictive approach performs can be obtained from Figure 4.6. Most of the replication origins in herpesviruses were predicted. The exact number of replication origins captured by this approach can be further examined.

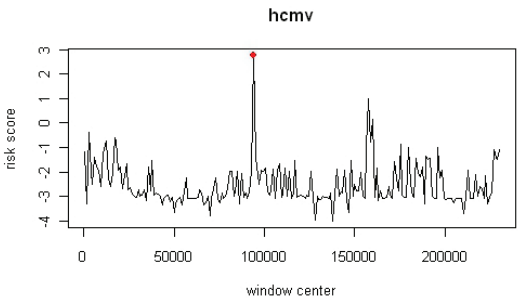
Alternatively, we can rank the windows in each herpesvirus according to their risk scores. We identify the windows with the highest ranks, which are really close to known replication origins. The highest ranks are listed in the last column titled

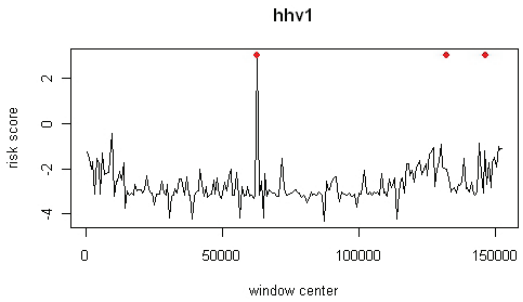

bohv1
risk score
window center
bohv4
risk score
window center
bohv5
risk score
window center
cehv1
risk score
window center
cehv2
risk score
window center
cehv9
risk score
window center
cehv16
risk score
window center
ebv
risk score
window center

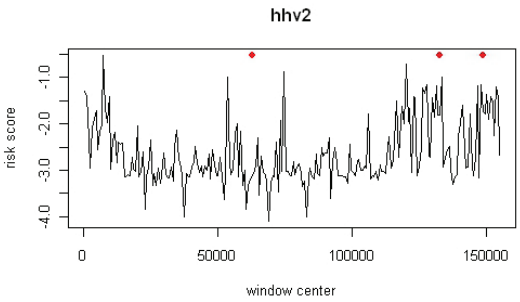

ehv1
risk score
window center

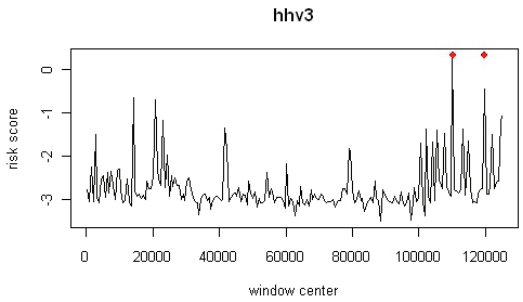

ehv4
risk score
window center

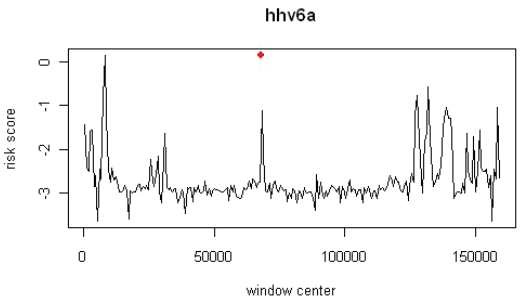

gahv1
risk score
window center


hcmv
risk score
window center


hhv1
risk score
window center


hhv2
risk score
window center


hhv3
risk score
window center


hhv6a
risk score
window center

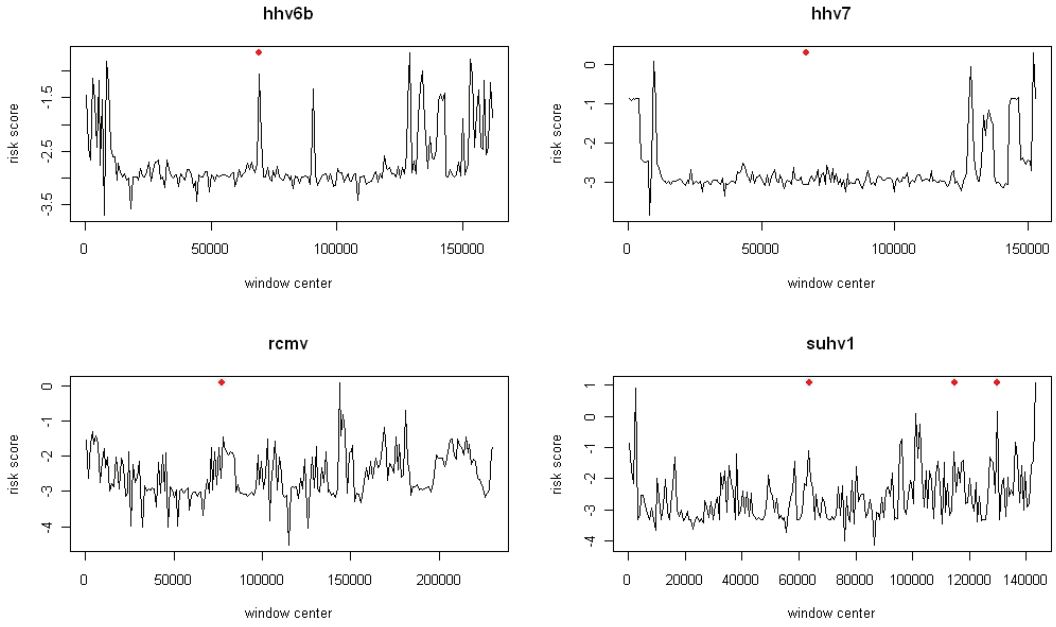


Figure 4.6: The plot of risk scores on the y-axis versus window centers along the x-axis for each herpesvirus genome sequence with known replication origins.

The risk scores for each of herpesvirus genome sequences are displayed in each graph. The red points correspond to the known locations of replication origins. The higher peaks of the risk score curves are more indicative of predicted locations of replication origins.

"Predictive Top Window" in Table 4.3. For example, "Predictive Top Window" for the virus bohv4 is 7. This means that the window, which was ranked seventh among windows in bohv4 genome sequence in terms of the fitted risk scores, successfully captured the replication origin; while the top 1-6 windows failed to predict its location. Table 4.3 shows that the 38 replication origins out of the total 43 were correctly predicted by the windows with the top 1-10 risk scores. The exact predictive accuracy of our generalized additive modeling approach will be discussed in the next section.

4.3 Prediction of Unknown Replication Origins in Herpesviruses

In this section, we will apply the generalized additive modeling approach to predict the locations of replication origins in 27 herpesviruses with unknown replication origins.

Because Model 5 $\log[p/(1-p)] = f_1(R) + f_2(P) + f_3(AT) + \beta_1 AT \cdot LM_{AT} + \beta_2 LM_{AT} + \beta_0$ is chosen as our ultimate model, it is used to predict potential locations of replication origins in each herpesvirus with unknown replication origins. We compute the risk score for each window, then choose the windows with the top 10 risk scores as the potential locations of replication origins. Table 4.4 provides a list of the top 10 risk scores for each of the 27 herpesviruses by our general-

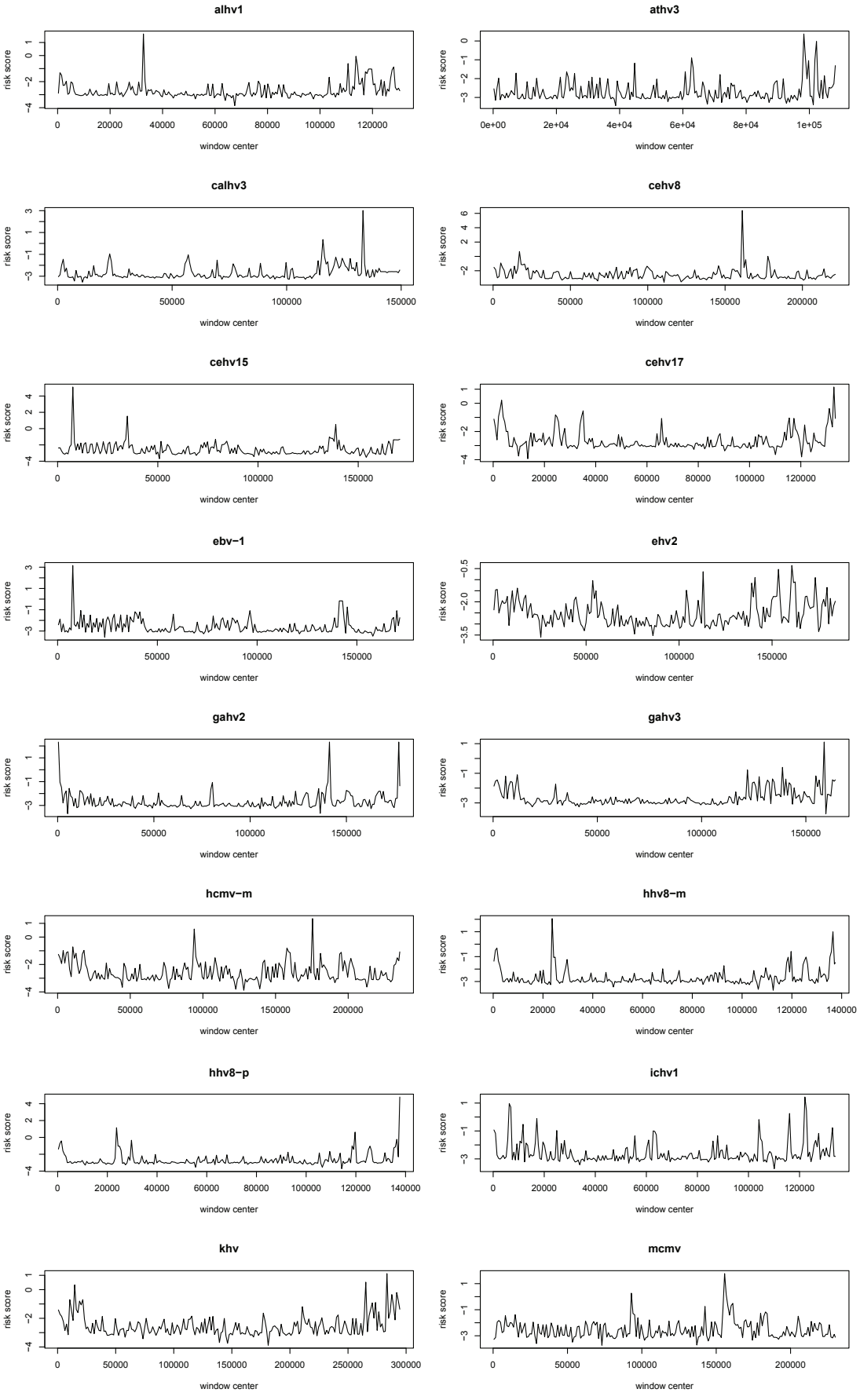
Table 4.3: Centers of known replication origins and the predictive top windows that captured replication origins. For example, for the virus hcmv, the top 1 risk scoring window correctly captured its replication origin.

Virus	Known Ori Center	Window Center	Family	Predictive Top Window
bohv1	111190	111301	alpha	1
bohv1	127028	126901	alpha	2
bohv4	97996.5	97751	gamma	7
bohv5	113312	113101	alpha	1
bohv5	129701	129901	alpha	2
cehv1	61690.5	61251	alpha	3
cehv1	61893.5	61951	alpha	3
cehv1	132795.5	132651	alpha	1
cehv1	132998.5	132651	alpha	1
cehv1	149425.5	149451	alpha	2
cehv1	149628.5	149451	alpha	2
cehv16	62981	62651	alpha	7
cehv16	133479	133351	alpha	6
cehv16	149824	150151	alpha	1
cehv2	61493.5	61251	alpha	9
cehv2	129537.5	129851	alpha	2
cehv2	144471.5	144551	alpha	1
cehv9	109636.5	109501	alpha	1
cehv9	118622.5	118501	alpha	3
ebv	8313.5	8401	gamma	1
ebv	40797	40401	gamma	11
ebv	143825.5	143601	gamma	2
ehv1	126262.5	126351	alpha	13
ehv4	73909.5	73851	alpha	5
ehv4	119471.5	119351	alpha	4
ehv4	138577.5	138251	alpha	134
gahv1	24871.5	24851	alpha	10
hcmv	93923.5	94051	beta	1
hhv6a	67805	67551	beta	6
hhv6b	69160.5	69201	beta	5
hhv7	66991.5	66851	beta	37
hhv1	62475	62651	alpha	1
hhv1	131999	131951	alpha	4
hhv1	146235	145951	alpha	3
hhv2	62930	62651	alpha	20
hhv2	132760	132651	alpha	4
hhv2	148981	148751	alpha	6
rcmv	77318	77551	beta	6
suhv1	63878	64051	alpha	7
suhv1	114701	114451	alpha	8
suhv1	129901	129851	alpha	3
hhv3	110218.5	110101	alpha	1
hhv3	119678.5	119701	alpha	2

ized additive model. The numbers in the table report the middle positions of the windows.

We also present the prediction in a graphical form, displayed in Figure 4.7, where the risk scores of the windows are plotted against the locations of the windows. The higher risk scoring windows are more likely to be close to replication origins, according to our generalized additive modeling approach.

Our predictive results may be useful to biology researchers. Based on information we provided, they may identify and confirm the exact locations of replication origins in these 27 herpesviruses genomes though experimentation.


alhv1
athv3
calhv3
cehv8
cehv15
cehv17
ebv−1
ehv2
gahv2
gahv3
hcmv−m
hhv8−m
hhv8−p
ichv1
khv
mcmv
risk score
window center

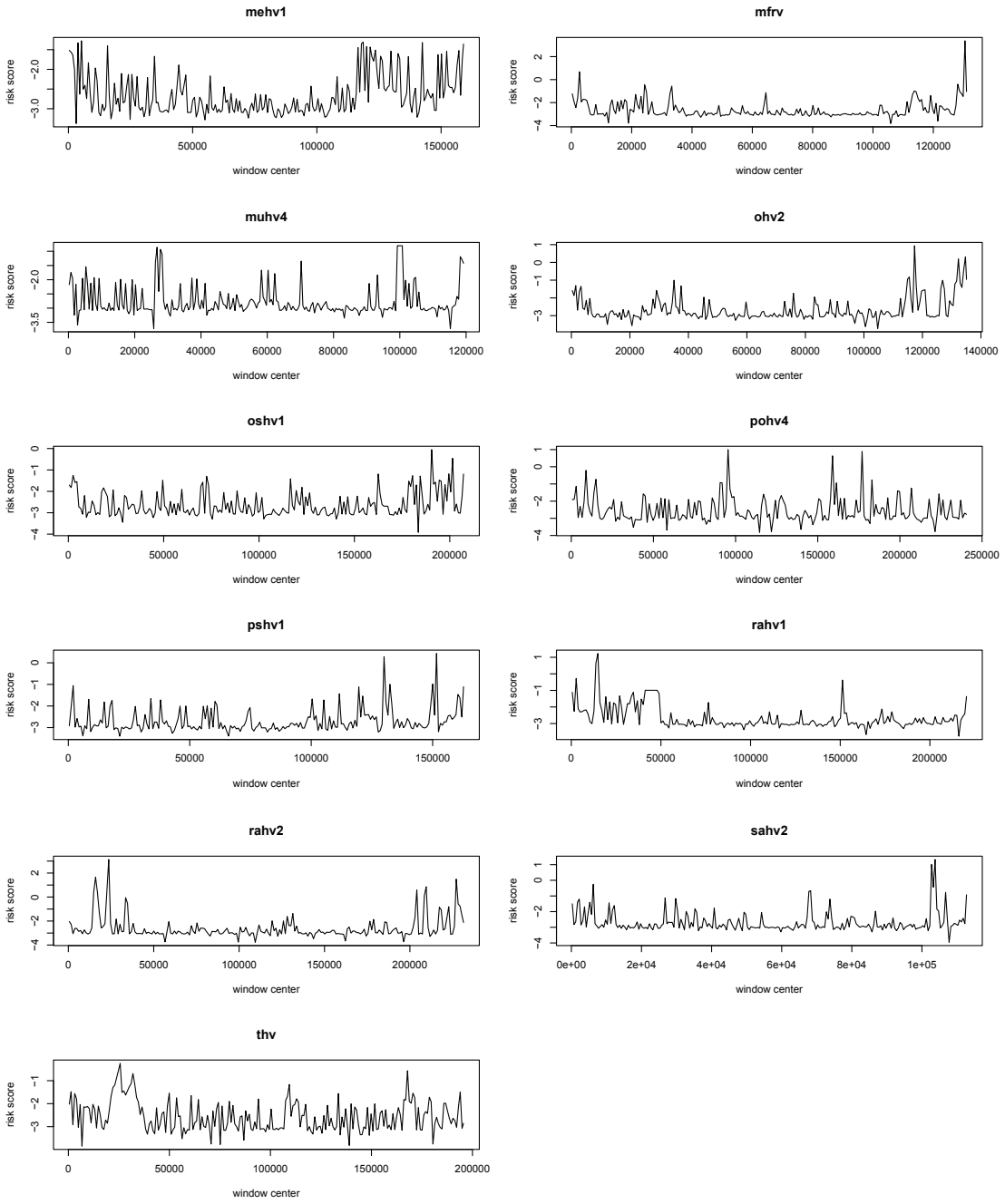


Figure 4.7: Window plots of risk scores for herpesviruses with unknown replication origins. The locations of the windows along the genome sequences are on the x-axis and the risk scores are on the y-axis.

Table 4.4: Predicted locations of replication origins in herpesviruses with unknown replication origins. The numbers in the table indicate the middle positions of the windows.

virus	top1	top2	top3	top4	top5	top6	top7	top8	top9	top10
alhv1	32701	113701	110701	114301	128101	119101	119701	118501	127501	117301
athv3	98251	102251	98751	62751	101751	99751	44751	108205	63251	60751
calhv3	133351	115851	22751	57051	116551	121451	124251	127751	2451	69651
cehv8	161151	17051	177651	163351	4951	178751	20351	18151	19251	145751
cehv15	7601	34801	138801	135601	136401	137201	78801	170749	34001	167601
cehv17	132901	3301	131101	2701	35101	24301	34501	2101	3901	24901
ebv-1	7601	141201	142001	142801	145201	11601	96401	170001	38801	41201
ehv2	160651	153451	112951	140851	173251	53551	162451	161551	139051	179551
gahv2	141201	177201	401	140401	1201	80401	177738	139601	2001	6001
gahv3	158801	138801	122001	11601	154801	6001	127601	134001	140401	162801
hcmv-m	175451	94051	10451	157851	18151	3851	158951	7151	235523	195251

virus	top1	top2	top3	top4	top5	top6	top7	top8	top9	top10
hhv8-m	23701	136501	1501	901	135901	119701	135301	125701	118501	24901
hhv8-p	137685	23701	119701	136501	29701	1501	901	118501	24301	125701
ichv1	122101	6301	6901	122701	116101	17101	104101	11701	132901	301
khv	283501	265301	14701	291901	287701	21701	10501	293301	18901	273701
mcmv	155651	92951	156751	161151	160051	157851	142451	154551	183151	179851
mehv1	5251	118651	142451	3851	117951	159031	15751	120051	121451	116551
mfrv	130501	2701	128101	24301	33301	113701	128701	131009	114301	24901
muhv4	99251	99751	100251	100751	26751	27751	28251	118251	118751	70251
ohv2	117301	134701	132301	134101	115501	114901	135068	35101	126901	131701
oshv1	190501	201501	199501	162501	207220	2501	180501	184501	72501	116501
pohv4	95401	177001	159001	9001	15001	183001	90601	91801	161401	3001
pshv1	151601	130001	150001	132401	2001	162713	119601	111601	160401	121201
rahv1	14851	13751	2751	151251	41251	47851	42351	44551	43451	45651
rahv2	23651	15951	227151	209551	204051	22551	14851	17051	208451	33551
sahv2	103751	102751	6251	103251	68251	67751	106751	112716	26751	29751
thv	25651	167851	24751	31951	23851	31051	32851	109351	22951	30151

4.4 Refined GAM Approach and Results

In order to improve our model, more explanatory variables were included in the model. We expected the models containing information of more sequence features to perform better in predicting replication origins. We explored 31 candidate variables to build the GAM. We first fitted 31 models with single variables, then proceeded to calculate AUC values for each model. The 31 variables were ranked by their AUC values and were listed in Table 4.5.

Table 4.5: AUC values of models with single variable.

Rank	Variable	AUC	Rank	Variable	AUC
1	win.no	0.812	17	AA	0.58
2	R·ATcontent	0.681	18	GA	0.571
3	ATcontent	0.654	19	AC	0.569
4	GC	0.632	20	Xalpha	0.567
5	GG	0.629	21	Xbeta	0.565
6	R	0.627	22	TT	0.565
7	CA	0.627	23	CT	0.564
8	CC	0.623	24	AG	0.559
9	ATcontent·P	0.615	25	P	0.557
10	AT	0.608	26	ATcontent·LMAT	0.542
11	TG	0.608	27	P·LMP	0.538
12	TA	0.604	28	R·LMR	0.538
13	R·P	0.604	29	LMR	0.523
14	TC	0.589	30	LMAT	0.511
15	CG	0.584	31	LMP	0.504
16	GT	0.581			

The forward stepwise variable selection approach by AUC criterion was applied

to seek the best model in terms of AUC values. The variables chosen in each step and the highest AUC values achieved at the corresponding step are given in Table 4.6. The largest AUC value achieved was 0.8771 at the 26th step of the stepwise variable selection procedure. As such the best generalized additive model in terms of AUC values include 26 variables selected in the top 26 steps. The AUC value of this model 0.8771 is much higher than 0.717, the AUC value of the previous Model 5 $\log[p/(1-p)] = f_1(R) + f_2(P) + f_3(AT) + \beta_1 AT \cdot LM_{AT} + \beta_2 LM_{AT} + \beta_0$. This indicates that the refined stepwise GAM approach GAM31 (AUC) improves the predictive accuracy of the general GAM approach (Model 5).

4.5 Comparing the Predictive Accuracy with Existing Methods

Chew *et al.* (2005) reported the predictive accuracy of their approaches in terms of sensitivity and positive predictive values. Here, sensitivity is the percentage of known origins that are close to the regions suggested by the prediction approach. Positive predictive value is the percentage of predicted regions that are close to the true known replication origins. For example, we compute the sensitivity and PPV of the generalized additive modeling approach by top 1 window. Because 11 replication origins of herpesviruses are captured by top 1 window and there

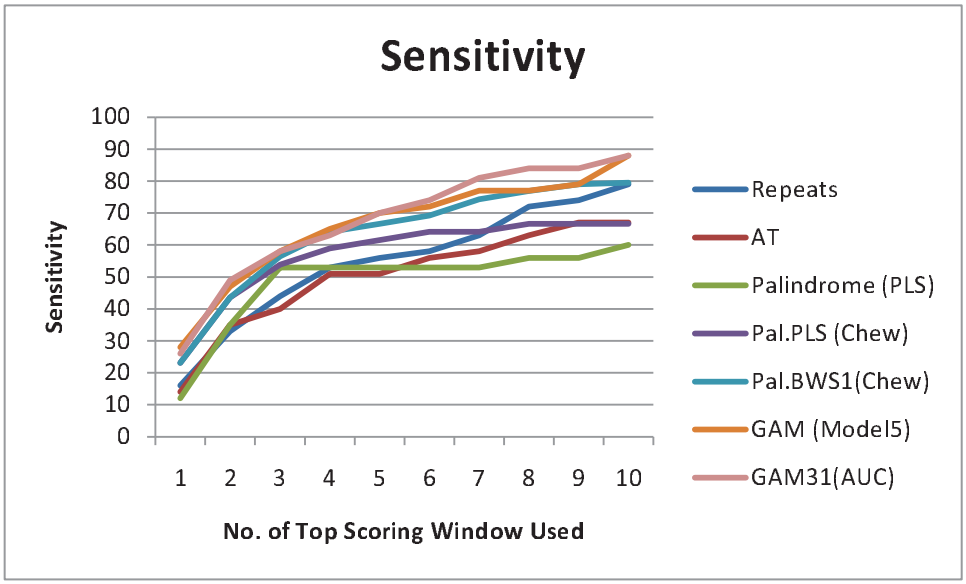
Table 4.6: The variables selected by the forward stepwise variable selection approach and the corresponding AUC values of the generalized additive model at each step in herpesviruses.

Step	Variable Selected	AUC	Step	Variable Selected	AUC
1	win.no	0.8120	17	R	0.8723
2	R·ATcontent	0.8308	18	R·LMR	0.8738
3	GG	0.8375	19	P·LMP	0.8743
4	TA	0.8425	20	AT	0.8746
5	CT	0.8471	21	TT	0.8756
6	ATcontent	0.8509	22	GC	0.8760
7	CA	0.8543	23	AG	0.8765
8	ATcontent·LMAT	0.8574	24	LMAT	0.8768
9	CC	0.8603	25	GA	0.8770
10	ATcontent·P	0.8626	26	LMR	0.8771
11	GT	0.8648	27	LMP	0.8769
12	AA	0.8667	28	Xalpha	0.8766
13	R·P	0.8682	29	Xbeta	0.8758
14	AC	0.8694	30	TG	0.8744
15	P	0.8707	31	CG	0.8739
16	TC	0.8715			

were 43 known replication origins in 20 herpesviruses in all, the sensitivity is $11/43 \times 100 = 26$ and PPV is $11/(20 \times 1) \times 100 = 55$.

We compare the sensitivity and positive predictive values of the GAM31 (AUC) approach to the palindrome-based approach with scoring scheme *palindrome length score* (PLS) and *base-pair weighted score of order 1* (BWS_1) introduced by Chew *et al.* (2005), our single sequence feature approaches based on repeats, AT content, palindromes (with PLS scoring scheme) and the GAM approach (Model 5 $\log[p/(1-p)] = f_1(R) + f_2(P) + f_3(AT) + \beta_1 AT \cdot LM_{AT} + \beta_2 LM_{AT} + \beta_0$). Figure 4.8 displays the comparison. Generally, as the number of top scoring windows used increases, sensitivity increases; however, positive predictive value decreases.

Although the GAM31 (AUC) approach achieved a higher AUC value than the GAM approach (Model 5), their sensitivity and positive predictive values are comparable. The sensitivity and positive predictive value of GAM31 (AUC) approach and GAM approach (Model 5) are higher than those of other approaches by using the top 1-10 windows. It shows that the GAM31 (AUC) approach outperforms Chew's palindrome-based approach by scoring schemes BWS_1 and PLS in terms of both the sensitivity and positive predictive values using the top 1-10 windows. The highest sensitivities attained by GAM31 (AUC), GAM (Model 5), repeats, AT content, and palindrome (PLS) approaches were 88%, 88%, 79%, 65%, and 60%, respectively. The positive predictive value achieved by GAM approach (Model 5) using the top 1 window was 60%, which is 13% higher than Chew *et al.*'s ap-



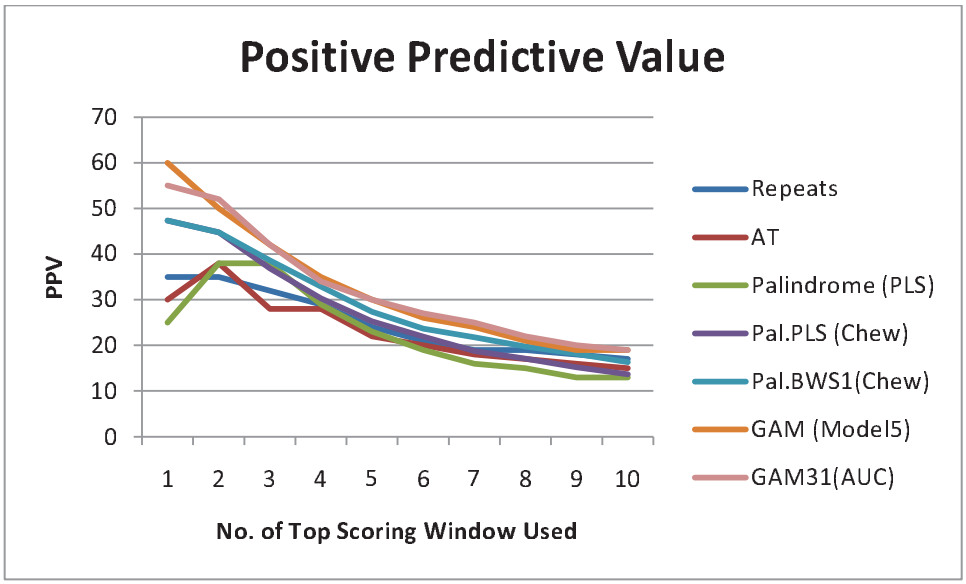


Figure 4.8: Sensitivity and positive predictive values of the GAM31 (AUC) approach, Chew *et al.*'s approaches (2005) and other approaches in this thesis. Repeats, AT, Palindrome (PLS) stand for close direct repeats, AT content and palindromes–based predictive approaches, respectively, where the palindromes–based approach is using PLS scoring scheme and non-overlapping windows. Pal.PLS (Chew) and Pal.BWS1 (Chew) denote the palindromes–based methods with PLS and $BWS1$ scoring schemes introduced by Chew *et al.* (2005). GAM (Model 5) and GAM31 (AUC) are two GAM approaches based on Model 5 and a stepwise model selection procedure by AUC criterion.

proaches (2005) using the top 1 window, a significant finding. In essence, because the GAM approach takes more information about the sequence into account in its prediction, predictive accuracy is expected to be better in terms of sensitivity and positive predictive value. Actually, from our results, it can be seen that predictive accuracy did indeed improve. Another observation from Figure 4.8 is that the predictive accuracies of the repeats-based approach are higher than those of AT content and palindromes (PLS)–based approaches. So the sequence feature of close direct repeats is a valuable feature to be incorporated in our GAM approach.

4.6 Applying the GAM Approach to Caudoviruses

Since the GAM approach works quite well in predicting replication origins in herpesviruses, we apply this approach to caudoviruses whose biological properties are similar to herpesviruses. The candidate variables of the model and model selection procedure are similar to those of herpesviruses. To assess the relative importance of different candidate variables for caudoviruses, we fit 31 models with a single variable. After fitting the models, AUC values for each model were calculated. AUC values are ranked in Table 4.7. This table shows that the standardized window number, the AT content, the interaction of repeats and AT content are considered the most important variables in terms of AUC values of univariate logistic models for herpesviruses.

Table 4.7: AUC values of models with single variable in caudoviruses.

Rank	Variable	AUC	Rank	Variable	AUC
1	win.no	0.711	17	TC	0.569
2	ATcontent	0.617	18	AA	0.566
3	R·ATcontent	0.617	19	GA	0.563
4	R	0.613	20	GT	0.558
5	ATcontent·P	0.603	21	CT	0.558
6	TG	0.600	22	CG	0.555
7	AT	0.585	23	R·LMR	0.554
8	GC	0.583	24	LMR	0.549
9	CA	0.582	25	AC	0.544
10	AG	0.581	26	Sipho	0.541
11	TT	0.579	27	P·LMP	0.534
12	CC	0.577	28	ATcontent·LMAT	0.533
13	TA	0.574	29	Podo	0.533
14	R*P	0.574	30	LMP	0.526
15	P	0.571	31	LMAT	0.515
16	GG	0.507			

The variables selected at each step by the forward stepwise variable selection are listed in Table 4.8. The highest AUC values achieved at each step can be found in this table as well. By including more variables, the AUC value increased at each step and reached 0.8885 when all the 31 candidate variable were selected.

Table 4.8: The variables selected by the forward stepwise variable selection approach and the corresponding AUC values of the generalized additive model at each step for caudoviruses.

Step	Variable Selected	AUC	Step	Variable Selected	AUC
1	win.no	0.7105	17	ATcontent	0.8646
2	P	0.7436	18	TC	0.8674
3	ATcontent·P	0.7645	19	R·ATcontent	0.8706
4	TG	0.7841	20	Podo	0.8729
5	R·P	0.7973	21	Sipho	0.8753
6	P·LMP	0.8078	22	R	0.8785
7	CC	0.8159	23	ATcontent·LMAT	0.8808
8	AG	0.8230	24	R·LMR	0.8826
9	GG	0.8277	25	GT	0.8842
10	CA	0.8329	26	LMAT	0.8853
11	CT	0.8400	27	GC	0.8864
12	AA	0.8443	28	LMP	0.8867
13	TA	0.8474	29	AT	0.8868
14	CG	0.8507	30	TT	0.8884
15	GA	0.8543	31	LMR	0.8885
16	AC	0.8583			

The sensitivity and positive predict value of this model were also calculated.

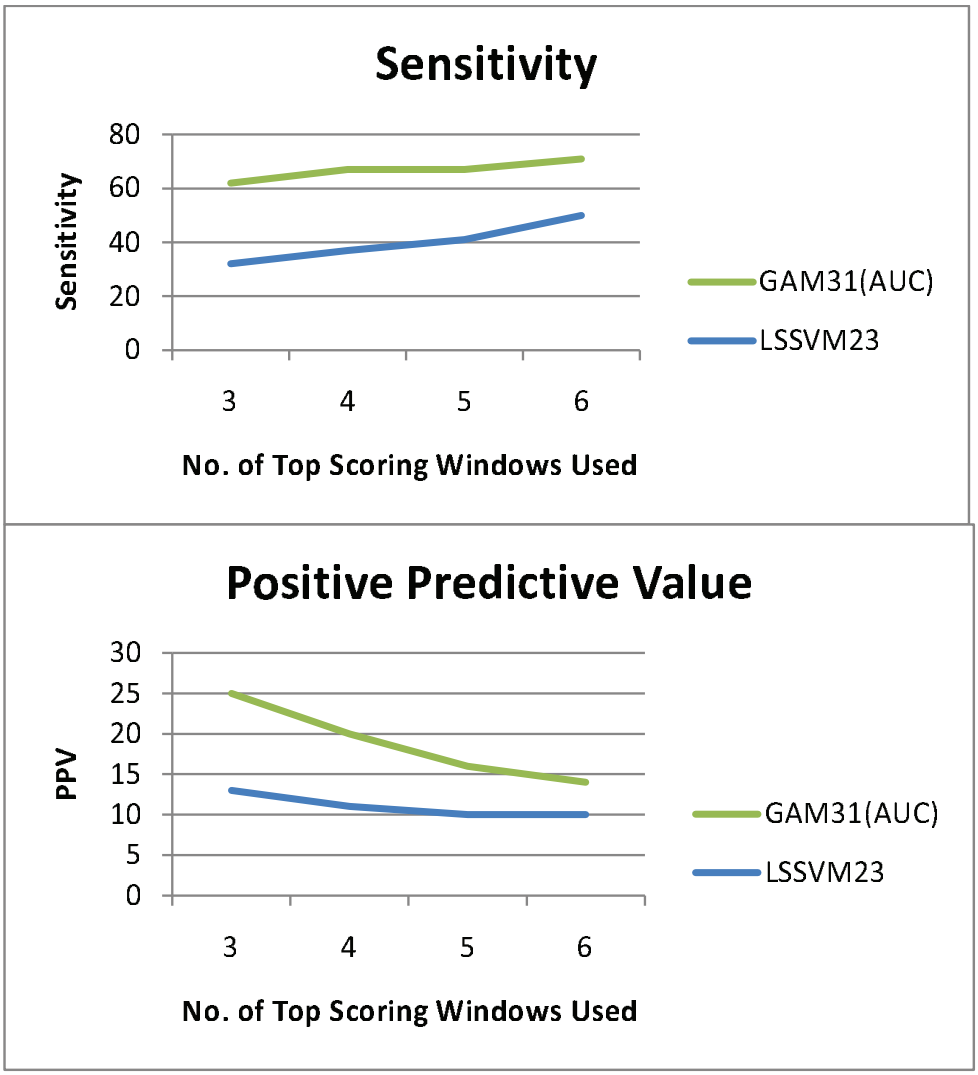


Figure 4.9: Sensitivity and positive predictive values of the GAM31 (AUC) approach and the LSSVM23 approach introduced by Cruz-Cano *et al.* (2010).

Figure 4.9 shows a comparison of the sensitivity and positive predictive values of the stepwise GAM31 (AUC) approach and the LSSVM23 approach introduced by Cruz-Cano *et al.* (2010). Cruz-Cano *et al.* (2010) compared several approaches (the palindrome-based approach using the BWS_1 score scheme (BWS_1), the artificial neural network approach (ANN), the least-squares support vector machines approach with 23 input variables (SVM23), the least-squares support vector machines approach with 16 input variables (SVM16), the least-squares support vector machines approach with 23 input variables in a set of artificial genomes (Art23)). The LSSVM23 approach was the best one among these methods. We compare our GAM31 (AUC) approach with the best existing approach LSSVM23. Both sensitivity and PPV of the GAM31 (AUC) approach are much higher than those of LSSVM23 approaches. The sensitivity and the positive predictive value achieved by the GAM31 (AUC) approach when we choose top 3 windows are 62% and 25% respectively, which are almost twice as the LSSVM23 approach. Since Cruz-Cano *et al.* (2010) only listed the sensitivity and PPV when the number of predictions goes from 3 to 6, we just show our results of the top 3 to 6 top windows. Compared with the LSSVM23 approach, our GAM31 (AUC) method with the top 3 windows correctly predicted 15 replication origins out of 24, while the LSSVM23 approach can only predict 7 replication origins successfully. So the GAM31 (AUC) approach identified 8 replication origins for caudoviruses which Cruz-Cano *et al.* (2010) failed to do so. Actually, if we choose the top 10 windows, sensitivity can be 83%. If we choose the top 1 window, PPV reaches as high as 35%. The results are

quite encouraging. Our GAM approach is a valuable approach to predict origins of replication in caudoviruses.

In both herpesviruses and caudoviruses, the standardized window number is considered the most important variable among 31 candidate variables, which is consistent with the findings of Cruz-Cano *et al.* (2010). So we conclude that the standardized window number provides much useful information of the location of the real replication origins in herpesviruses and caudoviruses.

4.7 Discussion

We recorded here some preliminary statistical analysis attempted before using the GAM approach. In a later subsection, we demonstrate that standardization is an essential pre-processing step in our analysis.

4.7.1 GLM Approach

The first natural approach in modeling binary response is logistic generalized linear models (GLM). Table 4.1 indicates that the GAM approach should be applied rather than the GLM approach in this thesis.

4.7.2 Boosting Approach

We tested a machine learning approach called boosting to predict replication origins in herpesviruses. In 2000, Friedman *et al.* described boosting as an important classification methodology, which is a way of combining the performance of many "weak" classifiers to produce a powerful "committee". However, the procedure failed to solve our prediction problem. It only successfully predicted 3 out of 43 replication origins in herpesviruses. Thus, we did not choose this method.

4.7.3 Predictive Accuracy for α-Herpesvriuses

We expected to improve the predictive accuracy by focusing on α subfamily, because members within the same subfamily share more similar biological properties. Model 5 $\log\left[p/(1-p)\right] = f_1(R) + f_2(P) + f_3(AT) + \beta_1 AT \cdot LM_{AT} + \beta_2 LM_{AT} + \beta_0$ was applied to do prediction for α herpesviruses subfamily. The sensitivity and positive predictive value were compared to those of predicting replication origins in all herpesviruses using the same Model 5. However, as shown in 4.10, the result was not as expected. The procedure that worked on the α subfamily did not improve the predictive accuracy.

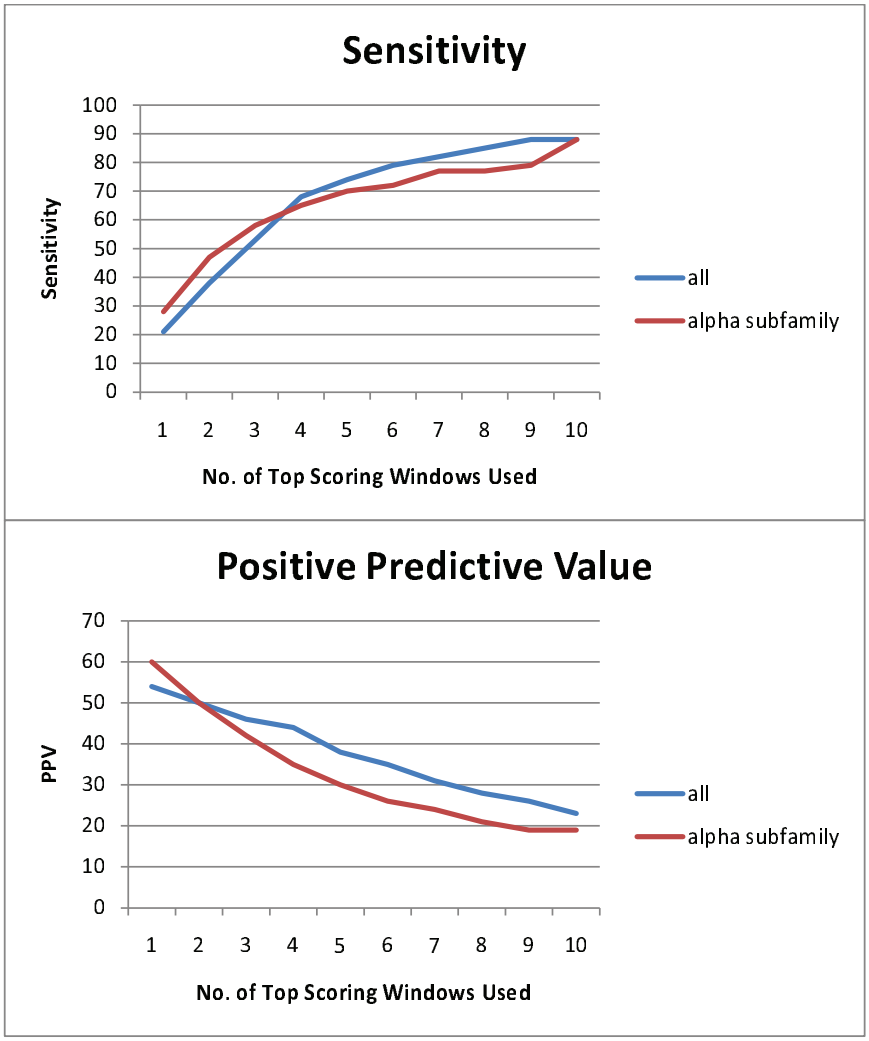


Figure 4.10: Sensitivity and positive predictive values of the GAM approach working on α subfamily and all genome sequences of herpesviruses.

4.7.4 Stepwise GAM Approach by the AIC Criterion

As for the model selection procedure, we tested another criterion, Akaike's information criterion (AIC; Akaike, 1974) instead of AUC. This model selection procedure was implemented in the function `step.gam` in software `R`. The function `step.gam` allows the user to step through arbitrary models along a pre-specified path. It builds a GAM model in a stepwise fashion. Using this approach, we chose the model:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \cdot R + \beta_2 \cdot X_\alpha + \beta_3 \cdot TG + \beta_4 \cdot ATcon \cdot P + s(ATcon) + s(R \cdot P) + s(swinno) + s(AT) + s(TA) + s(CA). \quad (4.2)$$

The AUC value of this model surpasses 0.851, which is higher than GAM approach (Model 5) whose AUC value was only 0.717, but lower than the AUC value of GAM31 (AUC) approach, which was 0.877.

4.7.5 Standardization in the Preprocessing Step

We mentioned in the previous chapter that the pooled window scores that were used to fit our generalized additive models were standardized due to the various ranges of window scores in different virus genome sequences. If we did not standardize our data, the AUC value of Model 5, which was chosen as our final model when the target variable set consisted of only four sequence features (palindromes,

repeats, AT content and local maxima) would be 0.664. This AUC value is lower than that of Model 5 fitted by standardized data, which was 0.717. Therefore, our original window scores should be standardized before they were used to fit the models.

In summary, the GAM is a valuable approach to predict replication origins in some double-stranded DNA viral genomes by integrating multiple sequence features. By comparing the existing approaches of predicting replication origins, we noted that the refined stepwise GAM approach GAM31 (AUC) gave the best result.

Chapter 5

Conclusion and Further Research

5.1 Conclusion

We developed a new computational method to integrate DNA sequence features for more accurate prediction of origins of replication in some double-stranded DNA viral genomes. Sequence features such as palindromes, AT content and close direct repeats are known to be associated with replication origins in viruses (Vlazny and Frenkel, 1981; Boehmer and Lehman, 1997; Hammarsten and Elias, 1997). Palindromes and AT content have been used to individually predict replication origins (Chew *et al.*, 2005; Chew *et al.*, 2007). This study introduced the method based on close direct repeats. Firstly, we introduced a scoring scheme (repeats length scheme) to quantify the spatial abundance of close direct repeats in a genomic sequence. The close direct repeats–based method using this scoring scheme

achieved better predictive accuracy than the palindrome-based and AT content–based methods, to some extent. This result suggests that this sequence feature is important around replication origins. This method is the first computational approach that quantifies close direct repeats in genome sequences to predict replication origins in herpesviruses.

There are three predictive approaches, each of them using only one of these three sequence features. By using the top 1–10 ranked windows based on these sequence features, we examined the numbers of replication origins that were correctly predicted, out of 43 known origins of replication in the herpesviruses. It was found that all three predictive approaches complement each other very well in predicting replication origins in herpesviruses. This result showed that suitably combining these sequence features should improve the performance of prediction. It was also found that our generalized additive model (GAM), a statistical model, which enabled us to take these features and their interactions into account, effective in identifying several replication origins that were not predicted previously. This approach also possessed good predictive accuracy, with both sensitivity and positive predictive values higher than those of existing methods (Chew *et al.*, 2005; Chew *et al.*, 2007).

In order to find out how the locations of replication origins are dependent on sequence features in herpesvirus genomes, we chose the best model among several generalized additive models with various covariates, which were scores

of quantified sequence features. The area under the Receiver Operating Curve (AUC) was adopted as a criterion for generalized additive model selection. The AUC criterion provides a novel guide for generalized additive model selection, which is a good summary measure to evaluate the overall classification accuracy for identifying the dichotomous response.

More sequence features (subfamily, standardized window number, and dinucleotide scores), which were associated with replication origins, were further integrated in the model. A stepwise model selection approach by AUC criterion GAM31 (AUC) was applied. The AUC value of the best generalized additive model selected by this approach was as high as 0.8771 for herpesviruses, which is better than all other approaches described in this thesis. The existing replication origin prediction methods did not perform well for caudoviruses, while the GAM31 (AUC) approach produced much better results. The good performance of this model can be attributed to the combination of information from several sequence features.

In addition, this approach also proved useful to predict replication origins in caudoviruses, which is another kind of double-stranded DNA virus. With the introduced models, the number of identified replication origins in herpesviruses and caudoviruses can be increased significantly. The key contribution of this study is that our GAM approach extends previous work on integrating DNA sequence features, rather than only considering one feature at a time, for more accurate

prediction of replication origins in double-stranded DNA viral genomes. The GAM approach is a valuable addition to existing predictive tools.

5.2 Topics for Further Research

Based on the predictive results obtained, discussion presented and conclusions drawn from this research work, some potential areas for further investigation related to the development of predictive approaches of replication origins in genomes are highlighted below.

5.2.1 Application of Generalized Additive Model to Replication Origins Prediction in Other Viral Genomes.

After building up models and assessing suitable approaches to predict origins of replication in herpesviruses and caudoviruses, we will apply the best model to predict origins of replication in other similar viral families; for instance, poxviruses, baculoviruses, and iridoviruses, which are all double-stranded DNA viruses. Poxviruses are slightly larger double-stranded DNA viruses (Hughes *et al.*, 2010). The genomes range from 130 to 380 kbp (Moss, 2001). Poxviruses can infect various animals (Hughes *et al.*, 2010). For example, the variola virus is a a member of the poxvirus family that causes the disease smallpox. More and more studies focus on poxviruses (Henderson, 1999; Miller, 2003). The baculoviruses, ranging

from 80 to 180 kbp, are a family of large, rod-shaped viruses that contain circular double-stranded genomes (Hyink et al., 2002). Iridoviridae is a family of virus with double-stranded DNA genomes ranging from 150 to 280 kbp (Eaton *et al.*, 2007). Vertebrate iridoviruses are found in fish, amphibians, and reptiles (Eaton *et al.*, 2007). Some iridoviruses infect fish and frogs, which is a serious problem in fish farming, modern aquaculture, and wildlife conservation (Tsai *et al.*, 2005). Since these viral families share similar physical characteristics, we hope that the approaches of predicting replication origins in herpesvirus can be extended to these viruses.

5.2.2 Further Potential Refinements

This study did not take into account the heterogeneity of the genomic sequence, although it was generally known that a genomic sequence is far from being homogenous. Therefore, the model developed in this study should be refined. Future research should attempt different approaches, such as, HMM (Churchill, 1989; Churchill, 1992), the change-point method (Braun and Muller, 1998), or the entropy method (Li, 2001), to segment the genomic sequence into homogenous segments. Then the issue of how to correct the window scores according to their background should be explored.

5.2.3 Exploration of Motifs around Replication Origins

Based on empirical studies, it is recommended that more related sequence patterns should also be considered besides the sequence features used to build up generalized additive models discussed in this thesis. One possible avenue for future work is the exploration of over- (or under-) represented motifs around replication origins.

Some over-(or under-) represented motifs in large sequences have been associated with various biological functions and mechanisms (Frith *et al.*, 2004). Under-represented motifs showed a harmful dysregulatory effect, while over-represented motifs often play an important role in biological function (Frith *et al.*, 2004).

Leung *et al.* (1996) found that clusters of some of the most over- and under-represented 4- and 5-words in some herpesvirus genomes were identified around functional sites such as replication origins and regulatory signals of individual viruses. Based on this finding, further research is therefore needed to identify over- (or under-) represented motifs around known replication origins. It is reasonable to guess that similar over- (or under-) represented motifs may be around unknown replication origins in other herpesvirus genomes. Therefore, similar motifs in other herpesvirus genomes should be explored.

One measurement of over- (or under-) representation is as follows. The frequency of the nucleotide X (A, C, G, or T) in the sequence is denoted by f_X.

Similarly, f_{XY} denotes the frequency of dinucleotide XY, f_{XYZ} denotes the frequency of trinucleotide XYZ, and so on. An odds ratio calculation that is used to assess the dinucleotide bias is through , namely $\rho_{XY} = f_{XY}/f_X f_Y$. If ρ_{XY} is sufficiently larger (or smaller) than 1, then the XY pair is considered over- (or under-) represented compared to a random association of mononucleotides. There are classical statistical tests of the contingency table genre in terms of ρ_{XY} (Hollander and Wolfe, 1973).

There are many approaches and tools to find over- (or under-) represented motifs (Apostolico *et al.*, 2000; Apostolico *et al.*, 2004; Schbath, 1997). VERBUMCULUS is a suite of software tools for the efficient and fast detection of over- or underrepresented words in nucleotide sequences (Apostolico *et al.*, 2004). This tool can find over- and under-represented words within both a single genetic sequence and a family of sequences. Thus, we can use this tool to search for over- (or under-) represented motifs to known replication origins in herpesvirus genome sequences.

5.2.4 Prediction of Replication Origins in Other Organisms

Our method can only predict replication origins in viral genomes, which may not be applicable to other organisms. Because DNA replication mechanisms are different in various organisms, the approach designed for predicting replication

origins in herpesvirus genomes may not work well on other organisms. Therefore, a necessary extension of our work is to develop methods to predict replication origins in other organisms, such as bacteria, eukaryote, etc.

Bibliography

Ackermann, H.W. (1998). Tailed bacteriophages: the order caudovirales. *Advances in virus research*, **51**, 135–201.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716-723.

Albrecht, J.C., Nicholas, J., Biller, D., Cameron, K.R., Biesinger, B., Newman, C., Wittmann, S., Craxton, M.A., Coleman, H. and Fleckenstein,B. (1992). Primary structure of the herpesvirus saimiri genome. *Journal of Virology*, **66**(8), 5047–5058.

Apostolico, A., Bock, M.E., Lonardi, S. and Xu, X. (2000). Efficient detection of unusual words. *Journal of Computational Biology*, **7**(1/2), 71–94.

Apostolico, A., Gong, F. and Lonardi, S. (2004). Verbumculus and the discovery of unusual words. *Journal of Computer Science and Technology*, **19**(1), 22–41.

Baker, M.L., Jiang, W., Rixon, F.J. and Chiu, W. (2005). Common Ancestry of Herpesviruses and Tailed DNA Bacteriophages. *Journal of Virology*, **79**(23), 14967–14970.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, **12**, 387–415.

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. *Nature Medicine*, **7**(7), 859–863.

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

Boehmer, P.E. and Lehman, I.R. (1997). Herpes Simplex Virus DNA Replication. *Annual Review of Biochemistry*, **66**,347–384.

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

Braun, J.V. and Muller, H.G. (1998). Statistical methods for DNA sequence segmentation. *Statistical Science*, **13**(2), 142–162.

Breier, A.M., Chatterji, S. and Cozzarelli, N.R. (2004). Prediction of Saccharomyces cerevisiae replication origins. *Genome Biology*, **5**,R22.

Brewer, B.J. and Fangman, W.L. (1987). The localization of replication origins on ARS plasmids in S.cerevisiae. *Cell*, **51**(3), 463-471.

Bridgen, A. (1991). A restriction endonuclease map for Alcelaphine herpesvirus 1 DNA, *In* S.J.O'Brien, ed., *Genetic Maps, Sixth Edition, Book 1, Viruses.* Cold Spring Harbor Laboratory Press.

Brodie of Brodie, E.B., Nicolay, S., Touchon, M., Audit, B., Aubenton Carafa, Y., Thermes, C., and Arneodo, A. (2005). From DNA sequence analysis to modeling replication in the human genome. *Physical Review Letters*, **94**(24), 248103.

Burge, C., Campbell, A.M. and Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 1358–1362.

Catalano, C.E. (2000). The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. *Cellular and Molecular Life Sciences*, **57**, 128-148.

Chalikian, T., Völker, J., Plum, G. and Breslauer, K. (1999). A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques. *Proceedings of the National Academy of Sciences of the United States of America*, **96**(14), 7853.

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 Research*, **33**(15), e134.

Chew, D.S.H., Leung, M.Y. and Choi, K.P. (2007). AT excursion: a new approach

to predict replicaiton origins by locating AT-rich regins. *BMC Bioinformatics*, **8**, 163.

Churchill, G.A. (1989). Stochastic models for heterogenous DNA sequences. *Bulletin of Mathematical Biology*, **51**, 79–94.

Churchill, G.A. (1992). Hidden Markov chains and the analysis of genome structure. *Computers in Chemistry*, **16**, 107–115.

Clausen-Schaumann, H., Rief, M., Tolksdorf, C. and Gaub, H. (2000). Mechanical stability of single DNA molecules. *Biophysical Journal*, **78**, 1997–2007.

Cruz-Cano, R., Chandran, D. and Leung, M.Y. (2007). Computational prediction of replication origins in herpesviruses. *'07 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology.* 283–290.

Cruz-Cano, R., Chew, D.S.H., Choi, K.P. and Leung, M.Y. (2010). Least-Squares Support Vector Machine Approach to Viral Replication Origin Prediction. *INFORMS Journal on Computing.* **22(3)**, 457–470.

Csorgo, M. and Revesz, P. (1978). Strong approximationis of the quantile process. *Annals of Statistics*, **6**, 882–894.

Davison, A.J., Trus, B.L., Cheng, N., Steven, A.C., Watson, M.S., Cunningham, C., Deuff, R.M.L. and Renault, T. (2005). A novel class of herpesvirus with bivalve hosts. *Journal of General Virology*, **86**, 41–53.

deHaseth, P. and Helmann, J. (1995). Open complex formation by Escherichia coli RNA polymerase: the mechanism of polymerase-induced strand separation of double helical DNA. *Molecular Microbiology*, **16**(5), 817–824.

Dutch, R.E., Bruckener, R.C., Mocarski, E.S. and Lehman, I.R. (1992). Herpes simplex virus type 1 recombination: role of DNA replication and viral a sequences. *Journal of Virology*, **66**(1), 277–285.

Eaton, H.E., Metcalf, J., Penny, E., Tcherepanov, V., Upton, C. and Brunetti, C.R. (2007). Comparative genomic analysis of the family Iridoviridae: re-annotating and defining the core set of iridovirus genes. *Virology Journal*, **4**, 11.

Fauquet, C.M., Mayo, M.A., Maniloff, J., Desselberger, U. and Ball, L.A. (2005). *Virus Taxonomy, Eighth Report of the international committee on taxonomy of viruses*. London: Elsevier/Academic Press.

Frenkel, N., Schirmer, E.C., Wyatt, L.S., Katsafanas, G., Roffman, E., Danovich, R.M. and June, C.H. (1990). Isolation of a new herpesvirus from human $CD4^+$ T cells. *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 748–752.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a additive statistical view of boosting. *The Annals of Statistics*, **28**(2), 337-407.

Friedman, K.L., Raghuraman, M.K., Fangman, W.L. and Brewer, B.J. (1995).

Analysis of the temporal program of replication initiation in yeast chromosomes. *Journal of Cell Science - Supplement*, **19**, 51–58.

Frith, M.C., Hansen, U., Spouge, J.L. and Weng, Z. (2004). Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, **32**(1), 189–200.

Ghosh, D. (2005). Nonparametric methods for analyzing replication origins in genomewide data. *Functional and Integrative Genomics*, **5**, 28–31.

Hammarsten, O. and Elias, P. (1997). Herpes simplex virus: selection of origins of DNA replication. *Nucleic Acids Research*, **25**(9), 1753–1760.

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. *The Journal of Virology*, **64**(12), 6184–6195.

Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under an ROC curve. *Radiology*, **143**, 29–36.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*. New York: Chapman and Hall.

Henderson, D.A. (1999). The looming threat of bioterrorism. *Science*, **283**, 1279–1282.

Hirsch, I., Cabral, G., Patterson, H. and Biswal, N. (1977). Studies on the intracellular replicating DNA of herpes simplex virus type I. *Virology*, **81**(1), 48–61.

Hollander, M. and Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. New York: Wiley.

Hsieh, F. and Turnbull, B.W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Annals of Statistics*, **24**, 25–40.

Hughes, A.L., Irausquina, S. and Friedmana, R. (2010). The evolutionary biology of poxviruses. *Infection, Genetics and Evolution*, **10**(1), 50–59.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.

Hyink, O., Dellow, R.A, Olsen, M.J., Caradoc-Davies, K.M.B., Drake, K., Herniou, E.A., Cory, J.S., OReilly, D.R. and Ward, V.K. (2002). Whole genome analysis of the Epiphyas postvittana nucleopolyhedrovirus. *Journal of General Virology*, **83**, 957–971.

Iyer, L.M., Balaji, S., Koonin, E.V. and Aravind, L. (2006). *Virus Research*, **117**(1), 156–184.

Josse, J., Kaiser, A.D. and Kornberg, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in

deoxyribonucleic acid. *The Journal of Biological Chemistry*, **236**(3), 864–875.

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

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, **11**(7), 283–290.

Komolos, J., Major, P. and Tusnady, G. (1975). An approximation of partial sums of independent RV's and the sample DF.I., *Z.Wahrsch. Werw. Gebiete* Probability Theory and Related Fields **32**, 111–131.

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

Kozhukhin, C.G. and Pevzner, P.A. (1991). Genome inhomogeneity is determined mainly by WW and SS dinucleotides. *Bioinformatics*, **7**(1), 39–49.

Kurtz, S. and Schleiermacher, C. (1999). REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**(5), 426–427.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic acids research*, **29**(22), 4633–4642.

Kurtz, S., Ohlebusch, E., Schleiermacher, C., Stoye, J. and Giegerich, R. (2000).

Computation and visualization of degenerate repeats in complete genomes. *In Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, 228–238.

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 Research*, **55**(1), 39–45.

Lehman, I.R. and Boehmer, P.E. (1999). Replication of herpes simplex virus DNA. *Journal of Biological Chemistry*, **274**(40), 28059–28062.

Leung, M.Y., Choi, K.P., Xia, A. and Chen, L.H.Y. (2005). Nonrandom clusters of palindromes in herpesvirus genomes. *Journal of Computational Biology*, **12**(3), 331–354.

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

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

Lewin, B. (2004). *Gene VIII*. Pearson Prentice Hall.

Li, W. (2001). DNA segmentation as a model selection process. *Proceedings of the fifth annual international conference on computational biology*, 204–210.

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. *Journal of Virology*, **77**(10), 5578–5588.

Lobry, J.R. (1996). A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**(5), 323-326.

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 Research*, **32**(13), 3781–3791.

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. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(12), 5246–5250.

Miller, S. E. (2003). Bioterrorism and electron microscopic differentiation of poxviruses from herpesviruses: dos and don'ts. *Ultrastruct Pathol*, **27**, 133–140.

Mizraji, E. and Ninio, J. (1985). Graphical coding of nucleic acid sequences. *Biochimie*, **67**, 445–448.

Moss, B. (2001). Poxviridae: The viruses and their replication. In: *Fields Virology*, Fourth Edition (D.M. Knipe and P.M. Howley, eds), 2849-2883. Philadel-

phia: Lippincott Williams and Wilkins.

Newcomb, W.W., Juhas, R.M., Thomsen, D.R., Homa, F.L., Burch, A.D., Weller, S.K. and Brown, J.C. (2001). The UL6 Gene Product Forms the Portal for Entry of DNA into the Herpes Simplex Virus Capsid. *Journal of Virology*, **75**(22), 10923–10932.

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.

Nguyen, H.K., Bonfils, E., Auffray, P., Costaglioli, P., Schmitt, P., Asseline, U., Durand, M., Maurizot, J.C., Dupret, D. and Thuong, N.T. (1998). The stability of duplexes involving AT and/or $G^{4Et}C$ base pairs is not dependent on their AT/$G^{4Et}C$ ratio content. Implication for DNA sequencing by hybridization. *Nucleic Acids Research*, **26**(18), 4249–4258.

Orlova, E.V. (2009). How viruses infect bacteria? *The EMBO Journal*, **28**, 797-798.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press.

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. *Molecular and Cellular Biology*, **5**(8), 1822–1832.

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

Ripley, B. D. (1996). *Pattern recognition and neural networks*. New York: Cambridge University Press.

Rocha, E.P.C. and Blanchard, A. (2002). Genomic repeats, genome plasticity and the dynamics of Mycoplasma evolution. *Nucleic Acids Research*, **30**(9), 2031–2042.

Roizman, B. and Baines, J. (1991). The diversity and unity of *Herpesviridae*. *Comparative Immunology, Microbiology and Infectious Disease*, **14**(2), 63–79.

Roizman, B., Carmichael L.E., Deinhard T.F., De The, G., Nahmias, A.N., Plowright, W., Rapp, F., Sheldrick, P., Takahashi M. and Wolf, K. (1981). Herpesviridae-definition, provisional nomenclature and taxonomy. *Intervirology*, **16**(4), 201–217.

Roy, A., Panigrahi, S., Bhattacharyya, M. and Bhattacharyya, D. (2008). Structure, stability, and dynamics of canonical and noncanonical base pairs: Quantum chemical studies. *Journal of Physical Chemistry B*, **112**(12), 3786–3796.

Russell, G.J. and Subak-Sharpe, J.H. (1977). Similarity of the general designs of protochordates and invertebrates. *Nature*, **266**(5602), 533-536.

Russell, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976). Doublet frequency analysis of fractionated vertebrate nuclear DNA. *Journal of*

Molecular Biology, **108**, 1–23.

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

Schbath, S. (1997). An efficient statistic to detect over- and under-represented words in DNA sequences. *Journal of Computational Biology*, **4**, 189–192.

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 Reports*, **4**(11), 1048–1053.

Sponer, J., Leszczynski, J. and Hobza, P. (1996). Structures and Energies of Hydrogen-Bonded DNA Base Pairs. A Nonempirical Study with Inclusion of Electron Correlation. *The Journal of Physical Chemistry*, **100**, 1965–1974.

Stillman, B. (1996). Comparison of DNA replication in cells from Prokarya and Eukarya. In:M.L. DePamphilis, ed. *DNA Replication in Eukaryotic Cells.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. pp. 435-460.

Stow, N.D. (1982). Localization of an origin of DNA replication within the TR_S/IR_S repeated region of the herpes simplex virus type 1 genome. *The EMBO Journal*, **1**(7), 863–867.

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

Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962). Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *The Journal of Biological Chemistry*, **237**, 1961–1967.

Swartzman, G., Silverman, E. and Williamson, N. (1995). Relating trends in walleye pollock (Theragra chalcogramma) abundance in the Bering Sea to environmental factors. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**, 369-380.

Touchon, M., Nicolay, S., Audit, B., Brodie of Brodie E.B., d'Aubenton-Carafa, Y., Arneodo, A. and Thermes, C. (2005). Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(28), 9836–9841.

Tsai, C.T., Ting, J.W., Wu, M.H., Wu, M.F., Guo, I.C. and Chang, C.Y. (2005). Complete genome sequence of the grouper iridovirus and comparison of genomic organization with those of other iridoviruses. *The Journal of Virology*, **79**, 2021–2023.

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 Neuropathologica*, **89**(1),

105–108.

Vlazny, D.A. and Frenkel, N. (1981). Replication of herpes simplex virus DNA: localization of replication recognition signals within defective virus genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **78**, 742–746.

Watson, J.D. and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, **171**, 737–738.

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. *Molecular and Cellular Biology*, **5**(5), 930–942.

Worning, P., Jensen, L.J., Hallin, P.F., Strfeldt, H.H. and Ussery, D.W. (2006). Origin of replication in circular prokaryotic chromosomes. *Environmental Microbiology*, **8**(2), 353-361.

Wyrick, J.J., Aparicio, J.G., Chen, T., Barnett, J.D., Jennings, E.G., Young, R.A., Bell, S.P. and Aparicio, O.M. (2001). Genome-Wide Distribution of ORC and MCM Proteins in S. cerevisiae: High-Resolution Mapping of Replication Origins. *Science*, **294**(5550), 2357–2360.

Yakovchuk, P., Protozanova, E. and Frank-Kamenetskii, M.D. (2006). Base-stacking and base-pairing contributions into thermal stability of the DNA

double helix. *Nucleic Acids Research*, **34**, 564–574.

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. *The Journal of Virology*, **72**(6), 4989–4996.