HIGH ACCURACY PREDICTION OF HUMAN PAPILLOMAVIRUS TYPES BY STATISTICAL CHAOS REPRESENTATION AND REDUCED DIMENSIONAL QUANTIZATION

Miss Watcharaporn Tanchotsrinon

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Science and Information Technology
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2015

การทำนายชนิดของไวรัสก่อมะเร็งปากมดลูกที่มีความแม่นยำสูงโดยใช้การแทนความอลวนเชิงสถิติและควอนไทเซชันที่ลดมิติ

นางสาววัชราภรณ์ ตันโชติศรีนนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

Thesis Title	HIGH ACCURACY PREDICTION OF HUMAN PAPILLOMAVIRUS TYPES BY STATISTICAL CHAOS REPRESENTATION AND REDUCED DIMENSIONAL QUANTIZATION
By	Miss Watcharaporn Tanchotsrinon
Field of Study	Computer Science and Information Technology
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.
Thesis Co-Advisor	Professor Yong Poovorawan, M.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

.....Dean of the Faculty of Science
(Associate Professor Polkit Sangvanich, Ph.D.)

THESIS COMMITTEE

.....Chairman
(Associate Professor Peraphon Sophatsathit, Ph.D.)

.....Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

.....Thesis Co-Advisor
(Professor Yong Poovorawan, M.D.)

.....Examiner
(Assistant Professor Saranya Maneeroj, Ph.D.)

.....Examiner
(Assistant Professor Suphakant Phimoltares, Ph.D.)

.....External Examiner
(Associate Professor Sartra Wongthanavasu, Ph.D.)

วัชราภรณ์ ตันโชติศรีนนท์ : การทำนายชนิดของไวรัสก่อมะเร็งปากมดลูกที่มีความแม่นยำสูงโดยใช้การแทนความอลวนเชิงสถิติและควอนไทเซชันที่ลดมิติ (HIGH ACCURACY PREDICTION OF HUMAN PAPILLOMAVIRUS TYPES BY STATISTICAL CHAOS REPRESENTATION AND REDUCED DIMENSIONAL QUANTIZATION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ. ดร. ชิดชนก เหลือสินทรัพย์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ศ. นพ. ยง ภู่วรวรรณ, 75 หน้า.

การทำนายชนิดของไวรัสเอชพีวี เป็นแนวทางสำคัญในการพัฒนากลยุทธ์สำหรับการวินิจฉัยโรค การรักษาและการป้องกันเพื่อต่อสู้กับมะเร็งปากมดลูก เริ่มแรก วิธีการสกัดข้อมูล 2 วิธี คือ วิธี ChaosCentroid และวิธี ChaosFrequency ได้ถูกนำเสนอสำหรับใช้ในการทำนายชนิดของไวรัสก่อมะเร็งปากมดลูกจากจีโนม วิธี ChaosCentroid สกัดข้อมูลโครงสร้างของลำดับนิวคลีโอไทด์ย่อยในรูปแบบของเซนทรอยด์ ในขณะที่วิธี ChaosFrequency จะสกัดค่าการกระจายตัวเชิงสถิติของลำดับย่อยในสายจีโนมของไวรัสแทน ในการทดลองนี้ ได้นำตัวจำแนกประเภทกลุ่มของข้อมูล คือ โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น โครงข่ายประสาทเทียมเรเดียลเบสิสฟังก์ชัน เทคนิคเพื่อนบ้านใกล้สุด และเทคนิคเพื่อนบ้านใกล้สุดแบบฟัซซี่ มาใช้ในการทำนายชนิดของไวรัสก่อมะเร็งปากมดลูก ผลการทดลองแสดงให้เห็นว่า ในระหว่างผลลัพธ์ที่ได้จากหลากหลายวิธีที่นำมาใช้เปรียบเทียบกัน ทุกวิธีสามารถให้ประสิทธิภาพการทำนายสูงสุดเหมือนกัน แต่วิธีการที่นำเสนอใช้เวลาในการทำนายชนิดของไวรัสน้อยกว่าวิธีอื่นอย่างมีนัยสำคัญ ต่อมา ได้นำเสนอวิธีการสกัดข้อมูลแบบใหม่อีก 1 วิธีคือ วิธี ChaosPoly สำหรับใช้ในการทำนายชนิดของไวรัสจากบางส่วนของยีน วิธี ChaosPoly จะให้ความสำคัญกับค่าการกระจายตัวของรูปแบบต่าง ๆ ของจุดในแต่ละบริเวณย่อยของการแทนความอลวนเชิงสถิติ ในรูปแบบของพหุนาม เทคนิคเพื่อนบ้านใกล้สุดแบบฟัซซี่ถูกนำมาใช้ในการทำนายชนิดของไวรัสนี้ และจากผลการทดลอง พบว่า วิธี ChaosPoly ให้ประสิทธิภาพที่สูงกว่าวิธี ChaosCentroid และวิธี ChaosFrequency

ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	ลายมือชื่อนิสิต .....
		ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
สาขาวิชา	วิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ	ลายมือชื่อ อ.ที่ปรึกษาร่วม .....
ปีการศึกษา	2558	

5373908623 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
KEYWORDS: HUMAN PAPILLOMAVIRUS / GENOTYPING / CERVICAL CANCER / CHAOS GAME REPRESENTATION / SINGULAR VALUE DECOMPOSITION

WATCHARAPORN TANCHOTSRINON: HIGH ACCURACY PREDICTION OF HUMAN PAPILLOMAVIRUS TYPES BY STATISTICAL CHAOS REPRESENTATION AND REDUCED DIMENSIONAL QUANTIZATION. ADVISOR: PROF. CHIDCHANOK LURSINSAP, Ph.D., CO-ADVISOR: PROF. YONG POOVORAWAN, M.D., 75 pp.

HPV genotyping is a significant approach to provide better diagnosis, medical treatment, and prevention strategies for fighting with cervical cancers. Firstly, ChaosCentroid and ChaosFrequency feature extraction techniques were proposed for HPV genotype prediction from whole genomes. ChaosCentroid captures the structure of nucleotide subsequences in terms of centroid, while ChaosFrequency extracts the statistical distribution of the subsequences along genomes. For predicting systems, multi-layer perceptron, radial basis function, k-nearest neighbor, and fuzzy k-nearest neighbor techniques were deployed. The experimental results showed that all methods yielded the highest prediction performance among the results obtained from several compared methods. But time complexity of the proposed techniques was considerably lower than the comparative alignment method. Secondly, ChaosPoly feature extraction technique was subsequently proposed for HPV genotype prediction from partial coding sequences. For each sub-region, ChaosPoly gives the precedence to the distribution of dot patterns in the chaos game representation in a form of polynomial. The fuzzy k nearest neighbor technique was deployed for identifying the corresponding HPV genotypes. The results showed that ChaosPoly outperforms ChaosCentroid and ChaosFrequency.

Department:	Mathematics and Computer Science	Student's Signature .....
		Advisor's Signature .....
Field of Study:	Computer Science and Information Technology	Co-Advisor's Signature .....
Academic Year:	2015	

ACKNOWLEDGEMENTS

I would like to express my immeasurable appreciation and deep gratitude to Professor Chidchanok Lursinsap and Professor Yong Poovorawan, my advisor and my co-advisor, respectively, for their supports, advices, guidance, encouragement, valuable comments, and provisions that are essential to the accomplishment and success of my dissertation. Without their guidance and persistent help, this dissertation would not have been possible.

I would like to acknowledge Thailand research fund for financial support through the Royal Golden Jubilee Ph.D. program (RGJ).

I would like to express deep gratitude to my dissertation committees and reviewers for the valuable comments and suggestions that led to the refinement of my dissertation.

I would like to express my sincere thanks to Dean Faculty of science, and all teachers and staffs of the Department of Mathematics and Computer science in the faculty of science at Chulalongkorn University, for teaching and providing necessary supports and permissions to conduct my dissertation.

I would like to thank the Advanced Virtual and Intelligent Computing (AVIC) Center for their material support in my research, and I also thank all my colleagues at AVIC Center for a lot of useful suggestion, generosity, and encouragement during my education. In addition, I would like to extend my sincere gratitude to everyone who directly and indirectly supports me in accomplishing this dissertation.

Finally, I would like to express my deep sense of gratitude to my family and my senior for endless supports with constant love, encouragement, and suggestions that motivated me to accomplish this dissertation.

CONTENTS

Page

Page

Page

CHAPTER 1
INTRODUCTION

1.1 Problem Identification

Human papillomaviruses (HPV) [1] are small double-stranded DNA viruses. The genetic sequence of the outer capsid protein L1 is used to differentiate the virus types. At present, there are more than 120 types of Human papillomaviruses that have been identified. Of those, approximately 40 HPV types infect the mucosal epithelium. Then, they are categorized according to their epidemiologic association with cervical cancer: non-oncogenic types (low risk) and oncogenic types (high risk). Infection with low risk HPV types, such as types 6 and 11, can cause benign or low-grade cervical cell abnormalities and genital warts. In contrast, high risk HPV types, such as 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 69, 73, and 82, act as carcinogens in the development of cervical cancer and other anogenital cancers.

Cervical cancer is the second most common cancer significantly causing morbidity and mortality in women worldwide, as claimed in [2]. There is an evidence that high risk HPV types are detected in 99% of cervical cancers. Especially, Type 16 is the cause of approximately 50% of cervical cancers worldwide, and types 16 and 18 together account for about 70% of cervical cancers. Then, HPV types 16 and 18 are responsible for the most HPV-caused cancers, and the infection with high risk HPV types is considered as a necessary factor for the development of cervical cancers. Each genotype of HPV has a different risk level in the cervical cancer. Furthermore, there is a wide variation in genotype distribution in different regions around the world. To better understand the relationship of HPV with carcinogenesis, many countries have investigated the HPV infection among women with cytological status by HPV genotyping methods, as revealed in Switzerland [3], in Italy [4], in Cambodia [5], and in Romania [6].

HPV genotyping is necessary for managing effective medical treatment strategies to patients with persistent HPV infection and for evaluating prevention

strategies to individual patients to be immunized with type-specific HPV vaccines, as elucidated in [7]. With the persistent infection, the risk of a precancerous lesion is in between 10% to 15% with HPV types 16 and 18 but below 3% for all other high risk types combined [8]. Additionally, the relevant diagnosis with cost effectiveness can be done based on epidemiological and prevalence studies from a wide variation in the genotype distribution in different regions around the world. The diversity of virus types and the incidence of multiple infections have made it necessary to develop reliable methods to identify the different genotypes for epidemiological studies and medical treatment.

Currently, there are various kinds of HPV genotyping tests used for detecting the genotypes of Human Papillomavirus, in clinical laboratories. Each genotyping test has focused on a different set of Human Papillomavirus types, and the nucleic acid targets and sizes are also various in the tests. The DNA target of PapilloCheck by Greiner Bio One is E1 gene with 350 base pairs. The DNA target of Linear Array HPV Genotyping Test by Roche Molecular Systems is L1 gene with 450 base pairs. The DNA target of PCR HPV Typing Set by Takara Bio INC. is E6 and E7 genes with 300 base pairs, etc.

Even though these HPV genotyping tests are beneficial and applicable for HPV diagnosis in patients nowadays, their some limitations should be considered. For instance, the HPV genotypes are hardly detected in some conditions, such as inadequate samples or low amplification signals of some genotypes. Furthermore, contamination with previously amplified material can lead to false positive results. In particular, mistaken classifications can be occurred through cross-reactivity among similar HPV types in the tests based on hybridization [9].

In summary, HPV genotyping can make a great contribution to the following aspects: HPV diagnosis in case of single and multiple infection, more information regarding risk stratification, a better understanding of the relationship of HPV with carcinogenesis, and prevention of the cancer though the development of type-specific vaccines. Consequently, HPV genotyping has become an important approach to fight with cervical cancer. For these reasons, this dissertation emphasized the development of new algorithms for predicting the HPV genotypes. Since the new algorithms were

proposed based on computational method, the problems that could be occurred by the genotyping tests in clinical laboratories can be limited.

Thus, this dissertation has concentrated on the prediction of HPV genotypes from two significant forms of DNA sequences, which are whole genomes and partial coding sequences.

1.2 Literature Review

Initially, discriminating whether the patients have been infected with the high risk types of Human papillomavirus is the most important and urgent aspect for diagnosis and medical treatment. Multiple perspectives were thus proposed to focus on classifying the HPV into high or low risk types. For instance, Wang and Xiao [10] presented multitudinous physicochemical and statistical features from the protein sequences using Fuzzy K nearest neighbor classifier for the risk type prediction of Human papillomaviruses. At the same year, they also subsequently developed the better algorithm based on geometric moments of protein distance matrix images using a Fuzzy K nearest neighbor classifier [11]. In addition, classification of HPV risk types was also proposed through algorithms based on decision tree [12], text mining [13], genetic mining of DNA sequence structures [14], support vector machines [15], gap-spectrum kernels [16], and ensemble support vector machines with protein secondary structures [17].

While the classification of HPV risk types is the urgent aspect for a diagnosis of cervical cancer as claimed by many researchers, the study on how to predict specific genotypes of this virus has not been significantly focused. In fact, an identification of HPV genotypes in infected patients is more essential than a rough classification of HPV risk types, as previously mentioned.

Chaos Game Representation (CGR) was proposed as a unique and scale-independent representation for genomic sequences by Jeffrey [18]. It is an iterative mapping technique assigning each nucleotides in a DNA or amino acids in a protein to a unique coordinates in a 2-dimensional space. It can be viewed as a 2-dimensional

image of distributed dots and captured in a form of 0-1 square matrix, where 1 represents a dot and 0 represents an empty coordinate. The distribution of positions has two properties of uniqueness and possibility to inverse a coordinate back to its corresponding nucleotide or amino acid [19]. Using graphic approaches to study biological systems can provide useful intuitive insights, as indicated by many previous studies on a series of important biological topics, such as DNA [20, 21], RNA [22], genome [23-27], protein [28-36], drug metabolism systems [37], protein-protein interactions [38], and analysis of protein sequence evolution [39].

Singular value decomposition (SVD) is a matrix factorization technique with various applications. For instance, it can be used to solve underdetermined and overdetermined systems of linear equations, find inverse and the pseudo-inverse matrices, compute the matrix condition number and calculate the vector system orthogonality and orthogonal complement [40]. SVD is also applied to several areas in gene expression data and microarray data, such as analysis [41-44], search [45], image compression [46], and classification [47, 48], etc.

Therefore, the new feature extraction techniques for the HPV genotype prediction were proposed by adapting Chaos game representation for utilizing its interesting properties of a unique and scale-independent representation of genomic sequences. Likewise, singular value decomposition (SVD) was deployed to reduce the size of CGR into a smaller number of feature matrices without losing any knowledge from the original data, for reducing the time complexity.

1.3 Problem Formulation

Problem: How can the algorithm classify the specific types of Human papillomaviruses?

Even though several researches in computational biology paid attention to classify the high and low risk types of Human papillomaviruses, those did not give precedence to further predict the specific types of the viruses. In contrast, in case of single or multiple type infection, recognizing the infected virus types in patients is essential for better medical diagnosis, treatment, and prevention. Thus, the main objective of this dissertation is to propose the algorithm for classifying the specific types of Human papillomaviruses.

Problem: How can the algorithm classify under independent lengths of viral genomes?

Some techniques of feature extraction can be conducted under some limitation, such as an equal length of all sequences in the experiment. The limitation is a problem in a classification, since viral genomes have different lengths even in the same type due to its mutation. To handle the limitation, the feature extraction without consideration of viral length should be used in the proposed algorithm to deal with independent lengths of viral genomes.

Problem: How can the optimum representative of the chaos representation be found out?

In the proposed algorithm, viral genome is transformed to coordinates by the statistical chaos representation, and the representation is further divided into several specified grids. Each grid has its own characteristics. Then, various kinds of aspects will be determined to represent the characteristics of all coordinates in each grid. Therefore, the appropriate representation will be investigated to find out the one that can optimize the prediction algorithm.

Problem: How can the optimum number of grids be identified?

In addition, the prediction performance in the preliminary results has shown that the accuracy values obtained from the proposed algorithm are different, when the number of dimensions in the input vectors has changed. Since the number of dimensions in the input vectors depends on the number of grids divided into the chaos representation, the optimum number of grids should be identified.

1.4 Research Objectives

The main objectives of this dissertation are as follows.

1. To predict the specific types of human papillomaviruses.
2. To reduce time complexity of the prediction algorithm.
3. To improve the prediction algorithm to deal with independent lengths of viral genomes.

1.5 Scopes of the Work

The scopes of this dissertation are constrained on the following issues.

1. The high risk types of Human papillomaviruses in this experiment are restricted according to hybrid capture 2, which is a nucleic acid hybridization assay for detection of human papillomavirus. Then, the data sets of high risk Human papillomaviruses of this experiment include 13 types, which are 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, and 68.
2. The data set of Human Papillomaviruses was collected from National Center for Biotechnology Information web site.
3. Our Prediction algorithm was proposed without using any direct comparison methods such as an alignment technique.

CHAPTER 2
THEORETICAL BACKGROUND

2.1 Human Papillomavirus

2.1.1 Physical Properties and Genome Organization

Papillomaviruses are small non-enveloped icosahedral viruses of approximately 50–60 nm in diameter that contain a circular double stranded DNA genome of approximately 7000–8000 base pairs (bp). The HPV genome can be divided into three functional regions: the early (E) region that encodes proteins (E1–E7) required for viral gene expression, replication and survival; the late (L) region that encodes the viral structural proteins (L1–L2) required for virion assembly; and the long control region (LCR) that is a largely non-coding part required for regulating viral gene expression and replication. The designations E and L refer to the phase in the viral life cycle when these proteins are first expressed. The genome organization of Human Papillomavirus type 16 is demonstrated in Figure 2-1, as below.

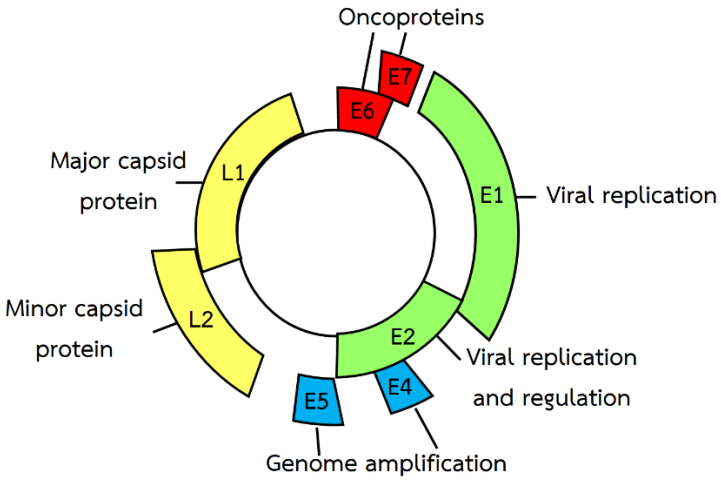

Figure 2-1 Genome Organization of Human Papillomavirus Type 16.

2.1.2 Functions of viral proteins

For Human Papillomavirus, the functions of the proteins are discussed and summarized in Table 2-1.

Table 2-1 Functions of proteins in Human Papillomavirus

Protein	Function
E1	- Adenosine triphosphatase (ATPase) and DNA helicase; - Recognizing and binding to the viral origin of DNA replication as a hexameric complex; - Necessity for viral DNA replication.
E2	- Main regulator of viral gene transcription; - Binding the viral transcriptional promoter as a dimer; - Involvement in viral DNA replication; - Interacting with and recruits E1 to the origin.
E4	- Acting late in the viral life cycle; - Interacting with the keratin cytoskeleton and intermediate filaments; - Localizing to nuclear domain 10; induces G2 arrest; - Being believed to facilitate virus assembly and release.
E5	- Inducing unscheduled cell proliferation; - Interacting with 16k subunit c of vacuolar ATPase; - Activating growth factor receptors and other protein kinases; - Inhibiting apoptosis; - Inhibiting traffic of major histocompatibility complexes to the cell surface.
E6	- Inducing DNA synthesis; induces telomerase; - Preventing cell differentiation; interacts with four classes of cellular proteins: transcriptional co-activators, proteins

	- Involvement in cell polarity and motility, tumour suppressors and inducers of apoptosis, primarily p53, and DNA replication and repair factors.
E7	- Inducing unscheduled cell proliferation; - Interacting with histone acetyl transferases; - Interacting with negative regulators of the cell cycle and tumour suppressors, primarily p105Rb.
L1	- Major viral structural protein; - Assembling in capsomeres and capsids; - Interacting with L2; - Interacting with cell receptor(s); - Encoding neutralizing epitopes.
L2	- Minor viral structural protein; interacts with DNA; - Interacting with nuclear domain 10s; - Being believed to facilitate virion assembly; - Interacting with cell receptor(s); - Encoding linear virus neutralizing epitopes.

2.2 Chaos Game Representation

Chaos Game Representation (CGR) was proposed as a unique and scale-independent representation for genomic sequences by Jeffrey [18]. It is an iterative mapping technique that each nucleotide in a DNA or each amino acid in a protein is assigned to a unique coordinates in a 2-dimensional space. Therefore, the distribution of positions in this representation has two main properties: uniqueness and possibility to inverse a coordinate back to its corresponding nucleotide or amino acid. Besides, this representation also has other interesting properties, as discussed below.

Properties of the CGR of a DNA Sequence

1. The k-th point plotted on the CGR of a sequence corresponds to the first k-long initial subsequence of the whole sequence. Thus, there is a one-to-one correspondence between the subsequences of a DNA and points of the CGR.

2. Therefore, any visible pattern in the CGR corresponds to some pattern in the nucleotide sequence.

3. Any portion of the picture may be magnified for revealing finer structure. Thus, if there is an area of interest in which suspected structure is obscured, it can be magnified to show the fine structure of the points. Therefore, it can reveal the structure of the sequences yielding the points.

4. Adjacent nucleotides in the sequence are not plotted adjacent to each other, except when the first point is close to a corner and the next nucleotide is the same. Being close in the CGR does not mean being close in the sequence.

5. In general, two close points may correspond to different sequences.

2.3 Singular Value Decomposition

Singular value decomposition (SVD) is a matrix factorization technique which has a beneficial property in reducing the size of data into a smaller number of features without losing any knowledge from the original data. The SVD theorem states as follows.

$$A_{m\times n} = U_{m\times m}\, S_{m\times n}\, V^{T}_{n\times n} \qquad \textit{Equation 2.1}$$

where $A_{m\times n}$ is a m-by-n matrix, $U_{m\times m}$ is an m-by-m unitary matrix, $S_{m\times n}$ is an m-by-n diagonal matrix with nonnegative real numbers on the diagonal and the diagonal elements s_i of $S_{m\times n}$ are the singular values of $A_{m\times n}$, and $V^{T}_{n\times n}$ is a conjugate transpose of $V_{n\times n}$, the n-by-n unitary matrix V.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 HPV Genotype Prediction from Genomes

The genomes of Human papillomavirus were collected from genotypes and their features were extracted by the proposed feature extraction techniques, i.e. ChaosCentroid and ChaosFrequency, as inputs for classification. These features were divided into the training and testing sets by a 2-fold cross validation technique. Accordingly, four different classification models were deployed to train and test the experimental data sets. Then, the prediction performance from the obtained results were evaluated and compared with a related method.

3.1.1 HPV Genome Data set

In this experiment, the collected HPV genotypes are those important genotypes detectable by Linear Array® HPV Genotyping Test. This HPV genotyping is a widely used qualitative test developed by Roche Molecular Diagnostics for detecting HPV genotypes associated with cervical cancer. This test can detect 37 high and low risk HPV genotypes, including those considered as a significant risk factor for HSIL progression to cervical cancer. To challenge the prediction, only HPV genotypes having genome diversity were concentrated in this experiment. Some of 37 genotypes containing few genomes were excluded. For this reason, only HPV genotypes 6, 11, 16, 18, 31, 33, 35, 45, 52, 53, 58 and 66 were involved. The genomes of these HPV genotypes were collected from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The HPV genome data set contains Human Papillomavirus genomes of 12 genotypes, including high, possible high, and low risk types. Table 3.1 demonstrates the number of genomes, the minimum and maximum genome lengths for each genotype of Human papillomavirus.

Table 3-1 The number of genomes, minimum and maximum genome lengths of HPV genotypes in the HPV genome data set.

HPV Genotypes	No. of genomes	Genome length (base pairs)	
		Minimum	Maximum
6	58	7954	8051
11	49	7931	10424
16	103	7881	7976
18	19	7824	7857
31	23	7878	7945
33	22	7830	7912
35	28	7820	7908
45	12	7841	7858
52	22	7933	7974
53	16	7856	7863
58	37	7814	7836
66	11	7816	7824

3.1.2 The Proposed Feature Extraction

ChaosCentroid and ChaosFrequency, were proposed to extract the features from the chaos game representation of HPV genomes. Therefore, the relations among subsets of HPV genomes must be clarified in order to identify an individual genotype. These relations are actually the local features. Since the CGR captures the information of the whole genome data, extracting the global features from the CGR may not be efficient enough to distinguish the HPV genotypes. In contrast, the local features hidden in various sub-regions of CGR must be more contemplated. Consequently, this research concentrated on extracting the local features rather than global features. The difference between ChaosCentroid and ChaosFrequency are the feature representation.

Prior to the discussion of ChaosCentroid and ChaosFrequency, the detail of how to construct the chaos game representation (CGR) is the following. Let x_i and y_i

be the coordinates of nucleotide η_i at the i^{th} position in the nucleotide sequence. Algorithm 3.1 illustrates how to construct a CGR for capturing a given nucleotide sequence.

Algorithm 3.1 Constructing Chaos Game Representation

1. Create a square with each corner representing Adenine (A), Cytosine (C), Guanine (G), and Thymine (T) at coordinates (-1,-1), (-1,1), (1,1), and (1,-1), respectively.
2. Case η_1 is
3. A: Place a dot at $x_1 = 0.5 \times (0 - 1)$; $y_1 = 0.5 \times (0 - 1)$.
4. C: Place a dot at $x_1 = 0.5 \times (0 - 1)$; $y_1 = 0.5 \times (0 + 1)$.
5. G: Place a dot at $x_1 = 0.5 \times (0 + 1)$; $y_1 = 0.5 \times (0 + 1)$.
6. T: Place a dot at $x_1 = 0.5 \times (0 + 1)$; $y_1 = 0.5 \times (0 - 1)$.
7. EndCase
8. For each other nucleotide η_i; $i > 1$ do
9. Case η_i is
10. A: Place a dot at $x_i = 0.5 \times (x_{i-1} - 1)$; $y_i = 0.5 \times (y_{i-1} - 1)$.
11. C: Place a dot at $x_i = 0.5 \times (x_{i-1} - 1)$; $y_i = 0.5 \times (y_{i-1} + 1)$.
12. G: Place a dot at $x_i = 0.5 \times (x_{i-1} + 1)$; $y_i = 0.5 \times (y_{i-1} + 1)$.
13. T: Place a dot at $x_i = 0.5 \times (x_{i-1} + 1)$; $y_i = 0.5 \times (y_{i-1} - 1)$.
14. EndCase

The chaos game representation (CGR) can be viewed as a square whose corners are at coordinates (-1, -1), (-1, 1), (1, 1), and (1, -1) representing A, C, G, and T nucleotides, respectively. Note that the size of CGR according to the coordinates of A, C, G, and T nucleotides is equal to 2 × 2 units. However, this unit size of original CGR is not appropriate for discussing the proposed algorithms. Therefore, the geometrical structure and the physical size of this CGR are re-defined as follows. The size of CGR square is set to $n \times n$ and $n \in R^+$. Its center is also located at the coordinates (0, 0). Each corner of this square represents the same nucleotide as that of the original CGR. After Algorithm 3.1, CGR can be viewed as an image of distributed dots. Figure 3.1 shows some examples of CGR of HPV genotypes 6, 16, 18, and 31.

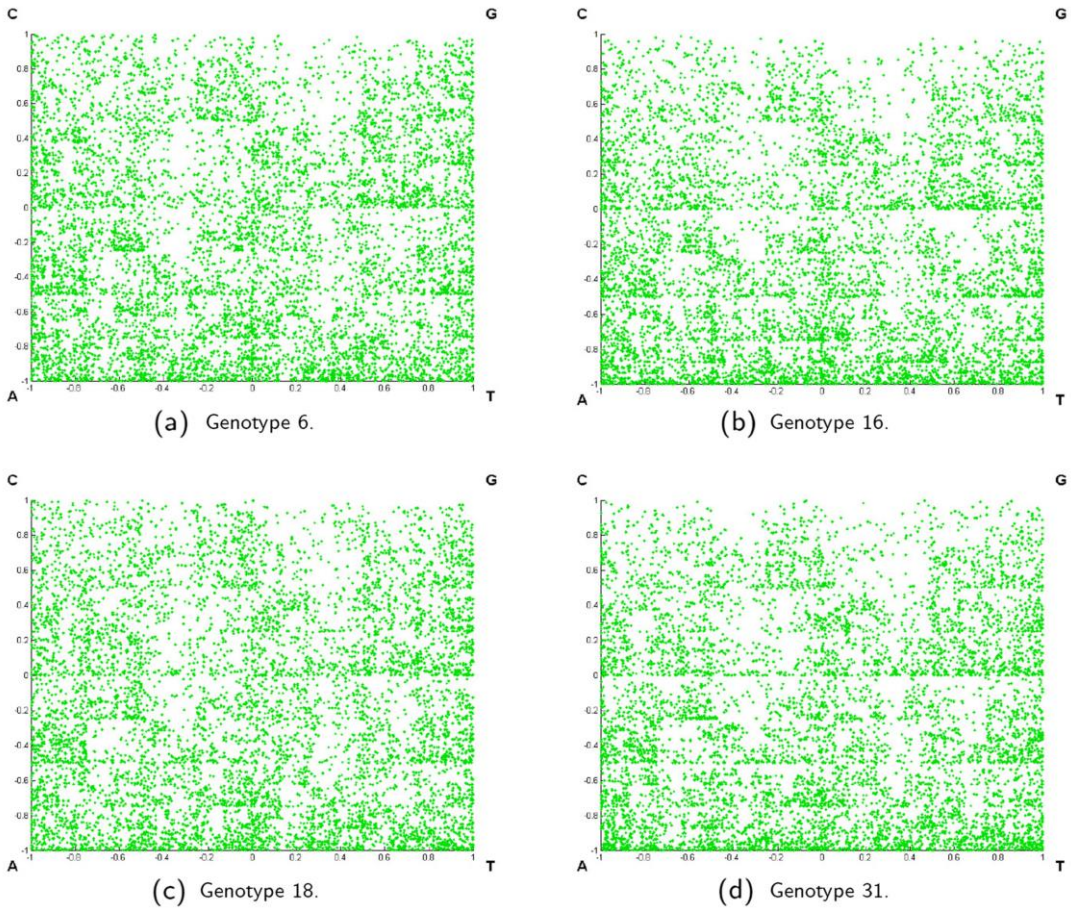

Figure 3-1 Chaos game representation (CGR) of HPV genotypes 6, 16, 18, and 31. (a) Genotype 6. (b) Genotype 16. (c) Genotype 18. (d) Genotype 31.

Obviously, the number of dots in a CGR is equal to the number of nucleotides in a given HPV sequence. Although this CGR image can be directly used in the prediction process, its computational time may be too high due to the large number of dots. Thus, it is necessary to extract only those relevant features from this set of dots to reduce the computational time complexity in the prediction process. In this dissertation, two different features as the representation of CGR image are proposed. The first feature is called ChaosCentroid and the second one is called ChaosFrequency. The detail of each feature is as follows.

3.1.2.1 ChaosCentroid

According to [18], the k-th dot plotted on the CGR of a sequence corresponds to the first k-long initial subsequence of the sequence. Therefore, any visible pattern of the CGR corresponds to some pattern of the nucleotide sequence. CGR represents the global information of the nucleotide sequence. Thus, partitioning the CGR into several sub-regions is implemented for revealing local information of the interested areas. If two dots are within the same quadrant, they correspond to sequences with the same last mononucleotide. But if they are in the same sub-quadrant, the sequences have the same last dinucleotides; and so on. This can demonstrate the structure of the sequences yielding the dots. ChaosCentroid utilizes this biological significance by computing the centroid of the distributed dots of each sub-region. Therefore, the centroid, which can be converted to specific structure of the sequence, is represented as local information of the sub-region.

For ChaosCentroid, the CGR is partitioned into $\frac{n}{g} \times \frac{n}{g}$ equal sub-regions, where $\frac{n}{g}$ is an element of positive integers ranging from 1 to 11. This range is derived from all possible numbers that can applied to the CGR. For instance, the CGR is not partitioned when $\frac{n}{g} = 1$, the CGR is partitioned into 4 equal sub-regions when $\frac{n}{g} = 2$, and so on. Furthermore, if the value of $\frac{n}{g}$ is greater than 11, some sub-regions do not contain any dots. So, 11 is the maximum value of $\frac{n}{g}$ in this experiment. For each of $\frac{n}{g}$ partitioned into the CGR, the centroid of each sub-region is computed first. Then all pairs of distances between the centroids and the center of CGR are computed and captured in a form of a matrix. This set of distances can be considered as the relation of information embedded in all sub-regions. However, the number of ChaosCentroids may be too large. Therefore, this matrix is decomposed by applying singular value decomposition (SVD) method to reduce information complexity. Finally, the $\frac{n}{g}$ diagonal elements from the $\frac{n}{g}$ - by - $\frac{n}{g}$ diagonal matrix of SVD are represented as the features of CGR and are subsequently used as the input vectors for prediction process. As a result, ChaosCentroid produces 11 formats of input vectors according to the value of $\frac{n}{g}$, i.e. the first format has 1 dimension, the second format has 2 dimensions, and so

on. Extracting ChaosCentroid consists of the following steps, as illustrated in Algorithm 3.2.

Algorithm 3.2 Extracting ChaosCentroid Feature

1. Represent the HPV genomes by chaos game representation (CGR) of size n × n.
2. Partition CGR into $\frac{n}{g} \times \frac{n}{g}$ equal sub-regions, each of size $g \times g$.
3. Let $r_{i,j}$ be the CGR region at row $1 \leq i \leq \frac{n}{g}$ and column $1 \leq j \leq \frac{n}{g}$.
4. Let $|r_{i,j}|$ be the number of dots in $r_{i,j}$.
5. **For** each sub-region $r_{i,j}$ **do**
6. Compute the centroid $c_{i,j} = \left(\frac{\sum_{k=1}^{|r_{i,j}|} x_k}{|r_{i,j}|}, \frac{\sum_{k=1}^{|r_{i,j}|} y_k}{|r_{i,j}|} \right)$.
7. **EndFor**
8. Compute a distance matrix $\mathbf{D} = \left[d_{i,j} \right]_{\frac{n}{g} \times \frac{n}{g}}$; $d_{i,j} = \left\| c_{i,j} \right\|$.
9. Let $\mathbf{S} = \left[s_{i,j} \right]_{\frac{n}{g} \times \frac{n}{g}}$ be the diagonal matrix of $\mathbf{D}$ computed by applying singular value decomposition.
10. Form vector $\mathbf{F} = \left[s_{i,i} \right]_{1 \leq i \leq \frac{n}{g}}^{T}$ as the feature of CGR.

A time complexity of ChaosCentroid algorithm is analyzed as follow.

To simplify this analysis, all HPV genomes will be considered as the nucleotide sequences of equal length L. Let $T(N)$ be the running time of ChaosCentroid on extracting features from N genomes of length L. Initially, constructing the CGR of N genomes takes $O(NL)$ time. Since the values of $\frac{n}{g}$ are set to constant number, partitioning the CGR into $\frac{n}{g} \times \frac{n}{g}$ equal sub-regions takes constant time. Hence, partitioning the CGRs of N genomes takes $O(N)$ time. Computing centroids in sub-regions takes $O(NL)$ time, and computing Euclidean distances in two dimensional spaces takes $O(N)$ time. Although extracting the singular value decomposition of $\frac{n}{g} - \text{by} - \frac{n}{g}$ matrices take $O\left(\left(\frac{n}{g}\right)^3\right)$ time, the maximum value of $\frac{n}{g}$ is set to 11 in this experiment. So, the SVD of N genomes takes $O(11^3 N)$ time. Finally, forming the feature vectors also takes $O(N)$ time. Therefore, the total time complexity of ChaosCentroid feature extraction is as follow.

$$T(N) = CGR + Partition + Centroid + Distance + SVD + Form\ vector$$

$$= O(NL) + O(N) + O(NL) + O(N) + O(11^3 N) + O(N)$$

$$= O(NL)$$

Additionally, Figure 3.2 shows an example of distances between the centroid of each sub-region and the center of CGR for HPV genotype 16 after being partitioned into sub-regions of size 2 × 2.

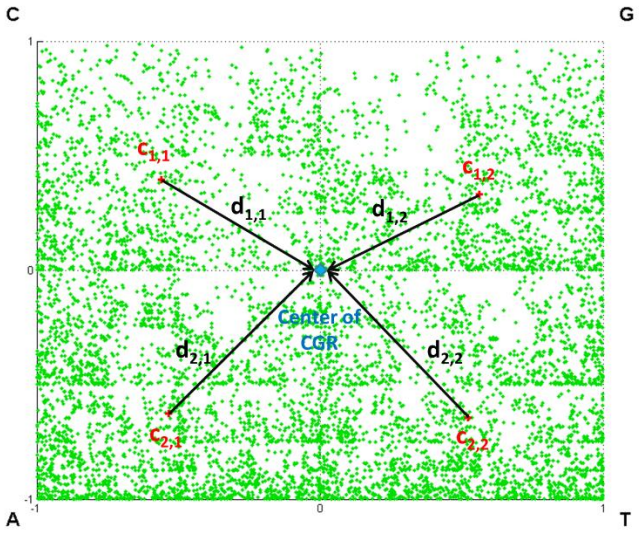

Figure 3-2 The distances between the centroids and the center of CGR for HPV genotype 16 after being partitioned into sub-regions of size 2 × 2.

3.1.2.2 ChaosFrequency

As elucidated in [20], the bias of distribution of different mono-, di-, tri-, or higher order nucleotides along the DNA/RNA sequences can generate different patterns in the CGR. This can be used as diagnostic patterns for different HPV genotypes. The CGRs of the HPV genomes of different genotypes tend to exhibit distinct patterns visually, as displayed in Figure 3.1. Therefore, ChaosFrequency has concentrated on the frequencies of subsequences occurred in the HPV genomes. Particularly, when $\frac{n}{g}$ is equal to 2^k for any positive integer $k \leq 3$, it represents the k-mer frequency occurred in the HPV sequences.

Accordingly, the ratio between the number of dots in the sub-region and the total number of dots in the CGR are computed and represented as the feature of each sub-region. This ratio can be interpreted as the *probability of distribution*. Suppose each sub-region is of size g × g. After extracting the ChaosFrequency of each sub-region, the whole CGR be viewed as a matrix of size $\frac{n}{g} \times \frac{n}{g}$. Likewise, this frequency matrix is decomposed by SVD to extract the $\frac{n}{g}$ diagonal elements used as the feature of CGR. Then this technique also produces 11 formats of input vectors, in accordance with those of ChaosCentroid. The detail of this proposed feature extraction technique is illustrated in Algorithm 3.3. Each sub-region is referred by its location according to the row and column after the partition of CGR. Let $m_{i,j}$ be the number of dots in sub-region at row *i* and column *j*. Suppose there are total M dots in CGR. Then the probability of distribution can be computed as $p_{i,j} = \frac{m_{i,j}}{M}$.

Algorithm 3.3 Extracting ChaosFrequency Feature

1. Represent the HPV genomes by chaos game representation (CGR) of size *n* × *n*.
2. Partition CGR into $\frac{n}{g} \times \frac{n}{g}$ equal sub-regions, each of size $g \times g$.
3. Let $r_{i,j}$ be the CGR region at row $1 \leq i \leq \frac{n}{g}$ and column $1 \leq j \leq \frac{n}{g}$.
4. **For** each sub-region $r_{i,j}$ **do**
5. Compute the probability of distribution $p_{i,j} = \frac{m_{i,j}}{M}$.
6. **EndFor**
7. Form matrix $\mathbf{D} = \left[d_{i,j}\right]_{\frac{n}{g} \times \frac{n}{g}}$; $d_{i,j} = p_{i,j}$.
8. Let $\mathbf{S} = \left[s_{i,j}\right]_{\frac{n}{g} \times \frac{n}{g}}$ be the diagonal matrix of $\mathbf{D}$ computed by applying singular value decomposition.
9. Form vector $\mathbf{F} = \left[s_{i,i}\right]^{T}_{1 \leq i \leq \frac{n}{g}}$ as the feature of CGR.

Likewise, all HPV genomes will be considered as the nucleotide sequences of equal length L. Let $T(N)$ be the running time of ChaosFrequency on extracting features from N genomes of length L. Initially, constructing the CGR of N genomes takes $O(NL)$ time, and partitioning the CGRs of N genomes takes $O(N)$ time. Then, computing probability of distribution in sub-regions takes $O(NL)$ time. Forming matrices D takes $O(N)$ time, extracting SVD takes $O(11^3 N)$ time, and forming the feature vectors takes $O(N)$ time. Therefore, the total time complexity of ChaosFrequency is as follows.

$$T(N) = CGR + Partition + Probability + Form\ matrix + SVD + Form\ vector$$

$$= O(NL) + O(N) + O(NL) + O(N) + O(11^3 N) + O(N)$$

$$= O(NL)$$

3.1.3 Predicting Systems

For evaluating the performance of the proposed feature extraction techniques, the testing sets were fed to four different types of predicting systems. Each system has its own principle and criteria for predicting the corresponding HPV genotypes. The predicting systems are multi-layer perceptron neural network, radial basis function network, k-nearest neighbor technique, and fuzzy k-nearest neighbor technique. For each of 400 HPV genomes, one of 12 genotypes which are types 6, 11, 16, 18, 31, 33, 35, 45, 52, 53, 58, and 66 was identified. The following is a detail of set-up for each predicting system in this experiment.

3.1.3.1 Multi-layer Perceptron Neural Network

After extracting features by Algorithms 3.2 and 3.3, each feature vector F was used as an input vector for multi-layer perceptron neural network. Therefore, the numbers of input neurons correspond to the sizes of the feature vector F, which are in the range of 1 - 11 dimensions. The number of hidden neurons was empirically varied from 1 to 24 neurons to find the most suitable number. From the experiments, 16 hidden neurons are the best number of neurons for producing the best prediction of HPV genotypes. Additionally, there are 12 output neurons, each of which corresponds to each HPV genotype. To make the testing efficient, the neuron 1 is for determining HPV genotype 6; neuron 2 for type 11; neuron 3 for type 16; neuron 4 for type 18; neuron 5 for type 31; neuron 6 for type 33; neuron 7 for type 35; neuron 8 for type 45; neuron 9 for type 52; neuron 10 for type 53; neuron 11 for type 58; and neuron 12 for type 66. As a result, the network deployed in these experiments consists of an input layer with $\frac{n}{g}$ neurons, a hidden layer with 16 neurons, and an output layer with 12 neurons. Furthermore, a backpropagation learning rule was adopted to adjust the weights of the network during the training process. Mean squared normalized error function was used as a terminating criterion in the training process.

In testing procedure, the predicted HPV genotype is determined by an Equation (3.1). Let o_i be the output value of output neuron i.

$$\text{HPV genotype} = \text{argtype} \max_{1 \le i \le 12} (o_i) \quad \textbf{(3.1)}$$

where argtype is the mapping from neuron index to its corresponding HPV genotype previously defined.

3.1.3.2 Radial Basis Function Network

For this predicting system, a spread distance was empirically varied from 0.1 to 1 with an interval of 0.1, in order to find the optimal distance that can yield the maximum average accuracy of all input dimensions. For each feature extraction technique, the optimal spread distances were subsequently set to the prediction systems based on radial basis function (RBF), i.e., 0.4 for ChaosCentroid and 0.1 for ChaosFrequency. The same network structure of multi-layer perceptron was adopted for this RBF network. The determination in Equation 3-1 of HPV genotypes for multi-layer perceptron was also used in this RBF predicting system.

3.1.3.3 K-nearest Neighbor Technique

In this technique, the determination of HPV genotypes depends upon the value of k nearest neighbors measured by Euclidean distance. For any tested feature vector, the HPV genotype of its nearest neighbor is assigned as the HPV genotype of the tested feature vector. Empirically, it was found that k = 1 gave the best performance in this experiment.

3.1.3.4 Fuzzy K-nearest Neighbor Technique

Fuzzy k-nearest neighbor technique was proposed by James M. Keller, Michael R. Gray, and James A. Givens [49]. It is a special variation of the k-nearest neighbor technique family. The algorithm of fuzzy k-nearest neighbor assigns class membership to a sample vector rather than assigning the vector to a particular class. An advantage is that no arbitrary assignments are made by the algorithm. Additionally, membership values of the vector should provide a level of assurance to accompany the resultant

classification. For this technique, k was also set to 1, which is similar to the previous k-nearest neighbor technique.

3.1.4 Performance Evaluation

Two-fold cross-validation technique was adopted in this experiment for evaluating the performance of HPV genotype prediction based on the proposed feature extraction techniques, i.e. ChaosCentroid and ChaosFrequency, with the different four predicting systems. Then, the reported prediction performance was obtained by the combination of both validating sets accordingly.

In this experiment, Equation 11 of [50] is adopted to formulate the set of four metrics, including Sensitivity (Sen), Specificity (Spec), Accuracy (Acc), and Matthew's Correlation Coefficient (MCC), for evaluating the prediction performance. The formulation of the four metrics is defined by the following equations.

$$Sensitivity = 1 - \frac{N_{-}^{+}}{N^{+}}, \quad 0 \leq Sen \leq 1 \tag{3.2}$$

$$Specificity = 1 - \frac{N_{+}^{-}}{N^{-}}, \quad 0 \leq Spec \leq 1 \tag{3.3}$$

$$Accuracy = 1 - \frac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}}, \quad 0 \leq Acc \leq 1 \tag{3.4}$$

$$MCC = \frac{1 - \left(\frac{N_{-}^{+}}{N^{+}} + \frac{N_{+}^{-}}{N^{-}}\right)}{\sqrt{\left(1 + \frac{N_{+}^{-} - N_{-}^{+}}{N^{+}}\right)\left(1 + \frac{N_{-}^{+} - N_{+}^{-}}{N^{-}}\right)}}, -1 \leq MCC \leq 1 \tag{3.5}$$

where N^{+} is the total number of HPV genomes of the investigated genotype whereas N_{-}^{+} is the number of HPV genomes of the investigated genotype that is incorrectly predicted as the other genotypes; N^{-} is the total number of HPV genomes of the other genotypes that are not investigated whereas N_{+}^{-} is the number of HPV genomes of the other genotypes that is incorrectly predicted as the investigated genotype. The investigated HPV genotype is 6, 11, 16, 18, 31, 33, 35, 45, 52, 53, 58, or 66. To illustrate

this point, if the investigated genotype is 6, N^+ is the total number of HPV genomes of genotype 6, while N^- is the total number of the genomes of the other genotypes, excluding genotype 6.

From Equation (3.2) to (3.5), the prediction performance can be evaluated in a meaningful explanation, as follows. The sensitivity is used for evaluating the performance of the predicting systems in identifying the investigated genotype. When N_-^+ = 0, none of HPV genomes of the investigated genotype was incorrectly predicted as the other genotypes, so the sensitivity is 1. In contrast, while $N_-^+ = N^+$, all HPV genomes of the investigated genotype were incorrectly predicted as the other genotypes, so the sensitivity is 0. The specificity is used for evaluating the performance of the systems in excluding the other genotypes. When N_+^- = 0, none of HPV genomes of the other genotypes was incorrectly predicted as the investigated genotype, so the specificity is 1; while $N_+^- = N^-$, all HPV genomes of the other genotype were incorrectly predicted as the investigated genotype, so the specificity is 0. The accuracy is used for evaluating the performance of the systems in classifying the investigated genotype and the other genotypes. When $N_-^+ = N_+^- = 0$, none of HPV genomes of the investigated genotype and none of HPV genomes of the other genotypes were incorrectly predicted, so the accuracy is 1; while $N_-^+ = N^+$and $N_+^- = N^-$ all HPV genomes of the investigated genotype and all HPV genomes of the other genotypes were incorrectly predicted, so the accuracy is 0. Typically, the Matthew's Correlation Coefficient (MCC) is used for measuring the quality of binary classification. When $N_-^+ = N_+^- = 0$, none of HPV genomes of the investigated genotypes and none of HPV genomes of the other genotypes were incorrectly predicted, so MCC is 1; when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$, MCC is 0 meaning no better than random prediction; when $N_-^+ = N^+$ and $N_+^- = N^-$, MCC is -1 indicating total disagreement between prediction and observation.

3.2 HPV Genotype Prediction from Partial Coding Sequences

The partial coding sequences of HPV genotypes were collected and their features were extracted by the proposed feature extraction technique, ChaosPoly, as inputs for classification. These features were divided into the training and testing sets by Leave-one-out cross validation technique. They were fed to the predicting system based on fuzzy k nearest neighbor technique for the corresponding HPV genotypes. Then, the prediction performance of ChaosPoly feature extraction technique was evaluated and compared with those of ChaosCentroid and ChaosFrequency feature extraction techniques.

3.2.1 HPV Partial Coding sequence data set

For the HPV genotyping tests used in clinical laboratories, their nucleic acid targets and sizes are various according to companies that develops these medical diagnostic products. These nucleic acid targets can be some regions of genes in viral genomes. To challenge and develop the prediction algorithm, this dissertation attempts to predict the HPV genotype from the DNA fragments, which may be small sizes and incomplete. Therefore, this experiment has paid attention to the HPV genotype prediction from partial coding sequences, which can be considered as incomplete genes. Accordingly, the partial coding sequences of HPV genotypes were collected from the web site of National Center for Biotechnology Information, as the HPV partial coding sequence data set of this experiment. For this data set, the HPV genotypes and the number of sequences in each genotype were revealed in Table 3-2.

Table 3-2 The HPV Partial Coding Sequence Data set.

HPV Genotype	The number of partial coding sequences
6	282
11	135
16	1423
18	135
31	176
33	57
35	26
39	41
42	25
45	21
51	33
52	90
53	145
56	66
58	221
59	34
66	140
68	32
70	20
71	23
81	23

3.2.2 The Proposed Feature Extraction

3.2.2.1 ChaosPoly

Partial coding sequences are incomplete genes having short nucleotide lengths, compared with the lengths of whole HPV genomes. The challenge of this experiment is how to find the optimum features for the prediction in case of less global information. So, extracting the local information from sub-regions of the Chaos game representation must be more contemplated in order to achieve the high performance of the prediction. Then, ChaosPoly feature extraction technique has been proposed for capturing more local information of the sub-regions in polynomial form. A procedure of ChaosPoly feature extraction technique is illustrated in Algorithm 3.4.

For ChaosPoly, the partial coding sequences are transformed into coordinates in the chaos game representation, and the CGRs are partitioned into $2^k \times 2^k$ equal sub-regions. This experiment sets k to 2, 3, and 4 in order to represent the sub-regions corresponding to 2-mer, 3-mer, and 4-mer, respectively. The k-mers refer to all the possible subsequences of length k obtained from DNA sequences. Each sub-region is referred by its location according to the row and column after the partition. Let $m_{i,j}$ be the number of dots in sub-region at row i and column j. Suppose there are total M dots in CGR. Then, the probability of distribution in sub-region at row i and column j can be computed as $p_{i,j} = \frac{m_{i,j}}{M}$. All of them are formed matrix P, where $\mathrm{P} = [p_{i,j}]_{2^k \times 2^k}$. The matrices P are subsequently partitioned into $2^{k-s} \times 2^{k-s}$ sub-matrices, where $1 \leq s \leq \mathrm{k}-1$. Let P_z be the z^{th} sub-matrix, where z are positive integers arranging in ascending order from left to right and top to bottom, and $1 \leq z \leq 2^{k-s} \times 2^{k-s}$. Each sub-matrix P_z contains $2^s \times 2^s$ elements. In this point of view, a procedure of the partition is depicted in Figure 3-3.

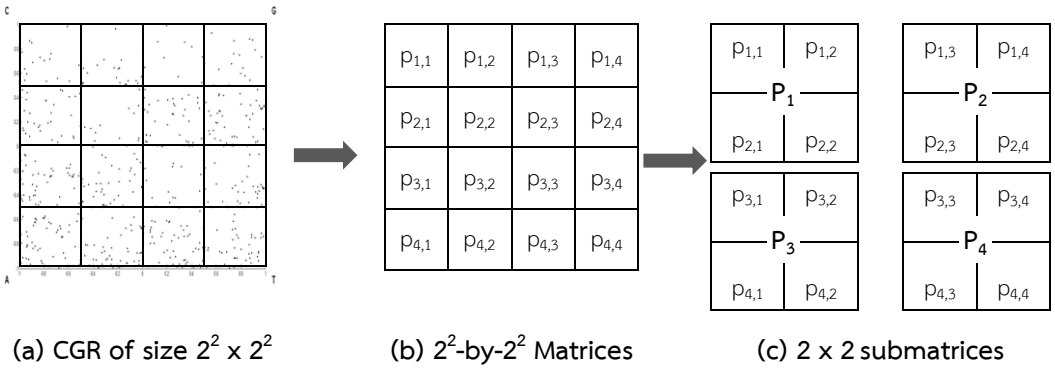

Figure 3-3 A procedure of the partition in ChaosPoly algorithm after setting k to 2, and s to 1. (a) The CGR after being partitioned into sub-regions of size 2^2 x 2^2. (b) The 2^2-by-2^2 Matrix P, each containing $p_{i,j}$, the probability of distribution in sub-region at row i and column j. (c) Partition Matrix P into 4 sub-matrices P_z, where z are positive integers ranging from 1 to 4.

For sub-matrices, the probability of distribution are captured in a form of polynomial. Then, these values of polynomial are represented as the feature of CGR.

Algorithm 3.4 Extracting ChaosPoly Feature

1. Represent the HPV genomes by chaos game representation (CGR) of size *n* × *n*.
2. Partition CGR into $2^k \times 2^k$ equal sub-regions, where $2 \leq k \leq 4$.
3. Let $r_{i,j}$ be the CGR region at row $1 \leq i \leq 2^k$ and column $1 \leq j \leq 2^k$.
4. Let $m_{i,j}$ be the number of dots in sub-region at row i and column j and there are total M dots in CGR.
5. **For** each sub-region $r_{i,j}$ **do**
6. Compute the probability of distribution $p_{i,j} = \frac{m_{i,j}}{M}$.
7. **EndFor**
8. Partition matrix $\mathrm{P} = [p_{i,j}]_{2^k \times 2^k}$ into $2^{k-s} \times 2^{k-s}$ submatrices, where $1 \leq s \leq \mathrm{k}-1$.
9. Let $P_z = [p_{x,y}]_{2^s \times 2^s}$ be the z^{th} submatrix, where z are positive integers arranging in ascending order from left to right and top to bottom, and $1 \leq z \leq 2^{k-s} \times 2^{k-s}$.
10. Let $p_{x,y}$ be the element of submatrix P_z at row $1 \leq x \leq 2^s$ and column $1 \leq y \leq 2^s$.
11. **For** each submatrix P_z **do**
12. Compute the value of polynomial

$$Poly_z = \sum_{x=1}^{2^s} \sum_{y=1}^{2^s} p_{x,y} \times 2^{2^{2s}-[2^s \times (x-1)]-y}$$

13. **EndFor**
14. Form vector $\mathbf{F} = [Poly_z]_{2^{k-s} \times 2^{k-s}}$ as the feature of CGR.

3.2.3 Predicting system

Since the predicting system based on fuzzy k nearest neighbor technique yielded the very high performance with stability in the HPV genotype prediction from whole genomes, this fuzzy technique was adopted for training and testing the HPV partial coding sequence data set. Likewise, k is set to 1 for this technique.

3.2.4 Performance Evaluation

Among the independent statistical accuracy testing methods for predicted results such as sub-sampling (e.g., 2, 5 or 10-fold cross-validation technique) and Leave-one-out cross validation technique, the Leave-one-out technique was deemed the most objective that can always yield a unique result for a given benchmark data set, as elucidated in [51] and demonstrated by Equations 28, 29 and 30 in [51]. Therefore, the Leave-one-out cross validation technique has been increasingly used and widely recognized by investigators to test the power of various prediction methods. Accordingly, this experiment adopted the Leave-one-out cross validation technique for dividing the training and testing sets of the HPV partial coding sequence data set.

Likewise, the set of four metrics, including Sensitivity (Sen), Specificity (Spec), Accuracy (Acc), and Matthew's Correlation Coefficient (MCC), is also adopted for evaluating the prediction performance. The formulation of the four metrics is defined by the Equation (3.2) to (3.5), as previously mentioned.

CHAPTER 4
EXPERIMENTAL RESULTS AND DISCUSSION

4.1 HPV Genotype Prediction from Genomes

The values of variable $\frac{n}{g}$ in Algorithms 3.2 and 3.3 were set from 1 to 11, thus the input vectors were in the range of 1-11 dimensions. For this HPV genome data set, the performance of HPV genotype prediction was separately summarized according to each predicting system and two proposed feature extracting schemes. The obtained experimental results are illustrated and discussed as follows.

4.1.1 Multi-layer Perceptron Neural Network

The experimental results of the HPV genotype prediction gained by ChaosCentroid and by ChaosFrequency feature extraction with the predicting system based on multi-layer perceptron neural network are summarized in Tables 4.1 and 4.2, respectively. These results were reported according to different values of $\frac{n}{g}$, which were in the range of 1-11.

When $\frac{n}{g}$ is equal to 1, the number of sub-regions of CGR is also equal to one. Thus there is only one centroid computed by ChaosCentroid and the probability of distribution of CGR computed by ChaosFrequency is equal to one. The overall performance of ChaosFrequency is much lower than those of ChaosCentroid. ChaosFrequency gain 0% of sensitivity and 100% of specificity in all genotypes, excepting genotype 16. It can be implied that the features of all genomes extracted by ChaosFrequency are totally predicted to genotype 16. In contrast, ChaosCentroid can obtain high performance metrics, including accuracy, sensitivity, specificity, and Matthew's Correlation Coefficient in almost all genotypes. This is because a centroid is computed from the coordinates of every dots. It is obvious that different HPV genotypes must have different distribution of dots and centroids. So, predicting HPV genotypes with high performance from these centroids is possible. But in case of ChaosFrequency, the probability of distribution of every HPV genotype is equal. This makes the feature of each HPV genotype indistinguishable.

In contrast, when the value of $\frac{n}{g}$ is greater than one, the local information regarding the frequency of subsequence among nucleotides in each sub-region is brought out and the performance is increased in proportion to the value of $\frac{n}{g}$. In addition, it is noticeable that there is no significant difference between the overall performance obtained from ChaosCentroid and ChaosFrequency when $3 < \frac{n}{g} \leq 11$. Therefore, it can be concluded that, to achieve high performance of prediction, the local information of each sub-region is more relevant than global information.

Table 4-1 The results of the HPV genotype prediction based on the features extracted by ChaosCentroid with multi-layer perceptron neural network.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	96.50	94.83	96.78	0.87	95.25	86.21	96.78	0.81	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	95.00	89.80	95.73	0.79	99.00	93.88	99.72	0.95
16	94.25	100.00	92.26	0.87	93.50	89.32	94.95	0.83	99.50	100.00	99.33	0.99
18	95.25	15.79	99.21	0.26	97.25	78.95	98.16	0.72	98.50	84.21	99.21	0.83
31	99.50	95.65	99.73	0.95	96.50	73.91	97.88	0.69	100.00	100.00	100.00	1.00
33	94.50	0.00	100.00	NaN	99.50	95.45	99.74	0.95	99.00	90.91	99.47	0.90
35	99.25	92.86	99.73	0.94	98.00	82.14	99.19	0.84	99.75	96.43	100.00	0.98
45	100.00	100.00	100.00	1.00	98.00	50.00	99.48	0.60	99.75	100.00	99.74	0.96
52	100.00	100.00	100.00	1.00	98.00	77.27	99.21	0.80	100.00	100.00	100.00	1.00
53	98.75	100.00	98.70	0.87	94.75	6.25	98.44	0.07	99.00	87.50	99.48	0.87
58	100.00	100.00	100.00	1.00	97.50	83.78	98.90	0.85	99.50	97.30	99.72	0.97
66	100.00	100.00	100.00	1.00	98.75	63.64	99.74	0.74	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	99.25	100.00	98.99	0.98	99.75	99.03	100.00	0.99	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	99.00	86.96	99.73	0.90	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	99.00	90.91	99.47	0.90	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	99.50	92.86	100.00	0.96	99.25	96.43	99.46	0.94	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	99.75	93.75	100.00	0.97	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	99.50	100.00	99.45	0.97	99.50	97.30	99.72	0.97	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.75	97.96	100.00	0.99
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.50	100.00	99.48	0.95
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.75	90.91	100.00	0.95

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	99.50	100.00	99.33	0.99
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	99.75	95.65	100.00	0.98
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	99.75	97.30	100.00	0.99
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

Table 4-2 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency with multi-layer perceptron neural network.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	85.50	0.00	100.00	NaN	100.00	100.00	100.00	1.00	98.25	93.10	99.12	0.93
11	87.75	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	97.96	100.00	0.99
16	25.75	100.00	0.00	NaN	97.75	97.09	97.98	0.94	100.00	100.00	100.00	1.00
18	95.25	0.00	100.00	NaN	99.25	89.47	99.74	0.92	98.00	84.21	98.69	0.79
31	94.25	0.00	100.00	NaN	99.75	95.65	100.00	0.98	100.00	100.00	100.00	1.00
33	94.50	0.00	100.00	NaN	97.75	72.73	99.21	0.77	100.00	100.00	100.00	1.00
35	93.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	97.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	94.50	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	96.00	0.00	100.00	NaN	99.25	93.75	99.48	0.91	100.00	100.00	100.00	1.00
58	90.75	0.00	100.00	NaN	99.75	100.00	99.72	0.99	99.50	100.00	99.45	0.97
66	97.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.50	81.82	100.00	0.90

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.75	97.96	100.00	0.99	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	99.25	100.00	98.99	0.98	99.75	100.00	99.66	0.99
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	99.00	86.96	99.73	0.90	99.75	100.00	99.73	0.98
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	99.75	96.43	100.00	0.98	99.50	92.86	100.00	0.96
45	99.75	100.00	99.74	0.96	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.50	90.91	100.00	0.95
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.50	100.00	99.45	0.97
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

4.1.2 Radial Basis Function Network

The experimental results of the HPV genotype prediction gained by ChaosCentroid and by ChaosFrequency feature extraction with the predicting system based on radial basis function network are summarized in Tables 4-3 and 4-4, respectively. According to these experimental results, the performance values obtained by this predicting system are unstable among input dimensions. This is because this experiment set only one optimal spread distance, which gain the maximum average accuracy of all dimensions, for each predicting system obtaining features extracted by ChaosCentroid and by ChaosFrequency, respectively. In fact, it is possible that each input dimension has its own proper spread distance, and one value of spread distance cannot fit for all input dimensions. Additionally, it is noticeable that ChaosFrequency with RBF at 4-dimensional input can achieve the best performance with minimum input dimension. The overall performance trend obtained from this predicting system is similar to those of multi-layer perceptron neural network. But the performance of multi-layer perceptron neural network is significantly higher than the performance of radial basis function network.

Table 4-3 The results of the HPV genotype prediction based on the features extracted by ChaosCentroid with radial basis function network.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	95.75	100.00	95.03	0.86	83.00	27.59	92.40	0.23	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	90.00	97.96	88.89	0.69	97.50	97.96	97.44	0.89
16	94.00	100.00	91.92	0.86	87.75	94.17	85.52	0.73	98.00	100.00	97.31	0.95
18	95.25	0.00	100.00	NaN	95.25	21.05	98.95	0.30	97.50	47.37	100.00	0.68
31	96.00	30.43	100.00	0.54	96.75	73.91	98.14	0.71	100.00	100.00	100.00	1.00
33	94.50	0.00	100.00	NaN	98.50	95.45	98.68	0.87	97.25	77.27	98.41	0.74
35	96.00	92.86	96.24	0.76	93.00	0.00	100.00	NaN	99.50	96.43	99.73	0.96
45	99.75	100.00	99.74	0.96	98.25	75.00	98.97	0.71	99.00	66.67	100.00	0.81
52	98.00	100.00	97.88	0.85	94.75	4.55	100.00	0.21	100.00	100.00	100.00	1.00
53	98.75	87.50	99.22	0.84	96.00	0.00	100.00	NaN	98.50	68.75	99.74	0.79
58	98.75	100.00	98.62	0.93	92.75	75.68	94.49	0.63	99.25	97.30	99.45	0.96
66	97.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.50	100.00	99.43	0.98	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	99.25	99.03	99.33	0.98	99.00	98.06	99.33	0.97	99.50	99.03	99.66	0.99
18	99.50	94.74	99.74	0.94	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	98.75	86.96	99.47	0.88	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	98.75	90.91	99.21	0.88	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	99.75	100.00	99.73	0.98	99.00	92.86	99.46	0.92	99.50	96.43	99.73	0.96
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	99.75	95.45	100.00	0.98	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	99.50	87.50	100.00	0.93	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	99.75	97.30	100.00	0.99	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.25	100.00	99.12	0.97
11	100.00	100.00	100.00	1.00	99.75	97.96	100.00	0.99	99.75	97.96	100.00	0.99
16	99.75	100.00	99.66	0.99	100.00	100.00	100.00	1.00	99.75	99.03	100.00	0.99
18	99.75	94.74	100.00	0.97	100.00	100.00	100.00	1.00	99.25	94.74	99.48	0.92
31	100.00	100.00	100.00	1.00	99.75	100.00	99.73	0.98	100.00	100.00	100.00	1.00
33	99.50	100.00	99.47	0.95	99.50	95.45	99.74	0.95	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	99.50	96.43	99.73	0.96	99.75	96.43	100.00	0.98
45	100.00	100.00	100.00	1.00	99.75	91.67	100.00	0.96	100.00	100.00	100.00	1.00
52	99.50	100.00	99.47	0.95	100.00	100.00	100.00	1.00	99.75	95.45	100.00	0.98
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	99.75	97.30	100.00	0.99	99.75	100.00	99.72	0.99	99.75	100.00	99.72	0.99
66	99.25	72.73	100.00	0.85	99.50	90.91	99.74	0.91	99.75	90.91	100.00	0.95

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	99.50	100.00	99.42	0.98	100.00	100.00	100.00	1.00
11	99.50	100.00	99.43	0.98	100.00	100.00	100.00	1.00
16	99.75	100.00	99.66	0.99	99.50	100.00	99.33	0.99
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	99.50	91.30	100.00	0.95	99.50	95.65	99.73	0.95
33	99.75	95.45	100.00	0.98	99.75	100.00	99.74	0.98
35	99.75	96.43	100.00	0.98	99.75	100.00	99.73	0.98
45	100.00	100.00	100.00	1.00	99.75	91.67	100.00	0.96
52	100.00	100.00	100.00	1.00	99.75	100.00	99.74	0.98
53	99.25	93.75	99.48	0.91	99.75	93.75	100.00	0.97
58	99.25	94.59	99.72	0.95	98.50	89.19	99.45	0.91
66	99.75	90.91	100.00	0.95	99.75	90.91	100.00	0.95

Table 4-4 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency with radial basis function network.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	85.50	0.00	100.00	NaN	100.00	100.00	100.00	1.00	98.50	98.28	98.54	0.94
11	87.75	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	97.96	100.00	0.99
16	25.75	100.00	0.00	NaN	97.25	99.03	96.63	0.93	100.00	100.00	100.00	1.00
18	95.25	0.00	100.00	NaN	97.75	84.21	98.43	0.77	98.25	73.68	99.48	0.79
31	94.25	0.00	100.00	NaN	99.75	95.65	100.00	0.98	99.50	100.00	99.47	0.96
33	94.50	0.00	100.00	NaN	97.25	54.55	99.74	0.70	99.50	90.91	100.00	0.95
35	93.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	97.00	0.00	100.00	NaN	98.50	91.67	98.71	0.79	100.00	100.00	100.00	1.00
52	94.50	0.00	100.00	NaN	99.50	100.00	99.47	0.95	99.50	100.00	99.47	0.95
53	96.00	0.00	100.00	NaN	96.75	37.50	99.22	0.49	100.00	100.00	100.00	1.00
58	90.75	0.00	100.00	NaN	99.25	94.59	99.72	0.95	99.75	100.00	99.72	0.99
66	97.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.25	72.73	100.00	0.85

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	99.75	100.00	99.66	0.99	99.75	100.00	99.66	0.99
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	99.50	95.65	99.73	0.95	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	99.75	95.45	100.00	0.98	99.75	95.45	100.00	0.98
35	100.00	100.00	100.00	1.00	99.75	96.43	100.00	0.98	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.75	95.45	100.00	0.98
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	99.75	100.00	99.72	0.99	99.75	100.00	99.72	0.99
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	99.75	100.00	99.74	0.97	99.75	94.74	100.00	0.97
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.75	100.00	99.74	0.96
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	99.75	90.91	100.00	0.95	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.75	97.96	100.00	0.99	99.75	97.96	100.00	0.99
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	99.50	94.74	99.74	0.94	100.00	100.00	100.00	1.00
31	99.75	100.00	99.73	0.98	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	99.50	100.00	99.46	0.96
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	99.75	95.45	100.00	0.98
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

4.1.3 K-nearest Neighbor Technique

The experimental results of the HPV genotype prediction gained by ChaosCentroid and by ChaosFrequency feature extraction with the predicting system based on k-nearest neighbor technique are summarized in Tables 4-5 and 4-6, respectively. The results obtained by this predicting system have shown the high performance of prediction. Thus, it can imply that, in each sub-region, the structure of sequence in a form of centroid by ChaosCentroid and the statistical distribution of mono-, di-, or higher order nucleotides in a form of frequency by ChaosFrequency, are closed to each other in the same genotype. The overall performance trend obtained by this predicting system is similar to those of multi-layer perceptron neural network. But the performance of this predicting system is slightly higher than the performance of multi-layer perceptron neural network.

Table 4-5 The results of the HPV genotype prediction based on the features extracted by ChaosCentroid with k-nearest neighbor technique.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	96.50	91.38	97.37	0.86	96.25	87.93	97.66	0.85	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	95.50	81.63	97.44	0.79	99.50	97.96	99.72	0.98
16	90.00	76.70	94.61	0.73	94.75	90.29	96.30	0.86	98.75	99.03	98.65	0.97
18	94.75	31.58	97.90	0.34	97.00	63.16	98.69	0.65	98.75	78.95	99.74	0.85
31	99.25	91.30	99.73	0.93	97.75	91.30	98.14	0.82	100.00	100.00	100.00	1.00
33	89.75	22.73	93.65	0.14	99.75	100.00	99.74	0.98	99.25	95.45	99.47	0.93
35	98.50	92.86	98.92	0.89	97.50	82.14	98.66	0.81	99.75	96.43	100.00	0.98
45	100.00	100.00	100.00	1.00	98.00	75.00	98.71	0.68	99.50	91.67	99.74	0.91
52	100.00	100.00	100.00	1.00	98.25	81.82	99.21	0.83	100.00	100.00	100.00	1.00
53	98.25	81.25	98.96	0.78	95.75	31.25	98.44	0.36	99.25	87.50	99.74	0.90
58	100.00	100.00	100.00	1.00	98.50	89.19	99.45	0.91	99.25	97.30	99.45	0.96
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.50	97.96	99.72	0.98	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	99.75	100.00	99.66	0.99	99.50	98.06	100.00	0.99	100.00	100.00	100.00	1.00
18	99.75	94.74	100.00	0.97	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	99.50	91.30	100.00	0.95	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	99.75	100.00	99.74	0.98	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	99.75	96.43	100.00	0.98	99.25	96.43	99.46	0.94	100.00	100.00	100.00	1.00
45	99.75	100.00	99.74	0.96	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	99.50	93.75	99.74	0.93	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	99.75	100.00	99.72	0.99	99.75	100.00	99.72	0.99	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

Table 4-6 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency with k-nearest neighbor technique.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	14.50	100.00	0.00	NaN	100.00	100.00	100.00	1.00	98.75	98.28	98.83	0.95
11	87.75	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	97.96	100.00	0.99
16	74.25	0.00	100.00	NaN	98.00	97.09	98.32	0.95	100.00	100.00	100.00	1.00
18	95.25	0.00	100.00	NaN	99.25	89.47	99.74	0.92	99.00	84.21	99.74	0.89
31	94.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	100.00	99.73	0.98
33	94.50	0.00	100.00	NaN	98.00	77.27	99.21	0.80	99.75	95.45	100.00	0.98
35	93.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	97.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	94.50	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	96.00	0.00	100.00	NaN	99.25	93.75	99.48	0.91	100.00	100.00	100.00	1.00
58	90.75	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	100.00	99.72	0.99
66	97.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	90.91	100.00	0.95

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.75	97.96	100.00	0.99	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.75	99.03	100.00	0.99
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	99.50	95.65	99.73	0.95	99.75	100.00	99.73	0.98
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	99.50	96.43	99.73	0.96	99.50	96.43	99.73	0.96
45	99.75	100.00	99.74	0.96	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

4.1.4 Fuzzy K-nearest Neighbor Technique

The experimental results of the HPV genotype prediction gained by ChaosCentroid and by ChaosFrequency feature extraction with the predicting system based on fuzzy k-nearest neighbor technique are summarized in Tables 4-7 and 4-8, respectively. Likewise, the overall performance trend obtained from this predicting system is similar to those of multi-layer perceptron neural network. But the overall performance of this predicting system is slightly higher than the performance of multi-layer perceptron neural network. Additionally, it is noticeable that the performance of this predicting system is statistically equal to the performance of k-nearest neighbor technique due to setting the same value of *k*.

Table 4-7 The results of the HPV genotype prediction based on the features extracted by ChaosCentroid with fuzzy k-nearest neighbor technique.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	96.50	91.38	97.37	0.86	96.25	87.93	97.66	0.85	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	95.50	81.63	97.44	0.79	99.50	97.96	99.72	0.98
16	90.00	76.70	94.61	0.73	94.75	90.29	96.30	0.86	98.75	99.03	98.65	0.97
18	94.75	31.58	97.90	0.34	97.00	63.16	98.69	0.65	98.75	78.95	99.74	0.85
31	99.25	91.30	99.73	0.93	97.75	91.30	98.14	0.82	100.00	100.00	100.00	1.00
33	89.75	22.73	93.65	0.14	99.75	100.00	99.74	0.98	99.25	95.45	99.47	0.93
35	98.50	92.86	98.92	0.89	97.50	82.14	98.66	0.81	99.75	96.43	100.00	0.98
45	100.00	100.00	100.00	1.00	98.00	75.00	98.71	0.68	99.50	91.67	99.74	0.91
52	100.00	100.00	100.00	1.00	98.25	81.82	99.21	0.83	100.00	100.00	100.00	1.00
53	98.25	81.25	98.96	0.78	95.75	31.25	98.44	0.36	99.25	87.50	99.74	0.90
58	100.00	100.00	100.00	1.00	98.50	89.19	99.45	0.91	99.25	97.30	99.45	0.96
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.50	97.96	99.72	0.98	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	99.75	100.00	99.66	0.99	99.50	98.06	100.00	0.99	100.00	100.00	100.00	1.00
18	99.75	94.74	100.00	0.97	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	99.50	91.30	100.00	0.95	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	99.75	100.00	99.74	0.98	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	99.75	96.43	100.00	0.98	99.25	96.43	99.46	0.94	100.00	100.00	100.00	1.00
45	99.75	100.00	99.74	0.96	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	99.50	93.75	99.74	0.93	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	99.75	100.00	99.72	0.99	99.75	100.00	99.72	0.99	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

Table 4-8 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency with fuzzy k-nearest neighbor technique.

HPV Genotype	Input Dimension											
	1				2				3			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	14.50	100.00	0.00	NaN	100.00	100.00	100.00	1.00	98.75	98.28	98.83	0.95
11	87.75	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	97.96	100.00	0.99
16	74.25	0.00	100.00	NaN	98.00	97.09	98.32	0.95	100.00	100.00	100.00	1.00
18	95.25	0.00	100.00	NaN	99.25	89.47	99.74	0.92	99.00	84.21	99.74	0.89
31	94.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	100.00	99.73	0.98
33	94.50	0.00	100.00	NaN	98.00	77.27	99.21	0.80	99.75	95.45	100.00	0.98
35	93.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	97.00	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	94.50	0.00	100.00	NaN	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	96.00	0.00	100.00	NaN	99.25	93.75	99.48	0.91	100.00	100.00	100.00	1.00
58	90.75	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	100.00	99.72	0.99
66	97.25	0.00	100.00	NaN	100.00	100.00	100.00	1.00	99.75	90.91	100.00	0.95

HPV Genotype	Input Dimension											
	4				5				6			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	99.75	97.96	100.00	0.99	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	99.75	99.03	100.00	0.99
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	99.50	95.65	99.73	0.95	99.75	100.00	99.73	0.98
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	99.50	96.43	99.73	0.96	99.50	96.43	99.73	0.96
45	99.75	100.00	99.74	0.96	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension											
	7				8				9			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

HPV Genotype	Input Dimension							
	10				11			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00	100.00	100.00	100.00	1.00

4.1.5 NCBI Viral Genotyping Tool

NCBI viral genotyping tool [52] is a web-based tool for identifying the genotype of a viral sequence. The algorithm of this tool is described as follows. Firstly, it works by sliding a window along the query sequence and processing each window (sequence segment) separately. Secondly, each segment is compared to a set of reference sequences using Basic Local Alignment Search Tool (BLAST), which returns the similarity scores for the local alignments. Then, the reference sequence genotype that matches the query with the highest similarity score is assigned to the query segment. The process is repeated for the next window until the whole length of the query sequence has been covered. Lastly, the results from all windows are combined. If the same genotype is assigned to most segments, then the query sequence is considered as the genotype. This tool is a web-based resource providing a reliable genotyping method based on alignment. Therefore, this experiment adopted this tool for identifying genotypes of the viral genomes in the HPV genome data set, and the experimental results of this tool was illustrated in Table 4-9.

Table 4-9 The results of the HPV genotype prediction obtained by NCBI viral genotyping tool

HPV Genotype	Accuracy	Sensitivity	Specificity	MCC
6	100.00	100.00	100.00	1.00
11	100.00	100.00	100.00	1.00
16	100.00	100.00	100.00	1.00
18	100.00	100.00	100.00	1.00
31	100.00	100.00	100.00	1.00
33	100.00	100.00	100.00	1.00
35	100.00	100.00	100.00	1.00
45	100.00	100.00	100.00	1.00
52	100.00	100.00	100.00	1.00
53	100.00	100.00	100.00	1.00
58	100.00	100.00	100.00	1.00
66	100.00	100.00	100.00	1.00

To evaluate the prediction performance, the result of this genotyping tool were compared with the best results of the proposed ChaosCentroid and ChaosFrequency feature extraction techniques with all predicting systems. The results have shown that all methods, excepting ChaosCentroid with radial basis function network, can achieve the best performance of the four metrics, including accuracy, sensitivity, specificity, and Matthew's Correlation Coefficient, in predicting the HPV genotypes from the HPV genome data set. It demonstrated that both of proposed techniques, i.e. ChaosCentroid and ChaosFrequency, with the predicting systems, and the NCBI genotyping tool can be used as methods for predicting the genotypes of HPV genomes.

Even though there is no significance between the proposed techniques and the NCBI genotyping tool, some issues should be considered.

Suppose there are $N+1$ genomes of length L. N genomes are for training and 1 genome is for testing.

In the prediction algorithms, ChaosCentroid and ChaosFrequency extract the features from HPV whole genomes, and identify the corresponding HPV genotypes by the predicting systems. Thus, the HPV genomes are firstly extracted features to be used as input vectors. Both feature extraction techniques take $O(NL)$ time, where N is the number of genomes of length L. The time complexity of HPV genotype prediction based on multi-layer perceptron neural network is discussed as below.

For training process, the time complexity of backpropagation is $O(N \cdot m \cdot h^k \cdot o \cdot i)$, where N is the number of training samples, m is the number of input dimensions, k is the number of hidden layers containing h neurons, o is output neurons, and i is the number of iterations. After substituting by parameters used in this experiment, the time complexity of backpropagation is approximately $O(N \cdot 11 \cdot 16^1 \cdot 12 \cdot 1000)$.

For testing process, the MLP computes the output of neuron b as

$$y_b(i) = \varphi(v_b(i))$$

where $v_b(i)$ is the induced local field of neuron b, defined by

$$v_b(i) = \sum_{i=0}^{m} w_{ba}(i) y_a(i)$$

where m is the total number of inputs applied to neuron b. $w_{ba}(i)$ is the synaptic weight connecting neuron a to neuron b. $y_a(i)$ is the input signal of neuron b or the function signal appearing at the output of neuron a.

Even though computing the output in testing process takes $O(mi)$ time, this experiment sets the maximum value of m to 11 and the number of iterations i to 1000. Then, the time complexity of testing one input vector takes constant time, which corresponds to $O(1)$.

Without artificial intelligence, NCBI viral genotyping tool provides a reliable genotyping method based on sequence homology searching with alignment procedure. This tool does not required for feature extraction and training process, then, the time complexity of this tool is only concerned with testing process. Although time complexity of this genotyping tool are not clearly revealed, a concept is based on Basic Local Alignment Search Tool (BLAST). So, it can analyze the time complexity according to BLAST algorithm. According to [53], the expected time complexity of nucleotide BLAST is approximately $c_1W + c_2N + c_3NW/4^w$, where W is the number of words generated, w is the length of words, N is the number of residues in the database, and c_1, c_2 and c_3 are constants. The W term accounts for compiling the word list, the N term covers the database scan, and the NW term is for extending the hits. In addition, this NCBI genotyping tool works by sliding a window along the query sequence and processing each window separately. This produces $\frac{L}{300}$ windows to proceed, where the window size is set to 300 in this experiment. Then, the time complexity of this tool should precisely be $\frac{L}{300}(c_1W + c_2N + c_3NW/4^w)$.

Therefore, the proposed prediction algorithm takes $O(Nn)$ times for extracting features and training process. It depends on the number of genomes and their lengths in training sample. The network learns the training samples only once, and it is ready for the prediction. It takes only constant time for every incoming input vector in testing process. In contrast, NCBI viral genotyping tool does not have training process so the running time spends for only the testing process. For every incoming tested genome, this tool takes $\frac{L}{300}(c_1W + c_2N + c_3NW/4^w)$, which depends on the length of query sequence, the number of generated words of length w, and the number of HPV

genomes in database. Then, the time consuming of this tool is considerably greater than the proposed prediction algorithms.

Interestingly, it was found that, in the proposed techniques, the large number of tested vectors can be altogether fed to the predicting systems in the same testing, while only one query sequence at a time can be processed by this tool. So, this tool is not appropriate for large scaled tasks.

In contrast, the proposed techniques, i.e. ChaosCentroid and ChaosFrequency, are based on Chaos game representation, which provides a unique and scale-independent representation of DNA sequences through the statistical distribution of mono-, di-, tri-, or higher order nucleotides along DNA sequences. The Advantage of CGR over alignment is that it neither requires prior knowledge of consensus sequences nor it involves exhaustive searches for sequences in databases.

However, the chaos game representation (CGR) also has some limitations. For instance, it spends some computational time to generate the representations from DNA sequences. In order to relieve this limitation, this experiment utilized the singular value decomposition to reduce the size of CGR into a smaller number of feature matrices so the computational time in the prediction process can also be reduced. From the experimental results, the proposed ChaosCentroid and ChaosFrequency can successfully extract the characteristic parameters of HPV genotypes for the prediction.

4.2 HPV Genotype Prediction from Partial Coding Sequences

For the HPV partial coding sequence data set, the features obtained by three proposed extraction techniques were fed to the predicting system based on the fuzzy k nearest neighbor technique for predicting the corresponding HPV genotypes. Accordingly, the performance of HPV genotype prediction was separately summarized according to each feature extraction technique and the size of sub-regions in the Chaos game representation after being partitioned. The explanation and discussion of these experimental results were as follows.

4.2.1 ChaosCentroid

The experimental results of the HPV genotype prediction gained by ChaosCentroid feature extraction with the predicting system based on fuzzy k-nearest neighbor technique are summarized in Tables 4-10. For the partial coding sequence data set, the lengths of nucleotide sequences are short and various, so some sub-regions may contain a very few dots or none of dots. Thus, the partition of the Chaos game representation is limited because it cannot compute centroids for the empty sub-regions. For this experiment, the maximum value of $\frac{n}{g}$ that can be used to partition the CGR is 2, and the results show the low performance of the HPV genotype prediction due to an insufficiency of the partition.

Table 4-10 The results of the HPV genotype prediction based on the features extracted by ChaosCentroid after partitioning the CGR into sub-regions of size 2 × 2.

HPV Genotype	Input Dimension			
	2			
	Acc	Sen	Spec	MCC
6	95.04	72.70	97.24	0.70
11	96.70	64.44	98.14	0.61
16	80.40	78.22	82.20	0.60
18	96.19	57.78	97.91	0.55
31	97.08	73.30	98.49	0.72
33	97.68	35.09	98.84	0.34
35	99.05	42.31	99.52	0.42
39	98.48	39.02	99.26	0.39
42	98.98	32.00	99.52	0.33
45	99.05	19.05	99.58	0.21
51	99.27	63.64	99.65	0.64
52	97.55	52.22	98.89	0.54
53	97.01	66.21	98.50	0.66
56	98.00	50.00	99.03	0.50
58	94.98	64.71	97.27	0.62
59	98.32	23.53	99.13	0.22
66	96.38	60.71	98.04	0.58
68	98.06	6.25	99.01	0.05
70	99.11	40.00	99.49	0.36
71	99.17	47.83	99.55	0.45
81	98.89	26.09	99.42	0.25

4.2.2 ChaosFrequency

The experimental results of the HPV genotype prediction gained by ChaosFrequency feature extraction with the predicting system based on fuzzy k-nearest neighbor technique are summarized in Tables 4-11 to 4-14. ChaosFrequency is more flexible than ChaosCentroid in a case that the frequencies would be zeros if the sub-regions contain none of dots. When $\frac{n}{g}$ is 2, the results of ChaosFrequency yielded the higher performance than those of ChaosCentroid. When $\frac{n}{g}$ is increased, the performances of ChaosFrequency are also moderately increased.

Table 4-11 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency after partitioning the CGR into sub-regions of size 2 × 2.

HPV Genotype	Input Dimension			
	2			
	Acc	Sen	Spec	MCC
6	96.16	77.66	97.98	0.76
11	97.20	68.15	98.51	0.66
16	85.55	84.12	86.72	0.71
18	96.98	66.67	98.34	0.64
31	96.06	63.64	97.98	0.62
33	98.32	49.12	99.22	0.51
35	99.68	76.92	99.87	0.80
39	98.86	60.98	99.36	0.58
42	99.40	64.00	99.68	0.62
45	98.89	23.81	99.39	0.22
51	99.11	66.67	99.45	0.61
52	97.74	58.89	98.89	0.59
53	97.62	73.79	98.77	0.73
56	98.25	57.58	99.12	0.57
58	95.87	72.40	97.64	0.69
59	99.24	61.76	99.65	0.63
66	97.55	70.71	98.80	0.71
68	98.73	31.25	99.42	0.33
70	99.11	35.00	99.52	0.33
71	99.97	95.65	100.00	0.98
81	99.59	69.57	99.81	0.71

Table 4-12 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency after partitioning the CGR into sub-regions of size 4 × 4.

HPV Genotype	Input Dimension			
	4			
	Acc	Sen	Spec	MCC
6	99.08	95.39	99.44	0.94
11	99.21	91.11	99.57	0.90
16	93.36	93.32	93.39	0.87
18	98.28	79.26	99.14	0.79
31	97.90	78.41	99.06	0.80
33	98.95	64.91	99.58	0.69
35	99.62	80.77	99.78	0.78
39	99.43	80.49	99.68	0.78
42	99.65	80.00	99.81	0.78
45	99.33	52.38	99.65	0.51
51	99.36	78.79	99.58	0.72
52	98.63	76.67	99.28	0.76
53	98.41	83.45	99.13	0.82
56	98.86	62.12	99.64	0.69
58	97.94	84.62	98.94	0.84
59	99.36	73.53	99.65	0.71
66	98.32	79.29	99.20	0.80
68	99.30	62.50	99.68	0.64
70	99.59	75.00	99.74	0.70
71	99.94	100.00	99.94	0.96
81	99.68	73.91	99.87	0.77

Table 4-13 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency after partitioning the CGR into sub-regions of size 8 × 8.

HPV Genotype	Input Dimension			
	8			
	Acc	Sen	Spec	MCC
6	98.98	93.26	99.55	0.94
11	99.40	94.07	99.63	0.93
16	95.30	95.29	95.30	0.91
18	98.89	88.89	99.34	0.87
31	98.25	83.52	99.13	0.83
33	99.02	73.68	99.48	0.73
35	99.78	84.62	99.90	0.86
39	99.68	85.37	99.87	0.87
42	99.71	72.00	99.94	0.80
45	99.75	85.71	99.84	0.82
51	99.81	87.88	99.94	0.91
52	99.05	82.22	99.54	0.83
53	98.79	86.90	99.37	0.86
56	99.21	77.27	99.68	0.80
58	98.38	90.95	98.94	0.88
59	99.56	82.35	99.74	0.80
66	99.08	86.43	99.67	0.89
68	99.08	50.00	99.58	0.52
70	99.78	80.00	99.90	0.82
71	99.84	95.65	99.87	0.90
81	99.81	82.61	99.94	0.86

Table 4-14 The results of the HPV genotype prediction based on the features extracted by ChaosFrequency after partitioning the CGR into sub-regions of size 16 × 16.

HPV Genotype	Input Dimension			
	16			
	Acc	Sen	Spec	MCC
6	99.21	95.39	99.58	0.95
11	99.56	97.04	99.67	0.95
16	95.58	95.50	95.65	0.91
18	99.02	90.37	99.40	0.88
31	98.44	83.52	99.33	0.85
33	99.08	77.19	99.48	0.75
35	99.87	88.46	99.97	0.92
39	99.56	75.61	99.87	0.82
42	99.71	80.00	99.87	0.82
45	99.68	76.19	99.84	0.76
51	99.71	93.94	99.78	0.87
52	99.21	84.44	99.64	0.85
53	99.11	86.90	99.70	0.90
56	99.33	83.33	99.68	0.84
58	98.60	90.95	99.18	0.89
59	99.56	76.47	99.81	0.79
66	99.02	90.71	99.40	0.89
68	99.40	62.50	99.78	0.68
70	99.75	80.00	99.87	0.80
71	99.97	95.65	100.00	0.98
81	99.62	78.26	99.78	0.75

4.2.3 ChaosPoly

The experimental results of the HPV genotype prediction gained by ChaosPoly feature extraction with the predicting system based on fuzzy k-nearest neighbor technique are summarized in Tables 4-15 to 4-17. According to the results, the performances obtained by ChaosPoly and by ChaosFrequency feature extraction techniques are in the same trend. That is, their performances are increased, in proportion to the number of sub-regions partitioning to the CGR. Additionally, the results also showed that almost all the performances of ChaosPoly are significantly higher than those of ChaosFrequency. Since both of ChaosFrequency and ChaosPoly were proposed based on computing the probability of distribution of different mono-, di-, tri-, or higher order nucleotides along the nucleotide sequences, it can be implied that, in the case of having less information that can be found in nucleotide sequences with short lengths, such as partial coding sequences, computing the distribution in polynomial form has an ability to capture competently with the characteristic of HPV genotypes, rather than applying the singular value decomposition. Besides, the relation between the number of sub-regions and the size of units in computing the polynomial form can influence on the achievement of the prediction. To illustrate this point, the performances gained by the more number of units have a tendency to gain the higher performance, even though the CGR is partitioned by the same numbers of sub-regions.

Table 4-15 The results of the HPV genotype prediction based on the features extracted by ChaosPoly after partitioning the CGR into sub-regions of size 4 × 4.

HPV Genotype	Input Dimension			
	4			
	Acc	Sen	Spec	MCC
6	98.63	93.26	99.16	0.92
11	99.56	94.81	99.77	0.95
16	95.84	95.92	95.77	0.92
18	99.17	91.85	99.50	0.90
31	98.38	83.52	99.26	0.84
33	99.27	80.70	99.61	0.80
35	99.59	69.23	99.84	0.73
39	99.46	80.49	99.71	0.79
42	99.68	76.00	99.87	0.79
45	99.75	80.95	99.87	0.81
51	99.75	87.88	99.87	0.88
52	98.92	76.67	99.57	0.80
53	99.27	93.10	99.57	0.92
56	98.76	66.67	99.45	0.69
58	98.76	90.95	99.35	0.90
59	99.65	82.35	99.84	0.83
66	98.98	88.57	99.47	0.88
68	99.05	50.00	99.55	0.51
70	99.75	80.00	99.87	0.80
71	99.84	100.00	99.84	0.91
81	99.65	65.22	99.90	0.74

Table 4-16 The results of the HPV genotype prediction based on the features extracted by ChaosPoly after partitioning the CGR into sub-regions of size 8 × 8.

HPV Genotype	Input Dimension							
	4				16			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	99.21	95.39	99.58	0.95	99.71	98.58	99.83	0.98
11	98.79	86.67	99.34	0.85	99.87	98.52	99.93	0.98
16	94.92	95.50	94.43	0.90	99.02	99.30	98.78	0.98
18	99.17	92.59	99.47	0.90	99.75	98.52	99.80	0.97
31	98.38	85.23	99.16	0.85	99.46	96.02	99.66	0.95
33	99.08	75.44	99.51	0.74	99.52	82.46	99.84	0.86
35	99.71	76.92	99.90	0.82	99.90	88.46	100.00	0.94
39	99.65	85.37	99.84	0.86	99.81	92.68	99.90	0.93
42	99.56	64.00	99.84	0.70	99.94	92.00	100.00	0.96
45	99.65	66.67	99.87	0.72	99.87	85.71	99.97	0.90
51	99.90	93.94	99.97	0.95	99.97	100.00	99.97	0.99
52	99.17	78.89	99.77	0.84	99.62	91.11	99.87	0.93
53	99.11	88.28	99.63	0.90	99.78	95.86	99.97	0.97
56	98.92	69.70	99.55	0.73	99.78	93.94	99.90	0.95
58	98.86	92.76	99.32	0.91	99.78	99.10	99.83	0.98
59	99.52	76.47	99.78	0.77	99.97	100.00	99.97	0.99
66	99.11	90.00	99.53	0.90	99.87	98.57	99.93	0.99
68	99.02	43.75	99.58	0.47	99.84	87.50	99.97	0.92
70	99.78	70.00	99.97	0.81	99.94	95.00	99.97	0.95
71	99.90	100.00	99.90	0.94	100.00	100.00	100.00	1.00
81	99.71	82.61	99.84	0.81	99.90	95.65	99.94	0.94

Table 4-17 The results of the HPV genotype prediction based on the features extracted by ChaosPoly after partitioning the CGR into sub-regions of size 16 × 16.

HPV Genotype	Input Dimension							
	4				16			
	Acc	Sen	Spec	MCC	Acc	Sen	Spec	MCC
6	99.46	95.74	99.83	0.97	99.78	97.87	99.97	0.99
11	99.27	92.59	99.57	0.91	99.75	96.30	99.90	0.97
16	95.49	95.64	95.36	0.91	98.60	99.02	98.26	0.97
18	99.43	93.33	99.70	0.93	99.65	96.30	99.80	0.96
31	98.38	82.95	99.29	0.84	99.24	92.05	99.66	0.93
33	99.14	70.18	99.68	0.74	99.43	82.46	99.74	0.84
35	99.68	84.62	99.81	0.81	99.94	92.31	100.00	0.96
39	99.62	82.93	99.84	0.85	99.81	92.68	99.90	0.93
42	99.52	68.00	99.78	0.69	99.90	96.00	99.94	0.94
45	99.62	66.67	99.84	0.70	99.94	90.48	100.00	0.95
51	99.84	93.94	99.90	0.92	99.90	96.97	99.94	0.95
52	99.24	86.67	99.61	0.86	99.62	92.22	99.84	0.93
53	98.98	88.28	99.50	0.88	99.71	95.86	99.90	0.97
56	99.30	84.85	99.61	0.83	99.81	93.94	99.94	0.95
58	99.17	95.02	99.49	0.94	99.68	99.10	99.73	0.98
59	99.49	79.41	99.71	0.77	99.94	100.00	99.94	0.97
66	99.14	92.14	99.47	0.90	99.94	98.57	100.00	0.99
68	99.36	62.50	99.74	0.66	99.75	84.38	99.90	0.87
70	99.84	85.00	99.94	0.87	99.94	95.00	99.97	0.95
71	99.90	86.96	100.00	0.93	99.97	100.00	99.97	0.98
81	99.78	86.96	99.87	0.85	99.94	95.65	99.97	0.96

HPV Genotype	Input Dimension			
	64			
	Acc	Sen	Spec	MCC
6	99.81	98.58	99.93	0.99
11	99.78	98.52	99.83	0.97
16	99.36	99.44	99.30	0.99
18	99.81	98.52	99.87	0.98
31	99.81	98.30	99.90	0.98
33	99.65	91.23	99.81	0.90
35	99.90	92.31	99.97	0.94
39	99.84	95.12	99.90	0.94
42	99.97	96.00	100.00	0.98
45	99.94	90.48	100.00	0.95
51	99.97	100.00	99.97	0.99
52	99.78	95.56	99.90	0.96
53	99.78	95.86	99.97	0.97
56	99.90	96.97	99.97	0.98
58	99.84	99.10	99.90	0.99
59	99.97	100.00	99.97	0.99
66	99.90	99.29	99.93	0.99
68	99.84	90.63	99.94	0.92
70	99.97	95.00	100.00	0.97
71	100.00	100.00	100.00	1.00
81	100.00	100.00	100.00	1.00

CHAPTER 5
CONCLUSION

Cervical cancer is the second most common cancer significantly causing morbidity and mortality in women worldwide, and a persistent Infection with high risk types of Human Papillomavirus is considered as a necessary cause of the cancer. HPV genotyping is essential to provide relevant information regarding risk stratification for diagnosis and medical treatment, in addition to reveal the better understanding of the relationship of HPV with carcinogenesis. Therefore, the objective of this dissertation is to develop the new algorithm for predicting the HPV genotypes. This experiment concentrates on predicting HPV genotypes from two significant forms of nucleotide sequences, namely, whole genomes and partial coding sequences.

For the prediction from whole genomes, two new feature extraction techniques, i.e. ChaosCentroid and ChaosFrequency, were proposed. In this techniques, a partitioned Chaos Game Representation (CGR) was deployed to represent HPV genomes. ChaosCentroid captures the structure of sequences in terms of centroid of each sub-region with Euclidean distances among the centroids and the center of CGR as the relations of all sub-regions. ChaosFrequency extracts the statistical distribution of mono-, di-, or higher order nucleotides along HPV genomes and forms a matrix of frequency of dots in each sub-region. For performance evaluation, four different types of classifiers, i.e. Multi-layer Perceptron, Radial Basis Function, K-Nearest Neighbor, and Fuzzy K-Nearest Neighbor Techniques were deployed, and the best results from each classifier were compared with the NCBI genotyping tool. The experimental results obtained by four different classifiers are in the same trend. ChaosCentroid gives considerably higher performance than ChaosFrequency when the input length is one but it is moderately lower than ChaosFrequency when the input length is two. Both techniques yielded almost and exactly the best performance when the input length is greater than three. In addition, when comparing these proposed techniques with the NCBI Viral genotyping tool, it reveals that there is no significance

in prediction performance. But the time complexity of the proposed techniques is considerably less than that of the genotyping tool for every incoming tested genomes.

For the prediction from partial coding sequences, ChaosPoly feature extraction technique was proposed. Since this data set contains the nucleotide sequences with short and various length, ChaosPoly gave more contemplation of extracting the local information hidden in sub-regions. In this technique, a partitioned Chaos Game Representation (CGR) was deployed to represent HPV genomes, and it extracted the statistical distribution of mono-, di-, or higher order nucleotides along HPV genomes and formed a matrix of frequency of dots in each sub-region, as ChaosFrequency. But this technique captured the relationship among sub-regions of the CGR in polynomial form. From the experimental results, ChaosCentroid was not appropriate for extracting the nucleotide sequences with short lengths because containing none of dots in some sub-regions can limit the numbers of sub-regions partitioning to the CGR. In addition, almost all the performances of ChaosPoly were significantly higher than those of ChaosFrequency. This implies that ChaosPoly has the ability to capture competently with the characteristic of HPV genotypes more than ChaosFrequency. Therefore, ChaosPoly is proper for the HPV genotype prediction in the case of having less information, which can be found in nucleotide sequences with short lengths, such as partial coding sequences.

REFERENCES

1. Centers for Disease Control and Prevention, *Human Papillomavirus*, U.S. Department of Health & Human Services, Editor. 2015.
2. Sheng, J. and W.-y. Zhang, *Identification of biomarkers for cervical cancer in peripheral blood lymphocytes using oligonucleotide microarrays* Chinese Medical Journal, 2010. **123**(8): p. 1000-1005.
3. Dobec, M., et al., *Human papillomavirus infection among women with cytological abnormalities in Switzerland investigated by an automated linear array genotyping test.* Journal of Medical Virology, 2011. **83**(8): p. 1370-1376.
4. Giorgi Rossi, P., et al., *Distribution of high and low risk HPV types by cytological status: a population based study from Italy.* Infectious Agents and Cancer, 2011. **6**(1): p. 1-8.
5. Couture, M.-C., et al., *Cervical human papillomavirus infection among young women engaged in sex work in Phnom Penh, Cambodia: prevalence, genotypes, risk factors and association with HIV infection.* BMC Infectious Diseases, 2012. **12**(1): p. 1-11.
6. Ursu, R., et al., *HPV prevalence and type distribution in women with or without cervical lesions in the Northeast region of Romania.* Virology Journal, 2011. **8**(1): p. 1-5.
7. Lee, S.H., et al., *Routine human papillomavirus genotyping by DNA sequencing in community hospital laboratories.* Infectious Agents and Cancer, 2007. **2**: p. 11-11.
8. Abreu, A.L., et al., *A review of methods for detect human Papillomavirus infection.* Virology Journal, 2012. **9**: p. 262.
9. Carvalho, N.d.O., et al., *Comparison of HPV genotyping by type-specific PCR and sequencing.* Memórias do Instituto Oswaldo Cruz, 2010. **105**: p. 73-78.
10. Wang, P. and X. Xiao. *Predicting the Risk Type of Human Papillomaviruses Based on Sequence-Derived Features*. in *2011 5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE)*. 2011.

11. Xiao, X. and P. Wang. *A new approach using geometric moments of distance matrix image for risk type prediction of human papillomaviruses*. in *2011 International Conference on Electronics, Communications and Control (ICECC)*. 2011.

12. Park, S.-B., S. Hwang, and B.-T. Zhang, *Classification of the Risk Types of Human Papillomavirus by Decision Trees*, in *Intelligent Data Engineering and Automated Learning*, J. Liu, Y.-m. Cheung, and H. Yin, Editors. 2003, Springer Berlin Heidelberg. p. 540-544.

13. Park, S.-B., S. Hwang, and B.-T. Zhang, *Classification of Human Papillomavirus (HPV) Risk Type via Text Mining.* Genomics & Informatics, 2003. **1**(2): p. 80-86.

14. Eom, J.-H., S.-B. Park, and B.-T. Zhang, *Genetic Mining of DNA Sequence Structures for Effective Classification of the Risk Types of Human Papillomavirus (HPV)*, in *Neural Information Processing*, N. Pal, et al., Editors. 2004, Springer Berlin Heidelberg. p. 1334-1343.

15. Kim, S. and B.-T. Zhang, *Human Papillomavirus Risk Type Classification from Protein Sequences Using Support Vector Machines*, in *Applications of Evolutionary Computing*, F. Rothlauf, et al., Editors. 2006, Springer Berlin Heidelberg. p. 57-66.

16. Kim, S. and J.-H. Eom, *Prediction of the Human Papillomavirus Risk Types Using Gap-Spectrum Kernels*, in *Advances in Neural Networks - ISNN 2006*, J. Wang, et al., Editors. 2006, Springer Berlin Heidelberg. p. 710-715.

17. Kim, S., J. Kim, and B.T. Zhang, *Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures.* Comput Biol Med, 2009. **39**(2): p. 187-93.

18. Jeffrey, H.J., *Chaos game representation of gene structure.* Nucleic Acids Research, 1990. **18**(8): p. 2163-2170.

19. Almeida, J.S., et al., *Analysis of genomic sequences by Chaos Game Representation.* Bioinformatics, 2001. **17**(5): p. 429-437.

20. Dutta, C. and J. Das, *Mathematical characterization of Chaos Game Representation: New algorithms for nucleotide sequence analysis.* Journal of Molecular Biology, 1992. **228**(3): p. 715-719.

21. JinLong, L., et al. *Predicting Thermophilic Nucleotide Sequences Based on Chaos Game Representation Features and Support Vector Machine*. in *Bioinformatics and Biomedical Engineering, (iCBBE) 2011 5th International Conference on*. 2011.
22. Qianjun, X., Z. Jinyu, and S. Long. *A novel 3D graphical representation of RNA secondary structures based on chaos game representation*. in *Natural Computation (ICNC), 2010 Sixth International Conference on*. 2010.
23. Deschavanne, P.J., et al., *Genomic signature: characterization and classification of species assessed by chaos game representation of sequences.* Molecular Biology and Evolution, 1999. **16**(10): p. 1391-1399.
24. Iman, T., et al. *Three Dimensional Chaos Game Representation of Genomic Sequences*. 2007.
25. Zu-Guo, Y., et al. *Chaos Game Representation of Genomes and their Simulation by Recurrent Iterated Function Systems*. in *Bioinformatics and Biomedical Engineering, 2008. ICBBE 2008. The 2nd International Conference on*. 2008.
26. Nair, V.V., et al. *ANN Based Classification of Unknown Genome Fragments Using Chaos Game Representation*. in *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*. 2010.
27. Messaoudi, I., A.E. Oueslati, and Z. Lachiri. *Genomic data visualization*. in *Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2012 6th International Conference on*. 2012.
28. Basu, S., et al., *Chaos game representation of proteins.* Journal of Molecular Graphics and Modelling, 1997. **15**(5): p. 279-289.
29. Yu, Z.-G., V. Anh, and K.-S. Lau, *Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses.* Journal of Theoretical Biology, 2004. **226**(3): p. 341-348.
30. Yang, J.-Y., Z.-G. Yu, and V. Anh, *Clustering structures of large proteins using multifractal analyses based on a 6-letter model and hydrophobicity scale of amino acids.* Chaos, Solitons & Fractals, 2009. **40**(2): p. 607-620.

31. Yang, J.-Y., et al., *Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation.* Journal of Theoretical Biology, 2009. **257**(4): p. 618-626.
32. Xuehai, H., et al. *Chaos Game Representation for Discriminating Thermophilic from Mesophilic Protein Sequences*. in *Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on*. 2009.
33. Li, N., et al. *Subcellular Locations Prediction of Proteins Based on Chaos Game Representation*. in *Bioinformatics and Biomedical Engineering , 2009. ICBBE 2009. 3rd International Conference on*. 2009.
34. Chaohong, S. and S. Feng. *Subcellular location of apoptosis proteins based on chaos game representation*. in *BioMedical Information Engineering, 2009. FBIE 2009. International Conference on Future*. 2009.
35. Yu, Z.-G., et al., *Chaos game representation of functional protein sequences, and simulation and multifractal analysis of induced measures.* Chinese Physics B, 2010. **19**(6): p. 068701.
36. Olyaee, M. and M. Yaghobi. *Improved Protein Structural Class Prediction Based on Chaos Game Representation*. in *Mathematical/Analytical Modelling and Computer Simulation (AMS), 2010 Fourth Asia International Conference on*. 2010.
37. Chou, K.-C., *Graphic Rule for Drug Metabolism Systems.* Current Drug Metabolism, 2010. **11**(4): p. 369 - 378.
38. Zhou, G.-P., *The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism.* Journal of Theoretical Biology, 2011. **284**(1): p. 142-148.
39. Wu, Z.-C., X. Xiao, and K.-C. Chou, *2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids.* Journal of Theoretical Biology, 2010. **267**(1): p. 29-34.
40. Alshalalfa, M., R. Alhajj, and J. Rokne. *Combining singular value decomposition and t-test into hybrid approach for significant gene extraction*

from microarray data. in *BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on*. 2008.

41. Zhong-Hui, D., et al. *Application of singular value decomposition and functional clustering to analyzing gene expression profiles of renal cell carcinoma*. in *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*. 2003.
42. Tomfohr, J., J. Lu, and T.B. Kepler, *Pathway level analysis of gene expression using singular value decomposition.* BMC Bioinformatics, 2005. **6**: p. 225-225.
43. Berger, J.A., et al., *Jointly analyzing gene expression and copy number data in breast cancer using data reduction models.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2006. **3**: p. 2-16.
44. Baty, F., et al., *Exploring the transcription factor activity in high-throughput gene expression data using RLQ analysis.* BMC Bioinformatics, 2013. **14**(1): p. 1-15.
45. Alireza Aghili, S., et al. *Efficient filtration of sequence similarity search through singular value decomposition*. in *Bioinformatics and Bioengineering, 2004. BIBE 2004. Proceedings. Fourth IEEE Symposium on*. 2004.
46. Peters, T.J., R. Smolikova-Wachowiak, and M.P. Wachowiak. *Microarray Image Compression Using a Variation of Singular Value Decomposition*. in *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*. 2007.
47. Hu, P., S. Bull, and H. Jiang, *Gene network modular-based classification of microarray samples.* BMC Bioinformatics, 2012. **13**(10): p. 1-7.
48. Holec, M., et al., *Comparative evaluation of set-level techniques in predictive classification of gene expression samples.* BMC Bioinformatics, 2012. **13**(10): p. 1-15.
49. Keller, J.M., M.R. Gray, and J.A. Givens, *A fuzzy K-nearest neighbor algorithm.* Systems, Man and Cybernetics, IEEE Transactions on, 1985. **SMC-15**(4): p. 580-585.

50. Guo, S.-H., et al., *iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition.* Bioinformatics, 2014. **30**(11): p. 1522-1529.

51. Chou, K.-C., *Some remarks on protein attribute prediction and pseudo amino acid composition.* Journal of Theoretical Biology, 2011. **273**(1): p. 236-247.

52. Rozanov, M., et al., *A web-based genotyping resource for viral sequences.* Nucleic Acids Research, 2004. **32**(suppl 2): p. W654-W659.

53. Altschul, S.F., et al., *Basic Local Alignment Search Tool.* J. Mol. Biol., 1990. **215**(3): p. 403–10.

APPENDIX

VITA

Name: Miss Watcharaporn Tanchotsrinon

Education:

2006-2009 M.Sc. in Computer science and Information (English Program), Department of Mathematics, Faculty of Science, Chulalongkorn University.

2002-2006 B.Sc. in Biochemistry, Department of Biochemistry, Faculty of Science, Chulalongkorn University.

Ph.D. Scholarship: Thailand Research Fund (TRF) under the Royal Golden Jubilee Scholarship.