# A topological characterization of DNA sequences based on chaos geometry and persistent homology

Dong Quan Ngoc Nguyen

Department of Applied and Computational Mathematics and Statistics,
University of Notre Dame,
Notre Dame, IN 46556, USA
email: dongquan.ngoc.nguyen@nd.edu

Phuong Dong Tan Le

Department of Applied Mathematics,
University of Waterloo,
Waterloo, Ontario, Canada N2L 3G1
email: pdle@uwaterloo.ca

Lin Xing

Department of Applied and Computational Mathematics and Statistics,
University of Notre Dame,
Notre Dame, IN 46556, USA
email: lxing@nd.edu

Lizhen Lin

Department of Applied and Computational Mathematics and Statistics,
University of Notre Dame,
Notre Dame, IN 46556, USA
email: lizhen.lin@nd.edu

February 1, 2021

**Abstract**

Methods for analyzing similarities among DNA sequences play a fundamental role in computational biology, and have a variety of applications in public health, and in the field of genetics. In this paper, a novel geometric and topological method for analyzing similarities among DNA sequences is developed, based on persistent homology from algebraic topology, in combination with chaos geometry in 4-dimensional space as a graphical representation of DNA sequences. Our topological framework for DNA similarity analysis is general, alignment-free, and can deal with DNA sequences of various lengths, while proving first-of-the-kind visualization features for visual inspection of DNA sequences directly, based on topological features of point clouds that represent DNA sequences. As an application, we test our methods on three datasets including genome sequences of different types of Hantavirus, Influenza A viruses, and Human Papillomavirus.

# 1 Introduction

The last few decades have witnessed a surge in the growth of methods that are devoted to analyzing the similarities among DNA sequences to obtain the corresponding genetic information. Despite these diverse methods, they can be classified into graphical representation of DNA sequences and other techniques based on numeric representations. Methods based on graphical representation of DNA sequences have contributed significantly to the general area of DNA similarity analysis. One of the main ideas used in graphical representation is to realize each nucleotide base, say A, C, G, T as a point in a Euclidean space, normally, $\mathbb{R}^n$ for some small values $n$ between 2 and 5. One then obtains a collection of points in $\mathbb{R}^n$ that represents a DNA sequence. Based on these collections of points, many papers are devoted, in combination with other techniques, such as signal processing (see, for example, [1]), to compare similarities among DNA sequences. The most frequently used similarity measures for analyzing differences or similarities between DNA sequences based their corresponding graphical representation are Euclidean distances or correlation angles. For papers that adopt graphical representation for DNA similarity analysis, the reader is referred to references [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21].

Other methods that do not involve graphical representation but instead use numeric representation are based on mapping each nucleotide base to a number, and thus each DNA sequence can be mapped to a number sequence. In order to compare differences or similarities among DNA sequences, other *ad hoc* methods are utilized to compare their corresponding number sequences. The reader is referred to, for example, work [22, 23, 24, 25, 26, 27, 28, 29], and their references therein for these non-graphical representation based methods.

A common theme in graphical representation based methods is to rely on *simply constructed geometric object*, say curves in Euclidean spaces that represent DNA sequences to apply for DNA similarity analysis. Although there were apparent successes in using such methods, the employed techniques pose some technical difficulties: (i) curves constructed via graphical representations may not truly represent the geometry of the corresponding DNA sequences due to degeneracy or self-crossings (ii) these methods may lead to poor performance due to not being able to effectively deal with short and long DNA sequences simultaneously. In an attempt to overcome these technical difficulties, our paper provides a completely novel method for DNA similarity analysis based on a combination of graphical representation with tools from algebraic topology in particular persistent homology. Our method also employs a graphical representation as the first step to transform each DNA sequence into a collection of points in a Euclidean space $\mathbb{R}^n$ which can be viewed as **point clouds** in topological data analysis. Instead of using simple geometry-based methods for analyzing these data points, we apply tools from algebraic topology such as persistent homology to obtain topological signatures of these point clouds that are signified via their persistence diagrams. Each DNA sequence, thus, is in correspondence with its unique persistence diagrams which encodes its topological signatures such as how many connected components, or how many 1-dimensional holes are present in the topology of DNA sequences. Using the well-known Wasserstein distance (which will be reviewed later) for such persistence diagrams, our key observation is that the similarities among DNA sequences are reflected by the Wasserstein distances among their corresponding persistence diagrams. One important feature of our method is its highly distinctive visuality of geometry and topology of DNA sequences. In other words, the DNA sequences, in many cases, can be immediately distinguished by highly distinctive visuality of their corresponding persistence diagrams which are diagrams in $\mathbb{R}^2$. Another outstanding feature in our method is that it can effectively deal with various lengths of DNA sequences. Our topological framework is general, alignment-free, and can deal with DNA sequences with only partial genome information while providing first-of-the-kind visualization features.

The rest of the paper is organized as follows. In Section 2, we introduce a new higher dimensional (4-dimensional) representation of DNA sequence based on chaos geometry. This new 4-D representation will serve as the basis for building our topological representation of the DNA sequences. In Section 3, we review some basic notions related to persistence homology and formally introduce out method in Section 4. We apply our method for analyzing three datasets in Section 5.

## 2 Chaos 4-dimensional Representation

In this section, we employ 4**-dimensional chaos** (see [30]) to transform a DNA sequence into a finite set of points in $\mathbb{R}^4$ that can be viewed as a 4-dimensional representation of the DNA sequence. To the best of our knowledge, this is the first time that chaos in higher dimensional space has been adopted to encode DNA sequences although chaos game in 2-dimensional space was already used to represent DNA sequences in previous work (see, for example, [31]). Let $\alpha$ be a DNA sequence of length $n$ of the form $\beta_1\beta_2\cdots\beta_n$, where the $\beta_i$ denotes one of 4 nucleotide bases A, C, G, T. We first map $\beta_1\beta_2\cdots\beta_n$ to a sequence of integers $(a_1, a_2, \ldots, a_n)$, where

$$a_i = \begin{cases} 1 & \text{if } \beta_i \text{ is nucleotide A,} \\ 2 & \text{if } \beta_i \text{ is nucleotide C,} \\ 3 & \text{if } \beta_i \text{ is nucleotide G,} \\ 4 & \text{if } \beta_i \text{ is nucleotide T.} \end{cases}$$

Set $e_1 = (1,0,0,0)$, $e_2 = (0,1,0,0)$, $e_3 = (0,0,1,0)$, and $e_4 = (0,0,0,1)$, the four standard unit vectors in $\mathbb{R}^4$. Using the 4-dimensional chaos, we construct the finite set of points $X_\alpha$ in $\mathbb{R}^4$, consisting of points $b_1, \ldots, b_n \in \mathbb{R}^4$ as follows.

(i) $b_1 = e_{a_1}$; and

(ii) for each $2 \leq k \leq n$, set $b_k = \frac{1}{2}b_{k-1} + \frac{1}{2}e_{a_k}$.

Intuitively, $b_k$ is the $k$-th point in the 3-simplex that is chosen to be the midpoint of the line segment that connects the $(k-1)$-th point and the vertex $e_{a_k}$. The map that transforms each DNA sequence $\alpha = \beta_1\beta_2\cdots\beta_n$ to a finite set of points $X_\alpha = \{b_1, \ldots, b_n\}$ in $\mathbb{R}^4$, is called the **Chaos 4-dimensional Representation** (C4DR).

# 3 Persistent homology and persistence diagrams

In this section, we recall the notion of persistent homology and persistence diagrams that we be used in our method for analyzing DNA sequences. One of the main references for persistent homology is [32].

## 3.1 Homology groups of a simplicial complex

**Definition 3.1.** ($k$-simplex)

Let $k$ be a nonnegative integer, and let $u_0, \ldots, u_k$ be $k+1$ points in $\mathbb{R}^{k+1}$. A $k$-simplex $\sigma$ generated by $\{u_0, \ldots, u_k\}$ is the convex hull of $\{u_0, \ldots, u_k\}$, i.e., the set consisting of all convex combinations of these points that is given by

$$\sigma = \{\sum_{i=0}^{k} \alpha_i u_i \mid \sum_{i=0}^{k} \alpha_i = 1 \ \text{ and } \ 0 \leq \alpha_i \leq 1\}.$$

Throughout this paper, we denote by $[u_0, \ldots, u_k]$ the $k$-simplex generated by the points $u_0, \ldots, u_k$.

Intuitively, a 0-simplex is a point in $\mathbb{R}$, a 1-simplex is a line in $\mathbb{R}^2$, and a 2-simplex is a triangle in $\mathbb{R}^3$ (see Figure for illustration of these simplices).

The convex hull of any subset of $\{u_0, \ldots, u_k\}$ with $d+1$ elements is also a $d$-simplex, and is called a **face of** $\sigma$.

**Definition 3.2.** (simplicial complex)

A simplicial complex $\Delta$ is a collection of simplices such that whenever $\sigma$ is a simplex in $\Delta$, all the faces of $\sigma$ are contained in $\Delta$.

A key notion for constructing persistent homology of a set of points in $\mathbb{R}^n$ is a formal sum of $j$-simplices. A **formal sum of $j$-simplices** is an object of the form $\sum_h a_h\sigma_h$, where the $a_h$ are real numbers in $\mathbb{R}$ such that all but finitely many $a_h$ are zero, and the $\sigma_h$ are $j$-simplices. Two formal sums $\sum_h a_h\sigma_h$ and $\sum_h b_h\sigma_h$ can be added in a natural way as

$$\sum_h a_h\sigma_h + \sum_h b_h\sigma_h = \sum_h (a_h + b_h)\sigma_h.$$

Equipped with this natural addition "+", the collection of all formal sums $\sum_{h\in H} a_h\sigma_h$ becomes a **group**–a mathematical structure in Algebra in which one can add, and subtract its elements in a similar way as the real numbers do.

Let $\Delta$ be a simplicial complex in $\mathbb{R}^n$. For $j \geq 0$, the **$j$-th chain group** $\mathcal{C}_j(\Delta)$ is the **group of all formal sums of $j$-simplices** $\sum_h a_h\sigma_h$, where the $\sigma_h$ are $j$-simplices in $\Delta$. Each formal sum in $\mathcal{C}_j(\Delta)$ is a **$j$-chain**.

There is a natural map $\partial_j$ which can send an element in $\mathcal{C}_j(\Delta)$ to $\mathcal{C}_{j-1}(\Delta)$ by removing one point from the generating set of a formal sum, and taking the alternating sum of them. It suffices to give the equation of $\partial_j$ for each $j$-simplex $[u_0, \ldots, u_j]$ in $\Delta$ that is given by

$$\partial_j([u_0, \ldots, u_j]) = \sum_{h=0}^{j} (-1)^h [u_0, \ldots, u_{h-1}, u_{h+1}, \ldots, u_j] \in \mathcal{C}_{j-1}(\Delta).$$

Thus one obtains the map, called the **$j$-th boundary map** $\partial_j : \mathcal{C}_j(\Delta) \to \mathcal{C}_{j-1}(\Delta)$.

The $j$-th chain group $\mathcal{C}_j(\Delta)$ contains two important subgroups $\mathcal{Z}_j(\Delta)$ and $\mathcal{B}_j(\Delta)$. The former, $\mathcal{Z}_j(\Delta)$, called the **$j$-th cycle group**, consists of all $j$-chains $\sigma$ in $\mathcal{C}_j(\Delta)$ such that $\partial_j(\sigma) = 0$. For example, for any three distinct points $u_0, u_1, u_2$ in $\mathbb{R}^2$, the 1-chain $\sigma = [u_0, u_1] + [u_1, u_2] + [u_2, u_3]$ is a 1-cycle since $\partial_1(\sigma) = 0$. The latter, $\mathcal{B}_j(\Delta)$, called the **$j$-th boundary group**, consists of all $j$-chains $\partial_{j+1}(\sigma)$, where $\sigma$ varies over the $(j+1)$-th chain group $\mathcal{C}_{j+1}(\Delta)$. The fundamental theorem of homology implies that $\mathcal{Z}_j(\Delta) \subset \mathcal{B}_j(\Delta)$.

The following is one of the most important notions that we will use in our method.

**Definition 3.3.** (homology groups)

Let $\Delta$ be a simplicial complex in $\mathbb{R}^n$. For each $j \geq 0$, the **$j$-th homology group** of $\Delta$ is the quotient group $\mathcal{H}_j(\Delta) = \mathcal{B}_j(\Delta)/\mathcal{Z}_j(\Delta)$.

## 3.2 Persistent Homology Groups and Persistence Diagrams

In this subsection, we recall the notion of persistent homology groups generated by a finite set of points in $\mathbb{R}^n$. Let $X$ be a finite set of points, say $u_1, \ldots, u_d$ in $\mathbb{R}^n$, and let $d$ denote the standard Euclidean distance in $\mathbb{R}^n$. For any $\alpha \geq 0$, let $\mathcal{VR}(X;\alpha)$ be the collection of all subsets $\sigma$ of $X$ such that the Euclidean distance between any two elements in $\sigma$ is at most $\alpha$, that is,

$$\mathcal{VR}(X;\alpha) = \{\sigma \subseteq X \mid d(a,b) \leq \alpha \text{ for any } a,b \in \sigma\}.$$

It is easy to verify that $\mathcal{VR}(X;\alpha)$ is a simplicial complex called the **$\alpha$-Vietoris-Rips complex of $X$**, and thus one can construct homology groups $\{\mathcal{H}_j(\mathcal{VR}(X;\alpha))\}_{j\geq 0}$ as in Subsection 3.1.

Now take an increasing sequence of nonnegative real numbers, say $\{\alpha_i\}_{i\geq 1}$ with $\alpha_i < \alpha_{i+1}$ for all $i$. Set $\alpha_0 = -\infty$. For each $i \geq 1$, set

$$\mathcal{VR}_i = \mathcal{VR}(X;\alpha_i),$$

and $\mathcal{VR}_0 = \mathcal{VR}(X;\alpha_0) = \emptyset$. Then one obtains a filtration of the form

$$\emptyset = \mathcal{VR}_0 \subseteq \mathcal{VR}_1 \subseteq \cdots \subseteq \mathcal{VR}_s = \mathcal{VR}_{s+1} = \cdots,$$

where $\alpha_s$ is large enough such that all pairs of points in $X$ are within $\alpha_s$. For each $0 \leq p \leq q \leq s$, there is a natural map $\partial_j^{p,q} : \mathcal{H}_j(\mathcal{VR}_p) \to \mathcal{H}_j(\mathcal{VR}_q)$ induced from the embedding $\mathcal{VR}_p \subseteq \mathcal{VR}_q$. We denote by $\mathrm{Im}(\partial_j^{p,q})$ the image of the map $\partial_j^{p,q}$.

**Definition 3.4.** (persistent homology groups)

For each $j \geq 0$, the **$j$-th persistent homology groups of $X$** are the images $\mathcal{H}_j^{p,q}(X) = \mathrm{Im}(\partial_j^{p,q})$ for $0 \leq p \leq q \leq s$.

By a **$j$-topological feature of $X$**, we mean an element $\gamma$ in $\mathcal{H}_j^{p,p}(X)$ for some $0 \leq p \leq s$. The **$j$-th persistence diagram of $X$** is a set of points $\{b,d) \ |0 \leq b < d\}$, where each point $(b,d)$ signifies the birth and death times of a $j$-topological feature $\gamma$ of $X$, i.e., $b$ is the radius in which $\gamma$ first appears in $\mathcal{VR}_b$ and $d$ is the radius in which $\gamma$ gets filled in with a lower dimensional simplex. We denote by $\mathcal{PD}_j(X)$ the $j$-th persistence diagram of $X$. In our methods, it suffices to consider only the 0-th and 1-th persistence diagrams, which correspond to topological features of connectedness and 1-dimensional holes of $X$, respectively.

Let $X, Y$ be two finite sets of points in $\mathbb{R}^n$. In order to compare topological features of $X, Y$ in our methods, we consider the **Wasserstein distance of degree** 1 between $\mathcal{PD}_1(X)$ and $\mathcal{PD}_1(Y)$, i.e.,

$$W_1(X,Y) = \inf_{\delta:\mathcal{PD}_1(X)\to\mathcal{PD}_1(Y)} \sum_{(b,d)\in\mathcal{PD}_1(X)} ||(b,d) - \delta(b,d)||_\infty,$$

where $||\cdot||_\infty$ denotes the $L_\infty$-distance between two points in $\mathbb{R}^2$.

# 4 Method

Our proposed method for reconstructing a phylogenetic tree of DNA sequences is described in the following algorithm:

(0) (Input) A collection of $n$ DNA sequences $\alpha_1, \dots, \alpha_n$.

(1) Construct Chaos 4-dimensional Representation (C4DR) of each DNA sequence $\alpha_i$ to obtain a finite set of points $X_{\alpha_i}$ in $\mathbb{R}^4$ (see Section 2).

(2) Compute the 1st persistence diagrams of the $X_{\alpha_i}$ to obtain the sets $\mathcal{PD}_1(X_{\alpha_i})$ in $\mathbb{R}^2$, using the notions in Section 3 [1].

(3) Compute the distance matrix of dimensions $n \times n$ whose $(i,j)$-entry is the Wasserstein distance $W_1(\mathcal{PD}_1(X_{\alpha_i}), \mathcal{PD}_1(X_{\alpha_j}))$.

(4) (Ouput) Construct the phylogenetic tree of the DNA sequences from the distance matrix in Step 3, using UPGMA algorithm (see [33]).

# 5 Results

In this section, we apply out method described in Section 4 to analyzing three datasets: Human Papillomavirus (HPV), Hantavirus, and Influenza A virus.

## 5.1 Human Papillomavirus (HPV)

Let us begin with the dataset of HPV. Human Papillomavirus is mostly responsible for cervical cancer which is the second most common cancer among women (see [34]). We apply our method on the data set of 400 HPV genomes that consist of 12 genotypes 6, 11, 16, 18, 31, 33, 35, 45, 52, 53, 58, and 66. Note that among these genotypes, there are low risk HPV types such as 6 and 11, and high risk HPV types such as 16 and 18. These high risk HPV types are responsible for about 70% of cervical cancer (see [35]). Thus it is an important problem to accurately classify HPV into low and high risk types. In addition, an ideal method should be able to identify HPV genotypes when only partial genomes are available. Our proposed method in Section 4 have all features to become a suitable and good candidate for efficiently classifying HPV genotypes. In addition it has a highly distinctive visuality that can clearly distinguish HPV genotypes in terms of visualization. For example, Figure 1 illustrates identical persistence diagrams of subtypes 11 and 15 of the same HPV genotype 6 whose highly identical visualization shows that they should belong in the same HPV genotype. On the other hand, the distinctive visualization between subtype 44 of HPV genotype 16 and subtype 18 of HPV genotype 18 in Figure 2 shows that they should belong in different genotypes of HPV. In fact the Wasserstein distance between them is 1.548–the maximum distance between any two genome sequences of HPV from the data set.

[1] We use Python packages from https://pypi.org/project/persim/ to compute persistence diagrams and Wasserstein distances.

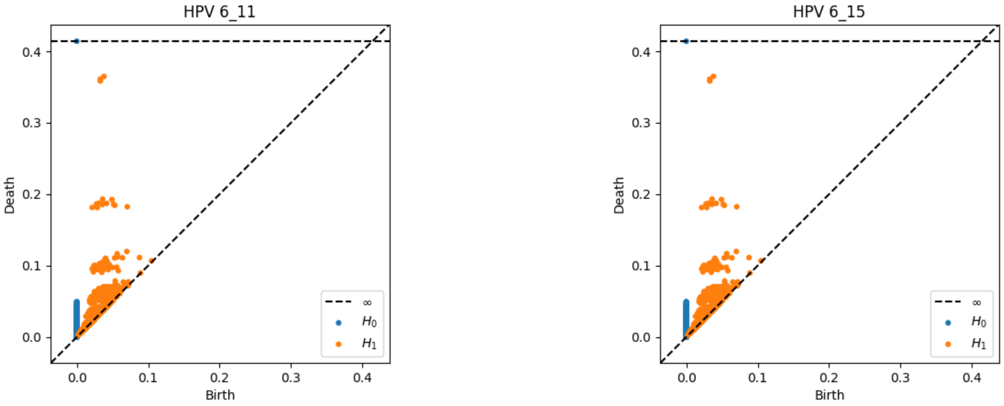

Figure 1: Persistence diagrams of HPV genomes with the minimum Wasserstein distance 0.

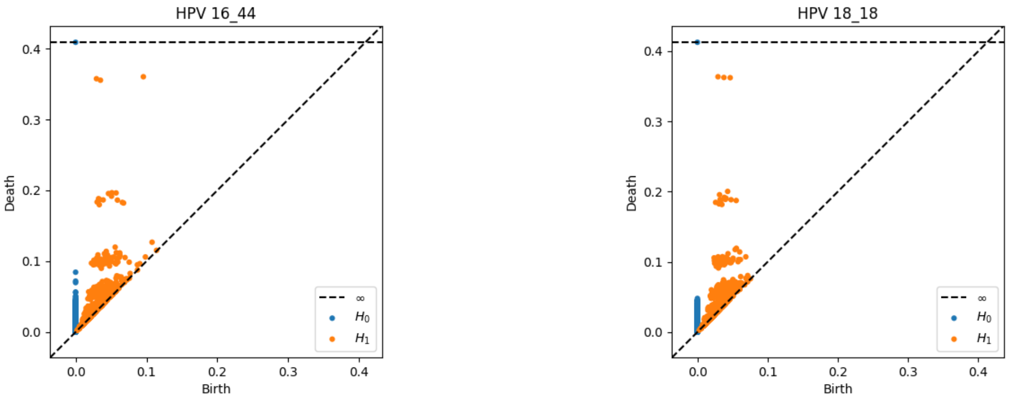

Figure 2: Persistence diagrams of HPV genomes with the maximum Wasserstein distance 1.548.

From the phylogenetic tree of HPV genomes based on our method in Figure 5, our method accurately classifies these HPV genomes into their corresponding genotypes. Note that in [1], their CGR method misplaced one genome of HPV type 11, and they further showed that Clustal Omega (see [36]) is able to classify the dataset of HPV correctly.

## 5.2 Hantavirus (HV)

Hantavirus, named after Hanta River in South Korea, is a recently discovered RNA virus in the family *Bunyaviridae*. HV may infect humans, and some strains of HV can cause possibly fatal diseases in humans such as *Hantavirus hemorrhagic fever with renal syndrome* (HFRS) or *hantavirus hemorrhagic fever with renal syndrome* (HPS). In Eastern Asia, the type of HV that causes HFRS, mainly include Hantaan (HTN) and Seoul (SEO) viruses. In Western European, Russia, and Northeastern China, Puumala (PUU) is the type of HV that causes HFRS. There is another genus of the family *Bunyaviridae* that is called Phlebovirus (PV). We apply our method on the data set of 34 HV genome sequences consisting of 4 different types HTN, SEO, PUU, and PV. The name of these strains are included in the phylogenetic tree (see Figure 3). From Figure 3, we find that our method accurately cluster CGRN strains together whose host is *Rattus norvegicus*. Similarly the method correctly clusters four CGAa strains together whose host is *Apodemus agrarius*. In addition, two CGHu strains whose host is Homo sapiens, is also grouped together.

Except strain Sotkamo (type PUU) forms an independent leave, we find that all leaves are classified accurately into their corresponding types. But branches of the phylogenetic tree constructed by our method contain both types as sub-branches which are quite different from the results in [37, 38, 39].

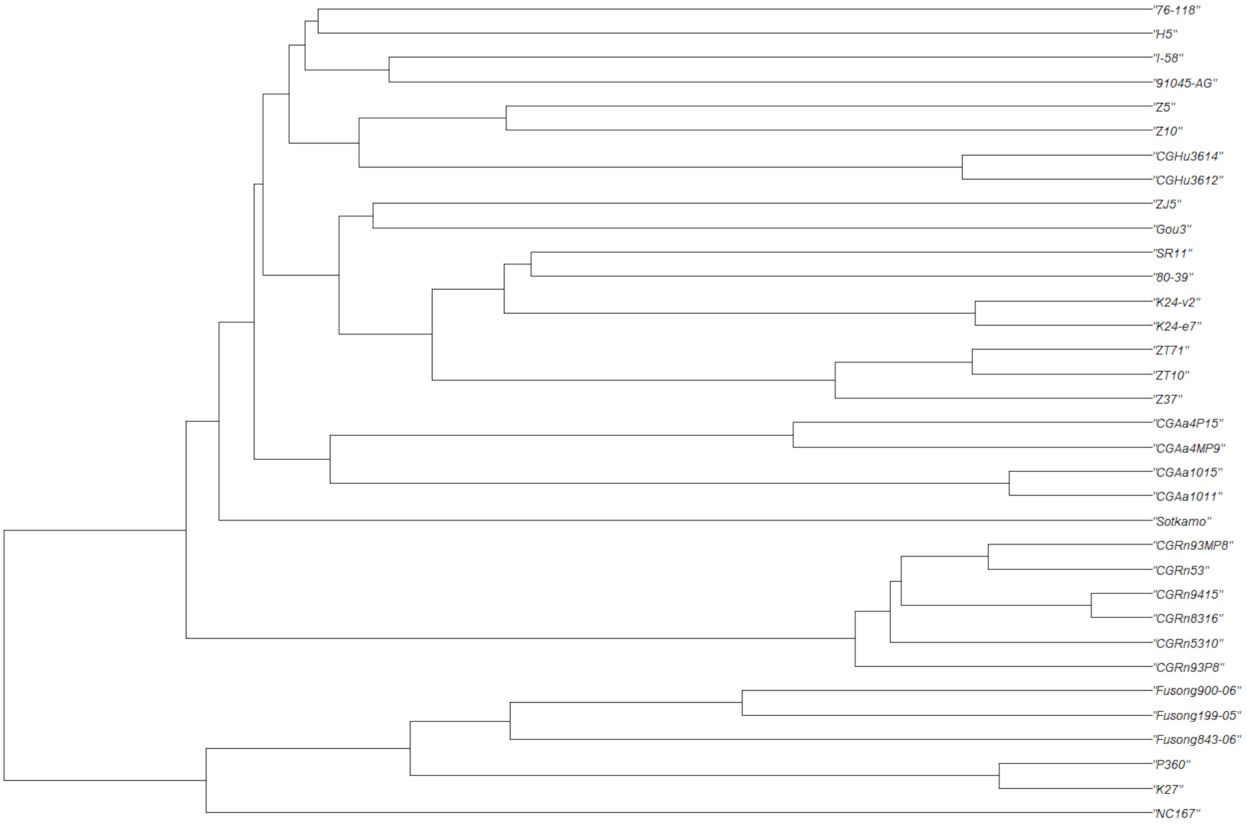

Figure 3: Phylogenetic tree of Hantavirus genomes

### 5.3 Influenza

Influenza A viruses are very dangerous because their hosts have a wide range including birds, horses, swine, and humans. These viruses have been a serious health threat to humans and animals (see [40]), are known to have high degree of genetic and antigenic variability (see [41, 42]). Some subtypes of Influenza A viruses are very lethal including H1N1, H2N2, H5N1, H7N3, and H7N9. We apply our method on the dataset consisting of 38 nfluenza A virus genomes. From Figure 4, we find that except A/emperor goose/Alaska/44297-260/2007 (H2N2) and A/turkey/VA/505477-18/2007 (H5N1), the *tips (or leaves)* of the phylogenetic tree of segment 6 of Influenza A virus genomes are clustered correctly into their types, but, for example, for type H1N1 genomes, our method clusters them into 3 distinct subbranches of the tree. We will address this problem in an ongoing work.

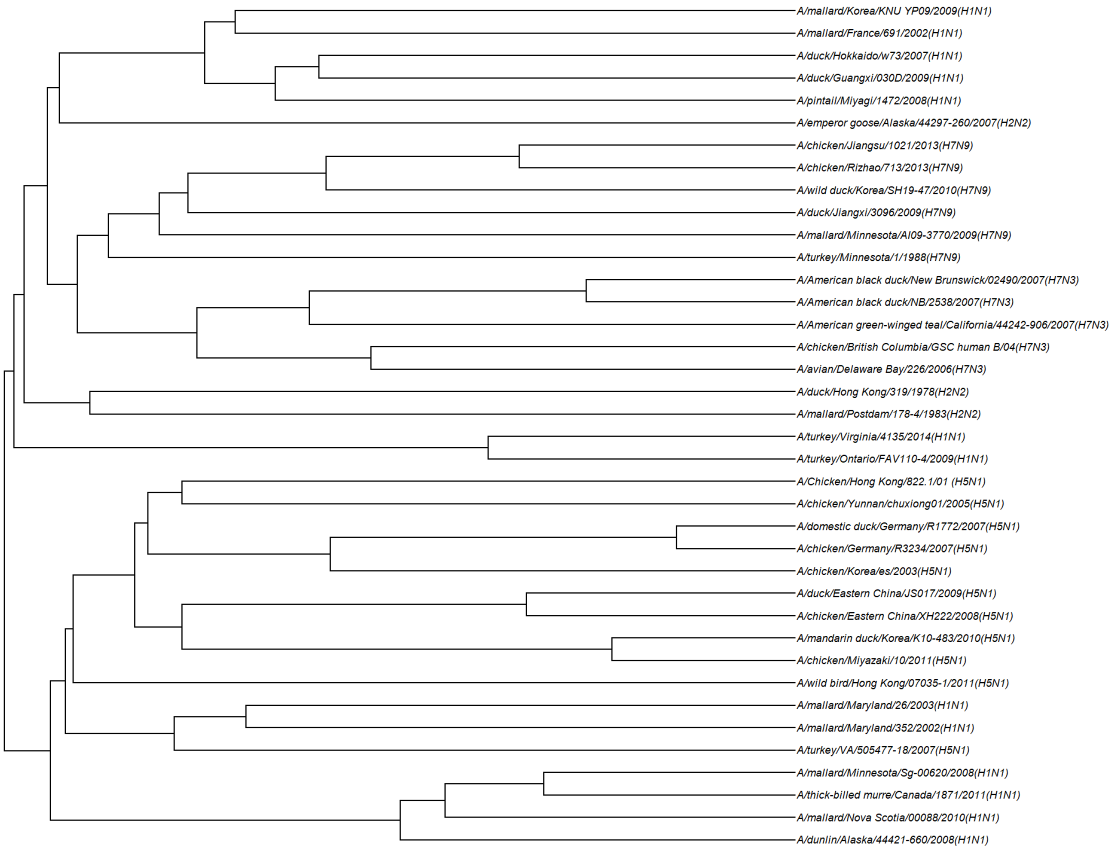
A/mallard/Korea/KNU YP09/2009(H1N1)
A/mallard/France/691/2002(H1N1)
A/duck/Hokkaido/w73/2007(H1N1)
A/duck/Guangxi/030D/2009(H1N1)
A/pintail/Miyagi/1472/2008(H1N1)
A/emperor goose/Alaska/44297-260/2007(H2N2)
A/chicken/Jiangsu/1021/2013(H7N9)
A/chicken/Rizhao/713/2013(H7N9)
A/wild duck/Korea/SH19-47/2010(H7N9)
A/duck/Jiangxi/3096/2009(H7N9)
A/mallard/Minnesota/AI09-3770/2009(H7N9)
A/turkey/Minnesota/1/1988(H7N9)
A/American black duck/New Brunswick/02490/2007(H7N3)
A/American black duck/NB/2538/2007(H7N3)
A/American green-winged teal/California/44242-906/2007(H7N3)
A/chicken/British Columbia/GSC human B/04(H7N3)
A/avian/Delaware Bay/226/2006(H7N3)
A/duck/Hong Kong/319/1978(H2N2)
A/mallard/Postdam/178-4/1983(H2N2)
A/turkey/Virginia/4135/2014(H1N1)
A/turkey/Ontario/FAV110-4/2009(H1N1)
A/Chicken/Hong Kong/822.1/01 (H5N1)
A/chicken/Yunnan/chuxiong01/2005(H5N1)
A/domestic duck/Germany/R1772/2007(H5N1)
A/chicken/Germany/R3234/2007(H5N1)
A/chicken/Korea/es/2003(H5N1)
A/duck/Eastern China/JS017/2009(H5N1)
A/chicken/Eastern China/XH222/2008(H5N1)
A/mandarin duck/Korea/K10-483/2010(H5N1)
A/chicken/Miyazaki/10/2011(H5N1)
A/wild bird/Hong Kong/07035-1/2011(H5N1)
A/mallard/Maryland/26/2003(H1N1)
A/mallard/Maryland/352/2002(H1N1)
A/turkey/VA/505477-18/2007(H5N1)
A/mallard/Minnesota/Sg-00620/2008(H1N1)
A/thick-billed murre/Canada/1871/2011(H1N1)
A/mallard/Nova Scotia/00088/2010(H1N1)
A/dunlin/Alaska/44421-660/2008(H1N1)

Figure 4: Phylogenetic tree of 38 Influenza A virus genomes

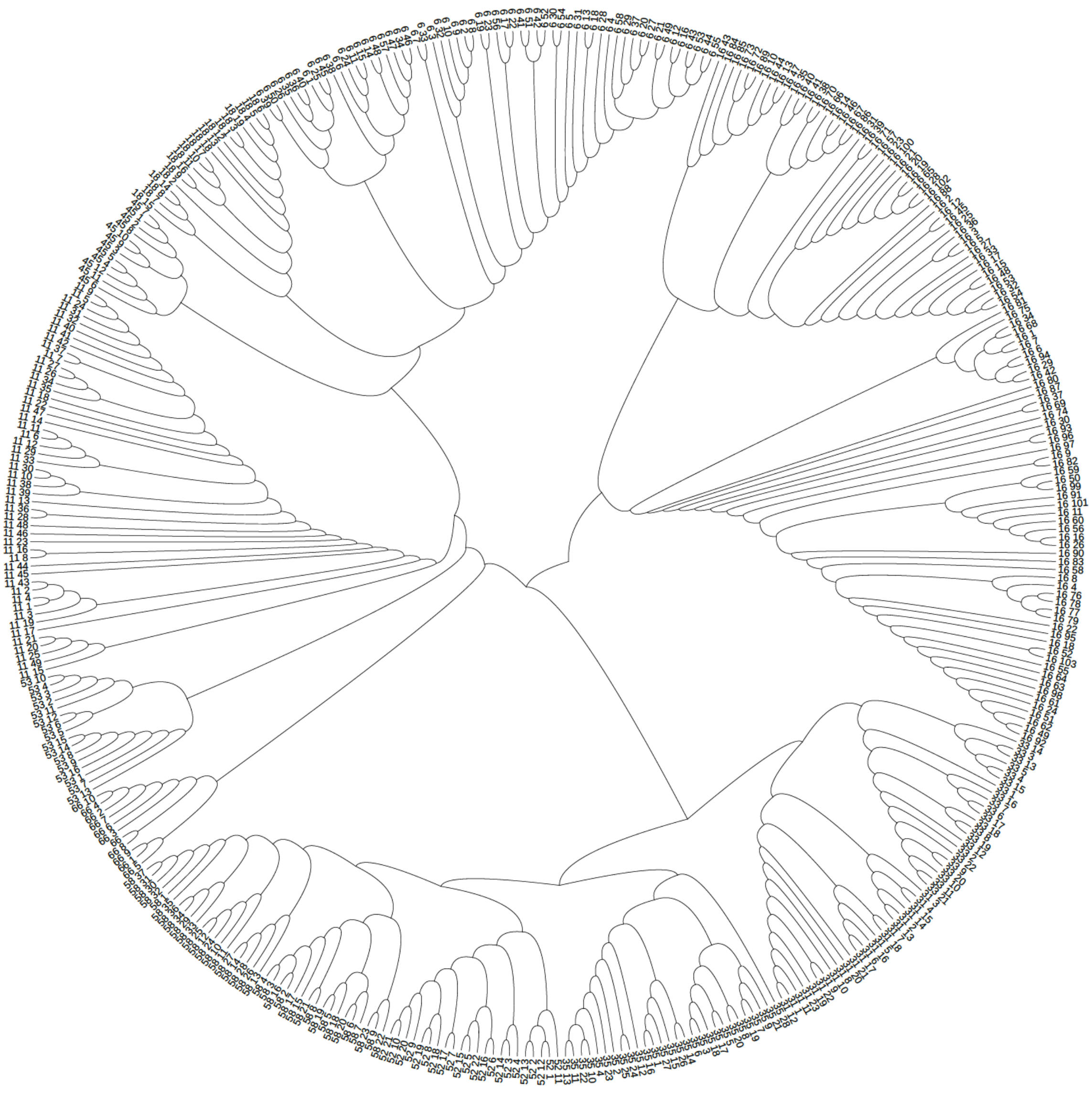

Figure 5: Phylogenetic tree of 400 HPV genomes of 12 genotypes

# References

[1] T. Hoang, C. Yin, and S. S.-T. Yau, "Numerical encoding of dna sequences by chaos game representation with application in similarity comparison.," *Genomics*, 2016.

[2] X. Jin, R. Nie, D. Zhou, S. Yao, Y. Chen, J. Yu, and Q. Wang, "A novel dna sequence similarity calculation based on simplified pulse-coupled neural network and huffman coding," *Physica A: Statistical Mechanics and its Applications*, vol. 461, pp. 325–338, 2016.

[3] J.-F. Yu, J.-H. Wang, and X. Sun, "Analysis of similarities/dissimilarities of dna sequences based on a novel graphical representation," *MATCH Commun. Math. Comput. Chem*, vol. 63, no. 2, pp. 493–512, 2010.

[4] S. Wang, F. Tian, Y. Qiu, and X. Liu, "Bilateral similarity function: A novel and universal method for similarity analysis of biological sequences," *Journal of theoretical biology*, vol. 265, no. 2, pp. 194–201, 2010.

[5] N. Jafarzadeh and A. Iranmanesh, "C-curve: a novel 3d graphical representation of dna sequence based on codons," *Mathematical Biosciences*, vol. 241, no. 2, pp. 217–224, 2013.

[6] B. Liao, Y. Zhang, K. Ding, and T.-M. Wang, "Analysis of similarity/dissimilarity of dna sequences based on a condensed curve representation," *Journal of Molecular Structure: THEOCHEM*, vol. 717, no. 1-3, pp. 199–203, 2005.

[7] E. Hamori and J. Ruskin, "H curves, a novel method of representation of nucleotide series especially suited for long dna sequences.," *Journal of Biological Chemistry*, vol. 258, no. 2, pp. 1318–1327, 1983.

[8] B. Liao, M. Tan, and K. Ding, "A 4d representation of dna sequences and its application," *Chemical Physics Letters*, vol. 402, no. 4-6, pp. 380–383, 2005.

[9] J. Wang and Y. Zhang, "Characterization and similarity analysis of dna sequences grounded on a 2-d graphical representation," *Chemical physics letters*, vol. 423, no. 1-3, pp. 50–53, 2006.

[10] X. Q. Liu, Q. Dai, Z. Xiu, and T. Wang, "Pnn-curve: A new 2d graphical representation of dna sequences and its application," *Journal of Theoretical Biology*, vol. 243, no. 4, pp. 555–561, 2006.

[11] X.-Q. Qi, J. Wen, and Z.-H. Qi, "New 3d graphical representation of dna sequence based on dual nucleotides," *Journal of Theoretical Biology*, vol. 249, no. 4, pp. 681–690, 2007.

[12] C. Li, X. Yu, and N. Helal, "Similarity analysis of dna sequences based on codon usage," *Chemical Physics Letters*, vol. 459, no. 1-6, pp. 172–174, 2008.

[13] N. Jafarzadeh and A. Iranmanesh, "A novel graphical and numerical representation for analyzing dna sequences based on codons," *Match-Communications in Mathematical and Computer Chemistry*, vol. 68, no. 2, p. 611, 2012.

[14] F. Kabli, H. R. Mohamed, and A. Abdelmalek, "Similarity analysis of dna sequences based on the chemical properties of nucleotide bases: frequency and position of group mutations," *Comput. Sci. Inf. Technol.*, vol. 6, no. 1, pp. 1–10, 2016.

[15] Q. Dai, X. Liu, and T. Wang, "A novel 2d graphical representation of dna sequences and its application," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 3, pp. 340–344, 2006.

[16] Y.-H. Yao, X.-Y. Nan, and T.-M. Wang, "A new 2d graphical representation—classification curve and the analysis of similarity/dissimilarity of dna sequences," *Journal of Molecular Structure: THEOCHEM*, vol. 764, no. 1-3, pp. 101–108, 2006.

[17] B. Liao, Q. Xiang, L. Cai, and Z. Cao, "A new graphical coding of dna sequence and its similarity calculation," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 19, pp. 4663–4667, 2013.

[18] P.-a. He and J. Wang, "Characteristic sequences for dna primary sequence," *Journal of Chemical Information & Modeling*, 2002.

[19] W. Hou, Q. Pan, and M. He, "A novel representation of dna sequence based on cmi coding," *PHYSICA A*, 2014.

[20] R. Zhang and C. Zhang, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," *J Biomol Struct Dyn*, vol. 11, no. 4, pp. 767–782, 1994.

[21] R. Zhang and C. Zhang, "A brief review: The z-curve theory and its application in genome analysis," *Curr Genomics*, vol. 15, no. 2, pp. 78–94, 2014.

[22] C. Yin, Y. Chen, and S. S.-T. Yau, "A measure of dna sequence similarity by fourier transform with applications on hierarchical clustering," *Journal of theoretical biology*, vol. 359, pp. 18–28, 2014.

[23] C.-K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. Stanley, "Fractal landscape analysis of dna walks," *Physica A: Statistical Mechanics and its Applications*, vol. 191, no. 1-4, pp. 25–29, 1992.

[24] S. Buldyrev, N. Dokholyan, A. Goldberger, S. Havlin, C.-K. Peng, H. Stanley, and G. Viswanathan, "Analysis of dna sequences using methods of statistical physics," *Physica A: Statistical Mechanics and its Applications*, vol. 249, no. 1-4, pp. 430–438, 1998.

[25] M. Randić and M. Vracko, "On the similarity of dna primary sequences," *Journal of chemical information and computer sciences*, vol. 40, no. 3, pp. 599–606, 2000.

[26] F. Bai, Y. Liu, and T. Wang, "A representation of dna primary sequences by random walk," *Mathematical biosciences*, vol. 209, no. 1, pp. 282–291, 2007.

[27] X. Jin, D. Zhou, S. Yao, R. Nie, Q. Wang, and K. He, "Analysis of similarity/dissimilarity of dna sequences based on pulse coupled neural network," in *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, 2016.

[28] J. Zheng, J. Zhang, F. Bai, and L. Ma, "Similarity analysis of dna sequences based on the mq-emd method," *Journal of Computational Information Systems*, vol. 8, no. 23, pp. 9823–9830, 2012.

[29] J. Zhang, R. Wang, F. Bai, and J. Zheng, "A quasi-mq emd method for similarity analysis of dna sequences," *Applied Mathematics Letters*, 2011.

[30] D. P. Feldman, *Chaos and fractals : an elementary introduction*. Oxford: Oxford University Press, 2012.

[31] J. H. Joel, "Chaos game representation of gene structure.," *Nucleic Acids Research*, no. 8, pp. 2163–2170, 1990.

[32] H. Edelsbrunner and J. Harer, *Computational Topology - an Introduction*. American Mathematical Society, 2010.

[33] S. Kumar, G. Stecher, and K. Tamura, "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets," *Molecular Biology and Evolution*, vol. 33, pp. 1870–1874, 03 2016.

[34] M. Arbyn, X. Castellsague, S. D. Sanjose, L. Bruni, M. Saraiya, F. Bray, and J. Ferlay, "Worldwide burden of cervical cancer in 2008," *Ann. Oncol.*, vol. 22, pp. 2675–2686, 2011.

[35] J. Smith, L. Lindsay, B. Hoots, J. Keys, S. Franceschi, R. Winer, and G. Clifford, "Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update," *Int. J. Cancer*, vol. 121, pp. 621–632, 2007.

[36] F. Sievers, A. Wilm, D. Dineen, T. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, and e. a. J. Söding, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," *Mol. Syst. Biol.*, vol. 7, p. 539, 2011.

[37] C. Li, W. Fei, Y. Zhao, and X. Yu, "Novel graphical representation and numerical characterization of dna sequences," *Applied Sciences*, vol. 6, no. 3, p. 63, 2016.

[38] L. Meng, C. Li, M. Jia, Y. Zhang, and Y. Yang, "Non-degenerate graphical representation of dna sequences and its applications to phylogenetic analysis," *Combinatorial Chemistry & High Throughput Screening*, vol. 16, no. 8, pp. 585–589, 2013.

[39] P.-P. Yao, H.-P. Zhu, X.-Z. Deng, F. Xu, R.-H. Xie, C.-H. Yao, J.-Q. Weng, Y. Zhang, Z.-Q. Yang, and Z.-Y. Zhu, "Molecular evolution analysis of hantaviruses in zhejiang province," *Bing du xue bao = Chinese journal of virology*, vol. 26, no. 6, p. 465, 2010.

[40] D. Alexander, "A review of avian influenza in different bird species," *Vet. Microbiol.*, vol. 74, pp. 3–13, 2000.

[41] R. Garten, C. Davis, C. Russell, B. Shu, S. Lindstrom, A. Balish, W. Sessions, E. S. X. Xu, and e. a. V. Deyde, "Antigenic and genetic characteristics of swine-origin 2009 a (h1n1) influenza viruses circulating in humans," *Science*, vol. 325, pp. 197–201, 2009.

[42] P. Palese and J. Young, "Variation of influenza a, b, and c viruses," *Science*, vol. 215, pp. 1468–1474, 1982.