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

* Type of submission: full research paper, acronym: CSCI-RTHI

1st Dong Quan Ngoc Nguyen
Yanqi Lake Beijing Institute of Mathematical Sciences and Applications
Beijing 101408, China
donguan.ngoc.nguyen@gmail.com

2nd Phuong Dong Tan Le
Department of Applied Mathematics
University of Waterloo
Waterloo, Canada
pdlle@uwaterloo.ca

3rd Lin Xing
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
Notre Dame, USA
lxing@nd.edu

4th Lizhen Lin
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
Notre Dame, USA
lizhen.lin@nd.edu

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 genome analysis. 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.

Index Terms—DNA similarity analysis, persistent homology, chaos geometry

I. 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, for example, references [2]–[21].

In addition to graphical representations, numeric representations are also used for DNA similarity analysis, which maps each DNA sequence to a sequence of numbers. 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, [22]–[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 (see [30] for this notion). Our method also employs a graphical representation by first transforming 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 encode 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 visualization of their corresponding persistence diagrams which are collections of points 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 II, we introduce a new higher dimensional (4-dimensional) representation of DNA sequence based on chaos geometry. This new 4-D representation will be used for building our topological representation of the DNA sequences. In Section III, we review some basic notions related to persistent homology and formally introduce our method in Section IV. We apply our method for analyzing three datasets in Section V, and compare with other state-of-the-art methods such as Clustal Omega (see [31]).

II. CHAOS 4-DIMENSIONAL REPRESENTATION

In this section, we employ 4-dimensional chaos (see [32]) 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. Note that chaos game in 2-dimensional space was already used to represent DNA sequences in previous work (see, for example, [33]). More specifically, 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) 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).

III. 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 [30] (see also [34] for a detailed account of persistent homology).

A. Homology groups of a simplicial complex

Definition III.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

\[
\sigma = \{ \sum_{i=0}^{k} \alpha_i u_i \mid \sum_{i=0}^{k} \alpha_i = 1 \text{ and } 0 \leq \alpha_i \leq 1 \}. \tag{2}
\]

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 III.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

\[
\sum_h a_h \sigma_h + \sum_h b_h \sigma_h = \sum_h (a_h + b_h) \sigma_h. \tag{3}
\]

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 $C_j(\Delta)$ is the group of all formal sums of $j$-simplices $\sum a_\sigma \sigma$, where the $\sigma$ are $j$-simplices in $\Delta$. Each formal sum in $C_j(\Delta)$ is a $j$-chain.

There is a natural map $\partial_j$, which can send an element in $C_j(\Delta)$ to $C_{j-1}(\Delta)$ 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

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

Thus one obtains the map, called the $j$-th boundary map $\partial_j : C_j(\Delta) \to C_{j-1}(\Delta)$.

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

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

**Definition III.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 $H_j(\Delta) = B_j(\Delta)/Z_j(\Delta)$.

**B. 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 $x_1, \ldots, x_n$ 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,

$$\mathcal{VR}(X; \alpha) = \{ \sigma \subseteq X : d(a, b) \leq \alpha \text{ for any } a, b \in \sigma \}. \quad (5)$$

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 $\{H_j(\mathcal{VR}(X; \alpha))\}_{j \geq 0}$ as in Subsection III-A.

Now take an increasing sequence of nonnegative real numbers, say $\{\alpha_i\}_{i \geq 0}$ 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 $\partial^p_q : H_j(\mathcal{VR}_p) \to H_j(\mathcal{VR}_q)$ induced from the embedding $\mathcal{VR}_p \subseteq \mathcal{VR}_q$. We denote by $\text{Im} (\partial^p_q)$ the image of the map $\partial^p_q$.

**Definition III.4.** (persistent homology groups)

For each $j \geq 0$, the $j$-th persistent homology groups of $X$ are the images $H_j^{p,q}(X) = \text{Im} (\partial^p_q)$ for $0 \leq p \leq q \leq s$.

By a $j$-topological feature of $X$, we mean an element $\gamma$ in $H_j^{p,q}(X)$ 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 $PD_j(X)$ 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 $PD_1(X)$ and $PD_1(Y)$, i.e.,

$$W_1(X,Y) = \inf_{\delta : PD_1(X) \to PD_1(Y)} \sum_{(b,d) \in PD_1(X)} ||(b,d) - \delta(b,d)||_\infty,$$

(8)

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

**IV. Proposed 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 $\alpha_1, \ldots, \alpha_n$.

(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 II).

(2) Compute the 1st persistence diagrams of the $X_{\alpha_i}$ to obtain the sets $PD_1(X_{\alpha_i})$ in $\mathbb{R}^2$, using the notions in Section III. We use Python packages from https://pypi.org/project/persim/ to compute persistence diagrams and Wasserstein distances.

(3) Compute the distance matrix of dimensions $n \times n$ whose $(i,j)$-entry is the Wasserstein distance $W_1(PD_1(X_{\alpha_i}), PD_1(X_{\alpha_j}))$.

(4) (Output) Construct the phylogenetic tree of the DNA sequences from the distance matrix in Step 3, using UPGMA algorithm (see [35]).

**Remark IV.1.** In the proposed method, and our experimental analysis, for Steps 2 and 3 above, we also try to use both the 0th and 1st persistence diagrams of $X_{\alpha_i}$, and hence use both Wasserstein distances $W_0$ for 0th persistence diagrams and $W_1$ for 1st persistence diagrams to construct the phylogenetic tree of the DNA sequences. The resulting phylogenetic trees are
very similar to those constructed using only 1st persistence diagrams. Intuitively 0th persistence diagrams only provide information about how many connected components there are from the topological space induced from cloud points, and thus they do not provide a great deal about the topological structure of the cloud points as 1st persistence diagrams. So in the proposed method above, and in our experimental analysis that we will describe in next section, we only use 1st persistence diagrams and the Wasserstein distance \( W_1 \) to construct the phylogenetic trees in order to save computational time.

V. Results

In this section, we apply our method described in Section IV to analyzing three datasets: Human Papillomavirus (HPV), Hantavirus, and Influenza A virus. All of the datasets are obtained from GenBank\(^1\), which is the NIH genetic sequence database. In addition, the accession numbers of HPV and Influenza datasets can be found in [36], the accession number of Hantavirus dataset can be found in [37].

A. Human Papillomavirus (HPV)

We first consider the dataset of HPV. Human Papillomavirus is mostly responsible for cervical cancer which is the second most common cancer among women (see [38]). We apply our method on the dataset 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 [39]). Thus it is an important problem to accurately classify HPV into low and high risk types. In addition, a reasonable method should be able to identify HPV genotypes when only partial genomes are available. Our proposed method in Section IV have all the above features, and is suitable for efficiently classifying HPV genotypes. In addition it allows one to visually inspect differences between HPV genotypes. For example, Fig. 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 Fig. 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 dataset.

From the phylogenetic tree of HPV genomes based on our method in Fig. 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 [31]) is able to classify the dataset of HPV correctly.

\(^1\)See https://www.ncbi.nlm.nih.gov/genbank/

B. 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, Paumala (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 Fig. 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 [40]–[42].

C. Influenza

Influenza A viruses are very dangerous because they have a wide range of hosts including birds, horses, swine, and humans. These viruses have been a serious health threat to humans and animals (see [43]), and are known to have high degree of genetic and antigenic variability (see [44], [45]). 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 influenza A virus genomes. From Fig. 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. However, for type H1N1 genomes, our method clusters them into 3 distinct subbranches of the tree. We will address this problem in an ongoing work.

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. 329–338, 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. Qu, 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. 241, no. 2, pp. 217–224, 2013.
Fig. 4: Phylogenetic tree of 38 Influenza A virus genomes

Fig. 5: Phylogenetic tree of 400 HPV genomes of 12 genotypes
[35] S. Kumar, G. Stecher, and K. Tamura, “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets,” Molecular Biology and Evolution, vol. 33, no. 7, pp. 1870–1874, 03 2016. [Online]. Available: https://doi.org/10.1093/molbev/msw054

[36] D. Q. N. Nguyen, L. Xing, P. D. T. Le, and L. Lin, “A graph-theoretical approach to dna similarity analysis,” Communications in Information and Systems, vol. 22, no. 3, pp. 383–400, 2022.

[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] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

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