A topological approach to DNA similarity analysis from 5-dimensional representation

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

Ziqing Hu

Department of Applied and Computational Mathematics and Statistics,
University of Notre Dame,
Notre Dame, IN 46556, USA
email: zhu4@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

March 5, 2021

1 Introduction

There are two main approaches to the problem of comparing DNA sequences: alignment methods and alignment-free methods. In alignment methods, the classical multiple sequence alignment (MSA) method is widely used. Although it has the highest accuracy among DNA similarity analysis methods, time complexity becomes too large for a large dataset of long DNA sequences.In alignment-free methods, several methods have been proposed to reduce time complexity while maintaining a high accuracy in comparing DNA sequences. In general, an alignment-free method based on geometry consists of two step, the first of which embed each DNA sequence into an Euclidean space, and the second of which is to compute the similarity matrix based on certain distance-based methods such as Euclidean distances or correlation angles to realize differences or similarities among DNA sequences. In a recent work, Nguyen, Le, Xing, and Lin [1] proposed a different alignment-free approach which combines geometry with topology of DNA sequences. In [1], a new 4D representation of DNA sequences was introduced using a chaos in the four-dimensional Euclidean space $\mathbb{R}^4$. Instead of computing similarities/dissimilarities between DNA sequences based on Euclidean distances or correlation angles as in other work, Nguyen, Le, Xing, and Lin [1] proposed to explore topological properties of the sets of 4-dimensional vectors that represent DNA sequences in order to obtain intrinsic geometrical and topological structures of DNA sequences for similarity analysis. The main mathematical tools used in [1] are chaos in high dimensions, and persistent homology which has recently gained an important role in topological data analysis and related areas. In this paper, we follow a similar strategy of using persistent homology as in [1] for DNA similarity analysis, but we use a different geometric 5-dimensional representation that is based on that of Liao, Li, and Zhu [2]. In [1], Nguyen, Le, Xing, and Lin used chaos representation in the 4-dimensional space $\mathbb{R}^4$ under which persistence diagrams of DNA sequences will contain nontrivial higher-dimensional homology groups, which in turn requires more powerful computation power for performing similarity analysis between DNA sequences, based on Wasserstein distances of order at least one. The main difference of our present method from [1] is that under our geometric 5D representation of DNA sequences, persistent homology only contains the zeroth homology group, which in turn greatly simplifies time complexity for analyzing similarities between DNA sequences, based on the Wasserstein distance of order zero. Furthermore the zeroth persistence diagrams of DNA sequences based on the 5D representation of DNA sequences provide a possibly simplest topological visualization of DNA sequences which in many cases provides a quick assessment of similarities/dissimilarities between DNA sequences.

2 Method

Our method consists of two steps. In the first step, we transform each DNA sequence into the 5-dimensional Euclidean space $\mathbb{R}^5$ so that each DNA sequence of length n can be represented by a collection of n vectors in $\mathbb{R}^5$. In the second step, we compute persistent homology for each such collection of vectors to obtain the persistence diagrams of

DNA sequences which contain intrinsic topological information of each DNA sequence. The 5D representation that we use only leads to the nontrivial zeroth persistence diagrams of DNA sequences which provide a greatly simple topological visualization of DNA sequences. It is known that the collection of zeroth persistence diagrams is a metric under the Wasserstein distance of order zero. In order to compare similarities between DNA sequences, we compute the similarity/dissimilarity matrix using the Wasserstein distance.

2.1 5-dimensional representation of DNA sequences

In this subsection, we introduce a map that transforms each DNA sequence of length n into a collection of n vectors in $\mathbb{R}^5$. Our construction is based on that in Liao, Li, and Zhu [2] with slight modification. Let $\alpha = a_1 a_2 \cdots a_n$ be a DNA sequence of length n, where the a_i denotes one of 4 nucleotide bases A, C, G, T. For each $1 \leq i \leq n$, set

$$\Gamma(a_i) = \begin{cases} (1,0,0,0,i) & \text{if } a_i = A, \\ (0,1,0,0,i) & \text{if } a_i = C, \\ (0,0,1,0,i) & \text{if } a_i = G, \\ (0,0,0,1,i) & \text{if } a_i = T. \end{cases}$$

In [2], Liao, Li, and Zhu maps $a_i = G$ to $(0,1,0,0,i)$, and $a_i = C$ to $(0,0,1,0,i)$ which is a permutation of the above construction.

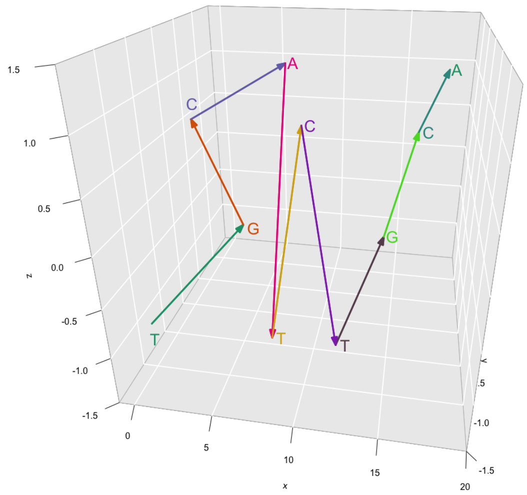

Figure 1: 5D representation of part of the DNA sequence of HRV 35-25 from nucleotides 4246 to 4250

2.2 Persistent homology and persistent diagrams

We briefly recall the notion of persistent homology and persistence diagrams that will apply for analyzing DNA sequences. The reader is referred to [3] for a detailed reference about persistent homology and persistence diagrams. For a given nonnegative integer $k \geq 0$ and a collection of $k+1$ points $u_0, \ldots, u_k$ in $\mathbb{R}^{k+1}$. One can create a **convex hull** of this collection in $\mathbb{R}^{k+1}$ by including all convex combinations of these points of the form $\sum_{i=0}^{k} \alpha_i u_i$, where the α_i are between 0 and 1 such that $\sum_{i=0}^{k} \alpha_i = 1$. We call such convex hull the k-simplex generated by the points $u_0, \ldots, u_k$, and denote it by $[u_0, \ldots, u_k]$. For a collection of simplexes in $\mathbb{R}^{k+1}$, say Δ. We call Δ a **simplicial complex** if whenever σ is a simplex in Δ, all d-simplexes contained in σ are also contained in Δ. For such a geometric object Δ in $\mathbb{R}^n$, based on algebraic topology, there exists, for each $j \geq 0$, an algebraic structure called the **j-th homology group of** Δ, denoted by $\mathcal{H}_j(\Delta)$ which behaves in a similar way as a vector space over $\mathbb{R}$. There is an analogue of dimensions of vector spaces over $\mathbb{R}$ in the setting of homology groups $\{\mathcal{H}_j(\Delta)\}_{j\geq 0}$ that we call the **rank of** $\mathcal{H}_j(\Delta)$, a positive integer, which signifies important geometric properties of Δ. For example, the rank of $\mathcal{H}_0(\Delta)$ equals the *number of connected components of* Δ in $\mathbb{R}^n$, and the rank of $\mathcal{H}_1(\Delta)$ denotes the *number of* 1*-dimensional holes* of Δ.

Let X be a finite set of points, say $a_1, \ldots, a_m$ in $\mathbb{R}^n$, and let d denote the standard Euclidean distance in $\mathbb{R}^n$. For each $\epsilon \geq 0$, set

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

One can verify that $\mathcal{VR}(X;\epsilon)$ is a simplicial complex called the **ϵ-Vietoris-Rips complex of** X. Let $\epsilon_0 = -\infty < 0 \leq \epsilon_1 \leq \cdots \leq \epsilon_h \leq \cdots$ be an increasing sequence of nonnegative real numbers. One can form a sequence of simplicial complexes $(\mathcal{VR}(X;\epsilon_k))_{k\in\mathbb{Z}_{\geq 0}}$, and one obtains a **filtration of the form**

$$\emptyset = \mathcal{VR}(X;\epsilon_0) \subseteq \mathcal{VR}(X;\epsilon_1) \subseteq \cdots \subseteq \mathcal{VR}(X;\epsilon_s) = \mathcal{VR}(X;\epsilon_{s+1}) = \cdots$$

which will stabilize at some point ϵ_s. For each $0 \leq p \leq q \leq s$, the embedding $\mathcal{VR}(X;\epsilon_p) \subset \mathcal{VR}(X;\epsilon_q)$ induces a sequence of natural maps $\partial_j^{p,q} : \mathcal{H}_j(\mathcal{VR}(X;\epsilon_p)) \to \mathcal{H}_j(\mathcal{VR}(X;\epsilon_q))$. For each $j \geq 0$, the **j-th persistent homology groups $\mathcal{H}_j^{p,q}(X)$ of** X are the images of $\partial_j^{p,q}$ which are $\partial_j^{p,q}(\mathcal{H}_j(\mathcal{VR}(X;\epsilon_p)))$.

Each element γ in $\mathcal{H}_j^{p,p}(X)$ is called a **j-topological feature of** X. The **j-th persistence diagram of** X is a set of points $\{(b,d) \mid 0 \leq b < d\} \subset \mathbb{R}^2$, where each point (b,d) signifies the **birth and death times of a j-topological feature** γ of X, i.e., b is the radius in which γ first appears in $\mathcal{VR}(X;\epsilon_b)$ and d is the radius in which γ 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 persistence diagrams, which correspond to topological features of connectedness of X.

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** 0 between $\mathcal{PD}_0(X)$ and $\mathcal{PD}_0(Y)$, i.e.,

$$W_0(X,Y) = \inf_{\delta:\mathcal{PD}_0(X)\to\mathcal{PD}_0(Y)} \sum_{(b,d)\in\mathcal{PD}_0(X)} ||(b,d) - \delta(b,d)||_\infty,$$

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

2.3 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 the 5-dimensional geometric representation of each DNA sequence α_i from Subsection 2.1 to obtain a finite set of points X_{α_i} in $\mathbb{R}^5$.

(2) Compute the 1st persistence diagrams of the X_{α_i} to obtain the sets of 0-th persistence diagrams $\mathcal{PD}_0(X_{\alpha_i})$ in $\mathbb{R}^2$, using the notions in Section 2.2. We use Python packages from `https://pypi.org/project/persim/` to compute persistence diagrams and Wasserstein distances. Note that using the 5-dimensional representation in Subsection 2.1, all j-th persistence diagrams with $j \geq 1$ are empty, which leads to a very simple way to compare DNA sequences based on the 0-th persistence diagrams.

(3) Compute the distance matrix of dimensions $n \times n$ whose (i, j)-entry is the Wasserstein distance $W_0(\mathcal{PD}_0(X_{\alpha_i}), \mathcal{PD}_0(X_{\alpha_j}))$.

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

3 Results

In this section, we apply our method described in Section 2 to analyzing three datasets: Human rhinovirus, Influenza A virus, and Human Papillomavirus (HPV).

3.1 Human rhinovirus (HRV)

HRV is the most common viral infectious agent in humans, and is the main cause of the common cold. In [5], using multiple sequence alignment, Palmenberg et al. [5] correctly classified the complete HRV genomes into three genetically distinct groups within the genus *Enterovirus* (HEV) and the family *Picornaviridae*. The dataset used in [5] consists of three groups HRV-A, HRV-B, HRV-C including 113 genomes, and three outgroup sequences HEV-C. The time complexity was very high because of the use of multiple sequence alignment. In this paper, we use the same dataset to test our method.

From the phylogenetic tree of HPV genomes based on our method in Figure 2, we find that except some genomes of type HRV-C inaccurately grouped together with type HRV-A, and some genomes of type HRV-A being away from the main branch of type HRV-A, all other genomes are correctly classified into their corresponding types.

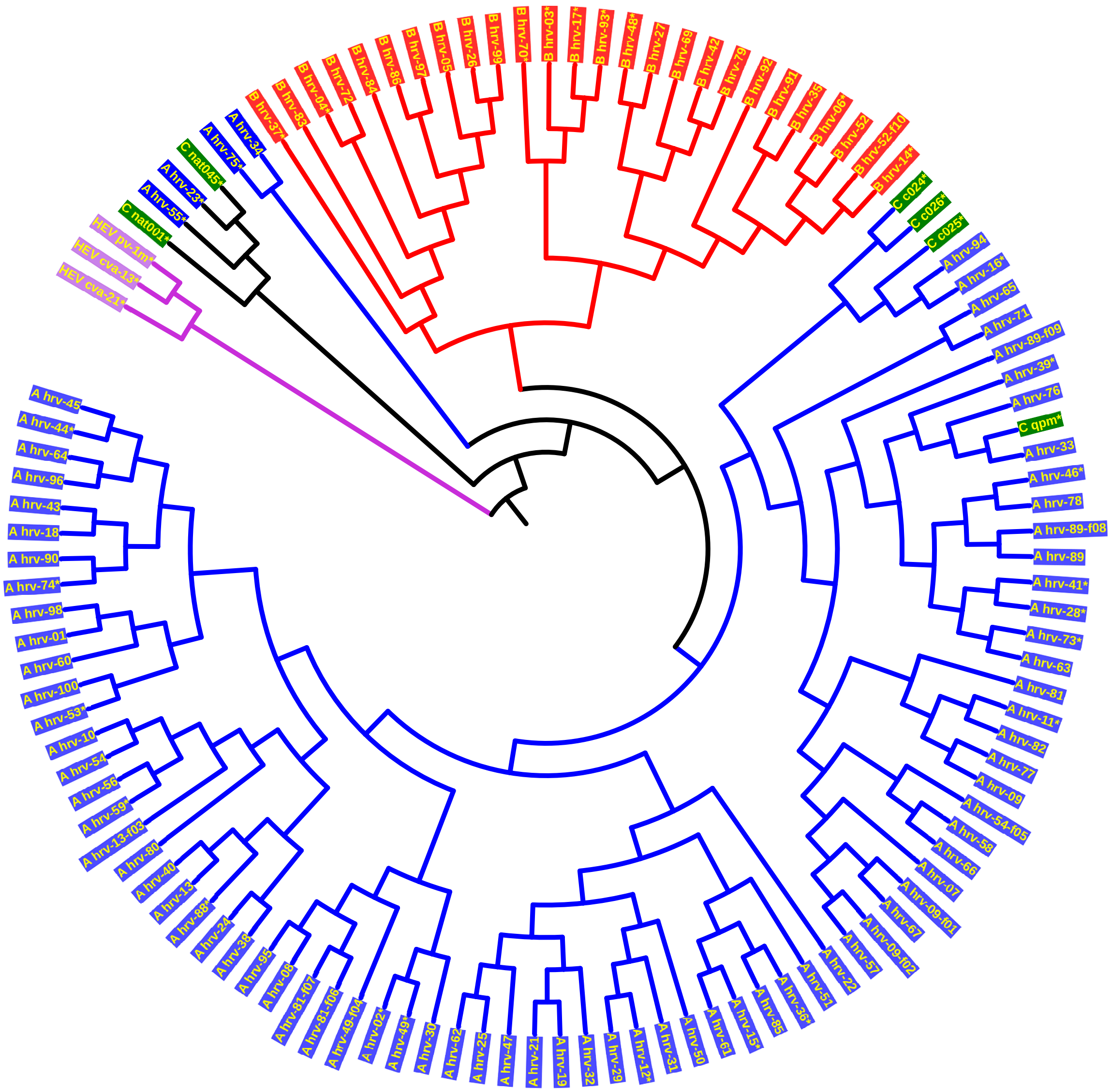

Figure 2: Phylogenetic tree of 116 HRV genomes of 4 genotypes

3.2 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 [6]), and are known to have high degree of genetic and antigenic variability (see [7, 8]). Some subtypes of Influenza A viruses are very dangerous, and lethal including H1N1, H2N2, H5N1, H7N3, and H7N9. We apply our method on the dataset consisting of 38 Influenza A virus genomes whose accession numbers in GenBank can be found in the Appendix A (Supplementary data) of [9]. From Figure 4, we find that except A/American black duck/NB/2538/2007-H7N3, A/chicken/British Columbia/GSC human B/04-H7N3, A/turkey/Minnesota/1/1988-H7N9 inaccurately misplaced in H2N2 group, all other Influenza A genomes are clustered correctly into their types.

In addition, our proposed method allows one to visually inspect differences between Influenza A virus genomes in the *simplest possible way* of persistent homology. For example, Figure 3 illustrates identical 0-th persistence diagrams of *A/mallard/Maryland/352/2002-H1N1* and *A/mallard/Maryland/26/2003-H1N1* whose highly identical visualization shows that they should belong in the same branch as indicated by Figure 4. Furthermore the 0-th persistence diagram *A/mallard/Maryland/352/2002-H1N1* shows that the geometric shape of its DNA sequence has exactly two connected components.

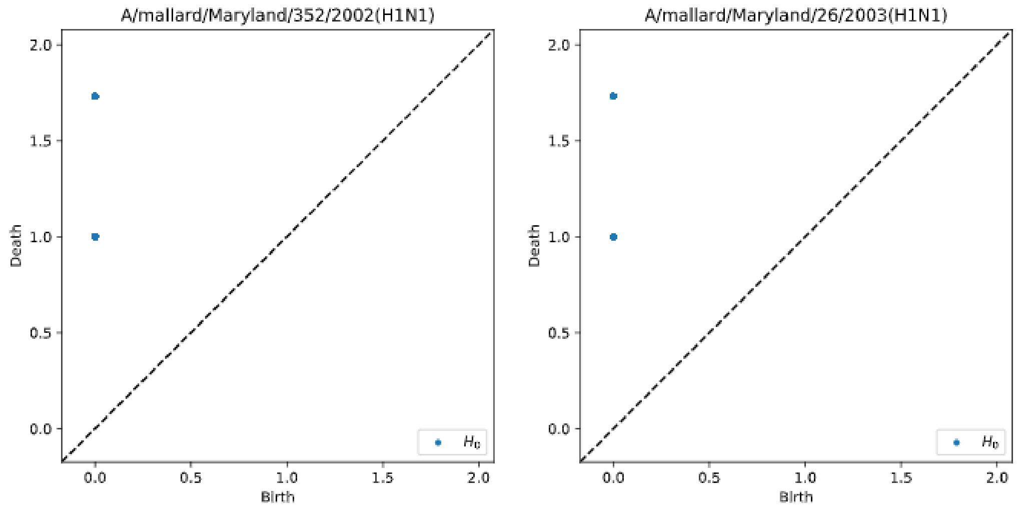

Figure 3: 0-th persistence diagrams of Influenza A virus genomes

References

[1] D. Q. N. Nguyen, P. D. T. Le, L. Xing, and L. Lin, “A topological characterization of dna sequences based on chaos geometry and persistent homology,” *Preprint, available at* https://www.biorxiv.org/content/10.1101/2021.01.31.429071v1.full, 2021.

[2] B. Liao, R. Li, and W. Zhu, “On the similarity of dna primary sequences based on 5-d representation,” *Journal of Mathematical Chemistry*, vol. 42, no. 1, pp. 47–57, 2007.

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

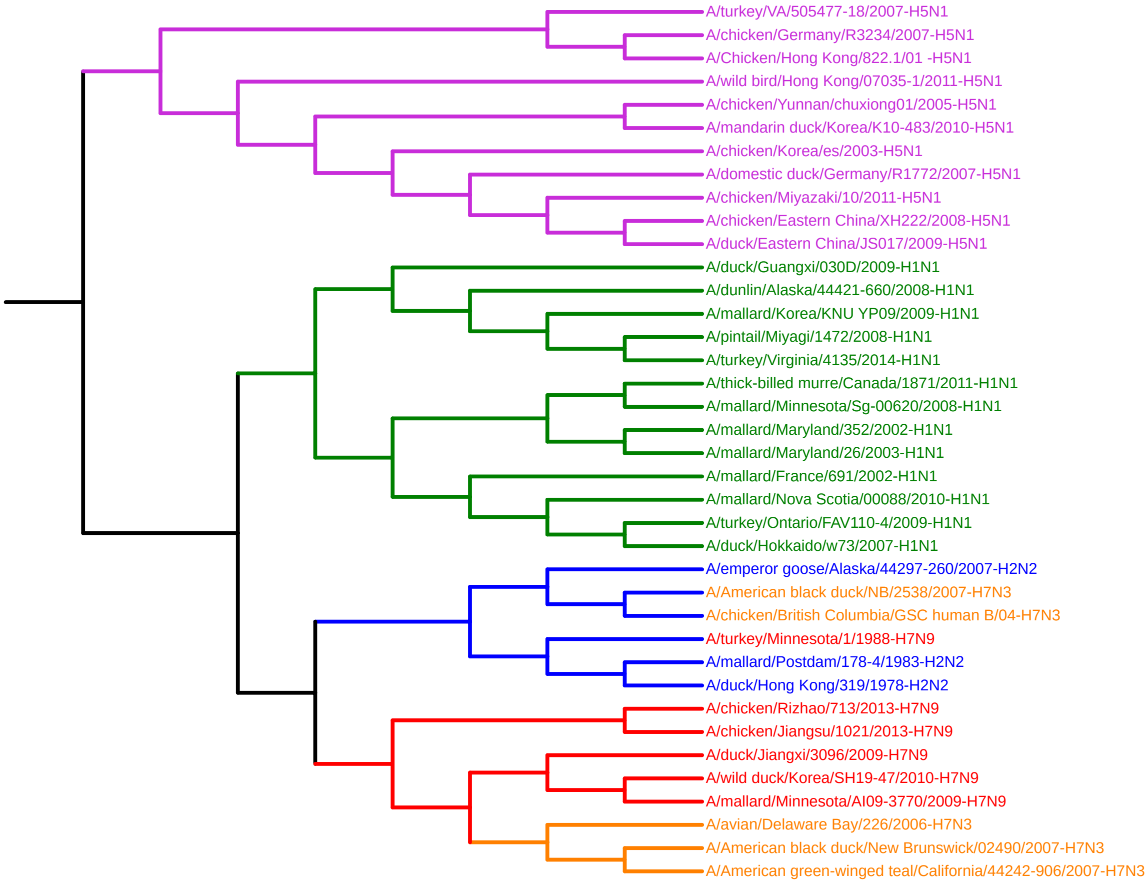
A/turkey/VA/505477-18/2007-H5N1
A/chicken/Germany/R3234/2007-H5N1
A/Chicken/Hong Kong/822.1/01 -H5N1
A/wild bird/Hong Kong/07035-1/2011-H5N1
A/chicken/Yunnan/chuxiong01/2005-H5N1
A/mandarin duck/Korea/K10-483/2010-H5N1
A/chicken/Korea/es/2003-H5N1
A/domestic duck/Germany/R1772/2007-H5N1
A/chicken/Miyazaki/10/2011-H5N1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

Figure 4: Phylogenetic tree of 38 Influenza A virus genomes

[4] 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.

[5] A. Palmenberg, D. Spiro, R. Kuzmickas, S. Wang, A. Djikeng, J. Rathe, C. Fraser-Liggett, and S. Liggett., “Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution,” *Science*, vol. 324, pp. 55–59, 2009.

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

[7] 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.

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

[9] 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.