Molecular Phylogenetics and Evolution 81 (2014) 29–36

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Global comparison of multiple-segmented viruses in 12-dimensional genome space

Hsin-Hsiung Huang [a], Chenglong Yu [b], Hui Zheng [c], Troy Hernandez [d], Shek-Chung Yau [e], Rong Lucy He [f], Jie Yang [c], Stephen S.-T. Yau [g,*]

[a] Department of Statistics, University of Central Florida, Orlando, FL 32816, USA
[b] Mind-Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, South Australia 5000, Australia
[c] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
[d] Civis Analytics, Chicago, IL, USA
[e] Information Technology Services Center, Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong
[f] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA
[g] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China

ARTICLE INFO

Article history:
Received 31 October 2013
Revised 11 July 2014
Accepted 3 August 2014
Available online 27 August 2014

Keywords:
Natural vector
Natural graphical representation
Phylogeny
Nucleotide sequence

ABSTRACT

We have recently developed a computational approach in a vector space for genome-based virus classification. This approach, called the "Natural Vector (NV) representation", which is an alignment-free method, allows us to classify single-segmented viruses with high speed and accuracy. For multiple-segmented viruses, typically phylogenetic trees of each segment are reconstructed for discovering viral phylogeny. Consensus tree methods may be used to combine the phylogenetic trees based on different segments. However, consensus tree methods were not developed for instances where the viruses have different numbers of segments or where their segments do not match well. We propose a novel approach for comparing multiple-segmented viruses globally, even in cases where viruses contain different numbers of segments. Using our method, each virus is represented by a set of vectors in R^{12}. The Hausdorff distance is then used to compare different sets of vectors. Phylogenetic trees can be reconstructed based on this distance. The proposed method is used for predicting classification labels of viruses with n-segments ($n \geqslant 1$). The correctness rates of our predictions based on cross-validation are as high as 96.5%, 95.4%, 99.7%, and 95.6% for Baltimore class, family, subfamily, and genus, respectively, which are comparable to the rates for single-segmented viruses only. Our method is not affected by the number or order of segments. We also demonstrate that the natural graphical representation based on the Hausdorff distance is more reasonable than the consensus tree for a recent public health threat, the influenza A (H7N9) viruses.

1. Introduction

In comparative genomics, at the sequence level, hundreds of thousands of genome sequences are produced and used for determining relatedness and ancestors. Searching for viral origins has also been an important issue in virology (Holmes, 2009). Viral sequence similarity plays a crucial role in revealing virus mutation history (Koonin et al., 2008). Multiple sequence alignment methods are popular, but computationally intensive. Additionally they may fail for diverse systems of different families of RNA viruses (Holmes, 2011). In the past one decade or so, alignment-free methods have attracted a lot of attention from researchers (Kantorovitz et al., 2007; Sims et al., 2009; Vinga and Almeida, 2003). One widely-used alignment-free method makes use of the frequencies of k-mers (Dai et al., 2008). Recently, the NV method was proposed as a fast and efficient way to characterize nucleotide sequences (Deng et al., 2011; Yu et al., 2010). Unlike the k-mer methods which ignore the position information of nucleotides, the NV representation provides both mean and variance of the positional information and establishes a genome space in a 12-dimensional Euclidean space. For many datasets from the real world including our working viral genome dataset, the map between the genome sequences and their 12-dimenasional NVs is one-to-one. It should be noted that the map between a set of arbitrary sequences in A, C, G, and T and their 12-dimensional NVs may not be one-to-one (see Deng et al. (2011) and Yu et al. (2013) for one-to-one correspondence theoretically constructed on higher-dimensional NVs). The

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).

http://dx.doi.org/10.1016/j.ympev.2014.08.003
1055-7903/

NV method is much faster and can derive more reasonable results than multiple sequence alignment methods for comparative genomic analysis or reconstructing phylogenetic trees (Yu et al., 2010). However, the NV method based on the Euclidean distance (Yu et al., 2013) can only deal with single-segmented viruses. To simultaneously compare viruses with multiple segments, we propose the use of the Hausdorff distance which measures the distance between two sets of vectors. It allows us to make a simultaneous comparison against all available multiple-segmented viruses at each taxonomic level (i.e., Baltimore class, family, subfamily, genus, and species) in a fast and efficient manner. The NV approach, which does not depend on model assumptions such as assumed mutation rates, uses the global sequence information of genomes. Furthermore, we apply the natural graphical representation (Yu et al., 2013) to display the viral phylogenetic relationships in the 12-dimensional genome space.

Using our NV method, all 2384 multiple-segmented referenced viral genomes in GenBank can be embedded in R^{12}, and we can use natural vectors to classify viral nucleotide sequences. Unlike other approaches, it allows us to determine phylogenetic relations for all viruses at any taxonomic level in real time. This approach is successfully used to predict and correct viral classifications, as well as to identify viral origins; e.g. a recent public health threat, the influenza A (H7N9) virus (Zhou et al., 2013).

2. Materials and methods

2.1. Summary of the procedure

We build a genome space in a 12-dimensional Euclidean space through the NV mapping which uses the quantity and global distribution of nucleotides in sequences. Each segment is uniquely represented by a single point in this space. The Hausdorff distance between two sets of segments represents the biological distance of the corresponding two viruses. Using the natural graphical representation, we undertake phylogenetic and cluster analysis for all the available reference viral genome sequences. The procedure is described below.

First, we use the NV algorithm to compute the natural vector of each segment of the viruses, and then calculate the Hausdorff distance between each pair of the viruses. Secondly, we perform the natural graphical representation to display the phylogenetic relationships. We illustrate using the H7N9 virus example (see Section Results) that our natural graphical representation based on the Hausdorff distance is more reasonable than the consensus trees (Barrett et al., 1991).

2.2. Natural vector and genome space

Using the NV, we can construct a viral genome space in a 12-dimensional Euclidean space, which only depends on the numbers and global distributions of nucleotides in the viral genome sequences. There are three reasons for a virus to be represented as a 12-dimensional vector in the viral genome space for predicting Baltimore and ICTV classification labels (Yu et al., 2013). First, the mapping between the 12-dimensional NVs and all the viruses examined is one-to-one. Secondly, the mapping from the original genome space into our 12-dimensional NV space keeps the phylogenetic relationships, that is, two viruses tend to remain in the same class if their corresponding NVs are close to each other. Thirdly, the classification results do not change if we keep more than the first 12 components of NVs. Our new approach of classifying viral genomes is not a partial-sequence-based method. It is constructed using the global sequence information of genomes.

Let $S = (s_1, \ldots, s_n)$ be a nucleotide sequence of length n, that is, $S_i \in \{A, C, G, T\}$, $i = 1, \ldots, n$. For $k = A, C, G, T$, define $w_k(\cdot):\{A, C, G, T\} \to \{0, 1\}$ such that $w_k(s_i) = 1$ if $s_i = k$ and 0 otherwise.

(1) Let $n_k = \sum_{i=1}^{n} w_k(s_i)$ denote the number of letter k in S.
(2) Let $\mu_k = \sum_{i=1}^{n} i \cdot \frac{w_k(s_i)}{n_k 2}$ be the mean position of letter k.
(3) Let $D_2^k = \sum_{i=1}^{n} \frac{(i-\mu_k)^2 w_k(s_i)}{n_k n}$ be a scaled variance of positions of letter k.

The 12-dimensional NV of a DNA/RNA sequence S is defined by $(n_A, \mu_A, D_2^A, n_C, \mu_C, D_2^C, n_G, \mu_G, D_2^G, n_T, \mu_T, D_2^T)$.

For NV defined for nucleotide sequences with ambiguous letters, see Yu et al. (2013).

We use the 12-dimensional NV to construct a viral genome space, which is a moduli space of viral genomes. Every virus segment corresponds to a point in this space. Using the Hausdorff distance (see next section) between two sets of points as a metric, we can perform phylogenetic and clustering analysis for the viral genome sequences consisting of any number of segments. We also provide the website: http://mathlab.math.uic.edu/dev for users who are interested in trying out our method for virus classification.

2.3. Hausdorff distance

Let X, Y denote two finite sets of 12-dimensional NVs. The Hausdorff distance is defined by $h(X, Y) = max\{max_{\{x\in X\}} min_{\{y\in Y\}} d(x, y), max_{\{y\in Y\}} min_{\{x\in X\}} d(x, y)\}$ where $d(x, y)$ is the Euclidean distance of two $(n_A, \mu_A, D_2^A, n_C, \mu_C, D_2^C, n_G, \mu_G, D_2^G, n_T, \mu_T, D_2^T)$ natural vectors x and y in X, Y respectively, *max* denotes the maximum, and *min* is the minimum (Morgan, 1987). Unlike many distances used in comparative genomics (e.g., the distances based on multiple alignment methods), the Hausdorff distance satisfies the triangle inequality $h(A, B) \leqslant h(A, C) + h(C, B)$.

For readers' reference, we provide a proof for the triangle inequality in the Appendix. In the following example, we illustrate the advantage of the Hausdorff distance using for comparing multiple-segmented viral genomes. Suppose virus X consists of four segments x_1, x_2, x_3, x_4, while virus Y consists of four segments y_1, y_2, y_3, y_4. Suppose the distance matrix $(d_{ij}) = (d(x_i, y_j))$ is given by

$$\begin{bmatrix} 16 & 7 & 2 & 23 \\ 1 & 10 & 15 & 8 \\ 25 & 18 & 7 & 2 \\ 19 & 3 & 37 & 22 \end{bmatrix}.$$

We find the smallest distance in each row (2, 1, 2, and 3, respectively) and in each column (1, 3, 2, and 2, respectively). The Hausdorff distance between X and Y is then $max(max\{2, 1, 2, 3\}, max\{1, 3, 2, 2\}) = 3$. The first advantage of the Hausdorff distance is that it remains the same even if we rearrange the order of the segments of the viruses. That is, it is unnecessary to arrange the segments of two viruses in a consistent order before measuring their distance. On the other hand, if $h(X, Y) = 0$, then the two viruses will be identical with matched segments. Another advantage is that the Hausdorff distance can compare two viruses with different numbers of segments. For example, if for some reason we miss segment x_4, we can still get $h(\{x_1, x_2, x_3,\}, Y) = 7 > 3 = h(X, Y)$. That is, in this case the Hausdorff distance may impose a penalty due to the missing segment.

It has been shown that the natural vector representation can be used to predict single-segmented viral classification labels quickly and accurately (Yu et al., 2013). Up to April 2013, NCBI kept reference sequences of 2384 viruses in its GenBank collection. Among the virus collection, 370 are multiple-segmented. We use the Hausdorff distance to compare all the 2384 viruses simultaneously, not just the single-segmented viruses. Using the Hausdorff

Table 1
Counts of viruses according to their numbers of segments (column labels) and the numbers of segments of their 1st nearest neighbor (row labels); the last row lists the total number of viruses with the corresponding number of segments (e.g. entry 5 at row 6 and column 1 indicates there are 5 viruses that have 1 segment but their 1st neighbor have 6 segments).

#Seg	1	2	3	4	5	6	7	8	9	10	11	12	16	20	24	30	56	105
1	2007	14	2												1			
2	2	209	3	1														
3		2	68			1					1						1	
4		1	1	6	1													
5				1	0													
6	5		1	2		2		2			1							
7							0											
8						1	1	6										
9									0									
10				1					1	16	1			1				1
11										1	6	1						
12											2	8						
16													0			1		
20										1				0				
24															0			
30													1			0		
56																	0	
105																		0
Total	2014	226	75	11	1	4	1	8	1	18	10	9	1	1	1	1	1	1

Table 2
Prediction inconsistency rates of 2384 viruses (including both single-segmented and multiple-segmented viruses).

Inconsistency rate	Baltimore (%)	Family (%)	Subfamily (%)	Genus (%)
With cutoff	3.5	4.6	0.3	4.4
Without cutoff	7.7	11.1	4.3	10.4

Table 3
Prediction inconsistency rates of the 2014 single-segmented viruses.

Inconsistency rate	Baltimore (%)	Family (%)	Subfamily (%)	Genus (%)
With cutoff	3.6	5.2	0.2	4.6
Without cutoff	7.5	11.1	4.0	10.9

Table 4
Names of the 28 H7N9 viruses from NCBI influenza database.

#	Short name	Strains name of influenza A (H7N9) virus
1	Czech 1	A/goose/Czech Republic/1848-K9/2009
2	Czech 2	A/goose/Czech Republic/1848-T14/2009
3	Spain 1	A/Anas crecca/Spain/1460/2008
4	China Fujian	A/Fujian/1/2013
5	China Hangzhou	A/Hangzhou/1/2013
6	China Nanjing 1	A/Nanjing/1/2013
7	China Shanghai 1	A/Shanghai/02/2013
8	China Shanghai 2	A/Shanghai/4664T/2013
9	Taiwan 1	A/Taiwan/S02076/2013
10	Taiwan 2	A/Taiwan/T02081/2013
11	China Zhejiang 1	A/Zhejiang/DTID-ZJU01/2013
12	China Zhejiang 2	A/Zhejiang/HZ1/2013
13	Guatemala 1	A/blue-winged teal/Guatemala/CIP049-01/2008
14	Guatemala 2	A/blue-winged teal/Guatemala/CIP049-02/2008
15	USA Ohio	A/blue-winged teal/Ohio/566/2006
16	China Zhejiang 3	A/duck/Zhejiang/SC410/2013
17	USA Alaska	A/emperor goose/Alaska/44063-061/2006
18	China Nanjing 2	A/environment/Nanjing/2913/2013
19	USA Mississippi	A/northern shoverl/Mississippi/11OS145/2011
20	USA Delaware	A/ruddy turnstone/DE/1638/2000
21	USA Delaware Bay	A/ruddy turnstone/Delaware Bay/220/1995
22	South Korea 1	A/spot-billed duck/Korea/447/2011
23	USA Minnesota 1	A/turkey/Minnesota/1/1988
24	USA Minnesota 2	A/turkey/Minnesota/38429/1988
25	South Korea 2	A/wild bird/Korea/A14/2011
26	South Korea 3	A/wild bird/Korea/A3/2011
27	South Korea 4	A/wild bird/Korea/A9/2011
28	Mongolia	A/wild duck/Mongolia/1-241/2008

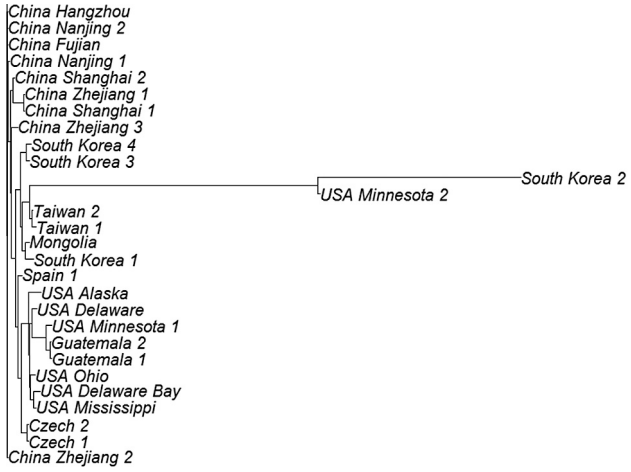

Fig. 1. Phylogenetic tree of the 28 H7N9 viruses using the NV's Euclidean distances of the HA segments.

distance, 97.7% of the 2384 viruses share the same number of segments with their first neighbors (see the diagonal entries in Table 1). That is, a multiple-segmented virus tends to have a virus with the same number of segments as its first neighbor under the Hausdorff distance. On the other hand, there are a few cases that a virus and its first neighbor have different numbers of segments (see the off-diagonal entries in Table 1). For example, based on the Hausdorff distance, the first neighbor of *Subterranean clover stunt* virus (SCSV), which consists of eight segments, is *Abaca bunchy top* virus (ABTV) which consists of six segments. Both viruses belong to the *Nanoviridae* family (Grigoras et al., 2007).

Due to the sparsity of currently available viral reference sequences, we may not be able to find any known virus in a reasonable distance from a given virus. Following (Yu et al., 2013), we choose the 75% quantile of the nearest distances as a cut-off by cross validation and predict the class label of the given virus only if we find a known virus whose distance from the given virus is smaller than the cut-off distance. The error rates of the reported predictions out of 2384 viruses at different class levels are 3.5% for Baltimore labels, 4.6% for family labels, 0.3% for subfamily labels, and 4.4% for genus labels (see Table 2), which are compara-

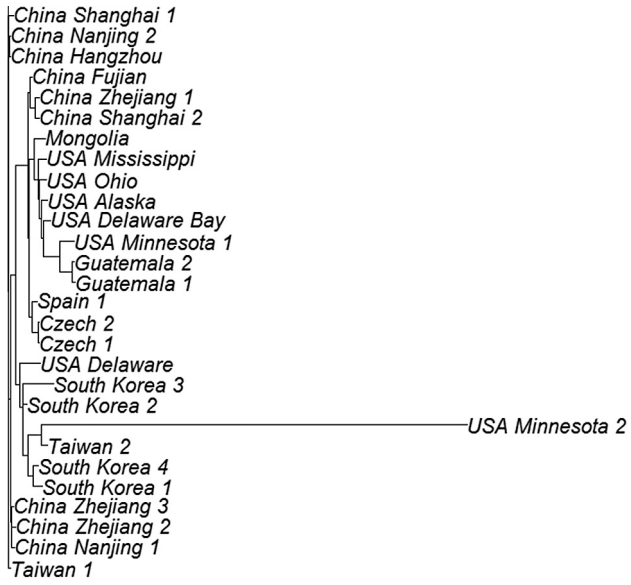

Fig. 2. Phylogenetic tree based on the Euclidean distances of the NA segments.

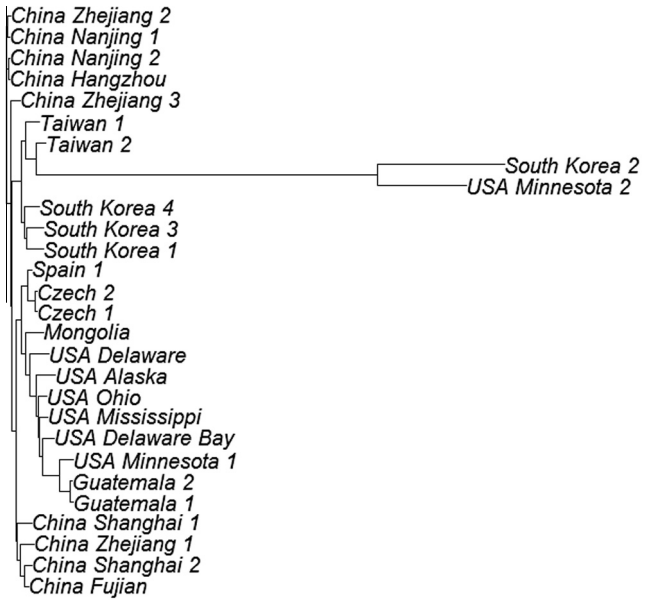

Fig. 4. Phylogenetic tree based on the Hausdorff distances of the NA and HA segments.

ble with the error rates for the 2014 single-segmented viruses using the Euclidean distance (Yu et al., 2013) (see Table 3 for the corresponding error rates based on the latest data).

We also apply our method to analyze the influenza A (H7N9) virus, a new public health threat, which consists of eight gene segments. The new H7N9 viruses are famous for being highly contagious, deadly, and rapid evolving (Liu et al., 2013). It is important to identify the newly evolved viruses quickly and track to place them within the phylogenetic tree of the H7N9 viruses.

The H7N9 virus, a subtype of *Orthomyxoviridae* virus, has typically been found in birds, but can now be found in humans according to the Centers for Disease Control and Prevention. Traditionally, the phylogenetic trees of these new strains are reconstructed segment by segment, typically based on the *hemagglutinin* (HA) and *neuraminidase* (*NA*) *gene segments only*. *We* apply our method to analyze H7N9 viruses based on these two gene segments. We download viral sequences of 28 strains of H7N9 from the NCBI Influenza virus database. Having computed the NVs of the segments of each virus and the Hausdorff distances among those viruses, we use the neighbor joining method to reconstruct a phylogenetic tree based on the HA and NA segments, as well as the natural graphical representation to find potential clusters of the new strains of H7N9. Our clustering results are better than those from the consensus tree, which combined the trees of the HA and NA segments by the majority rule (see Section Results). The complete R codes for computing both NVs and the Hausdorff distances with an application to the H7N9 data as an example can be found at http://www.math.uic.edu/~hhuang45/Natural_Vector/.

3. Results

To estimate the correctness rate of our method, we apply leave-one-out cross-validation on the 2384 referenced viruses from GenBank of NCBI updated on April 2013. That is, for each of the 2384 viruses, we predict its taxonomic labels based on the labels of all

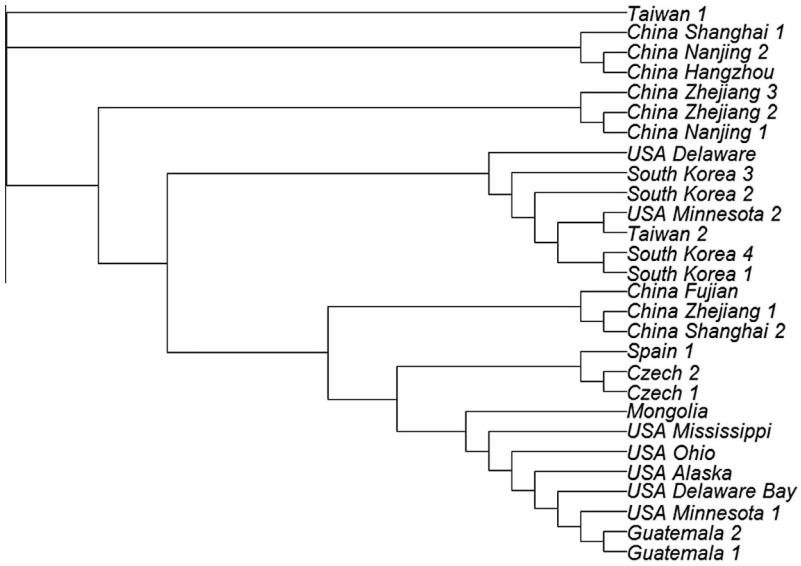

Fig. 3. Consensus tree of the phylogenetic trees of the NA and HA segments.

the other viruses except this one. Among these viruses, there are 2014 single-segmented viruses. We also apply the method used in Yu et al. (2013) as a comparison to predict the taxonomic labels of the single-segmented viruses in the same dataset. Using the 75% quantile as the cut-off point (Yu et al., 2013), the inconsistency rates of the multiple-segmented viruses are 3.5% for Baltimore class labels, 4.6% for family labels, 0.3% for subfamily labels and 4.4% for genus labels (see Table 2), whereas the inconsistency rates of the single-segmented viruses are 3.6% for Baltimore class labels, 5.2% for family labels, 0.2% for subfamily labels and 4.6% for genus labels (see Table 3). Therefore the correctness rates (1 – inconsistency rate) of the multiple-segmented virus classification are comparable or even better than those of the single-segmented viruses.

Epizootic hemorrhagic disease virus (EHDV), *Planaria asexual strain specific* virus-like element (PASSV), and *Rosellinia necatrix megabirna* virus (RNMV) are the three viruses having more than one segment without Baltimore classification labels. We use the first neighbor of the unlabeled viruses with respect to the Hausdorff distance of NVs for prediction. Since *Bluetongue* virus (BV) is a dsRNA virus, we predict that EHDV's Baltimore label is III. We can further compare biological traits of these two viruses. First, EHDV and BV are both RNA viruses with 10 segments of genome sequences. Secondly, their genomes sequences are both of linear shape. Thirdly, their ICTV classification labels are all the same. Furthermore, EHDV and BV are closest to each other. They share with three common nearest neighbors. Therefore, the prediction of Baltimore label of EHDV is reasonable and reliable.

The influenza A viral genomes consist of eight RNA segments. They are labeled as PB1, PB2, PA, HA, NP, NA, M1/M2, and NS1/NS2. Among these eight segments, the HA and NA segments are external glycoprotein antigens and the other six are internal. Influenza A viruses involve 17 HA types and 10 NA types. Moreover, according to virologists, the influenza A viruses are identified by their NA and HA segments.

Although the consensus tree method has not been used in classifying influenza viruses, consensus trees are useful for concatenating phylogenetic trees of multiple-segmented viruses which we do not know its reassortment clearly (Medina et al., 2009; Gao and Luo, 2011). For comparison purposes, we use H7N9 (see Table 4) as an example to illustrate that consensus trees may lead to a confusing result and our multiple-segmented virus classification method can get a reasonable phylogenetic relationship quickly.

We compute the NV of each segment of the H7N9 viruses and reconstruct the phylogenetic trees based on the HA segment only (Fig. 1), the NA segment only (Fig. 2), the consensus tree of the HA and NA segments (Fig. 3), the Hausdorff distance of the HA and NA segments' NVs with neighbor-joining method (Saitou and Nei, 1987) (Fig. 4), the Hausdorff distance of all the eight segments' NVs (Fig. 5), and the eight segments' consensus tree (Fig. 6) using an R package (Paradis et al., 2004), as well as the natural graphical representations based on the HA and NA segments only (Fig. 7) in addition to using all the eight segments (Fig. 8). The discussion below shows that the natural graphical representation based on the Hausdorff distances of all the eight segments (Fig. 8) is most reasonable.

4. Discussion

According to the recent studies of the new H7N9 virus strains isolated from humans, these virus strains are reassortants where the six internal genes were derived from avian H9N2 viruses (Zhou et al., 2013; Liu et al., 2013). However, the ancestors of their HA and NA segments have not been verified. Hence, the HA and NA segments of the new H7N9 may be very different from those of the old H7N9 viruses. Before the new H7N9 viruses reported in 2013, the H7N9 viruses were only found in birds. The new H7N9 virus was first discovered in China, and it was also found in Taiwan from an infected traveler who came back to the island from China. The two Taiwan strains, A/Taiwan/s02076/2013 (Taiwan 1) and A/Taiwan/T02081/2013 (Taiwan 2), were found in that traveler and mutated from the same origin, so that they should be closely related to each other. However, according to the phylogenetic tree (Fig. 2) of the NA segments, the strains of Taiwan 1 and Taiwan 2 are far away from each other, and the Taiwan 2 strain is close to the old strains found in birds from South Korea and USA. The results based on the NA segments are not consistent with the phylogenetic tree (Fig. 1) based on the HA segments in which the two Taiwan strains are the closest neighbors to each other. One problem with Fig. 1 is that it groups the two Taiwan strains with the ones from South Korea and Mongolia which is not reasonable since the host was traveling in a region close to Hangzhou in China.

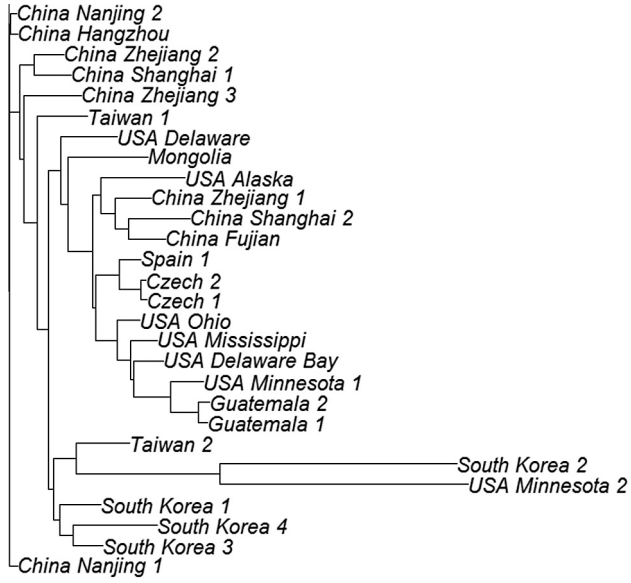

Fig. 5. Phylogenetic tree based on the Hausdorff distances of all the eight segments.

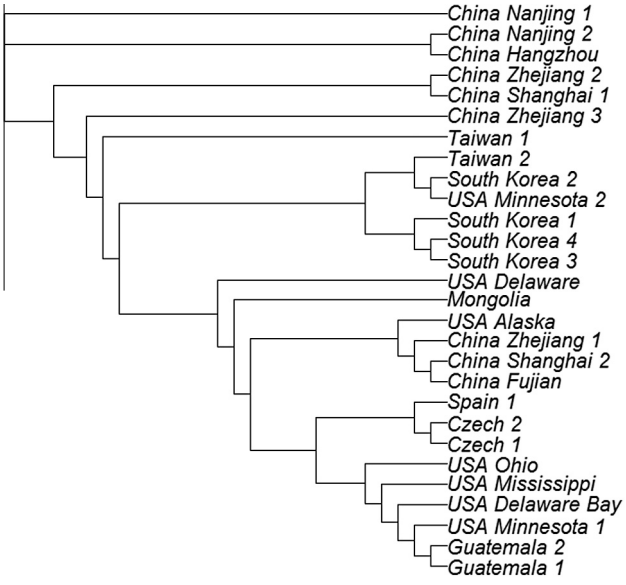

Fig. 6. Consensus tree of the phylogenetic trees of all the eight segments.

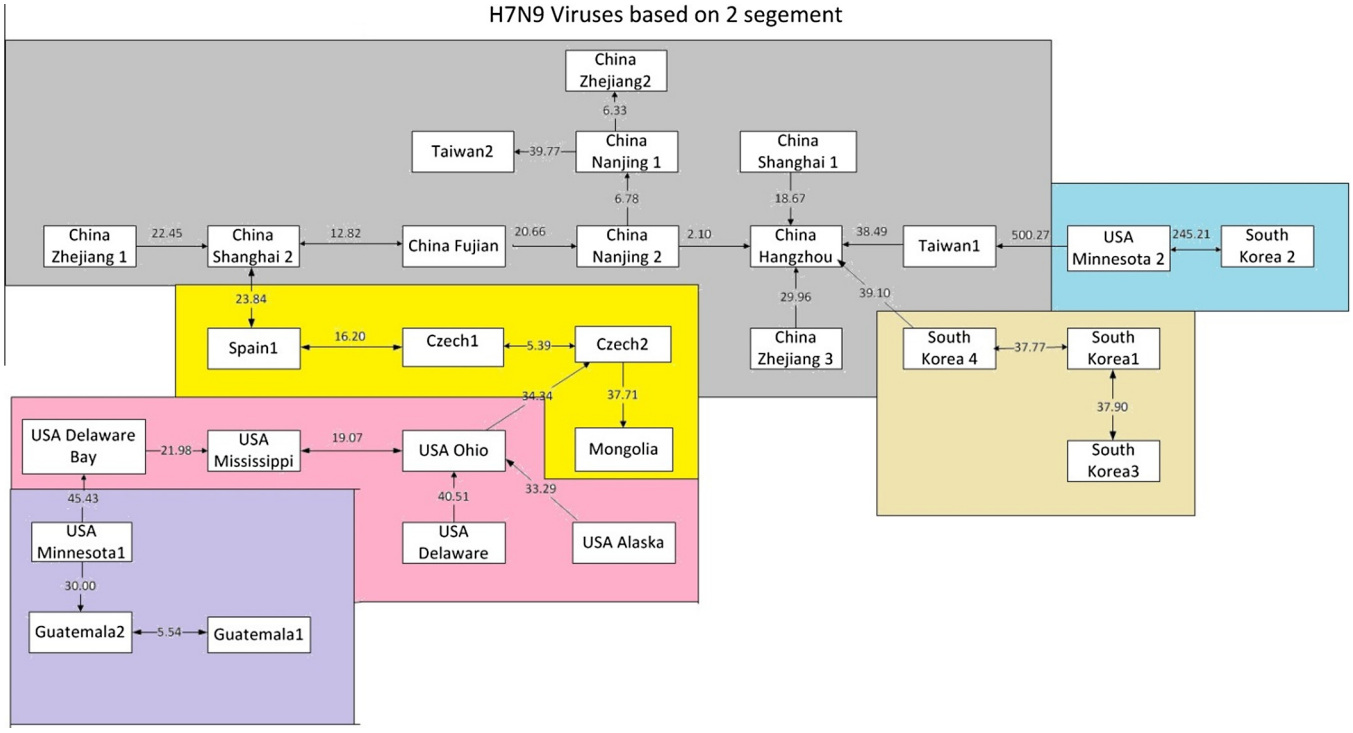

Fig. 7. Natural graphical representation of 28 H7N9 viruses based on the HA and NA segments' Hausdorff distances (boxes indicate viruses, lines between boxes indicate one virus is the 1st nearest neighbor of another, real numbers aside lines indicate the Hausdorff distances).

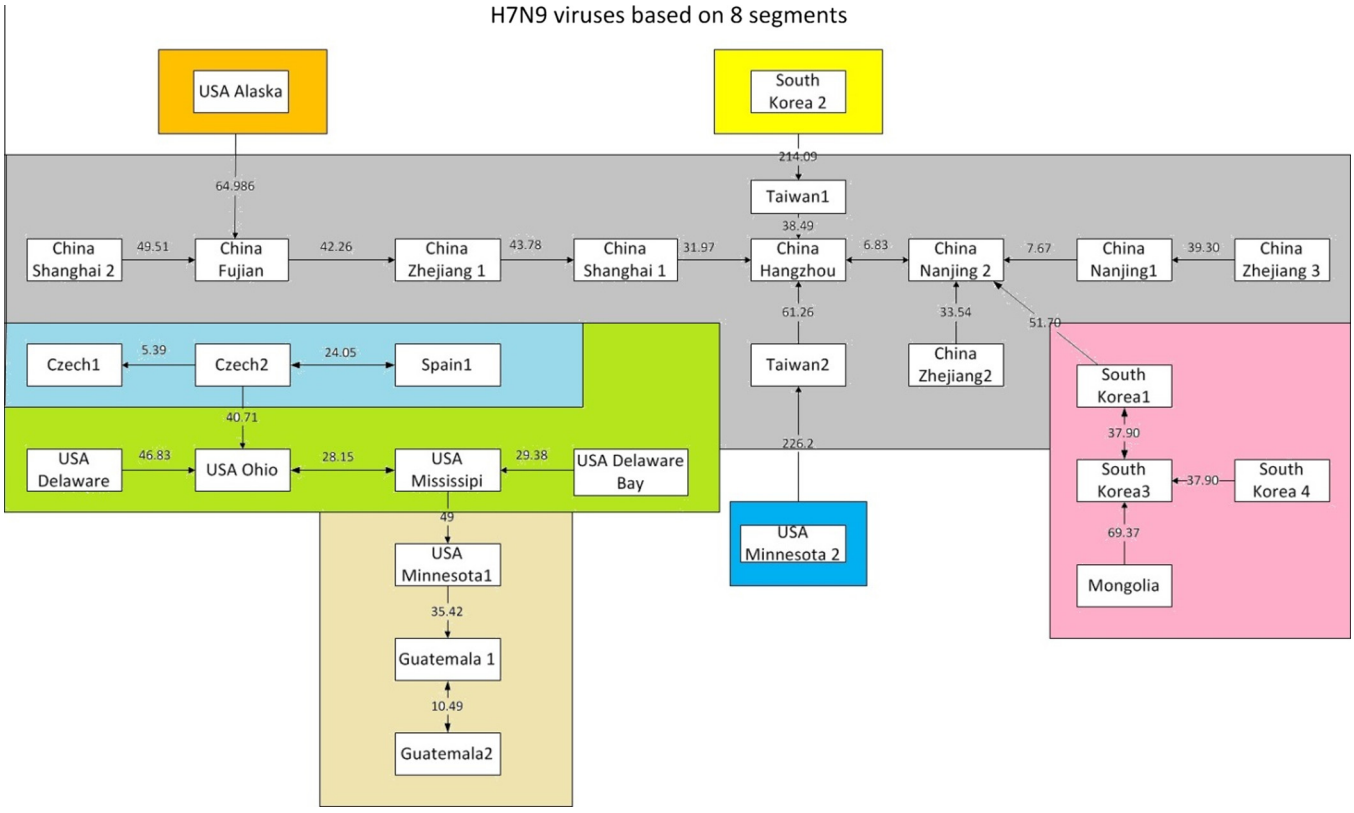

Fig. 8. Natural graphical representation of 28 H7N9 viruses based on the eight segments' Hausdorff distances.

When summarizing the information of the HA and NA segments, the majority-rule consensus tree (Fig. 3) of the HA and NA segments assigns these two Taiwan strains to different leaves, and the Taiwan 2 strain is even clustered with A/turkey/Minnesota/38429/1988 which is not reasonable. The argument above reflects that neither the phylogenetic tree based on a single segment nor the consensus tree based on two segments provides reliable biological classifications.

On the other hand, if restricted to the HA and NA segments, the phylogenetic visualization based on the Hausdorff distances (Figs. 4 and 7) still contains problems. Fig. 4 is a reconstructed phylogenetic tree using the neighbor joining method. It again clusters these two Taiwan strains along with those from South Korea instead of those from China. Fig. 5 is the phylogenetic tree based on the Hausdorff distances of all the eight segments. There are several unreasonable clusters such as two Taiwan strains on different branches and some new human strains being grouped with the old avian strains. In addition, Fig. 6 is the consensus tree of the phylogenetic trees of all the eight segments, and it has similar problems. For example, the two Taiwan strains are separated again. On the other hand, the natural graphical representation (Fig. 7) using the same Hausdorff distances clearly shows the two Taiwan strains are closer to A/Hangzhou/1/2013 and A/Nanjing/1/2013, respectively. Nevertheless, Fig. 7 shows the Mongolia one is closer to European strains (A/Anas crecca/Spain/1460/2008, A/goose/Czech Republic/1848-K9/2009, and A/goose/Czech Republic/1848-T14/2009) which is not reasonable.

Fig. 8 is the natural graphical representation based on the Hausdorff distances of all the eight segments, which is the most reasonable one. It illustrates that A/Hangzhou/1/2013 locates at the center surrounded by A/Shanghai/02/2013, Taiwan 1, Taiwan 2, and A/environment/Nanjing/2913/2013. It is reasonable because the cities Hangzhou, Suzhou, Shanghai, and Nanjing are close to each other (within 200 miles) and the two Taiwan strains were from the same traveler who visited Suzhou and Shanghai. It also implies that A/Hangzhou/1/2013 may be the origin of the other four strains. The phylogenetic analysis is most reasonable when we use their Hausdorff distances of all the eight segments.

5. Conclusion

In phylogenetic studies, finding origins and relationships are extremely important. The alignment-free NV method has been proven to be highly accurate in classifying single-segmented viruses. However, as seen from our analysis, the phylogenetic trees of different segments of multiple-segmented viruses can be inconsistent and misleading. In this paper, we propose an effective method to extend the NV method from handling single-segmented viruses to multiple-segmented viruses. It provides a more reliable cluster analysis result than the consensus trees used for multiple-segmented viruses. In practice, it is recommended to use the NV and the natural graphic representation, which is related to Borůvka's minimal spanning tree algorithm (Borůvka, 1926), along with the Hausdorff distance when comparing viruses involving multiple segments. However, the natural graphic representation may not result in a tree (Yu et al., 2013). For example, if a virus has two nearest neighbors that share the equal distance, the natural graphic representation will keep both neighbors for biological applications, while a minimal spanning tree would choose only one of the neighbors.

It should be noted that technically the Hausdorff distance can be combined with any alignment-free method. Some alignment-free methods including Comin and Verzotto (2012), Gao and Luo (2011), and Sims et al. (2009) are recently proposed. For comparison purpose, we applied the method of Sims et al. (2009) using their default setting along with the Hausdorff distance onto our NCBI reference dataset. The corresponding error rates are 11.9%, 14.8%, and 12.8% for Baltimore, Family, and Genus, respectively. Comparing to our error rates which are 7.7%, 11.1%, and 10.4%, the NV method performs better. We also tried the method of Comin and Verzotto (2012), which is computationally intensive. The time to analyze our NCBI dataset would last more than 30 days. Moreover, the phylogenetic tree of H7N9 virus data based on their method is not reasonable (the result is not shown here).

Acknowledgments

Funding: This research is supported by the U.S. National Science Foundation Grants DMS-1120824, DMS-1119612, China NSF Grant 31271408, Tsinghua University startup funding, and Tsinghua University independent research project grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A

Proof of the triangle inequality of the Hausdorff distance $h(A, B) \leqslant h(A, C) + h(C, B)$:

Define $d(A, B) = max_{\{a \in A\}} min_{\{b \in B\}} d(a, b)$, and thus $h(A, B) = \max\{d(A, B), d(B, A)\}$.

For each $a \in A$, we have

$$d(a, B) = min_{\{b \in B\}} d(a, b) \leq min_{\{b \in B\}} (d(a, c) + d(c, b)), \forall C \in C.$$

Then $\forall c \in C$,

$$\begin{aligned} min_{\{b \in B\}} (d(a, c) + d(c, b)) &= d(a, c) + min_{\{b \in B\}} d(c, b) \\ &= d(a, c) + d(c, B) \leq d(a, c) + d(C, B). \end{aligned}$$

The second term in the last expression does not depend on c, so taking minimization over c results in $d(a, B) \leqslant d(a, C) + d(C, B)$. Furthermore, taking maximization over a on the right leads to $d(a, B) \leqslant d(A, C) + d(C, B)$, and maximizing on the left gives the desired $d(A, B) \leqslant d(A, C) + d(C, B)$. Similarly, $d(B, A) \leqslant d(B, C) + d(C, A)$. Hence, $h(A, B) = \max\{d(A, B), d(B, A)\} \leqslant \max\{d(A, C) + d(C, B), d(B, C) + d(C, A)\} \leqslant \max\{d(A, C), d(C, A)\} + \max\{d(B, C), d(C, B)\} = h(A, C) + h(C, B)$.

References

Barrett, M., Donoghue, M.J., Sober, E., 1991. Against consensus. Syst. Zool. 40, 486–493.

Borůvka, O., 1926. O jistém problému minimálním (About a certain minimal problem). Práce mor. přírodověd. spol. v Brně III 3, 37–58 (in Czech, German summary).

Comin, M., Verzotto, D., 2012. Alignment-free phylogeny of whole genomes using underlying subwords. Algorithms Mol. Biol. 7 (1), 34.

Dai, Q., Yang, Y., Wang, T., 2008. Markov model plus k-word distributions: a synergy that produces novel statistical measure for sequence comparison. Bioinformatics 24, 2296–2302.

Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS ONE 6, e17293.

Gao, Y., Luo, L., 2011. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. Gene 492 (1), 309–314.

Grigoras, I., Timchenko, T., Gronenborn, B., 2007. Transcripts encoding the nanovirus master replication initiator proteins are terminally redundant. J. Gen. Virol. 89, 583–593.

Holmes, E.C., 2009. The comparative genomics of viral emergence. Proc. Natl. Acad. Sci. U.S.A. 107, 1742–1746.

Holmes, E.C., 2011. What does virus evolution tell us about virus origins? J. Virol. 86, 5247–5251.

Kantorovitz, R.M., Robinson, E.G., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics 23, i249–i255.

Koonin, E.V., Wolf, Y.I., Nagasaki, K., Dolja, V.V., 2008. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat. Rev. Microbiol. 6, 925–939.

Liu, D., Shi, W., Shi, Y., Wang, D., et al., 2013. Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. Lancet 381, 1926–1932.

Medina, R.A., Torres-Perez, F., Galeno, H., et al., 2009. Ecology, genetic diversity, and phylogeographic structure of andes virus in humans and rodents in chile. J. Virol. 83 (6), 2446–2459.

Morgan, F., 1987. Geometric Measure Theory. A Beginner's Guide. Academic Press, New York.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Sims, G.E., Jun, S.-R., Wu, G.A., Kim, S.-H., 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. USA 106 (8), 2677–2682.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison – a review. Bioinformatics 19, 513–523.

Yu, C., Liang, C., Yin, C., He, R.L., Yau, S.S.-T., 2010. A novel construction of genome space with biological geometry. DNA Res. 17, 165–168.

Yu, C., Hernandez, T., Zheng, H., Yau, S.-K., Huang, H.-H., He, L.R., Yang, J., Yau, S.S.-T., 2013. Real time classification of viruses in 12 dimensions. PLoS ONE 8, e64328.

Zhou, J., Wang, D., Gao, R., Zhao, B., et al., 2013. Biological features of novel avian influenza A (H7N9) virus. Nature. http://dx.doi.org/10.1038/nature12379.