Molecular Phylogenetics and Evolution 99 (2016) 53–62

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Virus classification in 60-dimensional protein space

Yongkun Li [a], Kun Tian [a], Changchuan Yin [b], Rong Lucy He [c], Stephen S.-T. Yau [a,*]

[a] *Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China*
[b] *Department of Mathematics, Statistics and Computer Science, The University of Illinois at Chicago, Chicago, IL 60607-7045, USA*
[c] *Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA*

ARTICLE INFO

Article history:
Received 14 June 2015
Revised 24 January 2016
Accepted 10 March 2016
Available online 15 March 2016

Keywords:
Hausdorff distance
Natural vector
Natural graphical representation
Virus classification

ABSTRACT

Due to vast sequence divergence among different viral groups, sequence alignment is not directly applicable to genome-wide comparative analysis of viruses. More and more attention has been paid to alignment-free methods for whole genome comparison and phylogenetic tree reconstruction. Among alignment-free methods, the recently proposed "Natural Vector (NV) representation" has successfully been used to study the phylogeny of multi-segmented viruses based on a 12-dimensional genome space derived from the nucleotide sequence structure. But the preference of proteomes over genomes for the determination of viral phylogeny was not deeply investigated. As the translated products of genes, proteins directly form the shape of viral structure and are vital for all metabolic pathways. In this study, using the NV representation of a protein sequence along with the Hausdorff distance suitable to compare point sets, we construct a 60-dimensional protein space to analyze the evolutionary relationships of 4021 viruses by whole-proteomes in the current NCBI Reference Sequence Database (RefSeq). We also take advantage of the previously developed natural graphical representation to recover viral phylogeny. Our results demonstrate that the proposed method is efficient and accurate for classifying viruses. The accuracy rates of our predictions such as for Baltimore II viruses are as high as 95.9% for family labels, 95.7% for subfamily labels and 96.5% for genus labels. Finally, we discover that proteomes lead to better viral classification when reliable protein sequences are abundant. In other cases, the accuracy rates using proteomes are still comparable to that of genomes.

Published by Elsevier Inc.

1. Introduction

With fast development of sequencing technology, an increasing number of viral sequences have been available. Phylogenetic and taxonomic studies on viral sequences become increasingly important for understanding diversities and origins of viruses (Holmes, 2010). Traditional approaches mostly are based on pairwise and multiple sequence alignment. There is high rate of divergence between different virus sequences due to gene mutation, horizontal gene transfer, gene duplication, gene insertion and deletion (Duffy et al., 2008). These features pose a challenge to phylogenetic investigation of viruses. Furthermore, whole genome sequences generally supply more comprehensive information for inferring the phylogeny of viruses than a few orthologous genes (Wong et al., 2008). Since genomes or proteomes include a lot of genes or proteins, the existing methods relating to multiple sequence comparison are computationally intensive (Vinga and Almeida, 2003). Thus they are not suitable for genome-wide phylogeny analysis.

In the past ten years, there has been a growing interest in genome based alignment-free methods for evolutionary studies. Among them, the *k*-mer related methods are all based on word frequencies, which ignore the position of nucleotides (Dai et al., 2008; Wu et al., 2009). In comparison, the natural vector method characterizes both the count and position information of nucleic acids (Deng et al., 2011). The NV method has succeeded in classifying viruses and reconstructing phylogenetic trees (Yu et al., 2013). The NV representation builds a one-to-one correspondence between a DNA sequence and a 12-dimensional numerical vector. Thus we establish a 12-dimensional genome space. Since the Euclidean distance between points in this space can represent their biological similarity to some extent, it allows comparing viruses simultaneously at family level, subfamily and genus levels. As some viral genomes are in the form of several segments, each segment corresponds to a point in R^{12} by the NV method, and then each virus corresponds to a set of points in R^{12}. Recall the general definition of Euclidean distance:

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

http://dx.doi.org/10.1016/j.ympev.2016.03.009
1055-7903/Published by Elsevier Inc.

$d(a,b)=\sqrt{(a_1-b_1)^2+\cdots+(a_d-b_d)^2}$, where $a=(a_1,\ldots,a_d)$, $b=(b_1,\ldots,b_d)$, d is the dimension of vectors a and b. So it is only used to measure the distance between two points. To solve this, the Hausdorff distance is used to measure the distance between point sets, which results in the global comparison of multiple segmented viruses, including single-segmented viruses as well (Huang et al., 2014).

It is of importance to determine whether virus classification using whole-proteomes is indeed better than classification using whole genome. Although one gets nucleotide sequences first, it is increasingly feasible to get corresponding protein sequences as many of gene annotations have been done automatically or manually. Moreover, proteome sequences may be directly involved in determining the variety of functions and the structure of viruses. Mutation in a protein may directly affect its functions which likely result in phenotype changes in evolution. However, changes to nucleic acids may not lead to a protein mutation, because of degeneracy of genetic codons and presence of introns. Even though it was suggested that using proteome sequences was better than using whole genome DNA sequences for genome-based phylogeny reconstruction (Xu and Hao, 2009; Yu et al., 2010a,b; Xie et al., 2015), these studies were only based on a specific Baltimore class or certain families of viruses in the National Center for Biotechnology Information database (NCBI). Virus classification by proteomes has not been systematically characterized. Therefore, we have done a large scale test using almost all viruses in RefSeq database, which is a reliable, non-redundant, and annotated reference subset of NCBI. We compare the results obtained through whole proteome sequences with those by whole DNA genomes.

The phylogenetic tree is a useful tool for classifying and inferring the origin of organisms. Traditionally, this tree has been constructed on the basis of a distance or dissimilarity matrix of species. Many algorithms such as the neighbor-joining algorithm (Saitou and Nei, 1987), have been designed to recover this tree from this matrix. But there are some disadvantages to the resulting phylogeny tree. For instance, the tree may not be unique if the dissimilarity matrix doesn't obey the triangle inequality (Buneman, 1974). To overcome these limitations, the natural graphical representation was proposed (Yu et al., 2013). It has been shown to perform well and can be computed efficiently.

In this paper, we determine the classification of 4021 viruses in seven Baltimore classes based on the NV representation of proteomes and Hausdorff distance. Additionally we also apply the natural graphical representation to show the viral phylogeny. To validate the advantages of proteomes in virus classification, we further process the single-segmented viruses in Baltimore class IV with k-mer method and NV approach based on genomes and Euclidean distance as comparison.

2. Materials and methods

2.1. Overview of the viral data sets

Viruses exhibit more biological diversity than the rest of bacterial, plant, and animal kingdoms. Genomes of viruses may be single-stranded or double-stranded, linear or circular, and in a single-segmented or multi-segmented configuration. In this work, we first downloaded all the referenced protein sequences corresponding to the 4021 viruses as well as their referenced genome sequences from RefSeq database release 69 (January 7, 2015) from NCBI. Traditionally, viruses are classified into seven Baltimore classes. The information for each class in the proteome data set is summarized in Table 1. The 4021 viruses consist of 91 families, 22 subfamilies and 523 genera in total. The viruses in Baltimore VII class have no subfamily labels, thus we use zero to denote the number of their subfamilies in this table. To be convenient, we number the viruses by integers.

2.2. Natural vector and protein space

Let $\mathscr{L}$ be the set of 20 types of amino acids, i.e., $\mathscr{L}=\{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$, and $S=(s_1,s_2,\ldots,s_n)$ be a protein sequence of length n, that is, $s_i\in\mathscr{L}$, $i=1,2,\ldots,n$. For $k\in\mathscr{L}$, define $w_k(\cdot):\mathscr{L}\to\{0,1\}$ 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 $\mu_k=\sum_{i=1}^{n}i\cdot\frac{w_k(s_i)}{n_k}$ be the mean position of letter k.
(3) Let $D_2^k=\sum_{i=1}^{n}\frac{(i-\mu_k)^2w_k(s_i)}{n_kn}$ be the normalized 2-nd central moment of positions of letter k.

For ambiguous amino acids, 1-letter B represents N or D; Z for E or Q; J for I or L; and X for all possible 20 types of amino acids. Thus for $k\in\mathscr{L}$ we define the weight $w_k(s_i)$ as the expected count of letter k in position i. For instance,

$$w_N(s_i)=\begin{cases}1, & s_i=N\\ 0.5, & s_i=B\\ 0.05, & s_i=X\\ 0, & \text{otherwise.}\end{cases}$$

The 60-dimensional NV of a protein sequence S is defined by $(n_A,n_R,\ldots,n_V,\mu_A,\ldots,\mu_V,D_2^A,\ldots,D_2^V)$. For nucleotide sequences, we have similarly defined NV, see Yu et al. (2013).

2.3. Hausdorff distance

Once each protein sequence is mapped to a unique point in the 60-dimensional NV space, each virus then corresponds to a set of points. But in our dataset, three viruses (#121, #297 and #700 in Baltimore class II) share the same set of proteins with other viruses. To ensure one-to-one correspondence between viruses and set of NVs, these three viruses were excluded from the subsequent study. The Hausdorff distance is utilized to measure the pairwise distance between point sets (Huttenlocher et al., 1993). This distance has been suitable to reconstruct the phylogenetic tree for multi-segmented viral genomes from different families when combined with Lempel–Ziv complexity or NV representation of nucleotide sequences (Yu et al., 2014; Huang et al., 2014). The extended version of it, Yau–Hausdorff distance, has achieved success in matching graphical curves of DNA or protein sequences (Tian et al., 2015).

To be precise, suppose A and B are two finite point sets in R^n. Their Hausdorff distance is defined by

$$h(A,B)=\max\left\{\max_{a\in A}\min_{b\in B}d(a,b),\max_{b\in B}\min_{a\in A}d(a,b)\right\},$$

where $d(a,b)$ is the Euclidean distance between two numeric vectors a and b. Note that when both the set A and B have only one member, the Hausdorff distance is reduced to the Euclidean distance. Additionally, unlike many similarity/distance measures in genomics, the Hausdorff distance is a true distance in the sense of mathematics, i.e. it is nonnegative, symmetric and satisfies triangle inequality. When comparing two viruses, this distance is free from the order of viral protein sequences in the form of NVs. The viral classification and phylogenetic tree can be built efficiently using the Hausdorff distance.

Table 1
Summary of seven Baltimore classes in our data set.

Baltimore class Name	I dsDNA	II ssDNA	III dsRNA	IV (+)ssRNA	V (−)ssRNA	VI ssRNA-RT	VII dsDNA-RT	Total
#species	1877	708	95	970	250	52	69	4021
#family	30	7	8	34	9	1	2	91
#subfamily	10	3	2	3	2	2	0	22
#genus	270	40	24	139	41	7	9	523

2.4. Prediction of viral label and the natural graphical representation

In order to predict the viral ranks in the taxonomic hierarchy, we search the nearest neighbor for each virus in the 60-dimensional protein space and infer that the viral has same rank as its neighbor. Due to the sparsity of data, practically to predict the family label, for each virus p in a Baltimore class, for example class I, we find its nearest neighbor q restricted to Baltimore class I which belongs to family Q. Then for each virus in Q, we compute its nearest distance to other virus in Q and collect all the derived distances to get their 95% quantile, denoted as L. Here 95% quantile is the distance value which is exactly larger than the 95% of all distances considered. If the Hausdorff distance $h(p, q)$ between p and q is smaller than L, then we predict that virus p is in family Q, otherwise we don't predict its label. For the subfamily and genus prediction, we perform the same process. It is important to note that the genus prediction is made given the family rank because most of viruses are not specified to the subfamily rank. Comparing to the 75% quantile used in the previous work (Yu et al., 2013), we choose the 95% quantile in this investigation, which may produce a relatively high accuracy rate and reduce the number of unpredicted viruses. Here we use a to denote the number of predicted viruses, b the number of correctly predicted viruses, we define the accuracy rate as b/a. It should be noted that those viruses with missing labels were excluded before accuracy rate predictions. To compare the difference between genome and proteome sequences of viruses, we also build the 12-dimensional genome space and make the prediction in the way mentioned above. We employ the natural graphical representation to display the phylogeny of viruses as well.

We can also infer the labels of viruses that are not assigned to any family or genus. For each virus without a label, we find its nearest neighbor among viruses with known labels. If their distance is less than a certain cutoff from the taxon of its neighbor, the unassigned virus can be considered part of the taxon. In fact, our results show that the rate of unpredicted viruses is more than 20% for 0.75-cutoff and less than 10% for 0.95-cutoff. But the 0.75-cutoff produces a slightly higher accuracy rates than the 0.95-cutoff, so here we choose the 0.75-cutoff to get reliable predictions.

3. Results

The accuracy rates of prediction for the seven Baltimore classes are shown in Table 2, in which the digits within angle brackets are the accuracy rates using genomes and the other digits are accuracy rates using proteomes.

3.1. Prediction of dsDNA viruses (class I)

There are 1877 viral species in this class composed of 30 families, 10 subfamilies, and 270 genera. The missing rates are 0.055 for family, i.e. 5.5 percent of viruses have no family labels. The rates are as high as 0.841, 0.599 for subfamily and genus respectively. The accuracy rates at different levels for proteome-based method and genome-based method are listed in Table 2. Using the proteome data set with the 0.95-cutoff, the accuracy rates at different levels are 0.884 for family labels, 0.961 for subfamily labels, 0.890 for genus labels, while for the genome data set, they are 0.819, 0.951 and 0.873 respectively for family, subfamily and genus (Table 2). Thus the result obtained by proteomes is generally better than that by genomes. We also predict the viruses with missing labels. As a result, within 104 viruses unspecified to families and 1125 viruses unspecified to genera, 54 viruses are classified into families and 292 are assigned to genera. For detail, one may refer to Table S2 and S3 in the supporting files.

3.2. Prediction of ssDNA viruses (class II)

The single-stranded DNA (ssDNA) viruses in our study have 708 members composed of 7 families, 3 subfamilies and 40 genera. The number of viruses remaining in each family are as follows: Anelloviridae 45; Circoviridae 76; Geminiviridae 371; Inoviridae 39; Microviridae 19; Nanoviridae 8; Parvoviridae 85. In addition, 65 viruses remain unassigned to any family. Using the proteome data set, the taxonomic accuracy rates for the 708 ssDNA viruses at different levels are 0.988 for family labels, 0.966 for subfamily labels, 0.960 for genus labels, while for the genome data set, they are as high as 0.983, 0.972, and 0.971 for family, subfamily and genus, respectively. The results of these two methods are comparable to each other. For this class of viruses, here we take the families Inoviridae, Nanoviridae and Microviridae as examples to reconstruct the evolutionary relationship among the viruses within the same families. For the viruses unassigned to any family or genus, using the 0.75-cutoff method, 51 viruses are assigned to families and 30 are assigned to genera. For detail, one may refer to Table S4 and S5 in the supporting files.

The phylogeny of virus in the family Inoviridae is illustrated in Fig. 1, which contains 28 viruses in Inovirus, 5 in Plectrovirus, and 6 without genus labels. In this graph, the genera are in different colors except the unassigned viruses in gray. The two genera are separated except the two viruses #10 (Acholeplasma phage MV-L1) and #514 (Spiroplasma phage SVTS2) from genus Plectrovirus. The viruses in Inoviridae are rod-shaped, non-enveloped, filamentous, and circular in DNA configuration. They contain the inoviruses of gram-negative bacteria and the plectroviruses of mollicutes, of which the SpV1 viruses, Spiroplasma phage 1-C74 (#511) and Spiroplasma phage 1-R8A2B (#512), are the representatives. For the genus Inovirus, we compute its 0.75-cutoff as 267.35, thus we make the reliable forecast that viruses #527, #442, #698, and #441 belong to this genus. The Spiroplasma phage SVTS2 (#514) is a SpV1-like virus infecting *Spiroplasma melliferum*, a honeybee pathogen. From our graph, its closest neighbor (#693) is in genus Inovirus with distance 262.37 which is less than the cutoff of the genus Inovirus. The second closest neighbor of virus #514 is virus #696 (distance 279.06) which is also in the same genus. Previous work (Sha et al., 2000) demonstrated that this virus shared nearly half of SpV1 genomes of other two representative spiroplasma plectroviruses and thus was tentatively paced in the genus Plectrovirus. Our results show that virus #514 likely belongs to genus Inovirus. For virus #10, its nearest neighbor is #450 and next closest neighbor is #696 both of which are in the genus Inovirus. Therefore it should be in the genus Inovirus. It is

Table 2
Comparison of accuracy rates of seven viral Baltimore classes between proteomes and genomes.

Baltimore class Name	I dsDNA	II ssDNA	III dsRNA	IV (+)ssRNA	V (−)ssRNA	VI ssRNA-RT	VII dsDNA-RT
Family							
Without cutoff							
Proteomes	0.859	0.942	0.953	0.942	0.943	1.00	1.00
Genomes	<0.812>	<0.949>	<0.977>	<0.919>	<0.963>	<1.00>	<1.00>
With cutoff							
	0.884	0.988	0.975	0.965	0.965	1.00	1.00
	<0.819>	<0.983>	<1.00>	<0.944>	<0.978>	<1.00>	<1.00>
Subfamily							
Without cutoff							
	0.950	0.957	0.943	0.987	1.00	0.960	
	<0.950>	<0.914>	<0.943>	<0.933>	<1.00>	<0.960>	
With cutoff							
	0.961	0.966	0.980	1.00	1.00	0.979	
	<0.951>	<0.972>	<0.940>	<0.943>	<1.00>	<0.979>	
Genus							
Without cutoff							
	0.855	0.960	0.841	0.902	0.885	0.771	0.949
	<0.862>	<0.957>	<0.921>	<0.872>	<0.901>	<0.813>	<0.966>
With cutoff							
	0.890	0.973	0.920	0.928	0.963	0.860	0.944
	<0.873>	<0.971>	<0.962>	<0.897>	<0.930>	<0.864>	<0.981>

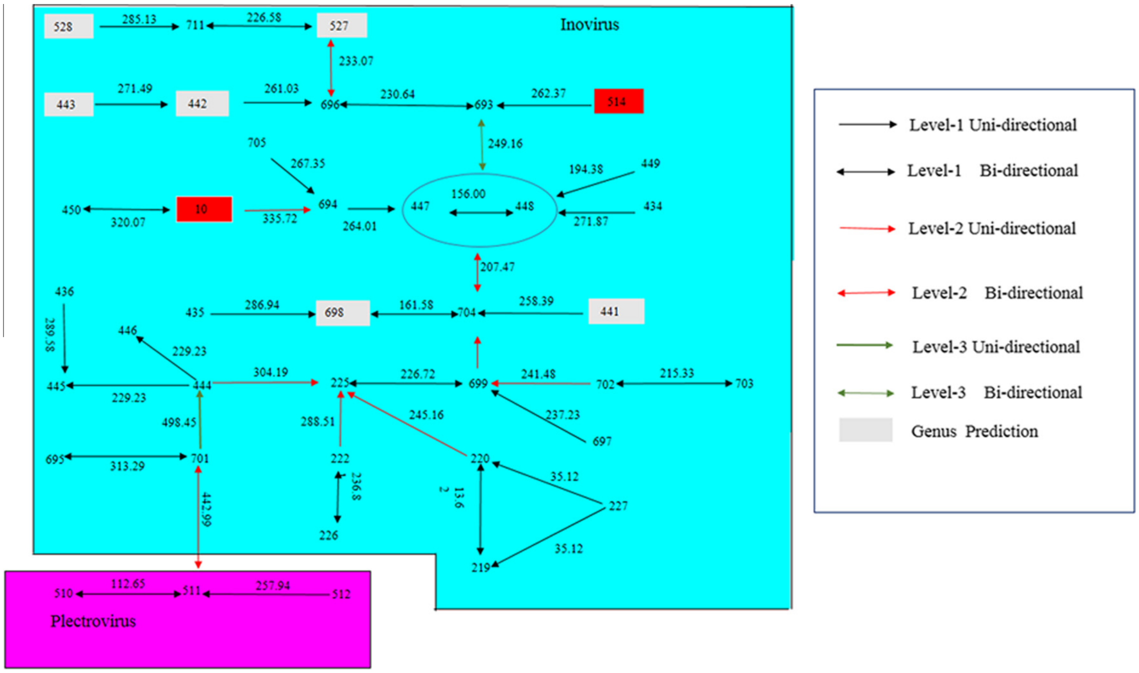

Fig. 1. The natural graphical representation of 39 viruses of family Inoviridae.

interesting to find that the distance from #447 to any other virus except #448 is equal to that from #448 to the same virus. This may be due to the high similarity of nucleotide sequences of the two viruses.

We draw the natural graphical representation for 8 viruses in family Nanoviridae in Fig. 2. They are divided into two monophyletic groups, the genera Babuvirus and Nanovirus. The distance between viruses #384 and #383 is small, so we infer that #384 also belongs to genus Nanovirus.

The natural graphical representation for 19 viruses in family Microviridae is shown in Fig. 3. Note that the four genera Microvirus, Chlamydiamicrovirus, Bdellomicrovirus, and Spiromicrovirus clearly form four clades. The nearest neighbor of #86 is #87 and the distance between them is only 126.77, which is far smaller than the second nearest distance (330.70) from #86 to other viruses, but 330.70 is dramatically larger than the distance of any two viruses among Microvirus genus. Therefore we conclude that the two viruses #87 and #86 form a new genus, which is consistent with the result of Karin et al. (2013).

3.3. Prediction of dsRNA viruses (class III)

In RefSeq database, there are 202 viral species in the Baltimore III class composed of 11 families, 2 subfamilies, and 33 genera.

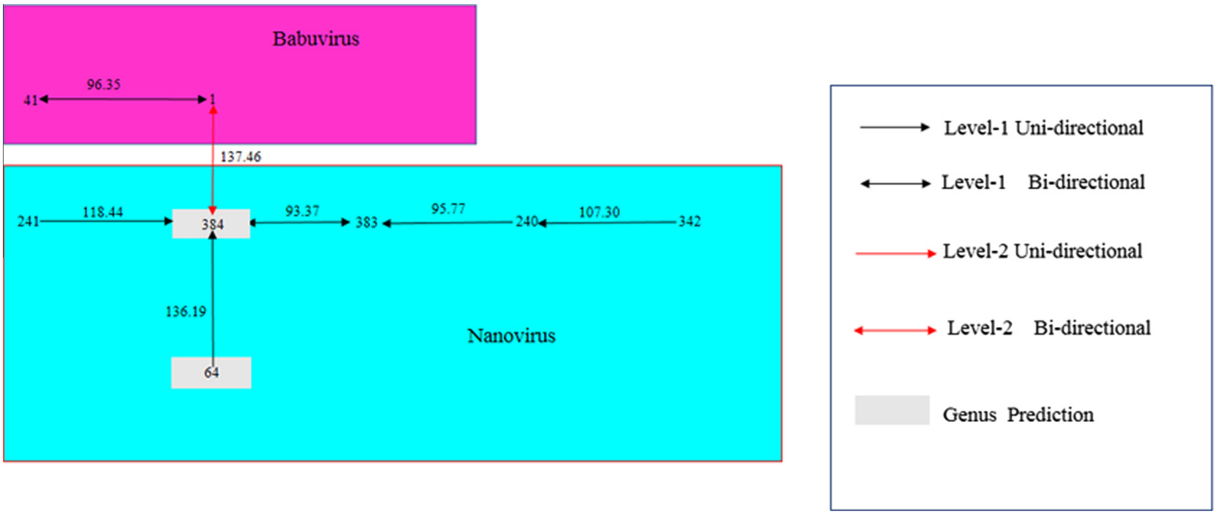

Fig. 2. The natural graphical representation of 8 viruses of family Nanoviridae.

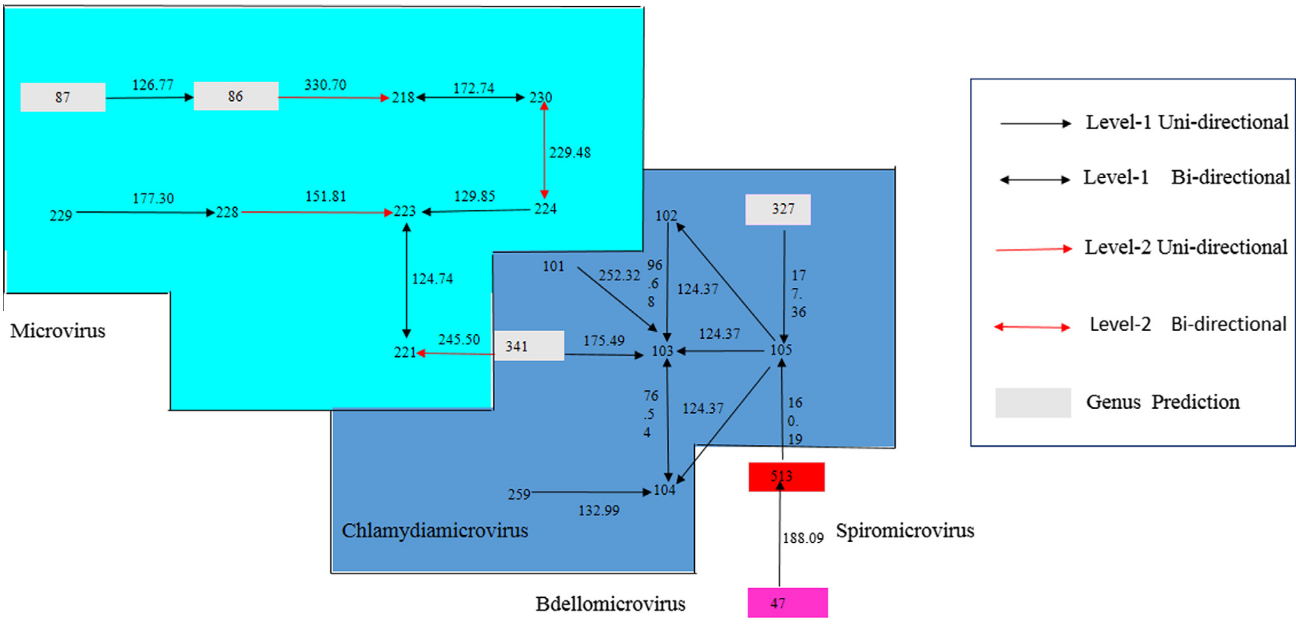

Fig. 3. The natural graphical representation of 19 viruses of family Microviridae.

Among these viruses, 107 species consist of no more than two proteins. The lack of referenced proteins may be caused by the limited amount of investigation done on them and the complexity of gene expression. To improve the forecast without loss of much information, we only analyzed the 95 viruses that remained after eliminating the 107 viruses. These viruses consist of 8 families, 2 subfamilies, and 24 genera. The missing labels for family, subfamily and genus are 0.074, 0.442, and 0.232 respectively. Using the proteome data set with the 0.95-cutoff, the accuracy rates at different levels are 0.975 for family labels, 0.980 for subfamily labels, 0.920 for genus labels, while for the genome data set, they are 1.00, 0.940 and 0.962 respectively. The results of these two methods are comparable to each other.

3.4. Prediction of (+)ssRNA viruses (class IV)

The Baltimore IV includes 970 viral species which form 34 families, 3 subfamilies, and 139 genera. The rates of viruses unassigned to any family, subfamily, and genus are 0.066, 0.923, and 0.141 respectively. Using the proteome data set with the 0.95-cutoff, the accuracy rates of classification at different levels are 0.965 for family labels, 1.00 for subfamily labels, 0.928 for genus labels, while for the genome data set, they are 0.944, 0.943 and 0.897 respectively (Table 2). Thus the result obtained by proteomes is uniformly better than that by genomes.

Within the 970 viral species, 104 viruses have no family labels and 137 viruses have no genus labels. Using the 0.75-cutoff prediction, 17 viruses are assigned to families and 59 are assigned to genera. For detail, one may refer to Table S6 and S7 in the supporting files.

The phylogeny of 45 viruses in family *Alphaflexiviridae* is illustrated in Fig. 4. Except the virus #1 and #10 which form a new clade, the other viruses within genus Potexvirus cluster together. For remaining genera, each forms its own clade regardless of the number of member in the genus. The virus #6 is not assigned to any genus, but it is next to the virus #14 at the distance 225.6.

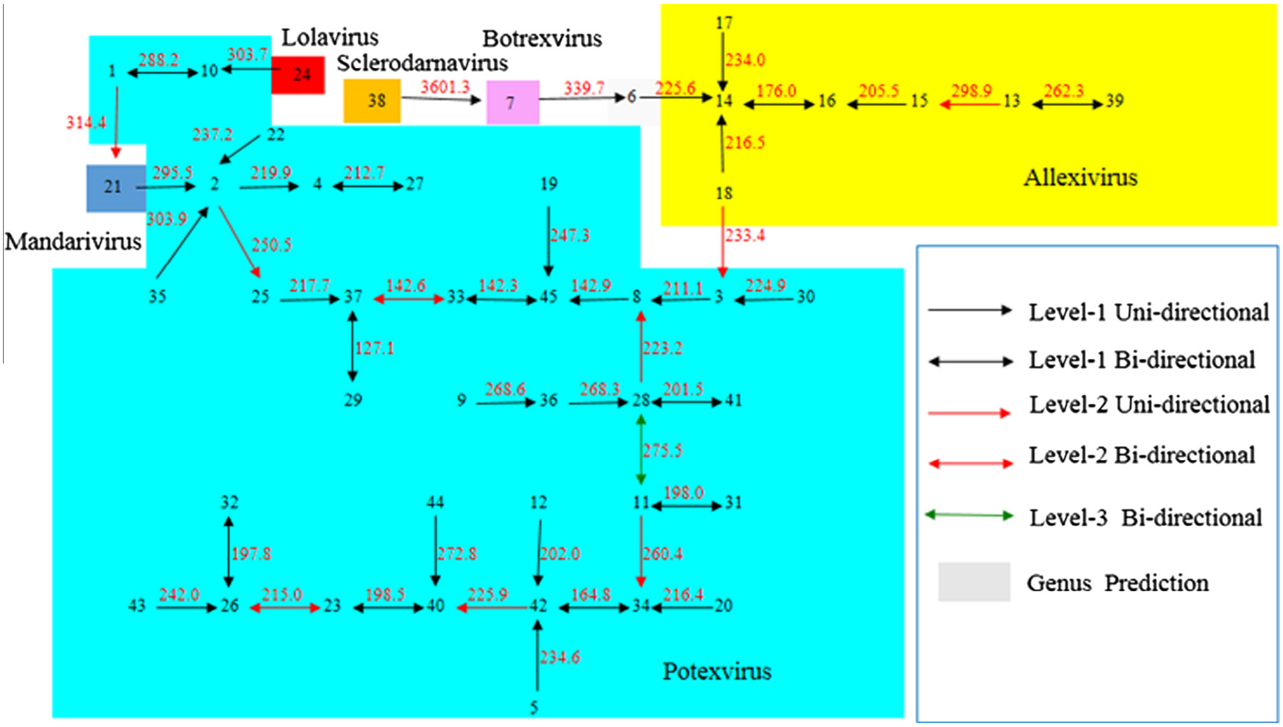

Fig. 4. The natural graphical representation of 45 viruses of family *Alphaflexiviridae*.

Since the median value of all nearest distances of viruses in genus Allexivirus is only 225.25, the virus #6 is thought to be in genus Allexivirus.

3.5. Prediction of (−)ssRNA viruses (class V)

There are 250 viral species in this class consisting of 9 families, 2 subfamilies, and 41 genera. For the family, subfamily and genus labels of viruses, the missing labels rates are 0.024, 0.772, and 0.168 respectively. Using proteome sequences, the accuracy rates of prediction are 0.965, 1.00, and 0.963 for family, subfamily and genus. For the genome data set, they are 0.978, 1.00 and 0.930 correspondingly (Table 2). Thus the result obtained by proteomes is comparable to that by genomes. For viruses without genus labels, using the 0.75-cutoff prediction, 14 viruses are classified into certain genera. For detail, see Table S8 in the supporting files.

3.6. Prediction of ssRNA-RT viruses (class VI)

In current RefSeq database, there are 62 viral species in the Baltimore VI class composed of one family, two subfamilies, and seven genera. In this small family, there exists ten viruses that comprise no more than two proteins. To get a reliable prediction without losing too much information, we only use the remaining 52 viruses for analysis. The missing rates are 0, 0.038, 0.058 for family, subfamily and genus labels. As the viral family labels are all known, the accuracy rate for family labels is 1. The accuracy rates of prediction are 0.979 and 0.860 for subfamily and genus labels. For the genome data set, the accuracy rates are 0.979 and 0.864 for subfamily and genus prediction. Thus the result obtained by proteomes is comparable to that by genomes.

3.7. Prediction of dsDNA-RT viruses (class VII)

There are 69 viral species in this class composed of two families and nine genera, but all these viruses have no subfamily labels. For the family and genus labels of viruses the missing rates are 0, 0.101 respectively and the accuracy rates of prediction are 1.00 and 0.944. For the genome data set, the accuracy rates are 1.00 and 0.981 correspondingly (Table 2). Thus both of the methods classify this class efficiently and comparatively.

In Fig. 5, the genus Caulimovirus (in yellow) forms a monophyletic clade. The genus Badnavirus basically groups together except two viruses #53 and #54. But the four viruses in genus Soymovirus are very divergent, among which the virus #12 and #45 form a clade, the other two viruses are phylogenetically distant. The unclassified virus #34 is next to #24 with a very small distance (37.53). Therefore this virus is considered in genus Badnavirus.

From Table 2, we observe that the accuracy rates of prediction with proteomes are higher than those with genomes at the family, subfamily level for the Baltimore class I. For Baltimore class IV, the proteins are definitely preferable to nuclei acids. As for the Baltimore class II, the proteins outperform nucleotides for family and genus prediction in the case using 0.95-cutoff. For the rest of classes, the accuracy rates based on proteomes are comparative to those using genomes due to the insufficiency of enough reliable proteins contained in these viruses. Generally speaking, the viral genomes are correctly sequenced with minor error in the large sequences. In our study, the 12-dimensional NV based on genomes gets slightly low accuracy rates for at the genus level for Baltimore I and VI, indicating that the genomes may not cover the information needed for accurate classification when sequence divergence is high.

In order to further assess the feasibility of our proteome based method, we investigate the influenza A (H7N9) virus. The virus consists of eight nucleic acid segments, but has a varied number of referenced proteins. Traditionally, the virus appears in birds, but now is found in humans, which poses a new public health threat as its high contagion and fast evolution (Liu et al., 2013). It is crucial to distinguish new H7N9 viruses rapidly and find their relationships in the phylogenetic tree of the viruses. We use the proteomes of 28 H7N9 strains analyzed by Huang (Huang et al.,

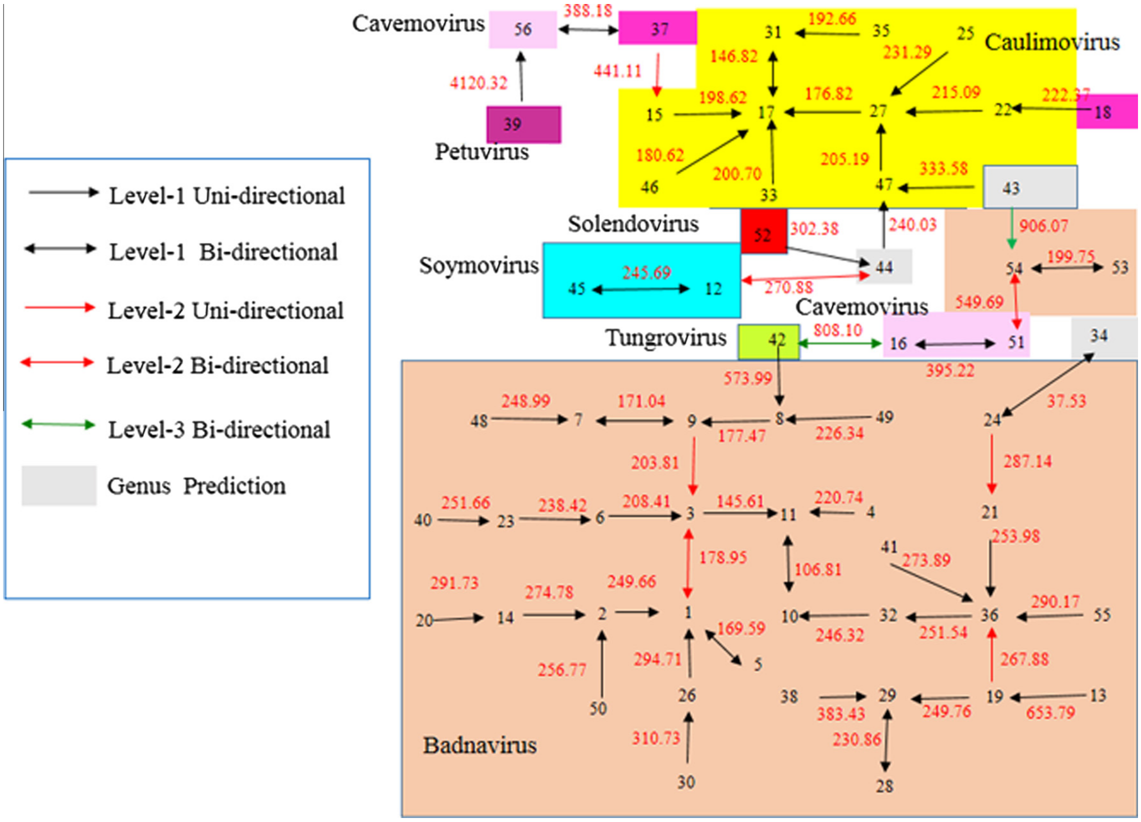

Fig. 5. The natural graphical representation of 56 viruses of family Caulimoviridae

2014). Their names are listed in Table S1 of the supporting files. Using the natural graphical representation with proteomes, their phylogeny is shown in Fig. 6.

The new H7N9 virus isolated from human was found in China and it then infected a Taiwan traveler who visited the region close to Zhejiang province in China mainland. The recent studies show that the six internal genes of this new virus are mutated from the old H9N2 virus. But the origin of other two segments, HA and NA, are unclear. According to our natural graph of the phylogenetic relationship, it is noted that Taiwan 2 has equal distance to Nanjing 1, Nanjing 2 and Taiwan 1, which indicates the Taiwan 2 virus may equally evolve from the three strains. Moreover the Taiwan 1 is also equally close to four viruses, which is very interesting and very reasonable as the host traveled around the Zhejiang. Since the Mongolia virus and Delaware virus are far away from the Nanjing 2 virus and Taiwan 1 virus respectively in the graph, it indicates that they are two phylogenetically unrelated clades. So are the South Korea 2, Minnesota 2. For USA Alaska, its first three neighbors are Guatemala 2, Mississippi, and Delaware Bay, with the corresponding distances 109.66, 109.97, 110.46, which are much bigger than the distances among countries or states in the gray box. It is illustrated that the Shanghai 2, Hangzhou, and Zhejiang 2 viruses are clustered together and they are closer to the avian strains in South Korea than the rest of those in mainland China. Therefore it is possible that some new viruses in China may be from South Korea by the migration of wild birds.

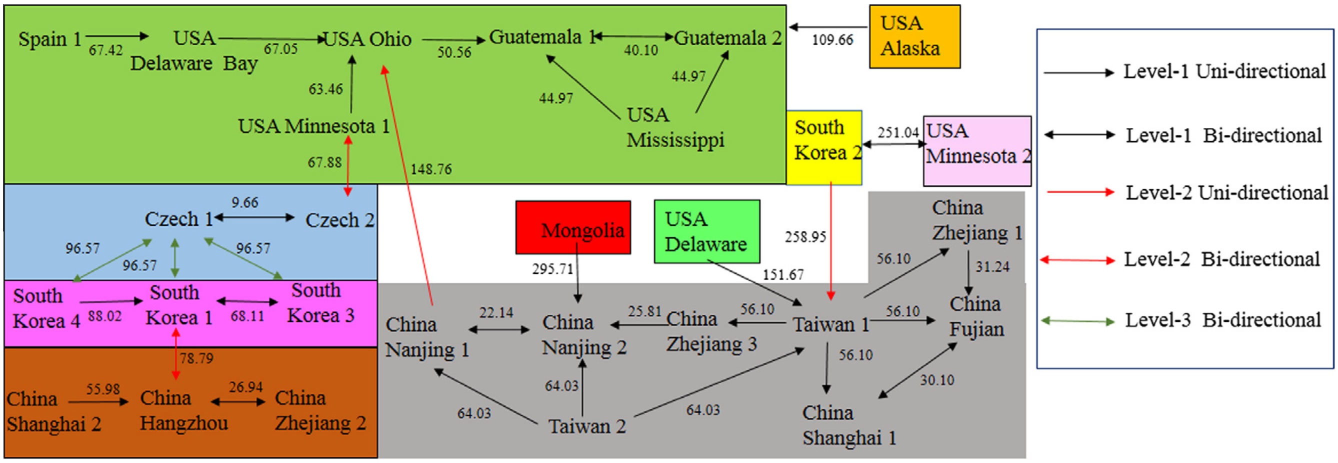

Fig. 6. The natural graphical representation of 28 strains of H7N9 based on the proteomes.

3.8. Comparison with other methods

The *k*-mer method has been extensively applied to analyze genomes of organisms including viruses. But the important parameter *k* is usually selected subjectively and dependent on typical data sets (Vinga and Almeida, 2003; Yu et al., 2010a; Wu et al., 2009). For another alignment-free method, the NV representation based on genomes and Euclidean distance proves to be efficient for classifying single-segmented viruses (Yu et al., 2013). In this work, we compare our method with these two approaches. The 970 viruses in Baltimore IV class (positive-sense ssRNA) are used as a test dataset. Since some RNA viruses have multiple segments in their genomes and each segment often codes for only one protein, to implement our comparison, we first removed the multiple-segmented viruses. Then for those single-segmented viruses, we excluded those with no family or genus labels and the genera with only one species. The resulting dataset contained 638 viruses, which is quite large to analyze. This dataset also included some viruses without enough reliable proteins. In particular, there were 84 with one protein, 21 with two proteins and 59 with three proteins. According to Sims et al. (2009), we chose the optimal resolution k as the minimum integer that is greater than or equal to $\log_4(n_{min})$, where n_{min} is the minimal length of DNA sequences studied. For this dataset, virus 'Ophiostoma novo-ulmi mitovirus 6-Ld' has a minimal genome size of 2343 base pairs, thus the optimum *k* is 6. The results of three methods are shown in Table 3. We notice that the 6-mer, 9-mer and 11-mer methods perform very differently when predicting the family label. The prediction rate for the family level is almost 100%. The rates of unpredicted viruses for the 6-mer, 9-mer and 11-mer methods are 35%, 91% and 92% respectively, while the rate for each NV-based method is about 6%. These results demonstrate that the proteome-based NV method outperforms the other two approaches even when about 26% viruses lack proteins.

To investigate the stability of our method when there are missing proteins, we further remove the viruses with less than five proteins among the 638 viruses and the families and genera with only one species to obtain a dataset containing 351 viruses. For the *k*-mer method, we use the same method to choose the optimal *k*. Beet black scorch virus has a minimal genome size of 3644 base pairs, thus the optimum *k* is 6. The results of the three methods are recorded in Table 4. Although the 6-mer, 9-mer and 11-mer methods perform well for the family level, respectively they leaves more than 30%, 86% and 86% of the 351 viruses unpredicted, while the rate of unpredicted viruses for the other two methods is about 6%. These results also imply that the proteome-based NV method outperforms other two approaches. In addition, compared with Table 3, the three methods basically provide improvement and our proposed method achieves very high accuracy rates, even almost 100% for family label prediction. This provides more evidence that our method will perform better than the genome-based NV method when enough reliable proteins are accessible.

Table 3
Accuracy rates of *k*-mer, genome-based NV and proteome-based NV methods for 638 viruses.

Methods	6-mer	9-mer	11-mer	Genome-based NV	Proteome-based NV
Family					
Without cutoff	0.668	0.096	0.085	0.934	0.961
With cutoff	0.998	1	1	0.953	0.980
Genus					
Without cutoff	0.865	0.605	0.544	0.875	0.903
With cutoff	0.924	0.615	0.556	0.901	0.929

Table 4
Accuracy rates of *k*-mer, genome-based NV and proteome-based NV methods for 351 viruses.

Methods	6-mer	9-mer	11-mer	Genome-based NV	Proteome-based NV
Family					
Without cutoff	0.712	0.148	0.142	0.954	0.977
With cutoff	1	1	1	0.966	0.991
Genus					
Without cutoff	0.883	0.769	0.672	0.858	0.912
With cutoff	0.899	0.794	0.704	0.883	0.940

4. Discussion and conclusion

In this article, we focus on the classification of 4021 viruses with their proteome sequences. This dataset is significantly larger than that in previous work which only contains 2418 virions (Yu et al., 2013; Huang et al., 2014). Based on the natural vector representation, we built a 60-dimensional protein space in which each vector corresponds to a protein sequence. Therefore each virus is mapped to a set of natural vectors (NVs) and the hausdorff distance is applied to measure the dissimilarity among NVs. The classification of viruses is achieved after the dissimilarity matrix is computed. To explore phylogenetic relationships, we used the natural graphic representation based on the Hausdorff distance, which has been proven to infer the phylogeny successfully. To further show the advantages of our natural graphical representation, we also supply the phylogenetic tree of 45 viruses of family *Alphaflexiviridae* (Fig. 7). The viruses of Potexvirus genus form one main clade in our natural graphical representation (Fig. 4). However, in Fig. 7, these viruses are scattered into three main clades, which is inconsistent with the classification of these viruses in NCBI. Moreover the Allexivirus of genus assigned by NCBI also clearly converges into one branch in Fig. 4, whereas it incorrectly diverges into two distant clades in Fig. 7.

Although the proteomes of many viruses are incomplete, the method by proteomes is comparable to that using genomes. For the Baltimore I, II and IV, the former is better than the latter. For the remaining four classes, their results are comparable to each other. But what is most important is that each of the classes I, II and IV is almost three times larger than the largest classes of class III, V, VI and VII. Moreover, the former three classes together contain 3555 of the total 4021 viruses, which provides sufficient justification for the superiority of our method. Thus the proteome of a virus includes more information than its genome for phylogenetic analysis. In practice, if the studied viruses have complete proteome sequences, we suggest the usage of them instead of the corresponding genomes. A noticeable advantage of our method is that it efficiently classifies seven obviously different classes of viruses simultaneously at the family, genus and species levels. The result of our method is of interest that the availability of only a part of the proteomes seems to be enough to reproduce the evolutionary relationships. Our approach is also convenient to build the databases of species, as once natural vectors are computed, it is unnecessary to compute them again. Besides, unlike many model-based methods with a large number of parameters to be estimated, our approach is model-free and natural.

Although our new method has these advantages, it may have some limitations. Even if several international projects have accumulated reliable protein sequences, the complexity of viral gene

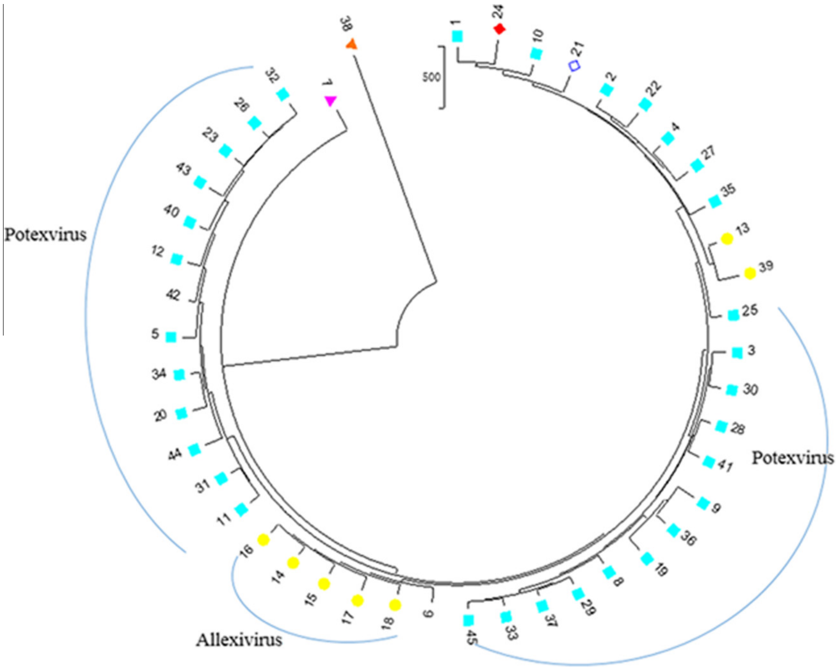

Fig. 7. The phylogenetic tree of 45 viruses of family *Alphaflexiviridae*.

expression such as open reading frame shift, the occurrence of coding genes in introns poses great difficulty in finding enough proteins. In addition, the viruses are highly varied and widespread in the world, resulting in the lack of detailed gene annotations. Since the RefSeq database only collects reliable proteins of viruses, many of the viruses analyzed contain less than 4 proteins which may reduce the efficiency of our method. Moreover, the 0.95-cutoff aims to quantify how divergent a group of viruses is. Thus it requires a number of members in this group, otherwise the cutoff is probably insufficient to represent the nearest distances among viruses. For instance, in the Baltimore I class, both the proteome-based and genome-based NV methods have low efficiency for prediction of family and genus labels. The main reason is that 18 out of 30 families and 255 out of 270 genera only incorporate less than 10 species, i.e. the majority of genera have a few species, which suggests that the dsDNA viruses are too divergent to generate effective cutoff.

5. Conflict of interest statement

The authors declare no competing financial interests.

Acknowledgements

This study is supported by the USA Natural Science Foundation (DMS-1120824 to S.S.-T. Yau), National Natural Sciences Foundation of China (31271408 to S.S.-T. Yau), Tsinghua University start up fund (to S.S.-T. Yau) and Tsinghua University independent research project grant (to S.S.-T. Yau). The funders did not take part in study design; in collection and analysis of data; in the writing of the manuscript; in the decision to publish this manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2016.03.009.

References

Buneman, P., 1974. A note on the metric properties of trees. J. Combin. Theory, Ser. B 17, 48–50.

Dai, Q., Yang, Y., Wang, T., 2008. Markov model plus k-word distributions: a synergy that produces novel statistical measures 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.

Duffy, S., Shackelton, L.A., Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9, 267–276.

Holmes, E.C., 2010. The comparative genomics of viral emergence. Proc. Natl. Acad. Sci. 107, 1742–1746.

Huang, H.H., Yu, C., Zheng, H., Hernandez, T., Yau, S.C., He, R.L., Yang, J., Yau, S.S.T., 2014. Global comparison of multiple-segmented viruses in 12-dimensional genome space. Molecul. Phylogenet. Evol. 81, 29–36.

Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J., 1993. Comparing images using the hausdorff distance. IEEE Trans. Pattern Anal. Mach. Intell. 15, 850–863.

Karin, H., Natalie, S., Manesh, S., Kristen, C., Lasse, R., Nathan, C., Matthew, B., 2013. Twelve previously unknown phage genera are ubiquitous in global oceans. Proc. Natl. Acad. Sci. 110, 12798–12803.

Liu, D., Shi, W., Shi, Y., Wang, D., Xiao, H., Li, W., Bi, Y., Wu, Y., Li, X., Yan, J., 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.

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

Sha, Y., Melcher, U., Davis, R.E., Fletcher, J., 2000. Common elements of spiroplasma plectroviruses revealed by nucleotide sequence of SVTS2. Virus Genes 20, 47–56.

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. 106, 2677–2682.

Tian, K., Yang, X., Kong, Q., Yin, C., He, R.L., Yau, S.S.T., 2015. Two dimensional Yau–Hausdorff distance with applications on comparison of dna and protein sequences. Plos One 10, e0136577.

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

Wong, K.M., Suchard, M.A., Huelsenbeck, J.P., 2008. Alignment uncertainty and genomic analysis. Science 319, 473–476.

Wu, G.A., Jun, S.R., Sims, G.E., Kim, S.H., 2009. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. Proc. Natl. Acad. Sci. 106, 12826–12831.

Xie, X.H., Yu, Z.G., Han, G.S., Yang, W.F., Anh, V., 2015. Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles. Molecul. Phylogenet. Evol. 89, 37–45.

Xu, Z., Hao, B., 2009. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. Nucl. Acids Res. 37, W174–W178.

Yu, Z.G., Chu, K.H., Li, C.P., Anh, V., Zhou, L.Q., Wang, R.W., 2010a. Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. BMC Evolution. Biol. 10, 192.

Yu, Z.G., Zhan, X.W., Han, G.S., Wang, R.W., Anh, V., Chu, K.H., 2010b. Proper distance metrics for phylogenetic analysis using complete genomes without sequence alignment. Int. J. Molecul. Sci. 11, 1141–1154.

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

Yu, C., He, R.L., Yau, S.S.T., 2014. Viral genome phylogeny based on Lempel–Ziv complexity and Hausdorff distance. J. Theoret. Biol. 348, 12–20.