International Journal of
Molecular Sciences

Article

Positional Correlation Natural Vector: A Novel Method for Genome Comparison

Lily He [1], Rui Dong [1], Rong Lucy He [2] and Stephen S.-T. Yau [1,*]

[1] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China; lilyhe6@163.com (L.H.); dongr15@mails.tsinghua.edu.cn (R.D.)
[2] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA; rhe@csu.edu
* Correspondence: yau@uic.edu

Received: 13 April 2020; Accepted: 26 May 2020; Published: 29 May 2020

Abstract: Advances in sequencing technology have made large amounts of biological data available. Evolutionary analysis of data such as DNA sequences is highly important in biological studies. As alignment methods are ineffective for analyzing large-scale data due to their inherently high costs, alignment-free methods have recently attracted attention in the field of bioinformatics. In this paper, we introduce a new positional correlation natural vector (PCNV) method that involves converting a DNA sequence into an 18-dimensional numerical feature vector. Using frequency and position correlation to represent the nucleotide distribution, it is possible to obtain a PCNV for a DNA sequence. This new numerical vector design uses six suitable features to characterize the correlation among nucleotide positions in sequences. PCNV is also very easy to compute and can be used for rapid genome comparison. To test our novel method, we performed phylogenetic analysis with several viral and bacterial genome datasets with PCNV. For comparison, an alignment-based method, Bayesian inference, and two alignment-free methods, feature frequency profile and natural vector, were performed using the same datasets. We found that the PCNV technique is fast and accurate when used for phylogenetic analysis and classification of viruses and bacteria.

Keywords: alignment-free; positional correlation natural vector; phylogenetic analysis; genome comparison

1. Introduction

Predicting the structures, functions, and evolutionary relationships of genes is a fundamental and vital aspect of modern biological research. Therefore, the comparison of genetic sequences is a pivotal step in many protocols and numerous approaches have been employed for this task. Most researchers use conventional alignment-based techniques for sequence comparison; these techniques involve sequence alignment based on selected scoring systems. The algorithms used are generally precise and highlight correlations among sequences. Several sequence alignment methods have been implemented via software packages, such as MrBayes [1]. However, alignment-based methods have disadvantages: they are slow and require a large amount of memory. Furthermore, based on previous studies, multiple sequence alignment (MSA)-based methods cannot be extended with using the huge datasets currently available [2]. Therefore, alignment-free (AF) methods may be used to overcome these problems [3]. Additionally, AF sequence comparison is drawing great interest driven by data-rich applications [4]. A notable common feature of AF approaches is the analysis of special numerical properties of the sequences being compared. High computational efficiency is observed when such techniques are applied to gene and protein data. A series of AF methods for sequence comparison has been developed. AF approaches include iterated-function systems [5], information theory [6], Fourier transformations [7], sequence representations based on chaos theory [8], and moments of the positions of the nucleotides [9,10]. The most widely used AF method is the k-mer-based method and has been published in many excellent journals [11–19]. This method involves the analysis of the frequency of strings of specific length k within sequences [20]. Several k-mer-based methods have been developed and applied for the phylogenetic analysis of bacteria and viruses. A notable example is feature frequency profiles (FFP) [21].

Although k-mer-based methods have been applied widely, they do not include positional correlations of nucleotides. However, it is significant to investigate the location for gene sequence comparison. Therefore, positional correlation is important for computational and analytical approaches. Recently, two methods based on moments of the positions of the nucleotides, namely natural vector (NV) [9] and multiple encoding vector (MEV) method [10], were proposed. They were successfully used for the classification and phylogenetic analysis of sequences. The NV method uses frequency, average site, and variance of site to compare sequences. Based on NV, the MEV method can add information about the chemical and physical properties of a nucleotide. The distribution of four bases is considered independently in these two methods. However, it has been reported that the four bases are correlated; in fact, the correlation of nucleotides is based on the widely applied hidden Markov model (HMM) [22]. In the present study, we propose a novel 18-dimensional numerical feature vector method to characterize DNA sequences. The method is named positional correlation natural vector (PCNV) to characterize DNA sequences. Our vector contains the frequency, average, and variance the locations of four bases. Furthermore, we added the position correlation of each pair of the four bases as important features. We tested the PCNV method using several datasets and compared it with the alignment-based Bayesian inference approach, which can be applied using MrBayes software [1], as well as two AF methods, FFP and NV.

2. Results

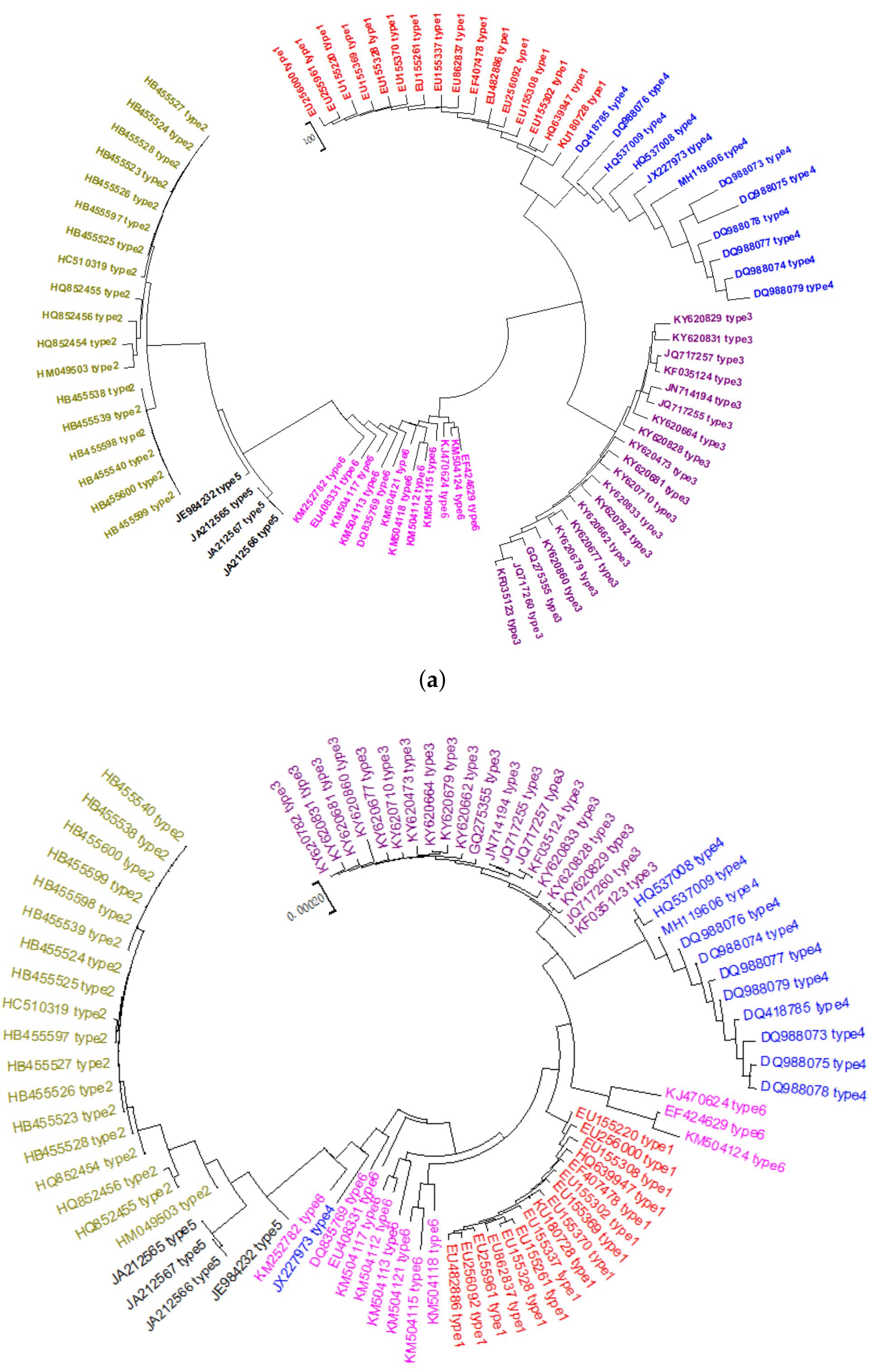
To demonstrate that PCNV is effective, we applied it to different datasets: the genomes of hepatitis C virus (HCV), hepatitis B virus (HBV), human papillomavirus (HPV), dengue virus (DENV), and 59 bacterial species. The length of the sequences studied ranged from thousands to millions of base pairs. For each dataset, the PCNVs of the sequences were computed using MATLAB R2016a and phylogenetic trees were reconstructed using MEGA 7. Finally, we evaluated the performance of our methods based on sensitivity, specificity, and accuracy. Computations were performed on a PC with Intel Core i7-6560U CPU @ 2.20 GHz and 8 GB RAM.

2.1. Phylogeny of HCV

Using our PCNV method, 82 HCVs are correctly clustered into six clades, as shown in Figure 1a [23]. Using the FFP method, the value of k for analysis of these viruses is 6. As shown in Figure 1b, some Genotype 6 and 1 HCVs are clustered together incorrectly in the FFP phylogenetic tree. Furthermore, a sequence in Genotype 4 is assigned to Genotype 6. PCNV produces better results for this dataset than FFP. The Bayesian inference method was also utilized for the evolutionary analysis of this dataset. Figure 1c shows that this method divides Genotype 3, shown in violet, into two groups.

2.2. Phylogeny of HBV

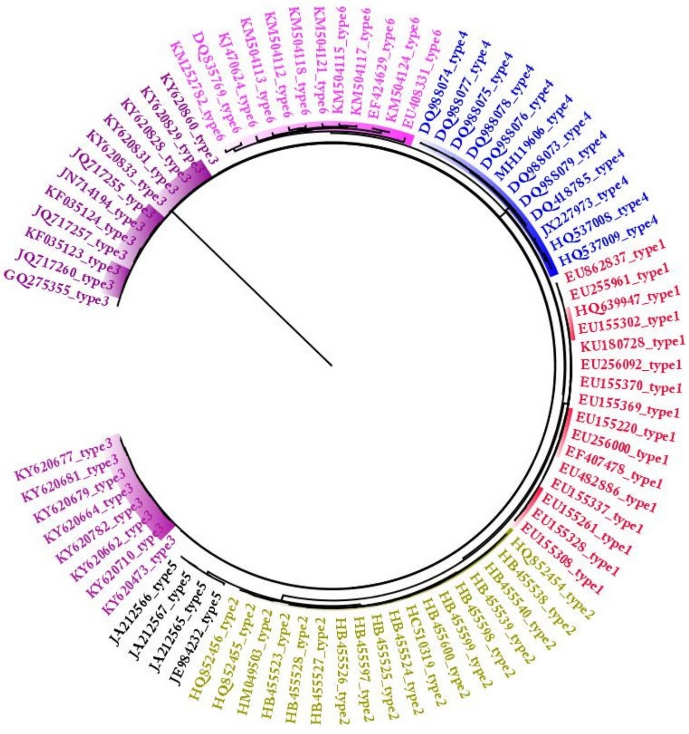
Using PCNV, 152 HBVs are correctly divided into eight lineages, as shown in Figure 2a. The phylogenetic tree created using the FFP method is shown in Figure 2b. According to the HBV database, “AJ627224” belongs to Genotype D. However, according to the FFP method, it is related to Genotype B. The FFP method also cluster “FJ356715” and “FJ356716”, belonging to Genotype H, to Genotypes F and G, respectively. Therefore, for this dataset, our PCNV method is superior to FFP. The phylogenetic tree created using the traditional NV method is shown in Figure 2c. However, in this tree, three Genotype C viruses are classified into other groups. The phylogenetic tree created using Bayesian inference is shown in Figure 2d. It shows that “AB371164” belongs to Genotype H, as separated from Genotype H. This is an indication that the positional correlation between nucleotides can improve the accuracy of classification.



(**a**)

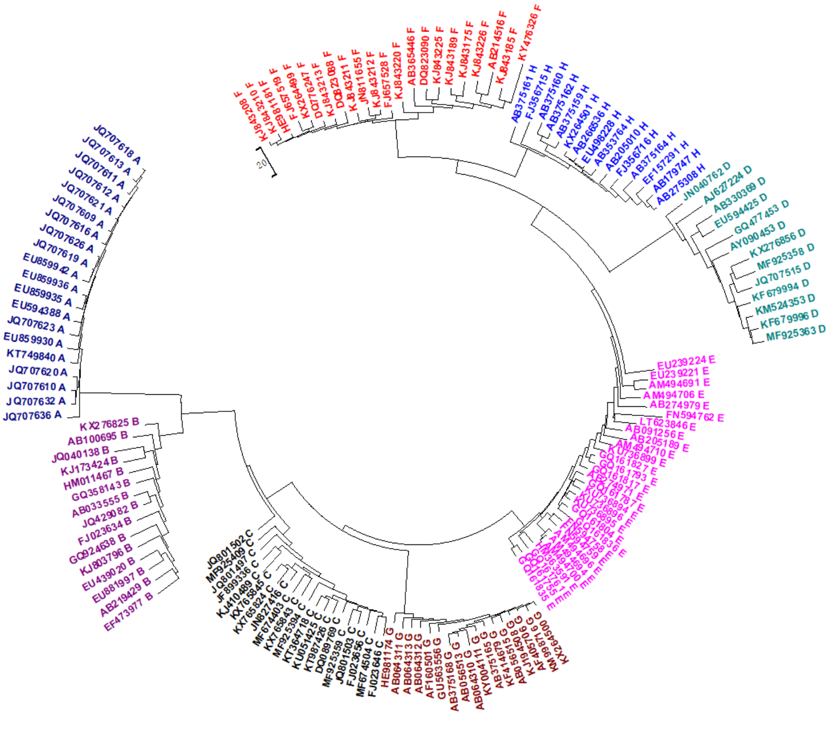
(**b**)

Figure 1. *Cont*.



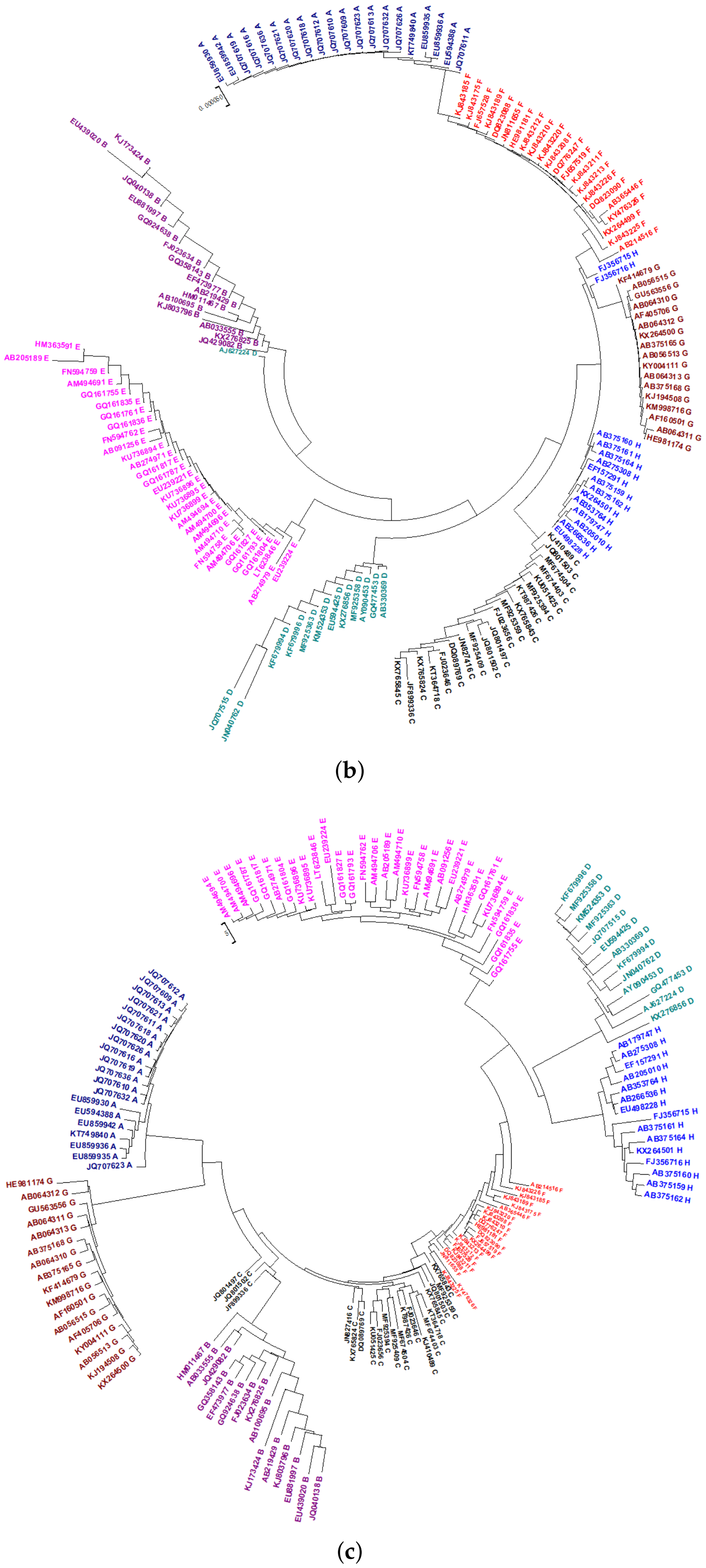
(**c**)

Figure 1. (**a**) The Neighbor-Joining phylogenetic tree of 82 HCV genome sequences based on PCNV method. (**b**) The Neighbor-Joining phylogenetic tree of 82 HCV genome sequences based on FFP method (k = 6). (**c**) The phylogenetic tree of 82 HCV genome sequences based on Bayesian inference method.



(**a**)

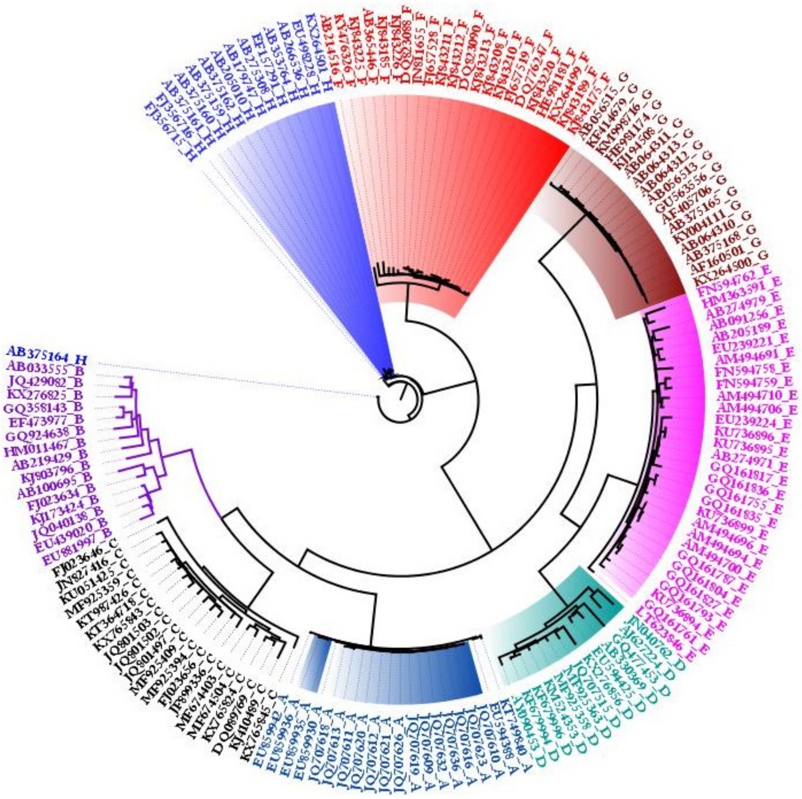
Figure 2. *Cont.*



(**b**)

(**c**)

Figure 2. *Cont.*



(**d**)

Figure 2. (**a**) The Neighbor-Joining phylogenetic tree of 152 HBV genome sequences based on PCNV method. (**b**) The Neighbor-Joining phylogenetic tree of 152 HBV genome sequences based on FFP method (k = 5). (**c**) The Neighbor-Joining phylogenetic tree of 152 HBV genome sequences based on NV method. (**d**) The phylogenetic tree of 152 HBV genome sequences based on Bayesian inference method.

2.3. *Phylogeny of DENV*

As shown in Figure 3a, the phylogenetic tree constructed using PCNV classifies all viruses into the correct categories. However, as shown in Figure 3b, the NV method divides Genotype 1 into two clusters. Therefore, once again, it is clear that positional correlation between nucleotides can effectively improve the NV method.

2.4. *Phylogeny of HPV*

We found that PCNV categorizes the dataset into the correct biological groups in 0.78 s (Figure 4a; Table 1); this is much faster than the FFP method, which takes 35 s (Table 1). The Bayesian inference method divide Genotype 11 into two parts, as highlighted in cyan in Figure 4b.

Table 1. Running time for PCNV, Bayesian inference, FFP, AFKS, and Muscle methods. "∼", unable to compute on laptop.

Method	**HCV (82)**	**HBV (152)**	**Dengue (330)**	**HPV (326)**	**Bacteria (59)**
PCNV	0.33s	0.27s	0.66s	0.78s	53.71s
Bayesian inference	1097s	263s	217,353s	217,512s	∼
FFP	11.11s (k = 6)	0.38s (k = 5)	49.40s (k = 6)	35.00s (k = 6)	larger than 1 day (k = 11)
AFKS	70.21s (k = 5)	29.62s (k = 4)	429.87s (k = 5)	413.79s (k = 5)	larger than 4 day (k = 9)
Muscle	753s	155s	3740s	4002s	∼

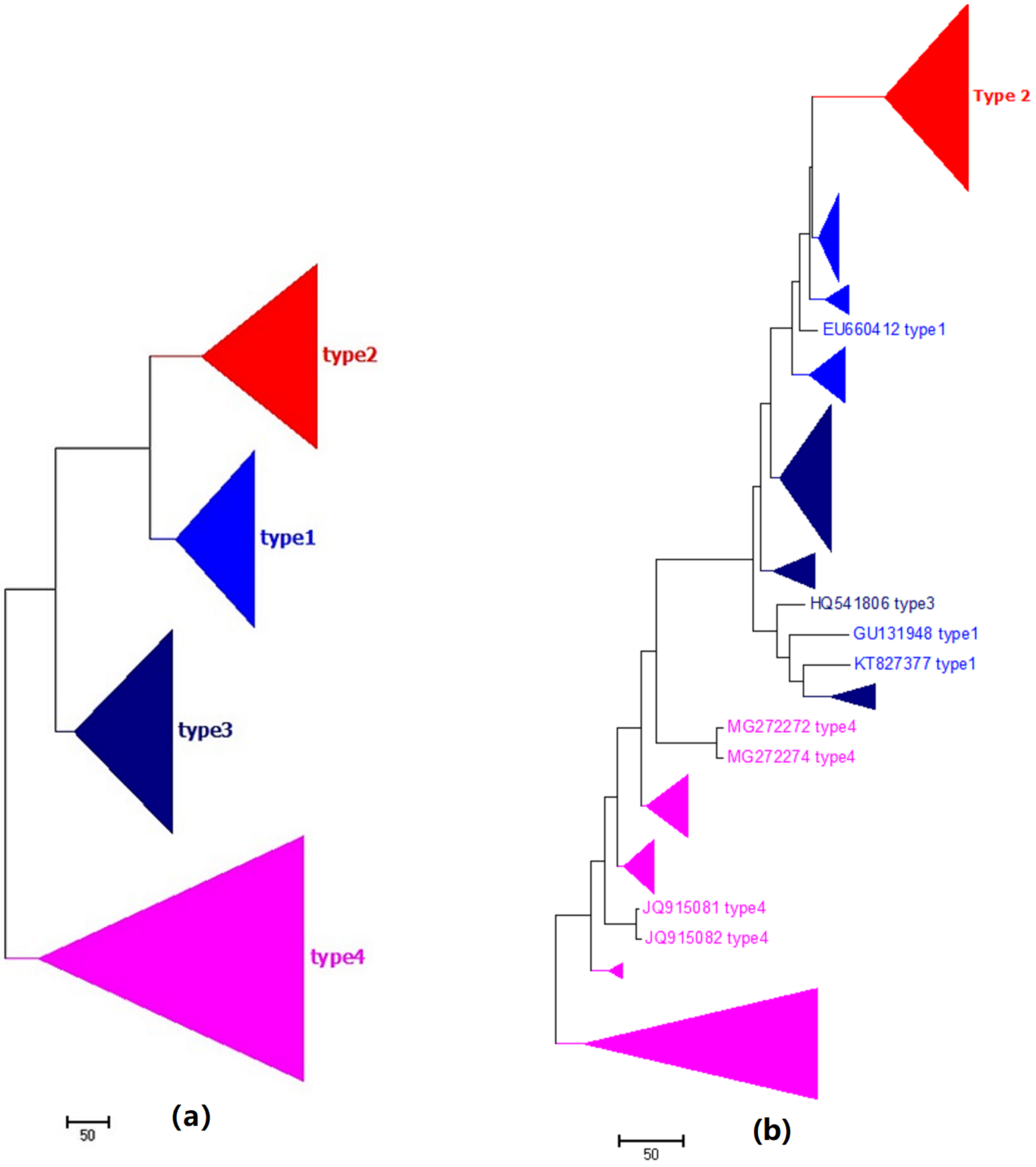
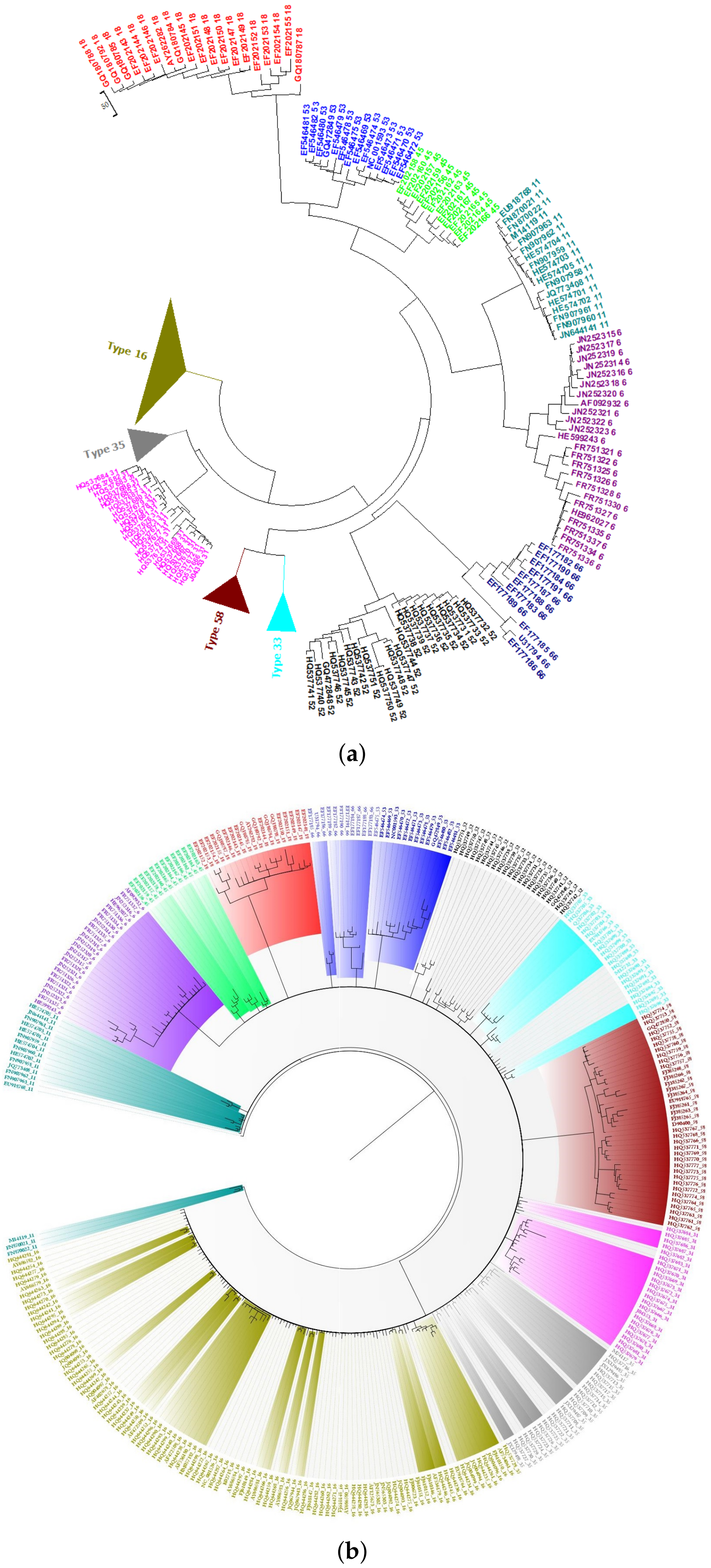


Figure 3. (**a**) The Neighbor-Joining phylogenetic tree of 330 dengue viruses genome sequences based on PCNV method. (**b**) The Neighbor-Joining phylogenetic tree of 330 dengue viruses genome sequences based on NV method.

2.5. Phylogeny of Bacteria

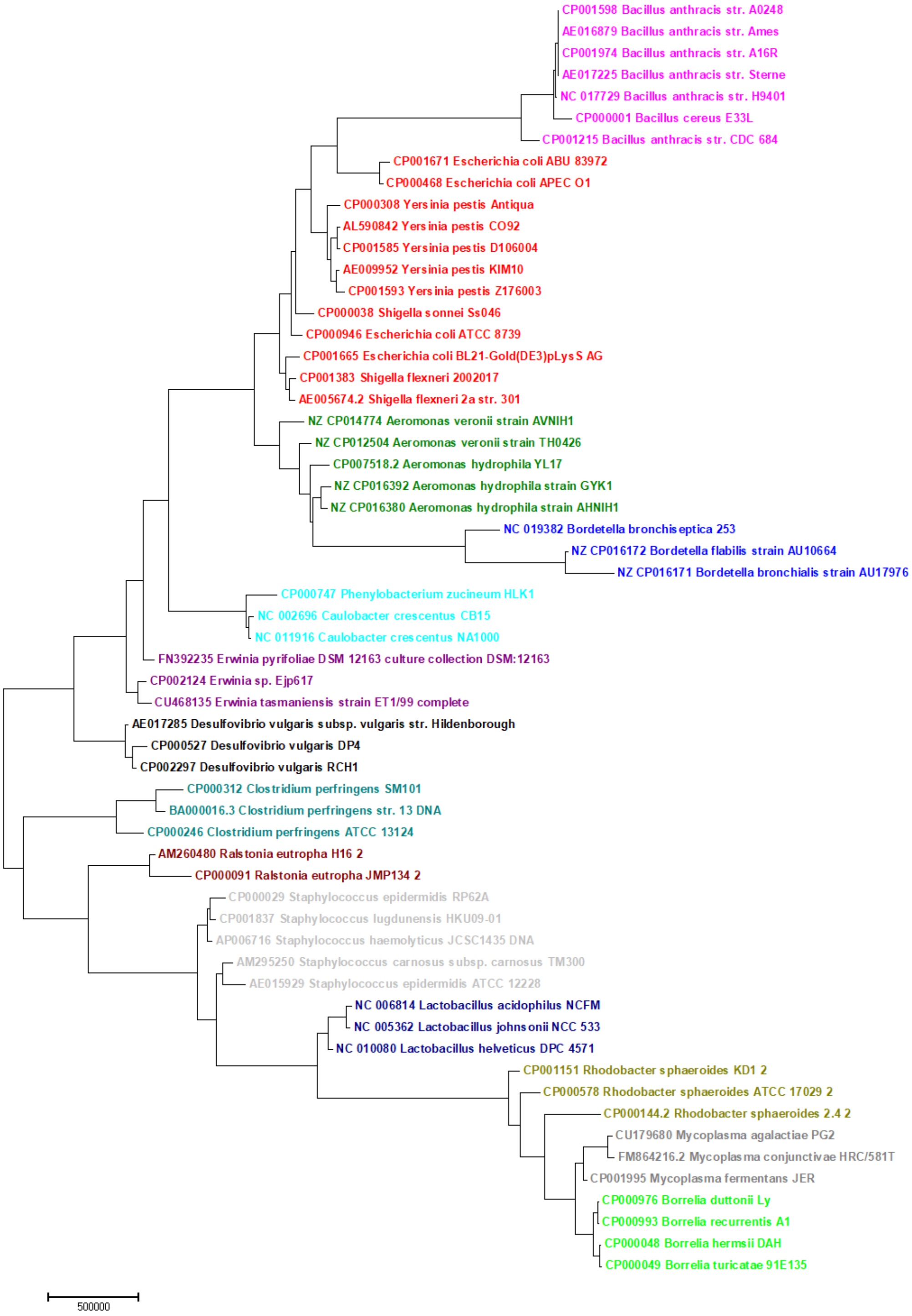
The dataset consisted of 14 families, as shown in Figure 5a,b, of bacterial species with long genomes that ranged from 0.8 to 5 million bp. Using the PCNV method, the phylogenetic tree of these organisms was reconstructed. As shown in Figure 5a, the 59 bacterial species are divided into 14 families that are separated from each other. The 11-mer FFP method mixed these families (Figure 5b). Additionally, the run time for FFP is more than a day, which is far longer than the time required for PCNV. Bayesian inference takes even longer, to the extent that it is not possible to complete the analysis using this method in Muscle on a server equipped with an Intel Xeon E5-2667 v3 Processor and Linux Home Premium with 384 GB RAM (Table 1).



(**a**)

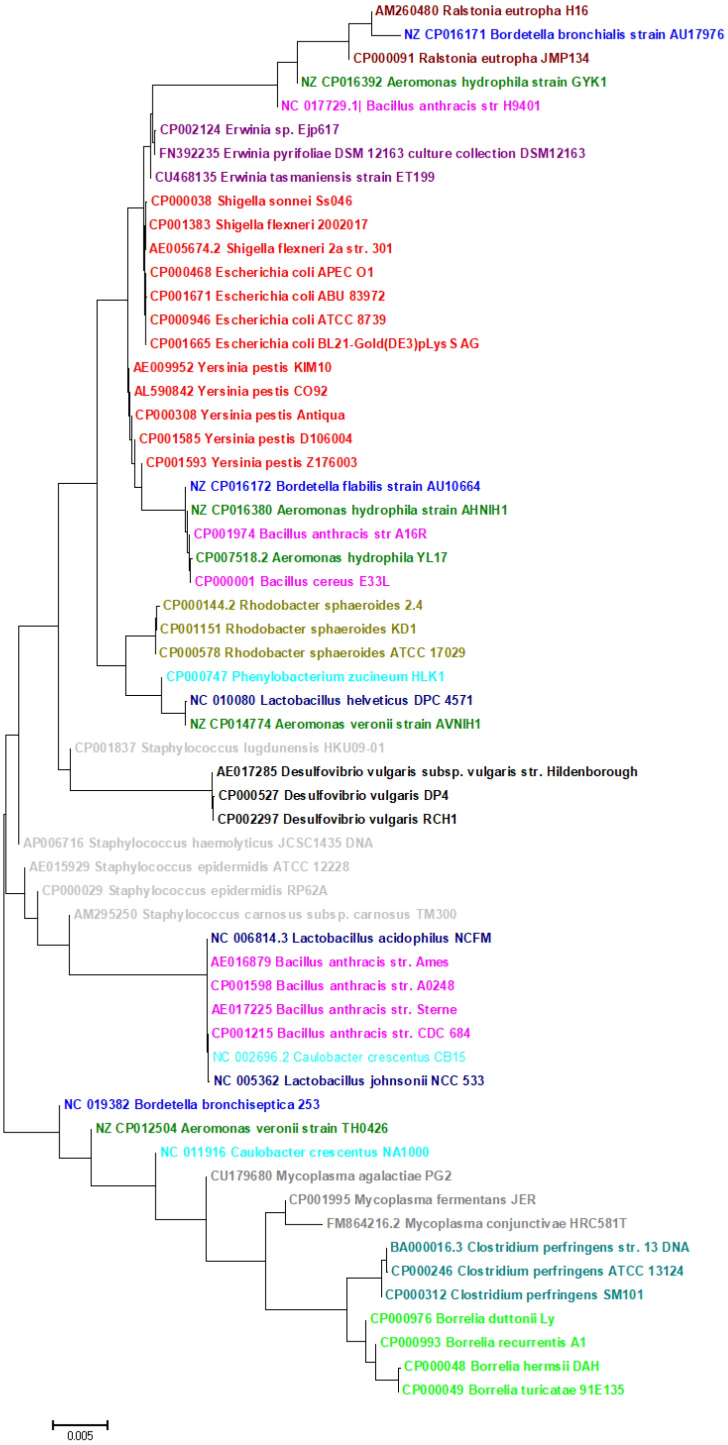
(**b**)

Figure 4. (**a**) The Neighbor-Joining phylogenetic tree of 326 HPV genome sequences based on PCNV method. (**b**) The phylogenetic tree of 326 HPV genome sequences based on Bayesian inference method.


CP001598 Bacillus anthracis str. A0248
AE016879 Bacillus anthracis str. Ames
CP001974 Bacillus anthracis str. A16R
AE017225 Bacillus anthracis str. Sterne
NC 017729 Bacillus anthracis str. H9401
CP000001 Bacillus cereus E33L
CP001215 Bacillus anthracis str. CDC 684
CP001671 Escherichia coli ABU 83972
CP000468 Escherichia coli APEC O1
CP000308 Yersinia pestis Antiqua
AL590842 Yersinia pestis CO92
CP001585 Yersinia pestis D106004
AE009952 Yersinia pestis KIM10
CP001593 Yersinia pestis Z176003
CP000038 Shigella sonnei Ss046
CP000946 Escherichia coli ATCC 8739
CP001665 Escherichia coli BL21-Gold(DE3)pLysS AG
CP001383 Shigella flexneri 2002017
AE005674.2 Shigella flexneri 2a str. 301
NZ CP014774 Aeromonas veronii strain AVNIH1
NZ CP012504 Aeromonas veronii strain TH0426
CP007518.2 Aeromonas hydrophila YL17
NZ CP016392 Aeromonas hydrophila strain GYK1
NZ CP016380 Aeromonas hydrophila strain AHNIH1
NC 019382 Bordetella bronchiseptica 253
NZ CP016172 Bordetella flabilis strain AU10664
NZ CP016171 Bordetella bronchialis strain AU17976
CP000747 Phenylobacterium zucineum HLK1
NC 002696 Caulobacter crescentus CB15
NC 011916 Caulobacter crescentus NA1000
FN392235 Erwinia pyrifoliae DSM 12163 culture collection DSM:12163
CP002124 Erwinia sp. Ejp617
CU468135 Erwinia tasmaniensis strain ET1/99 complete
AE017285 Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
CP000527 Desulfovibrio vulgaris DP4
CP002297 Desulfovibrio vulgaris RCH1
CP000312 Clostridium perfringens SM101
BA000016.3 Clostridium perfringens str. 13 DNA
CP000246 Clostridium perfringens ATCC 13124
AM260480 Ralstonia eutropha H16 2
CP000091 Ralstonia eutropha JMP134 2
CP000029 Staphylococcus epidermidis RP62A
CP001837 Staphylococcus lugdunensis HKU09-01
AP006716 Staphylococcus haemolyticus JCSC1435 DNA
AM295250 Staphylococcus carnosus subsp. carnosus TM300
AE015929 Staphylococcus epidermidis ATCC 12228
NC 006814 Lactobacillus acidophilus NCFM
NC 005362 Lactobacillus johnsonii NCC 533
NC 010080 Lactobacillus helveticus DPC 4571
CP001151 Rhodobacter sphaeroides KD1 2
CP000578 Rhodobacter sphaeroides ATCC 17029 2
CP000144.2 Rhodobacter sphaeroides 2.4 2
CU179680 Mycoplasma agalactiae PG2
FM864216.2 Mycoplasma conjunctivae HRC/581T
CP001995 Mycoplasma fermentans JER
CP000976 Borrelia duttonii Ly
CP000993 Borrelia recurrentis A1
CP000048 Borrelia hermsii DAH
CP000049 Borrelia turicatae 91E135
500000

(**a**)

Figure 5. *Cont.*


AM260480 Ralstonia eutropha H16
NZ CP016171 Bordetella bronchialis strain AU17976
CP000091 Ralstonia eutropha JMP134
NZ CP016392 Aeromonas hydrophila strain GYK1
NC 017729.1| Bacillus anthracis str H9401
CP002124 Erwinia sp. Ejp617
FN392235 Erwinia pyrifoliae DSM 12163 culture collection DSM12163
CU468135 Erwinia tasmaniensis strain ET199
CP000038 Shigella sonnei Ss046
CP001383 Shigella flexneri 2002017
AE005674.2 Shigella flexneri 2a str. 301
CP000468 Escherichia coli APEC O1
CP001671 Escherichia coli ABU 83972
CP000946 Escherichia coli ATCC 8739
CP001665 Escherichia coli BL21-Gold(DE3)pLysS AG
AE009952 Yersinia pestis KIM10
AL590842 Yersinia pestis CO92
CP000308 Yersinia pestis Antiqua
CP001585 Yersinia pestis D106004
CP001593 Yersinia pestis Z176003
NZ CP016172 Bordetella flabilis strain AU10664
NZ CP016380 Aeromonas hydrophila strain AHNIH1
CP001974 Bacillus anthracis str A16R
CP007518.2 Aeromonas hydrophila YL17
CP000001 Bacillus cereus E33L
CP000144.2 Rhodobacter sphaeroides 2.4
CP001151 Rhodobacter sphaeroides KD1
CP000578 Rhodobacter sphaeroides ATCC 17029
CP000747 Phenylobacterium zucineum HLK1
NC 010080 Lactobacillus helveticus DPC 4571
NZ CP014774 Aeromonas veronii strain AVNIH1
CP001837 Staphylococcus lugdunensis HKU09-01
AE017285 Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
CP000527 Desulfovibrio vulgaris DP4
CP002297 Desulfovibrio vulgaris RCH1
AP006716 Staphylococcus haemolyticus JCSC1435 DNA
AE015929 Staphylococcus epidermidis ATCC 12228
CP000029 Staphylococcus epidermidis RP62A
AM295250 Staphylococcus carnosus subsp. carnosus TM300
NC 006814.3 Lactobacillus acidophilus NCFM
AE016879 Bacillus anthracis str. Ames
CP001598 Bacillus anthracis str. A0248
AE017225 Bacillus anthracis str. Sterne
CP001215 Bacillus anthracis str. CDC 684
NC 002696.2 Caulobacter crescentus CB15
NC 005362 Lactobacillus johnsonii NCC 533
NC 019382 Bordetella bronchiseptica 253
NZ CP012504 Aeromonas veronii strain TH0426
NC 011916 Caulobacter crescentus NA1000
CU179680 Mycoplasma agalactiae PG2
CP001995 Mycoplasma fermentans JER
FM864216.2 Mycoplasma conjunctivae HRC581T
BA000016.3 Clostridium perfringens str. 13 DNA
CP000246 Clostridium perfringens ATCC 13124
CP000312 Clostridium perfringens SM101
CP000976 Borrelia duttonii Ly
CP000993 Borrelia recurrentis A1
CP000048 Borrelia hermsii DAH
CP000049 Borrelia turicatae 91E135
0.005

(**b**)

Figure 5. (**a**) The Neighbor-Joining phylogenetic tree of 59 bacteria genome sequences based on PCNV method. (**b**) The Neighbor-Joining phylogenetic tree of 59 bacteria genome sequences based on FFP method (k = 11).

2.6. Classification

Besides evolutionary analysis, PCNV can also be used for classification. Both FFP and Alignment-Free-Kmer-Statistics (AFKS) [16], based on the k-mer approach, can also be used for classification. However, the question of how to choose the value of k is not easily answered. In the present paper, for the FFP method, we set the k value as the minimum integer of $log_4(N)$, i.e., $k = floor(log_4(N))$, where N is the maximum length of the sequences studied [21]. For the AFKS method, we used $k = floor(log_4(\frac{1}{n}\sum_{i \in S} len(i)))$, where n is the number of sequences in the set S [16].

In the PCNV method, after computing the distance matrix using each approach, the one-nearest neighbor (1-NN) [24] method was used for predictions. The sensitivity, specificity, and accuracy of the predictions made using each method are shown in Table 2. It is clear that PCNV is superior to the other two algorithms in this study.

Table 2. Sensitivity (Sens), Specificity (Spec), and Accuracy (Acc) measures of classification are reported for the four virus datasets. For each dataset, the Ave. line displays average values for each measure.

	Nu-	**Sens**	**Sens**	**Sens**	**Spec**	**Spec**	**Spec**	**Acc**	**Acc**	**Acc**	
	Type	**mber**	**PCNV**	**FFP**	**AFKS**	**PCNV**	**FFP**	**AFKS**	**PCNV**	**FFP**	**AFKS**
			(%)	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**	**(%)**
HCV (82)	type1	16	100	62.5	50.0	100	87.9	86.4	100	62.5	50.0
	type2	18	100	55.6	94.4	100	93.8	98.4	100	55.6	94.4
	type3	20	100	80.0	90.0	100	93.5	96.8	100	80.0	90.0
	type4	12	100	50.0	33.3	100	97.1	90.0	100	50.0	33.3
	type5	4	100	50.0	75.0	100	96.2	97.4	100	50.0	75.0
	type6	12	100	50.0	83.3	100	91.4	98.6	100	50.0	83.3
	Ave.		100	58.0	71.0	100	93.3	94.6	100	58.0	71.0
HBV (152)	A	20	100	100	100	100	100	99.2	100	100	100
	B	15	100	100	40.0	100	100	96.4	100	100	40.0
	C	20	100	100	70.0	100	100	96.2	100	100	70.0
	D	13	100	100	76.9	100	100	97.1	100	100	76.9
	E	30	100	100	90.0	100	100	97.5	100	100	90.0
	F	22	100	100	72.7	100	100	93.8	100	100	72.7
	G	17	100	100	94.1	100	100	99.3	100	100	94.1
	H	15	100	100	80.0	100	100	97.1	100	100	80.0
	Ave.		100	100	78.0	100	100	97.1	100	100	78.0
Dengue (330)	type1	72	100	100	76.4	100	100	93.4	100	100	76.4
	type2	75	100	100	73.3	100	100	93.3	100	100	73.3
	type3	83	100	100	78.3	100	100	92.7	100	100	78.3
	type4	100	100	100	87.0	100	100	93.0	100	100	87.0
	Ave.		100	100	78.8	100	100	93.1	100	100	78.8
HPV (326)	6	24	100	100	75.0	100	100	97.7	100	100	75.0
	11	17	100	100	100	100	100	99.7	100	100	100
	16	99	100	100	92.9	100	100	96.5	100	100	92.9
	18	19	100	100	94.7	100	100	100	100	100	94.7
	31	23	100	100	82.6	100	100	99.0	100	100	82.6
	33	22	100	100	86.4	100	100	99.7	100	100	86.4
	35	26	100	100	88.5	100	100	99.0	100	100	88.5
	45	12	100	100	83.3	100	100	99.7	100	100	83.3
	52	22	100	100	81.8	100	100	98.4	100	100	81.8
	53	14	100	100	85.7	100	100	98.7	100	100	85.7
	58	37	100	100	94.6	100	100	99.7	100	100	94.6
	66	11	100	100	90.9	100	100	99.7	100	100	90.9
	Ave.		100	100	88.0	100	100	99.0	100	100	88.0

3. Discussion

In the present paper, we propose a novel 18-dimensional vector method for genome comparison. This PCNV method can be used to successfully define the distribution of nucleotides based on information on the frequency and position of DNA sequences. The correlation of position between two different bases is used in addition to the average position and variance of position of each base. As a result, a high-dimensional DNA genome sequence is reduced to an 18-dimensional numerical vector. Correlations in base distribution play a key role in sequence comparison. Usually, conventional alignment-based methods produce reasonable phylogenetic trees and are therefore widely applied. However, when the dataset volume is large or the sequences analyzed are very long, these methods become ineffective. The phylogenetic analysis results on several distinct datasets show that PCNV can quickly and accurately compare massive datasets of long DNA sequences. We also compared our method with three methods: the popular alignment-based Bayesian inference method, the alignment-free FFP, and Natural Vector methods. The results show that our method can construct more accurate evolutionary relationships among sequences.

To demonstrate the computational advantage of PCNV, we compared our running time constructing phylogenetic trees with FFP, AFKS, Bayesian inference, and Muscle. Compared with the two extensively applied alignment-free methods FFP and AFKS, the running time of PCNV is smallest for all datasets and even takes less than 1 s, except on the bacteria dataset. For bacteria dataset, PCNV is extremely fast and only takes about 53 s, while FFP and AFKS take more than one day. Compared with the two alignment-based methods Bayesian inference and Muscle, PCNV takes much less time for all datasets. For bacteria dataset, Bayesian inference and Muscle cannot obtain phylogeny tree within several days.

With long DNA sequences, particularly bacterial genomes, the Bayesian inference method was much slower than the PCNV approach, sometimes to the extent that it simply did not work, the main reason being that Bayesian inference method is based on alignment method, such as Muscle (Table 1). Even alignment-free methods such as FFP and AFKS were slow in comparison with PCNV, especially when the analyzed sequences were longer, as shown in Table 1.

Furthermore, the MEV was used for comparison. Although both PCNV and MEDV studies aim to solve the problem of genetic sequence comparison, they have significant differences in the features extracted from the sequences. MEV method did not consider the position correlation feature which is an important source of information for genetic sequences. The novelty of our new method is that it designs six suitable features to characterize the correlation among nucleotide position in sequences. The second difference is that our new method does not categorize four types of nucleotides into three groups according to their chemical properties. The third difference is that our new method applies the popular neighbor-joining algorithm to construct phylogenetic trees, while the old study used the UPGMA algorithm which may produce misleading trees.

To show the advantages of our new method (PCNV), MEV and PCNV were compared using all datasets studied used in for the present study. The Neighbor-Joining (NJ) algorithm is used in tree construction of both methods. The NJ trees built by the MEV method are shown in Figures S1–S5. As shown in Figure S1, there are six types of HCV dataset. Using the MEV method, the type 6 marked in pink is divided into two parts. In the NJ tree of the HBV dataset shown in Figure S2, the virus KX765843 belonging to clade C (marked in black) is incorrectly assigned to clade F (marked in red). Similarly, as shown in Figure S3, types 1 (marked in blue) and 3 (marked in navy) of dengue are mixed together. In Figure S4, two viruses from type 35 group (marked in gray) are categorized into other groups.

Horizontal (or lateral) gene transfer (HGT) is a common phenomenon in bacteria. Due to the problem of HGT, Koski and Golding [25] even stated that genes appearing to be the most similar based on BLAST hits are often not the closest relatives each other phylogenetically. It means when there are HGT, if alignment is used, there may be mistakes. For example, given two distantly related bacteria that have exchanged a gene, a phylogenetic tree including those species will show them to be closely related because that gene is the same, even though most other genes are dissimilar. For bacteria, due to the extensively existing HGT, the phylogenetic tree based on alignment may be misleading. To get correct phylogenetic trees of bacteria, one main method is the 16s rRNA-based method, which constructs trees according to the alignment of 16s ribosomal RNA, i.e., 16s rRNA. The 16s rRNA gene tends to be conserved among bacteria with close phylogenetic distances, and thus is often not affected by HGT, but has enough variable differences. However, the method of 16sRNA loses some information in the whole genome. Our PCNV method uses the whole genome sequence and needs no sequence alignment, thus has the potential to be not affected by the HGT and obtain the correct phylogenetic relationship among bacteria.

To show that our method may be not affected by HGT, we used another dataset of eight Yersinia genomes in a previous study [4]. The eight Yersinia genomes are too similar in sequence for classical phylogenetic inference, but share gene segments. The dataset includes two *Yersinia pseudotuberculosis* and six *Yersinia pestis* complete genomes. Using PCNV, we get the neighbor-joining tree of the eight bacteria shown in Figure 6. For the figure, we see that the two *Yersinia pseudotuberculosis* isolates form sisters and are separate from the six *Yersinia pestis* genomes.

Genetic distance is a measure of the genetic divergence between species or between populations within a species, whether the distance measures time from common ancestor or degree of differentiation. Several genetic distances have been proposed based on different evolutionary models. The genetic distance can only be applied to alignment results. A commonly used measure of genetic distance is the fixation index, which varies between 0 and 1. Our PCNA is an alignment-free approach; we measure the distance between species using Euclidean distance and we do not assume any evolutionary model. This distance is positively correlated with their genetic distance, since it can successfully measure the divergence between two species as well. In the neighbor-joining trees constructed by PCNV, sum of the length of the branches traversed from one species to another is equal to their distance and thus positively correlated with their genetic distance. Due to difference in mutation rates for species, for two given datasets, the average internal genetic distance in each dataset may be different enough. In this case, the average distance obtained with PCNV for two datasets may have a big difference, which leads to very different scales in the two derived phylogenetic trees. For the same dataset, different alignment-free method may produce different average distance and thus produce different scales of phylogenetic trees.

Our new method has several limitations, for example we cannot work out the time of evolution. These limitations will need to be studied in the future.

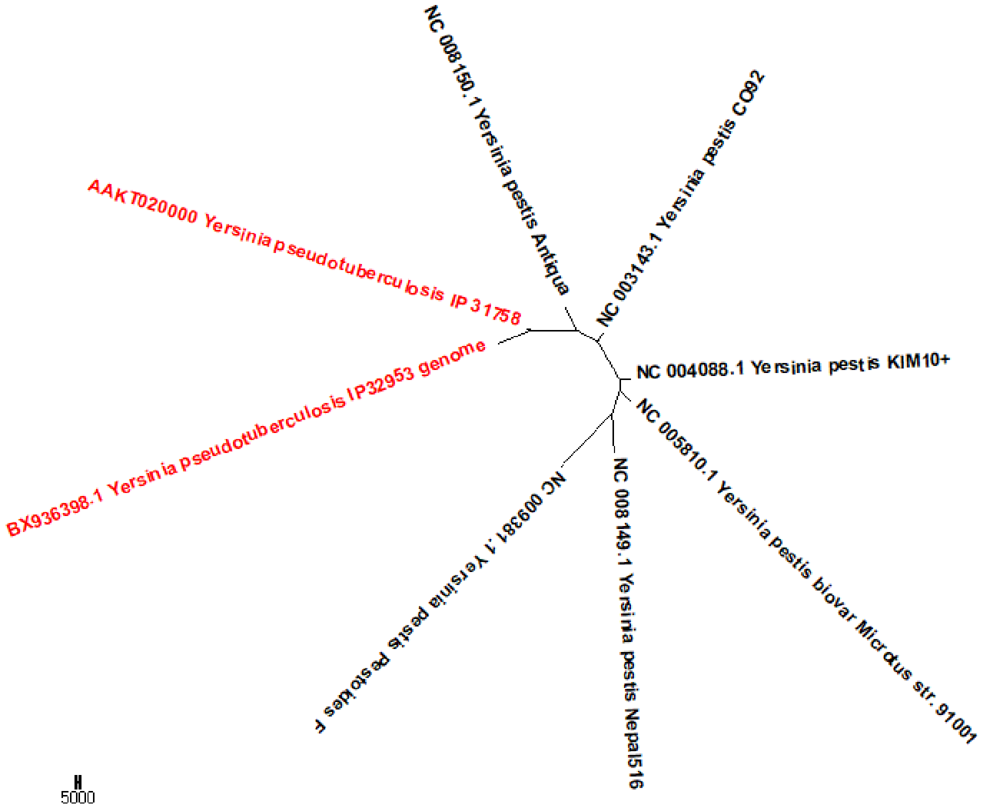

NC 008150.1 Yersinia pestis Antiqua
NC 003143.1 Yersinia pestis CO92
AAKT020000 Yersinia pseudotuberculosis IP 31758
BX936398.1 Yersinia pseudotuberculosis IP32953 genome
NC 004088.1 Yersinia pestis KIM10+
NC 005810.1 Yersinia pestis biovar Microtus str. 91001
NC 008149.1 Yersinia pestis Nepal516
NC 009381.1 Yersinia pestis Pestoides F
5000

Figure 6. The Neighbor-Joining phylogenetic tree of eight Yersinia genomes based on PCNV method.

4. Materials and Methods

4.1. Dataset

Five datasets were used to test and verify the new technique.

4.1.1. HCV

Hepatitis C is a liver infection caused by HCV. The World Health Organization (WHO) estimates that HCV infects 3% of the world's population [26]. Because this virus causes few symptoms, diagnosis is difficult in many cases [27]. In the present study, we acquired 82 complete HCV genomes from the Virus Pathogen Database and Analysis Resource (ViPR) [23]. This dataset has a genomic length of 8957–9666 nucleotides. The NCBI accession numbers are shown in Table S1.

4.1.2. HBV

HBV is a hepatotropic virus that can establish a persistent and chronic infection in humans through immune anergy. It exhibits formidable morbidity and mortality in humans and currently infects 3.5% of the global population [28,29]. HBV is genetically diverse and comprises 10 different genotypes, designated A–J [29,30]. Additional subgenotypes exist within Genotypes A–D and F [31]. The HBV genotypes differ in their geographic distributions. Identifying HBV genotypes quickly and accurately is very important for clinical diagnosis. In the present work, 152 complete HBV genomes including eight genotypes (A–H), were downloaded from the Hepatitis B Virus Database (HBVdb) [30]. The NCBI accession numbers are shown in Table S2.

4.1.3. DENV

DENVs are mosquito-borne aviviruses that have plagued humans for centuries [32]. Statistics show that DENVs infect up to 390 million people worldwide each year; 25% of these people suffer from clinical disease. With four antigenically distinct but immunologically cross-reactive serotypes (DENV-1–4), dengue has one of the most complex transmission processes of all infectious diseases in human populations [33]. Therefore, it is important to distinguish which group particular DENVs belong to. In the present study, 330 dengue viruses were used to demonstrate the effectiveness of the method. The NCBI accession numbers are shown in Table S3.

4.1.4. HPV

HPV is a circular double-stranded DNA virus that causes a variety of proliferative epithelial lesions at specific anatomical sites. It is also the most common sexually transmitted virus. There are many different types of HPV, several of which cause health problems such as genital warts and cancer [34]. For example, the low-risk HPV Types 6 and 11 can cause genital warts or benign cell changes, while the high-risk HPV Types 16 and 18 cause about 70% of cervical cancers [35]. Therefore, identification of HPV genotypes in infected patients is particularly important. In the present work, 326 complete genomes of 12 HPV strands (6, 11, 16, 18, 31,33, 35, 45, 52, 53, 58, and 66) were studied. All viral genomes in this HPV dataset are publicly available at GenBank or NCBI databases. The NCBI accession numbers are shown in Table S4.

4.1.5. Bacteria

There is a biomass of bacteria, the main representatives of the prokaryotes, on Earth. Researchers usually investigate evolutionary relationships among bacteria by building phylogenetic trees. Owing to the large genome size (>1 million bp) of bacteria, it is impossible to reconstruct a bacterial phylogenetic tree in a reasonable amount of time using traditional multiple sequence alignment methods with current computational technology. We used 59 bacterial species to test our method (Table 3). The NCBI accession numbers are shown in Table S5.

Table 3. Summary of the datasets HCV, HBV, Dengue, HPV, and Bacteria. The length distribution of each dataset validates that PCNV can work with long sequences.

Dataset	**Number**	**Min (bp)**	**Median (bp)**	**Mean (bp)**	**Max (bp)**
HCV	82	8957	9442	9427	9666
HBV	152	10161	10669	10606	10780
Dengue	330	10,161	10,669	10,606	10,780
HPV	326	7814	7,905	7895	8051
Bacteria	59	846,214	4,016,947	3,610,938	5,966,919

4.2. *Positional Distribution*

Let $S = s_1, s_2, ..., s_N$ be a DNA sequence of length N. Denote n_α as the number of nucleotides α in the sequence, where $\alpha \in \{A, C, G, T\}$. Here, p_{α_j} is the position of the nucleotide α at the jth appearance,$j = 1, 2, \cdots, n_\alpha$, and $p_{\alpha_0} = 0$. Obviously, $p_{\alpha_j} < p_{\alpha_{j+1}} (0 \le j < n_\alpha)$. For example, given the sequence "ACTGGCAAT", $n_A = 3, p_{A_1} = 1, p_{A_2} = 7, p_{A_3} = 8$. We first define the positional distribution of α $(U_\alpha(i), i = 1, 2, \cdots, N)$ as follows:

$$U_\alpha(i) = \begin{cases} \frac{p_{\alpha_j}}{p_{\alpha_{j+1}} - p_{\alpha_j}}; p_{\alpha_j} \le i < p_{\alpha_{j+1}}; j = 0, 1, \cdots, n_\alpha - 1. \\ \frac{p_{\alpha_{n_\alpha}}}{N - (p_{\alpha_{n_\alpha}} - 1)}; \quad p_{\alpha_{n_\alpha}} \le i \le N. \end{cases} \tag{1}$$

The positional distribution of $\alpha(\alpha \in A, C, G, T)$ is

$$\begin{gathered} \underbrace{0, 0, \cdots, 0}_{p_{\alpha_1}-1}, \underbrace{\frac{p_{\alpha_1}}{p_{\alpha_2} - p_{\alpha_1}}, \frac{p_{\alpha_1}}{p_{\alpha_2} - p_{\alpha_1}}, \cdots, \frac{p_{\alpha_1}}{p_{\alpha_2} - p_{\alpha_1}}}_{p_{\alpha_2}-p_{\alpha_1}}, \cdots, \\ \underbrace{\frac{p_{\alpha_{n_\alpha-1}}}{p_{\alpha_{n_\alpha}} - p_{\alpha_{n_\alpha-1}}}, \frac{p_{\alpha_{n_\alpha-1}}}{p_{\alpha_{n_\alpha}} - p_{\alpha_{n_\alpha-1}}}, \cdots, \frac{p_{\alpha_{n_\alpha-1}}}{p_{\alpha_{n_\alpha}} - p_{\alpha_{n_\alpha-1}}}}_{p_{\alpha_{n_\alpha}}-p_{\alpha_{n_\alpha-1}}}, \\ \underbrace{\frac{p_{\alpha_{n_\alpha}}}{N - (p_{\alpha_{n_\alpha}} - 1)}, \frac{p_{\alpha_{n_\alpha}}}{N - (p_{\alpha_{n_\alpha}} - 1)}, \cdots, \frac{p_{\alpha_{n_\alpha}}}{N - (p_{\alpha_{n_\alpha}} - 1)}}_{N-(p_{\alpha_{n_\alpha}}-1)} \end{gathered} \tag{2}$$

Example: Given the sequence "ACTGGCAAT", the positional distribution is as shown in Table 4. Here, we take $\alpha = C$ as an example to show the details of the calculation process. According to the above definition: $N = 9, n_C = 2; P_{C_0} = 0, P_{C_1} = 2, P_{C_2} = 6$.

(1) When $p_{C_0} \le i < p_{C_1}$, namely, $i = 1$, so $U_C(1) = \frac{p_{C_0}}{p_{C_1} - p_{C_0}} = 0$.

(2) When $p_{C_1} \le i < p_{C_2}$, namely, $i \in \{2, 3, 4, 5\}$, thus $U_C(2) = U_C(3) = U_C(4) = U_C(5) = \frac{p_{C_1}}{p_{C_2} - p_{C_1}} = \frac{2}{6-2} = \frac{2}{4}$.

(3) When $p_{C_2} \le i \le N$, namely, $i \in \{6, 7, 8, 9\}$, $U_C(6) = U_C(7) = U_C(8) = U_C(9) = \frac{p_{C_2}}{N - (p_{C_2} - 1)} = \frac{6}{9-(6-1)} = \frac{6}{4}$.

Similarly, we can get U_A, U_G, U_T.

Table 4. The positional distribution of "ACTGGCAAT".

Sequence **Position**	**A** **1**	**C** **2**	**T** **3**	**G** **4**	**G** **5**	**C** **6**	**A** **7**	**A** **8**	**T** **9**
$U_A(i)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{7}{1}$	$\frac{8}{2}$	$\frac{8}{2}$
$U_C(i)$	0	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{6}{4}$	$\frac{6}{4}$	$\frac{6}{4}$	$\frac{6}{4}$
$U_T(i)$	0	0	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{9}{1}$
$U_G(i)$	0	0	0	$\frac{4}{1}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$	$\frac{5}{5}$

4.3. Position Correlation Vector

4.3.1. Average Positional Distribution

The average positional distribution κ_α is defined as ($\alpha, \beta \in \{A, C, G, T\}$):

$$\kappa_\alpha = \frac{\sum_{i=1}^{N} U_\alpha(i)}{n_\alpha}, \quad \alpha \in \{A, C, G, T\}. \tag{3}$$

Therefore,

$$\kappa_\alpha = \frac{\sum_{i=1}^{N} U_\alpha(i)}{n_\alpha} = \frac{\sum_{j=0}^{n_\alpha - 1} \frac{p_{\alpha_i}}{p_{\alpha_{j+1}} - p_{\alpha_j}} \times (p_{\alpha_{j+1}} - p_{\alpha_j})}{n_\alpha} \\ + \frac{\frac{p n_\alpha}{N - (p n_\alpha - 1)} \times [N - (p_{n_\alpha} - 1)]}{n_\alpha} = \frac{\sum_{j=0}^{n_\alpha} p_{\alpha_j}}{n_\alpha}. \tag{4}$$

For the example sequence "ACTGGCAAT" above, $\kappa_A = \frac{\sum_{i=1}^{N} U_A(i)}{n_A} = \frac{\frac{1}{6} \times 6 + \frac{7}{1} + \frac{8}{2} \times 2}{3} = \frac{\frac{\sum_{j=0}^{n_A} p_{A_j}}{n_A}}{3} = \frac{1+7+8}{3} = \frac{16}{3}$, and likewise for C, G, and T.

4.3.2. Positional Covariance

The positional covariance of nucleotide α and β ($cov(\alpha, \beta)$) are defined as follows:

$$cov(\alpha, \beta) = \sum_{i=1}^{N} \frac{(U_\alpha(i) - \bar{U}_\alpha)(U_\beta(i) - \bar{U}_\beta)}{n_\alpha \cdot n_\beta}. \tag{5}$$

where

$$\bar{U}_\alpha = \frac{1}{N} \sum_{i=1}^{N} U_\alpha(i) = \frac{1}{N} \sum_{j=0}^{n_\alpha} p_{\alpha_j}. \tag{6}$$

For the sequence "ACTGGCAAT", the positional distributions of A and C are $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, 7, \frac{8}{2}, \frac{8}{2})$ and $(0, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{2}{4}, \frac{6}{4}, \frac{6}{4}, \frac{6}{4}, \frac{6}{4})$, respectively.

$\bar{U}_A = \frac{1}{N} \sum_{i=1}^{N} U_A(i) = \frac{\sum_{j=0}^{n_A} p_A}{N} = \frac{\frac{1}{6} \times 6 + \frac{7}{1} + \frac{8}{2} \times 2}{9} = \frac{16}{9}$.

$\bar{U}_C = \frac{1}{N} \sum_{i=1}^{N} U_C(i) = \frac{\sum_{j=0}^{n_C} p_C}{N} = \frac{0 + \frac{2}{4} \times 4 + \frac{6}{4} \times 4}{9} = \frac{8}{9}$.

$cov(A, C) = \sum_{i=1}^{N} \frac{(U_A(i) - \bar{U}_A)(U_C(i) - \bar{U}_C)}{n_A \cdot n_C} = \frac{1}{3*2}[(\frac{1}{6} - \frac{16}{9})(0 - \frac{8}{9}) + (\frac{1}{6} - \frac{16}{9})(\frac{2}{4} - \frac{8}{9}) \times 4 + (\frac{1}{6} - \frac{16}{9})(\frac{6}{4} - \frac{8}{9}) + (\frac{7}{1} - \frac{16}{9})(\frac{6}{4} - \frac{8}{9}) + (\frac{8}{2} - \frac{16}{9})(\frac{6}{4} - \frac{8}{9}) \times 2] = 1.4769$.

The same method is used to compute cov (A;G); cov (A; T); cov (C;G); cov (C; T); and cov (G; T).

4.3.3. Positional variance

The positional variance D_α is described as:

$$D_2^\alpha = cov(\alpha,\alpha) = \sum_{i=1}^{N} \frac{(U_\alpha(i) - \bar{U}_\alpha)(U_\alpha(i) - \bar{U}_\alpha)}{n_\alpha \cdot n_\alpha}. \tag{7}$$

Consequently, an 18-dimensional PCNV of a DNA sequence was constructed as follows:

$$(n_A, n_C, n_G, n_T, \kappa_A, \kappa_C, \kappa_G, \kappa_T, D_2^A, D_2^C, D_2^G, D_2^T, cov(A,C), cov(A,G), cov(A,T), cov(C,G),$$

$$cov(C,T), cov(G,T)).$$

Then, we used the Euclidean distance to compute the pairwise distance among the 18-dimensional vectors of the genome sequences. A phylogenetic tree can be built using the NJ algorithm in MEGA 7.0 software [36].

Additionally, the PCNV method can be used to classify organisms. Naturally, sensitivity, specificity, and accuracy were used to evaluate classification performance. The definitions of these measures are as follows:

$$Sensitivity = TP/(TP + FN). \tag{8}$$

and

$$Specificity = TN/(FP + TN). \tag{9}$$

where TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative predictions, respectively. The MATLAB source code in this paper is freely available to the public upon request.

Supplementary Materials: The information for all five datasets is shown in the Supplementary_File. The following are available at http://www.mdpi.com/1422-0067/21/11/3859/s1.

Author Contributions: S.S.-T.Y. and L.H. conceived the initial planning; L.H. implemented the code; and R.D. and R.L.H. analyzed the results. All authors wrote and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (91746119), Tsinghua University Education Foundation fund (042202008), and Tsinghua University start-up fund.

Acknowledgments: This study was supported by the National Natural Science Foundation of China (91746119), Tsinghua University Education Foundation fund (042202008), and Tsinghua University start-up fund. Stephen S.-T. Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done.

Conflicts of Interest: The authors declare no competing financial interests.

Abbreviations

The following abbreviations are used in this manuscript:

AFKS	Alignment-Free-Kmer-Statistics
DENV	Dengue virus
HCV	Hepatitis C virus
HBV	Hepatitis B virus
HMM	hidden Markov model
HPV	Human papillomavirus
PCNV	Positional Correlation Natural Vector
NV	Natural Vector
ViPR	Virus Pathogen Database and Analysis Resource
FFP	Feature Frequency Profiles

References

1. Fredrik, R.; Maxim, T.; van der Mark, P.; Ayres, D.L., Darling, A.; Höhna, S.; Liu, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542.
2. Bernard, G.; Chan, C.X.; Chan, Y.-B.; Chua, X.-Y.; Cong, Y.; Hogan, J.M. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Brief. Bioinform.* **2017**, *2*, 2.
3. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. AAlignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **2017**, *1*, 186. [CrossRef]
4. Zielezinsk,i A.; Girgis, H.Z.; Bernard, G. Benchmarking of alignment-free sequence comparison methods. *Genome Biol.* **2019**, *20*. [CrossRef]
5. Vinga, S. Information theory applications for biological sequence analysis. *Genome Biol.* **2014**, *15*, 376–89. [CrossRef]
6. Almeida, J.S. Sequence analysis by iterated maps, a review. *Brief. Bioinform.* **2014**, *15*, 369–75. [CrossRef]
7. Yin, C.; Yau, S.S.-T. An improved model for whole genome phylogenetic analysis by Fourier transform. *J. Theor. Biol.* **2015**, *382*, 99–110. [CrossRef]
8. Jeffrey, H.J. Chaos game representation of gene structure. *NAR* **1990**, *18*, 2163-70.
9. Deng, M.; Yu, C.L.; Liang, Q.; He, R.L.; Yau, S.S.-T. A novel method of characterizing genetic sequences: Genome space with biological distance and applications. *PLoS ONE* **2011**, *31*, e17293. [CrossRef]
10. Li, Y.; He, L.; Lucy, H.R.; Yau, S.S.-T. A novel fast vector method for genetic sequence comparison. *Sci. Rep.* **2017**, *7*, 12226. [CrossRef]
11. Ren, J.; Bai, X.; Lu, Y.Y.; Tang, K.; Wang, Y.; Reinert, G. Alignment-free sequence analysis and applications. *Annu. Rev. Biomed. Data Sci.* **2018**, *1*, 93–114. [CrossRef] [PubMed]
12. Wu, G.H.A. J.; Juna, S.-R.; Simsa, G.E.; S.; Kim, S.-H. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method.*PNAS* **2009**, *106*, 12826–12831. [CrossRef] [PubMed]
13. Dai, Q.; Yang, Y.C.; Wang, T.M. Markov model plus k-word distributions: A synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **2012**, *20*, 2296–2302. [CrossRef]
14. Xu, Z; Hao, B.L. CVTree update: A newly designed phylogenetic study platform using composition vectors and whole genomes. *NAR* **2009**, *37* (Suppl. 2), W174–W178. [CrossRef]
15. Hatje, K.; Kollmar, M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front. Plant Sci.* **2012**, *3*, 192. [CrossRef]
16. Luczak, B.B.; James, B.T.; Girgis, H.Z. A survey and evaluations of histogram-based statistics in alignment-ree sequence comparison. *Brief. Bioinform.* **2017**, *20*, 1222–1237. [CrossRef]
17. Leimeister, C.-A.; Boden, M.; Horwege, S.; Lindner, S.; Morgenstern, B. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics* **2014**, *30*, 1991–1999. [CrossRef]
18. Leimeister, C.-A.; Sohrabi-Jahromi, S.; Morgenstern, B. Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics* **2017**, *33*, 971–9. [CrossRef]
19. Murray, K.D.; Webers, C.; Ong, C.S.; Borevitz, J.; Warthmann, N. kWIP: The k-mer weighted inner product, a de novo estimator of genetic similarity. *PLoS Comput. Biol.* **2017**, *13*, e1005727. [CrossRef]
20. Blaisdell, B.E. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. *J. Mol. Evol.* **1989**, *29*, 538–547. [CrossRef]
21. Sims, G.E.; Jun, S.R.; Wu, G.A.; Kim, S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *PNAS* **2009**, *106*, 2677–2682. [CrossRef]
22. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genom.* **2009**, *10*, 402–415. [CrossRef]
23. Pickett, B.E.; Greer, D.S.; Zhang, Y.; Stewart, L.; Zhou, L.; Sun, G.Y.; Gu, Z.P.; Kumar, S.; Zaremba, S.; Larsen, C.N.; et al. Virus pathogen database and analy-sis resource (vipr): A comprehensive bioinformatics database and analysis resource for the coronavirus research community. *Viruses* **2012**, *4*, 3209–3226. [CrossRef]
24. Kramer, O. *K-Nearest Neighbors. Dimensionality Reduction with Unsupervised Nearest Neighbors*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 13–23.
25. Koski, L.B; Golding, G.B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **2001**, *52*, 540–542. [CrossRef] [PubMed]

26. Echeverría, N.; Moratorio, G.; Cristina, J.; Moreno, P. Hepatitis C virus genetic variability and evolution. *World J. Hepatol.* **2015**, *7*, 831–845. [CrossRef] [PubMed]
27. Campo, D.S.; Dimitrova, Z.; Mitchell, R.J.; Lara, J.; Khudyakov, Y. Coordinated evolution of the hepatitis C virus. *PNAS* **2008**, *105*, 9685–9690. [CrossRef] [PubMed]
28. Yuen, M.-F.; Chen, D.-S.; Dusheiko, G.M.; Janssen, H.L.A.; Lau, D.T.Y.; Locarnini, S.A.; Peters, M.G.; Lai, C.-L. Hepatitis B virus infection. *Nat. Rev. Dis.Prim.* **2018**, *4*, 18035. [CrossRef] [PubMed]
29. Patterson, R.Z.; Klunk, J.; Fornaciari, G.; Giuffra, V.; Duchěne, S.; Duggan, A.T.; Poinar, D.; Douglas, M.W.; Eden, J.-S.; Holmes, E.C.; et al. Correction: The paradox of hbv evolution as revealed from a 16th century mummy. *PLoS Pathog.* **2012**, *14*, e1006887-. [CrossRef]
30. Hayer, J.; Jadeau, F.; Deléage, G.; Kay, A.; Zoulim, F.; Combet, C. HBVdb: A knowledge database for Hepatitis B Virus. *NAR* **2013**, *41*, D566–D570. [CrossRef]
31. Norder, H.; Hammas, B.; Lee, S.D.; Bile, K.; Courouc, A.M.; Mushahwar, I.K.; Magnius, L.O. Genetic relatedness of hepatitis B viral strains of diverse geographical origin and natural variations in the primary structure of the surface antigen.*J. Gen. Virol.* **1993**, *74*, 1341–1348. [CrossRef]
32. Sirisena, P.D.N.N.; Noordeen, F. Evolution of dengue in Sri Lanka–changes in the virus, vector, and climate. *Int. J. Infect. Dis.* **2014**, *19*, 6–12. [CrossRef] [PubMed]
33. Tsang, T.K.; Ghebremariam, S.L.; Gresh, L.; Gordon, A.; Halloran, M.E.; Katzelnick, L.C.; Rojas, D.P.; Kuan, G.; Balmaseda, A.; Sugimoto, J.; et al. Effects of infection history on dengue virus infection and pathogenicity. *Nat. Commun.* **2019**, *10*, 1246. [CrossRef]
34. Eom, J.H.; Park, S.B.; Zhang, B.T. Genetic Mining of DNA Sequence Structures for Effective Classification of the Risk Types of Human Papillomavirus (HPV). *Lect. Notes Comput. Sci.* **2004**, *3316*, 1334–1343.
35. Smith, J.S.; Lindsay, L.; Hoots, B.; Keys, J.; Franceschi, S.; Winer, R.; Clifford, G.M. Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: A meta-analysis update. *Int. J. Cancer* **2007**, *121*, 621–632. [CrossRef] [PubMed]
36. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [CrossRef]