

Article

Phylogenetic Analysis of HIV-1 Genomes Based on the Position-Weighted K-mers Method

Yuanlin Ma ^{1,2}, Zuguo Yu ^{1,3,*}, Runbin Tang ¹, Xianhua Xie ⁴, Guosheng Han ¹ and Vo V. Anh ^{1,5}

¹ Hunan Key Laboratory for Computation and Simulation in Science and Engineering and Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan 411105, China; 201590110068@smail.xtu.edu.cn (Y.M); 201831510085@smail.xtu.edu.cn (R.T); hangsg@xtu.edu.cn (G.H); vanh@swin.edu.au (V.V.A.)

² School of Economics, Zhengzhou University of Aeronautics, Zhengzhou 450046, China

³ School of Electrical Engineering and Computer Science, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia

⁴ School of Mathematical and Computer Science, Gannan Normal University, Ganzhou 341000, China; xiexianhua@gnnu.edu.cn (X.X);

⁵ Faculty of Science, Engineering and Technology, Swinburne University of Technology, P.O. Box 218, Hawthorn, Victoria 3122, Australia

* Correspondence: yuzg@xtu.edu.cn

Received: 17 January 2020; Accepted: 20 February 2020; Published: 23 February 2020



Abstract: HIV-1 viruses, which are predominant in the family of HIV viruses, have strong pathogenicity and infectivity. They can evolve into many different variants in a very short time. In this study, we propose a new and effective alignment-free method for the phylogenetic analysis of HIV-1 viruses using complete genome sequences. Our method combines the position distribution information and the counts of the k -mers together. We also propose a metric to determine the optimal k value. We name our method the *Position-Weighted k -mers (PWkmer)* method. Validation and comparison with the Robinson–Foulds distance method and the modified bootstrap method on a benchmark dataset show that our method is reliable for the phylogenetic analysis of HIV-1 viruses. *PWkmer* can resolve within-group variations for different known subtypes of Group M of HIV-1 viruses. This method is simple and computationally fast for whole genome phylogenetic analysis.

Keywords: Alignment-free; HIV-1 virus; phylogenetic analysis; position-weighted k -mers; Robinson–Foulds distance

1. Introduction

Human Immunodeficiency Viruses (HIVs) are retroviruses which are the causative agents of the global pandemic of Acquired Immunodeficiency Syndrome (AIDS) [1]. There are two types of HIVs: Type 1 (HIV-1 viruses) and Type 2 (HIV-2 viruses). HIV-1 viruses are known to originate from the Simian Immunodeficiency Viruses (SIVs) found in central and eastern African chimpanzees, which form the most common pathogenic strain of HIV viruses and have a high mortality rate [2]. Usually, HIV-1 viruses are divided into a major group (Group M) and two or more minor groups, namely Groups N, O, and possibly Group P. Group M is further divided into subtypes A, B, C, D, E, F, J, K. The subtypes A and F are further divided into sub-subtypes (A1, A2) and (F1, F2) based on differential phylogenetic clustering, respectively. Two or more HIV-1 subtypes can recombine and form Circulating Recombinant Forms (CRFs) [3]. Obviously, classification of HIV-1 strains into subtypes, sub-subtypes, and CRFs is a complex issue, which leads to major problems in the development of vaccines against HIV-1. These problems include high genetic variation, the fast evolution of different variants, and sequence diversity. The first task to solve these problems is how

to obtain the phylogenetic relationships of HIV-1 genomes quickly and accurately. Traditional HIV-1 phylogenetic analysis methods are based on multiple sequence alignment. Although alignment-based methods generally yield excellent results when the sequences are closely related and can be reliably aligned, there are two limitations. Firstly, they lead to conflicting results by using different genes or genome fragments. Secondly, alignment-based methods are generally time-consuming and have high computational complexity when they are directly applied to whole-genome comparisons and phylogenetic studies [4]. Therefore, several alignment-free methods have been developed to overcome the critical limitations of alignment [5–14]. In particular, several alignment-free methods for HIV genome comparison have been developed in the past few decades. For example, Wu et al. [5] used the complete composition vector representation proposed by Hao and Qi [15] for the phylogenetic analysis of HIV-1 genomes, and obtained some acceptable results. Pandit et al. [16] used multifractal measures to capture the genomic variation in the different retroviral species. However, this multifractal method cannot resolve the subtle variations in the subtypes of Group M of HIV-1 viruses. The first usage of k -mers (substring of length k) counts for biological sequence comparison was implemented by Blaisdell [17]. Subsequently, a lot of alignment-free methods using k -mers emerged. Yang and Wang [7] proposed a novel statistical measure for sequence comparison on the basis of k -mers counts, which removes the influence of the length of sequences, and obtained some acceptable results for the phylogenetic analysis of HIV-1 genomes. Chang et al. [8] proposed a cumulative Markov mutual information (CMMI) method which was derived from several k -mers distributions in different genome sequences, and reported some computational results on the HIV-1 subtyping. These results are slightly different from those reported in the NCBI (National Center for Biotechnology Information Search database). In addition, there are other alignment-free methods that may also be used for HIV-1 genome comparisons, such as the gene content-based method [18], the data compression method [19], the fractal method [20], the CVTree method [21], the inter-amino-acid distance method [10], the higher-order Markov model [11], the dynamical language model [6,12], a method using spaced-word frequencies [9], and a method based on the distribution of k -mer intervals [22]. All these alignment-free methods for comparing biological sequences are intended to extract hidden information from the whole genomes, but from different angles.

In this study, we present a new alignment-free method based on position-weighted k -mers to capture the subtle variations from the complete genome sequences of HIV-1 viruses. In our method, the effects of k -mers counts and k -mers position distributions are combined to capture more evolutionary information. On the basis of the proposed method, we report and discuss the results on the HIV-1 subtyping. More importantly, the resulting phylogenetic tree of 44 HIV genome sequences is quite consistent with the accepted taxonomy from NCBI. Our results show that the new method works as well as the conventional alignment-based phylogenetic methods and other alignment-free methods, but is simpler and requires much less computational time and resources. Moreover, our approach can be applied to study the subtype clustering and phylogenetic relationships of a large volume of genome sequences. The source codes of our method can be downloaded from <https://github.com/myl446/HivStudy>. The detailed information please see the Supplementary Material.

2. Materials and Methods

2.1. Complete Genome Datasets

Twenty of the 21 genomes used in Chang et al. [8] are included in the 43 genomes used in Wu et al. [5]. For the phylogenetic analysis of HIV-1 complete genomes, we used a dataset which is composed of 44 HIV complete genomes (43 HIV complete genomic sequences used in the literature [5] and a misplaced sequence of the article categorization [8]). This dataset includes the subtypes A, B, C, D, F, G, J, K, H of the HIV-1 Groups M, O, and N, and a CPZ sequence. All of these sequences can be downloaded from the Los Alamos National Laboratory HIV Sequence Database (<http://www.hiv.lanl.gov/>). Specific accession, subtype, length (bp), and area are listed in Table 1.

Many studies suggested that all of the translated protein amino acid sequences from the genome is a better choice than whole genome DNA sequences and coding parts of complete genomes for genome-based phylogeny reconstruction [6,12,21,23]. However, after computational comparisons and theoretical analysis, we found that our present method is only suitable for whole genome DNA sequences.

Table 1. Labels of complete genome builds used for 44 HIV-1 genomes of the dataset.

No.	Accession	Subtype	Length (bp)	Area
1	U51190	A1	8999	Uganda
2	AF004885	A1	9160	Kenya
3	AF069670	A1	8813	Somalia
4	AF484509	A1	8807	Uganda
5	AF286237	A2	9060	Cyprus
6	AF286238	A2	8972	DRC
7	AY173951	B	8996	Thailand
8	AY331295	B	8834	USA
9	AY423387	B	9359	Netherlands
10	K03455	B	9719	France
11	AF146728	B	8887	Australia
12	AF067155	C	9002	India
13	AY772699	C	9011	South Africa
14	U46016	C	9031	Ethiopia
15	U52953	C	8959	Brazil
16	AY371157	D	8379	Cameroon
17	K03454	D	9176	DRC
18	U88824	D	8952	Uganda
19	AF005494	F1	8968	Brazil
20	AF075703	F1	8925	Finland
21	AF077336	F1	8903	Belgium (DRC)
22	AJ249238	F1	8614	France
23	AF377956	F2	8782	Cameroon
24	AJ249236	F2	8555	Cameroon
25	AJ249237	F2	8589	Cameroon
26	AY371158	F2	8349	Cameroon
27	AF061641	G	9047	Finland(Kenya)
28	AF061642	G	9074	Sweden (DRC)
29	AF084936	G	9707	Belgium (DRC)
30	AF005496	H	8953	Cent.Afr. Rep
31	AF190127	H	9056	Belgium
32	AF190128	H	9707	Belgium
33	AF082394	J	8943	Sweden
34	AF082395	J	8953	Sweden
35	AJ249235	K	8600	DRC
36	AJ249239	K	8604	Cameroon
37	AJ006022	N	9182	Cameroon
38	AJ271370	N	9045	Cameroon
39	AY532635	N	8938	Cameroon
40	AJ302647	O	9829	Senegal
41	AY169812	O	9110	Cameroon
42	L20571	O	9793	Cameroon
43	L20587	O	9754	Cameroon
44	AF447763	CPZ	9326	Tanzania

DRC: Democratic republic of Congo

2.2. The Measure of Position-Weighted K-mers

Assume that $s_1s_2 \dots s_k$ is a k -mer, where $s_i \in \{A, T, C, G\}$. If the k -mer $s_1s_2 \dots s_k$ occurs in a given nucleic acid sequence X , then we denote by $P_{s_1s_2 \dots s_k}$ the vector composed of the positions of

$s_1s_2 \dots s_k$ in X and by $P_{s_1s_2 \dots s_k}(i)$ its i th element. If $s_1s_2 \dots s_k$ does not exist in X , $P_{s_1s_2 \dots s_k}$ is a zero vector. For example, we consider the 2-mers position vectors for the following short nucleic acid sequence of length 20: $X = \text{TAAGCCGCATTAGCTGGTTT}$. We get $P_{AA} = (2), P_{GA} = (0), P_{GC} = (4, 7, 13) \dots$. These k -mers position vectors can effectively capture the distribution information of each k -mer in the given sequence. For a fixed k , we can reverse this sequence by some k -mers position vector. Furthermore, if a k -mer exists in the given sequence, the counts of this k -mer in the nucleic acid sequence are equal to the length of its corresponding position vector. Therefore, we can use the following 2-mers position vector to reconstruct the nucleic acid sequence used in this example:

$$\begin{aligned} P_{AA} &= (2), P_{AC} = (0), P_{AG} = (3, 12), P_{AT} = (9), P_{CA} = (0), P_{CC} = (5), \\ P_{CG} &= (6), P_{CT} = (14), P_{GA} = (0), P_{GC} = (4, 7, 13), P_{GG} = (16), \\ P_{GT} &= (17), P_{TA} = (1, 11), P_{TC} = (0), P_{TG} = (15), P_{TT} = (10, 18, 19). \end{aligned}$$

The 2-mers AC, CA, GA , and TC do not appear in this example. Now, we reverse the given nucleic acid sequence as follows:

$$\begin{array}{ll} P_{TA} = (1, 11) & TA \dots TA \dots \\ P_{AA} = (2) & TAA \dots TA \dots \\ P_{AG} = (3, 12) & TAAG \dots TA.AG \dots \\ P_{GC} = (4, 7, 13) & TAAGC.GCTA.AGC \dots \\ P_{CG} = (6) & TAAGCGGCTA.AGC \dots \\ P_{TA} = (1, 11) & TAAGCGGCTATAGC \dots \\ P_{TG} = (15) & TAAGCGGCTATAGCTG \dots \\ P_{GT} = (17) & TAAGCGGCTATAGCTGGT \dots \\ P_{TT} = (10, 18, 19) & TAAGCGGCTATAGCTGGTTT. \end{array}$$

Suppose

$$P_{s_1s_2 \dots s_k} = (p_1, p_2, \dots, p_m),$$

where m is the count of $s_1s_2 \dots s_k$ in the given nucleic acid sequence. The measure of $s_1s_2 \dots s_k$ based on its position in the sequence, denoted $f(s_1s_2 \dots s_k)$, is defined as

$$f(s_1s_2 \dots s_k) = \begin{cases} \frac{(\frac{p_1}{L} + \frac{p_2}{L} + \dots + \frac{p_m}{L})}{L-k+1}, & m \neq 0, \\ 0, & m = 0, \end{cases}$$

where L is the length of the given sequence.

After simplifying, the following form is obtained:

$$f(s_1s_2 \dots s_k) = \begin{cases} \frac{\sum_{i=1}^m p_i}{(L-k+1)L}, & m \neq 0, \\ 0, & m = 0. \end{cases} \quad (1)$$

To calculate the similarity distances between different sequences, we should assign a measure to each k -mer based on the k -mers position information. In this study, we use Formula (1) to extract evolutionary information from the nucleic acid sequence. As compared with the other k -mers-based methods, our method involves not only the counts of $s_1s_2 \dots s_k$, but also all the occurring positions of $s_1s_2 \dots s_k$. The method proposed here combines the position distribution information and the counts of the k -mers together, which can capture more phylogenetic information from sequences. For example, for two sequences $X_1 = \text{CCAGTTGCC}$, $X_2 = \text{CCCAGTTGCC}$, the counts of CC in X_1 and X_2 are both 3. If we only consider the frequency of CC , $N_{X_1}(CC) = N_{X_2}(CC)$, the phylogenetic information of

CC captured by $N(CC)$ is not sufficient. However, when we use our measure $f(s_1s_2 \dots s_k), f_{X_1}(CC) = 0.2, f_{X_2}(CC) = 0.133$. Hence, more phylogenetic information of CC can be captured by $f(CC)$.

2.3. Distance Calculations

There are a total of 4^k distinct k -mers for a fixed k . Sorting these k -mers in a fixed order, we can obtain a 4^k -dimensional feature representation vector denoted by $(S_1, S_2, \dots, S_{4^k})$. Then, according to the feature vector and our measure for k -mers, we obtain the corresponding vector $(f_1, f_2, \dots, f_{4^k})$. For given n nucleic acid sequences, we can get a $n \times 4^k$ feature matrix F ($f_{i,j}$ represents the j th feature of the sequence $i, i = 1, 2, \dots, n, j = 1, 2, \dots, 4^k, k$ is the length of k -mers):

$$\begin{bmatrix} f_{1,1} & f_{1,2} & \dots & f_{1,4^k} \\ f_{2,1} & f_{2,2} & \dots & f_{2,4^k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & \dots & f_{n,4^k} \end{bmatrix} \quad (2)$$

There are many methods to calculate the distance between two vectors. In this paper, we use the Manhattan distance [24,25], which was commonly used to analyze similarity of biological sequences. Assuming that $Y = (f_{Y_1}, f_{Y_2}, \dots, f_{Y_{4^k}})$ and $Z = (f_{Z_1}, f_{Z_2}, \dots, f_{Z_{4^k}})$ represent the feature vectors of the two sequences calculated by our method, we use the following formula to calculate the Manhattan distance:

$$d(Y, Z) = \sum_{l=1}^{4^k} |f_{Y_l} - f_{Z_l}|. \quad (3)$$

For the experimental dataset, we can obtain the pairwise distance matrix based on the Manhattan distance. The distance matrix can depict the similarity information of the nucleic acid sequences. After generating the distance matrix, we use it as an input to the MEGA7 [26] and use the Neighbor-Joining (NJ) program [27] to generate the phylogenetic tree. We name this method the *Position-Weighted k-mers (PWkmer)* method.

2.4. Selection of the k Value

The k value in our *PWkmer* method is very important to capture the subtle variation information of a genome sequence. Certainly, a larger value of k will give a vector containing finer evolutionary information. However, many k -mers with large value of k will not occur in the genome sequence. At the same time, some important information may be discarded and noise will dominate when a large value of k is considered. In order to determine the optimal k value, similar to the definition of the matrix in Shannon entropy by Zhao et al. [28], we consider a scoring scheme $score(k)$ to estimate the distribution of k -mers defined as

$$score(k) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{4^k} f_{ij} \log f_{ij}. \quad (4)$$

Note that the larger $score(k)$ is, the more information can be extracted by the k -mers distribution.

The relation between some $score(k)$ and k in our experiment using the dataset of HIV is given in Figure 1. It can be seen that $score(k)$ reaches the largest value at $k = 8$ and decreases after $k > 8$. This indicates that the difference between these genome sequences in 9-mers distributions is decreasing. At the same time, it will require a lot of memory to be computationally efficient when k increases. Therefore, we determine $k = 8$ as the optimal value in our *PWkmer* method to distinguish these genome sequences.

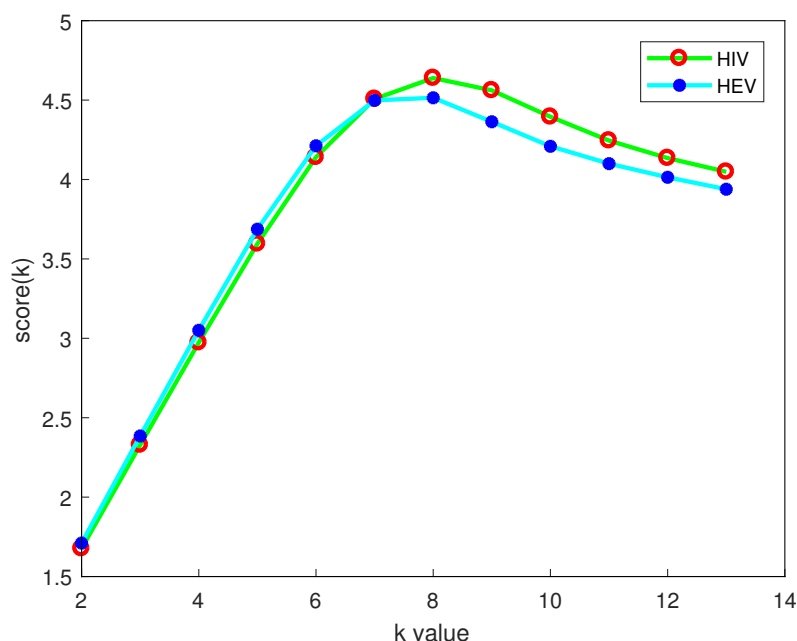


Figure 1. The trend chart of k value vs. scoring scheme $score(k)$. The red circles represent the score of the HIV dataset for different k values, and the blue dots represent the score of the HEV dataset for different k value.

2.5. Accuracy Test of the Phylogenetic Tree Based on the Robinson–Foulds Distance and Robustness Test Using the Modified Bootstrap Method

There are many methods to evaluate the accuracy of tree reconstruction methods. The Robinson–Foulds [29] metric is a way to measure the distance between unrooted phylogenetic trees. In this work, we use it to evaluate the accuracy of the trees we constructed. In general, subtyping of virus species is usually based on multiple sequence alignment in the field of virology. Therefore, we firstly find the reference tree of the species studied. Then, the Robinson–Foulds distance between our tree and the reference tree is implemented in the treedist program of the Phylip package [30]. The smaller the Robinson–Foulds distance is, the more accurate our tree is.

We also use the modified version of the bootstrap method proposed by Yu et al. [6] to evaluate the robustness of the trees we constructed. The workflow is as follows: first, we construct Matrix (2) with each row being the feature vector of each genome sequence. Second, we resample with repeats the 4^k columns to construct a new matrix. Third, we compute the Manhattan distances between any two row vectors based on the new matrix. Then, a distance matrix can be obtained based on the resampled matrix. Fourth, the same tree-building method is used to rebuild the tree. Finally, we repeat the above process a large number of times (usually 100 times). The frequency with which a particular phylogenetic branch emerges can be used as a measure of its reliability.

3. Results

3.1. Subtyping of HIV-1 Based on PWkmer Feature for Complete Genome Sequences

Using our PWkmer method, the phylogenetic analysis was performed on 44 HIV complete genome sequences listed in Table 1. We reconstructed the phylogenetic trees for $k = 2, 3, \dots, 10$. The phylogenetic tree for $k = 8$ is the best among these trees, which agree with our theoretical optimal value for k . The obtained phylogenetic tree for $k = 8$ is shown in Figure 2. It is seen in Figure 2 that the strains from the same subtype are closely clustered together. Forty-four HIV genomes are distinctly divided into four groups: Group M is the main group of viruses in the HIV-1 global pandemic, and it

contains multiple subtypes (A, B, C, D, F, G, H, J, K). Groups N and O are very distinctive forms of the viruses, which originate from other primates and then infect human beings. Group CPZ contains the closest non-human primate viruses related to HIV-1, which are the primate viruses isolated from chimpanzees. In this tree, all subtypes are clearly grouped together as distinct branches, and the closeness relationships among the subtypes are also well demonstrated. Namely, Subtypes B and D are closer to each other than to the others, and Subtype F(A) indeed contains two distinguishable Sub-subtypes F1 and F2 (A1 and A2). All these results are in very good agreement with those of previous studies [5,31].

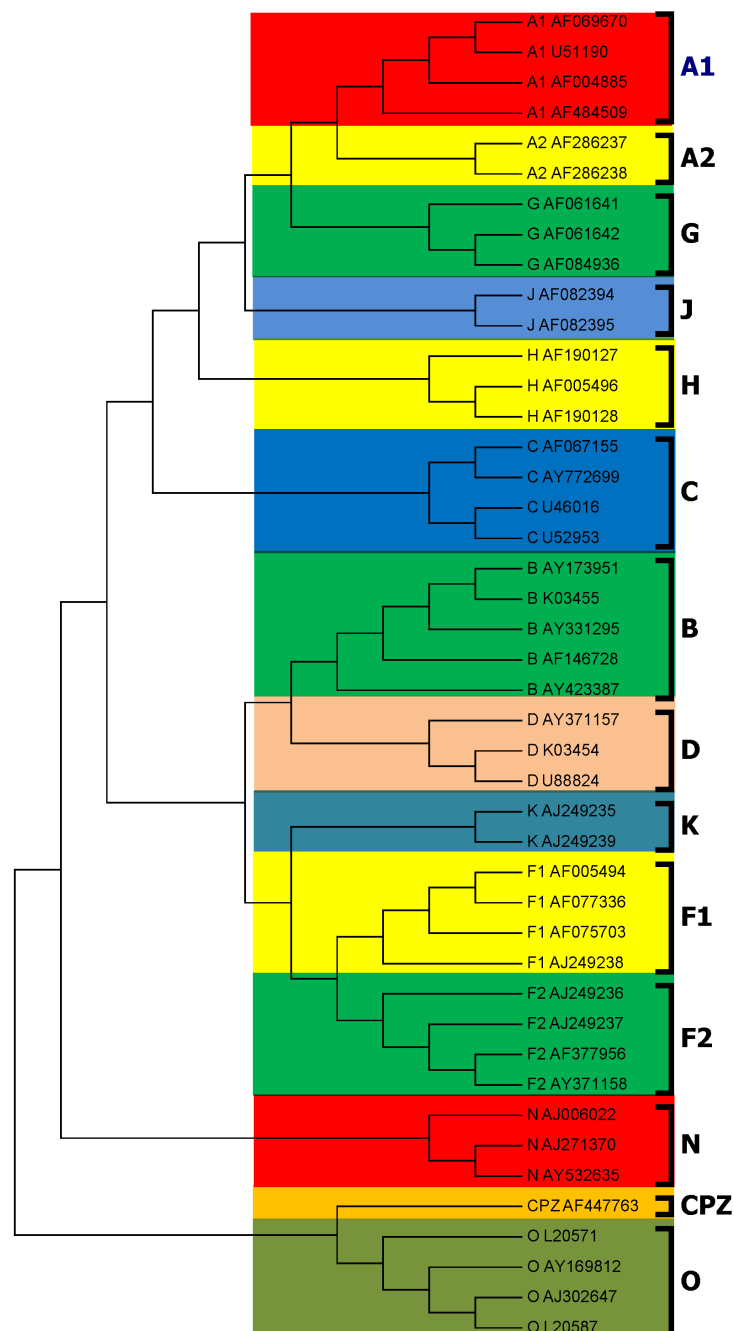


Figure 2. Subtyping of HIV based on position weighted k -mers feature for whole genome sequences. The Neighbor-Joining (NJ) tree of 44 HIV whole genomes is constructed by position weighted k -mers feature distance matrix ($k = 8$).

To verify the accuracy and reliability of the tree constructed by the *PWkmers* method, we used ClustalX [32], which is a multiple sequence alignment program, to construct a reference tree of 44 HIV complete genome sequences. As shown in Figure 3, this tree is quite consistent with the accepted taxonomy from NCBI. Moreover, we calculated the Robinson–Foulds distance between the tree constructed by the *PWkmer* method and the tree constructed by ClustalX. DLTREE [12] and CVTree [21] are the more classical alignment-free methods in the publicized existing software of phylogenetic analysis. We also used them to construct the phylogenetic trees for 44 HIV complete genome sequences. At the same time, we computed the Robinson–Foulds distance between these trees constructed by the *PWkmers* method, CVTree [21], DLTREE [12], and the tree constructed by ClustalX for 44 HIV complete genome sequences. The distances of the tree constructed by each method to the tree constructed by ClustalX are shown in Figure 4. The Robinson–Foulds distance of the *PWkmers* method is minimal, which illustrates that our results are the most closely consistent with the results of ClustalX.

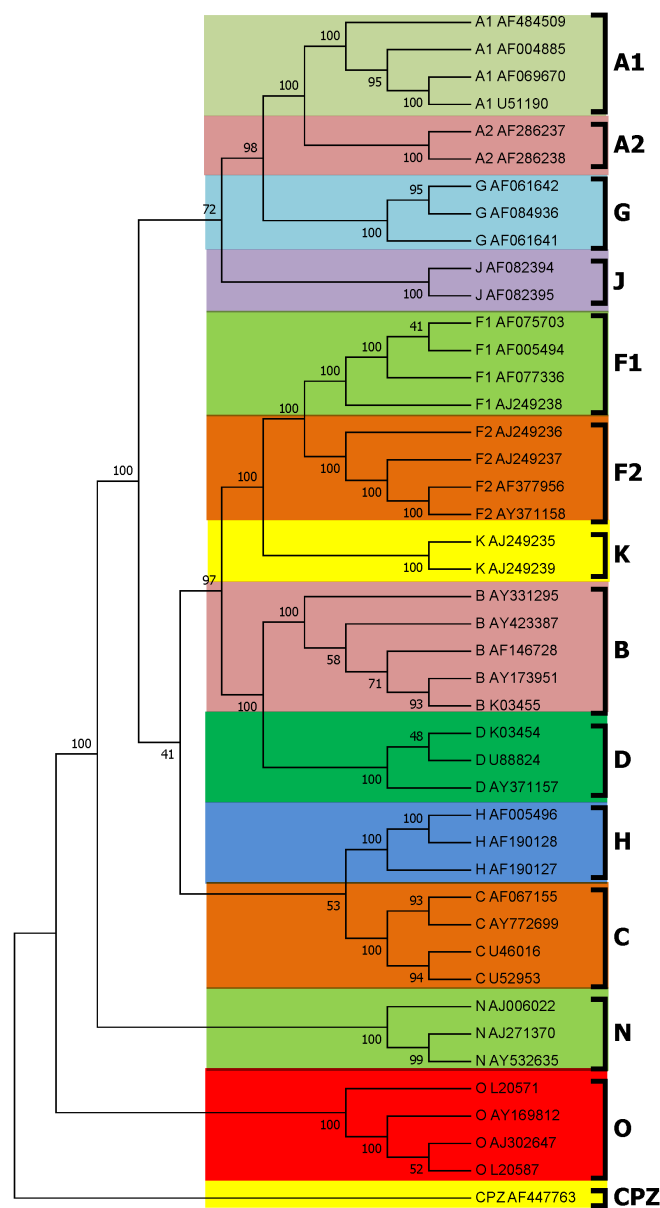


Figure 3. Subtyping of HIV based on alignment for whole genome sequences. The NJ tree of 44 HIV whole genomes is constructed by ClustalX.

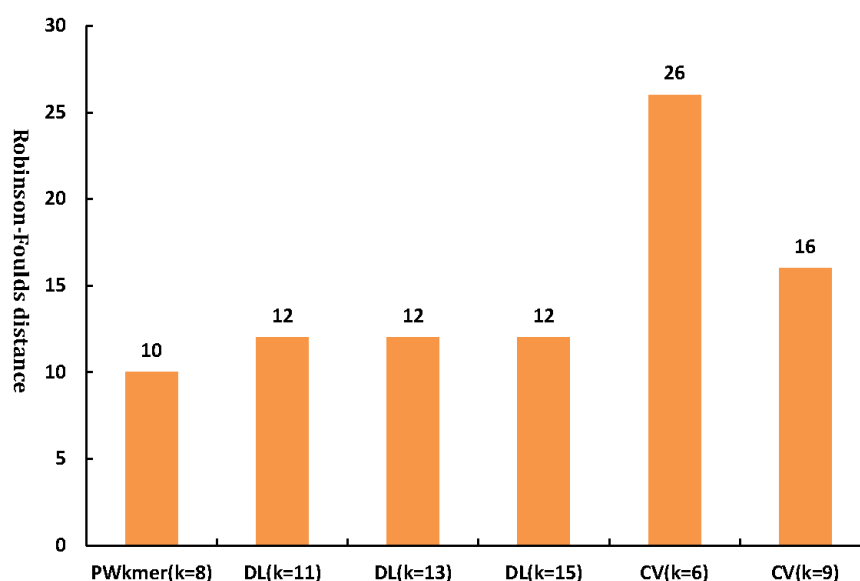


Figure 4. Robinson–Foulds distance between phylogenetic trees reconstructed by the *PWkmer* method, the *CVTree* method [20], the *DLTree* [12] method, and the tree reconstructed by *ClustalX* method for 44 HIV genome sequence in Table 1 (we selected their optimal result tree by *CVTree* and *DLTree*).

The modified bootstrap consensus tree for 44 HIV complete genome sequences is shown in Figure 5. As compared with Figure 2, the division of all HIV-1 genomes into Groups M, N, O, and CPZ is 100% supported. In Group M, each subtype branch is also 100% supported. In particular, in Subtype A and Subtype F of Group M, Sub-subtypes F1 and F2 (A1 and A2) are all 100% supported by the *PWkmers*. The branch of Subtype B and Subtype D is also supported by 100%. In Figure 2, Subtype C is divided into Group M, but in the consistent tree, as shown in Figure 5, Subtype C is divided out of Group M with a low supporting rate (44%).

We also compared the computational time required for our method in comparison to *ClustalX* [32] and *DLTree* [12]. On a modest PC (3.6 GHz quad core Intel Xeon processor, 4 GB RAM), for the whole genome sequences used in Table 1, it took 85 mins 54 secs for the alignment in *ClustalX* [32]. The *DLTree* model approach, which is a free-alignment method, used 20.3 secs of CPU time to get the distance matrices while the present *PWkmers* method only needs 5.8 secs of CPU time to get the distance matrices. This clearly shows the applicability of the *PWkmers* method for large datasets.

We also tested our method on three larger datasets: 867 HIV genomic sequences [5], 1625 HIV circulating recombinant form (CRF) genomic sequences, and 5596 pure subtype HIV genomic sequences from <http://www.hiv.lanl.gov/> for $k = 8$, respectively. We put these three datasets on <https://github.com/my1446/HivStudy>. Our method on our PC only takes 70secs, 244secs, and 46mins 52 secs for each dataset, respectively. For the two datasets including 867 HIV genomic sequences and 5596 pure subtype HIV genomic sequences, all HIV-1 sequences from the same subtype are clustered together with 100% accuracy, while for the dataset including 1625 HIV CRF genomic sequences, the accuracy is 88.35%.

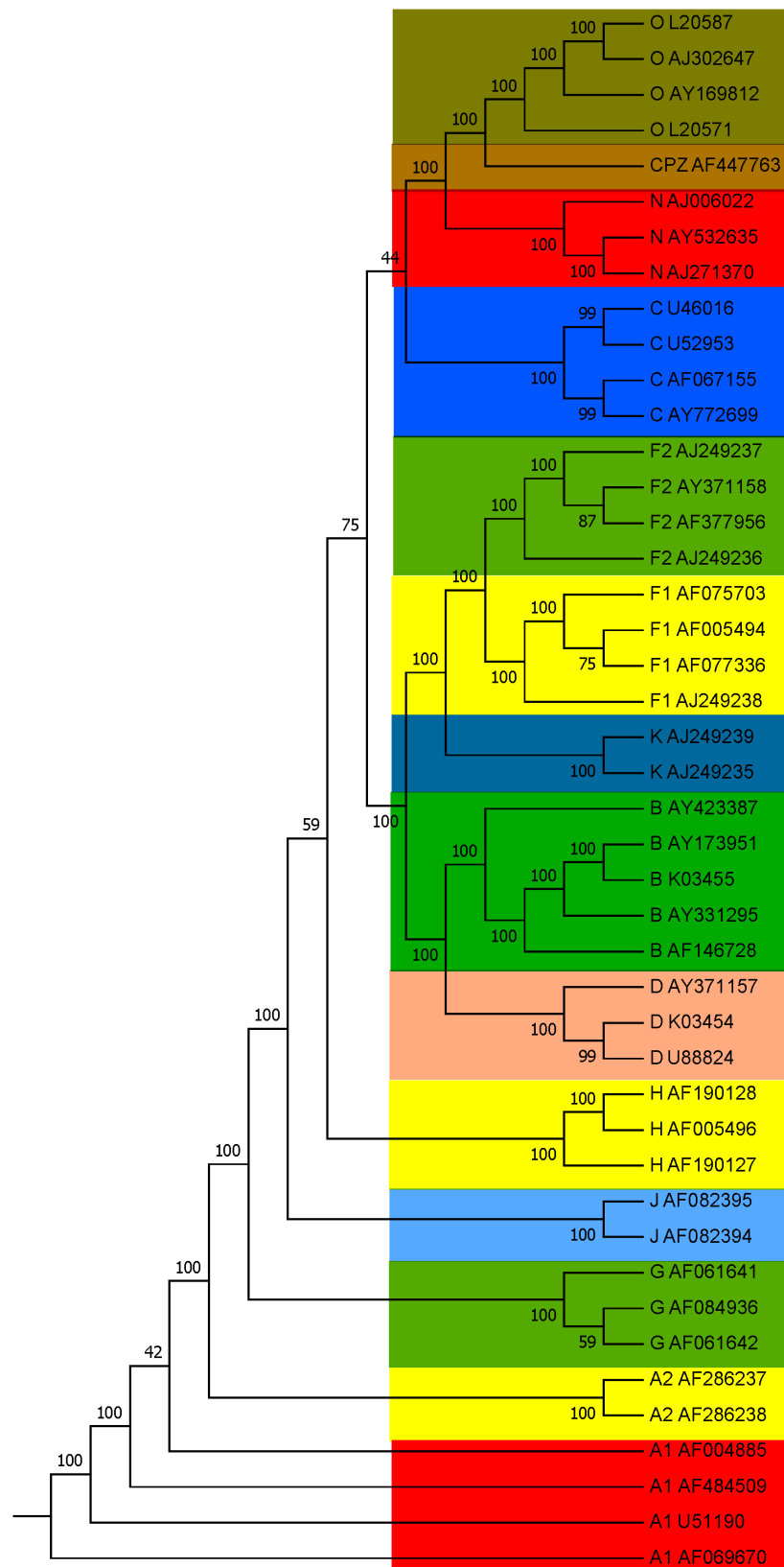


Figure 5. The modified bootstrap consensus tree for Figure 2 based on 100 replicates.

3.2. Application of Our Method on Other Datasets

We also used another benchmark dataset including 48 complete genome sequences used in previously published papers [7,8] to evaluate our *PWkmers* method. All these sequences can be downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). Details of these sequences can be found in [7,8]. Hepatitis E is an inflammation of the liver caused by infection by the HEV (hepatitis E viruses). Hepatitis E is divided into four genotypes, and classification is based on the nucleotide sequences of the complete genome. Genotype 1 has been classified into five subtypes, Genotype 2 into two subtypes, and Genotypes 3 and 4 into ten and seven subtypes [33], respectively.

The tree constructed by our *PWkmers* method (not shown here) indicates that 48 HEV genomes are grouped into four branches. Genotype 1 includes Subtypes Ia, Ib, Ic, Id, and Ie. Genotype 2 contains only a complete HEV genome M1. Genotype 3 includes Subtypes IIIa, IIIb, and IIIc. Genotype 4 includes Subtypes IVa, IVb, and IVc. This shows that our results are consistent with the accepted trees [34,35] and the reference tree constructed by ClustalX.

On the HEV dataset, we also compared the computational time of our method with ClustalX [32] and DLTREE [12]. For the whole genome used in 48 HEV sequences, it took 87 mins 34 secs on our computer for alignment in ClustalX [32]. The DLTREE model approach used 25.7 secs of CPU time to get the distance matrices while the present *PWkmers* method only needs 6 seconds of CPU time to get the distance matrices.

4. Discussion

Subtype classification has always been a focus in the field of virology, especially in the classification of HIV-1 viruses. Because of the wide range of viruses, sequence diversity, and rapid evolution, the development of HIV-1 vaccines is facing enormous challenges. In this work, we propose a new method to solve the problem of HIV-1 classification.

In our *PWkmer* method, we combined the number and position distribution of k -mers, and sequence length to capture more sequence information than traditional methods. In fact, our method records the average position of k -mers on the sequence. Ding et al. [36] presented an alignment-free method based on the normalized k -mers average interval distance to capture evolutionary information for sequence comparison. They only extracted the number and position distribution of k -mers. Tang et al. [37] presented the normalized k -mers average relative distance to improve the method of Ding et al. [36]. Nevertheless, in their methods, the determination of the k value requires empirical calculation, while we directly determine $k = 8$ by $score(k)$.

We computed the Robinson–Foulds distances between the phylogenetic trees reconstructed for different k by our method and the reference tree reconstructed by ClustalX on our HIV-1 dataset, which are shown in Table 2. It can be seen from Table 2 that when $k = 8$, the Robinson–Foulds distances decrease to a lower value, which means that, with the further increase of k , the trees of HIV become unstable and its topological structures change little. From Figure 1 and Table 2, we can see that the relative change in the score value and the Robinson–Foulds distance is the same, which further implies the rationality of the score value defined by us. Furthermore, when $k = 8$, the distance between the tree constructed by the *PWkmers* method and that constructed by ClustalX is the minimum. Therefore, in the subtyping of HIV-1 viruses, we recommend the k value of the string length to be 8.

Table 2. Robinson–Foulds distances between phylogenetic trees reconstructed by our method at $k = 2, 3, \dots, 9, 10$ in Manhattan distance and the tree reconstructed by ClustalX on the HIV dataset.

Species	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
HIV	74	54	38	26	20	14	10	12	14

The HIV subtype classification method based on sequence comparison mainly relies on three gene coding proteins: gag, pol, and env. There are controversies about the spread and origin of *SIV_{CPZ}*.

In this study, as can be seen from Figure 2, *SIV_{CPZ}* is more closely related to group O, and after the bootstrap test, it has a 100% support rate, which is consistent with the classification results based on the proteins env and pol in the HIV database (<http://www.hiv.lanl.gov/>). However, in the HIV database, the classification results based on the protein gag are consistent with the classification results of ClustalX, and *SIV_{CPZ}* is classified outside Groups N and O. As compared with the benchmark dataset used in many studies [5,7], we added a sequence (AF146728, Subtype B, HIV-1 isolated from Australia) which was obviously misclassified by Chang et al. [8]. In our method, we correctly grouped it in Subtype B and the cluster was 100% supported in the bootstrap test. In Pandit et al. [16], the authors concatenated the first and last of 10 sequences in the same subtype, and then classified them according to the fractal dimension. However, given a new sequence, this method cannot be used to determine which subtype it is attached to, or to which subtype it belongs. On the other hand, our method can directly calculate to determine which subtype or sub-subtype the new sequence belongs to. Our results show that the *PWkmer* method is useful and efficient.

5. Conclusions

The subtype classification of species in virology has always been a challenging problem. With the development of sequencing technology, more and more complete genome sequences become available. However, traditional sequence alignment tools and evolutionary models are not efficient in dealing with large-scale genome sequences. In this study, we proposed a new method to solve the problem of the subtype classification of HIV-1. Validation of the Robinson–Foulds distance method and the modified bootstrap method shows that the presented method is reliable for the phylogenetic analysis of HIV-1. At present, the common method for virus subtype classification is based on multi-sequence alignment. Compared with multi-sequence alignment, our method is fast and accurate, and can process large-scale data.

The selection of the *k* value is very important. Specifically, if the *k* value is too small, *k*-mers cannot capture the tiny differences in the genome of different strains; if the *k* value is too large, it takes too much time and computer memory space for function *f* of all *k*-mers. To determine the optimal *k* value, we proposed a new method, which provides a quantitative index for its determination. We then found that the *k* value is independent of the number of genome sequences in the dataset. In summary, our method can capture the *k*-mer distribution information and provide a fast tool for whole genome sequence comparison analysis. We hope that our method will be useful in the phylogenetic analysis of within-species variants using their complete genome sequences.

Supplementary Materials: The following are available online at <http://www.mdpi.com/1099-4300/22/2/255/s1>.

Author Contributions: Y.M. contributed to the conception and design of the study, developed the method and wrote the manuscript. Z.Y. gave the ideas and supervised the project. R.T. analyzed the data and results. All authors discussed the results and reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This project was supported by the Natural Science Foundation of China (Grant No. 11871061); Collaborative Research project for Overseas Scholars (including Hong Kong and Macau) of National Natural Science Foundation of China (Grant No. 61828203); the Research Foundation of Education Commission of Hunan Province of China (Grant No. 17K090), the innovation project of Hunan Province of China (Grant No. Cx2016B252), the Science and Technology Project of Jiangxi Provincial Education Department (GJJ170820), and partially by the Australian Research Council Grant DP160101366.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zachary, T.; Aboulafia, D.M. Review of screening guidelines for non-AIDS-defining malignancies: Evolving issues in the era of highly active antiretroviral therapy. *Aids Rev.* **2012**, *14*, 3–16.
2. Lemey, P.; Pybus, O.G.; Rambaut, A.; Drummond, A.J.; Robertson, D.L.; Roques, P.; Worobey, M.; Vandamme, A.M. The molecular population genetics of HIV-1 group O. *Genetics* **2004**, *167*, 1059–1068.

3. Tebit, D.M.; Nankya, I.; Arts, E.J.; Gao, Y. HIV diversity, recombination and disease progression: How does fitness “fit” into the puzzle? *Aids Rev.* **2015**, *9*, 75–87.
4. Herniou, E.A.; Luque, T.; Chen, X.; Vlak, J.M.; Winstanley, D.; Cory, J.S.; O’reilly, D.R. Use of whole genome sequence data to infer baculovirus phylogeny. *J. Virol.* **2001**, *75*, 8117–8126.
5. Wu, X.; Cai, Z.; Wan, X.F.; Hoang, T.; Goebel, R.; Lin, G. Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics* **2007**, *23*, 1744–1752.
6. Yu, Z.G.; Chu, K.H.; Li, C.P.; Vo, A.; Zhou, L.Q.; Wang, R.W. Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. *BMC Evol. Biol.* **2010**, *10*, 192.
7. Yang, X.; Wang, T. A novel statistical measure for sequence comparison on the basis of k-word counts. *J. Theor. Biol.* **2013**, *318*, 91–100.
8. Chang, G.; Wang, H.; Zhang, T. A novel alignment-free method for whole genome analysis: Application to HIV-1 subtyping and HEV genotyping. *Inf. Sci.* **2014**, *279*, 776–784.
9. 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.
10. Xie, X.H.; Yu, Z.G.; Han, G.S.; Yang, W.F.; Anh, V. Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles. *Mol. Phylogenet. Evol.* **2015**, *89*, 37–45.
11. Yang, W.F.; Yu, Z.G.; Anh, V. Whole genome/proteome based phylogeny reconstruction for prokaryotes using higher order Markov model and chaos game representation. *Mol. Phylogenet. Evol.* **2016**, *96*, 102–111.
12. Wu, Q.; Yu, Z.G.; Yang, J. Dltree: Efficient and accurate phylogeny reconstruction using the dynamical language method. *Bioinformatics* **2017**, *33*, 2214–2215.
13. Li, W.; Freudenberg, J. Alignment-free approaches for predicting novel Nuclear Mitochondrial Segments (NUMTs) in the human genome. *Gene* **2019**, *691*, 141–152.
14. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* **2017**, *18*, 186.
15. Hao, B.; Qi, J. Prokaryote phylogeny without sequence alignment: From avoidance signature to composition distance. *J. Bioinf. Comput. Biol.* **2004**, *2*, pp. 1–19.
16. Pandit, A.; Dasanna, A.K.; Sinha, S. Multifractal analysis of HIV-1 genomes. *Mol. Phylogenet. Evol.* **2012**, *62*, 756–763.
17. Blaisdell, B.E. A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc. Nat. Acad. Sci. USA* **1986**, *83*, 5155–5159.
18. Snel, B.; Bork, P.; Huynen, M.A. Genomes in flux: The evolution of archaeal and proteobacterial gene content. *Genome Res.* **2002**, *12*, 17–25.
19. Song, K.; Ren, J.; Reinert, G.; Deng, M.; Waterman, M.S.; Sun, F. New developments of alignment-free sequence comparison: Measures, statistics and next-generation sequencing. *Brief. Bioinf.* **2014**, *15*, 343–353.
20. Yu, Z.G.; Anh, V.; Lau, K.S. Multifractal and correlation analyses of protein sequences from complete genomes. *Phys. Rev. E* **2003**, *68*, 021913.
21. Zuo, G.; Hao, B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genom. Proteom. Bioinf.* **2015**, *13*, 321–331.
22. Han, G.-B.; Cho, D.-H. Genome classification improvements based on k-mer intervals in sequences. *Genomics* **2019**, *111*, 1574–1582.
23. Yu, Z.G.; Zhou, L.Q.; Anh, V.V.; Chu, K.H.; Long, S.C.; Deng, J.Q. Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment. *J. Mol. Evol.* **2005**, *60*, 538–545.
24. Krause, E.F. Taxicab geometry: Adventure in non-euclidean geometry. *Mathematical Gazette* **1988**, *72*, 255.
25. Solis-Reyes, S.; Avino, M.; Poon, A.; Kari, L. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-I genomes. *PLoS ONE* **2018**, *13*, e0206409.
26. Kumar S.; Stecher G.; Tamura K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874.
27. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.

28. Zhao, Z.Q.; Han, G.S.; Yu, Z.G.; Li, J. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. *Comput. Biol. Chem.* **2015**, *57*, 21–28.
29. Robinson, D.F.; Foulds, L.R. Comparison of phylogenetic trees. *Math. Biosci.* **1981**, *53*, 131–147.
30. Felsenstein, J. Mathematics vs. evolution: Mathematical evolutionary theory. *Science* **1989**, *246*, 941–942.
31. Foley, B.T.; Korber, B.T.M.; Leitner, T.K.; Apetrei, C.; Hahn, B.; Mizrachi, I.; Mullins, J.; Rambaut, A.; Wolinsky, S. HIV Sequence Compendium 2018. Available online: <https://www.osti.gov/biblio/1458915> (accessed on 22 February 2020)
32. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; Mcgettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948.
33. Manns, M.P.; Lohse, A.W.; Vergani, D. Autoimmune hepatitis-Update 2015. *J. Hepatol.* **2015**, *62*, S100–S111.
34. Liu, Z.; Meng, J.; Sun, X. A novel feature-based method for whole genome phylogenetic analysis without alignment: Application to HEV genotyping and subtyping. *Biochem. Biophys. Res. Commun.* **2015**, *368*, 223–230.
35. Ling, L.; Li, C.; Hagedorn, C.H. Phylogenetic analysis of global hepatitis E virus sequences: Genetic diversity, subtypes and zoonosis. *Rev. Med. Virol.* **2006**, *16*, 5–36.
36. Ding, S.; Li, Y.; Yang, X.; Wang, T. A simple k -word interval method for phylogenetic analysis of DNA sequences. *J. Theor. Biol.* **2013**, *317*, 192–199.
37. Tang, J.; Hua, K.; Chen, M.; Zhang, R.; Xie, X. A novel k -word relative measure for sequence comparison. *Comput. Biol. Chem.* **2014**, *53*, 331–338.



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).