frontiers in Genetics

ORIGINAL RESEARCH
published: 09 April 2019
doi: 10.3389/fgene.2019.00234

A Novel Approach to Clustering Genome Sequences Using Inter-nucleotide Covariance

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

[1] Department of Mathematical Sciences, Tsinghua University, Beijing, China, [2] Department of Biological Sciences, Chicago State University, Chicago, IL, United States

Classification of DNA sequences is an important issue in the bioinformatics study, yet most existing methods for phylogenetic analysis including Multiple Sequence Alignment (MSA) are time-consuming and computationally expensive. The alignment-free methods are popular nowadays, whereas the manual intervention in those methods usually decreases the accuracy. Also, the interactions among nucleotides are neglected in most methods. Here we propose a new Accumulated Natural Vector (ANV) method which represents each DNA sequence by a point in $\mathbb{R}^{18}$. By calculating the Accumulated Indicator Functions of nucleotides, we can further find an Accumulated Natural Vector for each sequence. This new Accumulated Natural Vector not only can capture the distribution of each nucleotide, but also provide the covariance among nucleotides. Thus global comparison of DNA sequences or genomes can be done easily in $\mathbb{R}^{18}$. The tests of ANV of datasets of different sizes and types have proved the accuracy and time-efficiency of the new proposed ANV method.

Keywords: accumulated natural vector, phylogenetic analysis, alignment-free, inter-nucleotide covariance, genomes

OPEN ACCESS

Edited by:
Alfredo Pulvirenti,
Università degli Studi di Catania, Italy

Reviewed by:
Cheong Xin Chan,
University of Queensland, Australia
Stefano Piotto,
University of Salerno, Italy

***Correspondence:**
Stephen S.-T. Yau
yau@uic.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *07 September 2018*
Accepted: *04 March 2019*
Published: *09 April 2019*

Citation:
Dong R, He L, He RL and Yau SS-T (2019) A Novel Approach to Clustering Genome Sequences Using Inter-nucleotide Covariance. Front. Genet. 10:234. doi: 10.3389/fgene.2019.00234

INTRODUCTION

With the rapid development of Next Generation Sequencing technology, more and more information of the genome sequences is available. Studying sequence similarity is a crucial question in research and can explain phylogenetic relationships by constructing trees. One of the most commonly used methods, Multiple Sequence Alignment (MSA) uses dynamic programming, a regression technique that finds an optimal alignment by assigning scores to different possible alignments and taking the one with the highest score (Yu et al., 2013a). However, the computational cost of MSA is extremely high and MSA may not produce accurate phylogeny for diverse systems of different families of RNA viruses (Yu et al., 2013b). Alignment-free approaches have been developed to overcome those limitations. Published alignment-free methods include Markov chain models (Apostolico and Denas, 2008), chaos theory (Hatje and Kollmar, 2012), and some other methods based on the statistics of oligomer frequency and associated with a fixed length segment, known as *k-mer* (Sims et al., 2009). Yau and his team proposed the natural vector method, which takes the position of each nucleotide into consideration. The natural vector method performs well on many datasets (Deng et al., 2011; Yu et al., 2013b; Hoang et al., 2016; Li et al., 2016), however, it only considers the number, average position and dispersion of positions of each nucleotide. Relationships between nucleotides are also important, especially when the functions may be related to interactions of nucleotides, such as the folding of a chromosome. In this paper, we propose a new

Accumulated Natural Vector (ANV) method, which not only considers the basic property of each nucleotide, but also the covariance between them. In the traditional Natural Vector (NV) method, each sequence is uniquely represented by a single point in $\mathbb{R}^{12}$. The traditional Natural Vector approach is firstly introduced in Deng et al. (2011): for a sequence of length N, n_α ($\alpha\epsilon\{A, C, T, G\}$) denotes the number of nucleotide α in the sequence. s [α] [v] is the distance from the first nucleotide (regarded as origin) to the v^{th} nucleotide α in the DNA sequence. $T_\alpha = \sum_{v=1}^{n_\alpha} \mathrm{s}\,[\alpha]\,[\mathrm{v}]$ denotes the total distance of each set of A,C,G,T from the origin, $\alpha\epsilon\{A, C, T, G\}$. $\mu_\alpha = \frac{T_\alpha}{n_\alpha}$, is the mean value of the distances of nucleotide α from the origin. $D_2^\alpha = \sum_{v=1}^{n_\alpha} \frac{(\mathrm{s}[\alpha][\mathrm{v}]-\mu_\alpha)^2}{n_\alpha \times N}$, is the normalized central moment of order 2, which can also be seen as the variance of the positions of nucleotide α. Therefore, a DNA sequence can be represented by a 12-dim vector:

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_c, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T)$$

In this paper, we propose an Accumulated Natural Vector approach, which projects each sequence into a point in $\mathbb{R}^{18}$, where the additional six dimensions describe the covariance between nucleotides. Obviously, ANV can provide more information than the traditional NV method, and doesn't include the human intervention, such as choosing the optimal value of k in the *k-mer* method. Therefore, it can distinguish different sequences and classify species into correct clusters with higher accuracy and less time cost.

MATERIALS AND METHODS

Materials

The following six datasets were used to validate the method. The Coronaviruses dataset includes 36 viral genomes, in which 34 viruses are from the exact same dataset with (Woo et al., 2005; Yu et al., 2010; Hoang et al., 2015) and the other two viruses are new members in Coronavirus. The second dataset consists of the genomes of 38 Influenza A viruses, which is a classic dataset to test if a new proposed method performs well. The third dataset includes 72 viruses from Zheng et al. (2015), which focuses on the classification of Ebolaviruses. The fourth one is from our colleagues' previous paper (Li et al., 2016) which includes 351 viruses chosen randomly under some criteria. The fifth one is the mitochondrial genomes of 31 mammals, which can be clustered into seven well-known categories. All the sequence materials can be found on NCBI with the reference number provided in the Appendices. We also generated different mutations by simulation in a DNA sequence and constructed phylogenetic trees of simulated sequences to test our ANV method.

All computations in this paper are done on a Dell laptop equipped with Intel i7 Processor under Windows 10 Home Premium with 8 GB RAM, together with the Matlab (version R2017a) and Mega X.

Methods

Indicator Function

For a given genomic sequence, we first define four Indicator Functions (u) for Adenine, Cytosine, Guanine and Thymine, respectively:

$$u_\alpha\,(i) = \begin{cases} 1, & if\ \alpha\ appears\ at\ the\ i^{th}\ position\ of\ the\ sequence \\ 0, & if\ \alpha\ doesn't\ appear\ at\ the\ i^{th}\ position\ of\ the\ sequence \end{cases} \quad (1)$$

where $\alpha\epsilon\{A, C, T, G\}$, and i = 1, 2, . . . , N. HereN is the length of the whole sequence.

For example, if the genomic sequence is "*ATCTAGCT*," then the four Indicator Functions are shown in **Table 1**.

Here are some simple properties about the Indicator Functions:

1. Each column has the sum of 1.

$$\sum\nolimits_{\alpha\in\{A,C,G,T\}} u_\alpha\,(i) = 1, for\ \mathrm{i} = 1, 2, \ldots, \mathrm{N} \quad (2)$$

2. Each row has the sum of the number of corresponding nucleotide.

$$n_\alpha = \sum\nolimits_{i=1}^{N} u_\alpha(i), \alpha \in \{A, C, G, T\} \quad (3)$$

Accumulated Indicator Function

Now we define four Accumulated Indicator Functions as the following:

$$\tilde{u}_\alpha\,(N) = \sum\nolimits_{i=1}^{N} u_\alpha\,(i) \quad (4)$$

The four Accumulated Indicator Functions for the example above ("*ATCTAGCT*"), are shown in **Table 2**.

Here are some properties about the Accumulated Indicator Functions:

1. The i^{th} column has the sum of i.

$$\sum\nolimits_{\alpha\in\{A,C,G,T\}} \tilde{u}_\alpha\,(i) = i \quad (5)$$

2. The last column is the total number of the nucleotide α in the sequence.

TABLE 1 | The Indicator Functions of the sequence "ATCTAGCT".

Sequence	A	T	C	T	A	G	C	T
Position(i)	1	2	3	4	5	6	7	8
u_A (i)	1	0	0	0	1	0	0	0
u_C (i)	0	0	1	0	0	0	1	0
u_G (i)	0	0	0	0	0	1	0	0
u_T (i)	0	1	0	1	0	0	0	1

TABLE 2 | The Accumulated Indicator Functions of the sequence "ATCTAGCT."

Sequence	A	T	C	T	A	G	C	T
Position(i)	1	2	3	4	5	6	7	8
$\tilde{u}_A(i)$	1	1	1	1	2	2	2	2
$\tilde{u}_C(i)$	0	0	1	1	1	1	2	2
$\tilde{u}_G(i)$	0	0	0	0	0	1	1	1
$\tilde{u}_T(i)$	0	1	1	2	2	2	2	3

$$\tilde{u}_\alpha(N) = n_\alpha \tag{6}$$

3. $$\sum_{i=1}^{N} \tilde{u}_\alpha(i) = n_\alpha \times (N + 1 - \mu_\alpha) \tag{7}$$

μ_k is the average position in the Natural Vector in Deng et al. (2011).

Property 1 and 2 can be easily proved by the definition of Indicator Function (u_α) and Accumulated Indicator Function ($\tilde{u}_\alpha$), now we prove the Property 3, which builds up the relationship between the Accumulated Indicator Function and the average position of a specific nucleotide.

If we assume that the positions of nucleotide α are $t_1, t_2, \ldots, t_{n_\alpha}$, where n_α is the number of nucleotide α in the sequence, then the basic form of Accumulated Indicator Function should be, which satisfies $1 \le t_1 < t_2 < \ldots < t_{i-1} < t_i < \ldots < t_{n_\alpha} \le N$,

$$\underbrace{[0, 0, \ldots, 0,}_{t_1 - 1} \underbrace{1, 1, \ldots, 1,}_{(t_2 - 1) - (t_1 - 1)} 2, 2, \ldots, 2, \ldots$$
$$\underbrace{(n_\alpha - 1), (n_\alpha - 1), \ldots (n_\alpha - 1),}_{(t_{n_\alpha} - 1) - (t_{n_\alpha - 1} - 1)} \underbrace{n_\alpha, \ldots, n_\alpha]}_{N - t_{n_\alpha} + 1}$$

If we add up those N elements above and denote the sum as Ω_α and $t_0 = 0$, we have:

$$\begin{aligned}
\Omega_\alpha &= \sum_{i=1}^{n_\alpha} (t_i - t_{i-1}) \times (i - 1) + \left(N - t_{n_\alpha} + 1\right) \times n_\alpha \\
&= (t_2 - t_1) \times 1 + (t_3 - t_2) \times 2 + (t_4 - t_3) \times 3 + \ldots \\
&\quad + \left(t_{n_\alpha} - t_{n_\alpha - 1}\right) \times (n_\alpha - 1) + \left(N - t_{n_\alpha} + 1\right) \times n_\alpha \\
&= -t_1 - t_2 - t_3 - \ldots - t_{n_\alpha - 1} + (n_\alpha - 1) \times t_{n_\alpha} + \ldots \\
&\quad + \left(N - t_{n_\alpha} + 1\right) \times n_\alpha \\
&= -t_1 - t_2 - t_3 - \ldots - t_{n_\alpha - 1} - t_{n_\alpha} + (N + 1) \times n_\alpha \\
&= -n_\alpha \times \mu_\alpha + (N + 1) \times n_\alpha
\end{aligned} \tag{8}$$

Thus, we have

$$\sum_{i=1}^{N} \tilde{u}_\alpha(i) = n_\alpha \times (N + 1 - \mu_\alpha) \tag{9}$$

Therefore, we use

$$\zeta_\alpha = \frac{\Omega_\alpha}{n_\alpha} \tag{10}$$

to describe the average position of nucleotide α, which indicates the distance of the average position to the end of the sequence.

Inter-nucleotide Covariance

For two finite point sets with equal number of elements: $A = \{a_1, a_2, \ldots, a_n\}$, $B = \{b_1, b_2, \ldots, b_n\}$ in $\mathbb{R}$, which satisfy $a_1 < a_2 < \ldots < a_n$ and $b_1 < b_2 < \ldots < b_n$, the covariance of two sets can be defined as follows:

$$\mathrm{cov}(A, B) = \sum_{i=1}^{n} \frac{(a_i - u_A) \times (b_i - u_B)}{a_n \times b_n} \tag{11}$$

where $u_A = \sum_{i=1}^{n} a_i / n$ and $u_B = \sum_{i=1}^{n} b_i / n$.

Now we apply the covariance formula above to the Accumulated Indicator Functions. A set is a collection of definite, distinct objects, known as the elements or members of the set. Now for each nucleotide, we have an array of N elements which is the Accumulated Indicator Function for the nucleotide $\alpha \in \{A, C, G, T\}$:

$$[0, 0, \ldots, 0, 1, 1, \ldots, 1, 2, 2, \ldots, 2, \ldots (n_\alpha - 1),$$
$$(n_\alpha - 1), \ldots, (n_\alpha - 1), n_\alpha, \ldots, n_\alpha]$$

However, those N elements cannot build up a set of N elements since many of them are replicated. Hence, we extend the definition of set to a generalized concept, where the elements in a set can be the same. In this generalized definition, each nucleotide has a set of N elements and they can be arranged in the ascending order, i.e., from the smallest to the biggest number. Thus, we can use the covariance formula (11). As the example of sequence *"ATCTAGCT,"* the covariance of nucleotide A and C can be computed in this way: the generalized set of nucleotide A is {1,1,1,1,2,2,2,2} and of C is {0,0,1,1,1,1,2,2}. Each generalized set has N = 8 elements and the generalized covariance would be

$$\theta_A = \sum_{i=1}^{N} \tilde{u}_A(i) / N = \frac{1 + 1 + 1 + 1 + 2 + 2 + 2 + 2}{8} = 1.5 \tag{12}$$

$$\theta_C = \sum_{i=1}^{N} \tilde{u}_C(i) / N = \frac{0 + 0 + 1 + 1 + 1 + 1 + 2 + 2}{8} = 1 \tag{13}$$

$$\begin{aligned}
\mathrm{cov}(A, C) &= \sum_{i=1}^{N} \frac{(\tilde{u}_A(i) - \theta_A) \times (\tilde{u}_C(i) - \theta_C)}{n_A \times n_C} \\
&= \frac{1}{2 \times 2} \times [(1 - 1.5) \times (0 - 1) + (1 - 1.5) \times (0 - 1) \\
&\quad + (1 - 1.5) \times (1 - 1) + (1 - 1.5) \times (1 - 1) \\
&\quad + (2 - 1.5) \times (1 - 1) + (2 - 1.5) \times (1 - 1) \\
&\quad + (2 - 1.5) \times (2 - 1) + (2 - 1.5) \times (2 - 1)] \\
&= \frac{1}{2}
\end{aligned} \tag{14}$$

Similarly, we can get cov (A,G), cov (A, T), cov (C, G), cov (C, T), cov (G, T).

Compatibility of Variance and Covariance

For two nucleotides like α and β, the covariance formula is

$$\mathrm{cov}(\alpha, \beta) = \sum_{i=1}^{N} \frac{(\tilde{u}_\alpha(i) - \theta_\alpha) \times \left(\tilde{u}_\beta(i) - \theta_\beta\right)}{n_\alpha \times n_\beta} \tag{15}$$

Then it is obvious that when $\alpha = \beta$, the corresponding formula should be

$$D_\alpha = \mathrm{cov}(\alpha, \alpha) = \sum_{i=1}^{N} \frac{(\tilde{u}_\alpha(i) - \theta_\alpha) \times (\tilde{u}_\alpha(i) - \theta_\alpha)}{n_\alpha \times n_\alpha}$$

$$= \sum_{i=1}^{N} \left(\frac{\tilde{u}_\alpha(i) - \theta_\alpha}{n_\alpha}\right)^2 \quad (16)$$

The formula above defines the variance of the positions of nucleotide α.

Accumulated Natural Vector

For a given nucleotide sequence, now we can build up its Accumulated Natural Vector. The first four dimensions describe the number of each nucleotide, denoted as n_A, n_C, n_G, n_T, which are the last column of the Accumulated Indicator Functions. The second four dimensions describe the average distance of nucleotides to the end of the sequence, denoted as $\zeta_A = \frac{\Omega_A}{n_A}, \zeta_c = \frac{\Omega_C}{n_C}, \zeta_G = \frac{\Omega_G}{n_G}, \zeta_T = \frac{\Omega_T}{n_T}$ as formula (10). The third four dimensions describe the divergence of each nucleotide, denoted as $D_A = \sum_{i=1}^{N} \left(\frac{\tilde{u}_A(i)-\theta_A}{n_A}\right)^2, D_C = \sum_{i=1}^{N} \left(\frac{\tilde{u}_C(i)-\theta_C}{n_C}\right)^2, D_G = \sum_{i=1}^{N} \left(\frac{\tilde{u}_G(i)-\theta_G}{n_G}\right)^2, D_T = \sum_{i=1}^{N} \left(\frac{\tilde{u}_T(i)-\theta_T}{n_T}\right)^2$ as formula (16). Please note that this D_α is a little different from the D_2^α in the traditional Natural Vector method since the previous definition of variance cannot be extended to a reliable definition of covariance. The last six dimensions describe the covariances between each two nucleotides, denoted as cov (A, G), cov (A, T), cov (C, G), cov (C, T), cov (G, T) as formula (15). And the universal form of Accumulated Natural Vector is

$$(n_A, n_C, n_G, n_T, \zeta_A, \zeta_c, \zeta_G, \zeta_T, D_A, D_C, D_G, D_T,$$
$$\text{cov (A, C)}, \text{cov (A, G)}, \text{cov (A, T)}, \text{cov (C, G)}, \text{cov (C, T)}, \text{cov (G, T)})$$

Euclidean Distances Between Accumulated Natural Vectors

From section 2.2.1 to section 2.2.5, we introduce how a DNA sequence is represented by a vector in $\mathbb{R}^{18}$ space. Therefore,

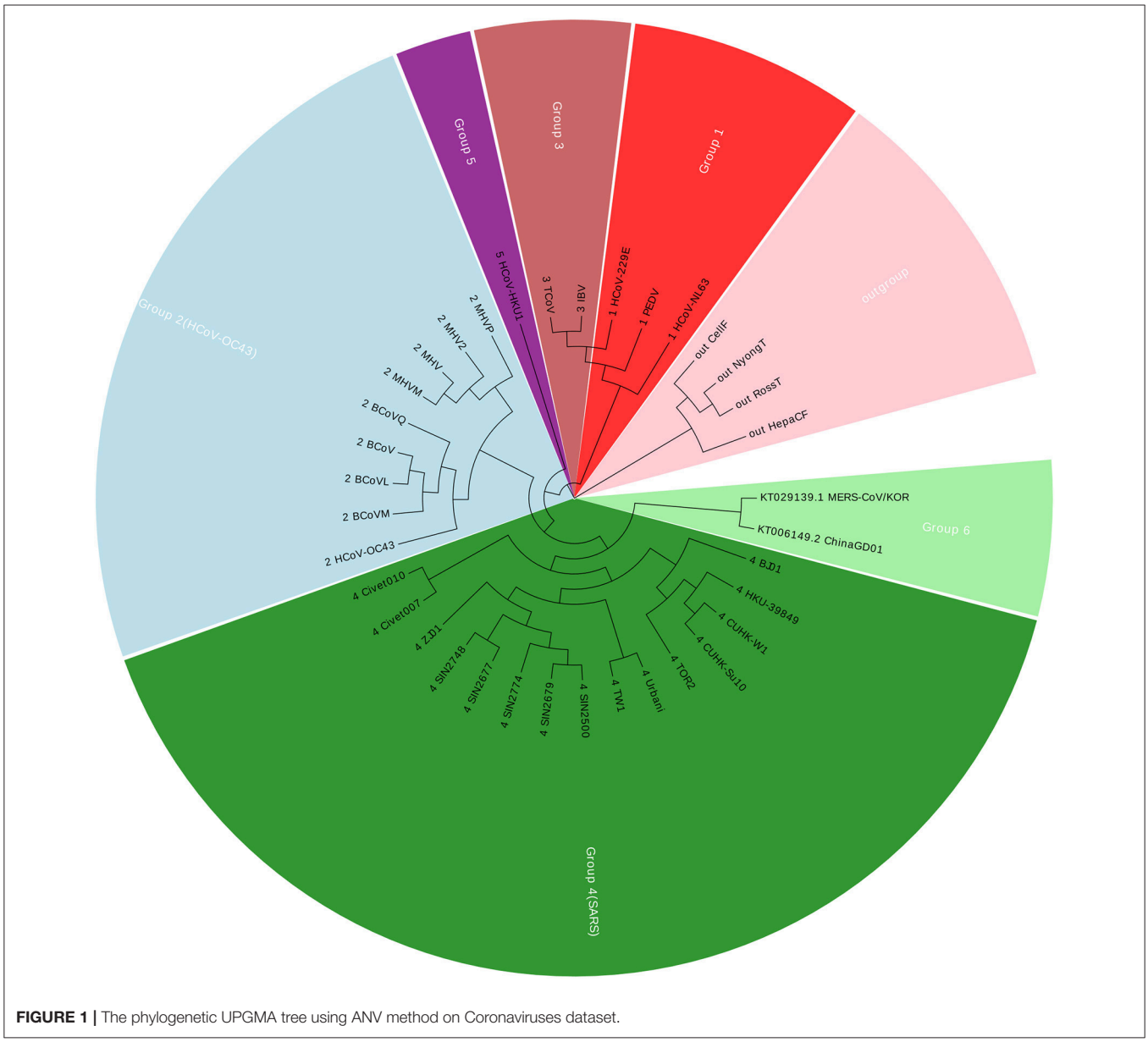

FIGURE 1 | The phylogenetic UPGMA tree using ANV method on Coronaviruses dataset.

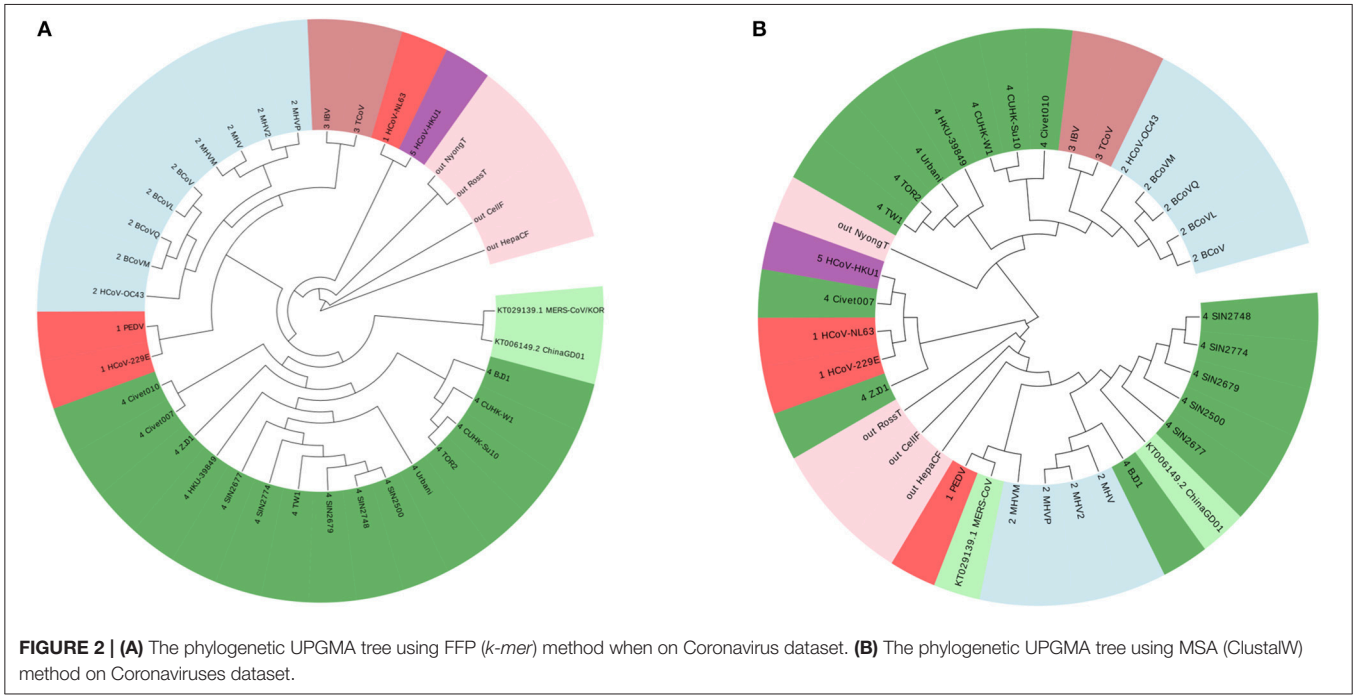

FIGURE 2 | (A) The phylogenetic UPGMA tree using FFP (*k-mer*) method when on Coronavirus dataset. **(B)** The phylogenetic UPGMA tree using MSA (ClustalW) method on Coronaviruses dataset.

the distance between two sequences can be measured by the Euclidean distance between two vectors. Suppose that now we have two sequences in $\mathbb{R}^w$ (in our case, w = 18), denoted as $x = (x_1, ..., x_w)$ and $(y_1, ..., y_w)$, the Euclidean distance between them is

$$d(x, y) = \left(\sum_{i=1}^{w} (x_i - y_i)^2\right)^{\frac{1}{2}} \quad (17)$$

For a dataset of m different sequences, we can construct a distance matrix $D = (d_{ij})_{m\times m}$, and $d_{ij}(\geq 0)$ represents the Euclidean distance between sequence i and sequence j. D is a symmetric matrix and the diagonal element is zero.

Constructing Phylogenetic Trees and Comparisons

In this research, we use Mega X to build up phylogenetic trees. In order to eliminate the influences of different algorithms of constructing trees, we apply the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm (Sneath and Sokal, 1973) for analysis on the four datasets.

For comparison with other common alignment or alignment-free method, we also perform *k-mer* and MSA (ClustalW or MUSCLE) on the same dataset. The Feature frequency profile (FFP) (Woo et al., 2005), which is based on *k-mer* frequency, calculates the frequency of each *k-mer* in the sequence and turns a DNA sequence into a vector in a 4^k-dimensional space. The Euclidean distance between two *k-mer* vectors can also be computed by formula (17). We apply MSA method, ClustalW on several datasets as well, with the default parameters in Mega X. ClustalW is much slower than another MSA algorithm, MUSCLE, while ClustalW can give a better result. MUSCLE is applied on the fourth dataset of 351 viruses and after we get the alignment result of the viruses, distance matrix is calculated using Hamming distance, to find the nearest neighbor of each virus. Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. It measures the minimum number of substitutions required to change one string into the other or the minimum number of errors that could have transformed one string into the other. Since alignment approaches are to arrange the sequences to identify regions of similarity between the sequences, the alignment would provide the performance of each sequence on a fixed number of positions. Therefore, the Hamming distance can be calculated by simply counting the number of pairwise differences in character states.

In the simulated dataset, we use the pairwise alignment distance by the "seqpdist" function inside MATLAB Bioinformatics toolbox, which uses the Jukes-Cantor algorithms as the correct tree, since the sequences are simulated according to a base sequence. Then the distance matrices are compared using Robinson-Foulds distances, which can measure the congruence to the reference topology.

RESULTS

We apply the Accumulated Natural Vector method on five datasets, and compare the results with common methods, such as MSA, *k-mer* (FFP) and the traditional Natural Vector method. From comparison, the results of Accumulated Natural Vector are more accurate and the calculation cost is very small compared to others. A dataset of 351 viruses has also been tested, and laptop cannot bear such a heavy burden of

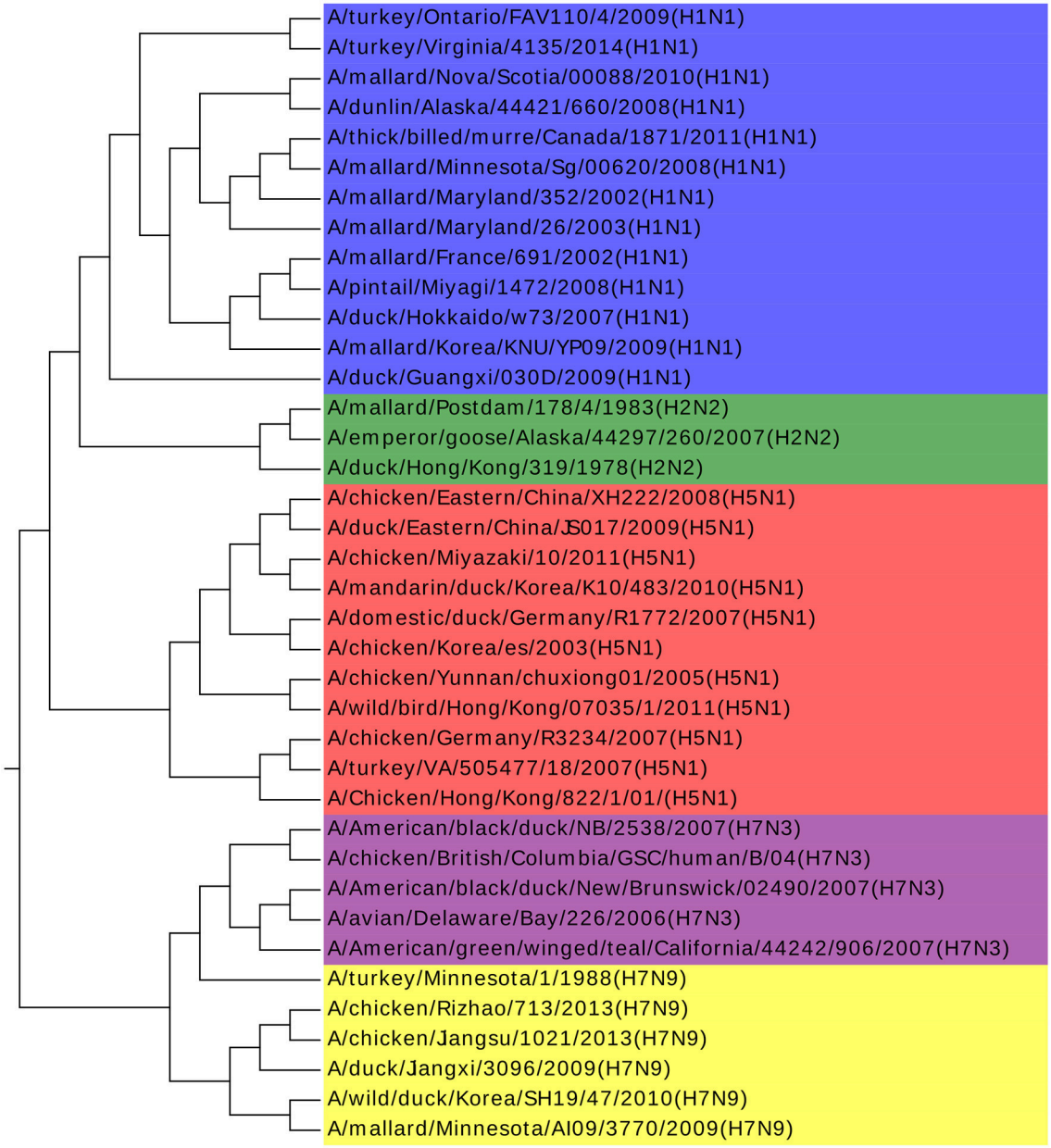

FIGURE 3 | The phylogenetic UPGMA tree using ANV method on Influenza A viruses dataset.

calculation of aligning them but alignment-free can still be done in a reasonable time. We also use a server to align segments of 351 sequences, to compare the results to ANV and other methods. ANV also gives the best performance on this dataset. Besides, we simulate another dataset of 20 sequences from a randomly generated sequence with length of 1,000bp, and test the phylogenetic trees from this and other methods.

We have chosen those datasets of different sizes (number of sequences, and lengths of sequences), which to test if ANV can be suitable in all cases. Most datasets have been analyzed by previous researches, therefore we can compare our results to others to evaluate the performances. Four datasets consist of viruses that are closely related to human health, and the mammal's dataset and simulated dataset show that this method can perform on other types of sequences as well.

Coronaviruses Dataset

Coronavirus belongs to the subfamily Coronavirinae in the family Coronaviridae, in the order Nidovirales. In this paper, we construct a dataset with 36 Coronaviruses, in which 34 viruses are from the exact same dataset with (Woo et al., 2005; Yu et al., 2010;

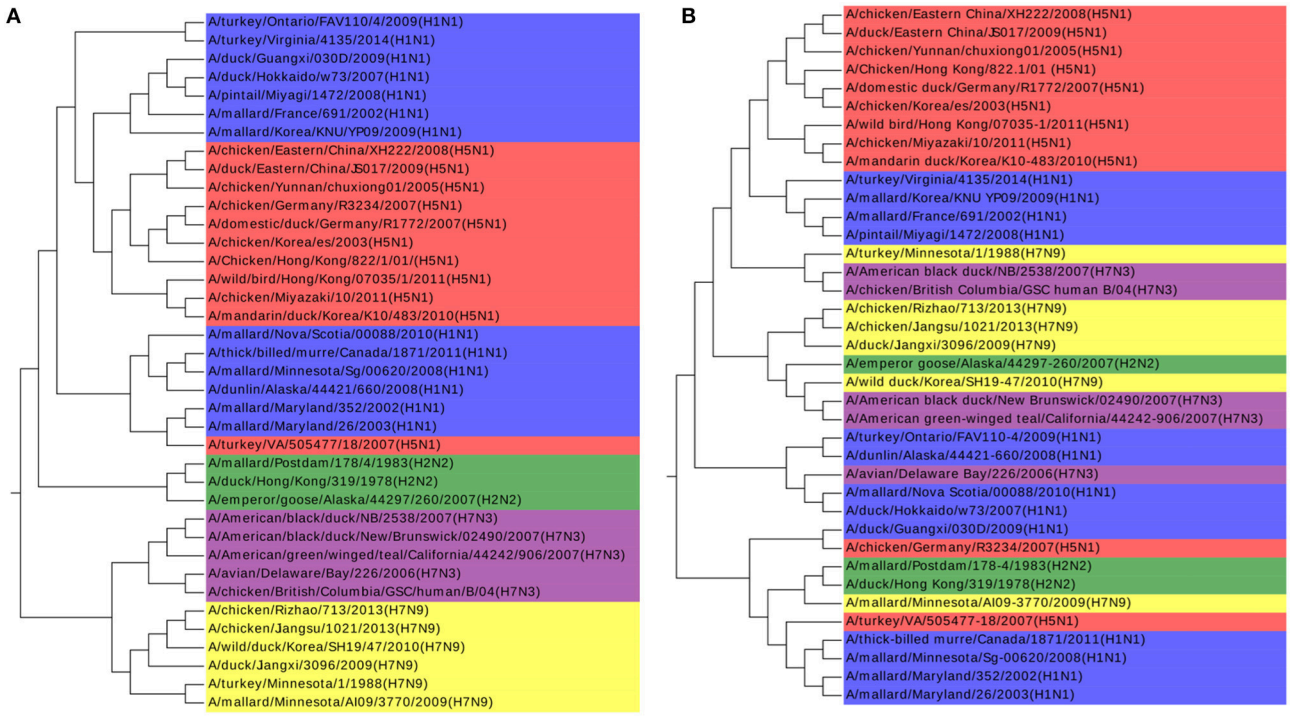

FIGURE 4 | (A) The phylogenetic UPGMA tree using the *k-mer* ($k = 5$) method on Influenza A viruses dataset. **(B)** The phylogenetic UPGMA tree using MSA (ClustalW) method on Influenza A viruses dataset.

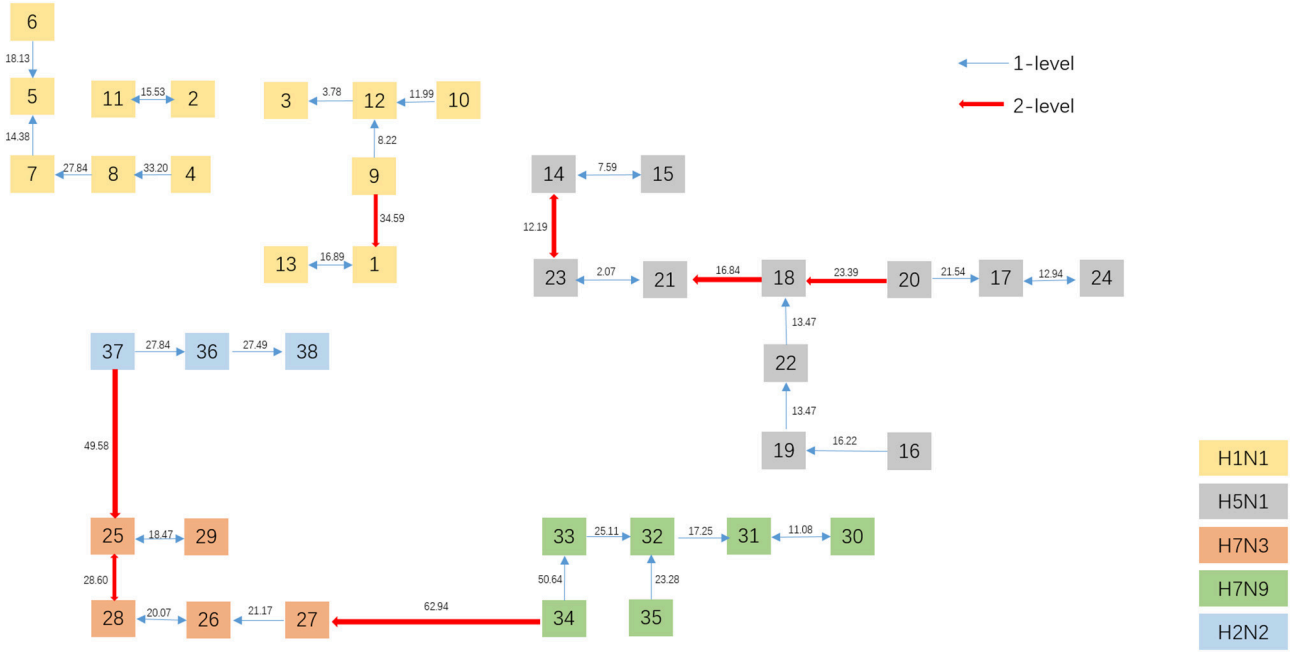

FIGURE 5 | The Natural Graph using ANV method on Influenza A virus dataset.

Hoang et al., 2015). The other two viruses are two new members in Coronavirus. Details of the Coronaviruses can be found in **Table S1**. The new ChinaGD01 (Lu et al., 2015) was identified in Guangdong Province (China) in 2015 and is an imported Middle East respiratory syndrome Coronavirus. The other one MERS-CoV/KOR is from South Korea (Kim et al., 2015). As of 15 June

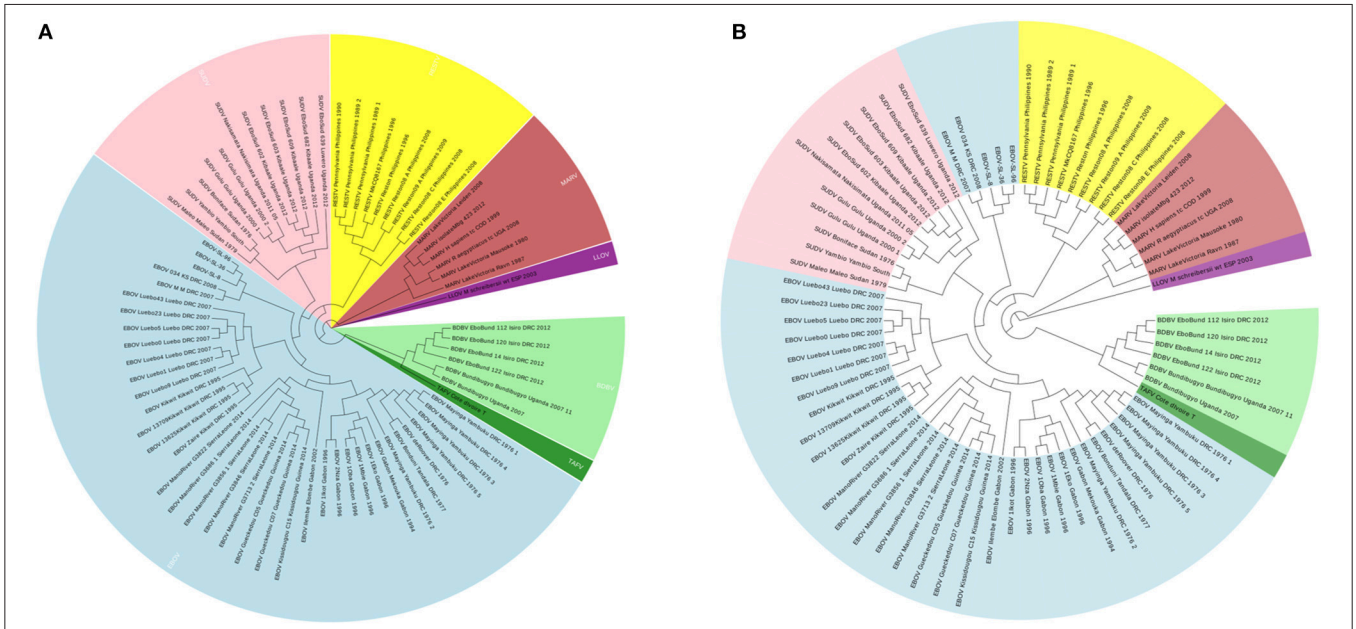

FIGURE 6 | (A) The phylogenetic UPGMA tree using ANV method on Ebolaviruses dataset. **(B)** The phylogenetic UPGMA tree using the traditional Natural Vector (NV) method on Ebolaviruses dataset.

TABLE 3 | Comparison of ANV and *k-mer* methods on 351 viruses dataset.

METHOD	ANV	6-MER	7-MER	8-MER	MSA (MUSCLE)
Family	94.87%	71.23%	25.36%	16.24%	72.08%
Genus	83.19%	65.24%	21.65%	12.25%	65.53%
Computing Time (seconds)	2466.73	4179.24	8636.13	24011.70	Unable to compute on laptop

2015, the MERS-COV was spreading in South Korea, and the ChinaGD01 case was a South Korean national who traveled to Guangdong in May 2015. Therefore, those two members were considered highly correlated with each other. The genomic size of Coronaviruses ranges from about 9 to 31 kbp, with the average of 27,567 nucleotides. Using our Accumulated Natural Vector and UPGMA method (Sneath and Sokal, 1973), we can build up a phylogenetic tree as shown in **Figure 1**.

Figure 1 shows that the two new members are clustered together with Group 4, which is also well-known as SARS (Severe Acute Respiratory Syndrome). Between November 2002 and July 2003, an outbreak of SARS in southern China caused an eventual 8,098 cases, resulting in 774 deaths reported in 37 countries. Both MERS-CoV and SARS viruses are beta-Coronaviruses, however, they belong to different lineages, for more details please see (Drexler et al., 2013; Hilgenfeld and Peiris, 2013). The phylogenetic tree indicates that the ChinaGD01 and MERS-CoV/KOR forms a monophyletic clade, sister to the SARS clade, which may possibly be a variant from some SARS viruses.

We also performed the same procedure with *k-mer* method on the Coronaviruses dataset. However, how to choose an optimal *k*-value is an interesting topic that requires manual intervention. Sims et al. showed in Woo et al. (2005) that the location of the peak in the distribution of k-mers, i.e., the k with the largest vocabulary, is related to the sequence length N. The k with maximum information is empirically determined but may be closely approximated by

$$k_{Hmax} = \log_4 N \tag{18}$$

where 4 is the alphabet size. They have shown in Sims et al. (2009) that reliable tree topologies are typically obtained with *k-mer* resolutions where $k > k_{Hmax}$ whereas lengths below $k_{H\max}$ yield unreliable trees. The upper limit of resolution can be empirically determined by a criterion that the tree topology for feature length k is equal to that of k+1, i.e., tree topologies converge.

According to this principle, we have $7 \leq k \leq 9$. We show the result of $k = 7$ in **Figure 2A**, and the results of $k = 8$ and $k = 9$ are in the **Figures S1**, **S2**. The four outgroup viruses cannot be clustered together as another branch from the tree of Coronaviruses, meanwhile the Group 1 was divided into smaller groups. The traditional ClustalW algorithm of Multiple Sequence Alignment (MSA) is also applied on the same dataset, and the result is shown in **Figure 2B**. MSA cannot cluster viruses from same groups together either. From this example, we can see that our ANV method is better than the *k-mer* and MSA method.

Influenza A viruses

Influenza A viruses are single-stranded RNA viruses, which have been a major health threat to both human society and animals (Hoang et al., 2015). Influenza A viruses' nomenclature is based on the surface glycoproteins: hemagglutinin (HA) and

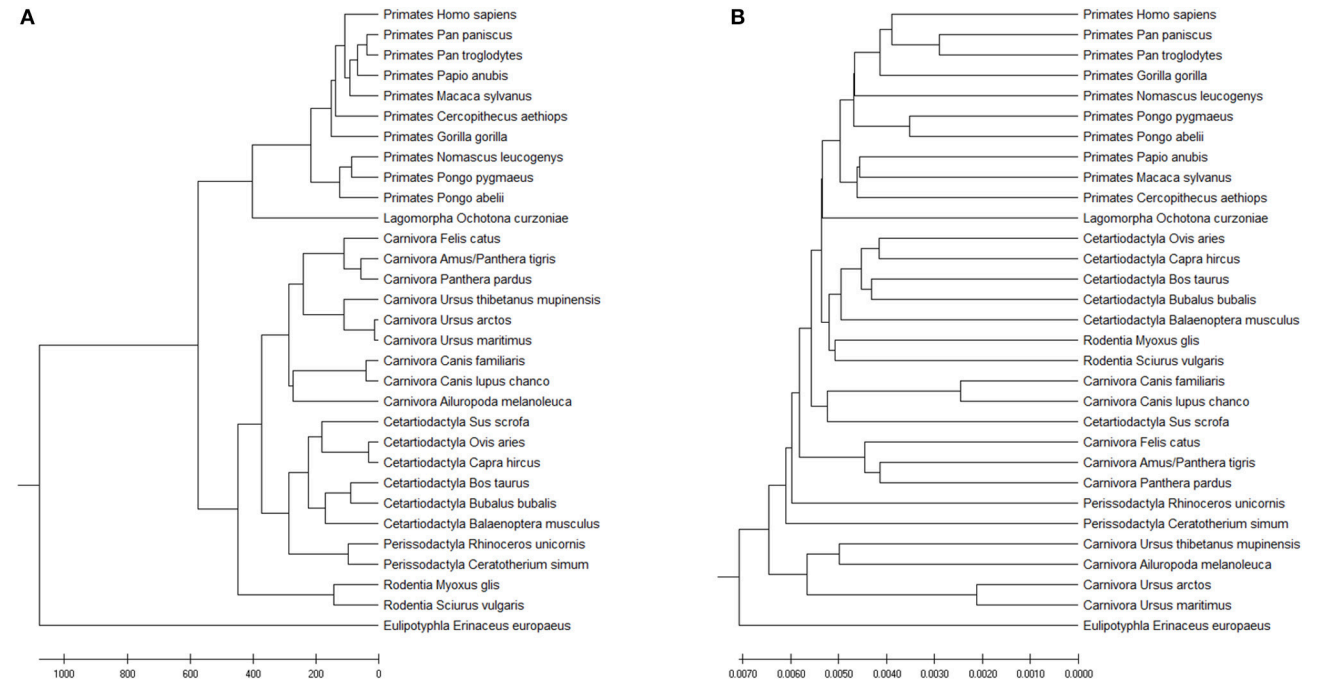

FIGURE 7 | (A) The phylogenetic UPGMA tree using ANV method on mammals mtDNA dataset. **(B)** The phylogenetic UPGMA tree using FFP (*k-mer*) method when K = 8 on mammals mtDNA dataset.

neuraminidase (NA) (Obenauer et al., 2006). HA has 15 subtypes and NA has 9 subtypes, which forms 135 different combinations. The NCBI number of the analyzed 38 Influenza A viruses can be found in **Table S2**. Our result agrees with previous work by Hoang et al. (2015). Furthermore, we find that all the Influenza A viruses are clustered with the same H and N type in **Figure 3**, with only one exception of A/turkey/Minnesota/1/1988(H7N9). There is no specific research about this virus and we infer that it may be the intermediate from H7N3 to H7N9. H7N3 had an outbreak in July 2012, causing millions of poultry's infection, but there is no report of infection from human to human yet. However, H7N9 was identified in Shanghai, China at the end of March 2013. Considering that the HA glycoprotein of those two subtypes are the same and the close outbreak date, we indicate that the H7N9 on March 2013 might be a variant from H7N3, and A/turkey/Minnesota/1/1988 (H7N9) plays a key role in this variation. We get the same conclusion in another work as well (Dong et al., 2018). More biological research on this virus should be done to deepen our understanding of Influenza A viruses to accelerate the invention of an effective vaccine and to prevent more dangerous variants.

The *k-mer* method and MSA are also performed on this dataset as shown in **Figures 4A,B**. The *k*-value is determined in the same procedure as in the Coronaviruses dataset as 5. In **Figure 4A**, the viruses from H1N1 and H5N1 are mixed up together with each other, while MSA has a worse result in **Figure 4B**. The results also indicate that *k-mer* and MSA cannot reveal the real relationships among the viruses.

To get a direct image of the relationships between Influenza A viruses, we draw the Natural Graph of them. Natural Graph was first introduced by Zheng et al. (2015). In **Figure 5**, the blue lines represent the 1-level connected components and the red ones 2-level. Classes are marked in different colors and it is obvious that after the construction of two levels, the Influenza A viruses with the same H and N are clustered together, including the A/turkey/Minnesota/1/1988(H7N9) which is Number 34 in **Figure 3**. H7N9 and H7N3 are clustered together in Level 2, indicating that they have a closer relationship, which accords with our previous conjecture.

72 Ebolaviruses Dataset

To illustrate that the new proposed ANV method is an important improvement of the traditional Natural Vector method, a 72 Ebolaviruses dataset is tested, which is a subset of the 163 viruses used in Zheng et al. (2015). It consists of 38 Ebola virus (EBOV), 11 Sudan virus (SUDV), 9 Reston virus (RESTV), 1 Taï Forest virus (TAFV), 6 Bundibugyo virus (BDBV), 6 Marburg virus (MARV) and 1 Lloviuvirus (LLOV). Details of this dataset are shown in **Table S3**. In **Figure 6A**, the phylogenetic tree shows that from the novel Accumulated Natural Vector method classifies all viruses into the right groups, however, in **Figure 6B**, the traditional Natural Vector method divides EBOV class into two clusters and SUDV is misclassified with some EBOV virus. This is an indication that including covariance between nucleotides helps improve the accuracy of classification. Hence this is an important improvement to the traditional Natural Vector and other alignment-free methods.

351 Viruses Dataset

We also test a large dataset of 351 viruses in Li et al. (2016), and the details of this dataset can be found in **Table S4**. The average length of them is 11,952 nucleotides and it makes alignment methods on a laptop impossible. Only server or cloud computing can finish such a task. Here we use 1-Nearest Neighbor (1-NN) method (Li et al., 2016) to see the accuracy of the prediction. This evaluation is inspired by the high rate of missing labels in many databases of viruses. For example, if a virus with missing family label has been added to the database, and it should share the same family label with the virus (stored in the database already) that is closest to it, then we can predict the missing family label according to the information of its nearest neighbor. Therefore, for a dataset with no missing labels, we can count how many viruses share the same label with its neighbor. "Nearest neighbor" of a specific virus can be defined as the virus that has the smallest Euclidean distance in the dataset to it for the alignment-free methods. For alignment results, we use the Hamming distance to measure the distance between two sequences. If the virus shares its distance with its neighbor, we consider it as a "correct" one, since even if its label is missing we can still predict it from its nearest neighbor. The accuracy can be computed by dividing the number of correct ones by the number of all viruses, in this case, by 351. We compare the result of ANV to the *k-mer* method since they are all alignment-free methods, and the results are shown in **Table 3**. The optimal choice of k is made by the same procedure in the other datasets. From **Table 3**, it is evident that ANV has much higher accuracy than the *k-mer* method, meanwhile using much less time. Thus, we have proved that ANV can apply to practical use with high time-efficiency and high-accuracy. For the alignment in this part, we tried to align all the sequences with full length on our server, but it fails to give a reliable result. Therefore, we extract 3,000 bp from the beginning and align 351 pieces of segments all with length of 3,000 bp. The results are shown in **Table 3** as well and the accuracy is still not as good as what ANV gives.

Mammals

Our Accumulated Natural Vector performs well not only on virus datasets, but also on other common species. We extract 31 mammalian mitochondrial genomes with the average length of 16,696 nucleotides, and the NCBI numbers of them can be found in **Table S5**. The genomes are from seven known clusters: Primates, Carnivora, Cetartiodactyla, Perissodactyla, Eulipotyphla, Lagomorpha, and Rodentia.

The Accumulated Natural Vector method can still distinguish the differences among the seven clusters, as shown in **Figure 7A**. FFP (*k-mer*) method has also been tested as well (the optimal *k*-value for this dataset is 8), as shown in **Figure 7B**. Since the species that includes in different paper are not all the same, it is hard to compare the whole topology of phylogenetic trees, however, our work still only has a small difference from the previous work in Murphy et al. (2001) and Tarver et al. (2016). The difference can be attributed to that mitochondrial genomes in mammals may not always reflect the organismal evolutionary history (Morgan et al., 2014), however, it still keeps more information than *k-mer* does in **Figure 7B**, since the distance within each group is smaller than the distances among groups, we can still distinguish clusters based on current dataset. In Ladoukakis and Zouros (2017), point out that most of the information researchers gained about the tree of life through the use of mtDNA remains valid, while we should pay more attention to its role in the function of the organism and its value as a tool in the study of major evolutionary novelties in the history of life. Therefore, the result implies that our ANV method can capture the key information hidden inside the DNA sequences and gives us a reliable topology among mammals.

TABLE 4 | Description of DNA sequence mutation in simulated tests.

Sequence Name	Description
A_original	200 point mutations from the randomly generated sequence with length 1,000 bp
A1	2 random nucleotide substitutions in A
A2	2 random nucleotide substitutions in A
A3	5 random nucleotide substitutions in A
A4	5 random nucleotide substitutions in A
A5	10 random nucleotide substitutions in A
A6	10 random nucleotide substitutions in A
B_original	200 point mutations from the randomly generated sequence with length 1,000 bp (different from A_original)
B1	2 random nucleotide substitutions in B_original
B2	2 random nucleotide substitutions in B_original
B3	5 random nucleotide substitutions in B_original
B4	5 random nucleotide substitutions in B_original
B5	10 random nucleotide substitutions in B_original
B6	10 random nucleotide substitutions in B_original
B7	10 bp Deletion from positions 51:60 in B_original
B8	10 bp Deletion from positions 601:610 in B_original
B9	20 bp Insertion at position 51 in B_original
B10	20 bp Insertion at position 601 in B_original
B11	50 bp Transposition from position 1 to 50 in B_original
B12	100 bp Transposition from position 601 to 700 in B_original

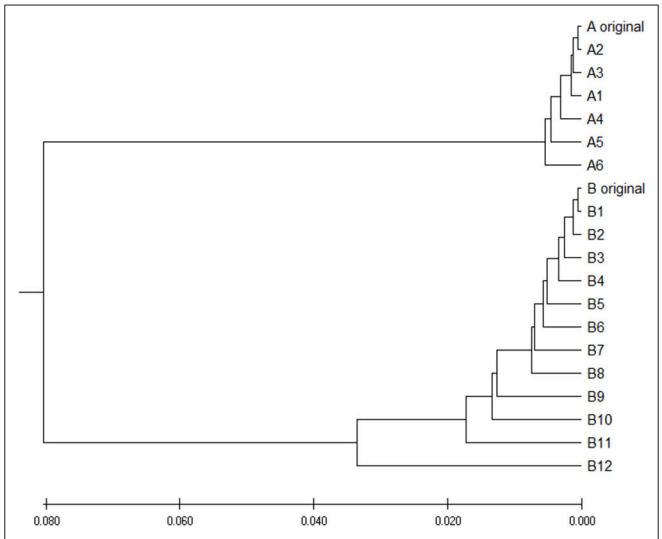

FIGURE 8 | The phylogenetic UPGMA tree using Jukes-Cantor pairwise alignment method on simulated dataset.

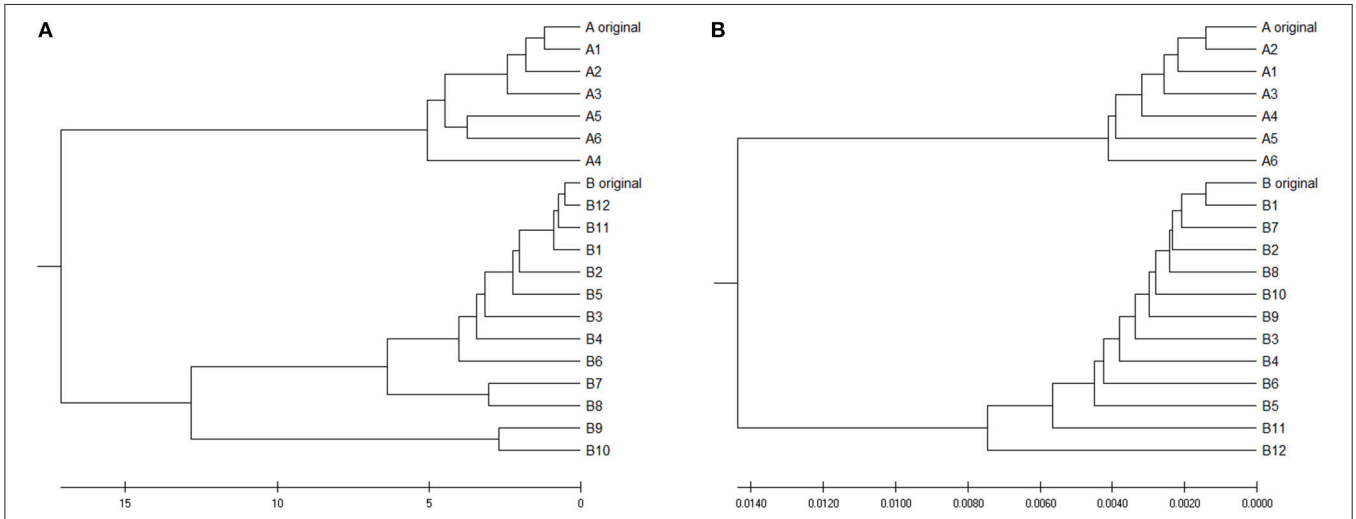

FIGURE 9 | (A) The phylogenetic UPGMA tree using ANV method on simulated dataset **(B)** The phylogenetic UPGMA tree using FFP (*k-mer*) method when K = 4 on simulated dataset.

TABLE 5 | Robinson-Foulds distances between trees by alignment-free methods and the reliable alignment tree.

Method	Alignment	ANV	4-mer	5-mer	6-mer
distance	0	23	27	29	29

Simulated Dataset

To verify is the similarity distance by our method can be used for clustering DNA sequences effectively, we also generated different mutations in DNA sequences and constructed phylogenetic trees by various methods. We simulated a sequence of length 1,000 bp as a base sequence, and generated two new sequences named "A_original" and "B_original" using point mutations. Both A and B have 100 nucleotides different from the original sequence. We then similarly evolved A and B into different mutants by four different mutations (substitutions, deletion, insertion, and transposition) as did in Yin et al. (2014). **Table 4** is the detailed description on the simulated DNA sequences with different mutations. Since the sequences are mutated slightly based on an exon sequence, we take the aligned result as the "correct" relationships among the sequences, and the alignment is done by the "seqpdist" function in MATLAB Bioinformatics toolbox. This function uses the classical Jukes-Cantor algorithm and we calculate the pairwise alignment distance. For comparison, we use the ANV method, FFP method (we test $k = 4,5,6$ in this case, since the lengths of sequences are about 1,000 bp). The UPGMA trees of alignment, ANV and FFP (k=4) methods are shown in **Figures 8**, **9A,B** separately. Among these trees, it is not very obvious which one is more similar to the alignment results, therefore we calculate the Robinson-Foulds distances between the distance matrix and the "correct" matrix and the results are shown in **Table 5**. Here we apply the program named "Robinson-Foulds" (Robinson and Foulds, 1981) when calculating **Table 5**. The simulated dataset is in **Table S6**. Actually, the differences among trees mainly lie in the branch of sequences generated from B, and ANV gives a more similar result, since the order is slightly disorganized by B5 and the transpositional sequences, while in **Figure 9B**, the whole branch of B is different from the alignment result.

DISCUSSION

In this paper, we propose a novel vector named Accumulated Natural Vector to analyze sequences, genomes and their phylogenetic relationships. Results from our analysis largely agree with the earlier studies, which indicates that our approach can detect the similarity and difference among sequences. Therefore, constructing phylogenetic trees only by sequence data could be done accurately in a very reasonable time, without using large computing platforms or conducting biological experiments of high cost. Our method can be applied in a global comparison of all genomes and provide a new powerful tool by including the correlations of nucleotides. We are working on extending the ANV method to protein sequences, nevertheless, for a protein sequence, it would produce an 1,830-dim vector for each sequence. The calculation cost for this is too large under the current technology. The covariance for three amino acids at a time may be more reasonable, since three consequent nucleotides can also become a codon in expression region of a sequence.

AUTHOR CONTRIBUTIONS

SS-TY and RH conceived the idea of covariance. RD implemented the idea and wrote the first draft of the manuscript. LH discussed and revised the first draft. RD, LH, RH, and SS-TY all contributed to the writing of the manuscript and agreed with the manuscript results and conclusions. They jointly developed the structure and arguments for the paper, made critical revisions and approved final version, and reviewed and approved the final manuscript.

FUNDING

This study is supported by the National Natural Science Foundation of China (91746119) (to SS-TY), Tsinghua University start-up fund (to SS-TY).

ACKNOWLEDGMENTS

The corresponding author would like to thank National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00234/full#supplementary-material

REFERENCES

Apostolico, A., and Denas, O. (2008). Fast algorithms for computing sequence distances by exhaustive substring composition. *Algorithm Mol. Biol.* 3:13. doi: 10.1186/1748-7188-3-13

Deng, M., Yu, C., Liang, Q., He, R. L., and Yau, S. S. T. (2011). A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS ONE* 6:e17293. doi: 10.1371/journal.pone.0017293

Dong, R., Zhu, Z., Yin, C., He, R. L., and Yau, S. S. T. (2018). A new method to cluster genomes based on cumulative Fourier power spectrum. *Gene* 673, 239–250. doi: 10.1016/j.gene.2018.06.042

Drexler, J. F., Corman, V. M., and Drosten, C. (2013). Ecology, evolution and classification of bat coronaviruses in the aftermath of SARS. *Antivir. Res.* 101, 45–56. doi: 10.1016/j.antiviral.2013.10.013

Hatje, K., and Kollmar, M. (2012). A phylogenetic analysis of the Brassicales clade on an alignmet-free sequence comparison method. *Front. Plant Sci.* 8:192. doi: 10.3389/fpls.2012.00192

Hilgenfeld, R., and Peiris, M. (2013). From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. *Antivir. Res.* 100, 286–295. doi: 10.1016/j.antiviral.2013.08.015

Hoang, T., Yin, C., and Yau, S. S. T. (2016). Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison. *Genomics* 108, 134–142. doi: 10.1016/j.ygeno.2016.08.002

Hoang, T., Yin, C., Zheng, H., Yu, C., He, R. L., and Yau, S. S. T. (2015). A new method to cluster DNA sequences using Fourier power spectrum. *J. Theor. Biol.* 372, 135–145. doi: 10.1016/j.jtbi.2015.02.026

Kim, Y. J., Cho, Y. J., Kim, D. W., Yang, J. S., Kim, H., Park, S., et al. (2015). Complete genome sequene of middle east respiratory syndrome Coronavirus KOR/KNIH/002_05_2016, isolated in South Korea. *Genome Announc.* 3, e00787–e00715. doi: 10.1128/genomeA.00787-15

Ladoukakis, E. D., and Zouros, E. (2017). Evolutionary and inheritance of animal mitochondrial DNA: rules and exceptions. *J. Biol. Res-Thessaloniki*. 24:2. doi: 10.1186/s40709-017-0060-4

Li, Y., Tian, K., Yin, C., He, R. L., and Yau, S. S. T. (2016). Virus classification in 60-dimensional protein space. *Mol. Phylogenet. Evol.* 99, 53–62. doi: 10.1016/j.ympev.2016.03.009

Lu, R., Wang, Y., Wang, W., Nie, K., Zhao, Y., Su, J., et al. (2015). Complete genome sequence of middle east respiratory syndrome Coronavirus (MERS-CoV) from the first imported MERS-CoV case in China. *Genome Announc.* 3, e00818–e00815. doi: 10.1128/genomeA.00818-15

Morgan, C. C., Creevey, C. J., and O'Connell, M. J. (2014). Mitochondrial data are not suitable for resolving placental mammals phylogeney. *Mamm Genome* 25, 636–647. doi: 10.1007/s00335-014-9544-9

Murphy, W., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A., and O'Brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature* 409, 614–618. doi: 10.1038/35054550

Obenauer, J. C., Denson, J., Mehta, P. K., Su, X., Mukatira, S., Finkelstein, D. B., et al. (2006). Large-scale sequence analysis of avian influenza isolates. *Science* 311, 1576–1580. doi: 10.1126/science.1121586

Robinson, D. F., and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147. doi: 10.1016/0025-5564(81)90043-2

Sims, G. E., Jun, S. R., Wu, G. A., and Kim, S. H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. U.S.A.* 106, 2677–2682. doi: 10.1073/pnas.0813249106

Sneath, P. H. A., and Sokal, R. R. (1973). *Numerical Taxonomy*. San Francisco: Freeman.

Tarver, J. E., Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'Reilly, J. E., et al. (2016). The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.* 8, 330–344. doi: 10.1093/gbe/evv261

Woo, P. C. Y., Lau, S. K. P., Chu, C., Chan, K., Tsoi, H., Huang, Y., et al. (2005). Characterization and complete genome sequence of a novel Coronavirus, Coronavirus HKU1, from patients with pneumonia. *J. Virol.* 79, 884–895. doi: 10.1128/JVI.79.2.884-895.2005

Yin, C., Chen, Y., and Yau, S S. T. (2014). A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. *J. Theor. Biol.* 359, 18–28. doi: 10.1016/j.jtbi.2014.05.043

Yu, C., He, R. L., and Yau, S. S. T. (2013a). Protein sequence comparison based on K-string dictionary. *Gene* 529, 250–256. doi: 10.1016/j.gene.2013.07.092

Yu, C., Hernandez, T., Zheng, H., Yau, S. C., Huang, H. H., He, R. L., et al. (2013b). Real time classification of viruses in 12 dimensions. *PLoS ONE* 8:e64328. doi: 10.1371/journal.pone.0064328

Yu, C., Liang, Q., Yin, C., He, R. L., and Yau, S. S. T. (2010). A novel construction of genome space with biological geometry. *DNA Res.* 17, 155–168. doi: 10.1093/dnares/dsq008

Zheng, H., Yin, C., Hoang, T., He, R. L., Yang, J., and Yau, S. S. T. (2015). Ebolavirus classification based on natural vectors. *DNA Cell Biol.* 34, 418–428. doi: 10.1089/dna.2014.2678

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.