SCIENTIFIC REPORTS

OPEN

A novel fast vector method for genetic sequence comparison

Yongkun Li[1], Lily He[1], Rong Lucy He[2] & Stephen S.-T. Yau[1]

Received: 23 March 2017
Accepted: 7 September 2017
Published online: 22 September 2017

With sharp increasing in biological sequences, the traditional sequence alignment methods become unsuitable and infeasible. It motivates a surge of fast alignment-free techniques for sequence analysis. Among these methods, many sorts of feature vector methods are established and applied to reconstruction of species phylogeny. The vectors basically consist of some typical numerical features for certain biological problems. The features may come from the primary sequences, secondary or three dimensional structures of macromolecules. In this study, we propose a novel numerical vector based on only primary sequences of organism to build their phylogeny. Three chemical and physical properties of primary sequences: purine, pyrimidine and keto are also incorporated to the vector. Using each property, we convert the nucleotide sequence into a new sequence consisting of only two kinds of letters. Therefore, three sequences are constructed according to the three properties. For each letter of each sequence we calculate the number of the letter, the average position of the letter and the variation of the position of the letter appearing in the sequence. Tested on several datasets related to mammals, viruses and bacteria, this new tool is fast in speed and accurate for inferring the phylogeny of organisms.

In the past few decades, a basic but crucial task in biology is comparing genes and proteins for predicting the structure and function of these biological sequences. Many methods have been proposed to compare genetic sequences. Traditionally, most of these approaches are the widely used alignment-based methods. In these methods, molecular sequences are optimally aligned based on selected scoring systems. The alignment-based methods often give high accuracy and may reveal the relationships among sequences. Some algorithms have been established and incorporated into softwares for sequence alignments[1,2]. However, one of the main drawback of these techniques is that they are very time-consuming and expensive in memory usage. As a result, alignment-free approaches have attracted more and more attention and have been applied to biological sequence comparison as well as phylogeny analysis recently[3–7]. The noticeable characteristic of these methods is to incorporate some *ad hoc* numerical feature of sequences. These methods are fast in computational speed when comparing genes and proteins. Generally, alignment-free are categorized into four groups: methods based on k-mer or word frequency, methods based on substrings, methods based on information theory, and methods based on graphical representation. Firstly, methods based k-mer or word frequency are very popular and studied extensively. The classic k-mer method was proposed to compare genetic sequences, in which it counts the frequencies of substrings with k letters appearing in respective sequences[8]. Afterwards, a lot of k-mer based methods have been developed and applied in sequence analysis and phylogeny for viruses and bacteria. Such as feature frequency profile (FFP)[9] which counts the number of substrings with fixed length k occurring in a genome[9] and a vector consisting of these numbers is formed for each genome, return time distribution (RTD)[10], frequency chaos game representation (FCGR)[11], an improved complete composition vector method (ICCV)[12] which was proposed by optimizing composition vector (CV)[13] and complete composition vector (CCV)[14]. ICCV is under the assumption of a uniform and independent model to estimate sequence information. Comparing with CCV and CV method, ICCV method is more robust and efficient in performing sequence comparison. Secondly, methods based on substrings employ the similarity of substrings in a pair of sequences. These algorithms are mostly used for string processing in computer science, like average common substring (ACS)[15], k-mismatch average common substring approach (kmacs)[16]. Thirdly, methods based on information theory include global and local characterization of sequences and estimate genome entrop to regions. Among them base-base correlation (BBC)[17], information correlation and partial information correlation (IC-PIC)[18] are representative articles. Finally, methods based on graphical representation also play important roles in sequence comparison in alignment-free area[19–21].

[1]Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, PR China. [2]Department of Biological Sciences, Chicago State University, Chicago, IL60628, USA. Correspondence and requests for materials should be addressed to S.S.-T.Y. (email: yau@uic.edu)

Although k-mer models are widely used in many biological studies, the position information and the properties of nucleotides is ignored in these methods. The second kind of methods are computationally expensive among alignment-free methods. For the third algorithms, most of them are alignment based. The graph based methods are often used to display some characteristics of sequences directly by curves. Therefore, it is better to use these information for sequences comparison. In this paper, we develop an 18 dimensional feature vector to characterize a DNA sequence. The vector includes the occurrence frequency of each of the four bases, the average position of nucleotides and the biochemical properties of nucleotides. To validate the advantages of our approach, we test it on several data sets and further make comparison with some state-of-the-art sequence analysis techniques. Our method performs as good as alignment methods in accuracy and spends far less time than them. For one viral dataset, it even provides better result than alignment methods. It also works better than k-mer based methods in accuracy.

Results

To test the effectiveness of our new method, we employ it to different datasets, which include mammalian mitochondrial genomes, virus, and bacteria genomes. Some datasets are small sized and others are medium sized. The length of sequences ranges from ten thousands to several millions base pairs. For each dataset, using our method the multiple encoding vectors of its genetic sequences are calculated. To verify the efficiency of our algorithms, the FFP method is used to compare. We use MEGA to draw the phylogeny of datasets for the two methods. Unlike FFP which applies Jensen-Shannon Divergence to compute the pairwise distance between any two vectors, our new method use Euclidean distance. In the process of using FFP method, we choose k to be the minimum integer of $log_4(n)$, i.e. $k = floor(log_4(n))$ (n is the minimum length of sequences studied). From the results we can see the efficiency of our new method is better than FFP on these datasets.

Phylogeny of mammals. The first benchmark dataset contains 41 mammalian complete mitochondrial genomes (mtDNA) with about 16500 base pairs. In these mammals, the structure of mtDNA is circular and double-stranded. One strand of each mtDNA has rich guanine referred to as the heavy strand and the other strand includes rich cytosine referred to as the light strand. In our study, the heavy strands of mtDNA are chosen. These sequences are not highly conserved and have a fast mutation rate[22]. Using our proposed method along with UPGMA algorithm, the phylogenetic tree of 41 mitochondrial genomes is constructed. As displayed in Fig. 1, the 41 species are correctly grouped into 8 clusters: *Primates* (red), *Cetacea* (green), *Artiodactyla* (pink), *Perissodactyla* (light green), *Rodentia* (black), *Lagomorpha* (dark red), *Carnivore* (blue), and *Erinaceomorpha* (grey). According to the FFP method, length of substrings is chosen to be 7 and thus the phylogeny of the 41 species is obtained. The phylogenetic tree is provided in the supplementary file. As displayed in Fig. S1, the 8 clusters are not classified well. The four species in *Perissodactyla* are improperly clustered into two clades. *Indus River Dolphin* from *Cetacea* is separated from other species from *Cetacea*. The *Carnivore*, *Primates* and *Artiodactyla* clades are all mistakenly divided into more than one group. Obviously, for this dataset our new method gets a greater result.

Phylogeny of influenza A viruses. Influenza A viruses may cause influenza disease in birds and mammals[23]. Influenza A viruses are single-stranded, negative-sense, segmented RNA viruses. The genome of flu virus contains around 13500 nucleotides and is divided into 8 segments. The subtypes of Influenza A viruses are classified according to an H number for the type of hemagglutinin (H) and an N number for the type of neuraminidase (N) in viral surfaces[24]. The viruses have 18 different H serotypes (H1-H18) and 11 different N (N1-N11) serotypes. For instance, the recently appearing H7N3 is an Influenza A strain with a type 7 hemagglutinin (H) and a type 3 neuraminidase (N). The several subtypes of Influenza A viruses are able to infect many species including wild birds, swines, dogs, horses, and even humans. It is known that H1N1, H1N2, and H3N2 subtypes can circulate among humans. In this study, we utilize Segment 6 gene encoding N (neuraminidase) of Influenza A virus to perform phylogenetic analysis. Using our method, the 38 flu viruses are correctly clustered into five groups as show in Fig. 2. Our result is consistent with previous study[25]. According to the FFP method, the choice for length of k-mer of these viruses is 5 and their phylogenetic tree is constructed. As shown in Fig. S2 (in the supplementary file), however, some of the H1N1 viruses and some H5N1 viruses are placed together. Therefor, on this dataset our method is also superior than FFP.

Phylogeny of Human rhinovirus. To test our proposed method, we perform the phylogeny of 116 Human rhinoviruses (HRVs). HRVs are the major cause of the common colds and lead to more than one-half of cold-like diseases. HRVs are single-stranded positive sense RNA viruses of about 7200 nucleotides in length. The genome of Human rhinovirus contains a single gene encoding one polyprotein, which is then cleaved to yield 11 proteins. HRVs belong to *Enterovirus* genus and *Picornaviridae* family. They form three phylogenetically distinct clusters: HRV-A, HRV-B, HRV-C[4,26]. In the previous work, 113 complete genomes composed of the three clusters are used to explore their evolutionary relationships. In addition, 3 HEV-C genomes were also added to the dataset as outgroup. Based on multiple sequence alignment (MSA), the evaluation was very time-consuming although the classification performed well. Unlike the MSA, our new approach is very fast to finish conversion from the genome sequences to 18 dimensional vectors. As shown in Fig. 3, the three groups of 116 genomes are identified and the remaining 3 outgroup viruses are clustered together alone. According to the FFP method, the length of words chosen for this dataset is 6 and the phylogenetic tree of 116 HRVs is obtained. As shown in Fig. S3 (in the supplementary file), two HRV-A viruses are mixed with the outlying HEV-C viruses.

Phylogeny of ebolavirus. The *Ebolavirus* genus includes five species: Ebola virus (formerly Zaire ebolavirus, EBOV), Sudan virus (SUDV), Bundibugyo virus (BDBV), Tai Forest virus (TAFV), and Reston virus (RESTV)[27]. Ebola viruses are single-strand negative sense RNA viruses whose genomes consist of about 18900

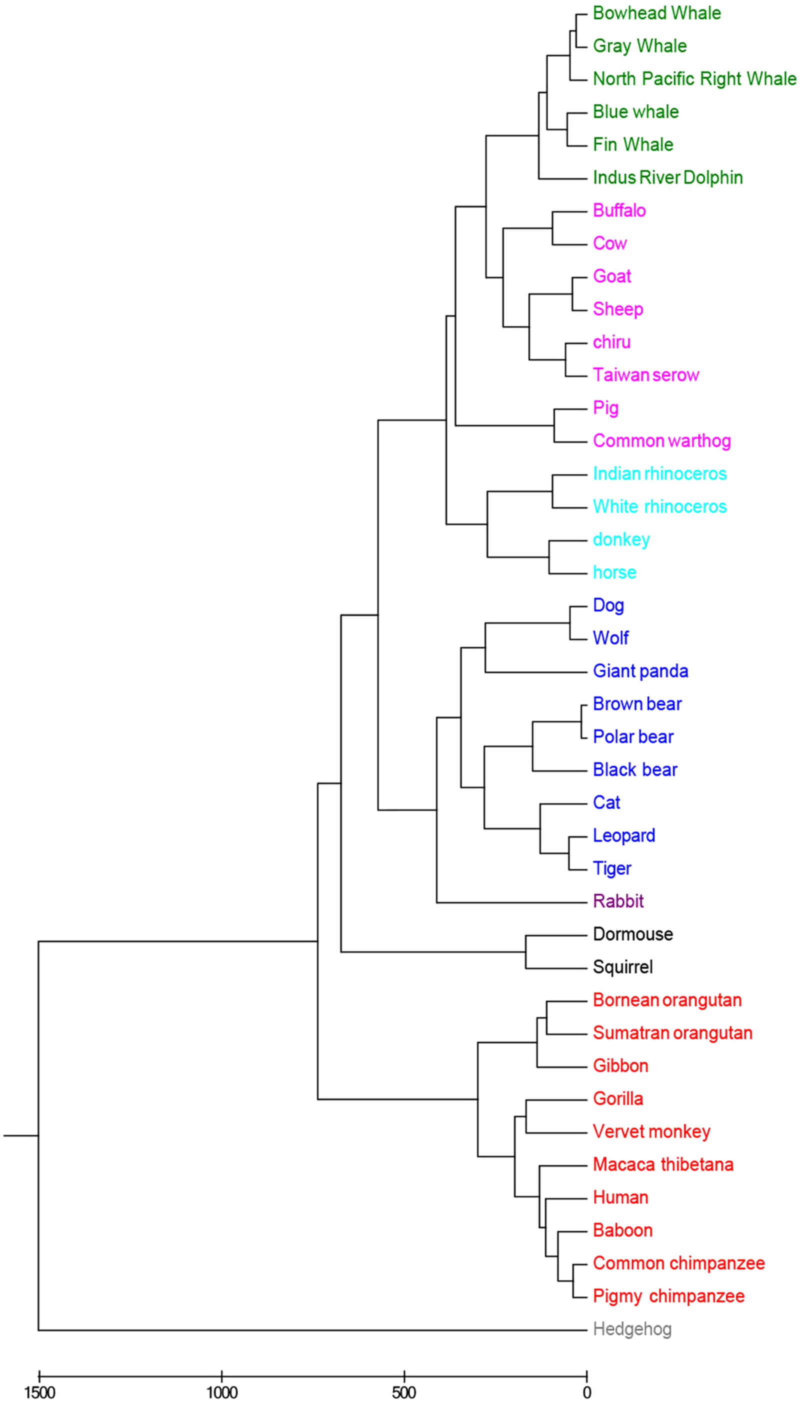

Figure 1. Phylogenetic tree of 41 mitochondrial genome sequences based on our multiple encoding vector method. The 41 species are correctly grouped into 8 clusters: *Primates* (red), *Cetacea* (green), *Artiodactyla* (pink), *Perissodactyla* (light green), *Rodentia* (black), *Lagomorpha* (dark red), *Carnivore* (blue), and *Erinaceomorpha* (grey).

bases. Each ebolavirus genome encodes seven proteins in which glycoprotein is the only viral protein on the surface of ebolavirus. The first case of human infected by EBOV was reported in 1976 in Zaire (currently the Democratic Republic of the Congo (DRC))[28]. Recently the epidemic starting in 2014 in Guinea is unprecedent because it is the first outbreak in West Africa and the magnitude is largest among known outbreaks.

To understand EBOV epidemics in humans and their relationships with other viruses in genus *Ebolavirus*, we apply the novel numerical vector method to construction of the phylogenetic tree of 59 viruses in *Ebolavirus* genus. As shown in Fig. 4, the five species are exactly separated. The EBOV strains from the confirmed pandemics form a lineage independent of the other four species in genus *Ebolavirus*. The EBOV strains in the 2014 epidemic in Guinea are clustered together. The EBOV strains in Zaire (DRC) pandemic in 1976–1977 are placed together as

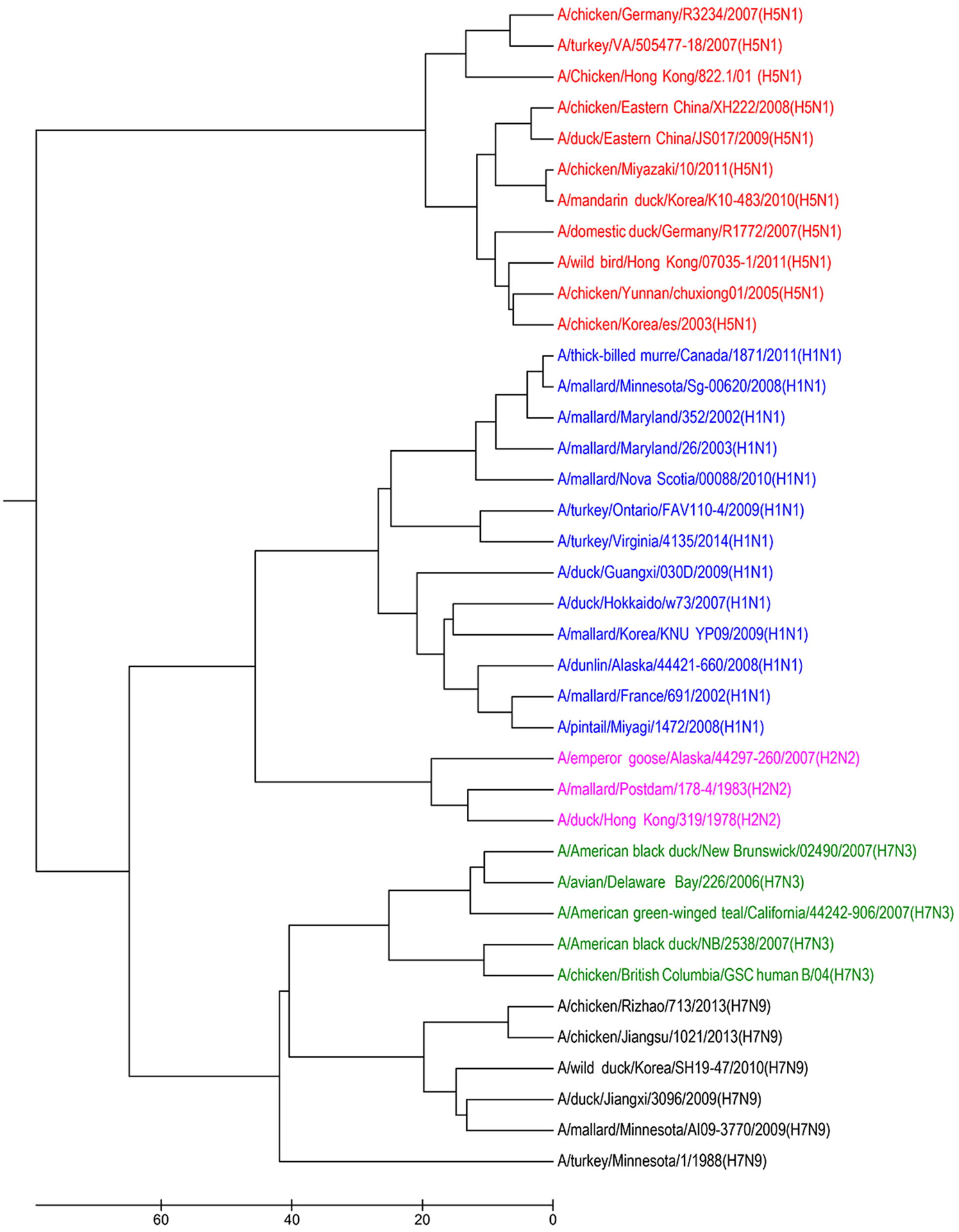
A/chicken/Germany/R3234/2007(H5N1)
A/turkey/VA/505477-18/2007(H5N1)
A/Chicken/Hong Kong/822.1/01 (H5N1)
A/chicken/Eastern China/XH222/2008(H5N1)
A/duck/Eastern China/JS017/2009(H5N1)
A/chicken/Miyazaki/10/2011(H5N1)
A/mandarin duck/Korea/K10-483/2010(H5N1)
A/domestic duck/Germany/R1772/2007(H5N1)
A/wild bird/Hong Kong/07035-1/2011(H5N1)
A/chicken/Yunnan/chuxiong01/2005(H5N1)
A/chicken/Korea/es/2003(H5N1)
A/thick-billed murre/Canada/1871/2011(H1N1)
A/mallard/Minnesota/Sg-00620/2008(H1N1)
A/mallard/Maryland/352/2002(H1N1)
A/mallard/Maryland/26/2003(H1N1)
A/mallard/Nova Scotia/00088/2010(H1N1)
A/turkey/Ontario/FAV110-4/2009(H1N1)
A/turkey/Virginia/4135/2014(H1N1)
A/duck/Guangxi/030D/2009(H1N1)
A/duck/Hokkaido/w73/2007(H1N1)
A/mallard/Korea/KNU YP09/2009(H1N1)
A/dunlin/Alaska/44421-660/2008(H1N1)
A/mallard/France/691/2002(H1N1)
A/pintail/Miyagi/1472/2008(H1N1)
A/emperor goose/Alaska/44297-260/2007(H2N2)
A/mallard/Postdam/178-4/1983(H2N2)
A/duck/Hong Kong/319/1978(H2N2)
A/American black duck/New Brunswick/02490/2007(H7N3)
A/avian/Delaware Bay/226/2006(H7N3)
A/American green-winged teal/California/44242-906/2007(H7N3)
A/American black duck/NB/2538/2007(H7N3)
A/chicken/British Columbia/GSC human B/04(H7N3)
A/chicken/Rizhao/713/2013(H7N9)
A/chicken/Jiangsu/1021/2013(H7N9)
A/wild duck/Korea/SH19-47/2010(H7N9)
A/duck/Jiangxi/3096/2009(H7N9)
A/mallard/Minnesota/AI09-3770/2009(H7N9)
A/turkey/Minnesota/1/1988(H7N9)
60
40
20
0

Figure 2. The UPGMA phylogenetic tree of 38 influenza A viruses based on our multiple encoding vector method.

a clade. The three EBOV strains from the 1995 outbreak in Zaire (DRC) form a clade. The EBOV strains in DRC pandemic in 2007 are clustered together. The strains from the two epidemics occurred in Gabon are divided into two different clusters according to the outbreak periods. The SUDV branch is clustered with the RESTV branch. BDBV and TAFV viruses are positioned together. Our results are consistent with those in previous work[27,29]. Since the viral genomes in this dataset have about 19 kilobases, the k-mer length is thus 7. The phylogenetic relationships among these viruses using FFP method are depicted in Fig. S4 (in the supplementary file). However, the SUDV branch is not clustered with RESTV branch in this figure, which is not consistent with the confirmed taxonomy of *Ebolavirus* genus.

Phylogeny of coronavirus. Coronaviruses are enveloped, single-stranded, positive-sense RNA viruses within the family *Coronaviridae*. Their genomes in length range from 25,000 to 32,000 nucleotides. Some subtypes of coronaviruses can also infect humans and cause many respiratory and gastrointestinal diseases with highly variable severity. Recently, the emergent Middle East respiratory syndrome coronavirus (MERS-CoV) had high case-fatality rate and caused outbreak of severe respiratory disease in Arabian Peninsula. As a result of pandemics from coronaviruses especially the SARS, the classification and evolutionary relationships among these viruses have been extensively investigated.

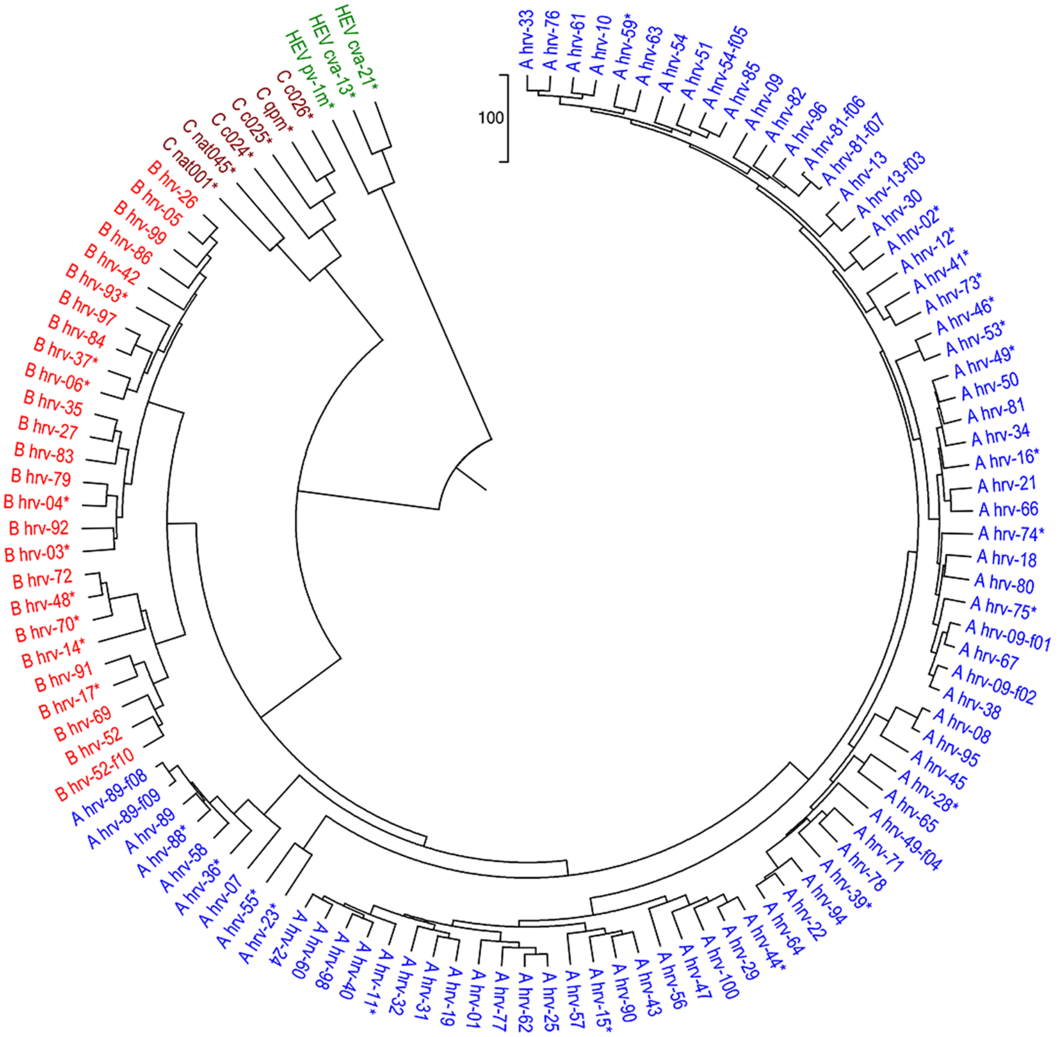

Figure 3. The UPGMA phylogenetic tree of 113 human rhinoviruses and 3 HEV-C viruses based on our multiple encoding vector method.

We apply our 18 dimensional vector method to a widely studied dataset containing 30 coronaviruses and 4 extra non-coronaviruses as outgroup. The newly discovered human coronavirus NL63 was clustered together with human coronavirus 229E (group 1)[30]. At the same time, it was positioned away from HCoV-OC43 (group 2) and SARS (group 4). The new isolated HCoV-HKU1 was thought as a distinct member of group 2[31]. However, the HCoV-HKU1 was designated as a member of a new group 5[32]. Using our novel method, the phylogeny of the coronaviruses is reconstructed based on the UPGMA algorithm. As shown in Fig. 5, HCoV-NL63, and HCoV-229E are grouped into group 1 as validated in the previous research[33]. Group 2, group 4, and new group 5 are separated in the tree, which is consistent with the studies in published papers[31,32]. According to the FFP method, the length of substrings chosen for these organisms is 6 and the phylogeny is constructed. As displayed in Fig. S5 (in the supplementary file) the HCoV-229E virus from group 1 is not correctly classified into group 2. Besides, the four viruses as outgroup are not clustered together. It means for this dataset our new method still has advantage over FFP.

Phylogeny of bacteria. Bacteria as one kind of prokaryotes are abundant on the earth. The phylogeny of these species has been an crucial topic in biology. As a result, the classification of bacteria based on whole genome sequences have attracted more and more attention. Due to long genomic sequences of bacteria with more than 1 million base pairs (Mb), the tradition multiple sequence alignment methods are computationally infeasible to perform phylogeny. To test our model, we analyze a dataset consisting of 59 bacterial species. This set contains 15 families: *Aeromonadaceae*, *Alcaligenaceae*, *Bacilleceae*, *Borreliaceae*, *Burkholderiaceae*, *Caulobacteraceae*, *Clostridiaceae*, *Desulfovibrionaceae*, *Enterobacteriaceae*, *Erwiniaceae*, *Lactobacillaceae*, *Mycoplasmataceae*, *Rhodobacteraceae*, *Staphylococcaceae*, *Yersiniaceae*. The genome length in this dataset mainly range from 3 to 10 Mb. As displayed in Fig. 6, the 15 families are well separated using our method. Based on the 9-mer FFP method, these families are also separated as illustrated in Fig. S6 (in the supplementary file). The two results are same for the phylogeny of 59 bacteria at the family level. However, our tree has some advantages at the phylum level. In Fig. 6, the three families *Lactobacillaceae*, *Clostridiaceae* and *Staphylococcaceae* from *phylum Bacilli* are

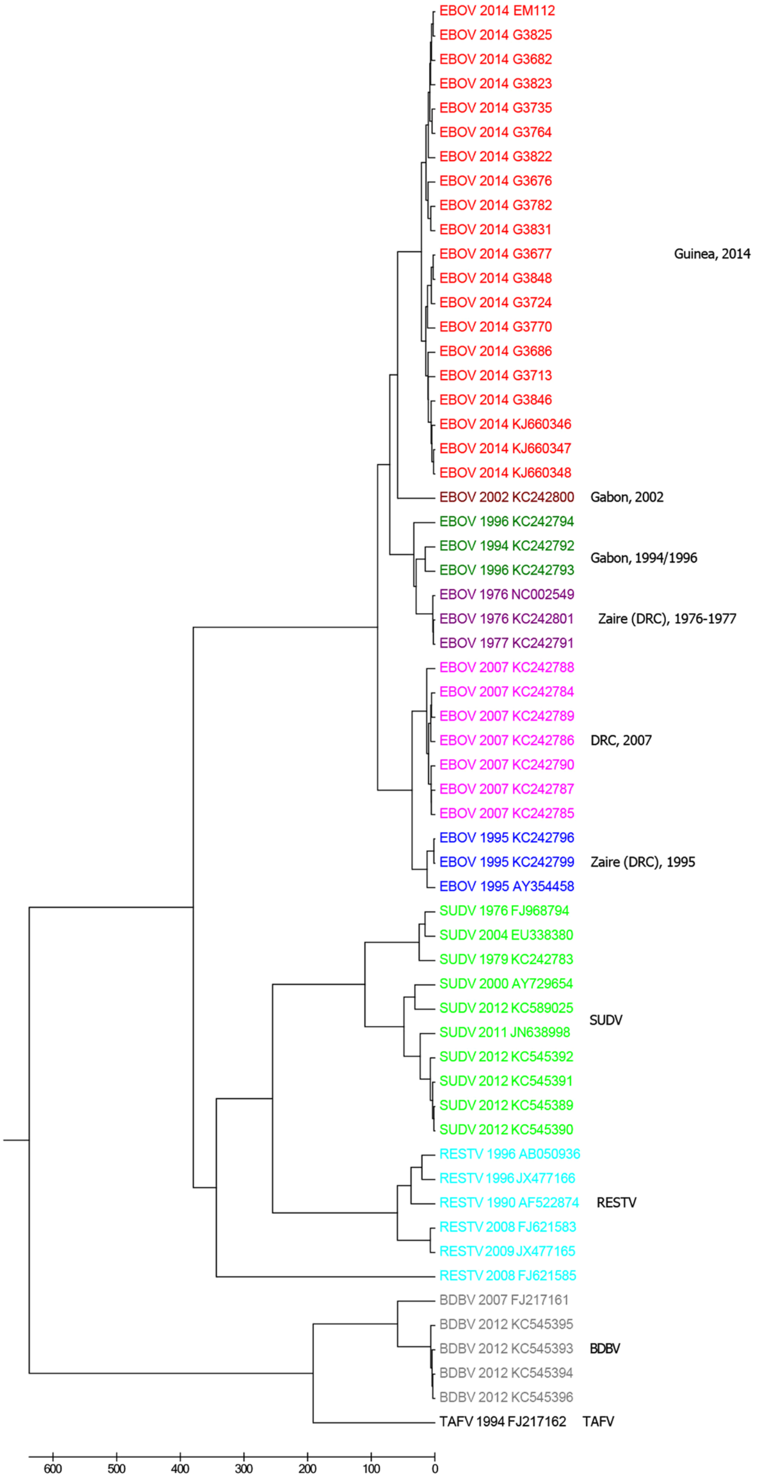

Figure 4. The UPGMA phylogenetic tree of 59 viruses in *Ebolavirus* genus based on our multiple encoding vector method. DRC = Democratic Republic of Congo, EBOV = Ebola virus, SUDV = Sudan virus, BDBV = Bundibugyo virus, TAFV = Tai Forest virus, RESTV = Reston virus.

clustered together. In the tree constructed by FFP model (Fig. S6), however, *Lactobacillaceae* are placed near to *Mycoplasmataceae* from phylum *Tenericutes*. Family *Desulfovibrionaceae* from phylum *Proteobacteria* are positioned in *phylum Bacilli*. Moreover, family *Clostridiaceae* from *phylum Bacilli* are grouped with *Borreliaceae* from *pylum Spirochaetes*. For the running time of algorithms, it is very time-consuming to calculate feature frequency profiles vectors with $4^9 = 262144$ dimensions. Since our vectors have 18 features, it takes only about 5 minutes to get these vectors.

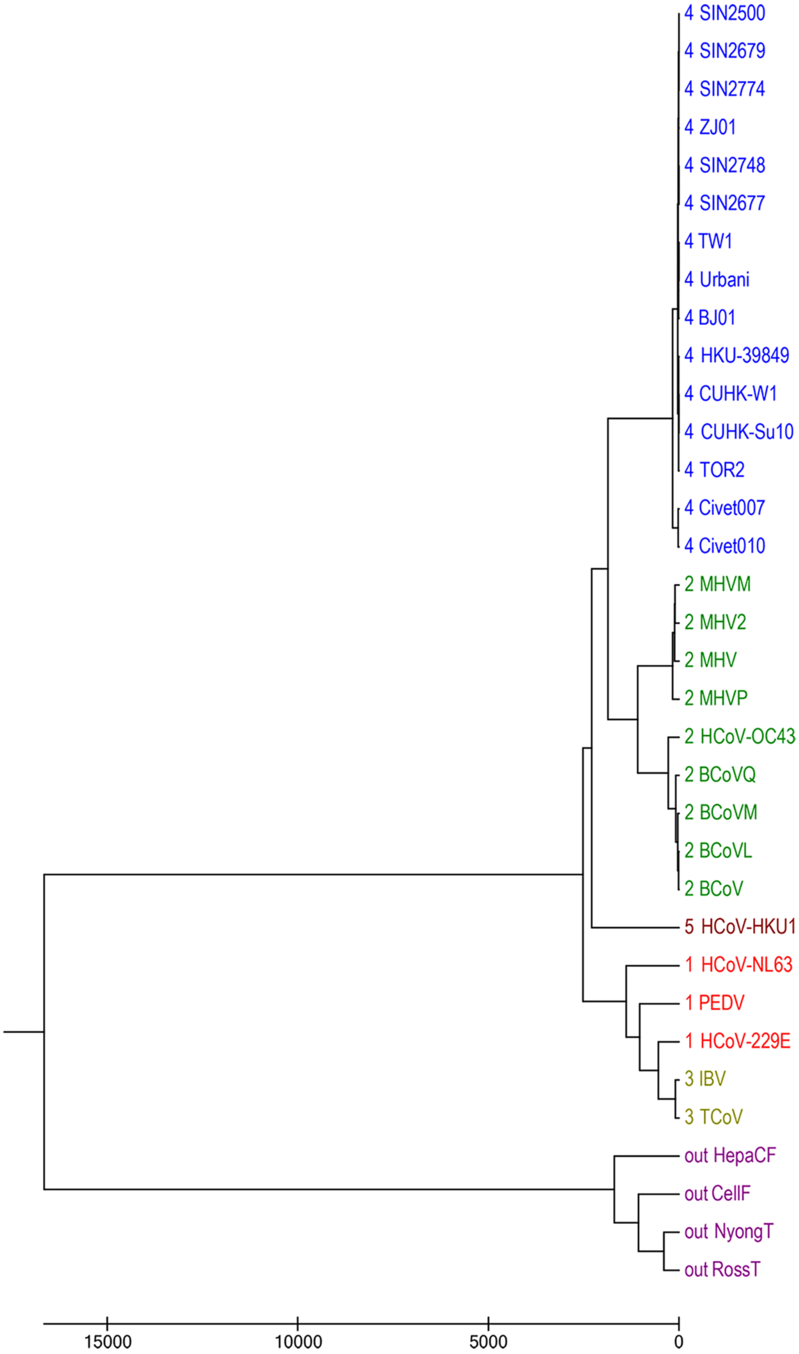

Figure 5. The UPGMA phylogenetic tree of 30 coronaviruses and 4 non-coronaviruses based on our multiple encoding vector method.

Discussion

Sequence comparison is crucial to understand the evolutionary relationships among organisms. Methods with alignment usually perform well if the genetic sequences are homologous. However, results obtained with alignment may be unreliable due to genetic recombination and high mutation rates. Besides, these approaches are time-consuming and thus are not suitable to align long sequences and large volume of data. According to the results mentioned above, our proposed alignment-free method shows its advantages on accuracy over alignment-based methods. Compared to a state-of-the-art method based on k-mer, the new technique still has the merit to construct correct evolutionary relationships. Our method is applicable to phylogeny construction of many kinds of species including mammals, viruses, and bacteria.

Additionally, we also compare our tool to ClustalW algorithm[2] which is one of the most popular alignment methods. All programs are implemented on a laptop with 8 GB RAM, Intel (R) Core (TM) i7-4500U CPU, and 1.80 GHz and the running time unit is minute. As listed in Table 1, our approach is very fast in computational

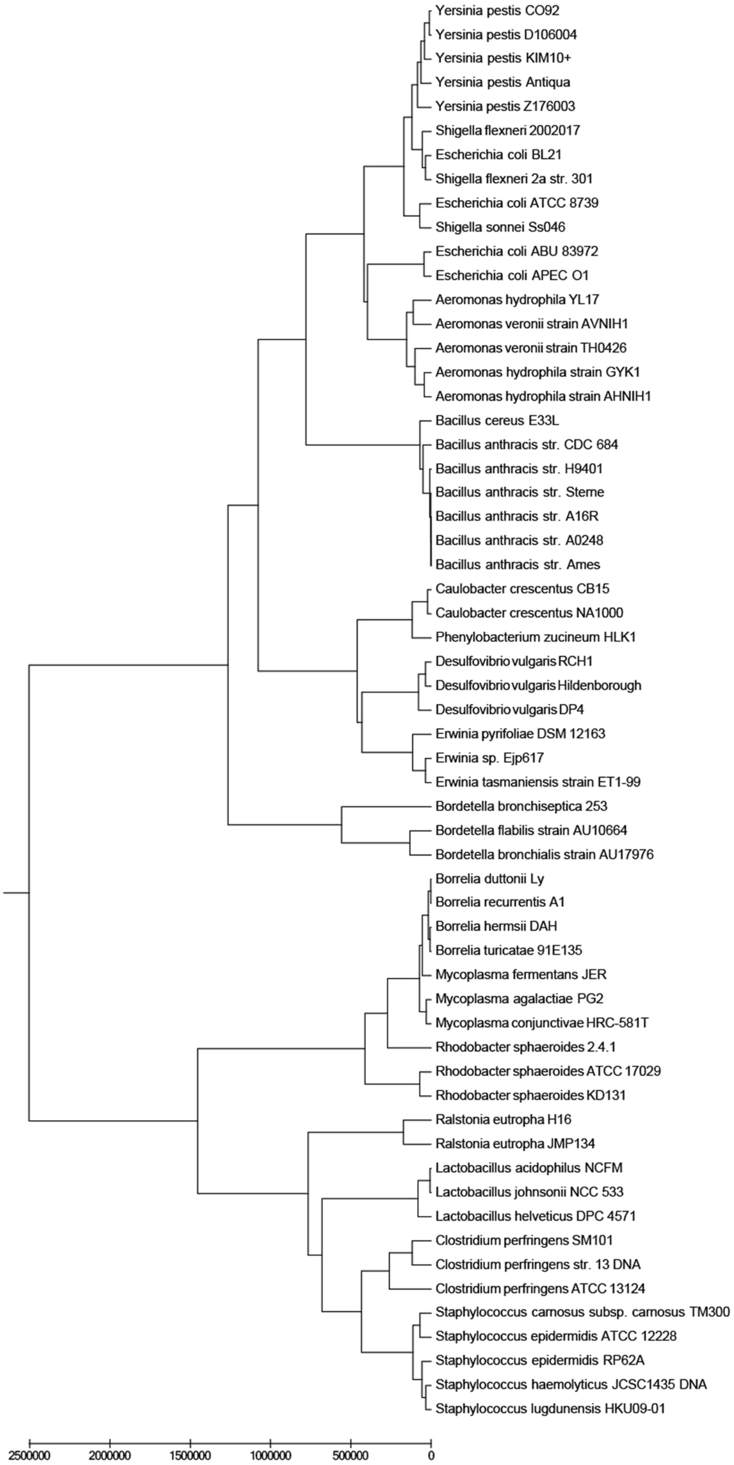
Yersinia pestis CO92
Yersinia pestis D106004
Yersinia pestis KIM10+
Yersinia pestis Antiqua
Yersinia pestis Z176003
Shigella flexneri 2002017
Escherichia coli BL21
Shigella flexneri 2a str. 301
Escherichia coli ATCC 8739
Shigella sonnei Ss046
Escherichia coli ABU 83972
Escherichia coli APEC O1
Aeromonas hydrophila YL17
Aeromonas veronii strain AVNIH1
Aeromonas veronii strain TH0426
Aeromonas hydrophila strain GYK1
Aeromonas hydrophila strain AHNIH1
Bacillus cereus E33L
Bacillus anthracis str. CDC 684
Bacillus anthracis str. H9401
Bacillus anthracis str. Sterne
Bacillus anthracis str. A16R
Bacillus anthracis str. A0248
Bacillus anthracis str. Ames
Caulobacter crescentus CB15
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Desulfovibrio vulgaris RCH1
Desulfovibrio vulgaris Hildenborough
Desulfovibrio vulgaris DP4
Erwinia pyrifoliae DSM 12163
Erwinia sp. Ejp617
Erwinia tasmaniensis strain ET1-99
Bordetella bronchiseptica 253
Bordetella flabilis strain AU10664
Bordetella bronchialis strain AU17976
Borrelia duttonii Ly
Borrelia recurrentis A1
Borrelia hermsii DAH
Borrelia turicatae 91E135
Mycoplasma fermentans JER
Mycoplasma agalactiae PG2
Mycoplasma conjunctivae HRC-581T
Rhodobacter sphaeroides 2.4.1
Rhodobacter sphaeroides ATCC 17029
Rhodobacter sphaeroides KD131
Ralstonia eutropha H16
Ralstonia eutropha JMP134
Lactobacillus acidophilus NCFM
Lactobacillus johnsonii NCC 533
Lactobacillus helveticus DPC 4571
Clostridium perfringens SM101
Clostridium perfringens str. 13 DNA
Clostridium perfringens ATCC 13124
Staphylococcus carnosus subsp. carnosus TM300
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Staphylococcus haemolyticus JCSC1435 DNA
Staphylococcus lugdunensis HKU09-01
2500000
2000000
1500000
1000000
500000
0

Figure 6. The UPGMA phylogenetic tree of 59 bacteria from 15 families based on our multiple encoding vector method.

	mtDNA	Influenza A	HRV	ebolavirus	coronavirus	bacteria
ClustalW	285	1.83	472	617	314	—
Our method	0.12	0.002	0.09	6.42	0.34	5.3

Table 1. Running time for our multiple encoding vector and ClustalW (minutes).

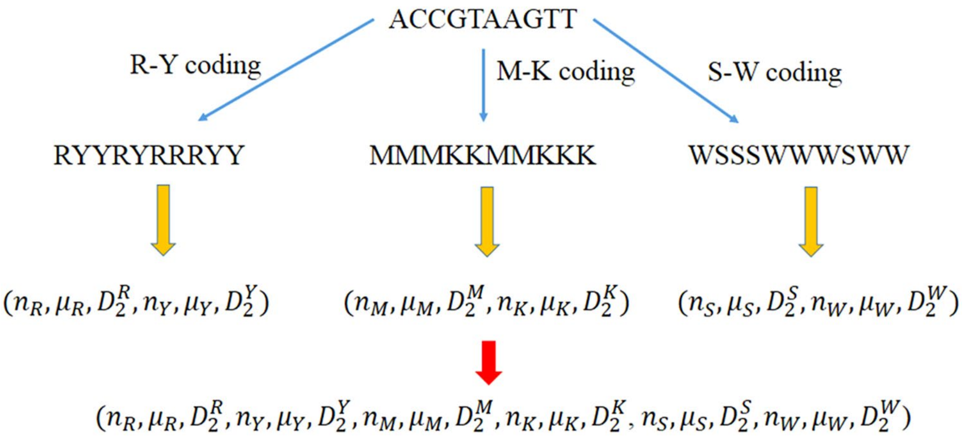
ACCGTAAGTT
R-Y coding
M-K coding
S-W coding
RYYRYRRRYY
MMMKKMMKKK
WSSSWWWSWW
$(n_R, \mu_R, D_2^R, n_Y, \mu_Y, D_2^Y)$
$(n_M, \mu_M, D_2^M, n_K, \mu_K, D_2^K)$
$(n_S, \mu_S, D_2^S, n_W, \mu_W, D_2^W)$
$(n_R, \mu_R, D_2^R, n_Y, \mu_Y, D_2^Y, n_M, \mu_M, D_2^M, n_K, \mu_K, D_2^K, n_S, \mu_S, D_2^S, n_W, \mu_W, D_2^W)$

Figure 7. Flow chart of constructing 18 dimensional multiple encoding vector.

speed. For the large data of 59 bacteria, ClustalW even can not finish the alignment, while our method takes only 5.61 minutes to produce those numerical vectors. We also find the result for influenza A data with clustalW is unfavourable. As shown in Fig. S7 (in the supplementary file), A/turkey/VA/505477-18/2007 (H5N1) virus is clustered with H1N1 viruses. This may be caused by reassortment events occurring among virions. Other results by ClustalW are fine and put in the supplemental document.

Our method is very fast and has the potential to construct phylogeny for whole genomes such large as mammalian. Note that mammalian genomes are often divided into several chromosomes. To compute the position of nucleotides, we only choose chromosome X of mammals to do phylogenetic analysis. Our dataset includes the species: chimpanzee (*Pan troglodytes*), human (*Homo sapiens*), monkey (*Macaca mulatta*), gorilla (*Gorilla gorilla*), dog (*Canis familiaris*), horse (*Equus caballus*), mouse (*Mus musculus*), opossum (*Monodelphis domesticus*), and platypus (*Ornithorhynchus anatinus*). The length of these X chromosomes ranges from about 6 to 147 Mb. The accession numbers are listed in the supporting file. Based on the new multiple encoding vector, the UPGMA phylogenetic tree of nine mammals is obtained. As shown in Fig. S8 in the supporting file, the primates are clustered together. Platypus and Opossum are placed in the base of tree.

In spite of the good performance in speed and accuracy of our approach, there are still some room for improvement. The method is sensitive to length of sequences analyzed. The sequences used for phylogenetic analysis should be approximately complete. For whole genome sequences studied, the use of partial genomes may produce incorrect evolutionary relationships. If we only use a gene or segment to infer phylogeny of species, the gene sequences or segments for all organism investigated should be intact. Thus we need to carefully prepare the sequences of organisms before constructing their phylogeny. Although the multiple sequence alignment does not have this requirement, gaps are inserted to sequences to make their length same automatically, which consumes much time.

In conclusion, we formulate a novel 18 dimensional vector method to compare biological sequences. The novelty of this method is that some important chemical and physical properties of sequences are incorporated into this method. Traditionally, the alignment-based method often produces phylogeny of species with high accuracy. However, these methods are time-consuming in speed and can not deal with large datasets and long sequences. As comparison, our approach is very fast and suitable for large sized biological sequences. Most importantly, the results tested on several different datasets show that this vector method can provide correct evolutionary relationships of different kinds of species.

Methods

Let $\mathcal{L}$ be the set of 4 bases, i.e., $\mathcal{L} = \{A, C, G, T\}$ and $Q = (s_1, s_2, ..., s_n)$ be a DNA sequence of length n, that is, $s_i \in \mathcal{L}$, $i = 1,2,\cdots, n$. The four bases are categorized into two groups according to three kinds of chemical and physical properties. The A and G bases are purine denoted by the letter R. The C and T bases are pyrimidine represented by Y. In another way of grouping, the A and C nucleotides are amino denoted by M and the G and T are keto denoted by K. According to the H bound, the G and C have strong H bond and they are represented by the letter S. The A and T bases contain weak H bond and they are represented by W. We define three numerical values for each of the letters R, Y, M, K, S and W. To characterize the distribution of R and Y in the sequence Q, we first replace the letters A and G by R and replace the C and T by Y in the sequence Q. Then the sequence only contains two kinds of letters: R and Y. For R, define $w_R(\cdot)$:$\{R, Y\} \to \{0, 1\}$ such that $w_R(s_i) = 1$ if $s_i = R$ and 0 otherwise. For R, we define the n_R, μ_R, D_2^R to describe the number of R, the average position of R and the variation of the position of R appearing in the sequence Q.

(1) Let $n_R = \sum_{i=1}^{n} w_R(s_i)$ denote the number of letter R occurring in Q.
(2) Let $\mu_R = \sum_{i=1}^{n} i \cdot \frac{w_R(s_i)}{n_R}$ be the mean position at which the letter R appears.
(3) Let $D_2^R = \sum_{i=1}^{n} \frac{(i - \mu_R)^2 w_R(s_i)}{n_R n}$ be the scaled 2-nd moment of positions of letter R.

Similarly, for Y we define $w_Y(\cdot)$:$\{R, Y\} \to \{0, 1\}$ such that $w_Y(s_i) = 1$ if $s_i = Y$, and 0 otherwise. Then we get three characteristics for Y: n_Y, μ_Y and D_2^Y. For this kind of classification for nucleotides, six values are used to present the distribution of the four bases with respect to this chemical property. In the same way, we define other triplet for M, K, S, and W. Thus an 18 dimensional vector of the DNA sequence Q is defined by $(n_R, \mu_R, D_2^R, \cdots, n_W, \mu_W, D_2^W)$. We call this multiple encoding vector because it utilizes three groups of letters to encode the DNA sequence. The construction of this vector is shown in Fig. 7. Pairwise distance among vectors derived from the new method are computed using Euclidean distance. Then the distance matrix of all biological sequences is constructed. Based on the UPGMA algorithm, the phylogenetic tree of organisms is built using MEGA 6.0 software[37,38].

Data Availability. The datasets analysed during the current study are available in the "supplementary file" and the R source code in this paper is freely available to the public upon request.

References

1. Altschul, S. F. *et al*. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–402 (1997).
2. Larkin, M. A. *et al*. Clustal w and clustal x version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
3. Vinga, S. & Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics* **19**, 513–523 (2003).
4. Deng, M., Yu, C., 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* **6**, e17293 (2011).
5. Yin, C., Chen, Y. & Yau, S. T. A measure of dna sequence similarity by fourier transform with applications on hierarchical clustering. *Journal of Theoretical Biology* **359**, 18–28 (2014).
6. Li, Y., Tian, K., Yin, C., He, R. L. & Yau, S. S.-T. Virus classification in 60-dimensional protein space. *Mol. Phylogenet. Evol.* **99**, 53–62 (2016).
7. Hoang, T., Yin, C. & Yau, S. S.-T. Numerical encoding of dna sequences by chaos game representation with application in similarity comparison. *Genomics* **108**, 134–142 (2016).
8. 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. *Journal of Molecular Evolution* **29**, 538–547 (1989).
9. Sims, G. E., Jun, S.-R., Wu, G. A. & Kim, S.-H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* **106**, 2677–2682 (2009).
10. Kolekar, P., Kale, M. & Kulkarni-Kale, U. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping. *Mol. Phylogenet. Evol.* **65**, 510–522 (2012).
11. Hatje, K. & Kollmar, M. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Front Plant Sci* **3**, 192 (2012).
12. Lu, G., Zhang, S. & Fang, X. An improved string composition method for sequence comparison. *BMC Bioinformatics* **9**(Suppl 6), S15 (2008).
13. Gao, L. & Qi, J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. *BMC Evol. Biol.* **7**, 41 (2007).
14. Wu, X., Wan, X.-F., Xu, D. & Lin, G. Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method. *Int J Bioinform Res Appl* **2**, 219–248 (2006).
15. Ulitsky, I., Burstein, D., Tuller, T. & Chor, B. The average common substring approach to phylogenomic reconstruction. *J. Comput. Biol.* **13**, 336–350 (2006).
16. Leimeister, C.-A. & Morgenstern, B. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* **30**, 2000–2008 (2014).
17. Cheng, J., Zeng, X., Ren, G. & Liu, Z. CGAP: a new comprehensive platform for the comparative analysis of chloroplast genome. *BMC Bioinformatics* **14**, 95 (2013).
18. Gao, Y. & Luo, L. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene* **492**, 309–314 (2012).
19. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic Acids Research* **18**, 2163–2170 (1990).
20. Goldman, N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Research* **21**, 2487–91 (1993).
21. Almeida, J. S., Carriço, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics* **17**, 429–437 (2001).
22. Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. Mitochondrial dna sequences of primates: Tempo and mode of evolution. *J. Mol. Evol.* **18**, 225–39 (1982).
23. Vijaykrishna, D. & Guan, Y. Reassortment of pandemic H1N1/2009 influenza A virus in swine. *Science* **328**, 1529 (2010).
24. Ghedin, E. *et al*. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* **437**, 1162–6 (2005).
25. Yin, C. & Yau, S. S. An improved model for whole genome phylogenetic analysis by fourier transform. *J. Theor. Biol.* **382**, 99–110 (2015).
26. Palmenberg, A. C. & David Spiro, R. K. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. *Science* **324**, 55–9 (2009).
27. Gire, S. K. *et al*. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).
28. Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. The evolution of ebola virus: Insights from the 2013–2016 epidemic. *Nature* **538**, 193–200 (2016).
29. Baize, S. *et al*. Emergence of zaire ebola virus disease in guinea. *N. Engl. J. Med.* **371**, 1418–1425 (2014).
30. Hoek, L. V. D. *et al*. Identification of a new coronavirus. *Nat. Med* **10**, 368–73 (2004).
31. Woo, P. C. Y. *et al*. Characterization and Complete Genome Sequence of a Novel Coronavirus, Coronavirus HKU1, from Patients with Pneumonia. *J. Virol.* **79**, 884–95 (2005).
32. Yu, C., Qian, L., Yin, C., He, R. L. & Yau, S. T. A novel construction of genome space with biological geometry. *DNA Res.* **17**, 155–68 (2010).
33. Berkhout, B. Identification of a new human coronavirus. *Nat. Med.* **10**, 368–73 (2004).
34. Dai, Q., Yang, Y. & Wang, T. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **24**, 2296–2302 (2008).
35. Wu, G. A., Jun, S.-R., Sims, G. E. & Kim, S.-H. Whole-proteome phylogeny of large dsdna virus families by an alignment-free method. *Proc. Natl. Acad. Sci. USA* **106**, 12826–12831 (2009).

36. Xu, Z. & Hao, B. Cvtree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Research* **37**, W174–W178 (2009).
37. Sokal, R. R. A statistical method of evaluating systematic relationships. *Univ. kansas Sci. bull* **38**, 1409–1438 (1958).
38. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. Mega6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729 (2013).

Acknowledgements

This study is supported by the National Natural Science Foundation of China (31271408) and Tsinghua University start up fund (to Stephen S.-T. Yau). Stephen S.-T. Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Author Contributions

S.S.Y. and Y.L. conceived the initial planning, Y.L. implemented the code, L.H. and R.L.H. analysed the results. All authors wrote and reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-12493-2.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.