Journal of Theoretical Biology 372 (2015) 135–145

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A new method to cluster DNA sequences using Fourier power spectrum

Tung Hoang [a], Changchuan Yin [a], Hui Zheng [a], Chenglong Yu [b,c], Rong Lucy He [d], Stephen S.-T. Yau [e,*]

[a] Department of Mathematics, Statistics and Computer Science, University of Ilinois at Chicago, Chicago, IL 60607, USA
[b] Mind and Brain Theme, South Australian Health and Medical Research Institute, North Terrace, Adelaide, SA 5000, Australia
[c] School of Medicine, Flinders University, Adelaide, SA 5001, Australia
[d] Department of Biological Sciences, Chicago State University, Chicago, IL, USA
[e] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

HIGHLIGHTS

- We propose to use Fourier power spectrum to cluster genes and genomes.
- We construct mathematical moments from the power spectrum.
- We perform phylogenetic analysis of genes and genomes based on moments.

ARTICLE INFO

Article history:
Received 12 September 2014
Received in revised form
15 January 2015
Accepted 23 February 2015
Available online 5 March 2015

Keywords:
Phylogenetic trees
Genes
Moments

ABSTRACT

A novel clustering method is proposed to classify genes and genomes. For a given DNA sequence, a binary indicator sequence of each nucleotide is constructed, and Discrete Fourier Transform is applied on these four sequences to attain respective power spectra. Mathematical moments are built from these spectra, and multidimensional vectors of real numbers are constructed from these moments. Cluster analysis is then performed in order to determine the evolutionary relationship between DNA sequences. The novelty of this method is that sequences with different lengths can be compared easily via the use of power spectra and moments. Experimental results on various datasets show that the proposed method provides an efficient tool to classify genes and genomes. It not only gives comparable results but also is remarkably faster than other multiple sequence alignment and alignment-free methods.

1. Introduction

In the last few decades, several methods to classify genes and proteins have been proposed. Most of these methods are alignment-based in which optimal alignments are obtained by using selected scoring systems. These methods provide accurate classification of biological sequences, and several algorithms have been developed and successfully applied (Katoh et al., 2002; Edgar, 2004; Larkin et al., 2007). Nevertheless, their major drawback is due to significantly high time and memory consumption which is not suitable when a quick clustering needs to be made, for example on a new deadly virus (Marra et al., 2003). Henceforth, an alignment-free technique is a trending method that often gives much faster classification on the same dataset (Vinga and Almeida, 2003; Yau et al., 2008; Yu et al., 2011, 2013). For example, the *k*-mer method is among the most popular alignment-free methods. In order to measure how different the two sequences are, the set of *k*-mers, or subsequences of length *k*, in the two biological sequences are collected and then the evolutionary distance between them is computed (Vinga and Almeida, 2003; Pandit and Sinha, 2010). The *k*-mer method gives comparable results to alignment-based methods while being computationally faster (Blaisdell, 1989).

Discrete Fourier Transform (DFT) is a powerful tool in signal and image processing. During recent years, DFT has been increasingly used in DNA research, such as gene prediction, protein coding region, genomic signature, hierarchical clustering, periodicity analysis (Tiwari et al., 1997; Anastassiou, 2000; Kotlar and Lavner, 2003; Vaidyanathan and Yoon, 2004; Afreixo et al., 2004, 2009; Tenreiro Machado et al., 2011). A DFT power spectrum of a DNA sequence reflects the nucleotide distribution and periodic

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

http://dx.doi.org/10.1016/j.jtbi.2015.02.026
0022-5193/

patterns of that sequence, and it has been applied to identify protein coding regions in genomic sequences (Fukushima et al., 2002; Yin and Yau, 2005, 2007). In this paper we provide a new alignment-free method to classify DNA sequences based on the DFT power spectrum. The method is tested and compared to other state-of-the-art methods on various datasets for speed and accuracy.

2. Materials and method

2.1. Mathematical background

In signal processing, sequences in time domain are commonly transformed into frequency domain to make some important features visible. Via that transformation, no information is lost but some hidden properties could be revealed (Oppenheim et al., 1989).

One of the most common transformations is Discrete Fourier Transform (Oppenheim et al., 1989). For a signal of length $N, f(n), n=0,\ldots,N-1$, the DFT of the signal at frequency k is

$$F(k)=\sum_{n=0}^{N-1} f(n)e^{-i(2\pi/N)kn}$$

for $k=0,\ldots,N-1$. The DFT power spectrum of a signal at frequency k is defined as

$$\mathrm{PS}(k)=|F(k)|^2,\quad k=0,\ldots,N-1$$

Notice that by definition, $\mathrm{PS}(0)=|F(0)|^2=|\sum_{n=0}^{N-1} f(n)|^2$.

The DFT is often used to find the frequency components of a signal buried in a noisy time domain. For example, let y be a signal containing a 60 Hz sinusoid of amplitude 0.8 and a 140 Hz sinusoid of amplitude 1. This signal can be corrupted by a zero-mean random noise:

$$y=0.8*\sin(2*\pi*60*t)+\sin(2*\pi*140*t)+\mathrm{random}$$

The frequencies can hardly be identified by looking at the original signal as in Fig. 1(a), but can be seen quite clearly when the signal is transformed to frequency domain by taking the DFT (Fig. 1(b)).

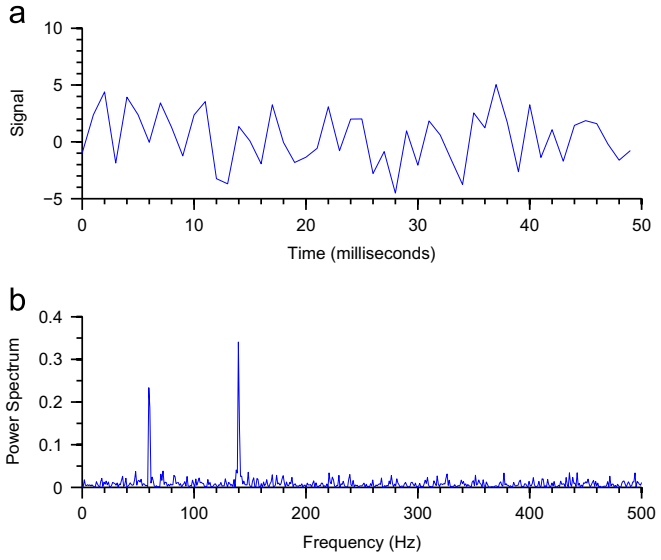

Fig. 1. Signal in time domain and frequency domain. (a) Signal Corrupted with Zero – Mean Random Noise. and (b) Single – Sided Power Spectrum

2.2. Moment vectors

For a DNA sequence composed of nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T), one typical way to get numerical representation is to use binary indicator sequences. The values of these sequences are either 0 or 1 indicating the absence or presence of a specific nucleotide. Specifically, for a given DNA sequence of length N, we define u_A of the same length as follows:

$$u_A(n)=\begin{cases}1 & \text{if } A \text{ is present at location } n \text{ of the sequence}\\ 0 & \text{otherwise}\end{cases}$$

u_C, u_G, u_T are defined similarly.

For example, for the sequence *AGTCTTACGA*, the corresponding indicator sequence of nucleotide A is $u_A=1000001001$.

The DFT of u_A is U_A where

$$U_A(k)=\sum_{n=0}^{N-1} u_A(n)e^{-i(2\pi/N)kn}$$

for $k=0,\ldots,N-1$.

The DFT power spectrum of u_A is PS_A where $\mathrm{PS}_A(k)=|U_A(k)|^2, k=0,\ldots,N-1$. The corresponding power spectrum for nucleotides C, G, T is defined similarly. In general, for a gene sequence of length N, let N_A, N_C, N_G, N_T be the number of nucleotide A, C, G, T in that sequence, respectively.

It is difficult to compare numerical sequences with different lengths, so we cannot cluster genes and genomes based on their power spectra sequences. One common approach to get over this problem is to use mathematical moments, e.g. for nucleotide A defines jth moment $M_j^A=\alpha_j^A\sum_{k=0}^{N-1}(\mathrm{PS}_A(k))^j, j=1,2\ldots$, where α_j^A be scaling factors. We want higher moments to converge to zero, i.e. essential information is kept in the first few moments. Thus, the chosen normalization factors α_j^A must reflect the nature of the sequences. Let us examine the binary indicator sequence of one nucleotide, A, in more detail.

By Parseval's theorem (Oppenheim et al., 1989),

$$\sum_{n=0}^{N-1}|u_A(n)|^2=\frac{1}{N}\sum_{k=0}^{N-1}\mathrm{PS}_A(k)\quad(\text{since } \mathrm{PS}_A(k)=|U_A(k)|^2)$$

The left side is actually N_A, i.e. the number of 1 in the A binary sequence. Hence, $\sum_{k=0}^{N-1}\mathrm{PS}_A(k)=N_AN$. So it is reasonable for α_j^A to be a power of N_{AN}. As stated above, we want moments converge to zero gradually so that information loss is minimal, thus $\alpha_j^A=1/(N_AN)^{j-1}$ is the best choice (which will be verified later). Therefore

$$M_j^A=\frac{1}{N_A^{j-1}N^{j-1}}\sum_{k=0}^{N-1}(\mathrm{PS}_A(k))^j$$

With this normalization, $M_1^A=\sum_{k=0}^{N-1}\mathrm{PS}_A(k)=N_AN$. Our experimental results on various datasets proved that this is a good normalization. However, by re-examining the formula, we find that a slight modification can be made to get better outcomes. From Section 2.1, we know $\mathrm{PS}_A(0)=|F_A(0)|^2=|\sum_{n=0}^{N-1}u_A(n)|^2=N_A^2$. Thus $\mathrm{PS}_A(0)$

Table 1
Running time comparison.

Datasets	Our method	MAFFT	*k*-mer	ClustalW
Mammals	4 s	NA	18 min 15 s	3 h 15 min
Influenza A	0.6 s	22 s	12 s	1 min 55 s
HRV	5 s	17 min 40 s	47 min 28 s	8 h 10 min
Coronavirus	6 s	NA	69 min 12 s	11 h 40 min
Bacteria	9 min 41 s	NA	NA	NA

might hold large weight compared to $PS_A(k)$ for other index k, which in turn leads to unnecessary memory consumption and computations for higher moments. Therefore, the terms $PS_A(0)$ are removed from the moments, and M_1^A becomes $\sum_{k=1}^{N-1} PS_A(k) = N_A N - PS_A(0) = N_A N - N_A^2 = N_A(N-N_A)$. From the first modified moment, we know how to adjust the scaling factor for the j-th moment in general, i.e. α_j^A must be a power of $N_A(N-N_A)$. Thus the new normalization is

$$M_j^A = \frac{1}{N_A^{j-1}(N-N_A)^{j-1}} \sum_{k=1}^{N-1} (PS_A(k))^j$$

The fact that higher moments tend to zero is verified as follows:

$$M_j^A = N_A(N-N_A) \sum_{k=1}^{N-1} \left(\frac{PS_A(k)}{N_A(N-N_A)} \right)^j = N_A(N-N_A) \sum_{k=1}^{N-1} z_k^j$$

where $z_k = PS_A(k)/N_A(N-N_A)$. Notice that $\sum_{k=1}^{N-1} z_k = 1$, thus it is obvious that $\lim_{j\to\infty} \sum_{k=1}^{N-1} z_k^j = 0$.

This fact also shows that $\alpha_j^A = 1/(N_A(N-N_A))^{j-1}$ is the best scaling factor, as for $\alpha_j^A = 1/(N_A(N-N_A))^j$, $M_1^A = 1$ so moments tend to zero very fast, thus much information can be lost, and for $\alpha_j^A = 1/(N_A(N-N_A))^{j-2}$, $M_1^A = N_A^2(N-N_A)^2$ so moments tend to zero much slower, thus more computational time and memory storage are needed. Additionally, due to symmetric property of DFT coefficients (Oppenheim et al., 1989), we only have to consider the first half of power spectrum. Therefore, the moments are improved as follows:

$$M_j^A = \frac{1}{N_A^{j-1}(N-N_A)^{j-1}} \sum_{k=1}^{\lfloor N/2 \rfloor} (PS_A(k))^j$$

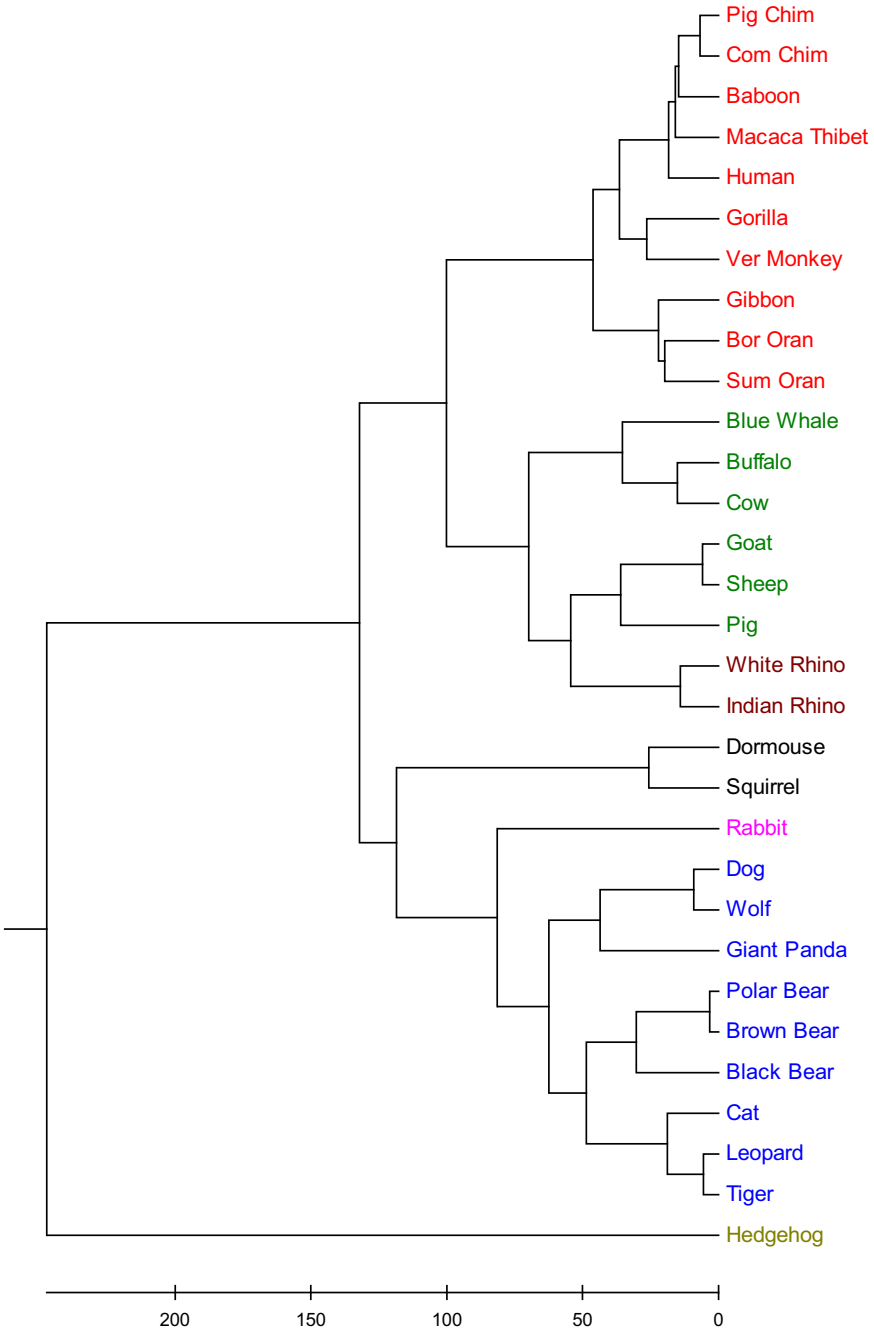

Fig. 2. Phylogenetic tree of 31 mammalian mitochondrial genomes by our method, drawn by Mega 6.

The moments for other nucleotides C, G, T are given similarly. Then the first few moments are used to construct vectors in Euclidean space. Our experimental results show that three moments are sufficient for an accurate clustering. Thus, each gene or genome sequence can be realized as a geometric point in the 12-dimensional Euclidean space, i.e. $(M_1^A, M_1^C, M_1^G, M_1^T, M_2^A, M_2^C, M_2^G, M_2^T, M_3^A, M_3^C, M_3^G, M_3^T)$. Pairwise Euclidean distances between these points are calculated to cluster the gene or genome sequences. We call this *Power Spectrum Moments* method, or *PS-M* method.

Power spectrum has been rarely used to study the phylogenetic analysis of DNA sequences because it is difficult to do comparison on sequences with different lengths. Zhao et al. (2011) came up with the idea of using normalized and centralized moments to compare sequences of different lengths. Motivated by that idea, we discovered a way to scale moments naturally, and only normalized moments are used to construct the Euclidean vectors. Discarding the first coefficient is another novelty of our *PS-M* method. The first coefficient holds significant weight and is highly dependent on the number of the respective nucleotide, which is redundant information. We also exclude the second half of the power spectrum due to its symmetry, and we give proof as to why the moment vectors converge to zero. Experimental results in the following section show that 12-dimensional moment vectors are enough to cluster genomes correctly.

3. Results

The *PS-M* method is tested on different datasets that range from small to medium size, as well as short to long genomes. In order to compare and analyze various genomic data, moment vectors are calculated and a matrix of Euclidean pairwise distances between those vectors is constructed. To cluster data into biological groups, a phylogenetic tree is built based on the distance matrix using the UPGMA method (Sokal, 1958).

The running time of our *PS-M* method is compared to three state-of-the-art methods. The first is the alignment-based method ClustalW (Larkin et al., 2007) implemented on MEGA 6 (Tamura

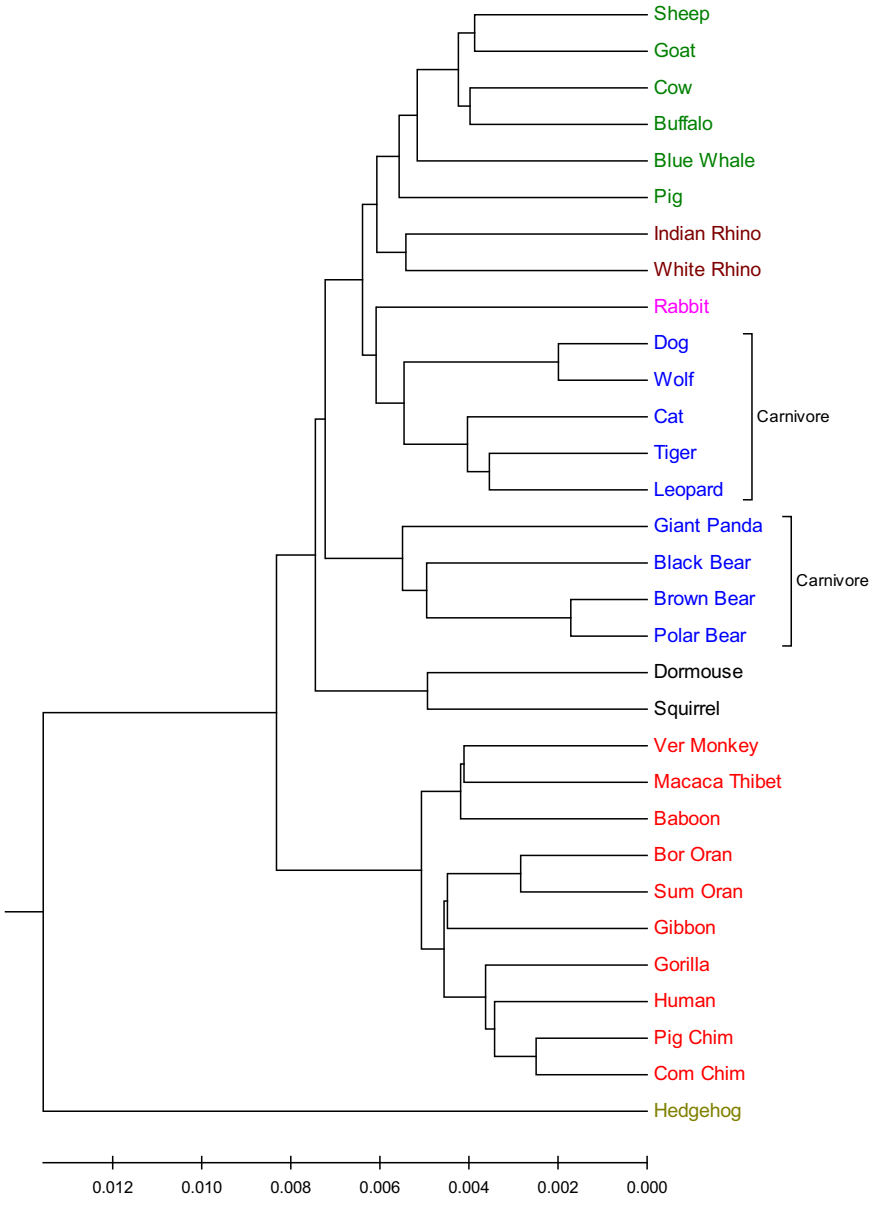

Fig. 3. Phylogenetic tree of 31 mammalian mitochondrial genomes by the k-mer method, $k=5$.

et al., 2013). The second is MAFFT (Katoh et al., 2002), another alignment-based method using fast Fourier transform. The last method is the alignment-free *k*-mer method (Vinga and Almeida, 2003) which is implemented on Matlab by Vinga and Almeida (2003). The running time is recorded in Table 1.

Even though reasonably accurate, ClustalW is significantly time consuming, as it aims to achieve the best possible results neglecting speed and efficiency. MAFFT runs much faster than ClustalW, but it sacrifices some accuracy in exchange. Moreover, errors occurred when we tested the datasets showing limitations of the MAFFT method. The errors are discussed in detail in the next section. Meanwhile, both *k*-mer and *PS-M* methods are alignment-free, and both attempt to improve speed and efficiency with little sacrifice of accuracy. Thus, our phylogeny results are directly compared to the *k*-mer method in this study. Phylogenies of the ClustalW method are included in the supplementary materials for reference (Figure S1–S4).

Phylogenies of our method and the *k*-mer method are drawn using Matlab and MEGA 6 respectively (Tamura et al., 2013). Computations in this research are implemented on a PC with configuration of Intel Core i7 CPU 2.40 GHz and 8 GB RAM.

3.1. Mammals

The mitochondrial genome is not highly conserved and has a rapid mutation rate, thus it is suitable for examining the mode and tempo of molecular evolution (Brown et al., 1982). The *PS-M* method

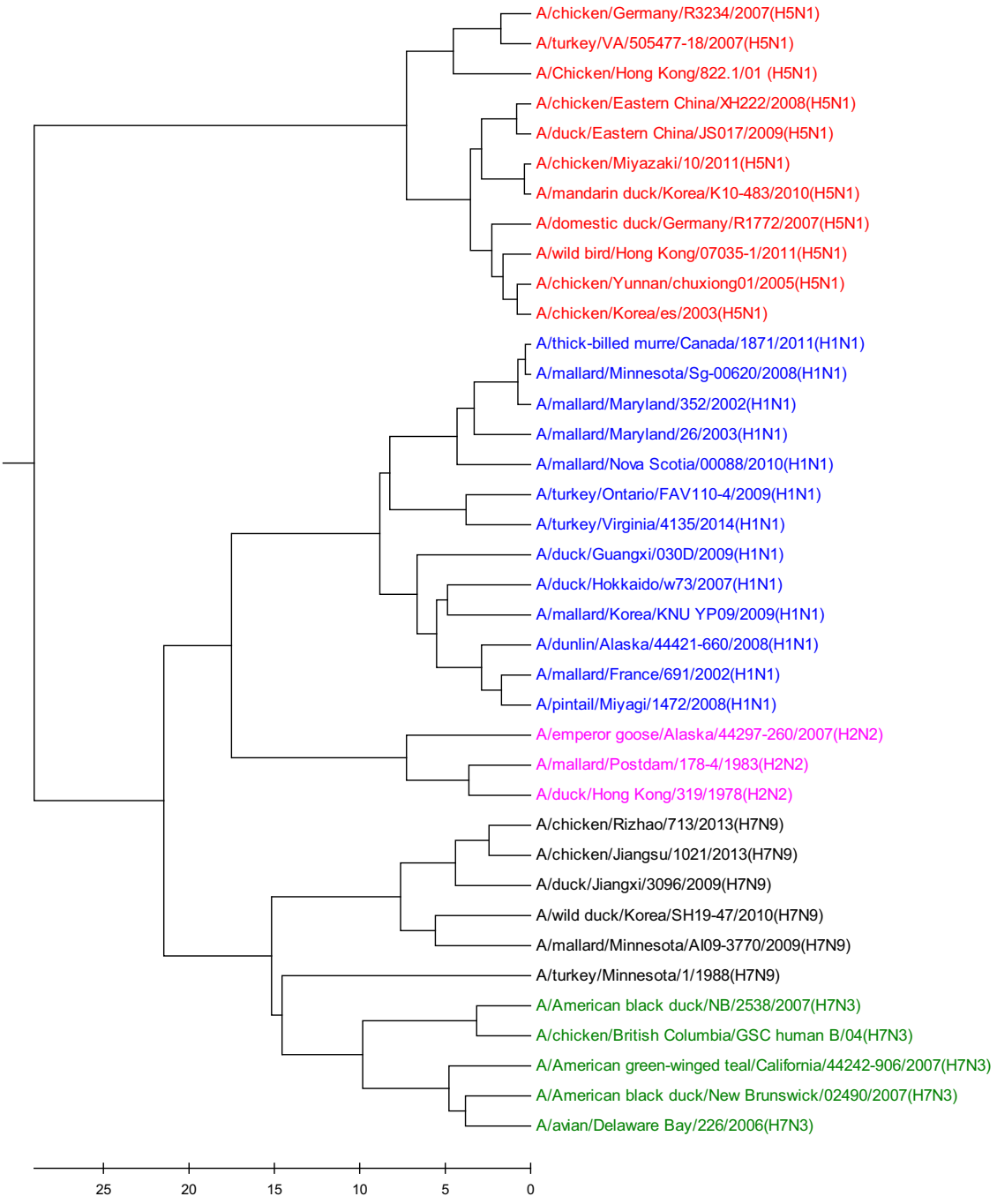

Fig. 4. Phylogenetic tree of 38 Influenza A viruses based on segment 6 by our method, drawn by Mega 6.

was tested on a mitochondrial DNA dataset of 31 mammalian genome sequences from GenBank, each sequence has a length range from 16,300 to 17,500 nucleotides. This dataset was analyzed by Deng et al. (2011) using the natural vector method. In our method, these 31 genomes are clustered correctly into 7 groups: *Primates, Cetacea and Artiodactyla, Perissodactyla, Rodentia, Lagomorpha, Carnivore, and Erinaceomorpha* (Fig. 2), which is consistent with the work of Deng et al. (2011). Meanwhile, as Fig. 3 illustrates, a majority part of Carnivore is misplaced by the *k*-mer method.

3.2. Influenza A viruses

We test the efficiency of the *PS-M* method at a gene level. Influenza A viruses have been a major health threat to both human society and animals (Alexander, 2000). Influenza A viruses are single-stranded, segmented RNA viruses, which are classified based on the viral surface proteins hemagglutinin (HA or H) and neuraminidase (NA or N) (Webster et al., 1992). Eighteen H serotypes (H1 to H18) and eleven N (N1 to N11) serotypes of Influenza A viruses have been identified. Influenza A viruses are the most dangerous due to their wide natural host range, including birds, horses, swines, and humans; and they are known to have high degree of genetic and antigenic variability (Palese and Young, 1982; Garten et al., 2009). Influenza A viruses have caused many pandemic flues, some of the most lethal subtypes are H1N1, H2N2, H5N1, H7N3, and H7N9. These subtypes will be chosen to test the efficiency of our method. Specifically, we will examine segment 6 of Influenza A virus genome, which encodes for neuraminidase (NA). As illustrated by the phylogenetic trees, the virus A/turkey/VA/505477-18/2007(H5N1) is not correctly clustered in the *k*-mer method (Fig. 5).

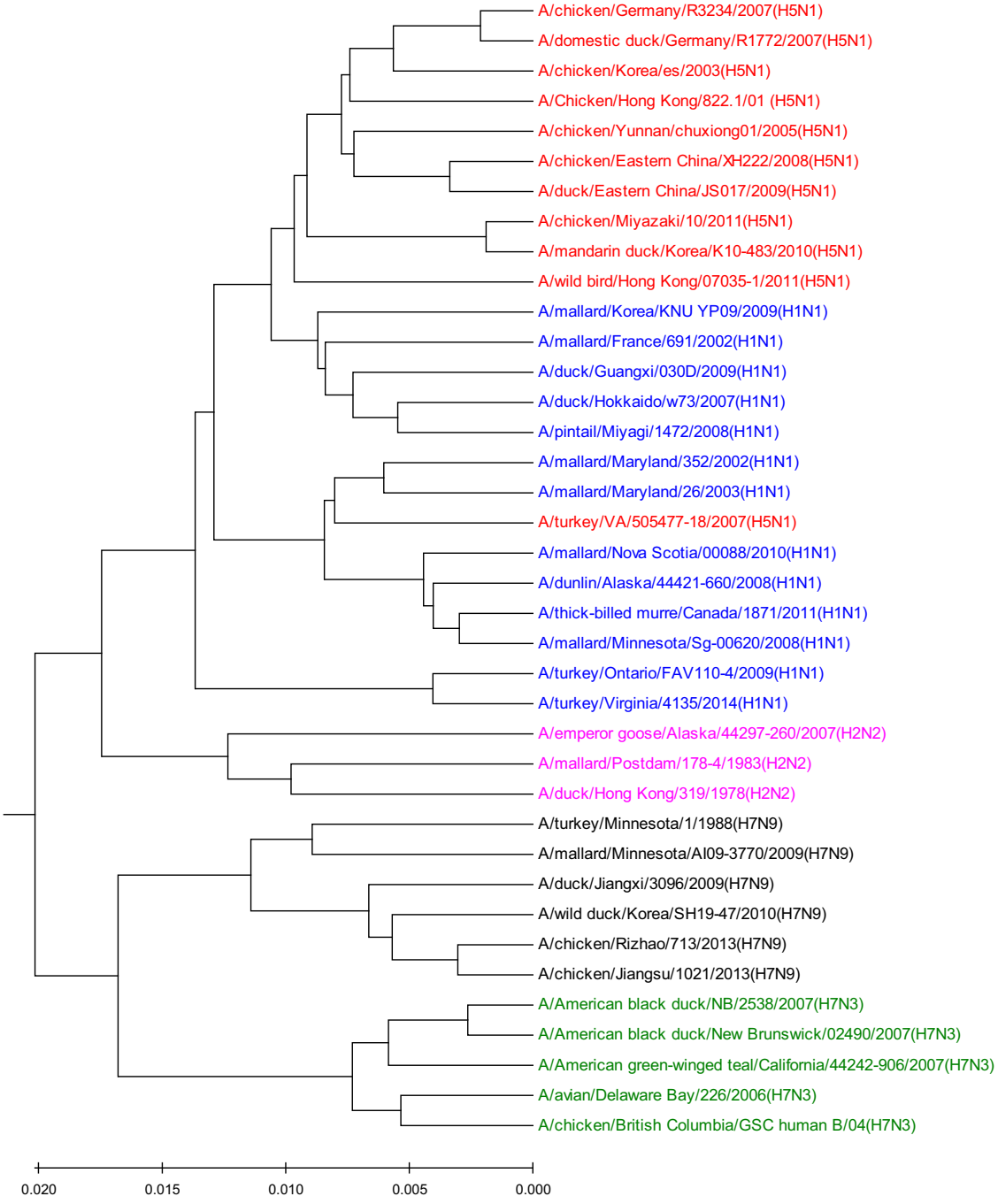

Fig. 5. Phylogenetic tree of 38 Influenza A viruses based on segment 6 by the *k*-mer method, $k=4$.

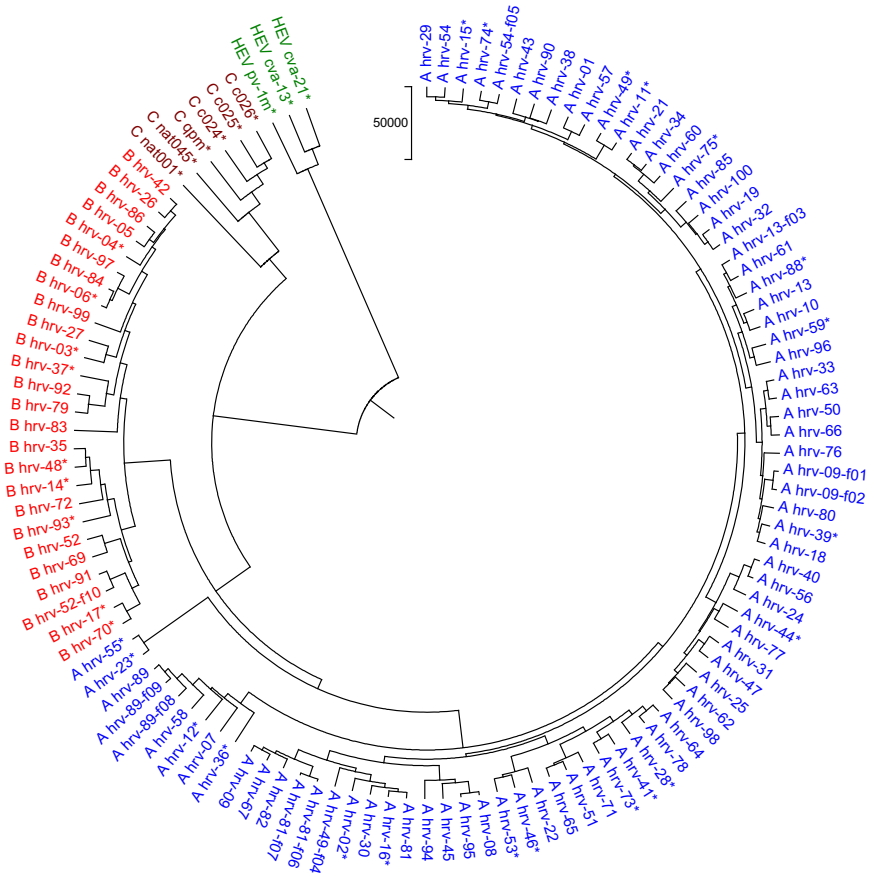

Fig. 6. Phylogenetic tree of HRV by our method, drawn by Mega 6.

Fig. 7. Phylogenetic tree of HRV by the *k*-mer method, $k=4$.

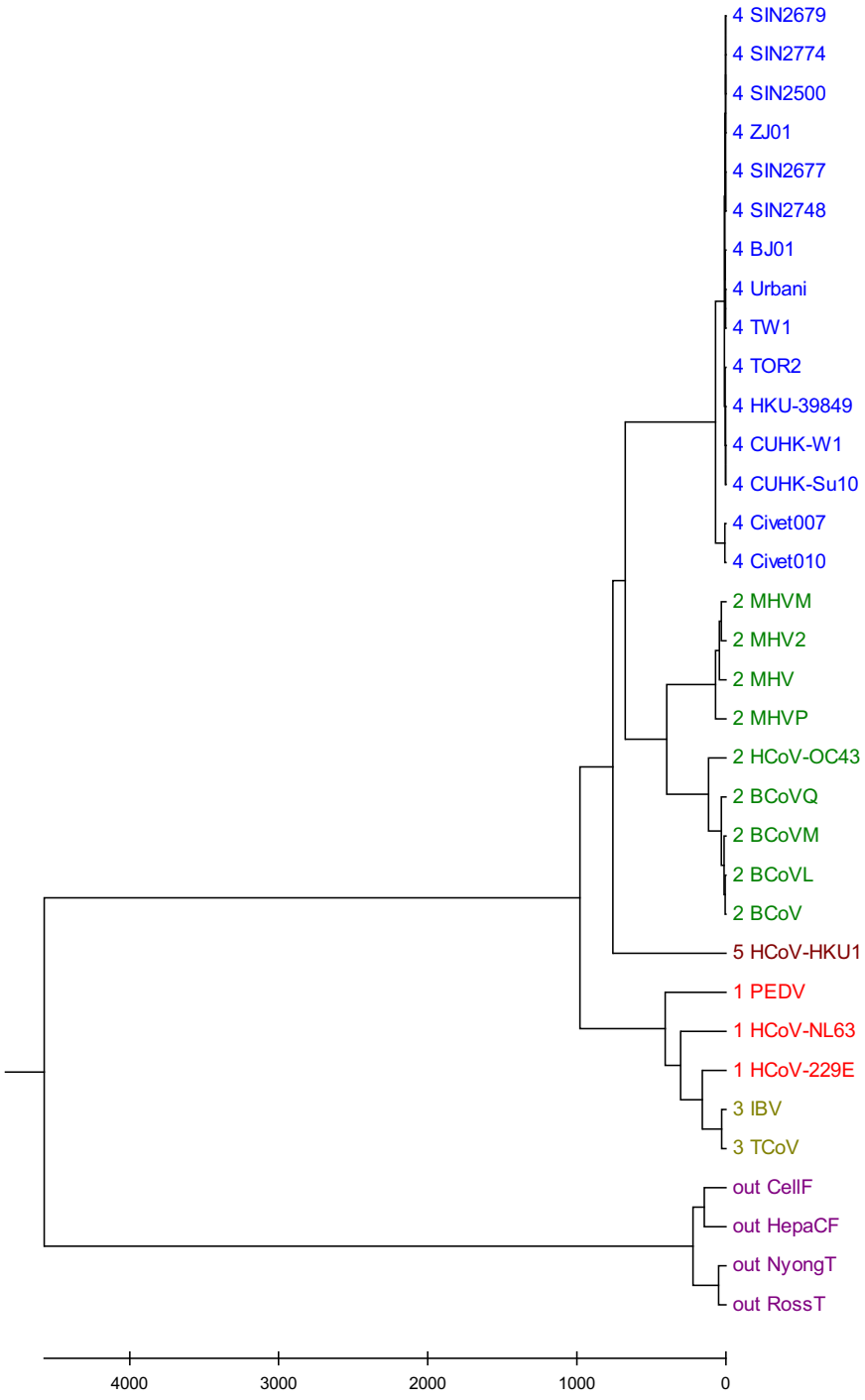

Fig. 8. Phylogenetic tree of coronavirus using our method, drawn by Mega 6.

On the other hand, the dataset is classified correctly into biological groups by our proposed *PS-M* method (Fig. 4).

3.3. *Human rhinovirus*

The efficiency of *PS-M* method was examined on a large size dataset, Human Rhinoviruses (HRV). HRV are associated with upper and lower respiratory diseases, particularly in patients with asthma. They are the predominant cause of the common cold and are responsible for more than one-half of cold-like illnesses. Past works have classified HRV into three genetically distinct groups within the genus Enterovirus and the family Picornaviridae (Palmenberg et al., 2009; Deng et al., 2011). In their paper, Palmenberg et al. (2009) clustered the complete HRV genomes, consisting of three groups HRV-A, HRV-B, HRV-C of 113 genomes and three outgroup sequences HEV-C. While the genomes were well classified, the running time was quite high due to the use of multiple sequence alignment to construct the evolutionary tree. By our method, the phylogenetic tree is reconstructed and the three groups of HRV are clearly separated and are distinguished from the outgroup HEV-C (Fig. 6). On the other hand, a small part of HRV-A is grouped on the same clade with HRV-B in the *k*-mer method (Fig. 7).

3.4. Coronavirus

Coronaviruses, a genus of the *Coronaviridae* family, can cause a variety of severe diseases in respiratory and gastrointestinal tract (HCoV-229E and HCoV-OC43), or even life-threatening pneumonia (severe acute respiratory syndrome, or SARS). Thus, identification and classification of coronaviruses, especially human coronaviruses, are important and have been extensively studied so far. We cluster the set of 30 coronaviruses and 4 outgroup non-coronavirus sequences using the *PS-M* method. This coronaviruses dataset has been used by various authors before, e.g. Woo et al. (2005) and Yu et al. (2010). In van der Hoek et al. (2004) the authors put the newly discovered human coronavirus NL63 into the same group with human coronavirus 229E (group 1 in our work), which is separated from other two human coronaviruses groups, HCoV-OC43 (group 2 in our work) and SARS (group 4 in our work). In Woo et al. (2005) the authors agreed to include the newfound HCoV-HKU1 in group 2 but also claimed that it is a distinct member within the group and it is not very closely related to the rest of the group. Yu et al. (2010) noticed that HCoV-HKU1 is an individual coronavirus between SARS group 4 and the traditional group 2, thus they proposed HCoV-HKU1 belongs to a new group 5. In our phylogenetic tree below (Fig. 8), HCoV-NL63 and HCoV-229E are clustered into group 1, which is similar to the work of van der Hoek et al. (2004). Moreover, group 5 is close to both group 4 and group 2 but separated from them, thus the result is consistent with the work of Woo et al. (2005) and Yu

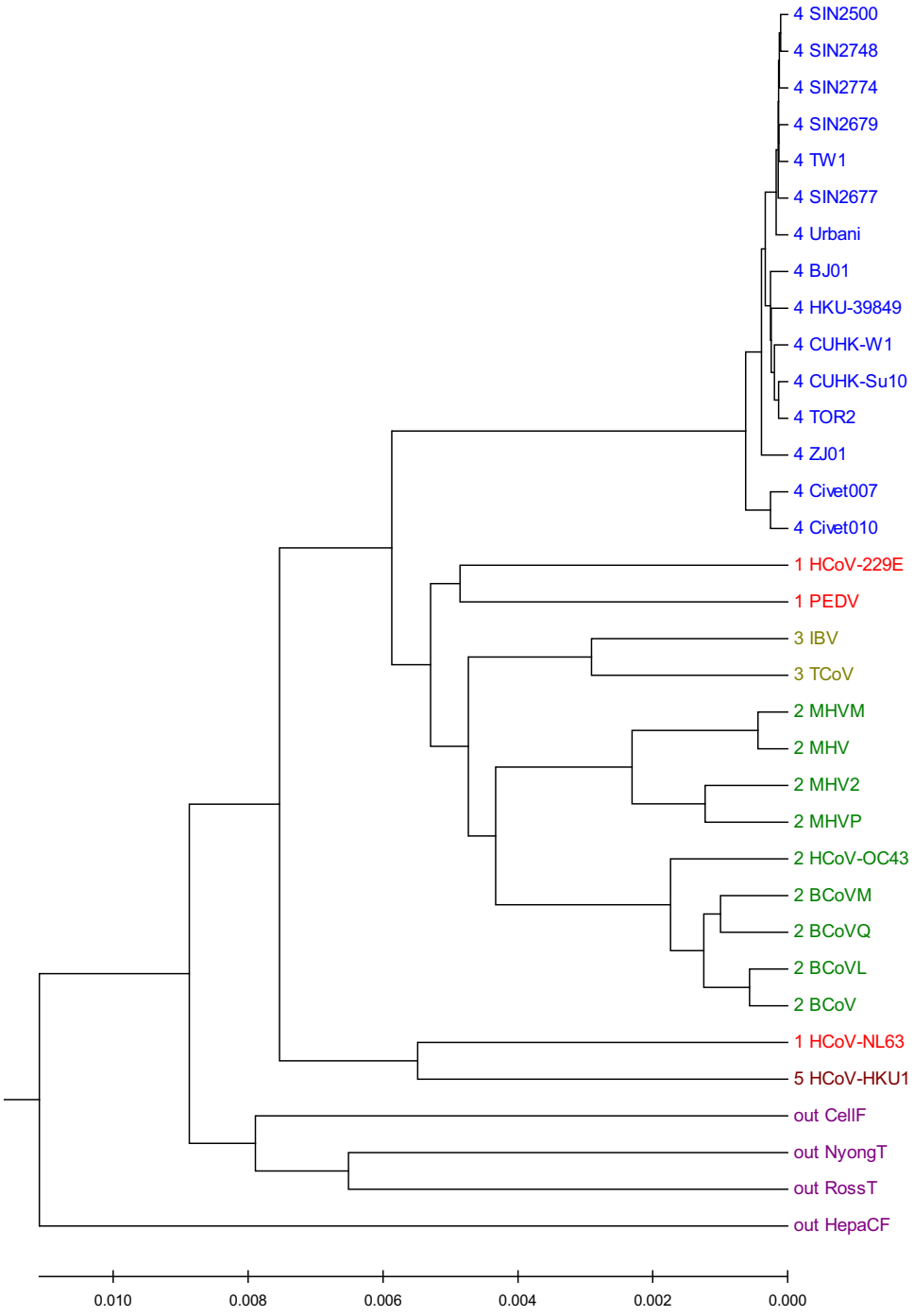

Fig. 9. Phylogenetic tree of coronavirus using the *k*-mer method, $k=6$.

et al. (2010). Based on Fig. 9, the HCoV-NL63 is misplaced from group 1 by the *k*-mer method.

3.5. Bacteria

Methods like multiple sequence alignment cannot handle large data. The bacteria genome, consisting of millions of base pairs, is a useful dataset for checking a method's efficiency. The *PS-M* method was tested on the set of 30 bacterial genomes, consisting of 8 families: *Enterobacteriaceae, Bacilleceae, Rhodobacteriaceae, Spirochaetaceae, Desulfovibrionaceae, Clostridiaceae, Burkholderiaceae, and Staphylococcaceae*. Most of the data has a length between 3 and 5 million base pairs. As illustrated by the phylogenetic tree in Fig. 10, the dataset is well clustered into 8 groups by the *PS-M* method in about 10 min. None of the three state-of-the-art methods (ClustalW, MAFFT, *k*-mer) are able to cluster these genomes.

4. Discussion

EMBL-EBI (http://www.ebi.ac.uk/Tools/msa/mafft/) is among the most common online tool for MAFFT. We tested our datasets on this site but there were errors for coronavirus and mammals genomes, namely "Raw Tool Output" as it appeared on the EBI website. Since the input DNA sequences worked for all other methods, the error seems to be the tool's problem. Therefore, despite MAFFT's reasonable biological classification and fast processing time, it has some drawbacks that is likely to require some restriction on the underlying dataset. Another obvious disadvantage of this online tool is that it does not allow us to save the phylogenetic tree directly to the computer, and it does not have circular tree option for dataset with a large number of elements. Thus, we do not include the phylogeny of HRV and Influenza A generated by MAFFT in this paper due to these reasons, even though the method can reasonably classify these datasets. However,

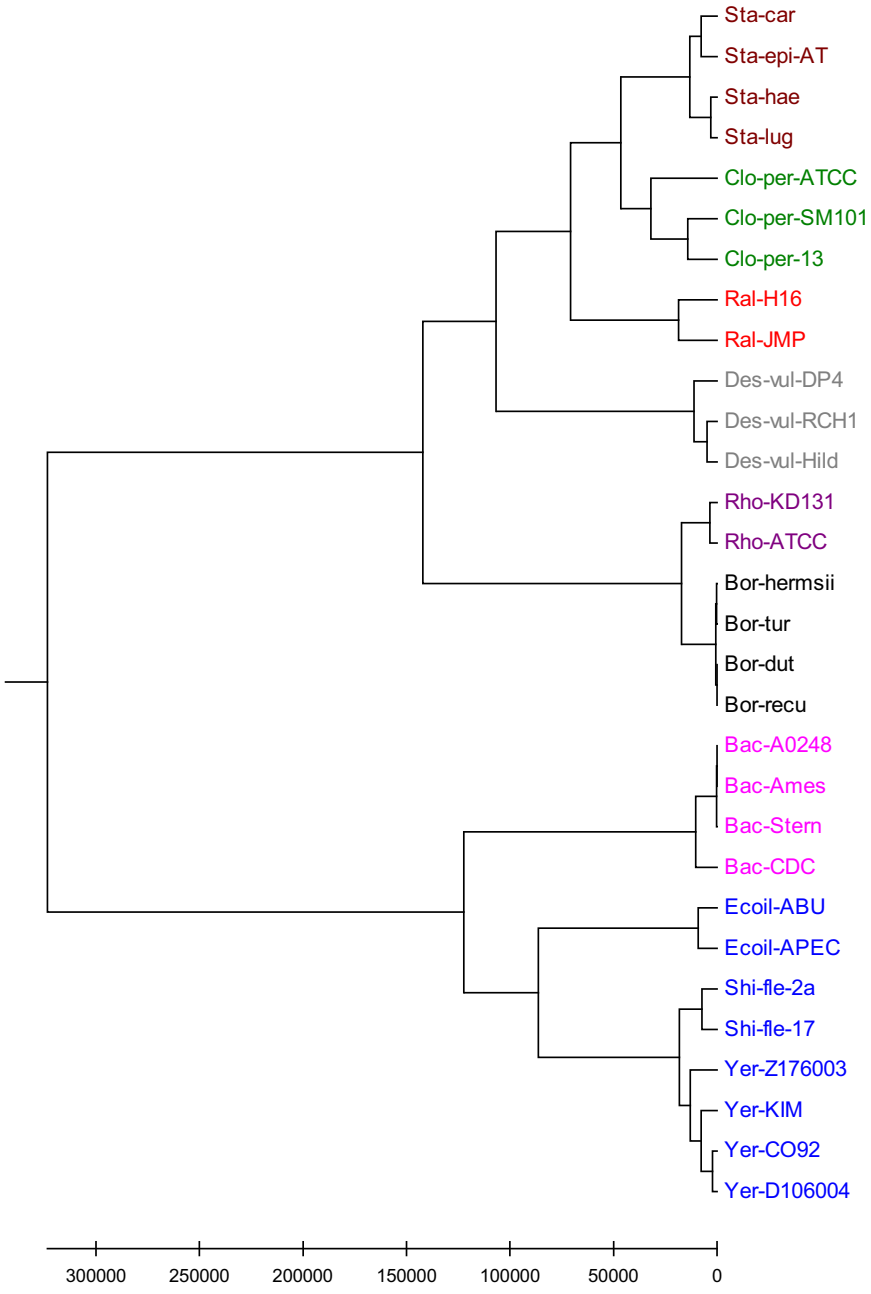

Fig. 10. Phylogenetic tree of 30 bacteria species by our method, drawn by Mega 6.

the running time of MAFFT for HRV and Influenza A is still recorded in Table 1.

The *k*-mer method is alignment-free and it gives fairly good classification with an acceptable running time. But its major disadvantage is that we do not know which value of *k* will produce the best classification. In our work, we had to try various values of *k* and then choose the value with the best outcome.

ClustalW is by far among the best alignment-based method. It often gives the most accurate biological classification, but its running time is significantly high in exchange. As a result, it is not able to perform well on large datasets.

From the performance comparison (Table 1), we can see that our method is much faster than other methods while still providing comparable biological classification. Most notably, none of the above three methods were able to cluster large genome like bacteria, while our method could do well in about 10 min. The MATLAB code for our method can be accessed from: http://www.mathworks.com/matlabcentral/fileexchange/49026.

Acknowledgment

This research is supported by the USA Natural Science Foundation (DMS-1120824 to S.S.-T.Y., 1119612 to R.L.H.), National Institutes of Health (5 SC3 GM098180-04 to R.L.H.), National Natural Sciences Foundation of China (31271408 to S.S.-T.Y.,), Tsinghua University startup fund and Tsinghua University independent research project grant.

Appendix A. Supporting information

Supplementary Material S1 contains tables of accession numbers and names of the underlying datasets (Table S1–S4) as well as the remaining phylogenetic trees generated by ClustalW (Figure S1–S3).

Appendix B. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2015.02.026.

References

Afreixo, V., Bastos, C.A., Pinho, A.J., Garcia, S.P., Ferreira, P.J., 2009. Genome analysis with inter-nucleotide distances. Bioinformatics 25 (23), 3064–3070.
Afreixo, V., Ferreira, P.J., Santos, D., 2004. Spectrum and symbol distribution of nucleotide sequences. Phys. Rev. E 70 (3), 031910.
Alexander, D.J., 2000. A review of avian influenza in different bird species. Vet. Microbiol. 74 (1), 3–13.
Anastassiou, D., 2000. Frequency-domain analysis of biomolecular sequences. Bioinformatics 16 (12), 1073–1081.
Blaisdell, B.E., 1989. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a computer-generated model system. J. Mol. Evol. 29 (6), 538–547.
Brown, W.M., Prager, E.M., Wang, A., Wilson, A.C., 1982. Mitochondrial dna sequences of primates: tempo and mode of evolution. J. Mol. Evol. 18 (4), 225–239.
Deng, M., Yu, C., Liang, Q., He, R.L., Yau, S.S.-T., 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PloS One 6 (3), e17293.
Edgar, R.C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32 (5), 1792–1797.
Fukushima, A., Ikemura, T., Kinouchi, M., Oshima, T., Kudo, Y., Mori, H., Kanaya, S., 2002. Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. Gene 300 (1), 203–211.
Garten, R.J., Davis, C.T., Russell, C.A., Shu, B., Lindstrom, S., Balish, A., Sessions, W.M., Xu, X., Skepner, E., Deyde, V., et al., 2009. Antigenic and genetic characteristics of swine-origin 2009 a (h1n1) influenza viruses circulating in humans. Science 325 (5937), 197–201.
Katoh, K., Misawa, K., Kuma, K.-i., Miyata, T., 2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. Nucleic Acids Res. 30 (14), 3059–3066.
Kotlar, D., Lavner, Y., 2003. Gene prediction by spectral rotation measure: a new method for identifying protein-coding regions. Genome Res. 13 (8), 1930–1937.
Larkin, M.A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al., 2007. Clustal w and clustal x version 2.0. Bioinformatics 23 (21), 2947–2948.
Marra, M.A., Jones, S.J., Astell, C.R., Holt, R.A., Brooks-Wilson, A., Butterfield, Y.S., Khattra, J., Asano, J.K., Barber, S.A., Chan, S.Y., et al., 2003. The genome sequence of the sars-associated coronavirus. Science 300 (5624), 1399–1404.
Oppenheim, A.V., Schafer, R.W., Buck, J.R., et al., 1989. Discrete-Time Signal Processing, vol. 2. Prentice-hall, Englewood Cliffs.
Palese, P., Young, J.F., 1982. Variation of influenza a, b, and c viruses. Science 215 (4539), 1468–1474.
Palmenberg, A.C., Spiro, D., Kuzmickas, R., Wang, S., Djikeng, A., Rathe, J.A., Fraser-Liggett, C.M., Liggett, S.B., 2009. Sequencing and analyses of all known human rhinovirus genomes reveal structure and evolution. Science 324 (5923), 55–59.
Pandit, A., Sinha, S., 2010. Using genomic signatures for hiv-1 sub-typing. BMC Bioinf. 11 (Suppl 1), S26.
Sokal, R.R., 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38, 1409–1438.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. Mega6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. 30 (12), 2725–2729.
Tenreiro Machado, J., Costa, A.C., Quelhas, M.D., 2011. Fractional dynamics in dna. Commun. Nonlinear Sci. Numer. Simul. 16 (8), 2963–2969.
Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S., Ramaswamy, R., 1997. Prediction of probable genes by fourier analysis of genomic sequences. Bioinformatics 13 (3), 263–270.
Vaidyanathan, P., Yoon, B.-J., 2004. The role of signal-processing concepts in genomics and proteomics. J. Frankl. Inst. 341 (1), 111–135.
van der Hoek, L., Pyrc, K., Jebbink, M.F., Vermeulen-Oost, W., Berkhout, R.J., Wolthers, K.C., Wertheim-van Dillen, P.M., Kaandorp, J., Spaargaren, J., Berkhout, B., 2004. Identification of a new human coronavirus. Nat. Med. 10 (4), 368–373.
Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison—a review. Bioinformatics 19 (4), 513–523.
Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M., Kawaoka, Y., 1992. Evolution and ecology of influenza a viruses. Microbiol. Rev. 56 (1), 152–179.
Woo, P.C., Lau, S.K., Chu, C.-m., Chan, K.-h., Tsoi, H.-w., Huang, Y., Wong, B.H., Poon, R.W., Cai, J.J., Luk, W.-k., et al., 2005. Characterization and complete genome sequence of a novel coronavirus coronavirus, hku1, from patients with pneumonia. J. Virol. 79 (2), 884–895.
Yau, S.S.-T., Yu, C., He, R., 2008. A protein map and its application. DNA Cell Biol. 27 (5), 241–250.
Yin, C., Yau, S.S.-T., 2005. A fourier characteristic of coding sequences: origins and a non-fourier approximation. J. Comput. Biol. 12 (9), 1153–1165.
Yin, C., Yau, S.S.-T., 2007. Prediction of protein coding regions by the 3-base periodicity analysis of a dna sequence. J. Theor. Biol. 247 (4), 687–694.
Yu, C., Deng, M., Yau, S.S.-T., 2011. DNA sequence comparison by a novel probabilistic method. Inf. Sci. 181 (8), 1484–1492.
Yu, C., Hernandez, T., Zheng, H., Yau, S.-C., Huang, H.-H., He, R.L., Yang, J., Yau, S.S.-T., 2013. Real time classification of viruses in 12 dimensions. PloS One 8 (5), e64328.
Yu, C., Liang, Q., Yin, C., He, R. L., Yau, S. S.-T., 2010. A novel construction of genome space with biological geometry. DNA Res., dsq008.
Zhao, B., Duan, V., Yau, S.S.-T., 2011. A novel clustering method via nucleotide-based fourier power spectrum analysis. J. Theor. Biol. 279 (1), 83–89.