Gene 673 (2018) 239–250

Contents lists available at ScienceDirect

Gene

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

Research article

A new method to cluster genomes based on cumulative Fourier power spectrum

Rui Dong[a,1], Ziyue Zhu[a,1], Changchuan Yin[b], Rong L. He[c], Stephen S.-T. Yau[a,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
[b] Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, IL 60607, USA
[c] Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA

ARTICLE INFO

Keywords:
Cumulative Fourier power spectrum
Moment vectors
DNA sequences
Phylogenetic trees

ABSTRACT

Analyzing phylogenetic relationships using mathematical methods has always been of importance in bioinformatics. Quantitative research may interpret the raw biological data in a precise way. Multiple Sequence Alignment (MSA) is used frequently to analyze biological evolutions, but is very time-consuming. When the scale of data is large, alignment methods cannot finish calculation in reasonable time. Therefore, we present a new method using moments of cumulative Fourier power spectrum in clustering the DNA sequences. Each sequence is translated into a vector in Euclidean space. Distances between the vectors can reflect the relationships between sequences. The mapping between the spectra and moment vector is one-to-one, which means that no information is lost in the power spectra during the calculation. We cluster and classify several datasets including Influenza A, primates, and human rhinovirus (HRV) datasets to build up the phylogenetic trees. Results show that the new proposed cumulative Fourier power spectrum is much faster and more accurately than MSA and another alignment-free method known as k-mer. The research provides us new insights in the study of phylogeny, evolution, and efficient DNA comparison algorithms for large genomes. The computer programs of the cumulative Fourier power spectrum are available at GitHub (https://github.com/YaulabTsinghua/cumulative-Fourier-power-spectrum).

1. Introduction

In molecular biology, mathematical methods are often used to interpret biological sequence information. Mathematics can transform biological sequences into numerical representations to analyze them quantitatively. Genetic recombination and, in particular, genetic shuffling are at odds with sequence comparison by alignment, which assumes conservation of contiguity between homologous segments (Vinga and Almeida, 2003). Apart from the common Multiple Sequence Alignment method (MSA), which is usually accurate but can take much time to compute large genomic data, many alignment-free methods have been proposed in recent years such as the Feature Vector model (Delibas and Seker, 2017), Chaos Game Representation (Almeida et al., 2001; Jeffrey, 1990), and the Maximum entropy method (Chan et al., 2010). Another new method via nucleotide-based Fourier power spectrum (PS) was proposed by Yau (Zhao et al., 2011). In this method, discrete Fourier transformation (DFT) and the moment vectors are used. Given a DNA sequence, indicator functions are involved to decide a corresponding mathematical representation. The indicator functions consist of four separate sequences which show the distribution of the four nucleotides respectively. Then by DFT, some frequency properties of these sequences can be observed. By moment vectors, sequences are transformed to points in the space. Less storage space is needed compared to the MSA method, which is a significant benefit when large genomic data is analyzed.

Despite achieving some accurate classification by the previous PS method, in the process of transforming the power spectrum to moment vectors, the mapping between the moment vectors and power spectra is not one-to-one, namely, one cannot recover full power spectra given the information of moment vectors. To improve the PS method, we now propose cumulative Fourier power spectrum (CPS). This new method still holds the advantage that large genomic datasets can be handled. Moreover, because the CPS is increasing, full power spectra can be recovered directly from the moment vectors of CPS. Therefore, more

Abbreviations: MSA, Multiple Sequence Alignment; HRV, human rhinovirus; PS, Fourier power spectrum; DFT, discrete Fourier transform; CPS, cumulative Fourier power spectrum; COI, cytochrome oxidase subunit I; HA, hemagglutinin; NA, neuraminidase

[*] Corresponding author at: Department of Mathematics, Tsinghua University, Beijing 100084, China.
E-mail address: yau@uic.edu (S.S.-T. Yau).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.gene.2018.06.042
Received 7 March 2018; Received in revised form 12 June 2018; Accepted 14 June 2018
Available online 20 June 2018

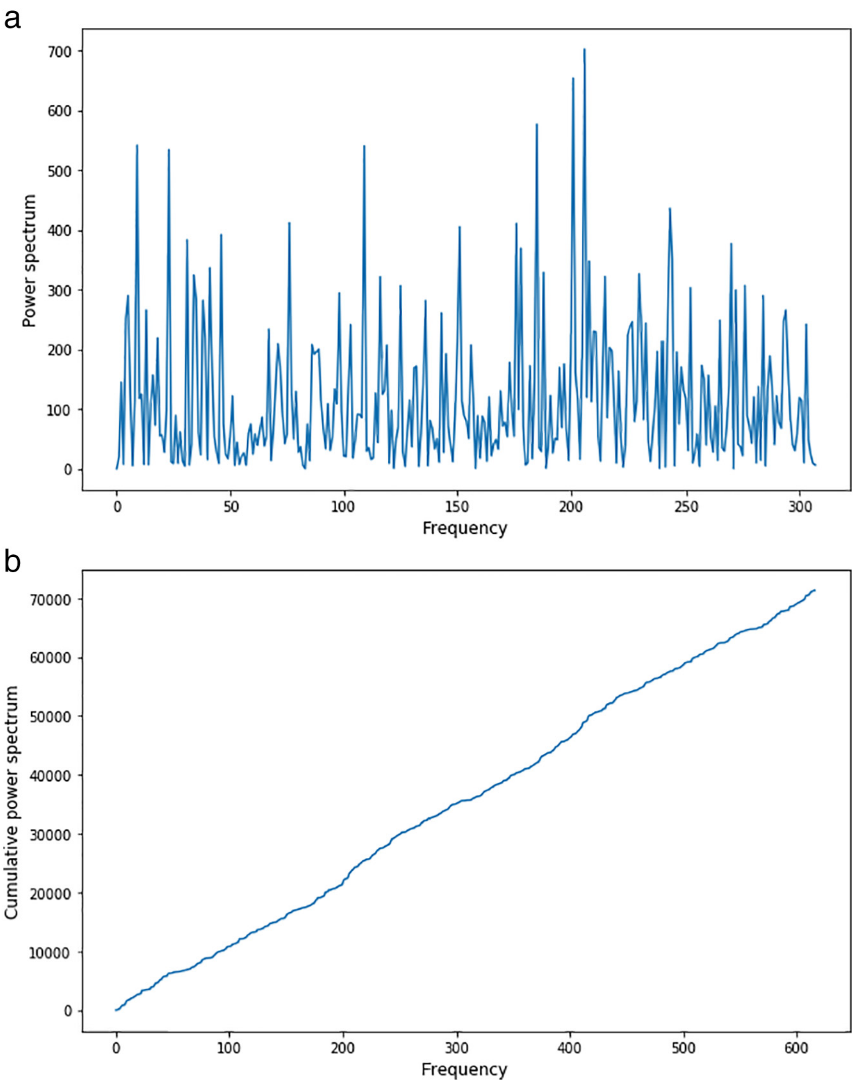


Fig. 1. (a) Fourier power spectrum and (b) cumulative Fourier power spectrum of nucleotide A of *Homo sapiens* cytochrome oxidase subunit I (COI) gene. The GenBank access number is EU834863.

information is preserved. When applied to genomic data, we find that CPS gives better results than other methods (MSA, k-mer and Power Spectrum method) for Influenza A, primates, and human rhinovirus (HRV) datasets.

This new proposed CPS method converts each DNA/RNA sequence into a point in the 16-dimensional space, and the comparison among sequences can be performed by calculating the distances among the points in Euclidean space. The transformation keeps the important information hidden in the original sequence, therefore reflects the real relationships among the sequences. Therefore, the CPS method provides us new insights in the study of phylogeny and evolution and efficient DNA comparison algorithms for large genomes.

2. Method

2.1. Indicator functions

Suppose we have some genomic sequences and use mathematical methods to assign them corresponding points in the Euclidean space. The first step is to represent the distribution of the four nucleotides. We define four indicator functions for adenine (A), cytosine (C), guanine (G) and thymine (T), respectively:

$$u_\alpha(n) = 1(\alpha \text{ appears at the } n-th \text{ location of the sequence})$$

where $\alpha = A, C, G, T$; n = 0, 1, 2, …, N − 1, and N is the length of the sequence.

For example, if the sequence is TTAAAACTGGAT, then it is represented by four separate indicator sequences as follows:

u_A: 001111000010
u_C: 000000100000
u_G: 000000001100
u_T: 110000010001

2.2. Cumulative Fourier power spectrum

Next, we apply the DFT to the four indicator sequences:

$$U_\alpha(k) = \sum_{n=0}^{N-1} u_\alpha(n) e^{-2\pi ikn/N}, \ \mathrm{k} = 0, 1, 2, \ldots, \mathrm{N} - 1.$$

then, consider the power spectrum

$$PS_\alpha(k) = |U_\alpha(k)|^2$$

To use the cumulative Fourier power spectrum, we delete PS(0) since it is a constant term. Also, it is much larger than the other terms. Now we have

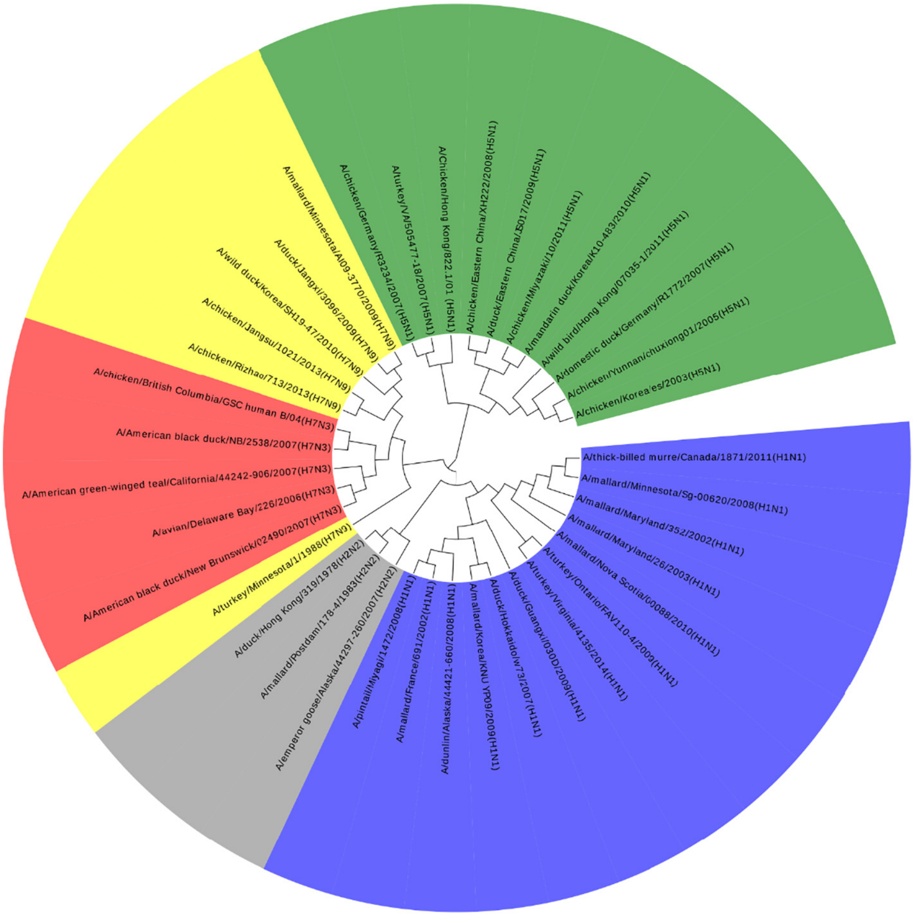

A/mallard/Minnesota/AI09-3770/2009(H7N9)
A/chicken/Germany/R3234/2007(H5N1)
A/turkey/VA/505477-18/2007(H5N1)
A/Chicken/Hong Kong/822.1/01 (H5N1)
A/chicken/Eastern China/XH222/2008(H5N1)
A/duck/Eastern China/JS017/2009(H5N1)
A/chicken/Miyazaki/10/2011(H5N1)
A/mandarin duck/Korea/K10-483/2010(H5N1)
A/wild bird/Hong Kong/07035-1/2011(H5N1)
A/domestic duck/Germany/R1772/2007(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/turkey/Minnesota/1/1988(H7N9)
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/Jangsu/1021/2013(H7N9)
A/wild duck/Korea/SH19-47/2010(H7N9)
A/duck/Jangxi/3096/2009(H7N9)

Fig. 2. Phylogenetic tree of 38 Influenza A viruses by the cumulative Fourier power spectrum and UPGMA method. The viruses with same HA and NA proteins are in the same color.

$$CPS_\alpha(k) = \sum_{n=1}^{k} PS_\alpha(n),\ k = 1, 2, \ldots, N-1$$

2.3. Moment vectors

Different sequences have different lengths, so the numbers of their CPS terms are not the same and thus the Euclidean distance between two DNA sequences of unequal lengths cannot be defined. Therefore, we do not use the power spectrum directly. Instead, we use moment vectors of the power spectra of DNA sequences so that the sequences can be assigned by special points in the same dimensional space:

$$M_j^\alpha = a_j^\alpha \sum_{k=1}^{N-1} (CPS_\alpha(k))^j$$

$where\ j = 1, 2, \ldots and\ \alpha = A, C, G, T.$

Here, a_j^α is the scale factor to decide the definition of the moment vectors, where j is the moment ordinal. Previous studies (Zhao et al., 2011; Hoang et al., 2015) suggest that the new scale factor should be related to the characteristics of the sequences and make the moment vectors converge to zero as j increases.

According to Parseval's theorem (Oppenheim and Schafer, 1989), we have

$$\sum_{k=0}^{N-1} PS_\alpha(k) = N \sum_{n=0}^{N-1} |u_\alpha(n)|^2$$

We also have

$$\sum_{n=0}^{N-1} |u_\alpha(n)|^2 = N_\alpha$$

where N_α is the number of the nucleotide α in the sequence.

Therefore, we have

$$\sum_{k=0}^{N-1} PS_\alpha(k) = N \sum_{n=0}^{N-1} |u_\alpha(n)|^2 = NN_\alpha$$

Besides, the following equation can be easily obtained:

$$PS_\alpha(0) = \left| \sum_{n=0}^{N-1} u_\alpha(n) \right|^2 = N_\alpha^2$$

Now we have

$$\sum_{n=1}^{N-1} CPS_\alpha(n) = \textstyle\sum_{n=0}^{N-1} \displaystyle\sum_{k=0}^{n} PS_\alpha(k) - NPS_\alpha(0) \le \mathrm{N} \sum_{k=0}^{N-1} PS_\alpha(k) - NPS_\alpha(0)$$
$$= NN_\alpha(N - N_\alpha)$$

So we may choose the scale factor as a power of $\frac{1}{NN_\alpha(N-N_\alpha)}$.

We now prove $\left(\frac{1}{NN_\alpha(N-N_\alpha)}\right)^{j-1}$ is the optimal scale factor if we consider a power of $\frac{1}{NN_\alpha(N-N_\alpha)}$. In this case, we have

$$M_j^\alpha = a_j^\alpha \sum_{k=1}^{N-1} (CPS_\alpha(k))^j = (NN_\alpha(N - N_\alpha)) \sum_{k=1}^{N-1} \left(\frac{CPS_\alpha(k)}{NN_\alpha(N - N_\alpha)}\right)^j$$

Because

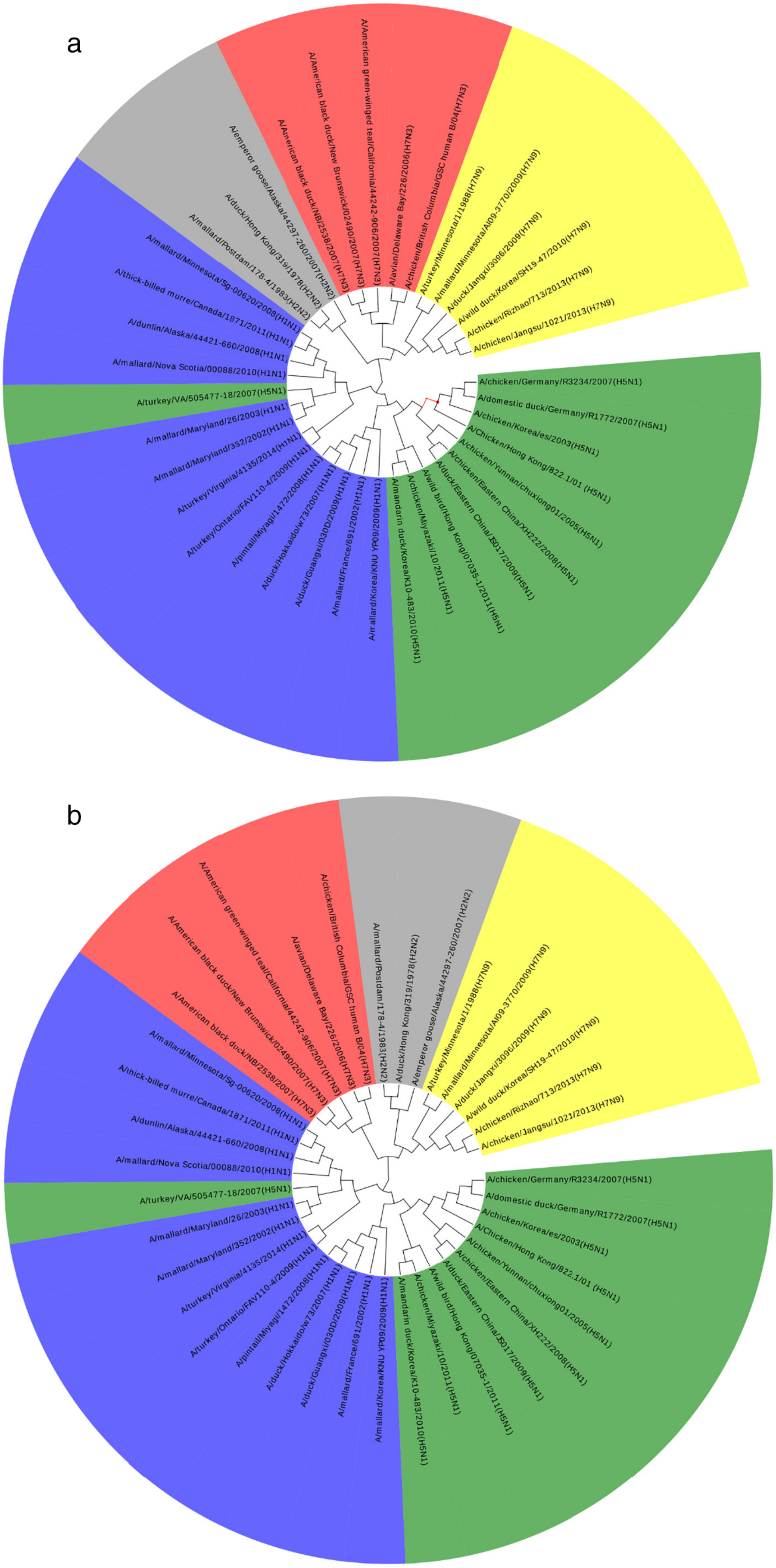

a
A/emperor goose/Alaska/44297-260/2007(H2N2)
A/duck/Hong Kong/319/1978(H2N2)
A/mallard/Postdam/178-4/1983(H2N2)
A/American black duck/NB/2538/2007(H7N3)
A/American black duck/New Brunswick/02490/2007(H7N3)
A/American green-winged teal/California/44242-906/2007(H7N3)
A/avian/Delaware Bay/226/2006(H7N3)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

Fig. 3. Phylogenetic tree of 38 Influenza A viruses by the k-mer and UPGMA method. (a) k = 6 (b) k = 7.

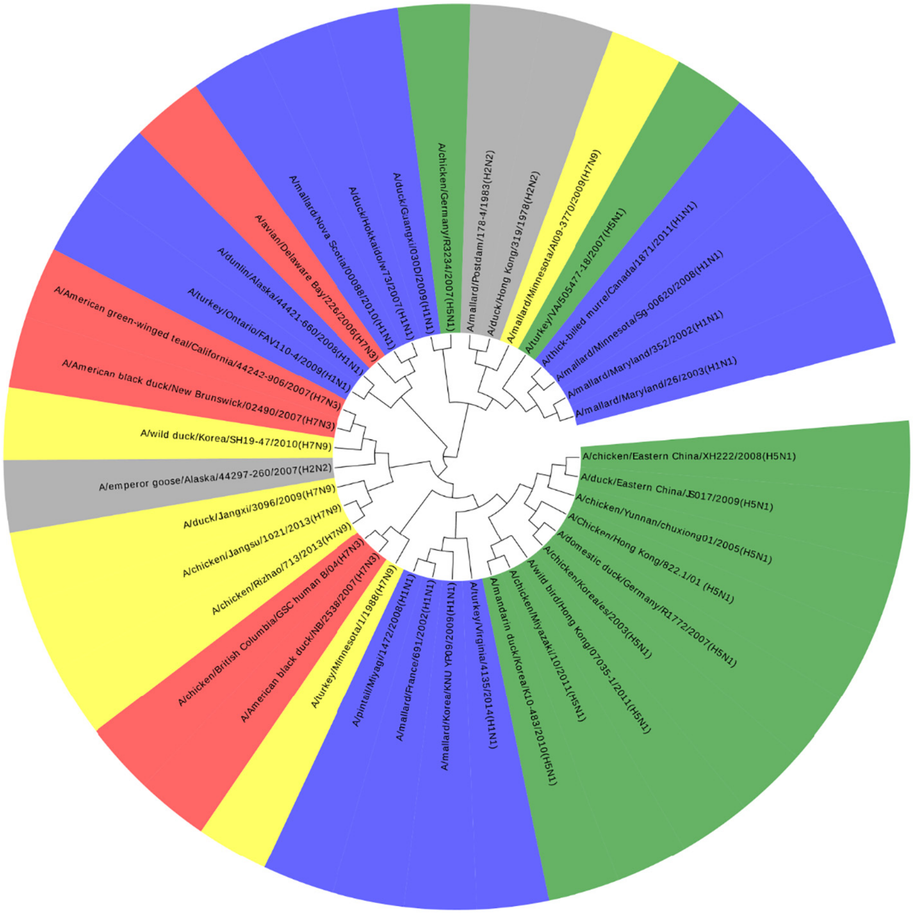

A/American green-winged teal/California/44242-906/2007(H7N3)
A/American black duck/New Brunswick/02490/2007(H7N3)
A/turkey/Ontario/FAV110-4/2009(H1N1)
A/dunlin/Alaska/44421-660/2008(H1N1)
A/avian/Delaware Bay/226/2006(H7N3)
A/mallard/Nova Scotia/00088/2010(H1N1)
A/duck/Hokkaido/w73/2007(H1N1)
A/duck/Guangxi/030D/2009(H1N1)
A/chicken/Germany/R3234/2007(H5N1)
A/mallard/Postdam/178-4/1983(H2N2)
A/duck/Hong Kong/319/1978(H2N2)
A/mallard/Minnesota/AI09-3770/2009(H7N9)
A/turkey/VA/505477-18/2007(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/chicken/Eastern China/XH222/2008(H5N1)
A/duck/Eastern China/JS017/2009(H5N1)
A/chicken/Yunnan/chuxiong01/2005(H5N1)
A/Chicken/Hong Kong/822.1/01 (H5N1)
A/domestic duck/Germany/R1772/2007(H5N1)
A/chicken/Korea/es/2003(H5N1)
A/wild bird/Hong Kong/07035-1/2011(H5N1)
A/chicken/Miyazaki/10/2011(H5N1)
A/mandarin duck/Korea/K10-483/2010(H5N1)
A/turkey/Virginia/4135/2014(H1N1)
A/mallard/Korea/KNU YP09/2009(H1N1)
A/mallard/France/691/2002(H1N1)
A/pintail/Miyagi/1472/2008(H1N1)
A/turkey/Minnesota/1/1988(H7N9)
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/Jangsu/1021/2013(H7N9)
A/duck/Jangxi/3096/2009(H7N9)
A/emperor goose/Alaska/44297-260/2007(H2N2)
A/wild duck/Korea/SH19-47/2010(H7N9)

Fig. 4. Phylogenetic tree of 38 Influenza A viruses by the MSA and UPGMA method.

$$\sum_{k=1}^{N-1}\frac{CPS_\alpha(k)}{NN_\alpha(N-N_\alpha)}\le 1$$

it is not difficult to find that

$$\sum_{k=1}^{N-1}\left(\frac{CPS_\alpha(k)}{NN_\alpha(N-N_\alpha)}\right)^j$$

converges to zero.

If $a_j^\alpha=\left(\frac{1}{NN_\alpha(N-N_\alpha)}\right)^j$, then $M_1{}^\alpha\le 1$ and the moment vectors converge to zero too quickly. Therefore, a lot of information is lost.

If $a_j^\alpha=\left(\frac{1}{NN_\alpha(N-N_\alpha)}\right)^{j-2}$, then

$$M_1^\alpha=(NN_\alpha(N-N_\alpha))\sum_{n=1}^{N-1}CPS_\alpha(n)\ge(NN_\alpha(N-N_\alpha))\sum_{n=1}^{N-1}PS_\alpha(n)$$

$$=(NN_\alpha(N-N_\alpha))\left(\sum_{n=1}^{N-1}PS_\alpha(n)-PS_\alpha(0)\right)$$

$$=(NN_\alpha(N-N_\alpha))(NN_\alpha-N_\alpha^2)=NN_\alpha^2(N-N_\alpha)^2$$

and the moment vectors converge to zero slowly. This equation suggests that a large number of moments are needed to preserve information. But a large number of moments may take much storage space and computational time.

If $\left(\frac{1}{NN_\alpha(N-N_\alpha)}\right)^{j-1}$ is chosen, we have

$$M_1^\alpha=a_1^\alpha\sum_{k=1}^{N-1}CPS_\alpha(k)=\sum_{k=1}^{N-1}CPS_\alpha(k).$$

Since CPS_α is cumulative, $M_1{}^\alpha$ becomes large if $\alpha_1{}^\alpha=1$.

Therefore, the scale factor is chosen as

$$\frac{1}{N(NN_\alpha(N-N_\alpha))^{j-1}}=\frac{1}{(N_\alpha(N-N_\alpha))^{j-1}N^j}$$

for our CPS method.

So we have

$$M_j^\alpha=\frac{1}{(N_\alpha(N-N_\alpha))^{j-1}N^j}\sum_{k=1}^{N-1}(CPS_\alpha(k))^j$$

In (Zhao et al., 2011), both moment vectors and central moment vectors were used in the PS method. For the CPS method, we also consider both.

The mean value of the CPS is defined as

$$Mean_\alpha=\frac{1}{N-1}\sum_{n=1}^{N-1}CPS_\alpha(n)$$

For central moment vectors, we consider the absolute value because otherwise, the first central moment vector would be zero. Using the same scale factor, the central moment vectors are defined as follows,

$$CM_j^\alpha=\frac{1}{(N_\alpha(N-N_\alpha))^{j-1}N^j}\sum_{k=1}^{N-1}|CPS_\alpha(k)-Mean_\alpha|^j$$

When calculating the moments of the cumulative Fourier power spectra of genomic sequences, we found that the moment vectors and central moment vectors from the third are very small compared to the first and second moments. Due to this observation, we only consider the initial two moment vectors and the first two central moment vectors, giving every sequence its 16-dimensional point

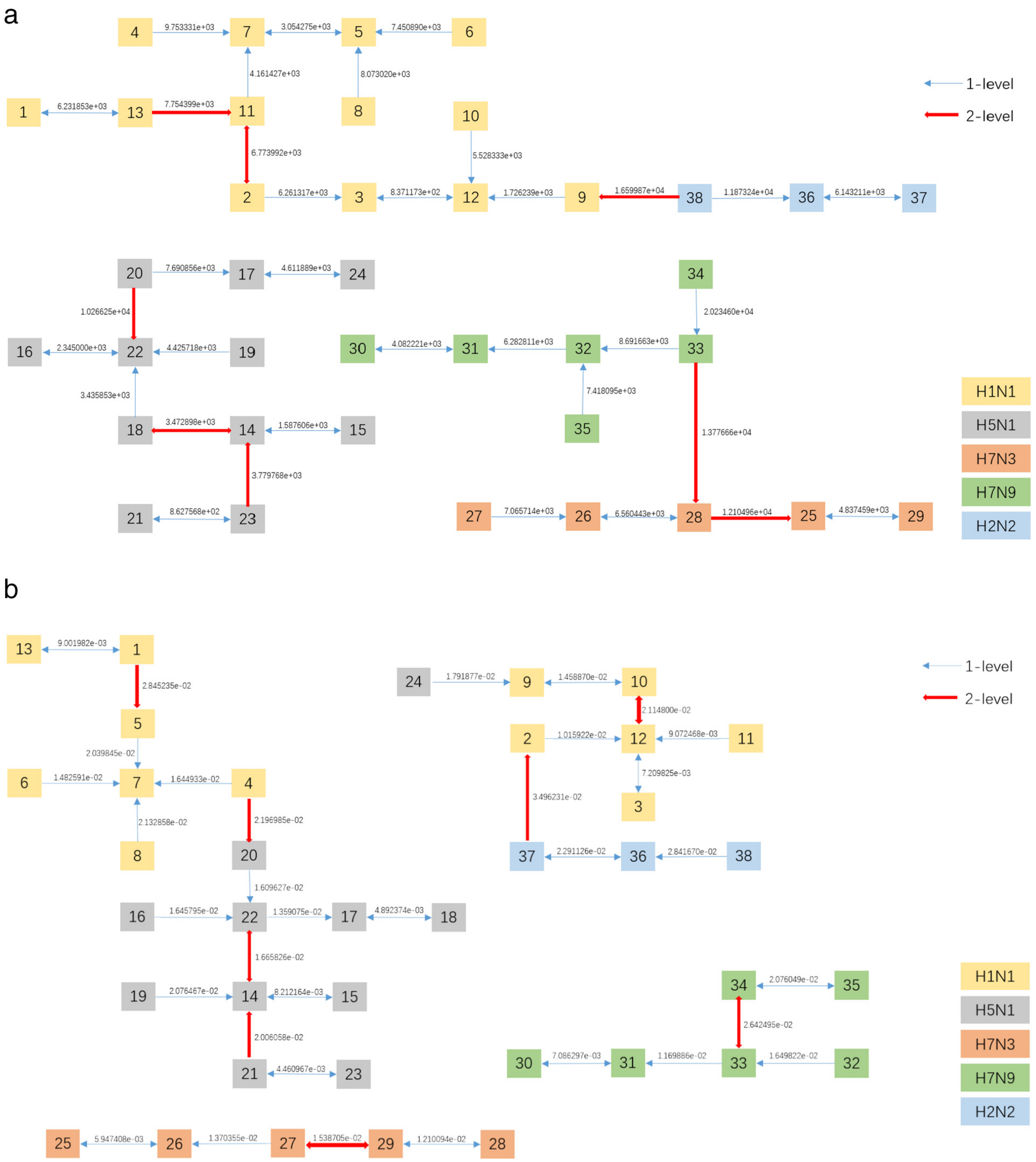


Fig. 5. Natural graph of 38 Influenza A viruses. Five classes can be distinguished by different colors as labeled on the right. (a) CPS (b) k-mer (k = 6).

$$(M_1^A, M_2^A, CM_1^A, CM_2^A, M_1^C, M_2^C, CM_1^C, CM_2^C, M_1^G, M_2^G, CM_1^G, CM_2^G, M_1^T, M_2^T, CM_1^T, CM_2^T)$$

in the Euclidean space. We name this method as the cumulative Fourier power spectrum to distinguish it from the traditional Power Spectrum. The major improvement from the traditional Power Spectrum method is that, mathematically, CPS and moment vectors can be computed from each other, while the power spectrum cannot achieve this. Thus we keep more important information in the original sequence during the transformation from the original sequence to numerical sequences. For more details, please see Appendix. B.

Algorithm 1 below shows the whole procedure.

Algorithm 1. Calculate the moment and central moment vectors.

```
Input: a DNA sequence consisting of A,C,G,T

1. Calculate the indicator functions $u_A$, $u_C$, $u_G$, $u_T$
2. Apply DFT on the $u_A$, $u_C$, $u_G$, $u_T$ and get $U_A$, $U_C$, $U_G$, $U_T$
3. Calculate the power spectrum of $U_A$, $U_C$, $U_G$, $U_T$ as $PS_A$, $PS_C$, $PS_G$, $PS_T$
4. Add up the power spectra up and get $CPS_A$, $CPS_C$, $CPS_G$, $CPS_T$
5. Calculate the Moment vectors of four nucleotides and the central
   moment vectors

Output: the moment vectors and central moment vectors of four
nucleotides
```

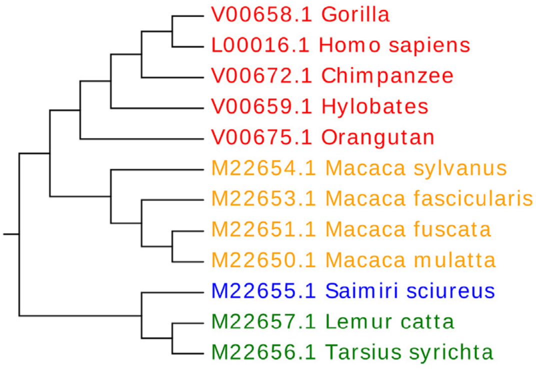


Fig. 6. Phylogenetic tree of 12 primates by the cumulative Fourier power spectrum and UPGMA method. All species from same group are clustered together.

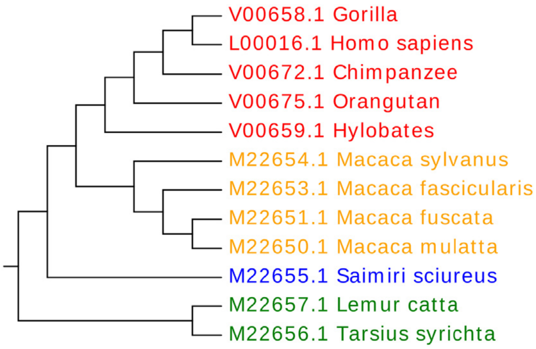


Fig. 7. Phylogenetic tree of 12 primates by the k-mer and UPGMA method. All species from the same group are clustered together.

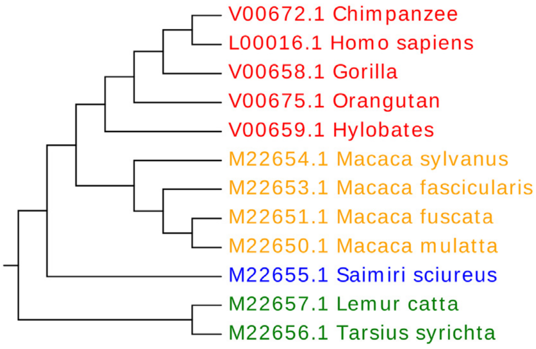


Fig. 8. Phylogenetic tree of 12 primates by the MSA and UPGMA method. All species from the same group are clustered together.

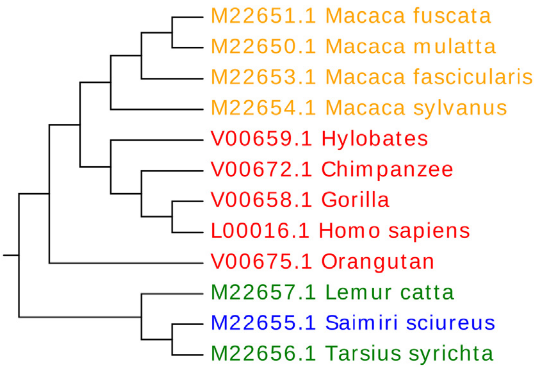


Fig. 9. Phylogenetic tree of 12 primates by the original PS and UPGMA method.

2.4. An example of Homo sapiens cytochrome oxidase subunit I (COI) gene

Here we take *Homo sapiens* cytochrome oxidase subunit I (COI) gene as an example to show the Fourier power spectrum and cumulative power spectrum. The original sequence of this gene can be found in Appendix A. Fig. 1 shows the result of nucleotide A. In Fig. 1(a), it reaches the highest peak at a frequency around 205, which indicates the periodic signature of the sequence. Fig. 1(b) shows that the cumulative Fourier power spectrum is increasing as frequency increases, and ascension is a crucial advantage compared to the traditional power spectrum method.

2.5. Clustering method

We calculate the Euclidean distances between every two points. Based on the distance matrix, we use the unweighted pair-group method with arithmetic means (UPGMA) method to draw phylogenetic trees in Mega 7 (Sneath and Sokal, 1973; Kumar et al., 2016).

3. Results

We apply our cumulative Fourier power spectrum method on several datasets, and compare the results with MSA (Larkin et al., 2007; Edgar and Batzoglou, 2006), k-mer (Yu et al., 2010) and power spectrum proposed in (Hoang et al., 2015). The results below show that our method outperforms other popular methods with higher accuracy on species clustering.

3.1. Influenza A virus

Influenza A viruses are single-stranded RNA viruses, which have been a major health threat to both human society and animals. Influenza A viruses' nomenclature is based on the surface glycoproteins: hemagglutinin(HA) and neuraminidase(NA). HA has 15 subtypes and NA has 9 subtypes, which forms 135 different combinations. Here we test a dataset of 38 sequences from segment 6 of Influenza A virus genomes, which are from some of the most lethal subtypes like H1N1, H2N2, H5N1, H7N9, and H7N3. The results are shown in Fig. 2 and agree with previous work in (Hoang et al., 2015).

We also tested the k-mer method and MSA on this dataset and the different results are shown in Figs. 3 and 4, respectively. The optimal choice of k is 6 or 7, according to the principle in (Sims et al., 2008). In Fig. 3(a) and Fig. 3(b), the H5N1 and H1N1 viruses are mixed up in the phylogenetic tree. There are five H1N1 viruses *A/pintail/Miyagi/1472/2008(H1N1), A/duck/Hokkaido/w73/2007(H1N1), A/duck/Guangxi/030D/2009(H1N1), A/mallard/France/691/2002(H1N1), A/mallard/Korea/KNU YP09/2009(H1N1)* that are in the branch of H5N1 viruses. There is also an *A/turkey/VA/505477-18/2007 (H5N1)* in the branch of H1N1. Meanwhile, all viruses are mixed up in the phylogenetic tree constructed by the MSA method as shown in Fig. 4.

To get a direct image of the relationships between Influenza A viruses, we make the Natural Graph of them. Natural Graph was first introduced in (Yu et al., 2013). Distance matrices are usually used for phylogenetic analysis of DNA and proteins. Many algorithms may produce either rooted or unrooted phylogenetic trees based on the distance matrices. For example, the neighbor-joining algorithm produced unrooted trees, while the UPGMA algorithm produces rooted trees. Given a distance matrix, the resulting trees are not unique for any existing tree construction methods. Thus, the phylogenetic results are inconsistent due to the above two basic problems. However, the Natural Graph representation is unique and the direction in the graph can show the closest elements of each element based on their biological distance. Thus we apply the Natural Graph representation on the proposed CPS method and k-mer method. In Fig. 5, the blue lines represent the 1-level connected components and the red ones represent 2-level. Virus classes are marked in different colors and it is obvious that in Fig. 5(a), after

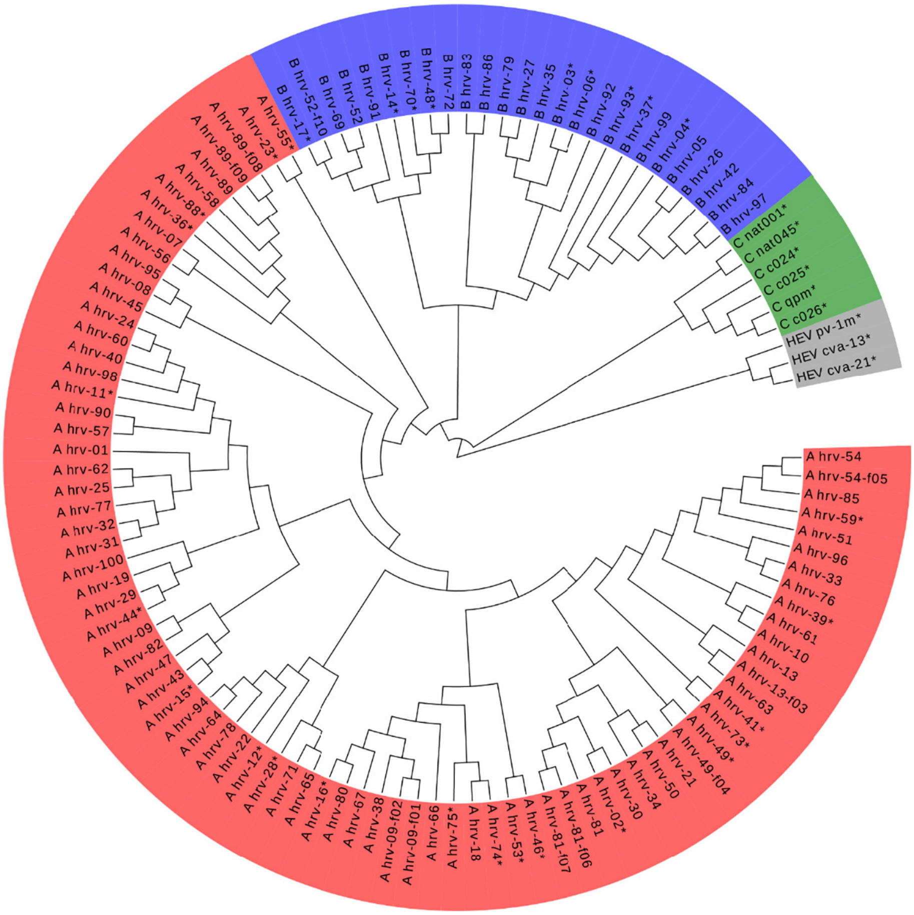


Fig. 10. Phylogenetic tree of 113 HRV and 3 HEV dataset by the cumulative Fourier power spectrum and UPGMA method. HRV-A viruses are colored in red, HRV-B in purple, HRV-C in green and the outgroup HEV viruses are in grey. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the construction of two levels, the Influenza A viruses with the same H and N are clustered together. The construction of 2-level Natural Graph requires each group in 1-level to find its nearest neighbor group, which results in the arrow from *No.38* and *No.9*, and from *No.33* to *No.28*. However, the distances between the different virus classes are all larger than the distances within each class, such as the distance between *No.14* and *No.18*, *No.14* and *No.23*.

However, in Fig. 5(b), H5N1 and H1N1 are not distinguished by the k-mer method. The H1N1 is divided into two parts, and *No.24* from H5N1 class is clustered in the wrong class after 2-level construction of Natural Graph.

3.2. Primates

The dataset, which is also used in (Yin et al., 2014) consists of four species of old-world monkeys (*Macaca fascicular, Macaca fuscata, Macaca sylvanus, Macaca mulatta*), one species of new-world monkeys (*Saimiri scirueus*), two species of prosimians (*Lemur catta, Tarsisus syrichta*), and five hominoid species (*Human, Chimpanzee, Gorilla, Orangutan, and Hylobates*). We use the NADH dehydrogenase subunit 4 genes of 12 species of four different groups of primates.

The results by the CPS, k-mer (k = 6), and MSA method are shown in Fig. 6, Fig. 7, and Fig. 8, respectively. All three methods can distinguish the four classes, while the CPS only requires about one-sixth of the time that k-mer takes, less than one-tenth of the time that MSA takes. The statistics of time required by each method are shown in Section 3.5.

The traditional PS method gives a less satisfying result as shown in Fig.9, where the prosimians are clustered together with the new-world monkey (*Saimiri scirueus*). Since the corresponding numerical vectors should reflect the most important information hidden in the original sequence, the wrong positions of phylogenetic tree indicate that the corresponding vectors cannot capture useful information. Therefore, the previous PS method must have lost some useful information during the transformation from the original sequence to numerical vectors. This shows that the new proposed CPS method has collected more information than the PS method.

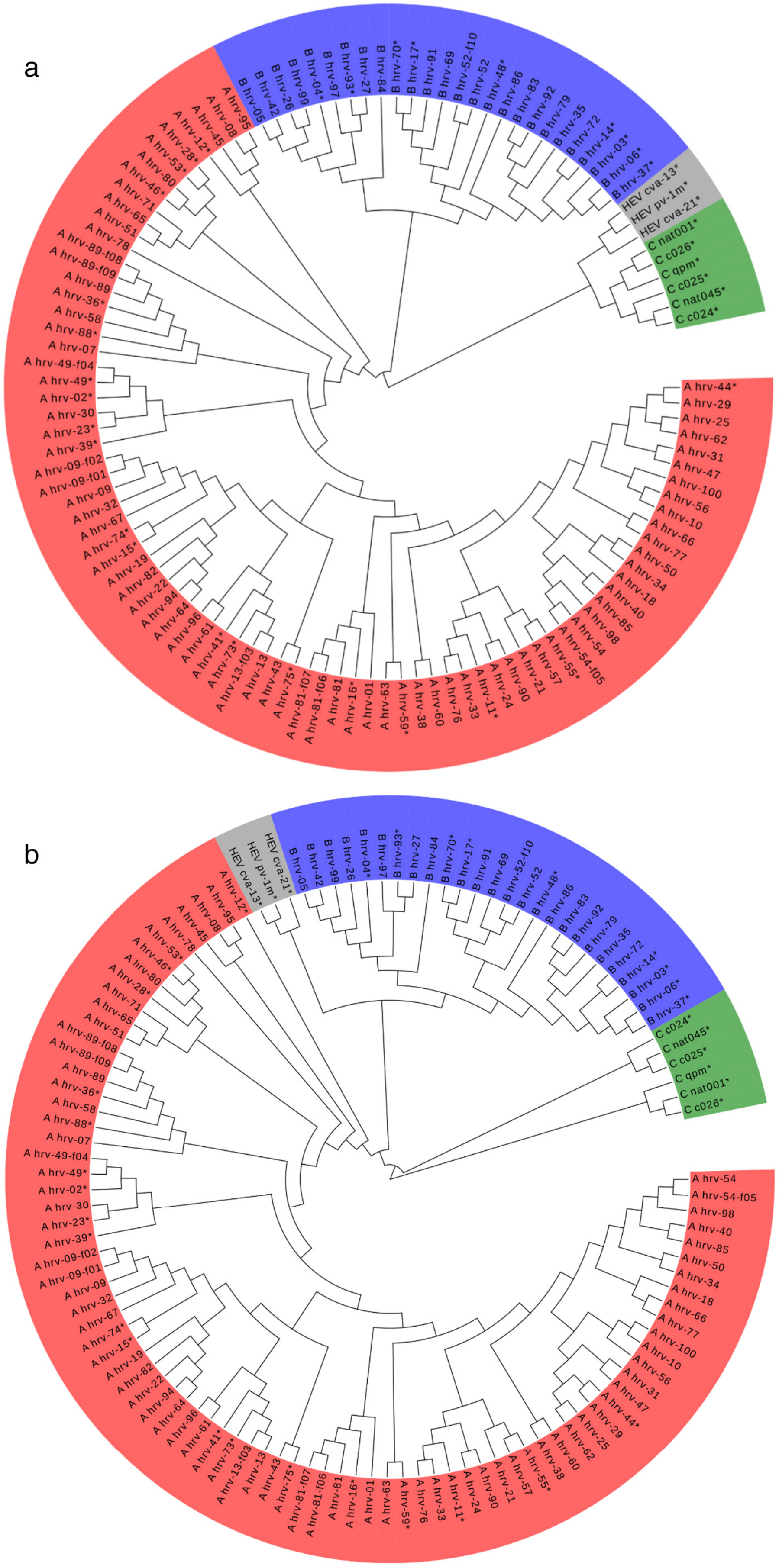
3.3. HRV (human rhinovirus)

HRVs (human rhinoviruses) are the major cause of common cold. Like influenza A viruses, HRVs are also classified based on their serotypes. There are 99 types of HRVs found now.

The complete HRV genomes consist of three different groups, i.e. HRV-A, HRV-B and HRV-C, 113 genomes in total. We also consider three HEV-Cs (Hepatitis E virus-C) which serve as the outgroup sequences. Thus the ideal phylogenetic tree should first classify the outgroup and HRV group, then divide the HRV-A, HRV-B and HRV-C. The results by the CPS method is shown in Fig. 10, where each class is marked with different colors. Meanwhile, the k-mer and MSA methods do not perform well on this dataset, as shown in Fig. 11 and Fig. 12. In Fig. 11, we take k = 7 and k = 8 and it fails to distinguish the HRVs and HEV group. MSA captures no useful information in the original sequence, since all the species are shuffled in the phylogenetic tree in Fig. 12.

3.4. Identification of exon/introns

Our CPS method can also be applied on the identification of exon/introns. We select the CCR9 gene from chromosome 3 of *Homo sapiens*, which encodes the C-C chemokine receptor type 9 protein. This protein is a member of the beta chemokine receptor family and chemokines and



(caption on next page)

Fig. 11. Phylogenetic tree of 113 HRV and 3 HEV dataset by the k-mer and UPGMA method. The outgroup viruses cannot be separated from the other viruses in this figure. (a) k = 7 (b) k = 8.

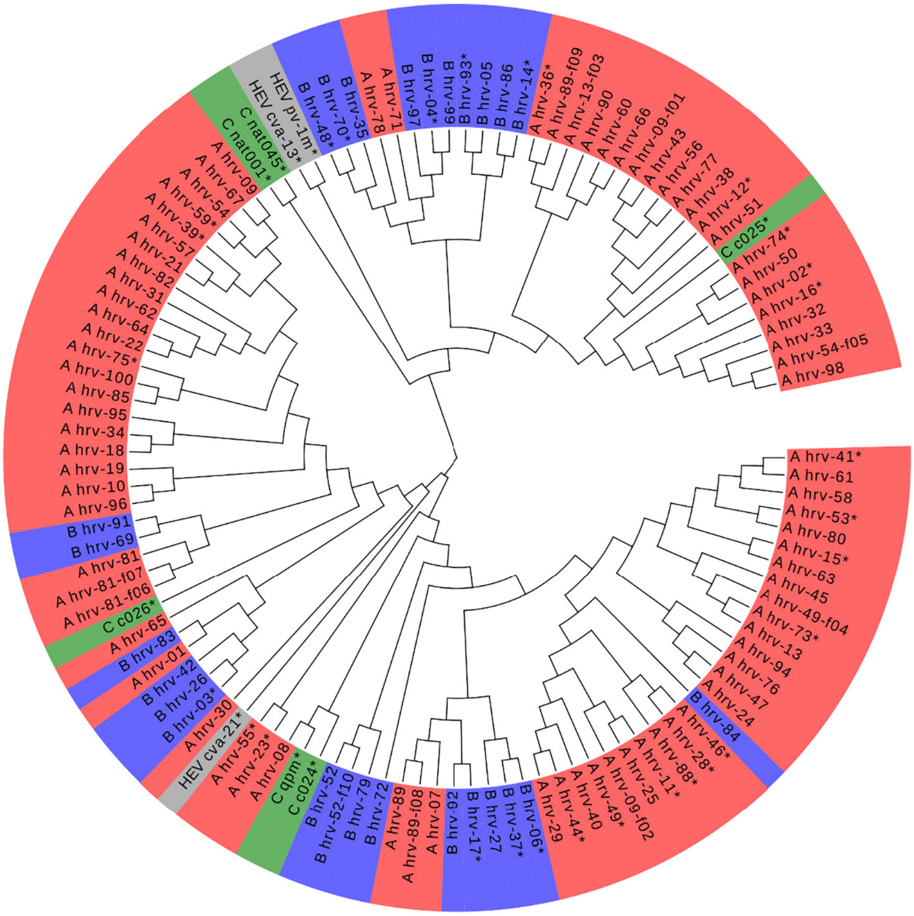


Fig. 12. Phylogenetic tree of 113 HRV and 3 HEV dataset by the MSA and UPGMA method. All viruses are mixed together and MSA cannot distinguish any class.

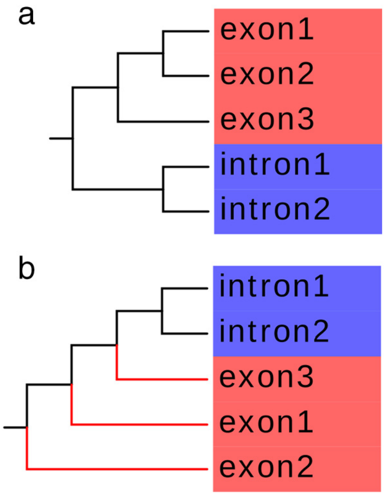


Fig. 13. Phylogenetic tree of exons and introns in gene CCR9 by the (a) CPS and (b) k-mer (k = 5) and Neighbor-Joining method. The CPS can distinguish exons and introns while the k-mer method clusters all segments together. (a) CPS (b) k-mer.

Table 1
Benchmark performance comparison of the three methods.

Dataset	Average length (bp)	Number of species	CPS (seconds)	k-mer (seconds)	MSA (seconds)
Influenza A	1406.8	38	3.76	62.77 (k = 6) 221.59 (k = 7)	195.32
Primates	895.5	12	0.93	6.28 (k = 5) 15.32 (k = 6)	11.28
HRV	7153.7	116	163.05	1778.26 (k = 7) 7777.52 (k = 8)	36,598

their receptors are key regulators of thymocyte migration and maturation in normal and inflammatory conditions. CCR9 gene consists of three exons and two introns. Both the proposed CPS method and k-mer method are applied on those five segments. The classification results are shown in Fig. 13. Considering that UPGMA assumes all the species involve at the same speed, which is not the case in this part, we use Neighbor-Joining tree (Saitou and Nei, 1987) to construct the phylogenetic trees. Clearly in Fig. 13(a), the three exons are clustered together and the two introns are together. However, exons and introns cannot be distinguished by the k-mer method.

3.5. Time statistics and complexity analysis

We record the time that the CPS, k-mer and MSA methods require on the three datasets we tested above and the result is shown in Table 1. We perform all the calculations on the same machine and clear the memory each time to avoid redundancy and influence on the next-step calculation. The computation environment is Intel(R) Core(TM) i7-5500U CPU @2.40 GHz Windows10 PC with 8.00 GB RAM.

From the table, we can draw the conclusion that the alignment-free methods are much more time-efficient than the alignment method. However, the time that k-mer requires depends on the value of k and becomes unacceptable when k becomes too large. The k-mer method produces a vector of length 4^k for each sequence, and the cost on the storage and memory on the computer is very heavy when $k \geq 5$, especially when the whole genomes are input for analysis. Meanwhile, the proposed CPS assign a vector in 16-dim space to each sequence, which saves much time and storage for the next-step computation. The CPS method is based on the Fourier transform and we apply the Fast Fourier Transform(FFT) algorithm in our calculation. The FFT algorithm is mature and widely used for many applications in engineering, science, and mathematics. If we denote n as the product of number of species and the average length, the complexity of our proposed cumulated Fourier power spectrum method is O(nlog*n*) using the Fast Fourier Transform, while the time for alignment will increase exponentially as n increases. n can be considered as the scale of original data since each nucleotide must be input and read by any algorithm. The MSA method is actually an NP-hard computational optimization problem which is implausible for a huge amount of sequences (Wang and Jiang, 1994).

When a new unknown species is added into the dataset, we only need to calculate the distance between the new species and the previous ones to get the relationship between them. In other words, the new-added species doesn't affect the relationships of previous species. However, the MSA method requires the reconstruction of alignment result, because the alignment depends on the identification of homologous positions (Edgar and Batzoglou, 2006).

4. Conclusion and discussion

In the Method section, we prove that CPS is better than PS theoretically. In the process of using mathematical methods to analyze biological data, we always want to preserve more information in raw data. CPS makes it possible to recover the power spectra of the DNA sequences while PS cannot, and that illustrates its better performance on real datasets.

The performance on datasets of various species and scales proves that the new proposed CPS method can be applied on large genomes and produce accurate results with high time-efficiency. It provides a new quantitative way of analyzing evolutionary relationships among species in molecular biological study.

MSA algorithm can be seen as a generalization of pairwise sequence alignment, in which, instead of aligning two sequences, k sequences are aligned simultaneously. However, the alignment itself requires a given scoring matrix, which represents the penalty functions for gap and mismatch. The assigning of scores influences the result of alignment directly. However, there is no clear biological explanation on the scoring matrix yet. Inappropriate scoring matrix will cause errors in alignment, which means that an ancestral position has not been identified correctly, and consequently inferences of the number of substitutions will be incorrect. Alignment is the first step in many evolutionary studies, and the errors can amplify in later computational stages. We consider this as the main cause for the bad performance on large datasets such as the Influenza A and HRV datasets. For smaller datasets like primates, alignment is very easily obtained without much confusion. When sequences get longer and the species become more, the performance of alignment methods will get worse. Besides, the MSA method is an NP-hard algorithm, which deeply weakens its strength in the genomics analysis.

The main reason that the CPS method performs better than the k-mer method is that, k-mer only captures the frequency of each word in the k-dictionary, while it ignores the positions of the k-string word. In biology, the nucleotides are arranged on the genome orderly, thus the position information is essential when analyzing sequence similarity. Two different sequences may show no difference when calculating the frequency of each k-mer. Here we take the VIPR1-AS1 gene as an example. This gene has only 1711 bp and we generate two artificial mutations in the sequence. We add one small segment of 'AAAAACCCCC GGGGGTTTTT' at the position 5 (near the start of the gene) in the original sequence and denote this as mutation-1. The same segment is added to the original sequence at position 1663 (near the end of the gene) and denoted as mutation-2. Since the segment is added at different positions, the numerical vector should be able to detect the difference between the two mutations. However, the corresponding k-mer ($k = 5$) cannot capture this difference, since the two mutational sequences produce the same k-mer vector. This is one of the main information that k-mer fails to collect, since it only captures the frequency of each k-mer, without recording the positions of each mutation. However, their corresponding CPS vectors are different and the distance between them is not zero. The sequences of the original VIPR1-AS1 gene and two mutations can be found in Appendix A.

Although our CPS method performs well on several datasets, there are still some small issues. For example, in Fig. 2, we suspect the *A/turkey/Minne-sota/1988(H7N9)* as the main cause of the outbreak of H7N9 in 2013, since it locates on the cluster of H7N9, and H7N3. We may further deduce that it could be a variant from the H7N3 viruses on the surface neuraminidase(NA), and further produces the H7N9 variants since 2009. This conjecture needs more biological evidence to prove it correct. In the primates' dataset, four classes are all differentiated in Fig. 6, while *Homo sapiens* is considered to have the closest relationship with *Chimpanzee*, where the result given by MSA agrees with common sense in Fig. 8 (Gibbons, 2012).

Another issue is that during Fourier Transformation, the information is kept in both the power spectrum and the angle information. The proposed CPS method only captures information in the power spectrums while there must be some other useful information left in the angles. We expect higher accuracy when taking the angle information into consideration, which is also the focus of our future work.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 91746119) and Tsinghua University startup fund. Prof. Stephen Yau is grateful to National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done.

Also, we would like to thank Mr. Steve Yau for his technology help when running the files in the server; Yusi Zhang and Weifeng Sun for their help in proving the theorem; Ms. Xiujuan Yu and all the members in Professor Stephen S.-T. Yau's bioinformatics seminar for their help during the whole process of doing the project.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2018.06.042.

References

Almeida, J., Carrico, J., Maretzek, A., Noble, P., Fletcher, M., 2001. Analysis of genomic sequences by Chaos Game Representation. Bioinformatics 17 (5), 429–437.

Chan, R., Wang, R., Wong, J., 2010. Maximum entropy method composition vector method. In: Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, pp. 599–622.

Delibas, E., Seker, A., 2017. A new feature vector model for alignment-free similarity analysis of DNA sequences. In: Conference: 2nd World Conference on Technology, Innovation and Entrepreneurship.

Edgar, R., Batzoglou, S., 2006. Multiple sequence alignment. Curr. Opin. Struct. Biol. 16 (3), 368–373.

Gibbons, A., 2012. Bonobos join chimps as closest human relatives. Science 6, 13.

Hoang, T., Yin, C., Zheng, H., Yu, C., He, R., Yau, S., 2015. A new method to cluster DNA sequences using Fourier power spectrum. J. Theor. Biol. 372, 135–145.

Jeffrey, J., 1990. Chaos game representation of gene structure. Nucleic Acids Res. 18 (8), 2163–2170.

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.

Larkin, M., Blakshields, G., Brown, N., Chenna, R., Mcgettigan, P., Mcwilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., Higgins, D., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23 (21), 2947–2948.

Oppenheim, A., Schafer, R., 1989. Discrete-Time Signal Processing. 23(2). Prentice-Hall, pp. 157.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4 (4), 406–425.

Sims, G., Jun, S., Wu, G., Kim, S., 2008. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. 106 (8), 2677–2682.

Sneath, P., Sokal, R., 1973. Numerical Taxonomy. Freeman, San Francisco.

Vinga, S., Almeida, J., 2003. Alignment-free sequence comparison–a review. Bioinformatics 19 (4), 513–523.

Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1 (4), 337–348.

Yin, C., Chen, Y., Yau, S., 2014. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. J. Theor. Biol. 359, 18–28.

Yu, Z., Chu, K., Li, C., Anh, V., Zhou, L., Wang, R., 2010. Whole-proteome phylogeny of large dsDNA viruses and parvoviruses through a composition vector method related to dynamical language model. BMC Evol. Biol. 10, 192.

Yu, C., Deng, M., Cheng, S., Yau, S., He, R., Yau, S., 2013. Protein space: a natural method for realizing the nature of protein universe. J. Theor. Biol. 318, 197–204.

Zhao, B., Duan, V., Yau, S., 2011. A novel clustering method via nucleotide-based Fourier power spectrum analysis. J. Theor. Biol. 279, 83–89.