

Large-Scale Genome Comparison Based on Cumulative Fourier Power and Phase Spectra: Central Moment and Covariance Vector

Shaojun Pei^a, Rui Dong^a, Rong Lucy He^b, Stephen S.-T. Yau^{a,*}

^a Department of Mathematical Sciences, Tsinghua University, Beijing, PR China

^b Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA

ARTICLE INFO

Article history:

Received 1 May 2019

Received in revised form 24 June 2019

Accepted 10 July 2019

Available online 11 July 2019

Dedicate to Professor Micheal Waterman on the occasion of his 77th birthday.

Keywords:

Cumulative Fourier transform

Power and phase spectra

Central moments

Covariance

ABSTRACT

Genome comparison is a vital research area of bioinformatics. For large-scale genome comparisons, the Multiple Sequence Alignment (MSA) methods have been impractical to use due to its algorithmic complexity. In this study, we propose a novel alignment-free method based on the one-to-one correspondence between a DNA sequence and its complete central moment vector of the cumulative Fourier power and phase spectra. In addition, the covariance between the four nucleotides in the power and phase spectra is included. We use the cumulative Fourier power and phase spectra to define a 28-dimensional vector for each DNA sequence. Euclidean distances between the vectors can measure the dissimilarity between DNA sequences. We perform testing with datasets of different sizes and types including simulated DNA sequences, exon-intron and complete genomes. The results show that our method is more accurate and efficient for performing hierarchical clustering than other alignment-free methods and MSA methods.

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

The comparison of DNA sequences is an important approach for establishing the phylogenetic relationships of DNA sequences in bioinformatics research [1,2]. The Multiple Sequence Alignment (MSA) method has been used for classifying DNA and protein sequences, such as Clustal Omega [3], MAFFT [4] and MSAProbs [5]. While accurate results for some species can be obtained using the MSA methods, when the amount of biological sequence information increases, the processing time and memory requirements become excessive. Thus, as a more effective approach to handle large data, numerical alignment-free methods have gained increasing attention in analyzing biological sequences [6–10]. These methods require the transformation or mapping of biological sequences, usually represented as a string of characters (i.e., A, C, G, and T) to a numerical representation (i.e., a signal) that can be processed using mathematical functions [11]. For example, Voss (1992) proposed the binary indicators to convert a DNA sequence to four sequences of 0 and 1 that represent A, C, G, and T [12]. This technique has been applied repeatedly in several recent studies [8,13]. Another approach is to use feature (or k-mer) frequency profiles (FFP) of whole genomes for comparison. Upon choosing of appropriate value k , the optimum FFP method is applicable for comparing whole genomes or large

genomic regions even when there are no common genes with high homology [14,15]. However, with the long k -mer lengths, the k -mer type number increases exponentially, which exceeds the storage capacity of the computer [16].

Discrete Fourier Transform (DFT), which is one of the most common digital signal processing methods, has also been used in genome comparisons [17–20]. Based on DFT, cumulative Fourier power spectrum (CFPS) method was proposed [21]. However, it has two main shortcomings. First, this method only used the power spectrum to calculate the moment vectors, but it lacked the information of phase spectrum. As a result, the mapping between moment vectors and genome sequences is not one-to-one. So it cannot reflect all the biological properties of the original genome sequences. And the relationship between nucleotides is important, particularly which may be associated with the domain structure of genomes, such as intron and exon. However, the definition of moment vector in CFPS method cannot measure the correlation between different nucleotides.

The method presented here is a new alignment-free method based on the cumulative Fourier power and phase spectra, which overcomes these limitations mentioned above. We define a 28-dimensional vector to characterize a DNA sequence, where the dissimilarity of DNA sequences is taken as the Euclidean distances between the vectors. As an improvement over previous work, the power and phase spectra and their covariances are included as part of the vector. So, we can measure the correlation between the four nucleotides in the power and phase spectra by the vectors. We

* Corresponding author.

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

discover that the distribution of covariance in exons and introns is not the same, so our method can be used to identify exons and intron. Since we add the phase spectrum to the moment vectors, we can prove that the mapping between DNA sequence and its complete central moment vector of the cumulative Fourier power and phase spectra is one-to-one. As a result, our method is more accurate than CFPS method.

This study is structured into three parts. First, we describe the cumulative Fourier transform algorithm and define the 28-dimensional vector using the power and phase spectra. Next, we define a similarity metric using the Euclidean distance and test whether the distances between vectors can measure the DNA sequence similarity by simulated DNA sequences. Finally, we apply our method to the identification of exons and introns in *S. cerevisiae* and *S. pombe*, and the comparison of genome sequences from different species including viral genomes, bacterial genomes. We show that our method is highly accurate and effective for the hierarchical clustering of a variety of DNA sequences and genomes compared with the CFPS, FFP(k-mer), Clustal Omega, MAFFT and MSAProbs.

2. Materials and Methods

2.1. Materials

The following three datasets were used to validate the method. The first dataset consists of the segment 6 neuraminidase (NA) genes of 38 Influenza A viruses. The second dataset includes 341 viruses from [22], which focuses on the classification of Human papilloma virus (HPV). The third one includes 56 bacterial genomes which can be clustered into 14 well-known families. All the accession numbers of sequences are provided in the Appendices. We also simulated some mutations in a DNA sequence and constructed phylogenetic trees of simulated DNA sequences to test our method.

2.2. Methods

Our alignment-free method consists of three major steps. First, we transform a DNA sequence into four binary sequences for A, C, G, and T. Then we perform Discrete Fourier Transform on four binary sequences and acquire the cumulative Fourier power and phase spectra. Next, we use the cumulative Fourier power and phase spectra to calculate the central moments and the covariance of four nucleotides. Finally, we can obtain a 28-dimensional vector for each genome sequence.

2.2.1. Indicator Function

For a DNA sequence $s_0s_2\dots s_{N-1}$, we define four indicator functions for A, C, G, and T:

$$u_\alpha(n) = \begin{cases} 1, & s_n = \alpha \\ 0, & s_n \neq \alpha \end{cases} \quad n = 0, \dots, N-1, \quad \alpha = A, C, G, T. \quad (1)$$

For instance, for the sequence ACCGATTAG, four indicator functions are as follows:

$$u_A : 100010010; u_C : 011000000; \\ u_G : 000100001; u_T : 000001100.$$

2.2.2. DFT and Cumulative Power and Phase Spectrum

Discrete Fourier transform (DFT) is a broadly used digital signal processing approach, which transforms data from time space to frequency space and reveals periodicities that are hidden in time space [20]. The frequency domain vector contains all the information about the signal in the time domain.

For a sequence of length N , the DFT of four indicator functions at frequency k is:

$$F_\alpha(k) = \sum_{n=0}^{N-1} u_\alpha(n) e^{-\frac{2\pi}{N}kn}, \quad k = 0, 1, 2, \dots, N-1 \quad (2)$$

The DFT power spectrum at frequency k is defined as:

$$PS_\alpha(k) = |F_\alpha(k)|^2 \quad (3)$$

Then the DFT phase spectrum at frequency k is defined as:

$$AS_\alpha(k) = \arg(F_\alpha(k)) \quad (4)$$

By definition, $AS_\alpha(k) \in [0, 2\pi]$. Note that $AS_\alpha(0) = 0$, $PS_\alpha(0) = |\sum_{n=0}^{N-1} u_\alpha(n)|^2$, so we delete $PS_\alpha(0)$, $AS_\alpha(0)$ to calculate the cumulative Fourier power and phase spectrum. Now, we get:

$$CPS_\alpha(k) = \sum_{n=1}^k PS_\alpha(n), \quad k = 1, 2, \dots, N-1 \quad (5)$$

$$CAS_\alpha(k) = \sum_{n=1}^k AS_\alpha(n), \quad k = 1, 2, \dots, N-1 \quad (6)$$

Because all the $AS_\alpha(k)$ and $PS_\alpha(k)$ are non-negative, $CPS_\alpha(k)$ and $CAS_\alpha(k)$ are non-decreasing.

2.2.3. Central Moment Vector

Because of different lengths of DNA sequences, cumulative Fourier power spectrum (CPS) and cumulative Fourier phase spectrum (CAS) series have clear differences in number. Consequently, the Euclidean distance between two DNA sequences with different lengths cannot be defined. This means we do not use CPS and CAS directly. To solve this issue, we use the central moment vector of CPS and CAS. Thus we transform the CPS and CAS series of different sequences into the points in the same dimensional space. We define the mean value of CPS and CAS as follows:

$$Mean_\alpha(CPS) = \frac{1}{N-1} \sum_{k=1}^{N-1} CPS_\alpha(k) \quad (7)$$

$$Mean_\alpha(CAS) = \frac{1}{N-1} \sum_{k=1}^{N-1} CAS_\alpha(k) \quad (8)$$

To measure the distribution of Fourier power and phase spectra of different genomes, we follow the statistical method to introduce the central moments. To make sure that the first moment vector would not be zero, we chose the absolute value:

$$CM_j^\alpha(CPS) = \frac{\sum_{k=1}^{N-1} |CPS_\alpha(k) - Mean_\alpha(CPS)|^j}{(N_\alpha(N - N_\alpha))^{j-1} N^j} \quad (9)$$

$$CM_j^\alpha(CAS) = \frac{\sum_{k=1}^{N-1} |CAS_\alpha(k) - Mean_\alpha(CAS)|^j}{(N_\alpha(N - N_\alpha))^{j-1} N^j} \quad (10)$$

here $j = 1, 2, \dots, N - 1$, $\alpha = A, C, G, T$ and N_α is the number of the nucleotide α in the sequence and the scale factor $1/(N_\alpha(N - N_\alpha))^{j-1} N^j$ is chosen as CFPS method [21].

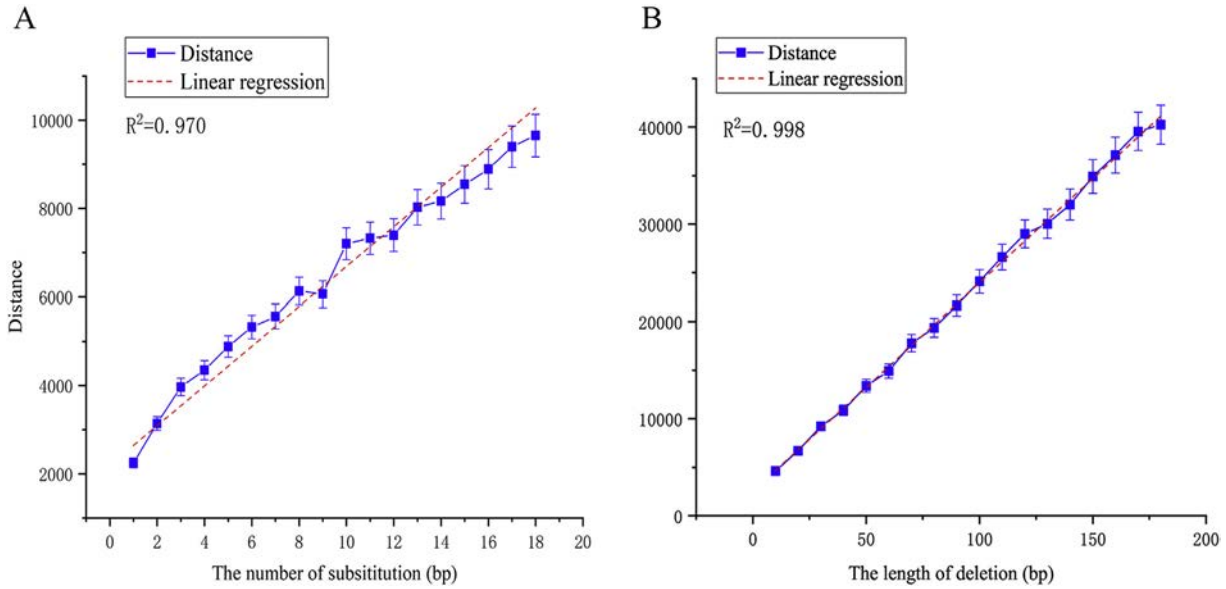


Fig. 1. Correlation between Euclidean distance and biological distance. The error bar is standard deviation for 10 random experiments. (A) Correlation between distances of 28-dim vector and the number of substitutions of DNA sequence. (B) Correlation between distances of 28-dim vector and the length of deletions of DNA sequence.

Now we get the complete central moment vector:

$$\begin{aligned}
 & (Mean_A(CPS), Mean_C(CPS), Mean_G(CPS), Mean_T(CPS), \\
 & Mean_A(CAS), Mean_C(CAS), Mean_G(CAS), Mean_T(CAS), \\
 & CM_1^A(CPS), CM_2^A(CPS), \dots, CM_{N-1}^A(CPS), \dots, \\
 & CM_1^T(CPS), CM_2^T(CPS), \dots, CM_{N-1}^T(CPS), \\
 & CM_1^A(CAS), CM_2^A(CAS), \dots, CM_{N-1}^A(CAS), \dots, \\
 & CM_1^T(CAS), CM_2^T(CAS), \dots, CM_{N-1}^T(CAS)).
 \end{aligned} \quad (11)$$

By this definition, we can prove that Fourier power and phase spectra can be recovered by the complete central moment vector. This means the vector and the spectra of Fourier transform is one-to-one [23]. Then, we can use Inverse Discrete Fourier Transform (IDFT) to recover the original DNA sequence. Thus, we keep all the information in the original DNA sequence during the transformation from the DNA sequence to numerical sequence. We provide proof of this in the appendices A.

When calculating the central moment vector, we find that the central moment converges to zero as j increases. Compared with $Mean(CPS)$, $Mean(CAS)$, $CM_1(CPS)$ and $CM_2(CPS)$, other central moments are very small, which has no effect on the classification and phylogenetic

Table 1
DNA sequence mutations description in simulation tests.

A	Generated from gene 574,406 (Gene ID)
A_substitution_2_1	2 random nucleotide substitutions in A
A_substitution_2_2	2 random nucleotide substitutions in A
A_substitution_5_1	5 random nucleotide substitutions in A
A_substitution_5_2	5 random nucleotide substitutions in A
A_substitution_10_1	10 random nucleotide substitutions in A
A_substitution_10_2	10 random nucleotide substitutions in A
B	Generated from gene 574,406 (Gene ID)
B_substitution_2_1	2 random nucleotide substitutions in B
B_substitution_2_2	2 random nucleotide substitutions in B
B_insertion_5_1	5 bp insertion at position 51 in B
B_insertion_5_2	5 bp insertion at position 101 in B
B_deletion_5_1	5 bp deletion from position 51:55 in B
B_deletion_5_2	5 bp deletion from position 101:105 in B
B_transposition_10_1	10 bp transposition from position 1001 to 3001 in B
B_transposition_10_2	10 bp transposition from position 1251 to 2001 in B

results. Owing to this observation, we only consider $Mean$ for the power and phase spectra and the first two central moments for the power spectrum to get a truncated central moment vector. The truncated central moment vector can greatly save storage space and computational time. Then we give a 16-dimensional point in the Euclidean space of every sequence:

$$\begin{aligned}
 & (Mean_A(CPS), Mean_C(CPS), Mean_G(CPS), Mean_T(CPS), \\
 & Mean_A(CAS), Mean_C(CAS), Mean_G(CAS), Mean_T(CAS), \\
 & CM_1^A(CPS), CM_1^C(CPS), CM_1^G(CPS), CM_1^T(CPS), \\
 & CM_2^A(CPS), CM_2^C(CPS), CM_2^G(CPS), CM_2^T(CPS))
 \end{aligned} \quad (12)$$

This 16-dimensional vector contains almost all the information of central moments for the power and phase spectra.

2.2.4. Covariance

However, the 16-dimensional point can only reveal the distribution of A, C, G, T respectively. To measure the relationship of four nucleotides, we add covariances to the point. We define the covariance as follows:

$$COV_{CPS}(\alpha, \beta) = \frac{N-1}{N^2 N_{\alpha+\beta} (N - N_{\alpha+\beta})} cov(CPS_{\alpha}, CPS_{\beta}) \quad (13)$$

$$COV_{CAS}(\alpha, \beta) = \frac{N-1}{N^2 N_{\alpha+\beta} (N - N_{\alpha+\beta})} cov(CAS_{\alpha}, CAS_{\beta}) \quad (14)$$

$$COV(\alpha, \beta) = (COV_{CPS}(\alpha, \beta), COV_{CAS}(\alpha, \beta)) \quad (15)$$

where

$$\begin{aligned}
 & cov(CPS_{\alpha}, CPS_{\beta}) \\
 & = \frac{1}{N-1} \sum_{k=1}^{N-1} |CPS_{\alpha}(k) - Mean_{\alpha}(CPS)| |CPS_{\beta}(k) - Mean_{\beta}(CPS)|
 \end{aligned} \quad (16)$$

$$\begin{aligned}
 & cov(CAS_{\alpha}, CAS_{\beta}) \\
 & = \frac{1}{N-1} \sum_{k=1}^{N-1} |CAS_{\alpha}(k) - Mean_{\alpha}(CAS)| |CAS_{\beta}(k) - Mean_{\beta}(CAS)|
 \end{aligned} \quad (17)$$

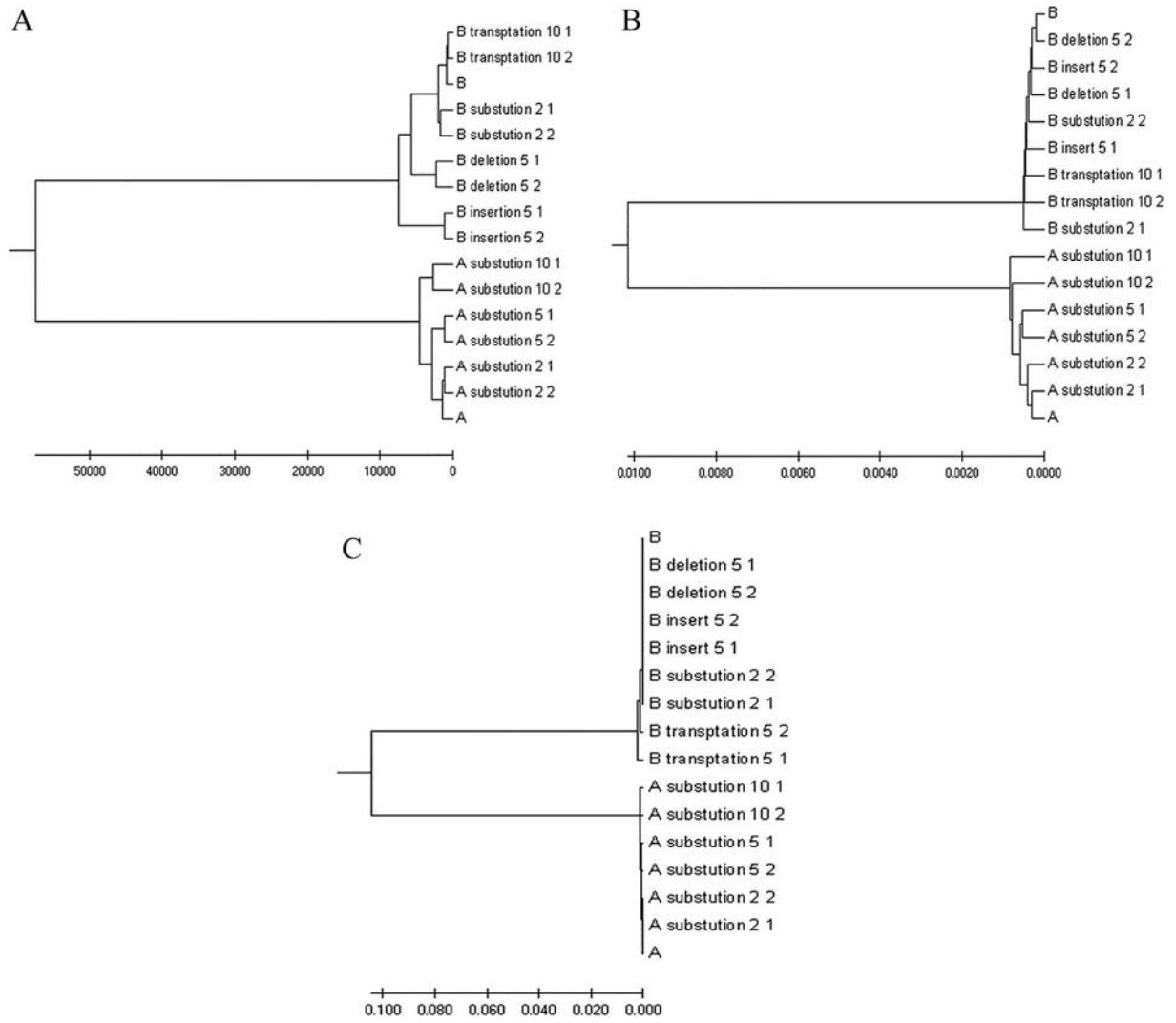


Fig. 2. Clustering analysis of different mutations by phylogenetic trees of simulated DNA sequences in Table 1. (A) our method, (B) the FFP (k-mer) method, (C) Clustal Omega.

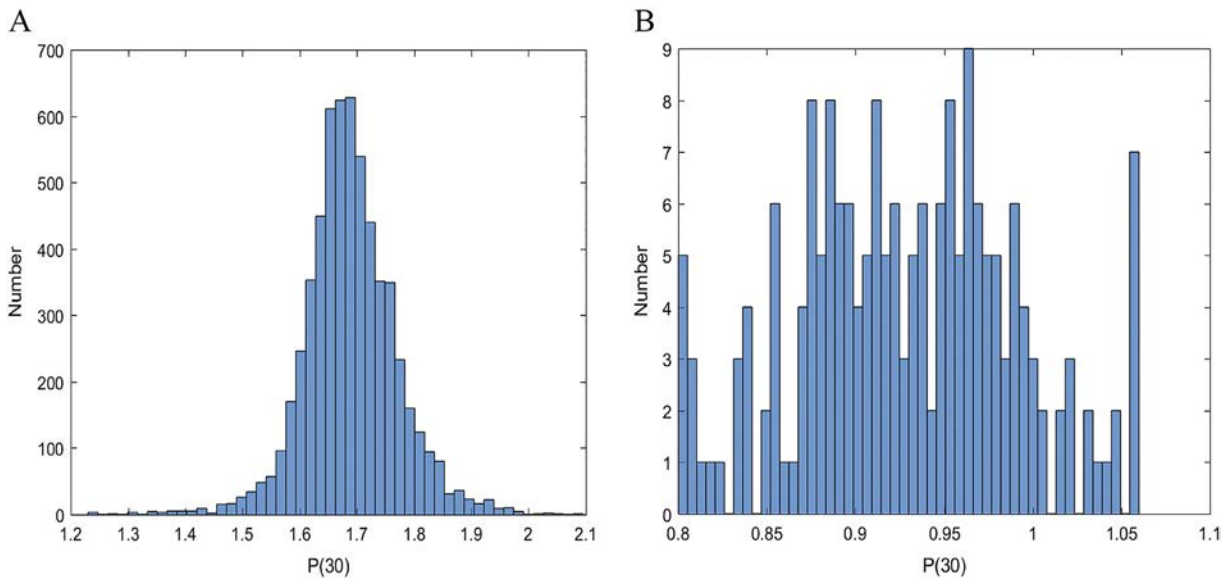


Fig. 3. (A) distribution for exons in all chromosomes in *S. cerevisiae*. (B) distribution for introns in all chromosomes in *S. cerevisiae*.

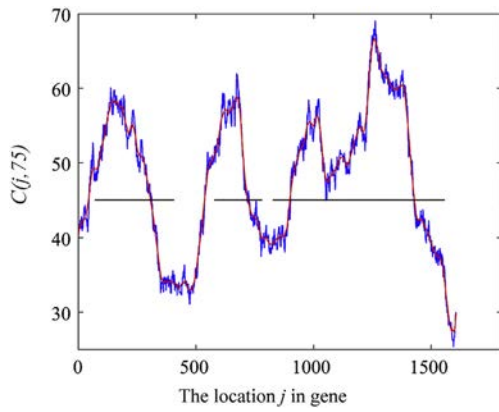


Fig. 4. Graph of $C(j,75)$ SPBC1685.08 in chromosome 2 of *S. pombe*, using a sliding window of 75 bp. The horizontal segments represent the actual location of the three exons. The red curve is the smoothing curve of $C(j,75)$.

and

$$N_{\alpha+\beta} = \frac{N_{\alpha} + N_{\beta}}{2} \tag{18}$$

From this definition, we can see $COV(\alpha, \alpha) = (CM_2^{\alpha}(CPS), CM_2^{\alpha}(CAS))$. Then we add covariance to the 16-dimensional vector (12). Consequently, a 28-dimensional point of the DNA sequence is constructed,

which is

$$\begin{aligned} &(Mean_A(CPS), Mean_C(CPS), Mean_G(CPS), Mean_T(CPS), \\ &Mean_A(CAS), Mean_C(CAS), Mean_G(CAS), Mean_T(CAS), \\ &CM_1^A(CPS), CM_1^C(CPS), CM_1^G(CPS), CM_1^T(CPS), \\ &CM_2^A(CPS), CM_2^C(CPS), CM_2^G(CPS), CM_2^T(CPS), \\ &COV(A, C), COV(A, G), COV(A, T), \dots, COV(G, T)) \end{aligned} \tag{19}$$

3. Results

3.1. The Accuracy of Similarity Distance Metric

In our method, we use the Euclidean distance to measure the biological relationship between DNA sequences. A series of artificial deletion and substitution mutations of a DNA sequence (Gene ID:574406) are used to assess the accuracy of the similarity distances. Then we calculate the Euclidean distances between 28-dimensional vectors of the mutants and the original sequence. Thus, we can obtain the correlation between similarity distances and the mutation numbers. The results in Fig. 1 show a sound linear relationship between the distances and the number of substitutions and length of deletions. These results demonstrate the accuracy of the 28-dimensional vector distance metric for different types of nucleotide mutations in the same DNA sequence, meaning the distance of the 28-dimensional vector is linearly correlated to the edit distance for DNA sequences.

3.2. Construction of Phylogenetic Trees on Different Simulated DNA Mutations

To determine whether the Euclidean distances between 28-dimensional vectors can be used to cluster DNA sequences, we generated different mutations in a DNA sequence (Gene ID: 574406, 5188 bp) and constructed phylogenetic trees by our method, the FFP (k-mer) method and the MSA method. We generated two new sequences A and B from the original sequence by making 10% of substitutions randomly. Then, we similarly transformed A and B into different mutants by four different types of mutations (substitutions, deletions, insertions, and transpositions). Table 1 is the description of the simulated DNA sequences with different mutations. Unweighted Pair Group Method with Arithmetic Mean (UPGMA) phylogenetic trees of the mutations were constructed from the distance metric using the

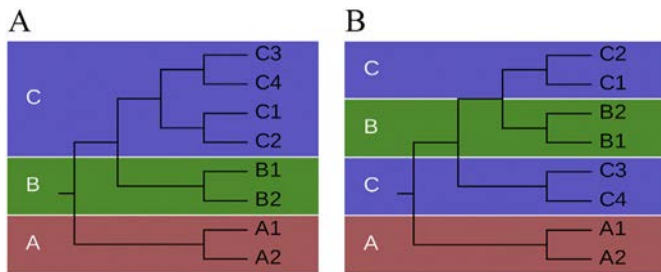


Fig. 5. UPGMA tree of 8 HIV sequences by 12-dimensional vector and 16-dimensional vector. (A) 16-dimensional vector (Mean (CPS), Mean (CAS), $CM_1(CPS)$, $CM_2(CPS)$), (B) 12-dimensional vector (Mean (CPS), $CM_1(CPS)$, $CM_2(CPS)$).

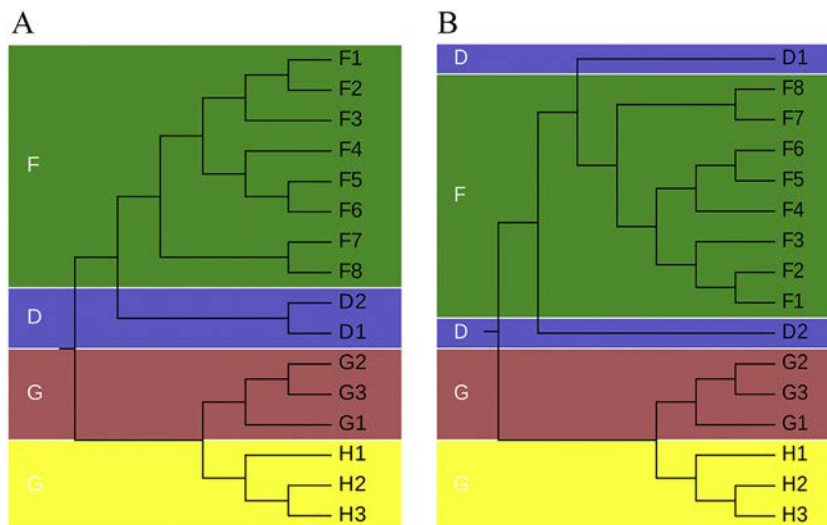


Fig. 6. UPGMA tree of 16 HIV sequences by 16-dimensional vector and 28-dimensional vector. (A) 28-dimensional vector (Mean (CPS), Mean (CAS), $CM_1(CPS)$, $CM_2(CPS)$, $COV(CPS)$), (B) 16dimensional vector (Mean (CPS), Mean (CAS), $CM_1(CPS)$, $CM_2(CPS)$).

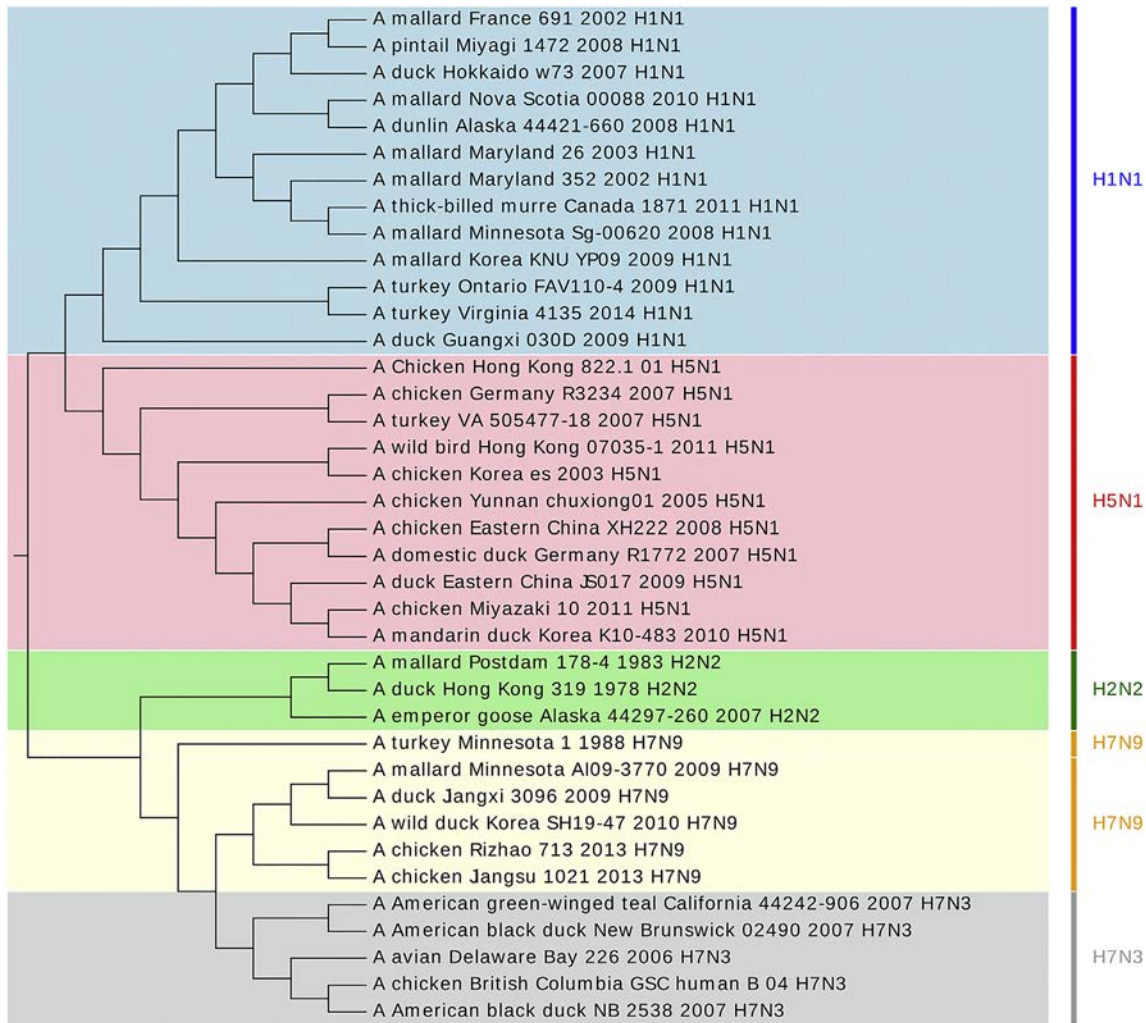


Fig. 7. UPGMA tree of 38 influenza A viruses by our method. It is divided into 5 clusters: H1N1(red), H2N2(green), H7N9(black), H7N3(blue), H5N1(purple) correctly.

proposed method, the alignment-free FFP (k-mer) method and Clustal Omega, as shown in Fig. 2(A–C), respectively.

For the different substitution mutations of sequence A, Fig. 2 shows that all the three methods can cluster different numbers of substitutions. This result indicates that the similarity measure of 28-dimensional vector, the FFP (k-mer) method and Clustal Omega have the same capacity in identifying and measuring the distances between substitutions. Fig. 2 shows the topological differences among the 28-dimensional vector measure, the FFP (k-mer) method and Clustal Omega for deletion, insertion and transposition mutations of sequence B. Deletion and insertion are two types of mutation with phenotypic effects that often more pronounced than those of substitutions. Fig. 2 (A) and (C) show that the substitutions can be separated from deletion or insertion mutations by the 28-dimensional vector method and Clustal Omega; however, the FFP (k-mer) method cannot identify these substitutions from deletion/insertion mutations and mix them in the same branches (Fig. 2(B)). For transposition mutations, Fig. 2 (A) and (C) show that the 10-bp transpositions can be clearly separated from both substitutions and insertion/deletion mutations by 28-dimensional vector and Clustal Omega, but the FFP (k-mer) method cannot separate transposition mutants from substitutions, classifying them in the same branches as shown in Fig. 2(B). Our method contains the nucleotide distribution at all the positions in DNA sequences and the relationship of four nucleotides, so it can identify different types of mutations. The FFP (k-mer) method is mainly based on the frequencies of the k-mers in the sequence but does not contain the information

of position and relationship of four nucleotides, so similarity measure from the FFP (k-mer) method is less reliable for sequence rearrangements. This result shows that our similarity measure may have special capacity to distinguish different mutations, while the FFP (k-mer) method may not recognize these differences.

3.3. Distribution of Covariance Between Exons and Introns

We first use exon and intron sequences to test the performance of our method in genome comparison. The covariances of Fourier transform power and phase spectra can measure the relationship of the four nucleotides. As a result, we calculated the covariances between all experimental exons and introns in *S. cerevisiae*.

Using the covariances of Fourier power and phase spectra in Eq. (15), for a sequence with length N , we calculate all $C(j, L)$ for its subpieces with window size L at position j from 1 to $N - 1$. $P(L)$ is the mean of the value of $C(j, L)$:

$$COV_{CPS} = (COV_{CPS}(A, C), \dots, COV_{CPS}(G, T)) \quad (20)$$

$$COV_{CAS} = (COV_{CAS}(A, C), \dots, COV_{CAS}(G, T)) \quad (21)$$

$$C(j, L) = \|(COV_{CPS}, COV_{CAS})\|_2 \quad (22)$$

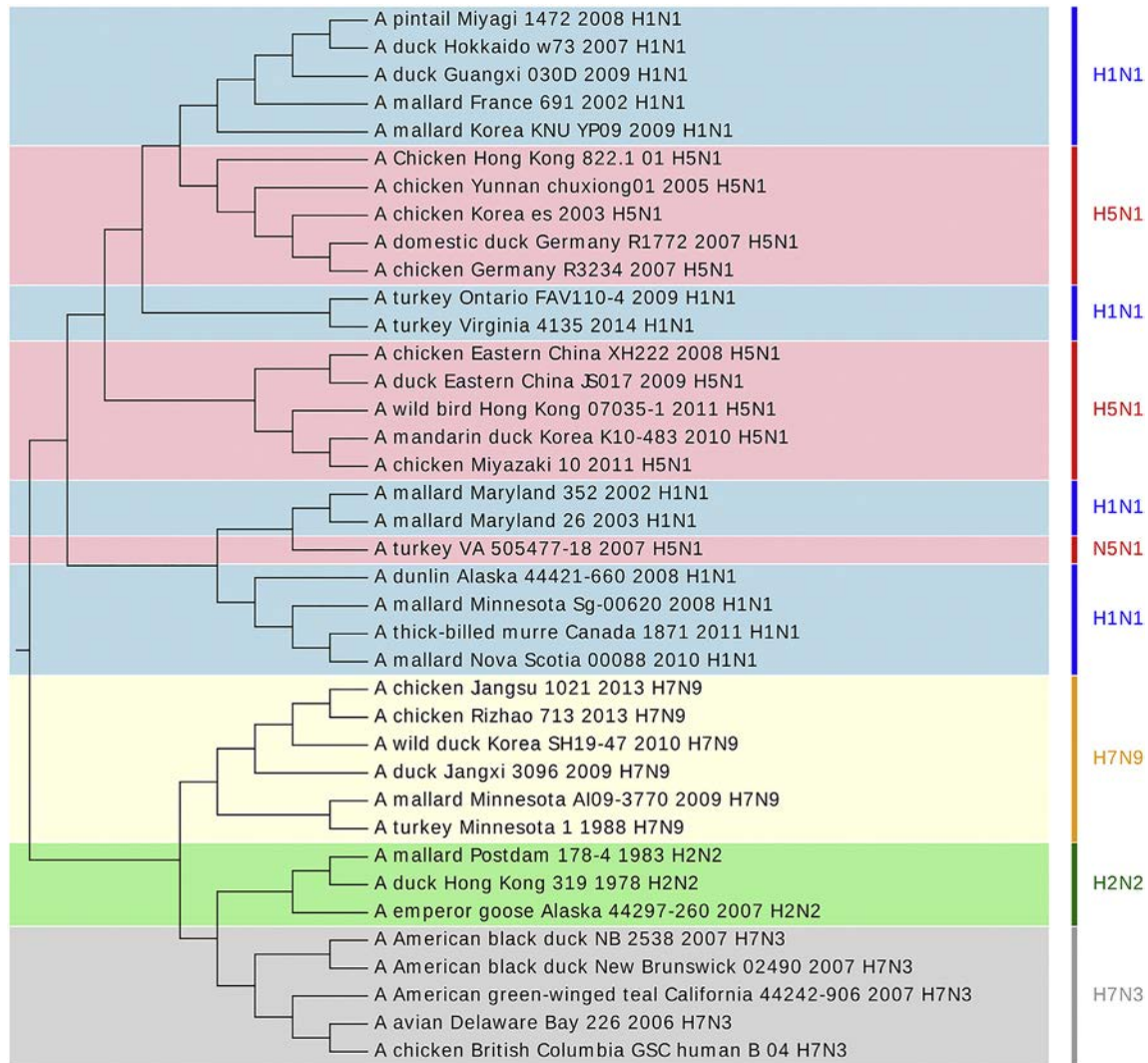


Fig. 8. UPGMA tree of 38 influenza A viruses by the FFP method ($k = 5$). It misplaces the virus 'A turkey VA 505477-18 2007 H5N1' into the H1N1 group, and a part of H1N1 viruses are incorrectly grouped with H5N1 subtype.

$$P(L) = \frac{\sum_{j=1}^{N-L} C(j, L)}{N-L} \quad (23)$$

We chose $L = 30$ which is about the shortest length of all the exons and introns of *S. cerevisiae* and calculate $L(30)$ for all exons and introns. Fig. 3(A) shows the histogram describing these distributions for all experimental exons in the 16 chromosomes of *S. cerevisiae*. As this figure reveals, the covariance of exons is distributed around a central value as a normal distribution. Fig. 3(B) shows the corresponding histogram for intron regions in the 16 chromosomes. The distribution for intron regions is close to uniform, which is different from the distribution that was obtained for exon regions.

To study how the difference between exons and introns in terms of argument distribution can be applied to gene prediction, we observed the changes of covariance at different positions on a typical split gene of *S. pombe* (gene SPBC1685.08 in chromosome 2). Table A.1 information of exons on this gene is shown. We take $L = 75$, which is about the shortest length of all the exons of *S. pombe* and calculate $C(j, 75)$. Fig. 4 shows the curve of $C(j, 75)$. The horizontal lines represent the actual locations of the three exons. Note that $C(j, 75)$ for exon positions has a higher value than for introns.

3.4. Construction of Phylogenetic Trees With/Without Phase Spectrum and Covariance

In our previous paper, we proposed CFPS method using only the central moments of power spectrum [21]. The novelty of the current method is that we add the phase spectrum as well as the covariance of the A, C, G and T power spectrum in our analysis. This two additional information give us significant improvement in correctly annotating genome sequences than our previous method. To illustrate this, we evaluate the performance of our new method with phase spectrum and covariance in two steps. First, we use 8 human immunodeficiency virus (HIV) whole genome sequences of A, B, C sub-types to construct the phylogenetic tree by central moments of power spectrum with/without central moments of phase spectrum. In order to balance the magnitude of central moments and covariance, we first normalize all the components of 28-dimensional vector. Then we use Mega7 to construct the UPGMA phylogenetic tree of HIV sequences [24]. From Fig. 5 (A), we can see that the three sub-types of HIV genomes are completely classified in the phylogenetic tree in the case of our method with 16-dimensional vectors: (including central moments of power and phase spectra). However, HIV genomes of C sub-types aren't gathered in one branch by CFPS method which only contains central moments of power spectrum (Fig. 5(B)). Second, we use 16 HIV whole genome

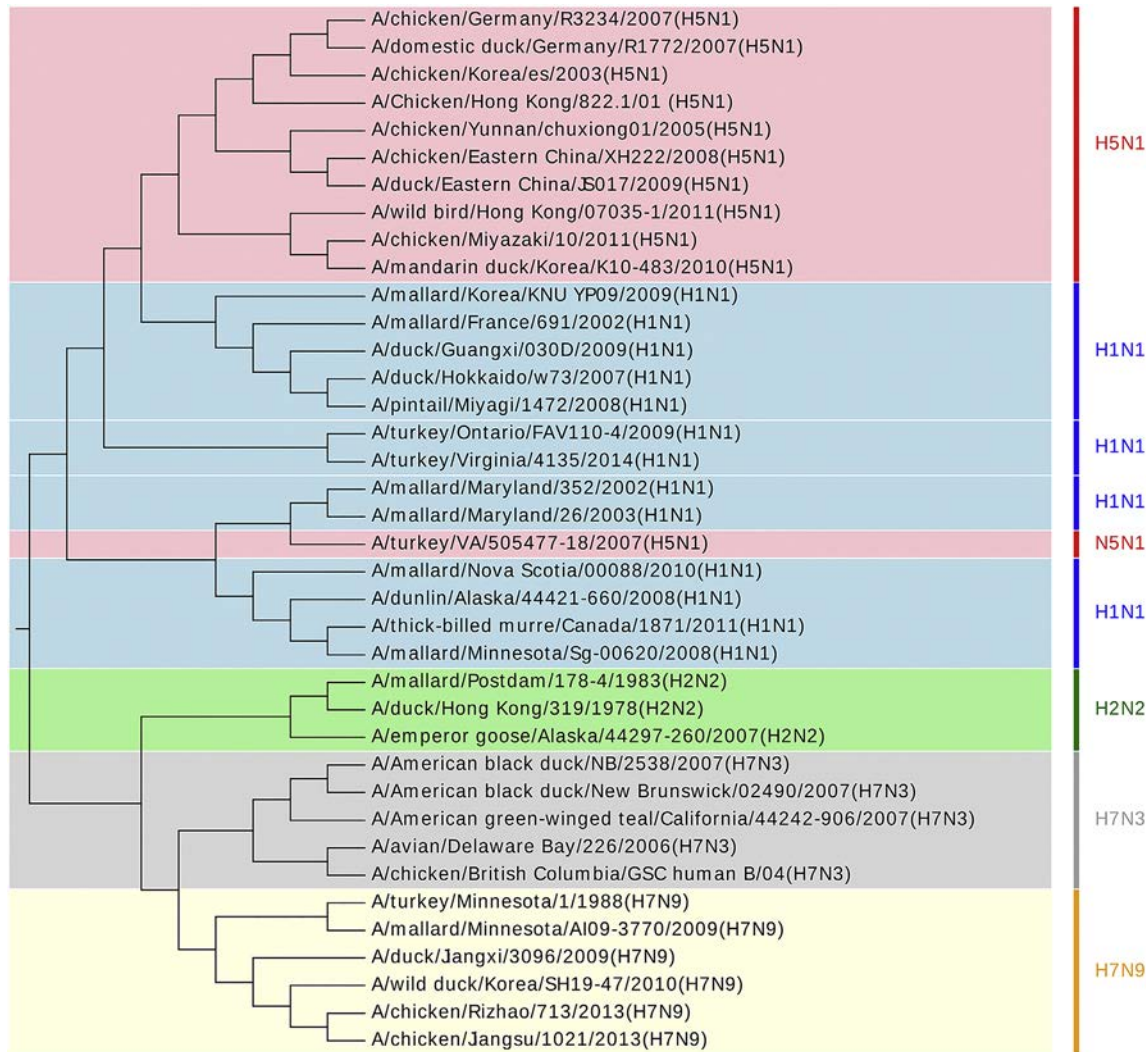


Fig. 9. UPGMA tree of 38 influenza A viruses by Clustal Omega. It misplaces the virus ‘A turkey VA 505477-18 2007 H5N1’ into the H1N1 group.

sequences of D, F, G, H sub-types to construct the phylogenetic tree by central moments of power and phase spectra with/without their covariances. From Fig. 6, our 28-dimensional vectors can distinguish the four sub-types, while 16-dimensional vectors without covariance divide D1 and F sub-type into one branch. Thus, we can see that our method contains more information and the performance is significantly improved compared to the previous CFPS method.

3.5. Construction of Phylogenetic Trees

Here, we compare our method to FFP (k-mer) method, Clustal Omega, MAFFT and MSAProbs, for computational efficiency and accuracy. MATLAB R2017b and MEGA 7 are used to draw the phylogenies of genomes [24]. To verify our method, we apply it to various data sets including viral genome sequence and bacterial genome sequences (DNA GenBank information is shown in Table A.2–A.4) to construct a UPGMA tree.

3.5.1. Influenza A Viruses

Influenza A viruses are a constant threat to both human and animal health because of their high mutation rate. They are negative-sense, single-stranded, segmented RNA viruses that are classified by their surface glycoproteins: hemagglutinin (HA) and neuraminidase (NA) [25]. The dataset used in this work consists of 38 of the most lethal subtypes

of Influenza A viruses, such as H1N1, H2N2, H5N1, H7N9, and H7N3. As Fig. 7, the dataset is divided correctly into five groups, which are consistent with the biological taxonomy, except ‘A turkey Minnesota 11,988 H5N1’. We find that the ‘A turkey Minnesota 1988 H7N9’ is earlier than the rest H7N9 and H7N3 viruses, so it locates at the root of H7N9, and H7N3. Regarding FFP (k-mer) method (Fig. 8), we choose $k = 5$. It misplaces the virus ‘A turkey VA 505477-18 2007 H5N1’ into the H1N1 group, and a part of H1N1 viruses are incorrectly grouped with H5N1 subtype. The Clustal Omega may only classify ‘A 258 turkey VA 505477-18 2007 H5N1’ with some uncertainty (Fig. 9). To investigate the reason of this exception, using sequence alignment by MEGA, we found that there is an 8 bases insertion mutation at position 13 in ‘A turkey VA 505477-18 2007 H5N1’ compared with other H5N1 sequences. Thus, the tree by our method can display clear levels of hierarchy and relationship among different viruses, but MSA and FFP (k-mer) methods cannot have clear spatial separation of similar species in the tree. Therefore, the results obtained using our method are better than MSA and FFP (k-mer) methods.

3.5.2. Human Papilloma Virus (HPV)

Human papillomavirus (HPV) causes cervical cancer, which is the fourth most common cancer in women, with an estimated 266,000 deaths and 528,000 new cases in 2012. Virtually all cervical cancer cases (99%) are linked to genital infection with HPV and it is the most

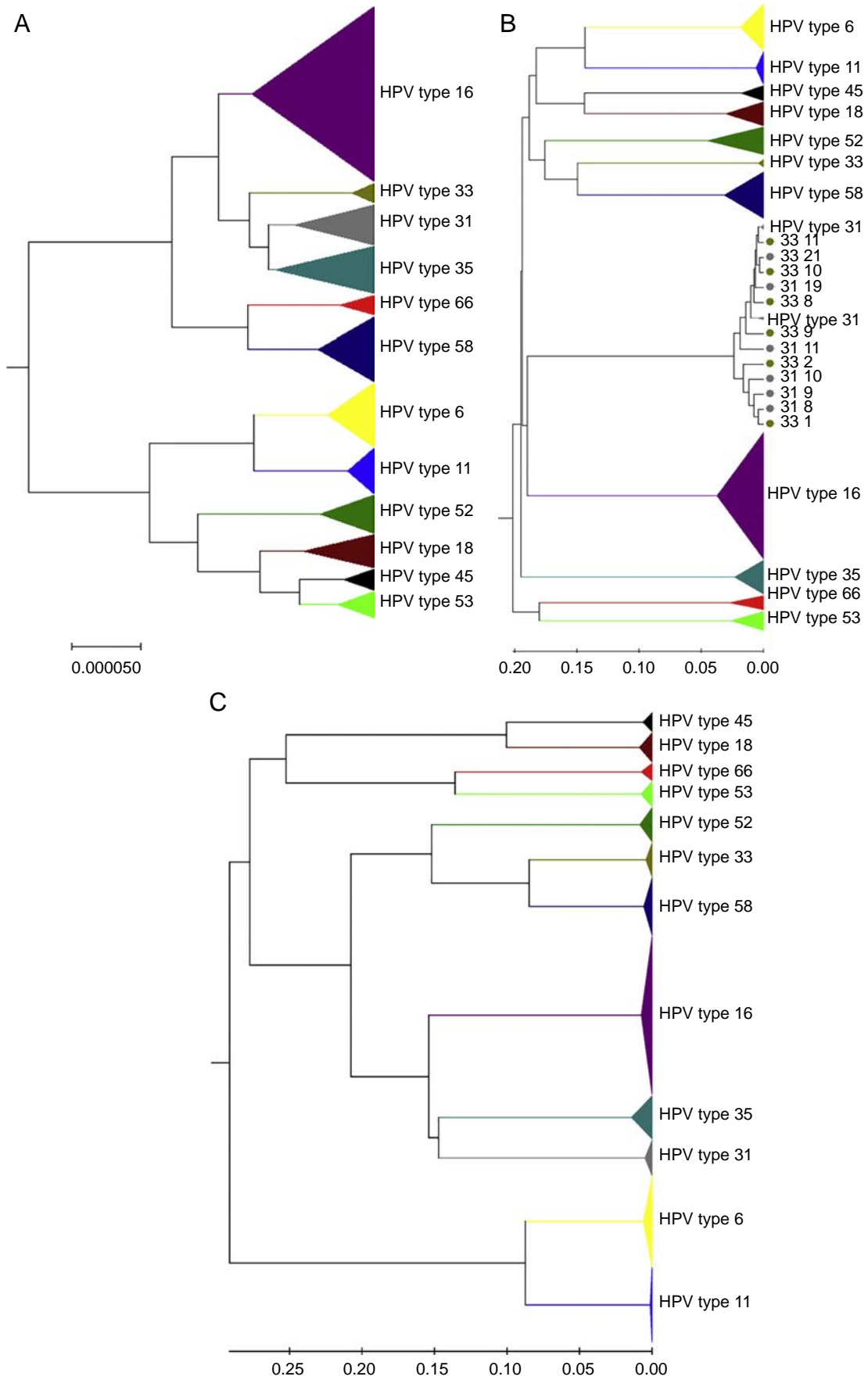


Fig. 10. UPGMA tree of 341 HPV sequences by three methods. (A) our method, (B) FFP method (k = 6), (C) Clustal Omega.

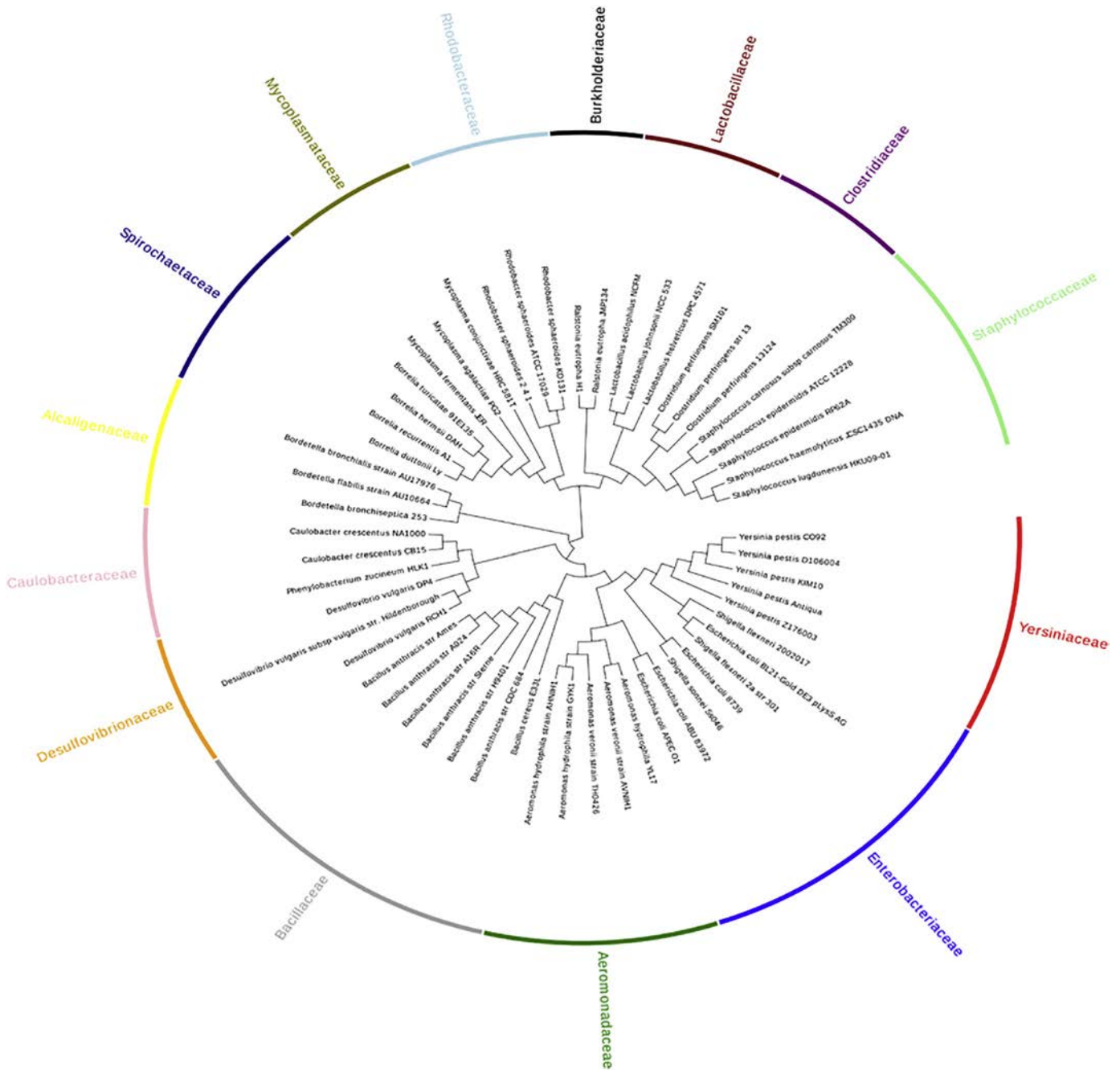


Fig. 11. UPGMA tree of 56 bacteria by our method.

common viral infection of the reproductive tract. HPV is a group of more than 150 related viruses [22]. Each HPV virus in this large group is given a number which is called its HPV type. Low risk HPV types such as 6 and 11 can cause genital warts or benign. High risk HPV types such as 16 and 18 account for about 70% of cervical cancer [9]. As a result, how to quickly and accurately predict HPV risk types has become a hot span, such as chaos game representation [22], support vector machines [26], decision tree [27], and ensemble support vector machines with protein secondary structures [28].

In this work, our method classifies the dataset of 341 HPV genomes into 12 genotypes 6, 11, 16, 18, 31, 33, 35, 45, 52, 53, 58, and 66 correctly in less than 20 s (Fig. 10(A)). However, FFP (k-mer) method misplaces a part of 33 types into 31 types, obviously worse than results by our method (Fig. 10(B)). Although Clustal Omega can classify the dataset correctly, the time is longer than our methods. (Fig. 10(C)).

3.5.3. Bacteria

Bacteria are important in many stages of the nutrient cycle by recycling nutrients such as the fixation of nitrogen from the atmosphere. While bacterial fossils exist, such as stromatolites, their lack of distinctive morphology prevents them from being used to examine the history of bacterial evolution, or to date the time of origin of a bacterial species [29]. Hence, it is vital to reconstruct the bacterial phylogeny by DNA sequences. Nevertheless, the length of bacterial whole genome sequence is over 1 million bp, thus MSA methods will consume a large amount of time. The dataset of 56 bacteria is used to test our method, including 14 families: *Aeromonadaceae*, *Alcaligenaceae*, *Bacillaceae*, *Burkholderiaceae*, *Caulobacteraceae*, *Clostridiaceae*, *Desulfovibrionaceae*, *Enterobacteriaceae*, *Lactobacillaceae*, *Mycoplasmataceae*, *Rhodobacteriaceae*, *Spirochaetaceae*, *Staphylococcaceae* and *Yersiniaceae*. The length of genome sequence ranges from 3 to 5 million bp. As illustrated by the phylogenetic

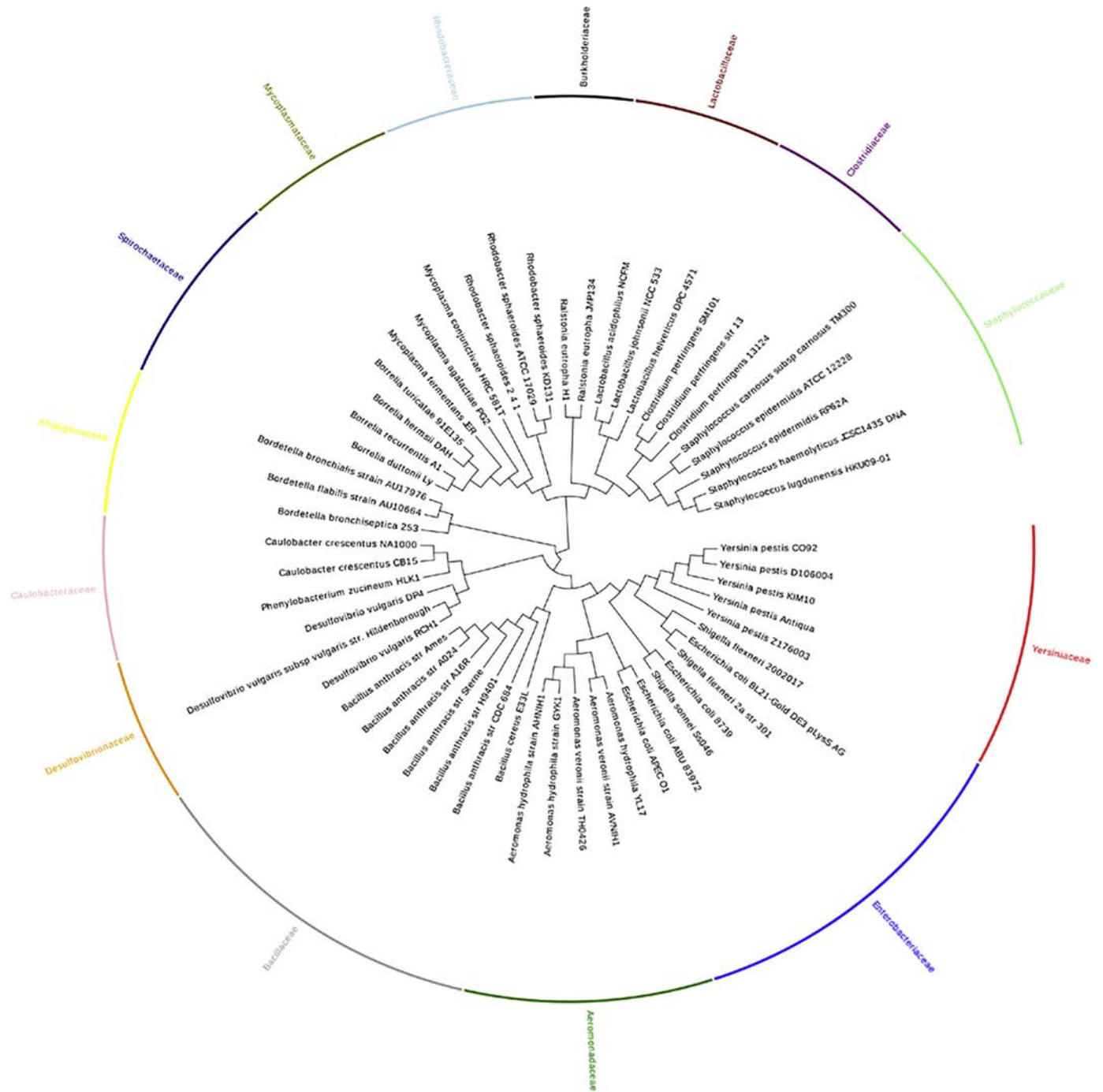


Fig. 12. UPGMA tree of 56 bacteria by FFP method ($k = 6$).

tree (Fig. 11), it is well separated into 14 biological families in less than 700 s. Though FFP method ($k = 6$) also classifies correctly, its running time is about 30,000 s, which is significantly longer than our method (Fig. 12).

3.6. Time Statistics and Algorithm Complexity

We performed all the calculations on the same machine and clear the memory each time to avoid redundancy and influence on the

Table 2
Time comparison of the six methods.

Datasets	Average length (bp)	Number of species	Our method (seconds)	CFPS (seconds)	FFP (seconds)	Clustal Omega (seconds)	MAFFT (seconds)	MSAProbs (seconds)
Influenza A	1497	38	0.38	0.29	8.68	13.45	0.87	32.24
HPV	7915	341	15.28	13.06	99.30	4170.70	22.52	–
Bacteria	3,610,982	56	659.77	641.83	32,394.79	–	–	–

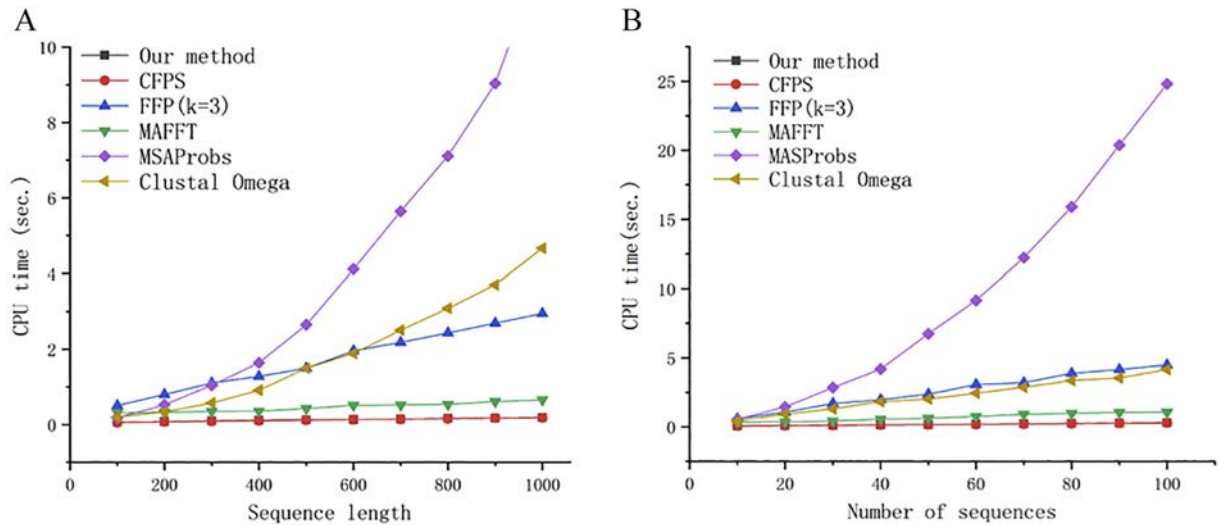


Fig. 13. (A) The plot of CPU time versus the average length of input sequences. The number of sequences is 30. (B) The plot of CPU time versus the number of input sequences. The average length of sequences is 500.

next-step calculation. The computation environment is CentOS 7 Linux Server running on Dell PowerEdge R730 with Dual Intel Xeon E5–2670 v3 12C/24 T CPU @2.30GHz and 384 GB RAM. We recorded the time and memory of our method, CFPS, FFP, Clustal Omega, MAFFT and MSAProbs methods. When measuring running times, each execution was repeated three times and averaged.

From Table 2, we can conclude that alignment-free methods are much more time-efficient than MSA methods. Because our method need calculate the covariance of power and phase spectra, it is a bit slower than CFPS method. In order to evaluate the complexity of the present methods, we analyze the relationship between CPU time and the length and number of sequences. Fig. 13(A) shows the dependence of CPU time on the sequence length (L) and the number of sequences is 30. And Fig. 13(B) shows the dependence of CPU time on the number (N) of sequences, and the average length of sequences is 500. We can conclude the time consumption of MSAProbs is $O(N^2L^3)$ [5]. Clustal Omega consumes $O(N\log NL^2)$ CPU time [3]. Other methods require approximately $O(NL)$ CPU times [4,16].

The memory requirements are shown in Table 3 and the unit is the memory requirement as a percentage of the total memory of the Linux server. We can see that alignment-free methods require more memory than alignment methods on the dataset of *Influenza A*, while alignment-free methods require less memory than alignment methods on larger datasets. The memory of FFP (k -mer) method depends on the value of k and becomes unacceptable when k becomes too large, especially when the whole genomes are input for analysis [16]. For a computer with 16G of memory, theoretically it can only calculate up to $k = 16$. Therefore, our method is more efficient than other methods, especially when the dataset is large.

Table 3
Memory comparison of the six methods.

Datasets	Average length (bp)	Number of species	Our method (%)	CFPS (%)	FFP (%)	Clustal Omega (%)	MAFFT (%)	MSAProbs (%)
Influenza A	1497	38	0.2	0.2	0.2	0.1	0.1	0.2
HPV	7915	341	0.2	0.2	0.2	0.8	0.1	–
Bacteria	3,610,982	56	0.6	0.6	0.8	–	1.4	–

4. Conclusions

In this work, we establish a novel method for genome comparison based on the cumulative Fourier power and phase spectra. In this method, we use power and phase spectra to create a 28-dimensional vector to represent a DNA sequence and define the Euclidean distances between the vectors as the similarity metric.

Our method has three main advantages. First, it contains all the information of the Fourier transform. Second, the mapping between DNA sequence and its complete central moment vector of the cumulative Fourier power and phase spectra is one-to-one. Although we only use truncated central moment vector in this study, the mapping between DNA sequence and its truncated central moment vector of the cumulative Fourier power and phase spectra is also one-to-one in our tested dataset. What's more, the covariance between spectra can measure the relationship of four nucleotides, with the distribution of this covariance differing between exons and introns. The results showed that our method is highly accurate and computationally effective at identifying different mutations (substitutions, insertions/deletions, and transpositions), exon-intron and for large-scale genome comparisons.

In addition, we found that the covariances of the power and phase spectra of the cumulative Fourier transform in exons is approximately normal, whereas in introns, the distribution is close to uniform in *S. cerevisiae*. Next, we used a sliding window to calculate the covariance at different positions for genes in *S. pombe*. We observed that there is generally a peak in exons. Therefore, this study also provides a new concept for predicting coding regions for future research.

The comparison of multiple-segmented genomes is also the focus of our future work. For multiple-segmented genomes, each segment is corresponds to a 28-dimensional vector by our method. So, the vectors of different segments can form a set. Then, we may use Hausdorff distance to calculate the distance between two sets as the similarity metric [30,31]. In further improvements, we will test the performance of our method on multiple-segmented genomes datasets.

Funding

This work was supported by Tsinghua University start-up fund and National Natural Science Foundation of China grant (#91746119) and Tsinghua University Education Foundation fund (042202008).

Declaration of Competing Interest

None declared.

Acknowledgements

The corresponding author would like to thank National Center for Theoretical Sciences (NCTS) for providing excellent research environment while part of this research was done. Stephen S.-T. Yau and Rong Lucy He designed the research; Shaojun Pei and Rui Dong performed the research; Stephen S.-T. Yau guided the paper and is the corresponding author; and Shaojun Pei wrote the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2019.07.003>. The Matlab program of ENV is available at GitHub (<https://github.com/YaulabTsinghua/The-central-moments-and-covariance-vector-of-cumulative-Fourier-Transform-power-and-phase-spectra>).

References

- [1] Initiative A. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 2005;408(6814):796–815.
- [2] Riechmann J, Heard J, Martin G, Reuber L, Jiang ZC, et al. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 2000;290(5499):2105–10. <https://doi.org/10.1126/science.290.5499.2105>.
- [3] Sievers F, Higgins D. Clustal omega for making accurate alignments of many protein sequences. *Protein Sci* 2018;27:135–45. <https://doi.org/10.1002/pro.3290>.
- [4] Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 2005;33(2):511–8. <https://doi.org/10.1093/nar/gki198>.
- [5] Liu Y, Bertil S, Douglas L. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* 2005;26(16):1958–64.
- [6] Yau S, Wang J, Niknejad A, Lu C, Jin N, Ho Y. DNA sequence representation without degeneracy. *Nucleic Acids Res* 2003;31(12):3078. <https://doi.org/10.1093/nar/gkg432>.
- [7] Vinga S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics* 2003;19(4):513. <https://doi.org/10.1093/bioinformatics/btg005>.
- [8] Hoang T, Yin C, Zheng H, Yu C, He R, Yau S. A new method to cluster DNA sequences using Fourier power spectrum. *J Theor Biol* 2015;372:135–45. <https://doi.org/10.1016/j.jtbi.2015.02.026>.
- [9] Hoang T, Yin C, Yau S. Numerical encoding of DNA sequences by Chaos game representation with application in similarity comparison. *Genomics* 2016;108(3–4):134–42.
- [10] Deng M, Yu C, Liang Q, He R, Yau S. Correction: a novel method of characterizing genetic sequences: genome space with biological distance and applications. *Plos One* 2012;6(3):e17293.
- [11] Kwan H, Arniker S. Numerical representation of DNA sequences. *IEEE Int Conf Electro/Inf Technol* 2009:307–10.
- [12] Voss R. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys Rev Lett* 1992;68(25):3805–8. <https://doi.org/10.1103/PhysRevLett.68.3805>.
- [13] Yin C, Yau S. An improved model for whole genome phylogenetic analysis by Fourier transform. *J Theor Biol* 2015;382:99–110. <https://doi.org/10.1016/j.jtbi.2015.06.033>.
- [14] Qi J, Luo H, Hao B. CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Res* 2004;32(2):W45–7.
- [15] Sims G, Jun S, Wu G, Kim S. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *PNAS* 2009;106(8):2677–82. <https://doi.org/10.1073/pnas.0813249106>.
- [16] Yu C, He R, Yau S. Protein sequence comparison based on K-string dictionary. *Gene* 2013;529(2):250–6. <https://doi.org/10.1016/j.gene.2013.07.092>.
- [17] Sharma D, Issac B, Raghava G, Ramaswamy R. Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics* 2004;20(9):1405–12. <https://doi.org/10.1093/bioinformatics/bth103>.
- [18] Dougherty E, Huang Y, Kim S, Cai X, Rui Y. Genomic signal processing signal processing. *10(6)*; 2009; 364.
- [19] Yin C, Yau S. A Fourier characteristic of coding sequences: origins and a nonFourier approximation. *J Comput Biol* 2005;12(9):1153–65. <https://doi.org/10.1089/cmb.2005.12.1153>.
- [20] Yin C, Yau S. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *J Theor Biol* 2007;247(4):687–94. <https://doi.org/10.1016/j.jtbi.2007.03.038>.
- [21] Dong R, Zhu Z, Yin C, He R, Yau S. A new method to cluster genomes based on cumulative Fourier power spectrum. *Gene* 2018;673(5):239–50.
- [22] Tanchotsrinon W, Lursinsap C, Poovorawan Y. A high-performance prediction of HPV genotypes by chaos game representation and singular value decomposition. *BMC Bioinforma* 2015;16(1):71.
- [23] Yu C, Liang Q, Yin C, He R, Yau S. A novel construction of genome space with biological geometry. *DNA Res* 2019;17(3):155–68.
- [24] Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 2016;33(7):1870. <https://doi.org/10.1093/molbev/msw054>.
- [25] Alexander D. A review of avian influenza in different bird species. *Vet Microbiol* 2000;74(1):3–13.
- [26] Sun K, Zhang B. Human papillomavirus risk type classification from protein sequences using support vector machines. Applications of evolutionary computing, Budapest, Hungary, April 10–12, proceedings; 2006. p. 57–66.
- [27] Park S, Hwang S, Zhang B. Classification of the risk types of human papillomavirus by decision trees. Intelligent data engineering and automated learning, international conference, ideal 2003, Hong Kong, China; March 21–23, 2003. p. 540544 [Revised Papers].
- [28] Sun K, Kim J, Zhang B. Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures. *Comput Biol Med* 2009;39(2):187–93.
- [29] Brown J, Doolittle W. Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* 1997;61(4):456–502.
- [30] Huang H, Yu C, Zheng H, Hernandez T, Yau S, He R, et al. Global comparison of multiple-segmented viruses in 12-dimensional genome space. *Mol Phylogenet Evol* 2014;81:29–36. <https://doi.org/10.1016/j.ympev.2014.08.003>.
- [31] Yu C, He R, Yau S. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. *J Theor Biol* 2014;348:12–20. <https://doi.org/10.1016/j.jtbi.2014.01.022>.