An efficient numerical representation of genome sequence: natural vector with covariance component

Nan Sun[1], Xin Zhao[2] and Stephen S.-T. Yau[1,3]

[1] Department of Mathematical Sciences, Tsinghua University, Beijing, China
[2] Beijing Electronic Science and Technology Institute, Beijing, China
[3] Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, China

ABSTRACT

Background: The characterization and comparison of microbial sequences, including archaea, bacteria, viruses and fungi, are very important to understand their evolutionary origin and the population relationship. Most methods are limited by the sequence length and lack of generality. The purpose of this study is to propose a general characterization method, and to study the classification and phylogeny of the existing datasets.

Methods: We present a new alignment-free method to represent and compare biological sequences. By adding the covariance between each two nucleotides, the new 18-dimensional natural vector successfully describes 24,250 genomic sequences and 95,542 DNA barcode sequences. The new numerical representation is used to study the classification and phylogenetic relationship of microbial sequences.

Results: First, the classification results validate that the six-dimensional covariance vector is necessary to characterize sequences. Then, the 18-dimensional natural vector is further used to conduct the similarity relationship between giant virus and archaea, bacteria, other viruses. The nearest distance calculation results reflect that the giant viruses are closer to bacteria in distribution of four nucleotides.

The phylogenetic relationships of the three representative families, Mimiviridae, Pandoraviridae and Marsellieviridae from giant viruses are analyzed. The trees show that ten sequences of Mimiviridae are clustered with Pandoraviridae, and Mimiviridae is closer to the root of the tree than Marsellieviridae. The new developed alignment-free method can be computed very fast, which provides an effective numerical representation for the sequence of microorganisms.

Subjects Bioinformatics, Genomics, Mathematical Biology, Microbiology, Data Mining and Machine Learning

Keywords Bacteria, Virus, Giant virus, Archaea, Fungi, Convex hull classification, The nearest neighbor classification, Phylogeny, Natural vector with covariance component

Submitted 16 November 2021
Accepted 16 May 2022
Published 16 June 2022

Corresponding author
Stephen S.-T. Yau, yau@uic.edu

Academic editor
Alexander Bolshoy

Additional Information and Declarations can be found on page 20

DOI **10.7717/peerj.13544**

INTRODUCTION

With the increasingly close relationship between microorganisms and human beings, a deeper understanding of microorganisms becomes important (*Wessner, Dupont & Charles, 2013*). Comparing sequence similarity and inferring phylogenetic relationship is helpful to understand their properties, so as to reduce the harm of microorganisms and let them serve human beings better. Molecular biologists believe that similar sequences have

similar functions. If the function of microbial species is known, the function of microorganisms with similar sequences can be inferred. Similar sequences can be obtained by sequence comparison. Two methods, alignment-free and alignment methods, are known to compare sequences. Traditional commonly used alignment approaches include MUSCLE (*Edgar, 2004a*; *Edgar, 2004b*), ClustalW (*Larkin, Blackshields & Brown, 2007*), and BLAST (*Altschul et al., 1990*; *Altschul et al., 1997*), but they are very time-consuming and require a lot of memory. Alignment-free methods are developed to overcome these limitations, such as chaos game representation (*Jeffrey, 1990*; *Hatje & Kollmar, 2012*), Fourier transform (*Yin, Chen & Yau, 2014*), information theory (*Vinga, 2014*; *Almeida, 2014*), k-mer theory (*Dai, Yang & Wang, 2008*; *Leimeister & Morgenstern, 2014*), *etc.* The alignment-free methods do not require the neutral theory assumption and can process a large number of microbial sequences.

Yau and his team proposed a powerful alignment-free method, a 12-dimensional natural vector, to describe the nucleotides distribution within the DNA sequence (*Deng et al., 2011*). This natural vector consists of the count, average position and central moment of each nucleotide. It has been successfully applied to many research fields, especially the clustering (*Zhao, Tian & Yau, 2018*; *Zhao et al., 2019*; *Sun et al., 2021*) and classification (*He et al., 2020*; *Pei et al., 2020*) of biological sequences. The 12-dimensional natural vector only considers the respective distribution of single nucleotide but ignores the relationship between the two nucleotides. This inspires us to extend the definition and further consider the correlation of each two nucleotides. There are four nucleotides for a sequence, and the new natural vector is 18-dimensional, of which the covariance component is six dimensional.

Microorganisms are diverse and live in every part of the biosphere, which mainly include archaea, bacteria, virus, fungi, and some other protozoa (*Wessner, Dupont & Charles, 2013*). As a vital part of microorganisms, the study of the giant virus has attracted much attention. Their evolutionary origin and the relationship with other viruses, bacteria and archaea remain controversial (*Bichell, 2017*). Giant viruses have extremely large genomes, some of which are even larger than bacterial genome (*Ogata, Toyoda & Tomaru, 2009*). Recently discovered giant viruses have even longer genome sequences and more encoding genes (*Van Etten, 2011*; *Legendre, Arslan & Abergel, 2012*). Researchers claim two hypotheses that they evolve either from small viruses or from very complex organisms. These enlighten us to analyze microbial sequences using our method.

In this article, we improve the 12-dimensional natural vector by defining the covariance between nucleotides, and add the six-dimensional covariance vector to the 12-dimensional vector. The 18-dimensional natural vector can effectively represent a sequence, and has been tested on five genomic sequence datasets and one DNA barcode sequence dataset. The results of convex hull classification show that the six-dimensional covariance vector is necessary to characterize sequences. The study of giant virus based on our method gives more stable results than other alignment-free methods. Our new 18-dimensional natural vector shows outstanding ability in classification and phylogeny of biological sequences.

MATERIALS AND METHODS

Natural vector

Natural vector is a 12-dimensional numerical representation of a nucleotide sequence (*Deng et al., 2011*), which is defined as follows. Let $S = s_1,\ s_2,\ s_3,\ \ldots s_n$ be a genomic sequence of length n, and $L = \{A,\ C,\ G, \text{T or U}\}$. For $k\ \in\ L$, the indicator functions $w_k(\cdot)$: $L \rightarrow \{0, 1\}$ is defined as:

$$w_k(s_i) = \begin{cases} 1, & if\ s_i = k, \\ 0, & otherwise. \end{cases} \tag{1}$$

where $s_i\ \in\ L,\ i = 1,\ 2,\ 3,\ \ldots,\ n$. Let $n_k = \sum_{i=1}^{n} w_k(s_i)$ denote the counts of nucleotide k in S, $\mu_k = \sum_{i=1}^{n} i\frac{w_k(s_i)}{n_k}$ specify the average location of letter k, $D_j^k = \sum_{i=1}^{n} \frac{(i-\mu_k)^j w_k(s_i)}{n_k^{j-1} n^{j-1}}$ be the j-th central moment of position of letter k. If $j = 2$, we get the traditional 12-dimensional natural vector:

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T) \tag{2}$$

Here we give an example to calculate the vector. If the genomic sequence is ACGGTAGTCC, the indicator functions are $w_A = 1000010000,\ w_C = 0100000011,\ w_G = 0011001000,\ w_T = 0000100100$. Each component of the vector is calculated as follow:

- $\mathrm{n_A} = 2, \mathrm{n_C} = 3, \mathrm{n_G} = 3, \mathrm{n_T} = 2,$
- $\mu_\mathrm{A} = 1 \cdot \frac{1}{2} + 6 \cdot \frac{1}{2} = 3.5,$
- $\mu_C = 2 \cdot \frac{1}{3} + 9 \cdot \frac{1}{3} + 10 \cdot \frac{1}{3} = 7,$
- $\mu_\mathrm{G} = 3 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 7 \cdot \frac{1}{3} = 4.67,$
- $\mu_T = 5 \cdot \frac{1}{2} + 8 \cdot \frac{1}{2} = 6.5,$
- $D_2^A = \frac{(1-\frac{7}{2})^2}{2 \cdot 10} + \frac{(6-\frac{7}{2})^2}{2 \cdot 10} = 0.625,$
- $D_2^C = \frac{(2-7)^2}{3 \cdot 10} + \frac{(9-7)^2}{3 \cdot 10} + \frac{(10-7)^2}{3 \cdot 10} = 1.27,$
- $D_2^G = \frac{(3-\frac{14}{3})^2}{3 \cdot 10} + \frac{(4-\frac{14}{3})^2}{3 \cdot 10} + \frac{(7-\frac{14}{3})^2}{3 \cdot 10} = 0.289,$
- $D_2^T = \frac{(5-\frac{13}{2})^2}{2 \cdot 10} + \frac{(8-\frac{13}{2})^2}{2 \cdot 10} = 0.225.$

Then the 12-dimensional natural vector in formula (2) is:

$(2,\ 3,\ 3,\ 2,\ 3.5,\ 7,\ 4.67,\ 6.5,\ 0.625,\ 1.27,\ 0.289,\ 0.225).$

A novel definition of covariance between nucleotides

The 12-dimensional natural vector definition reveals the respective distribution of four nucleotides, but ignores the relationship of each two nucleotides. We improve the 12-dimensional vector by defining the covariance between nucleotides. For the same genomic sequence $S = s_1,\ s_2,\ s_3, \ldots s_n$, and $L = \{A,\ C,\ G,\ \mathrm{T\ or\ U}\}$, we redefine the indicator function, $w_{kl}(\cdot) = w_{lk}(\cdot)$: $L \to \{0, 1\},\ k,\ l\ \in\ L$:

$$w_{kl}(s_i) = w_{lk}(s_i) = \begin{cases} 1, & if\ s_i = k\ or\ l, \\ 0, & otherwise. \end{cases} \tag{3}$$

And the covariance between k and l is defined as:

$$Cov(k, l) = \sum_{i=1}^{n} \frac{[i - \mu_k][i - \mu_l] w_{kl}(s_i)}{n\sqrt{n_k}\sqrt{n_l}}. \tag{4}$$

The above formula reflects the correlation relationship of the position of each two nucleotides. The beauty of the definition is $Cov(k, k) = D_2^k$.

Thus, we get a novel 18-dimensional natural vector with covariance component:

$$\begin{aligned} &(n_A,\ n_C,\ n_G,\ n_T,\ \mu_A,\ \mu_C,\ \mu_G, \mu_T,\ D_2^A,\ D_2^C,\ D_2^G,\ D_2^T, \\ &Cov(A, C),\ Cov(A, G),\ Cov(A, T),\ Cov(C, G),\ Cov(C, T),\ Cov(G, T)). \end{aligned} \tag{5}$$

For the same sequence ACGGTAGTCC, the indicator functions are $w_{AC} = 1100010011$, $w_{AG} = 1011011000$, $w_{AT} = 1000110100$, $w_{CG} = 0111001011$, $w_{CT} = 0100100111$, $w_{GT} = 0011101100$. The corresponding components are calculated as follows:

- $Cov(A, C) = \sum_{i \in \{1,\ 2,\ 6,\ 9,\ 10\}} \frac{[i - 3.5][i - 7]}{10 \cdot \sqrt{2} \cdot \sqrt{3}} = 2.06,$
- $Cov(A, G) = \sum_{i \in \{1,\ 3,\ 4,\ 6,\ 7\}} \frac{[i - 3.5][i - 4.67]}{10 \cdot \sqrt{2} \cdot \sqrt{3}} = 0.864,$
- $Cov(A, T) = \sum_{i \in \{1,\ 5,\ 6,\ 8\}} \frac{[i - 3.5][i - 6.5]}{10 \cdot \sqrt{2} \cdot \sqrt{2}} = 0.85,$
- $Cov(C, G) = \sum_{i \in \{2,\ 3,\ 4,\ 7,\ 9,\ 10\}} \frac{[i - 7][i - 4.67]}{10 \cdot \sqrt{3} \cdot \sqrt{3}} = 1.56,$
- $Cov(C, T) = \sum_{i \in \{2,\ 5,\ 8,\ 9,\ 10\}} \frac{[i - 7][i - 6.5]}{10 \cdot \sqrt{3} \cdot \sqrt{2}} = 1.735,$
- $Cov(G, T) = \sum_{i \in \{3,\ 4,\ 5,\ 7,\ 8\}} \frac{[i - 4.67][i - 6.5]}{10 \cdot \sqrt{3} \cdot \sqrt{2}} = 0.54.$

Then the nucleotide distribution of sequence ACGGTAGTCC can be described by an 18-dimensional natural vector with covariance component:

$$\begin{aligned} &(2,\ 3,\ 3,\ 2,\ 3.5,\ 7,\ 4.67,\ 6.5,\ 0.625,\ 1.27,\ 0.289,\ 0.225, 2.06,\ 0.864,\ 0.85, \\ &1.56,\ 1.735,\ 0.54). \end{aligned}$$

Table 1 The summary of the datasets in this study.

	Dataset	Type	Number of sequences	Number of families	Database access link and description
1	Archaea	genomic sequence	298	20	ftp.ncbi.nih.gov/refseq/release/archaea/
2	Bacteria	genomic sequence	16,375	178	ftp.ncbi.nih.gov/refseq/release/bacteria/
3	Virus	genomic sequence	7,382	83	ftp.ncbi.nlm.nih.gov/genomes/Viruses
4	Fungi	genomic sequence	387	22	ftp.ncbi.nih.gov/refseq/release/fungi/
5	Giant virus	genomic sequence	677	16	Giant virus toplist: https://pitgroup.org/giant-virus-toplist/; The largest known viral genomes (completely sequenced, >170 kb): http://www.giantvirus.org/2014-04-14top.html; Reference sequences from Virus dataset, which genome size is over 170 kb, and the CDS counts are over 100.
6	Fungi	DNA barcode	95,542	467	Barcode of Life Data System (BOLD): http://www.barcodinglife.org

The biological similarity between two sequences can be measured using the Euclidean distance of their corresponding 18-dimensional natural vectors, which is commonly used in our previous studies (*Deng et al., 2011*; *Zhao, Tian & Yau, 2018*; *Zhao et al., 2019*; *Sun et al., 2021*; *He et al., 2020*; *Pei et al., 2020*).

Datasets and tools

The newly proposed alignment-free method is tested on six microorganism datasets, including the genome and gene sequences of archaea, bacteria, virus and fungi, as shown in Table 1. The genomes of archaea, bacteria, virus and fungi are downloaded from National Center for Biotechnology Information (NCBI). Giant viruses in this study are regarded as the viruses with genome size greater than 170 kb, which are collected to study their relationship with archaea, bacteria, and other viruses. This giant virus dataset is collected from two public databases (Giant virus toplist: https://pitgroup.org/giant-virus-toplist/; The largest known viral genomes (completely sequenced, >170 kb): http://www.giantvirus.org/2014-04-14top.html) and virus dataset (Viruses with genome size greater than 170 kb). In order to compare with the previous nucleotide covariance study (*Zhao, Tian & Yau, 2018*), we re-download DNA barcode of fungi from Barcode of Life Data System (BOLD) according to the dataset record (*Zhao, Tian & Yau, 2018*).

To ensure the reliability of the data, we remove three types of sequences from the original datasets: (1) sequences without family taxonomic information; (2) families with less three sequences. (3) sequences of plasmid for bacteria dataset. DNA barcode uniquely identifies species using a short fragment of DNA sequence from specific genes (*iBOL, 2022*). For this DNA barcode of fungi dataset, we only remain the sequences pertaining to internal transcribed spacer (ITS) region of fungi (*Conrad, Keith & Sabine, 2012*; *Naturvetenskapliga, 2010*). At last, there are 298 genomic sequences belonging to 20 families for archaea dataset, 16,375 genomic sequences belonging to 178 families for bacteria dataset, 7,382 genomic sequences belonging to 83 families for virus dataset, 387 genomic sequences belonging to 22 families for fungi genome dataset, 677 sequences belonging to 16 families for giant virus dataset, and 95,542 sequences belonging to 467

families for fungi DNA barcode dataset. All accession numbers can be found in Datas S1–S6.

All the programs in this article are written in MATLAB R2020a and run on the same laptop (MacBook Air, 1.8 GHz Intel Core i5, 8 GB 1,600 MHz DDR3).

Convex hull principle for genomes

Convex hull principle for genomes has been demonstrated to be a wonderful classification tool (*Zhao, Tian & Yau, 2018*; *Zhao et al., 2019*; *Sun et al., 2021*). In Mathematics, the convex hull of a point set $A = \{a_1,\ a_2,\ \ldots, a_m\}, a_i\ \in\ R^k$ is the minimal convex set that contains these m points, where R^k is the k-dimensional Euclidean space, and a_i is a k-dimensional point. The convex hull of finite set A is defined as the set of convex combinations of all points in A:

$$\begin{aligned} CovA =&\{\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_m a_m : a_i\ \in\ A, \lambda_1 + \lambda_2 + \ldots + \\ &\lambda_m = 1, \lambda_i \geq 0, \quad i = 1, 2, \ldots, m\} \end{aligned} \tag{6}$$

The boundary of the convex hull is spanned by some points of A (vertexes), and the rest points of A are lying inside the hull. The convex hull is a convex polygon if all $a_i (i = 1, 2, \ldots, m)$ are two-dimensional vectors, and the convex hull is a convex polytope if all $a_i (i = 1, 2, \ldots, m)$ are high dimensional vectors. Particularly, triangles and tetrahedrons are convex hulls.

Here $a_i (i = 1, 2, \ldots, m)$ are 18-dimensional natural vectors, and they are divided into several classes (families). Intuitively, the distribution of nucleotides from the same family is similar, and the 18-dimensional points from the same families lie closely. The sequence is transformed into an 18-dimensional numerical vector first, and the points from the same family can form a convex hull. Convex hull principle states that convex hulls corresponding to different families are mutually disjoint (*Zhao et al., 2019*).

Linear programming (LP) can be used to check whether two convex hulls are disjoint (*Sun et al., 2021*). If $A = Cov\{a_1, a_2, \ldots, a_m\}$ and $B = Cov\{b_1, b_2, \ldots, b_n\}$ intersect, then the convex combination of these points satisfy the formula: $\sum_{i=1}^{m} \lambda_i a_i = \sum_{j=1}^{n} \beta_j b_j$, where $\sum_{i=1}^{m} \lambda_i = 1, \sum_{j=1}^{n} \beta_j = 1$. a_i or b_j is an 18-dimensional natural vector. It means that the following linear programming problem has a feasible solution (That is, there are non-zero coefficients $\{\lambda_1, \lambda_2, \ldots, \lambda_m; \beta_1, \beta_2, \ldots, \beta_n\}$ such that the minimum value of the optimization problem is 0):

$$\begin{aligned} &\min\ 0. \\ &s.t. \sum_{i=1}^{m} \lambda_i a_i = \sum_{j=1}^{n} \beta_j b_j \\ &\sum_{i=1}^{m} \lambda_i = 1, \lambda_i \geq 0, i = 1, 2, \ldots, m \\ &\sum_{j=1}^{n} \beta_j = 1, \beta_j \geq 0, j = 1, 2, \ldots, n \end{aligned} \tag{7}$$

The above problem can be implemented through *linprog* function built in MATLAB.

Table 2 The number of disjoint convex hull pairs based on traditional 12-dimensional natural vector and our new 18-dimensional natural vector of adding the six-dimensional covariance vector. For 178 families of bacteria, there are 15,753 convex hull pairs. The number of disjoint convex hull pairs based on 18-dimensional natural vector is 15,160, which is more than that based on traditional 12-dimensional natural vectors (14,565). The other four datasets have similar conclusions. The results show that the six-dimensional covariance vector is necessary to characterize sequences. C_n^k represents the combinatorial number.

Dataset	Convex hull pairs	12-dim	18-dim
Archaea	$C_{20}^2 = 190$	184	190
Bacteria	$C_{178}^2 = 15{,}753$	14,565	15,160
Virus	$C_{83}^2 = 3{,}403$	3,321	3,322
Fungi (Genome)	$C_{22}^2 = 231$	207	227
Fungi (DNA barcode)	$C_{467}^2 = 108{,}811$	75,237	88,719

RESULTS

Convex hull classification of archaea, bacteria, virus and fungi

The five datasets, archaea, bacteria, virus, fungi (genome), fungi (DNA barcode), are used for convex hull classification, and the necessity of adding extra six-dimensional nucleotide covariance component to 12-dimensional natural vector is verified. For each dataset, we first calculate the 12-dimensional or 18-dimensional natural vector to describe the distribution of each sequence, and construct convex hull for each family. Then the LP method is utilized to check whether the convex hulls of different families are disjoint. The comparison of classification performance between 12-dimensional natural vector and 18-dimensional natural vector is displayed in Table 2. For archaea dataset, there are 190 convex hull pairs, and 184 convex hull pairs are mutually disjoint in 12-dimensional space, while all pairs are disjoint in 18-dimensional space. For bacteria dataset, there are $C_{178}^2 = 15{,}753$ convex hull pairs, and 15,160 pairs are not intersected using 18-dimensional vector method. The non-intersection ratio in 18-dimensional space is 96.24%, which is larger than that in 12-dimensional space (92.46%). For virus dataset, there are 3,403 convex hull pairs. A total of 3,322 pairs are disjoint in 18-dimensional space, and 3,321 pairs are disjoint in 12-dimensional space. For both fungi (genome) and fungi (DNA barcode) datasets, the non-intersection ratios in 18-dimensional space are higher than those in 12-dimensional space. The results demonstrate that our new 18-dimensional natural vector performs better classification results than 12-dimensional natural vector. The 18-dimensional vector contains more sequence information because it considers the correlation relationships of the two nucleotides. The results of fungi (DNA barcode) dataset further show that covariance can be used not only for sequence classification at genome level, but also for sequence classification at gene sequence level.

To visualize the convex hull classification results, we use the traditional support vector machine method (SVM) to reduce the dimension (*Sun et al., 2021*). For two disjoint convex hulls in 18-dimensional space, there is a hyperplane to separate them:

$$w^T x + b = 0, \tag{8}$$

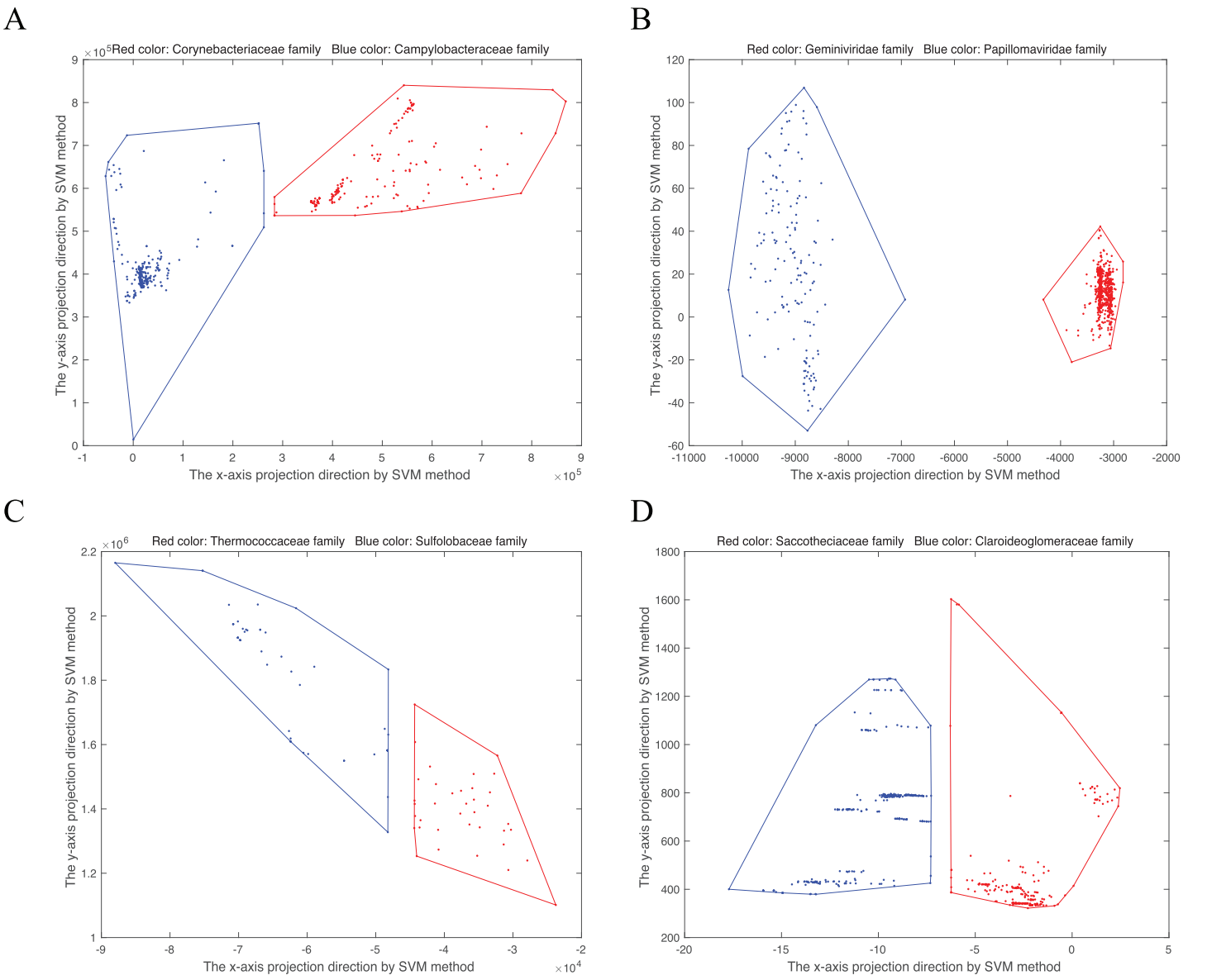

Figure 1 Convex hull pairs after dimension reduction by SVM method. (A) Convex hull pair constructed by two bacterial families, Corynebacteriaceae (243 points) and Campylobacteraceae (305 points). (B) Convex hull pair constructed by two virus families, Geminiviridae (565 points) and Papillomaviridae (154 points). (C) Convex hull pair constructed by two archaea families, Thermococcaceae (37 points) and Sulfolobaceae (51 points). (D) Convex hull pair constructed by two fungi (DNA barcode) families, Saccotheciaceae (274 points) and Claroideoglomeraceae (301 points).
Full-size DOI: 10.7717/peerj.13544/fig-1

$w = (w_1, w_2, \ldots, w_k)^T$ is the normal vector, b is the offset item. There is at least a vector v perpendicular to w. For an 18-dimensional point, named NV in the convex hull, the new point projected into two-dimensional space is $(w^T \cdot NV, v^T \cdot NV)$. Then the convex hull can be visualized in two-dimensional space. As shown in Fig. 1, we randomly select four disjoint convex hull pairs. Points from the same family gather together.

Preliminary statistical analysis of the genomes of giant viruses

The research on phylogenetic relationship between giant virus and bacteria, archaea and other viruses is still active (*Claverie & Abergel, 2013*). The discovery of Mimivirus comes as a shock in biological field for a long time (*Birtles, Rowbotham & Storey, 1997*), and its genome is about 118 kbp (*Didier, Stéphane & Catherine, 2004*). Later viral genomes larger than 118 kbp started to accumulate quickly. This suggests that viruses with genomes larger than 118 kbp are not as exceptional as previously thought, especially Pandoravirus, which is discovered in recent years, with a length of more than 2,500 kbp (*Philippe et al., 2013*). Recent discoveries are not sufficient to erase the tremendous gap in genome size between giant viruses (*Brandes & Linial, 2019*), which is also reflected in Fig. 2. The x-axis represents the sequence number, the y-axis represents the genome size, and each dot represents a sequence. There are many gaps between genomes sizes of "Giant Virus" (blue color). Here "Giant Virus" only indicates the sequences of giant virus collected from two public databases (genome size > 170 kb). "Giant & RefSeq" indicates the sequences of giant virus collected from virus dataset (genome size > 170 kb). "Other Virus RefSeq" indicates the rest sequences in virus dataset except for giant virus. In Fig. 2A, there is a significant slope change at about 300 kbp, and the genome size of giant viruses is larger than that of other viruses (green color). In Fig. 2B, it is found that the genome sizes of several giant viruses and bacteria overlap at about 2,000 kbp, and the genome sizes of some viruses are larger than those of some bacteria. In Fig. 2C, the genome sizes of several giant viruses also overlap at about 2,000 kbp. Fig. 2 gives an intuitive understanding of the genome size relationship between giant virus and bacteria, archaea, other viruses. There is not yet any reason to believe that the genome size of giant viruses has reached the upper limit, and viruses with different genome lengths may emerge continuously (*Claverie & Abergel, 2013*). The origin of the giant virus still remains mysterious, so we can study it through the relationship between genomes of similar length.

We next make preliminary statistics on the gene distribution to understand the giant virus. All known giant viruses belong to the phylum Nucleocytoviricota (nucleocytoplasmic large DNA viruses, NCLDVs), which contains many unique genes that cannot be found in other life forms (*Ogata, Toyoda & Tomaru, 2009*). Here we only consider the coding sequence. We plot the number of NCLDV coding sequences on the y-axis, and the genome size on the x-axis. The results are shown in Fig. 3. The representative families of giant viruses, Mimiviridae and Pandoraviridae have larger genome size and more coding sequences (Mimiviridae: purple star, Pandoraviridae: black triangle). In addition, the G+C content of Pandoraviridae is the highest among NCLDVs (average 61.3%), and the earliest discovered virus group, Mimiviridae has the lowest G+C content (average 26.3%, Table S1).

Nearest neighbor classification of the genomes of giant viruses

The similarity of genomic distribution between giant viruses and bacteria, archaea, other viruses is of great importance to understand their phylogenetic relationship. We use a new 18-dimensional natural vector with covariance to measure it. There are 24,250 different biological sequences in these four datasets (Genomes of giant virus, bacteria, archaea, and

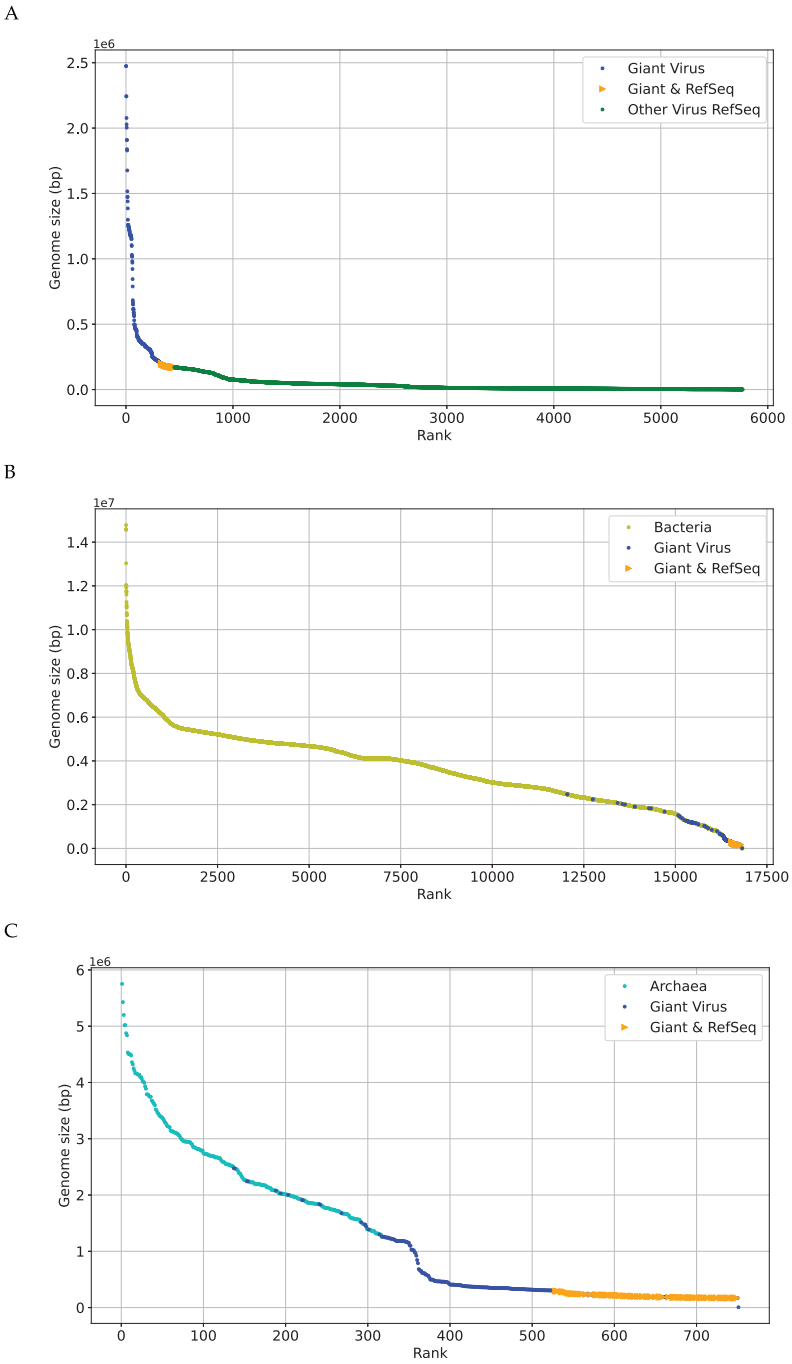

Figure 2 Genome size distribution of giant viruses. X-axis represents the number of the sequence, and Y-axis represents the genome size. Here "Giant Virus" only represents the sequences (genome size > 170 kb) of giant virus collected from two public databases. "Giant & RefSeq" represents the sequences (genome size > 170 kb) of giant virus collected from virus dataset. "Other Virus RefSeq" represents the rest sequences in virus dataset except for giant virus. (A) Genome size distribution of "Giant Virus", "Giant & RefSeq" and "Other Virus RefSeq". (B) Genome size distribution of bacteria, "Giant Virus" and "Giant & RefSeq". (C) Genome size distribution of archaea, "Giant Virus" and "Giant & RefSeq".

Full-size DOI: 10.7717/peerj.13544/fig-2

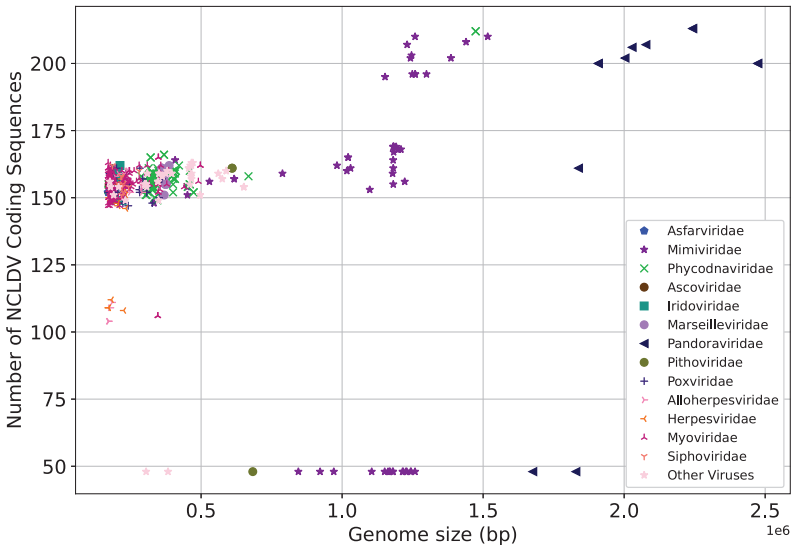

Figure 3 Number of coding sequences of nucleocytoplasmic large DNA viruses. Each dot represents a sequence. X-axis represents the genome size, and Y-axis represents the number of coding sequences of NCLDV. The representative families of giant viruses, Mimiviridae and Pandoraviridae have larger genome size and more coding sequences. Full-size DOI: 10.7717/peerj.13544/fig-3

virus), of which 677 sequences belong to 31 groups for giant virus dataset. Each sequence is converted into an 18-dimensional natural vector first. For each sequence in giant virus dataset, the closest sequence is then calculated and the sequence type is checked. The Euclidean distance between two 18-dimensional natural vectors is used to measure the biological similarity of the corresponding sequences. The nearest neighbor classification results are shown in Table 3.

Except for Polydnaviridae and Myoviridae, the nucleotide distribution of most sequences is similar to that of bacteria $\left(\frac{285}{309} \times 100\% = 92.23\%\right)$. About half of the 138 Myoviridae sequences are close to the bacterial genomes (67 sequences) and half to the viral genomes (71 sequences). One of the reasons for this phenomenon is that the sequence length of Myoviridae is similar to that of some bacteria. Of the 230 Polydnaviridae sequences, 229 are closest to the virus sequence. These 230 sequences belong to two genera, *Ichnovirus* (IV) and *Bracovirus* (BV), and five species (Table S2). *Glypta fumiferanae ichnovirus* (GfIV) has 105 segments, which is the virus with the most segments among all segmented viruses. The segment length is about 2,777 bp, which is shorter than bacterial sequence. In addition, all sequences of three representative giant virus families, Mimiviridae, Marseilleviridae, and Pandoraviridae are closer to bacterial sequences.

Phylogenetic analysis of the genomes of giant viruses

Phylogenetic relationships between giant virus and bacteria, archaea, other viruses are studied using our new alignment-free method. We select three representative families from giant virus (Mimiviridae, Pandoraviridae, and Marseilleviridae) and randomly select two or three families from other three datasets to construct the phylogenetic tree. Figure 4 shows the phylogenetic tree of three other virus families (Adenoviridae, Anelloviridae, and

Table 3 Statistics of the closest sequence types for giant virus dataset.

	Group name of large genome virus	Seq Nu.	Nearest type: Bacteria	Nearest type: Virus	SeqLen			
					Max	Min	Mean	Median
1	Polydnaviridae	230	1	229	41,573	1,533	6,684	3,836
2	Myoviridae	138	67	71	4,97,513	170,286	229,450	191,652
3	Phycodnaviridae	82	76	6	1,473,573	171,045	362,340	329,946
4	Mimiviridae	59	59	–	1,516,267	317,278	1,048,973	1,181,042
5	Poxviridae	39	34	5	359,853	170,560	244,824	224,499
6	Alloherpesviridae	20	17	3	295,146	171,096	224,435	223,830
7	Herpesviridae	17	12	5	233,501	171,823	203,228	204,237
8	Marseilleviridae	13	13	–	386,631	360,610	370,058	369,360
9	Pandoraviridae	12	12	–	2,473,870	1,676,110	2,058,604	2,015,816
10	Iridoviridae	10	10	–	220,222	186,250	202,184	201,674
11	Siphoviridae	10	10	–	279,967	185,683	223,445	219,073
12	Faustovirus	8	8	–	470,659	455,803	464,660	465,984
13	Nimaviridae	7	7	–	309,286	300,223	305,918	305,119
14	Cedratvirus	4	4	–	589,068	560,887	578,546	582,115
15	Pithoviridae	3	3	–	683,254	610,033	634,440	610,033
16	*Apis mellifera* filamentous virus	2	2	–	496,396	496,396	496,396	496,396
17	Ascoviridae	2	–	2	186,262	174,059	180,161	180,161
18	Kaumoebavirus	2	2	–	350,731	350,731	350,731	350,731
19	Lausannevirus	2	2	–	346,754	346,754	346,754	346,754
20	Mollivirus	2	2	–	651,523	651,523	651,523	651,523
21	Pacmanvirus	2	2	–	395,405	395,405	395,405	395,405
22	Bamfordvirae	2	2	–	380,011	349,275	364,643	364,643
23	Baculoviridae	2	–	2	178,733	176,677	177,705	177,705
24	Asfarviridae	1	–	1	170,101	170,101	170,101	170,101
25	Brazilian cedratvirus	1	1	–	460,038	460,038	460,038	460,038
26	Brazilian Marseilleviridae	2	2	–	362,276	362,276	362,276	362,276
27	Cannes 8 virus	1	1	–	374,041	374,041	374,041	374,041
28	Glossinavirus	1	1	–	190,032	190,032	190,032	190,032
29	Nudiviridae	1	1	–	231,621	231,621	231,621	231,621
30	Insectomime	1	1	–	382,785	382,785	382,785	382,785
31	Malacoherpesviridae	1	1	–	207,439	207,439	207,439	207,439

Closteroviridae) and the three families of giant virus, different groups are marked in different colors. All groups are separate except Mimiviridae. To clearly observe the relationships between the three giant virus families, two bacterial families, Pseudomonadaceae and Alteromonadaceae are selected for further analysis.

The phylogenetic result in Fig. 5 shows that all sequences from the same family are clustered except Mimiviridae. Beyond that, two archaea families, Methanosarcinaceae and Natrialbaceae are selected to construct the phylogenetic tree, the same family groups together except Mimiviridae is shown in Fig. 6. The algorithm of constructing the

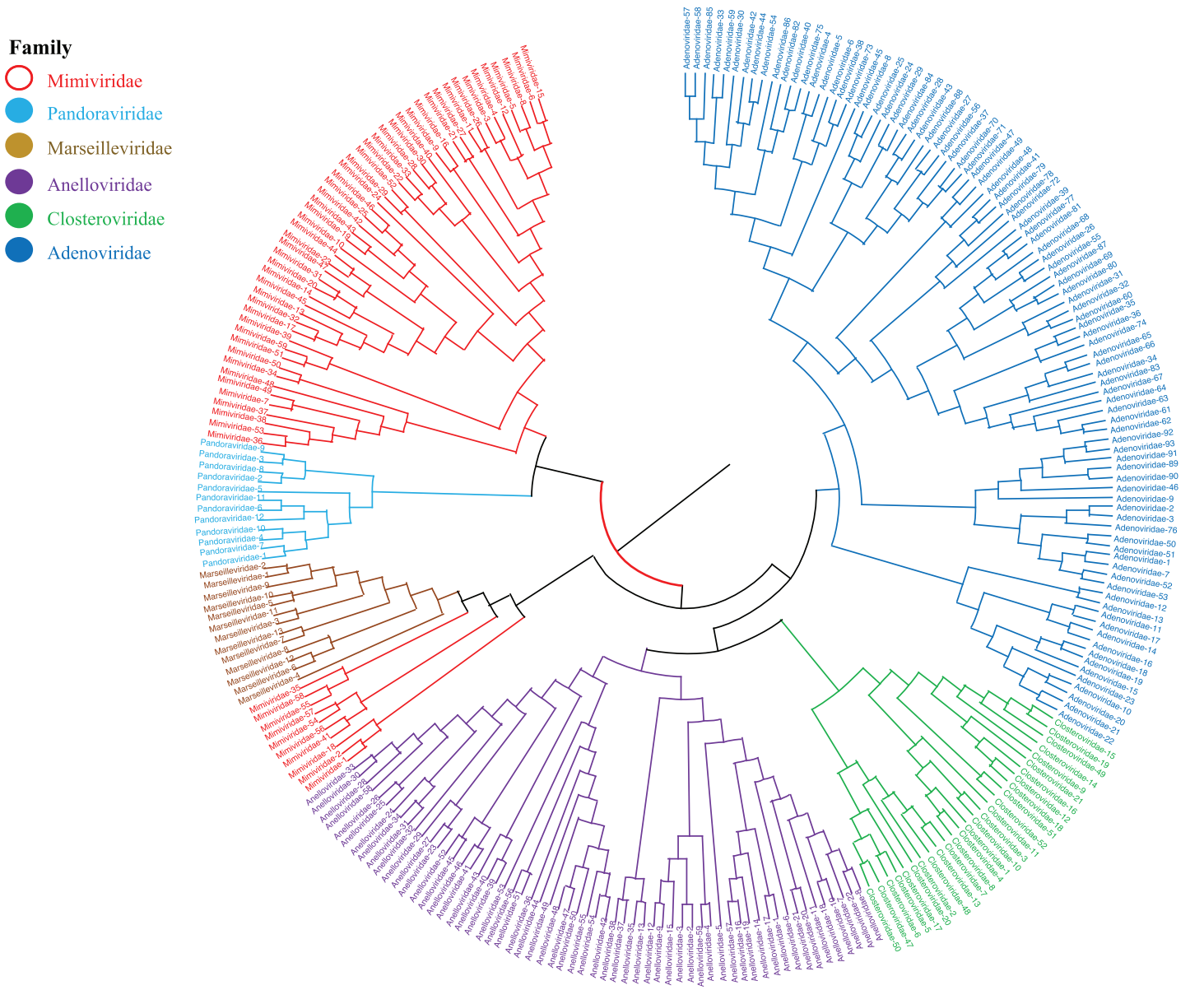

Figure 4 Phylogenetic tree for the three representative families (Mimiviridae, Pandoraviridae, and Marseilleviridae) of giant virus and three other viral families (Adenoviridae, Anelloviridae, and Closteroviridae). Different colors represent different families.

Full-size DOI: 10.7717/peerj.13544/fig-4

phylogenetic tree is unweighted pair-group method with arithmetic mean (UPGMA), which is an approach of constructing rooted tree based on distance matrix. The distance matrix is calculated using the commonly used Euclidean distance. In Figs. 4–6, the same 10 Mimiviridae species and Marseilleviridae species are clustered together. The sequence information is displayed in Table S3. The length of the 10 sequences is shorter than that of other viruses of Mimiviridae, but similar to some viruses of Marseilleviridae.

The average sequence length of Marseilleviridae is about 370 kbp, that of Mimiviridae is about 1,000 kbp, and the length of 10 sequences is hundreds of thousands of bp.

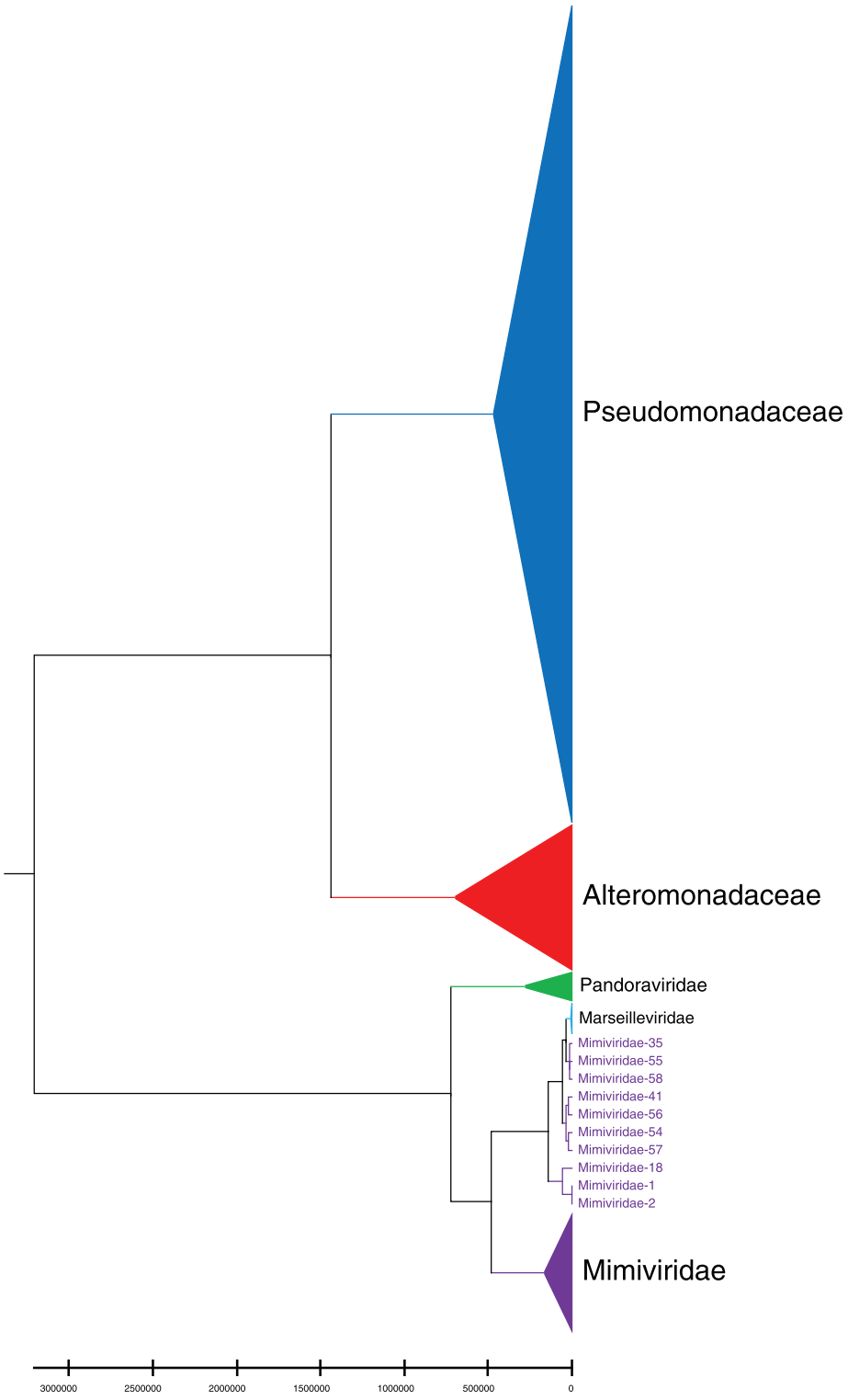

Figure 5 Phylogenetic tree for the three representative families (Mimiviridae, Pandoraviridae, and Marseilleviridae) of giant virus and two bacterial families (Pseudomonadaceae and Alteromonadaceae). Different colors represent different families.

Full-size DOI: 10.7717/peerj.13544/fig-5

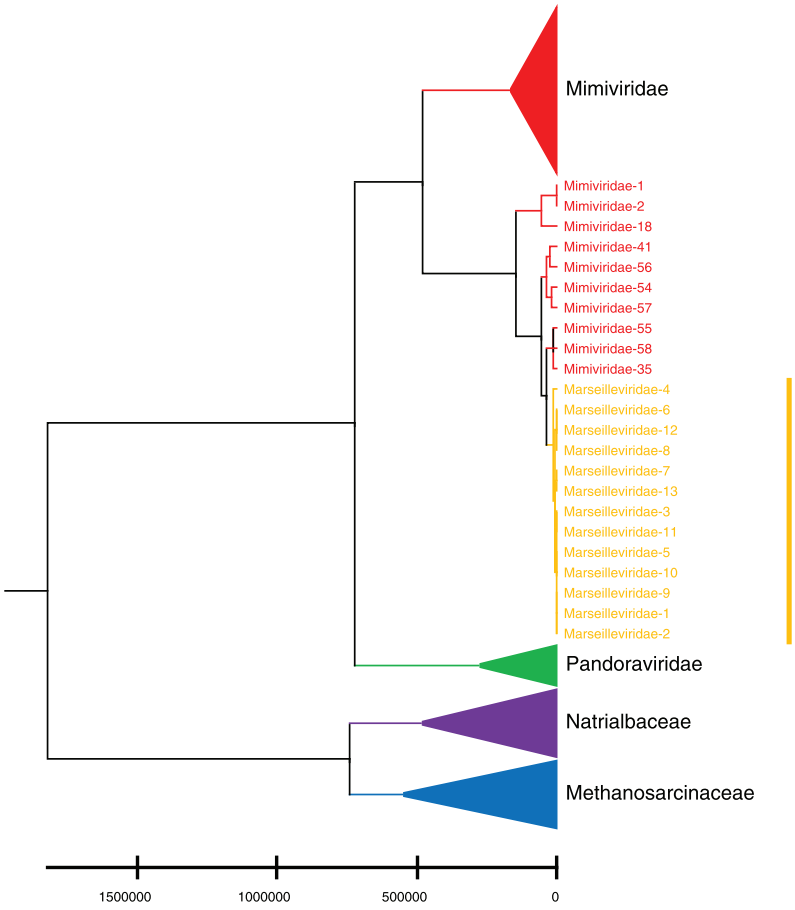

Figure 6 Phylogenetic tree for the three representative families (Mimiviridae, Pandoraviridae, and Marseilleviridae) of giant virus and two archaea families (Methanosarcinaceae and Natrialbaceae). Different colors represent different families. Full-size DOI: 10.7717/peerj.13544/fig-6

To demonstrate the importance of natural vector with six-dimensional covariance component in characterizing biological sequence, we compare it with the traditional 12-dimensional natural vector and another alignment-free method, position natural vector (*He et al., 2020*) on the same three small datasets. The phylogenetic trees based on 12-dimensional natural vector are shown in Figs. S1 to S3. In Fig. S1, the sequences of Adenoviridae colored navy blue are not clustered together. In Fig. S2, there are two extra sequences from Mimiviridae clustered with Marseilleviridae. In Fig. S3, the sequences of Methanosarcinaceae are not grouped together. Additionally, the results of position natural vector method are shown in Figs. S4 to S6. In Fig. S4, a sequence from Marseilleviridae is incorrectly clustered with Mimiviridae. Figure S5 shows an unstable result compared with Fig. 5. In Fig. S6, the sequences of three families Methanosarcinaceae, Pandoraviridae, and Natrialbaceae are mixed. In conclusion, the tree based on our new 18-dimensional natural vector gives more reasonable and interpretable results.

Haursdorff distance can reflect the genetic distance between groups, and we want to analyze the interspecific differences of the above 10 families. Mathematically, Haursdorff

distance is used to calculate the distance between point sets. If X and Y are two non-empty point sets, then the Haursdorff distance is defined as:

$$d(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}. \tag{9}$$

The distance calculation can be implemented through the code of the MathWorks community: https://ww2.mathworks.cn/matlabcentral/fileexchange/26738-hausdorff-distance. Each sequence is transformed into an 18-dimensional natural vector, then Haursdorff distance between groups is calculated, and UPGMA algorithm based on distance matrix is used to construct the tree, as shown in Fig. S7. Mimiviridae is marked with circles, and Marsellieviridae with triangles. Mimiviridae is closer to the root of the tree than Marsellieviridae.

Time and phylogeny comparison with previous covariance method

The previous study has shown that nucleotide covariance is related to the phylogenetics of fungi (*Zhao, Tian & Yau, 2018*), we redefine the nucleotide covariance to make it simpler and more general.

The previous covariance definition is:

$$Cov(k,l) = \frac{Cov(A,B)}{N}, \tag{10}$$

$k,l \in \{A,C,G,\text{T or U}\}$. $A = \{a_1,a_2,\ldots,a_n\}, B = \{b_1,b_2,\ldots,b_m\}$ are the position of k and l, respectively.

- If $m = n$, $Cov(A,B) = \sum_{i=1}^{n} \frac{(a_i - \sum_{i=1}^{n}\frac{a_i}{n})(b_i - \sum_{i=1}^{m}\frac{b_i}{m})}{n}$.
- If $m \neq n$ (assume $m > n$), the covariance between A and any n values in B is computed and the average of these C_m^n results is $Cov(A,B)$.

For a bacterial sequence (Accession number in Genbank is AP018515, sequence length is 3,041,114 bp), the number of nucleotide T is m = 723,764 and the number of nucleotide C is n = 799,865, the covariance of T and C using the previous definition is:

$$Cov(T,C) = \frac{Cov(A,B)}{3041114}, \tag{11}$$

Suppose the positions of T and C are $A = \{a_1,a_2,\ldots,a_{723764}\}, B = \{b_1,b_2,\ldots,b_{799865}\}$, respectively, then the covariance between A and any 723,764 values in B is computed and the average of these C_{799865}^{723764} results is:

$$\begin{aligned} Cov(A,B) = &\frac{1}{C_{799865}^{723764}} \sum_{\{j_1,\, j_2,\, \ldots j_{723764}\} \in B = \{b_1,\, b_2,\, \ldots, b_{799865}\}} \\ &\sum_{i=1}^{723764} \frac{(a_i - \sum_{i=1}^{723764}\frac{a_i}{n})(b_i - \sum_{i=1}^{723764}\frac{b_{j_i}}{m})}{723764}, \end{aligned} \tag{12}$$

If the number of two nucleotides is quite different, the above formula (12) is too complicated and difficult to compute, which will take a lot of time to calculate. Our new

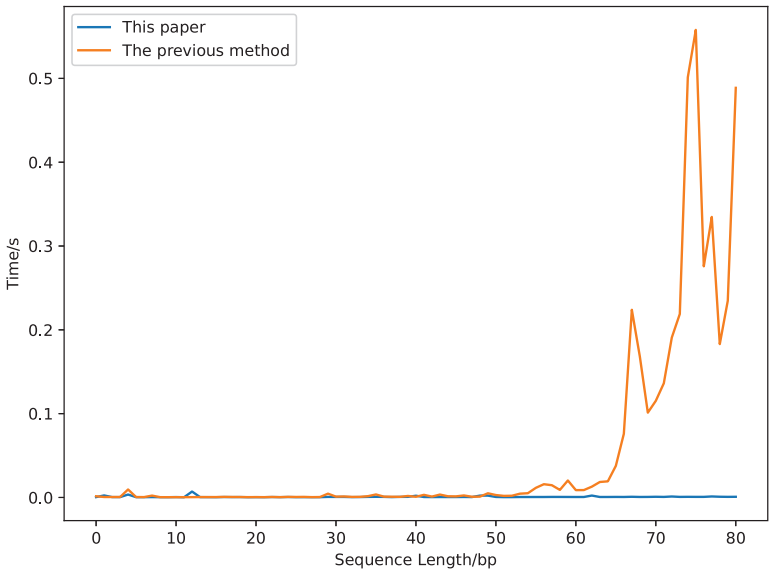

Figure 7 The calculation time comparison of 18-dimensional natural vectors under the covariance definitions of this article and the previous study. Full-size DOI: 10.7717/peerj.13544/fig-7

nucleotide covariance definition is $Cov(k,l) = \sum_{i=1}^{N} \frac{[i-\mu_k][i-\mu_l]w_{kl}(s_i)}{N\sqrt{n_k}\sqrt{n_l}}$, there is no need to compute C_{799865}^{723764} values, so the calculation of covariance takes less time.

We compare the running time of the two definitions using a sequence segment with a length of 90 bp (gaggaagtaaaagtcgtaacaaggtttccttccgggtgtagcacctgccgaagcctcccgc agcgactctaaagaaactgcgcagtctgc, a segment of a sequence whose accession number is DQ525472). The results are displayed in Fig. 7. The calculation time increases greatly with the increase of the sequence length for the previous covariance method, but is robust for our new method. It costs about 0.5 s to compute the previous covariance vector of a sequence with length 75 bp, which is much longer than that of our method (Fig. 7). While it only costs 0.017 s to compute our covariance vector of a sequence with length 813 bp (GenBank accession number is DQ525472).

The previous covariance method is tested on the fungi DNA barcode dataset, so we also apply our method on the same dataset. To show that our method can also be used for phylogenetic analysis, we choose 311 nucleotide sequences of DNA barcode from nine species of fungi (*Monosporascus cannonballus, Peniophorella praetermissa, Biatora helvola, Ceratobasidium cereale, Rhodotorula glutinis, Geosmithia landonii, Fusarium cf. solani, Acarospora smaragdula, Acaulospors kentinensis*) to construct phylogenetic tree, as shown in Fig. S8. The nine groups can be separated based on our method, while a sequence from *Peniophorella praetermissa* doesn't cluster together with *Peniophorella praetermissa* based on the previous covariance method.

DISCUSSION

As a phylogenetic researcher with mathematical background, we propose a new alignment-free method to compare sequences from the perspective of statistics, which overcomes the shortcomings of high demand for computer experimental setup of the

traditional alignment method. The new 18-dimensional natural vector method improves the 12-dimensional natural vector, and successfully numerically characterizes a large number of microbial genome data. In this way, each sequence corresponds to a point in 18-dimensional Euclidean space. The sequence similarity can be compared quickly and accurately. The traditional 12-dimensional natural vector only considers the distribution of a single nucleotide, but ignores the relationship between nucleotides. The new 18-dimensional natural vector method takes all these features into account, and contains the counts, average position and central moment of single nucleotide as well as the covariance between nucleotides. The rationality of the method is verified by testing on five datasets, archaea, bacteria, virus, fungi (genome), fungi (DNA barcode) using the convex hull classification. The results show that the six-dimensional covariance vector is a necessary condition to characterize the sequence. A further verification is performed on a special dataset, giant virus. The results show that the new 18-dimensional vector plays an important role in classification and evolution.

Many details of our article are worth discussing. First, the previous study has shown that the covariance between nucleotides is related to the phylogeny of fungi (*Zhao, Tian & Yau, 2018*). However, the limitation of the old covariance definition makes it difficult to calculate in a reasonable time if the number of two nucleotides in a sequence is quite different. Therefore, we propose a new and more natural covariance definition, which can be computed quickly. The new proposed method has obvious advantages in processing a large number of sequences and detecting the similarity and difference between sequences. Second, our proposed method has important practical significance in sequence alignment. For some sequence similarity search tools (For example, BLAST, MUSCLE), searching and aligning sequences will be very time-consuming (Figs. S9 and S10). The new method overcomes these shortcomings. It can not only naturally and effectively describe the distribution of four nucleotides, but also need less memory to store the 18-dimensional numerical vectors. Our method is promising for studying sequence alignment problems with large-scale data. Third, the microbial nucleotide sequences are utilized to verify the rationality of our new covariance definition. The data is abundant and easy to process, which is convenient for other researchers to repeat the experiment. The method can also be used for sequence comparison of other organisms, such as plants, vertebrates and invertebrates. It is beyond the scope of this article, we will study it in future work. Fourth, we use the convex hull classification results to show the necessity of six-dimensional covariance vectors in characterizing sequence. The results are reliable because the convex hull principle has strict theoretical supports: optimization background (*Zhao, Tian & Yau, 2018*), protein classification (*Zhao et al., 2019*) and the geometry construction of virus genome space (*Sun et al., 2021*). Fifth, we only use 18-dimensional natural vector to study the relationship of giant virus and bacteria, archaea, other viruses. Further studies can also be explored, for example, the geometric construction of genome space, evolution analysis. Sixth, the covariance definition of nucleotide can be extended to protein sequences. There are 20 amino acids: A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V. The natural vector with the covariance component of a protein sequence consists of four kinds of features including the counts (n_k), the average positions (μ_k) and the central

moment of position (D_2^k) of the 20 amino acids as well as the covariance between different amino acids ($Cov(k,l)$, k or l $\in$ {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}):

$[n_A, n_R, n_N, n_D, n_C, n_Q, n_E, n_G, n_H, n_I, n_L, n_K, n_M, n_F, n_P, n_S, n_T, n_W, n_Y, n_V,$

$\mu_A, \mu_R, \mu_N, \mu_D, \mu_C, \mu_Q, \mu_E, \mu_G, \mu_H, \mu_I, \mu_L, \mu_K, \mu_M, \mu_F, \mu_P, \mu_S, \mu_T, \mu_W, \mu_Y, \mu_V,$

$D_2^A, D_2^R, D_2^N, D_2^D, D_2^C, D_2^Q, D_2^E, D_2^G, D_2^H, D_2^I, D_2^L, D_2^K, D_2^M, D_2^F, D_2^P, D_2^S, D_2^T, D_2^W, D_2^Y, D_2^V,$

$Cov(A,R), Cov(A,N), Cov(A,D), Cov(A,C), Cov(A,Q), Cov(A,E), Cov(A,G), Cov(A,H), Cov(A,I), Cov(A,L),$

$Cov(A,K), Cov(A,M), Cov(A,F), Cov(A,P), Cov(A,S), Cov(A,T), Cov(A,W), Cov(A,Y), Cov(A,V), \ldots Cov(Y,V)]$.

So, the natural vector for a protein sequence is 250-dimensional, of which the covariance component is 190-dimensional. Seventh, we emphasize that dealing with the massive data sets and viewing biological problems from a mathematical perspective will lead to a deeper and more rapid understanding of their nature than relying solely on expensive experimentation. Because mathematicians may see some profound and unexpected structures and invent new mathematical methods to understand the high-dimensional properties and complex dynamics of biological problems (*Zhao, Pei & Yau, 2020*). Therefore, we propose a new mathematical numerical representation to describe the nucleotide distribution of sequences. Compared with direct (linguistic) method that avoid text conversion, for example, k-mer method, our method has two advantages: (1) It doesn't depend on any assumption because there is no need to determine k. For a genome sequence, k-mer is a sequence segment of length k (*Dai, Yang & Wang, 2008*; *Leimeister & Morgenstern, 2014*). For each given k, the number of k-mer is fixed: 1-mers indicate A, C, G, T. Our method only considers the distribution of 1-mers; (2) It is more statistically significant than k-mers because we consider three features of 1-mers, even we can extend the definition by combining 18-dimensional vector and k-mers: k-mer natural vector. The distribution of 2-mers (2-mers include AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT) can be described as:

$[n_{AA}, n_{AC}, n_{AG}, n_{AT}, n_{CA}, n_{CC}, n_{CG}, n_{CT}, n_{GA}, n_{GC}, n_{GG}, n_{GT}, n_{TA}, n_{TC}, n_{TG}, n_{TT},$

$\mu_{AA}, \mu_{AC}, \mu_{AG}, \mu_{AT}, \mu_{CA}, \mu_{CC}, \mu_{CG}, \mu_{CT}, \mu_{GA}, \mu_{GC}, \mu_{GG}, \mu_{GT}, \mu_{TA}, \mu_{TC}, \mu_{TG}, \mu_{TT},$

$D_2^{AA}, D_2^{AC}, D_2^{AG}, D_2^{AT}, D_2^{CA}, D_2^{CC}, D_2^{CG}, D_2^{CT}, D_2^{GA}, D_2^{GC}, D_2^{GG}, D_2^{GT}, D_2^{TA}, D_2^{TC}, D_2^{TG}, D_2^{TT},$

$Cov(AA,AC), Cov(AA,AG), Cov(AA,AT), Cov(AA,CA), Cov(AA,CC), \ldots, Cov(TG,TT)]$.

Therefore, our method gives a new mathematical framework to describe the distributions of k-mers, and it may be of great significance for the application of efficient and large-scale sequence alignment in the future.

CONCLUSIONS

In this article, a new six-dimensional covariance vector is proposed to reflect the correlation between nucleotides, which improves the traditional 12-dimensional natural vector. The new 18-dimensional natural vector is tested on six datasets, including five genome sequence datasets (archaea, bacteria, virus, fungi and giant virus) and one gene sequence dataset (fungi). First of all, we perform the convex hull classification. The results

show that the classification performance of 18-dimensional natural vector is better than that of 12-dimensional vector, which verifies the necessity of 6-dimensional covariance vector in characterizing biological sequences. Furthermore, we study a special virus, giant virus. Statistical analysis gives us a preliminary understanding of the representative families of giant virus, Mimiviridae, Pandoraviridae and Marsellieviridae: Pandoravirus has the largest genome size and the most of the G+C content, Mimiviridae and Pandoraviridae have more coding sequences than other families. Then we analyze the relationship between giant viruses and bacteria, archaea, other viruses. Except for Polydnaviridae and Myoviridae, the nucleotide distribution of most sequences is similar to that of bacteria. And the phylogenetic trees show a stable result, that is, 10 sequences of Mimiviridae cluster with Marseilleviridae, and Mimiviridae is closer to the root of the tree than Marsellieviridae. While another alignment-free method, position natural vector, and 12-dimensional natural vector applied to the same dataset do not show stable results. Finally, we compare the time and phylogeny with the previous covariance definition. Our method needs less computation and memory, but gets accurate results. Our 18-dimensional natural vector is a powerful alignment-free method to characterize biological sequences.

ACKNOWLEDGEMENTS

We are grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done. We thank the researchers who sequenced and shared the nucleotide sequences in NCBI. We thank the reviewers for their insightful suggestions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is supported by the National Natural Science Foundation of China (NSFC) grant (12171275), the Tsinghua University Spring Breeze Fund (2020Z99CFY044), the Tsinghua University start-up fund, and the Tsinghua University Education Foundation fund (042202008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China (NSFC) Grant: 12171275.
Tsinghua University Spring Breeze Fund: 2020Z99CFY044.
Tsinghua University start-up fund.
Tsinghua University Education Foundation fund: 042202008.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Nan Sun conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Xin Zhao performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Stephen S.-T. Yau conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data are provided in the Supplemental Files.

The Archaea (ftp.ncbi.nih.gov/refseq/release/archaea), Bacteria (ftp.ncbi.nih.gov/refseq/release/bacteria) and Fungi (ftp.ncbi.nih.gov/refseq/release/fungi) datasets are available at NCBI Reference Sequence Database: https://www.ncbi.nlm.nih.gov/refseq/.

The Virus Genomes dataset (ftp.ncbi.nlm.nih.gov/genomes/Viruses) is available at NCBI Genome: https://www.ncbi.nlm.nih.gov/genome.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.13544#supplemental-information.

REFERENCES

Almeida JS. 2014. Sequence analysis by iterated maps, a review. *Briefings in Bioinformatics* **15(3)**:369–375 DOI 10.1093/bib/bbt072.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403–410 DOI 10.1016/S0022-2836(05)80360-2.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25(17)**:3389–3402 DOI 10.1093/nar/25.17.3389.

Bichell R. 2017. In giant virus genes, hints about their mysterious origin. All Things Considered. *Available at https://www.npr.org/sections/health-shots/2017/04/06/522478901/in-giant-virus-genes-hints-about-their-mysterious-origin*.

Birtles R, Rowbotham TJ, Storey C. 1997. Chlamydia-like obligate parasite of free-living Amoebae. *The Lancet* **349(9056)**:925–926 DOI 10.1016/S0140-6736(05)62701-8.

Brandes N, Linial M. 2019. Giant viruses-big surprises. *Viruses* **11(5)**:404 DOI 10.3390/v11050404.

Claverie J, Abergel C. 2013. Open questions about giant viruses. *Advances in Virus Research* **85**:25–56 DOI 10.1016/B978-0-12-408116-1.00002-1.

Conrad LS, Keith AS, Sabine H. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the National Academy of Sciences of the United States of America* **109(16)**:6241–6246 DOI 10.1073/pnas.1117018109.

Dai Q, Yang YC, Wang TM. 2008. Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison. *Bioinformatics* **24(20)**:2296–2302 DOI 10.1093/bioinformatics/btn436.

Deng M, Yu C, Liang Q, He RL, Yau SST. 2011. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLOS ONE* **6(3)**:e17293 DOI 10.1371/journal.pone.0017293.

Didier R, Stéphane A, Catherine R. 2004. The 1.2-Megabase genome sequence of Mimivirus. *Science* **306(5700)**:1344–1350 DOI 10.1126/science.1101485.

Edgar RC. 2004a. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32(5)**:1792–1797 DOI 10.1093/nar/gkh340.

Edgar RC. 2004b. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5(1)**:113 DOI 10.1186/1471-2105-5-113.

Hatje K, Kollmar M. 2012. A phylogenetic analysis of the Brassicales clade based on an alignment-free sequence comparison method. *Frontiers in Plant Science* **3**:192 DOI 10.3389/fpls.2012.00192.

He L, Dong R, He RL, Yau SST. 2020. Positional correlation natural vector: a novel method for genome comparison. *International Journal of Molecular Sciences* **21(11)**:3859 DOI 10.3390/ijms21113859.

iBOL. 2022. What is DNA barcoding? *Available at https://ibol.org/about/dna-barcoding/* (accessed 05 May 2022).

Jeffrey HJ. 1990. Chaos game representation of gene structure. *Nucleic Acids Research* **18(8)**:2163–2170 DOI 10.1093/nar/18.8.2163.

Larkin MA, Blackshields G, Brown NP. 2007. ClustalW and ClustalX version 2.0. *Bioinformatics* **23(21)**:2947–2948 DOI 10.1093/bioinformatics/btm404.

Legendre M, Arslan D, Abergel C. 2012. Genomics of Megavirus and the elusive fourth domain of life. *Communicative and Integrative Biology* **5(1)**:102–106 DOI 10.4161/cib.18624.

Leimeister C, Morgenstern B. 2014. Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* **30(14)**:2000–2008 DOI 10.1093/bioinformatics/btu331.

Naturvetenskapliga F. 2010. Current state and perspectives of fungal DNA barcoding and rapid identification procedures. *Applied Microbiology and Biotechnology* **87(1)**:99–108 DOI 10.1007/s00253-010-2585-4.

Ogata H, Toyoda K, Tomaru Y. 2009. Remarkable sequence similarity between the dinoflagellate-infecting marine virus and the terrestrial pathogen African swine fever virus. *Virology Journal* **6(1)**:178 DOI 10.1186/1743-422X-6-178.

Pei SJ, Dong R, Bao YM, He RL, Yau SST. 2020. Classification of genomic components and prediction of genes of Begomovirus based on subsequence natural vector and support vector machine. *PeerJ* **8**:e9625 DOI 10.7717/peerj.9625.

Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, Lescot M, Arslan D, Seltzer V, Bertaux L, Bruley C, Garin J, Claverie J-M, Abergel C. 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 mb reaching that of parasitic eukaryotes. *Science* **341(6143)**:281–286 DOI 10.1126/science.1239181.

Sun N, Pei SJ, He L, Yin CC, He RL, Yau SST. 2021. Geometric construction of viral genome space and its applications. *Computational and Structural Biotechnology Journal* **19**:4226–4234 DOI 10.1016/j.csbj.2021.07.028.

Van Etten J. 2011. Giant viruses. *American Scientist* **99(4)**:304–311 DOI 10.1511/2011.91.304.

Vinga S. 2014. Information theory applications for biological sequence analysis. *Briefings in Bioinformatics* **15(3)**:376–389 DOI 10.1093/bib/bbt068.

Wessner DR, Dupont C, Charles T. 2013. *Microbiology*. Hoboken: Wiley.

Yin CC, Chen Y, Yau SST. 2014. A measure of DNA sequence similarity by Fourier transform with applications on hierarchical clustering. *Journal of Theoretical Biology* **359(14)**:18–28 DOI 10.1016/j.jtbi.2014.05.043.

Zhao RZ, Pei SJ, Yau SST. 2020. New genome sequence detection via natural vector convex hull method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19(3)**:1782–1793 DOI 10.1109/TCBB.2020.3040706.

Zhao X, Tian K, He RL, Yau SST. 2019. Convex hull principle for classification and phylogeny of eukaryotic proteins. *Genomic* **111(6)**:1777–1784 DOI 10.1016/j.ygeno.2018.11.033.

Zhao X, Tian K, Yau SST. 2018. A new efficient method for analyzing fungi species using correlations between nucleotides. *BMC Evolutionary Biology* **18(1)**:200 DOI 10.1186/s12862-018-1330-y.