Computational and Structural Biotechnology Journal 19 (2021) 4226–4234

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Geometric construction of viral genome space and its applications

Nan Sun [a,1], Shaojun Pei [a,1], Lily He [b], Changchuan Yin [c], Rong Lucy He [d], Stephen S.-T. Yau [a,e,]*

[a] *Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China*
[b] *Department of Mathematics, School of Science, Beijing University of Civil Engineering and Architecture, Beijing, PR China*
[c] *Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60628, USA*
[d] *Department of Biological Sciences, Chicago State University, Chicago, IL 60628, USA*
[e] *Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China*

ARTICLE INFO

Article history:
Received 25 May 2021
Received in revised form 24 July 2021
Accepted 24 July 2021
Available online 27 July 2021

Keywords:
Genome space
Geometry
Natural metric
Virus

ABSTRACT

Understanding the relationships between genomic sequences is essential to the classification and characterization of living beings. The classes and characteristics of an organism can be identified in the corresponding genome space. In the genome space, the natural metric is important to describe the distribution of genomes. Therefore, the similarity of two biological sequences can be measured. Here, we report that all of the viral genomes are in 32-dimensional Euclidean space, in which the natural metric is the weighted summation of Euclidean distance of k-mer natural vectors. The classification of viral genomes in the constructed genome space further proves the convex hull principle of taxonomy, which states that convex hulls of different families are mutually disjoint. This study provides a novel geometric perspective to describe the genome sequences.

1. Introduction

A genome space consists of all known genomes and provides insights into their relationships, reflecting the important nature of the genomic universe [1]. Mathematically, the genome space can be considered to be the moduli space and constructed as a subspace in a high-dimensional Euclidean space. In this space, a genome sequence is uniquely represented as a point, yet how sequences are arranged in the genome space is unknown. Another difficult task is to find a proper natural metric for describing the geometry of the genome space. The metric should reflect the structural and functional proximity of biological sequences [1]. It is essential for measuring the nucleotide distribution and inferring similar properties among genomic sequences. Briefly, the genome space with a proper metric is a powerful means of determining the phylogenetics and classification of genomes.

The methods to analyze biological sequence similarity can be alignment-based or alignment-free. Traditional alignment-based methods are inefficient at handling massive amounts of sequence because of the computational complexity and memory. However, alignment-free methods can overcome these limitations, such as traditional Natural Vector [2], k-mer theory [3], power spectrum [4], and density-based method [5]. Notably, the traditional Natural Vector, a probabilistic approach, illustrates the 12-dimensional nucleotide distributions, including the counts, mean locations, and normalized central moments of each nucleotide. The Natural Vector method and its extended versions have been applied to many studies and achieve high accuracy in sequence classification and phylogeny [6–8]. Here we apply the Natural Vector method with high order central moments to construct the genome space and combine k-mer theory and Natural Vector to define the new metric.

Each genome sequence is transformed into a natural vector in the genome space and corresponds to a point. The key characters of the genome space are the spatial patterns of the sequence points. The protein space based on the Natural Vector method has been proposed [9]. In the 250-dimensional protein space, the convex hulls corresponding to different families are disjoint. Therefore, the convex hull principle of taxonomy by protein sequences is devised [10], and the protein sequence arrangement in the protein space has been unfolded. However, the scarcity of studies on genomic space prompts us to develop a similar approach to infer the genome space by the similarity and diversity of sequences. Genomes contain all genes that specify the morphological and physiological characteristics of organisms [11–14], and sequences from the same family have similar nucleotide distribution. The convex hull principle for genome states that the points

* Corresponding author at: Department of Mathematical Sciences, Tsinghua University, Beijing 100084, PR China.

E-mail addresses: yau@uic.edu, yau@tsinghua.edu.cn (S.S.-T. Yau).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2021.07.028
2001-0370/

of one family are located in different spatial regions from points belonging to other families. In other words, the convex hull formed from natural vectors from the same family does not intersect with the convex hulls formed from natural vectors from other families. This fact inspired us to calculate the dimension of natural vectors when the convex hulls from different families for genomic sequences are mutually disjoint. Then, the genome space exists, and the subspace of the Euclidean space under this dimension is the genome space.

A virus is small in size and simple in structure, with only one kind of nucleic acid (DNA or RNA). We downloaded all reference viral genomes in NCBI to construct the genome space. The reference genomes are of high quality and reliable for genome space construction. We find that the viral genome space is located in a 32-dimensional Euclidean space, which means that the convex hull principle for vial genomes holds in a 32-dimensional space. This study shows that the Euclidean distance of the natural vectors cannot reflect the biological similarity of genome sequences according to the results of the nearest neighborhood classification. Under multiple attempts for the metric definitions, we propose a new natural metric that contains the differences in the genome distributions of 1-mer to n-mer [15–17]. We define the metric as the weighted summation of Euclidean distance of k-mer Natural Vectors. The uncertainty of k gives the space to adjust the weights and improve the classification accuracy using the metric definition. The classification and phylogenetic results of virus families demonstrate the performance of the metric definition. The construction of genome space with the novel natural metric makes it possible to characterize the huge genome universe and solve the fundamental problems of genome sequences.

2. Materials and methods

2.1. Virus genomic sequences dataset and the statistic information

There are 9603 viral reference sequences in NCBI (National Center for Biotechnology Information) up to March 2020 (ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses). We download all sequences and update our lab database VirusDB [18]. In this study, we remove three types of sequences: (1) viruses without Baltimore class label; (2) viruses without family label; and (3) families including one or two sequences. And 7382 sequences are retained, which belong to 83 families, 304 genera, and 7 Baltimore classes (dsDNA, ssDNA, dsRNA, (+) ssRNA, (−) ssRNA, ssRNA-RT and dsDNA-RT). The sequence statistical information is shown in Fig. A.1. Baltimore class I contains the most sequences, as well as the most families and genera, the average sequences length is also the longest. Baltimore class VI only has 1 family and Baltimore class VII has 2 families, the two classes only account for 2% of the total number of reference sequences. It is worth noting that some viruses from Baltimore class I ~ V have multiple segment genomic sequences. The detailed accession numbers are shown in Data A.1, and families and genera information are shown in Data A.2 and A.3.

2.2. Natural vector with high order central moments

Let $S = s_1s_2s_3\cdots s_n$ be a genomic sequence of length n, and $L = \{A, C, G, T/U\}$. For $k \in L$, we define the indicator functions: $w_k(\cdot) : L \rightarrow \{0, 1\}$, i.e.:

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

Where $s_i \in L, i = 1, 2, 3, \cdots, n$.

- Let $n_k = \sum_{i=1}^{n} w_k(s_i)$ denote the counts of nucleotide k in S.
- Let $\mu_k = \sum_{i=1}^{n} i\frac{w_k(s_i)}{n_k}$ specify the average location of letter k.
- Let $D_j^k = \sum_{i=1}^{n} \frac{(i-\mu_k)^j w_k(s_i)}{n_k^{j-1} n^{j-1}}$ be the j-th central moment of position of letter k.

Then we can get (8 + 4n)-dimensional Natural Vector: $\left(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T, \cdots, D_{n+1}^A, D_{n+1}^C, D_{n+1}^G, D_{n+1}^T, \cdots\right)$

Here we give an example. If the genomic sequence is ACGGTAGTCC, the indicator functions are shown in Table A.1.

The corresponding components of distribution vector are calculated as follows:

- $n_A = 2, n_C = 3, n_G = 3, n_T = 2.$
- $\mu_A = 1\cdot\frac{1}{2} + 6\cdot\frac{1}{2} = 3.5;\ \mu_C = 2\cdot\frac{1}{3} + 9\cdot\frac{1}{3} + 10\cdot\frac{1}{3} = 7;$
  $\mu_G = 3\cdot\frac{1}{3} + 4\cdot\frac{1}{3} + 7\cdot\frac{1}{3} = 4.67;\ \mu_T = 5\cdot\frac{1}{2} + 8\cdot\frac{1}{2} = 6.5.$
- $D_2^A = \frac{\left(1-\frac{7}{2}\right)^2}{2\cdot10} + \frac{\left(6-\frac{7}{2}\right)^2}{2\cdot10} = 0.63;$
- $D_2^C = \frac{(2-7)^2}{3\cdot10} + \frac{(9-7)^2}{3\cdot10} + \frac{(10-7)^2}{3\cdot10} = 1.27;$
- $D_2^G = \frac{\left(3-\frac{14}{3}\right)^2}{3\cdot10} + \frac{\left(4-\frac{14}{3}\right)^2}{3\cdot10} + \frac{\left(7-\frac{14}{3}\right)^2}{3\cdot10} = 0.29;$
- $D_2^T = \frac{\left(5-\frac{13}{2}\right)^2}{2\cdot10} + \frac{\left(8-\frac{13}{2}\right)^2}{2\cdot10} 0.23;$

Then the 12-dimensional Natural Vector is: $(2, 3, 3, 2, 3.5, 7, 4.67, 6.5, 0.63, 1.27, 0.29, 0.23)$.

2.3. k-mer Natural vector

K-mer l_i is a string of length k composed of four nucleotides. If genomic sequence is still $S = s_1s_2s_3\cdots s_n$, $s_i \in \{A, C, G, T/U\}$, $l_i[j]$ is the location of the j-th occurrence of a k-mer l_i in S, $i = 1, 2, \cdots, 4^k$. For each given k, the distributions of a k-mer l_i can be described by three quantities.

- n_{l_i} denotes the counts of k-mer l_i occurrences in S;
- μ_{l_i} specify the average location of k-mer l_i;
- $D_m^{l_i} = \sum_{m=1}^{n_{l_i}} \frac{\left(l_i[j]-\mu_{l_i}\right)^m}{n_{l_i}^{m-1}(n-k+1)^{m-1}} (m = 1, 2\cdots, n_{l_i})$ is the m-th central moment of emergence position of letter k-mer l_i

Thus, high order k-mer Natural Vector for sequence S is defined by:

$$\left(n_{l_1}, \ldots, n_{l_{4^k}}, \mu_{l_1}, \cdots, \mu_{l_{4^k}}, D_2^{l_1}, \cdots, D_2^{l_{4^k}}, \cdots, D_n^{l_1}, \cdots, D_n^{l_{4^k}}\right).$$

And its dimensional is $4^k\cdot(n+1)$.k-mer Natural Vector with second central moment has been verified to be enough to represent the sequence and satisfies one-to-one mapping, so the k-mer Natural Vector is $4^k\cdot3$ dimension:

$$\left(n_{l_1}, \ldots, n_{l_{4^k}}, \mu_{l_1}, \cdots, \mu_{l_{4^k}}, D_2^{l_1}, \cdots, D_2^{l_{4^k}}\right).$$

2.4. Convex hull principle

Convex hull is one of the most fundamental concepts in computational geometry [19]. The geometric structure is widely used in many application domains, such as image processing [20,21] and pattern recognition [22,23]. Mathematically, the convex hull of a point set $\{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_k\}, \mathrm{x}_i \in \mathrm{R}^n$ is the minimal convex set that contains these points. Note that a convex set C is the region such that straight line segment connecting any two points within C is also located in C. Any region which has hollowness, dent or extended vertices are not convex. Particularly a triangle is composed of all convex combinations of its three vertexes and a tetrahedron consists of the convex combinations of its four vertexes in three dimensions. By the concept of convex combinations, the convex hull of a finite point set C is equivalently defined as the set of all convex combinations of points in C:

$$convC = \{\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_k x_k | x_i \in C, \theta_1 + \theta_2 + \cdots + \theta_k = 1, \\ \theta_i \geq 0, i = 1, 2, \cdots, k\}.$$

One of the important properties for the convex hull is that its boundary is spanned by some points of C, called vertexes and the rest points of C are lying inside the hull. When all x_i are two dimensional vectors, the convex hull is a convex polygon. In general, the convex hull is called convex polytope in high dimensional space. We use the *convhull* function incorporated in MATLAB to find the convex hull of a finite point set.

In this study, x_i is the natural vector and we propose a convex hull principle of molecular biology for viruses, pointing out that convex hulls corresponding to different virus families or genera do not overlap with each other. For those viruses with a single segment sequence, we directly calculate the natural vectors and then establish a convex hull. For those viruses with multi-segment sequences, we first calculate the natural vector of each segment of the virus and establish a small convex hull for these segment sequences, and then build a large convex hull with the remained viruses of the family to which the virus belongs. In this way each family corresponds to a point cloud, which reflects the genetic variety of this family.

2.5. Linear programming method

Determining the separateness of two convex polyhedrons is a significant problem. Most of the popular methods are capable in low dimensional space [24]. While these approaches fail to work if the dimension is high. Calculating the distance between two convex polytopes is an efficient way to judge whether two convex hulls intersect, which can be implemented by quadratic optimization regardless of the dimension. If A is the convex hull of point set $\{a_1, a_2, \cdots, a_m\}$ and B is the convex hull of point set $\{b_1, b_2, \cdots, b_n\}$. The method to prove the separateness between A and B is the linear programming (LP) method, it can be solved through *linprog* function in MATLAB. The mathematical principle is that if there exists non-zero coefficients $\{\lambda_1, \lambda_2, \cdots, \lambda_m, \beta_1, \beta_2, \cdots, \beta_n\}$ in feasible domain such that the optimization value of the following LP problem is 0, then A and B intersect:

$$\frac{\min \quad 0}{s.t. \sum_{i=1}^{m} \lambda_i a_i = \sum_{j=1}^{n} \beta_j b_j}$$
$$\sum_{i=1}^{m} \lambda_i = 1, \lambda_i \geq 0, i = 1, 2, \cdots, m$$
$$\sum_{j=1}^{n} \beta_j = 1, \beta_j \geq 0, j = 1, 2, \cdots, n$$

2.6. The projection method

If two convex hulls do not intersect in high dimension, the corresponding projected 2-dimensional convex hulls do not intersect either. To visualize the disjoint convex hulls, we project the high dimensional convex hulls into 2-dimensional space. We use the idea of support vector machine (SVM) and Linear Discriminate Analysis (LDA) as the projection method to achieve our goal.

SVM is a famous method to do classification [25]. The easiest situation is the linear kernel, that is to say, if two sets of points in high dimensional space are linearly separable, there exists a separating hyperplane between these two sets. Then we can take the normal vector and the vector perpendicular to it as the direction of the new coordinate axis, and project natural vectors in these two directions. Then the convex hulls of these two sets are disjoint in 2-dimensional space. The mathematical method to determine the normal vector and offset item of the hyperplane is as follows. There is a dataset $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}, x_i \in R^d, y_i \in \{+1, -1\}$ including two classes of samples. The separating hyperplane is $w^T x + b = 0, w = (w_1, w_2, \cdots, w_d)^T$ is the normal vector, b is the offset item. To find the separating hyperplane with the maximum margin, it is equivalent to solve the following convex quadratic programming problem:

$$\frac{\min\limits_{w,b} \frac{1}{2} ||w||^2}{s.t. y_i \left(w^T x_i + b\right) \geq 1, i=1,2,\cdots,m.}$$

The dual problem is easier to solve, so the dual algorithm is usually used to find the solution of the primal problem. First, Lagrange multiplier $\alpha_i (i = 1, 2, \cdots, m)$ for each constraint is introduced and the Lagrange function is defined as: $L(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^{m} \alpha_i y_i (w^T x_i + b) + \sum_{i=1}^{m} \alpha_i$. According to the Lagrange duality, the dual problem of the primal problem is maximal-minimum problem: $\max\limits_{\alpha} \min\limits_{w,b} L(w, b, \alpha)$. To find the optimal solution is equivalent to solve the following dual problem:

$$\min\limits_{\alpha} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{m} \alpha_i$$
$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \cdots, m.$$

From the above derivation steps and our dataset is discrete point sets, the KKT (Karush–Kuhn–Tucker) conditions hold, so α^* is the optimal solution of the dual problem:

$$\begin{cases} \nabla_w L(w, b, \alpha) = 0; \\ \nabla_b L(w, b, \alpha) = 0; \\ \alpha_i \left(y_i \left(w^T x_i + b\right) - 1\right) = 0; . \\ y_i \left(w^T x_i + b\right) - 1 \geq 0; \\ \alpha_i \geq 0, (\mathrm{i} = 1, 2, \cdots, \mathrm{m}); \end{cases}$$

We conclude that $w^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i$ and $b^* = y_j - \sum_{i=1}^{m} \alpha_i^* y_i (x_i \cdot x_j)$. There is a vector v^* being perpendicular to vector w. For vector $(V, y) \in D$, we can project it into 2-dimensional space, and the new coordinates are $V \cdot w^*$ and $V \cdot v^*$. Then the points in D can be separated into 2 clusters.

Above prime and dual problems are both quadratic programming, and they can be solved by *quadprog* function in build-in MATLAB or MOSEK toolbox. The size of the quadratic problem relies on sample numbers, which will be time-consuming in real operations, so there is an efficient algorithm, which can be implemented by *libsvm* toolbox [26].

Linear Discriminate Analysis (LDA) is a dimension reduction technology of supervised learning. The label of each sample in the dataset is known before, which is different from Principal Component Analysis (PCA). The high dimensional vectors are projected into low dimensional points such that the points from the same group are as close as possible, and reverse for the different group [27].

We use SVM or LDA to project the high dimensional vectors into 2-dimensional space, then the classification result can be visualized in a low dimensional space.

3. Results

3.1. Convex hull principle for genomes and viral genome space construction

All of the reference viral genome sequences in NCBI up to March 2020 were downloaded, and we excluded sequences that have no

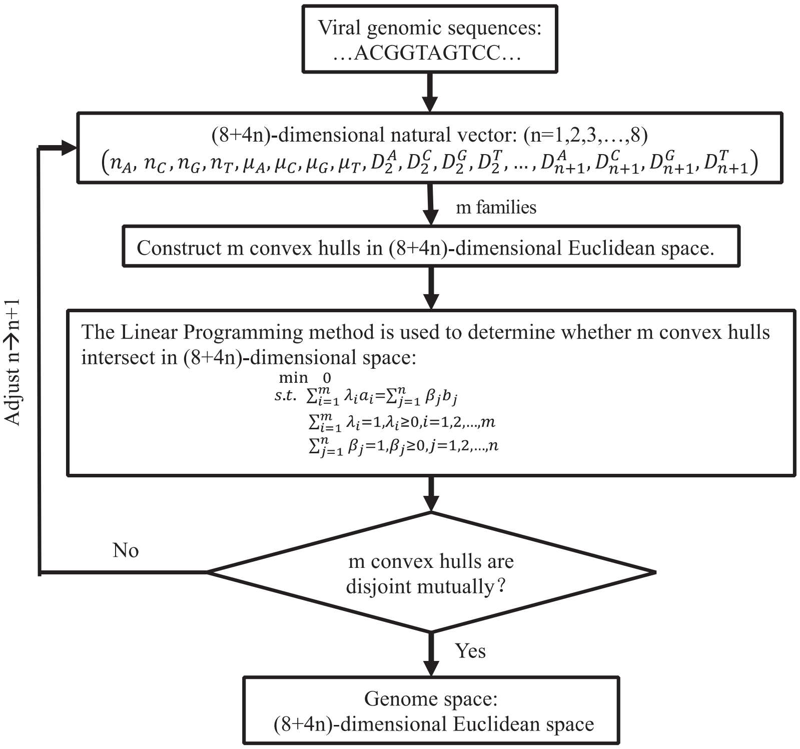
Viral genomic sequences:
…ACGGTAGTCC…
(8+4n)-dimensional natural vector: (n=1,2,3,…,8)
$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T, \dots, D_{n+1}^A, D_{n+1}^C, D_{n+1}^G, D_{n+1}^T)$
m families
Construct m convex hulls in (8+4n)-dimensional Euclidean space.
The Linear Programming method is used to determine whether m convex hulls intersect in (8+4n)-dimensional space:
min 0
$s.t.\ \sum_{i=1}^m \lambda_i a_i = \sum_{j=1}^n \beta_j b_j$
$\sum_{i=1}^m \lambda_i = 1, \lambda_i \geq 0, i=1,2,\dots,m$
$\sum_{j=1}^n \beta_j = 1, \beta_j \geq 0, j=1,2,\dots,n$
m convex hulls are disjoint mutually?
No
Adjust n→n+1
Yes
Genome space:
(8+4n)-dimensional Euclidean space

Fig. 1. The flowchart for constructing the viral genome space. The genome space is constructed based on 83 families. All convex hulls in a 32-dimensional space are mutually disjoint.

Baltimore classes or family labels. We also excluded the sequences from these families that have only one or two sequences. The dataset contains 7,382 sequences of 83 families. We used these viral sequences to construct genome space, the flowchart of constructing the genome space is illustrated in Fig. 1. Each viral genomic sequence S was mapped into a (8 + 4n)-dimensional natural vector first:

$$\left(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T, \cdots, D_{n+1}^A, D_{n+1}^C, D_{n+1}^G, D_{n+1}^T\right),$$

where n = 1, 2, 3, . . ., 8. n_k denotes the count of nucleotide k in S, μ_k specifies the average location of letter k. D_j^k is the j-th central moment of the position of the letter k, $k \in \{A, C, G, T/U\}$. The natural vectors are located in R^{8+4n}. The convex hull for each virus family in this high dimensional Euclidean space is constructed based on the (8 + 4n)-dimensional natural vectors, and there are $C_{83}^2 = 3403$ convex hull pairs. The convex hull principle of genome states that convex hulls corresponding to different families are mutually disjoint. Therefore, we checked whether all convex hull pairs intersect in $R^{12}, R^{16}, R^{20}, R^{24}, R^{28}, R^{32}, R^{36}, R^{40}$ $(n = 1, \cdots, 8)$, respectively. A simple way to determine the separation between two convex hulls is the linear programming [28], in which $\sum_{i=1}^m \lambda_i a_i = \sum_{j=1}^n \beta_j b_j$ is satisfied if the convex hull pair corresponding to two point sets $\{a_1, a_2, \cdots, a_m\}$ and $\{b_1, b_2, \cdots, b_n\}$ intersect, where $\sum_{i=1}^m \lambda_i = 1$, $\sum_{j=1}^n \beta_j = 1$. The numbers of disjoint convex hull pairs in different spaces are shown in Table 1. With the increase in the dimension of natural vectors, disjoint convex hull pairs also increase. When no convex hull intersects another one, the convex hull principle for viral genomes holds in R^{32}. Therefore, the viral genome space is located in a 32-dimensional Euclidean space. Our results suggest that viruses with a similar nucleotide distribution lie in the same convex hull, and all convex hulls show the global landscape of viruses at the family level.

To visualize this result, we propose to use support vector machine (SVM) [29] to project the 32-dimensional convex hull into 2-dimensional space. Because each convex hull pair has been confirmed not to intersect another in R^{32}, we can find a hyperplane $\omega^T x + b = 0$ to separate them. We take the normal vector ω of the hyperplane and a random vector v on the hyperplane being perpendicular to vector ω as two directions of the new axis. We then projected the natural vector Vinto the hyperplane of these two vectors v and ω. The new 2-dimensional coordinates are $V \cdot \omega$ and $V \cdot v$, respectively. Through SVM projection, the dimension of natural vectors is reduced to 2, then the convex hull based on the new 2-dimensional vectors for each family is formed. Every two convex hulls from two viral families do not overlap, and the complete results are stored in https://github.com/sunn19/Virus_Genome_Space.git.

Table 1
The number of disjoint convex hull pairs changes with the increase in the dimension of the Euclidean space. Total convex hull pairs of family are 3404. When the dimension of the natural vector is more than 32 ($n \geq 6$), there are no intersecting convex hull pairs. According to the definition of embedding dimension of the moduli space, we chose the space with the lowest dimension, which indicates that the viral genome space is sitting in a 32-dimensional Euclidean space.

Euclidean space	n = 1 R^{12}	n = 2 R^{16}	n = 3 R^{20}	n = 4 R^{24}	n = 5 R^{28}	n = 6 R^{32}	n = 7 R^{36}	n = 8 R^{40}
No. of disjoint convex hull pairs	3221	3291	3338	3354	3395	3403	3403	3403
No. of intersecting convex hull pairs	182	112	65	49	8	0	0	0

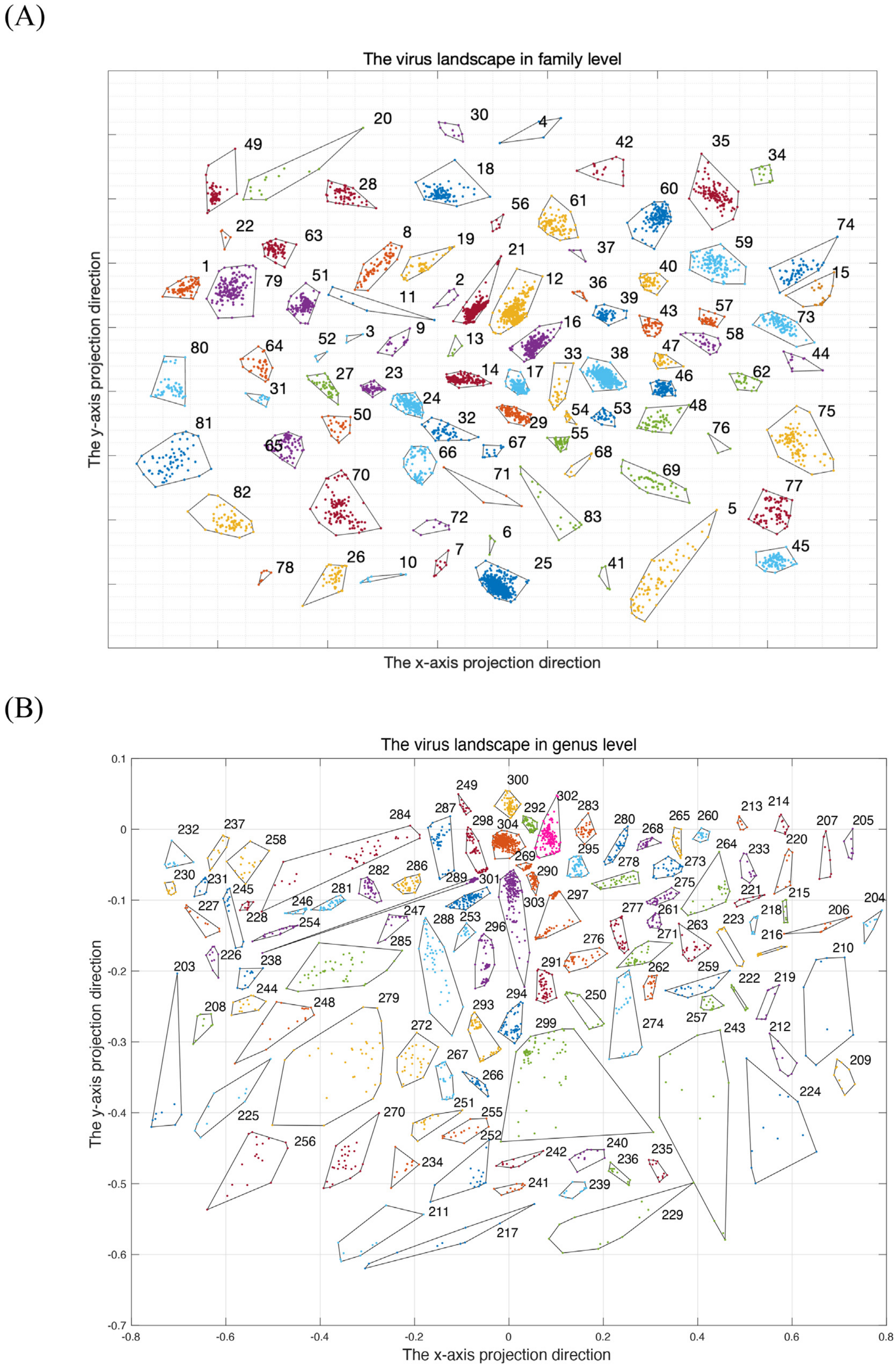

Fig. 2. Virus convex hull landscape projection in R^2. The numbers represent groups of viruses, and group name can be found in Data A.2 and A.3. The boundary of each convex hull is marked in black color.

To combine all convex hulls in one figure, we used the linear discriminant analysis (LDA) [30] method to transform the high-dimensional convex hull into a 2-dimensional space. The 2-dimensional landscape at the family level is shown in Fig. 2A. Here, we only consider the hull shape instead of size and location. The convex hulls of families for each Baltimore class are also mutually disjoint, and the results are presented in Fig. A.3.

Notably, convex hulls are also mutually disjoint at the genus level in the 32-dimensional genome space. We removed three types of sequences: the sequences of the genus that have less than two sequences, or have no genus label, or are not classified. Therefore, total 304 sequences of genera are remained. There are $C_{304}^2 = 46056$ convex hull pairs. Similarly, we built the virus landscape at the genus level. The 2-dimensional projection results are stored in https://github.com/sunn19/Virus_Genome_Space.git. We displayed the convex hulls of genera in multiple pictures and, due to the limitation of picture size, there can be an overlapping genus in different pictures. We only show part of the virus landscape in Fig. 2B. The remaining part of the landscape is exhibited in Fig. A.2 A and A.2 B. Three pictures constitute the 2-dimensional landscape of viruses at the genus level. There are 102 genera in three pictures, respectively. Genus #203 is in both Fig. 2B and A.2 A, genus #102 is in both Fig. A.2 A and A.2 B.

3.2. Novel natural metric

To show the geometry of the viral genome space, a descriptive metric on this space shall be provided. We used the nearest neighborhood (1NN) classification accuracy to determine the metric. The 1NN definition here is as follows, for a virus sequence V1, we calculated the virus sequence V2 nearest to V1, and if these two sequences have the same family label, the classification result is correct, and the accuracy equals the number of correct labels divided by the total number of labels. To find a reliable metric, we removed virus sequences containing characters other than ACGT; a total of 6916 viral reference sequences remained. Intuitively, the Euclidean metric can be put on the 32-dimensional space, but the accuracy based on natural vector was only 79.9%, which indicates that the Euclidean distance is not a proper metric for this space. This requires us to define a new metric on the space.

K-mer natural vector combines the frequency of k-mer and traditional natural vector, which reflects the distributions of strings of length k in the genome sequences. Each genome sequence can be mapped into a $4^k(n+1)$-dimensional vector:

$$\left(n_{l_1},\ldots,n_{l_{4^k}},\mu_{l_1},\cdots,\mu_{l_{4^k}},D_2^{l_1},\cdots,D_2^{l_{4^k}},\cdots,D_n^{l_1},\cdots,D_n^{l_{4^k}}\right),$$

k-mer l_i is a string of length k composed of four nucleotides. n_{l_i} denotes the counts of k-mer l_i in S, μ_{l_i} specifies the average location of k-merl_i, and $D_j^{l_i}$ is the j-th central moment of emergence position of letter k-merl_i. The correspondence between a genetic sequence and its associated k-mer natural vector is one-to-one [31]. The 1-mer natural vector is the main component representing the sequence distribution. The new natural metric based on k-mer natural vector is defined as:

$$d = d_1 + \frac{1}{2}d_2 + \cdots + \frac{1}{2^{n-1}}d_n,$$

where d_k is the Euclidean distance between k-mer natural vectors of two genomic sequences. The beauty of our new natural metric definition is that it contains the distribution differences from 1-mer to n-mer. The accuracies of virus family classification based on the new metric are shown in Table 2. When $n = 9$, the accuracy is 88.3%. We found that with the increase in n, the classification accuracy increased. We believe that the natural metric should involve all the k-mers for k≥ 1. Consequently, we conclude that, when n is large enough, the new metric can truly reflect the relationship between viral sequences.

3.3. Natural graph for a small viral dataset

To illustrate that the new metric is meaningful, we used a small dataset to draw a natural graph, which is a distance-based classification method and a direct image of the relationships between viral sequences could be obtained [32]. The dataset includes eight families with fewer than ten sequences from Baltimore class I III IV V, which are *Bicaudaviridae, Tectiviridae, Picobirnaviridae, Quadriviridae, Hepeviridae, Mesoniviridae, Filoviridae*, and *Ophioviridae*. The virus accession numbers are in Data A.4. The natural graphical representation is shown in Fig. 3. The number in the graph represents a viral sequence, and the sequences from the same family are marked in the same color. The arrow from sequence #2 to #3 indicates that #3 is the closest sequence to #2. Two-way arrow indicates the two sequences are the closest sequence to each other. The blue arrow shows the closest distance (1-level), and the red arrow shows the sub closest distance (2-level). For virus #2 (GenBank accession number: NC_029316), it is from *Bicaudaviridae* of Baltimore class I. In the natural graphical representation, it is closest to virus #1 (GenBank accession number: NC_007409), virus #1 is closest to virus #2 as well, the distance based on the k-mer natural vectors of the two viral sequences is 115349.60. Virus #2 and virus #4 are the next closest to each other, and the distance is 125452.95. There may be some viral sequences missing in our dataset, which could be located in virus #2 and virus #4; it is a challenging job to find these members. The unique natural graph gives an accurate classification result and shows the direct phylogenetic relationships between these eight families. We also constructed a natural graph based on Euclidean distance for comparison, as shown in Fig. A.4 and viral sequences from the same family are lying together, which further demonstrates the meaningfulness of our new metric definition. Our metric contains more information about the distribution difference of two sequences than Euclidean distance.

3.4. Phylogenetic analysis for each Baltimore class

As a further application of natural metric, we performed a phylogenetic analysis for each Baltimore class. The distance matrix was computed based on the new metric, and the phylogenetic tree was constructed by UPGMA algorithm [33] of MEGAX [34,35]. Fig. 4A shows the phylogenetic tree of 5 families from Baltimore class I, which consists of 399 viral sequences and they are divided into 5 clusters. The number of sequences per leaf is displayed next to its right. Fig. 4B reveals the clustering result of four families, *Circoviridae, Nanoviridae, Inoviridae* and *Parvoviridae* from Baltimore

Table 2
The nearest neighborhood classification accuracies of virus family based on the new natural metric for different n. For weight $\frac{1}{2^k}$ $(d = \sum_{k=1}^{n} \frac{1}{2^{k-1}} d_k)$, the classification is more accurate with the increase in n. For weight $\frac{1}{k^2}$ $(d = \sum_{k=1}^{n} \frac{1}{k^2} d_k)$, the accuracy decreases when n = 9, indicating that this definition is unstable. The natural metric is defined as $d = d_1 + \frac{1}{2}d_2 + \cdots + \frac{1}{2^{n-1}}d_n$.

Weight	n	1	2	3	4	5	6	7	8	9
$\frac{1}{2^k}$	Accuracy	79.9%	82.8%	83.3%	83.3%	84.1%	85.8%	86.9%	87.4%	88.3%
$\frac{1}{k^2}$	Accuracy	79.9%	82.8%	83.3%	83.3%	84.4%	86.3%	87.7%	88.0%	85.6%

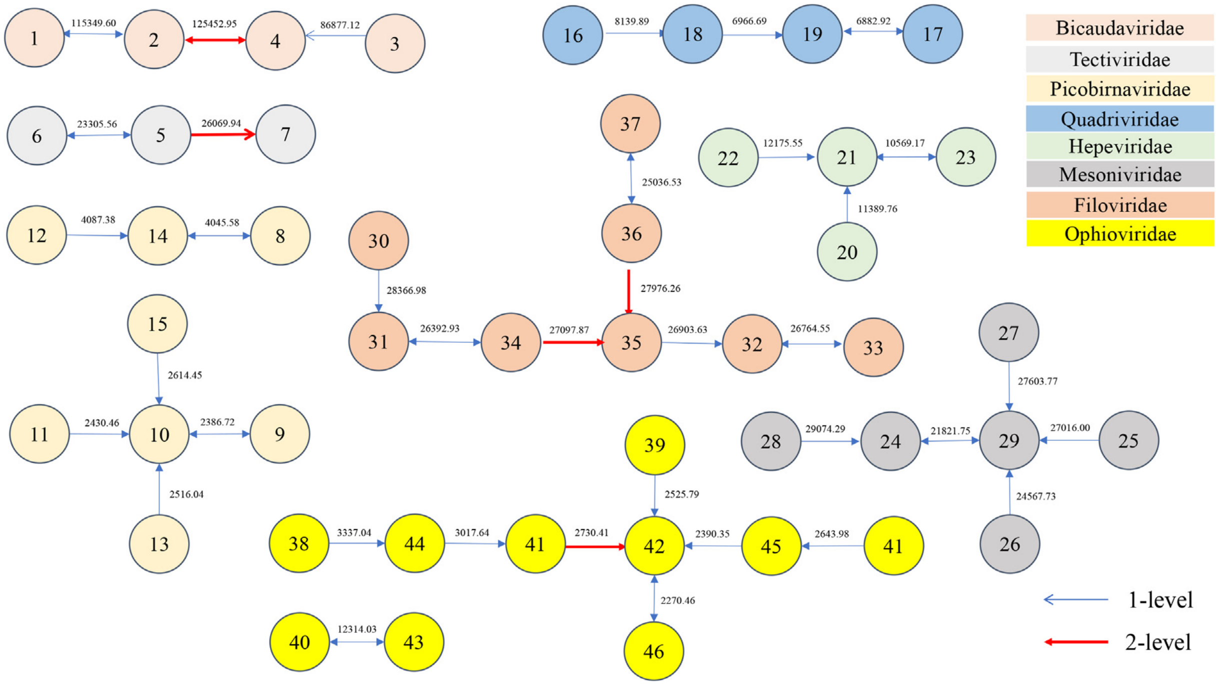

Fig. 3. Natural graph of nine families based on the new natural metric. Each node represents a viral genome. The nodes marked in the same color are from the same family. The distance between each two nodes is tagged on the arrow. The arrow from sequence #2 to #3 indicates that #3 is the closest sequence to #2. Two-way arrow indicates the two sequences are the closest sequence to each other. The blue arrow shows the closest distance (1-level), and the red arrow shows the sub closest distance (2-level). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

class II, which shows that the genomes are divided into four subgroups, and it is in agreement with the old taxonomy. The phylogenetic trees for other Baltimore classes can be found in Figs. A.5–A.8, which all gave perfect clustering results.

As a comparison, we also drew the tree based on Euclidean distance for each Baltimore class. The clustering trees are shown in Figs. A.9–A.14. Fig. A.9 is the phylogenetic tree based on Euclidean distance for Baltimore class I, where *Siphoviridae* and *Adenoviridae* cannot be separate. Viral sequences from *Inoviridae* and *Parvoviridae* are mixed in Fig. A.10. In Fig. A.11, a sequence from *Totiviridae* clusters together with *Chrysoviridae*, and two sequences from *Partitiviridae* do not cluster with the other sequences in *Partitiviridae*. In Figs. A.12–A.14, families from Baltimore class IV, V, and VII are all separate. The above results demonstrate that the clustering trees based on the new metric outperform those of the Euclidean distance method and reveal the rationality of the new metric.

4. Discussion and conclusion

We addressed two problems proposed in the comparative genomics of 23 mathematical challenges proposed by the Defense Advanced Research Projects Agency (DARPA) in 2008 [36], namely, “The Geometry of Genome Space” and “What are the Fundamental Laws of Biology?”. Through the convex hull principle, we found that the viral genome space is located in a 32-dimensional Euclidean space. In this space, we defined a novel natural metric, which is the weighted summation of the Euclidean distance. It contains the differences in the genome distributions of 1-mer to n-mer natural vectors. The new natural metric can reflect biological similarity. Many methods based on the k-mer character have been developed. However, most of them are based only on frequency, without considering the distribution of k-mers, and the ordinary k-mer methods lose a lot of information since they cannot recover the sequence. The k-mer natural vector method contains both the frequency and the distribution of k-mers, which does not lose information and produces a one-to-one correspondence between genome sequences and vectors in a finite dimensional space. It is a classical dilemma in k-mer methods to choose a proper k. For each k, we get a metric, but which only gives partial information. Thus, we weighed the distance of k-mer natural vector and calculated the nearest neighborhood accuracy to determine which metric is better. We tested other metrics, such as the Manhattan distance $\left(d(x,y)=\sum_{i=1}^{n}|y_i-x_i|\right)$, Chebyshev distance $\left(d(x,y)=\max_{1\le i\le n}|y_i-x_i|\right)$ and cosine similarity $\left(\cos(\angle(x,y))=\frac{x\cdot y}{|x||y|}=\frac{\sum_{i=1}^{n}x_i\times y_i}{\left(\sum_{i=1}^{n}x_i^2\right)^{\frac{1}{2}}\times\left(\sum_{i=1}^{n}y_i^2\right)^{\frac{1}{2}}}\right)$, and the Euclidean distance $\left(d(x,y)=\left(\sum_{i=1}^{n}(y_i-x_i)^2\right)^{\frac{1}{2}}\right)$ had the best classification performance. Moreover, we check several metrics with different weights, for example, $d=\sum_k\frac{1}{k^k}d_k$ and $d=\sum_k\frac{1}{k^2}d_k$, while the metric $\left(d=\sum_k\frac{1}{2^{k-1}}d_k\right)$ performs best on the one label classification, where d_k is the Euclidean distance of k-mer natural vector. The beauty of our new metric definition is that all k-mers are involved. Unfortunately, the limitation of our computer hardware makes it difficult to compute the k-mer natural vectors when k goes to greater than 9. The classification and phylogenetic results still imply the new metric with weight $\frac{1}{2^k}$ is very powerful.

The geometry of genome space shows that the convex hull principle is fundamental in genome analysis because the distribution of genome sequence determines its property. The underlying principle is that species close to each other have a similar distribution of nucleotides in their complete genomes. The natural vector is used to describe the distribution of nucleotides mathematically, and each genome sequence is represented as a point uniquely in high dimensional Euclidean space. Then using all these points, one can form a convex hull in this space, which is helpful to describe the similarity of the distribution among species. The convex hull principle as a fundamental law of molecular biology for

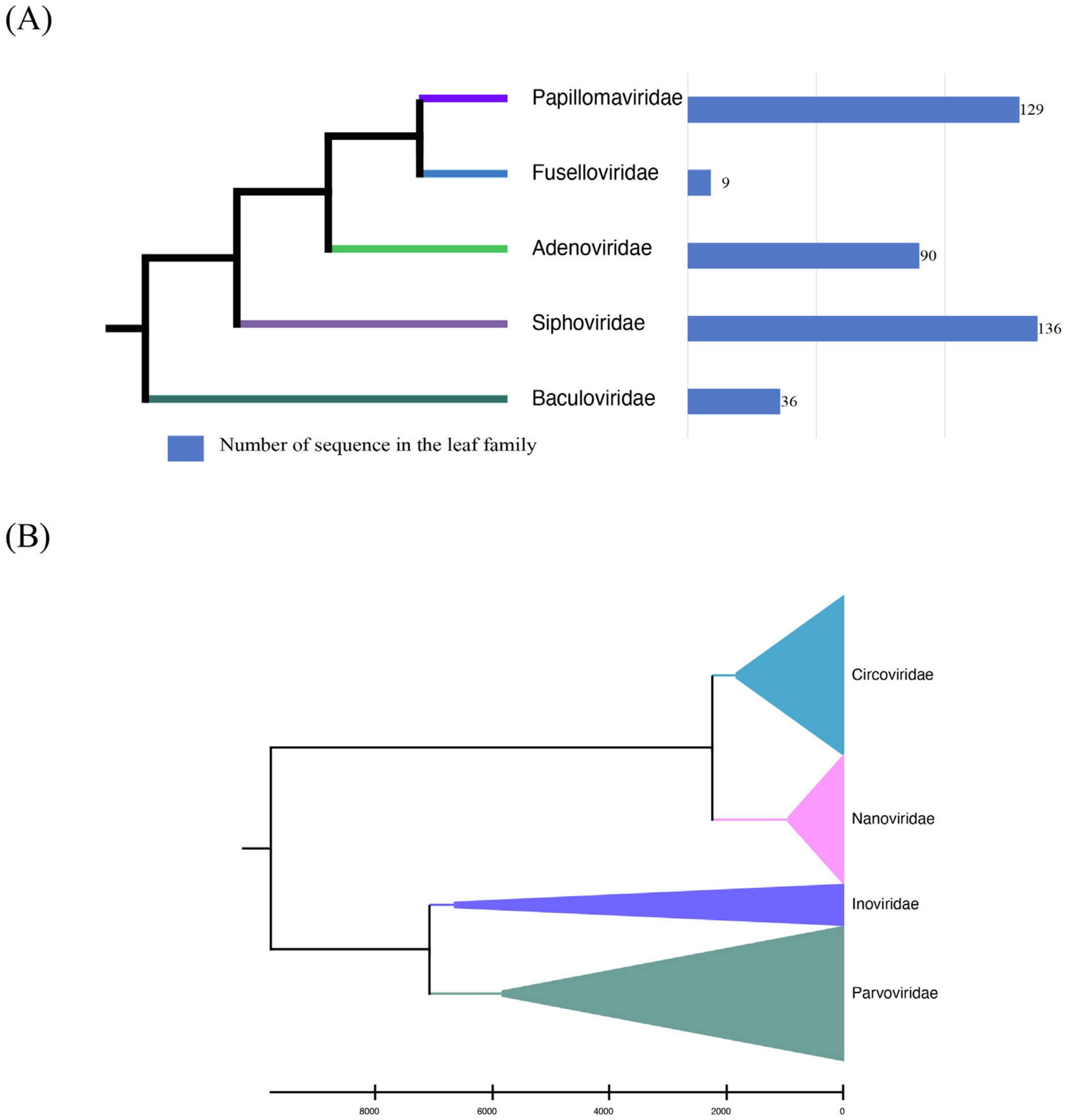

Fig. 4. Phylogenetic trees of viruses from Baltimore class I and II, respectively. In a 32-dimmensional genome space, we can use the new metric to perform phylogenetic analysis. (A) Tree of five families from Baltimore I. Sequence number of each group is presented besides the tree. (B) Tree of four families from Baltimore II. Genome sequences are clustered into four clades.

genome states that the convex hulls corresponding to different families are mutually disjoint. Besides, there are no two species that give the same point in the convex hull. Since the convex hull delimits and delineates the boundary of the same family or genus among the genome universe, if we can find a nucleotide sequence whose natural vector lies in the convex hull, then we have found a new, undiscovered species in this family [37]. Most phylogenetic analysis is mainly based on known sequences. Convex hull principle makes it possible to detect unknown but possible existent sequences and conduct further analysis. Moreover, it can create genome space and can be used to sequence classification and genome comparison with the same topological structure globally. Thus, we established the fundamental laws of genomes from a mathematical perspective.

There are still a few remaining goals to be accomplished. First, the dimension determination of the natural vector is associated with the size and category of the genome sequence dataset. If we use other genome datasets, such as bacteria or archaea, we may need to recalculate the dimensions of the space. Second, the boundaries of protein convex hull have been demonstrated to be basically stable [10]. However, the resulting convex hull boundaries of viruses may become bigger as more viral sequences are discovered. We will test the stability of the boundaries of virus family in future studies.

Funding

This work is supported by Tsinghua University Spring Breeze Fund (2020Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008). Professor Stephen S.-T. Yau is grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done.

Author contributions

Stephen S.-T. Yau conceived the project and designed the studies with Rong Lucy He. Nan Sun and Lily He collected data. Nan Sun

and Shaojun Pei carried out the data analysis including figures drawing and wrote the preliminary version of the paper. All authors participated in writing up the paper. The final version was done by Nan Sun, Shaojun Pei, Changchuan Yin and Stephen S.-T. Yau;

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript. We thank NCBI database for supporting transparent sharing of viral genomic data. The extract data can be found in Data A.1. We thank Chih-Jen Lin's lab for developing a library for support vector machines.

Data and materials availability

All data is available in the main text, supplementary materials, and the projections of convex hull pairs have been stored in https://github.com/sunn19/Virus_Genome_Space.git.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.07.028.

References

[1] Deng M, Yu CL, Liang Q, He RL, Yau SST. A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One 2011;6:e17293.
[2] Yu CL, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, et al. Real time classification of viruses in 12 dimensions. PLoS One. 2013;8:E64328.
[3] Wen J, Chan RHF, Yau SC, He RL, Yau SST. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene 2014;546:25–34.
[4] Yin C, Chen Y, Yau SST. A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. J Theor Biol 2014;359:18–28.
[5] Sun N, Dong R, Pei S, Yin C, Yau SST. A new method based on coding sequence density to cluster bacteria. J Comput Biol 2020;27:1688–98.
[6] Yau SST, Mao WG, Benson M, He RL. Distinguishing proteins from arbitrary amino acid sequences. Sci Rep 2015;5:7972.
[7] Zheng H, Yin C, Hoang T, Yau SST. Ebolavirus classification based on natural vectors. DNA Cell Biol 2015;34:418–28.
[8] Dong R, He L, He RL, Yau SST. A novel approach to clustering genome sequences using inter-nucleotide covariance. Front Genet 2019;10:234.
[9] Yu CL, Deng M, Cheng SY, Yau SC, He RL, Yau SST. Protein space: a natural method for realizing the nature of protein universe. J Theor Biol 2013;318:197–204.
[10] Zhao X, Tian K, He RL, Yau SST. Convex hull principle for classification and phylogeny of eukaryotic proteins. Genomics 2019;111:1777–84.
[11] The arabidopsis genome initiative. analysis of the genome sequence of the flowering plant arabidopsis thaliana. Nature 2000;408:796–815.
[12] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, et al. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 2001;29:22–8.
[13] International Human Genome Sequencing Consortium., Whitehead institute for Biomedical Research, Center for Genome Research., Lander, E. et al. Initial sequencing and analysis of the human genome. Nature. 409, 860–921 (2001).
[14] Himmelreich R, Hilber H, Plagens H, Pirkl E, Li BC, Herrmann R. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 1996;24:4420–49.
[15] Blaisdell BE. A measure of the similarity of sets of sequences not requiring sequence alignment. PNAS 1986;83:5155–9.
[16] Sims GE, Jun SR, Wu GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. PNAS 2009;106:2677–82.
[17] Liu S, Pei SJ, Yau SST, Wu Q. Assessment of kmer degeneration method for complicated genomes. Commun. Inf. Syst 2019;19:17–35.
[18] Dong R, Zheng H, Tian K, Yau SC, Mao WG, Yu WP, et al. Virus database and online inquiry system based on natural vectors. Evolutionary Bioinformatics. 2017;13. 1176934317746667.
[19] Mark DB, Marc VK, Mark O, Otfried S. Computational geometry. Berlin, Heidelberg: Springer; 1997. 1–17.
[20] Sun M, Zhang D, Wang Z, Ren J, Jin JS. Monte Carlo convex hull model for classification of traditional Chinese paintings. Neurocomputing. 2016;171:788–97.
[21] Singh N, Arya R, Agrawal RK. A convex hull approach in conjunction with Gaussian mixture model for salient object detection. Digital Signal Process 2016;55:22–31.
[22] Das N, Pramanik S, Basu S, Saha PK, Sarkar R, Kundu M, et al. Recognition of handwritten Bangla basic characters and digits using convex hull based feature set. arXiv. 2014;1410:0478.
[23] Cupec R, Vidović I, Filko D, Đurović P. Object recognition based on convex hull alignment. Pattern Recogn 2020;102:107199.
[24] Muller DE, Preparata FP. Finding the intersection of two convex polyhedra. Theoret Comput Sci 1978;7:217–36.
[25] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 1992;92:144.
[26] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1–27.
[27] Barker M, Rayens W. Partial least squares for discrimination. Journal of Chemometrics. 2003;17:166–73.
[28] Boyd S, Lieven V. Convex optimization. Cambridge 2004.
[29] Cortes C, Vapnik V. Support vector networks. Machine Learning. 1995;20:273–97.
[30] Martinez AM, Kak AC. PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 2001;23:228–33.
[31] Deng M, Yu C, Liang Q, He R, Yau SST. A novel method of characterizing genetic sequences: genome space with biological distance and applications. Plos one. 2011;6:E17293.
[32] Zheng H, Yin CC, Hoang T, He RL, Yang J, Yau SST. Ebolavirus classification based on natural vectors. DNA Cell Biol 2015;34:418–28.
[33] Sneath PHA, Sokal RR. Numerical taxonomy. Freeman, San Francisco.
[34] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGAX: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.
[35] Stecher G, Tamura K, Kumar S. Molecular evolutionary genetics analysis (MEGA) for macOS. Mol Biol Evol 2020.
[36] Defense Advanced Research Projects Agency (DARPA) 2008 proposal of the 23 mathematical challenges. http://www.darpa.mil/dso/personnel/mann.htm.
[37] Zhao R, Pei S, Yau SST. New genome sequence detection via natural vector convex hull method. IEEE/ACM Transactions on Computational Biology and Bioinformatics, doi: 10.1109/TCBB.2020.3040706.