RESEARCH ARTICLE

A Novel Natural Graph for Efficient Clustering of Virus Genome Sequences

Harris Song[1,#], Nan Sun[2,#], Wenping Yu[3,*] and Stephen S.-T. Yau[2,4,*]

[1]Walnut High School, Los Angeles, USA; [2]Department of Mathematical Sciences, Tsinghua University, Beijing, P.R. China; [3]College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, P.R. China; [4]Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, P.R. China

ARTICLE HISTORY

Received: June 26, 2023
Revised: September 13, 2023
Accepted: September 26, 2023

DOI:
10.2174/0115748936269106231025064143

Abstract: ***Background*:** This study addresses the need for analyzing viral genome sequences and understanding their genetic relationships. The focus is on introducing a novel natural graph approach as a solution.

***Objective*:** The objective of this study is to demonstrate the effectiveness and advantages of the proposed natural graph approach in clustering viral genome sequences into distinct clades, subtypes, or districts. Additionally, the aim is to explore its interpretability, potential applications, and implications for pandemic control and public health interventions.

***Methods*:** The study utilizes the proposed natural graph algorithm to cluster viral genome sequences. The results are compared with existing methods and multidimensional scaling to evaluate the performance and effectiveness of the approach.

***Results*:** The natural graph approach successfully clusters viral genome sequences, providing valuable insights into viral evolution and transmission dynamics. The ability to generate directed connections between nodes enhances the interpretability of the results, facilitating the investigation of transmission pathways and viral fitness.

***Conclusion*:** The findings highlight the potential applications of the natural graph algorithm in pandemic control, transmission tracing, and vaccine design. Future research directions may involve scaling up the analysis to larger datasets and incorporating additional genetic features for improved resolution.

The natural graph approach presents a promising tool for viral genomics research with implications for public health interventions.

Keywords: Viral genome analysis, genetic relationships, natural graph approach, phylogenetic clustering.

1. INTRODUCTION

Nucleic acid sequences serve as the fundamental constituents of all organisms and play a crucial role in understanding biological components [1, 2]. Composed of nucleotides that intricately interact, they give rise to complex systems essential for life [3]. Within the vast nucleotide universe, each sequence possesses a distinct spatial arrangement of nucleotides, suggesting potential evolutionary relationships [4-6]. Unraveling these connections is of utmost importance as it offers valuable insights into genetic variation [7, 8], evolutionary processes [9, 10], and the functioning of biological systems [11, 12]. Investigating the relationships between nucleotide sequences is a pivotal endeavor in biological research, contributing to advancements in genetics, genomics, and evolutionary biology.

Nucleotide sequences are strings composed of four nucleotide bases (A, C, G, and T or U), and a k-mer is a substring of length k, and for a fixed k, there are 4^k possible k-mers for a DNA sequence [13, 14]. To facilitate sequence comparison, nucleotide sequences can be transformed into k-mer-based numerical vectors [15-18]. Many studies have demonstrated that k-mer-based methods perform well on sequence clustering problems [19, 20]. We also proposed a k-mer-based method, the k-mer Natural Vector method, in 2014 [21], which reflects the distribution of each k-mer, including its count, average position, and normalized central moment in the sequences [13, 21, 22]. It has been applied in various studies, such as measuring mutation rates [23], defining genome or protein space theory [5, 6, 24-27], and explor-

*Address Correspondence to these authors at the Department of Mathematical Sciences, Tsinghua University, Beijing, P.R. China; E-mail: yau@uic.edu (SS-TY); *Co-correspondence author, College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, P.R. China; E-mail: yuwenping@tust.edu.cn (WY)

[#]*These authors contributed equally to this work.*

ing sequence minimal models [28]. In this study, we employ this method to transform nucleotide sequences into vectors. The mathematical distance between these vectors serves as a measure of sequence similarity, allowing for comparative analysis of the sequences [29-33].

The distance-based natural graph representation method elucidates the relationships between sequences [5]. Natural graphs provide a mathematical representation of interconnected elements observed in the natural world. These graphs capture intricate patterns and relationships in natural systems, enabling a deeper understanding of complex phenomena. Our team used to develop a natural graph approach that considers only the minimum distance between nodes [4, 5, 34] while disregarding other distances. However, it is important to consider additional distances to capture a more comprehensive view of the relationships within the graph.

By improving upon traditional natural graph methods, we proposed a novel approach to uncover significant connections between nucleotide sequences. Our method does not rely on the dimensionality of sequence vector representations but instead relies on a computed distance matrix. When representing the data in a 2D plane using the distance matrix, some distance values are inevitably lost. However, our method takes into account as many distance values as possible. We analyzed five subsets within the nucleotide sequence universe and determined the correlations between sequences. The results demonstrate the effectiveness of our method.

2. MATERIALS AND METHODS

2.1. K-mer Natural Vector

The k-mer Natural Vector is an encoding method that maps a biological sequence in the sequence space $\mathcal{S}$ to a vector space $\mathcal{E}$ [21].

The definition of k-mer should be introduced first. $S = (s_1 s_2 \dots s_n) \in \mathcal{S}$ is a genome sequence of length n, $s_i \in \mathcal{L} = \{A, C, G, T \text{ or } U\}$; k-mer, $kstring_i$, is a subsequence of length k, which has 4^k types (*i.e.* $i = 1, 2, \dots, 4^k$). Specifically, if $k = 1$, $kstring_i \in \mathcal{L}_1 = \{A, C, G, T\}$; if $k = 2$, $kstring_i \in \mathcal{L}_2 = \{\mathrm{AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT}\}$. There are n-k+1 k-mers for sequence S: $kstring_1 = s_1 s_2 \dots s_k, kstring_2 = s_2 s_3 \dots s_{k+1}, \dots, kstring_{n-k+1} = s_{n-k+1} s_{n-k+2} \cdots s_n$.

The distribution of the k-mer, $kstring_i$, can be characterized by three components: $n_{kstring_i}$, $\mu_{kstring_i}$ and $D^{\mathrm{j}}_{\mathrm{kstring}_i}$. The definition of each component is as follows:

- $n_{kstring_i}$: It represents the frequency of occurrence of the k-mer $kstring_i$ in a given sequence S;
- $\mu_{kstring_i} = \sum_{l=1}^{n_{kstring_i}} \frac{w_{kstring_i}[l]}{n_{kstring_i}}$: It represents the average position of the k-mer $kstring_i$ in sequence S, where $w_{kstring_i}[l]$ is the l-th position of $kstring_i$ in the sequence;
- $D^j_{kstring_i} = \sum_{l=1}^{n_{kstring_i}} \frac{\left(w_{kstring_i}[l]-\mu_{kstring_i}\right)^j}{n^{j-1}_{kstring_i}(n-k+1)^{j-1}}$: It represents the j-order normalized central moment of k-mer $kstring_i$ occurring in sequence S.

Previous studies have demonstrated that the second order is sufficient to represent the sequence S and there is a one-to-one correspondence between the vector and the sequence [5]. The second-order k-mer Natural Vector of a DNA sequence is defined as follows:

$$\Big(n_{kstring_1}, n_{kstring_2}, \cdots\cdots, n_{kstring_{4^k}},$$
$$\mu_{kstring_1}, \mu_{kstring_2}, \cdots\cdots, \mu_{kstring_{4^k}},$$
$$D^2_{kstring_1}, D^2_{kstring_2}, \cdots\cdots, D^2_{kstring_{4^k}}\Big).$$

2.2. Sequence Similarity Measurement

Each sequence can be converted into a k-mer Natural Vector, and the similarity between sequences can be measured by the geometric distance between their corresponding vectors. The most commonly used distance includes Euclidean distance. The Euclidean distance of the k-mer Natural Vector is defined as $d_k: \mathcal{E} \times \mathcal{E} \to R$, $d_k(x, y) = (\sum_{i=1}^{n} (y_i - x_i)^2)^{\frac{1}{2}}$, and the weighted summation of the distance yields better performance [4]: $D_n(x, y) = a_1 d_1(x, y) + a_2 d_2(x, y) + a_3 d_3(x, y) + \cdots + a_n d_n(x, y)$, $a_k \in R^+, k = 1, 2, \dots, n$. The weighted distance is a true metric and satisfies the following conditions in the metric space: (1) Nonnegativity: $d(x, y) \geq 0$, $d(x, y) = 0$ if and only if $x = y$; (2) Symmetry: $d(x, y) = d(y, x)$; (3) Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$, $\forall x, y, z \in \mathcal{E}$. The weight a_k reflects the contribution of the corresponding k-mer Natural Vector to the description of sequence similarity, and for virus sequences, $a_k = \frac{1}{2^{k-1}}$ [4].

2.3. A Novel Graphical Representation for Phylogeny

Distance-based phylogenetic algorithms primarily include Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [35-37], Weighted Pair Group Method with Arithmetic Mean (WPGMA) [35, 38, 39], Neighbor-Joining method (NJ) [40-42], Minimum Evolution method (ME) [43-45], and Fitch-Margoliash method (FM) [46-48]. These methods generate rooted or unrooted phylogenetic trees and provide insights into the evolutionary relationships among DNA or protein sequences [49-51]. However, different algorithms based on distance matrices can yield non-unique results, leading to ongoing debates and controversies. To address this limitation, we have developed a novel approach called the Natural Graph algorithm for inferring phylogenetic relationships in biological sequences. Unlike the previous version [5], our new approach considers not only the most similar sequence for each sequence but also the second or even the third most similar sequence. This expanded consideration helps to better capture the relationships between sequences. The algorithm is as follows:

Algorithm 1: The novel Natural Graph

Input: Distance Matrix D

Step 1: For each element i, find the nearest element j to i. Then draw a directional line from i to j.

Step 2: If i points to j and j points to k, the three elements are connected. Repeat this step until no other elements are connected. Several undirected connected subgraphs are obtained.

Step 3: Determine the distance matrix for each subgraph. The positions of elements within each subgraph are not random. We aim to ensure that the distance matrix of the connected subgraph is similar to the distance matrix of the original dataset. To achieve this, the second or third nearest element for each element is considered. More computational details can be found in Section 3.1, which provides examples of randomly generated points. The undirected subgraphs acquire directions.

Step 4: Define the distance between two subgraphs as the minimum distance between any element in one subgraph and any element in the other subgraph.

Step 5: Finally, obtain a connected directed graph representing all elements. This serves as the final directed graphical representation. The new graphical representation provides a low-dimensional visualization that preserves the distances of the high-dimensional vectors.

Output: Natural Graph and n subgraphs.

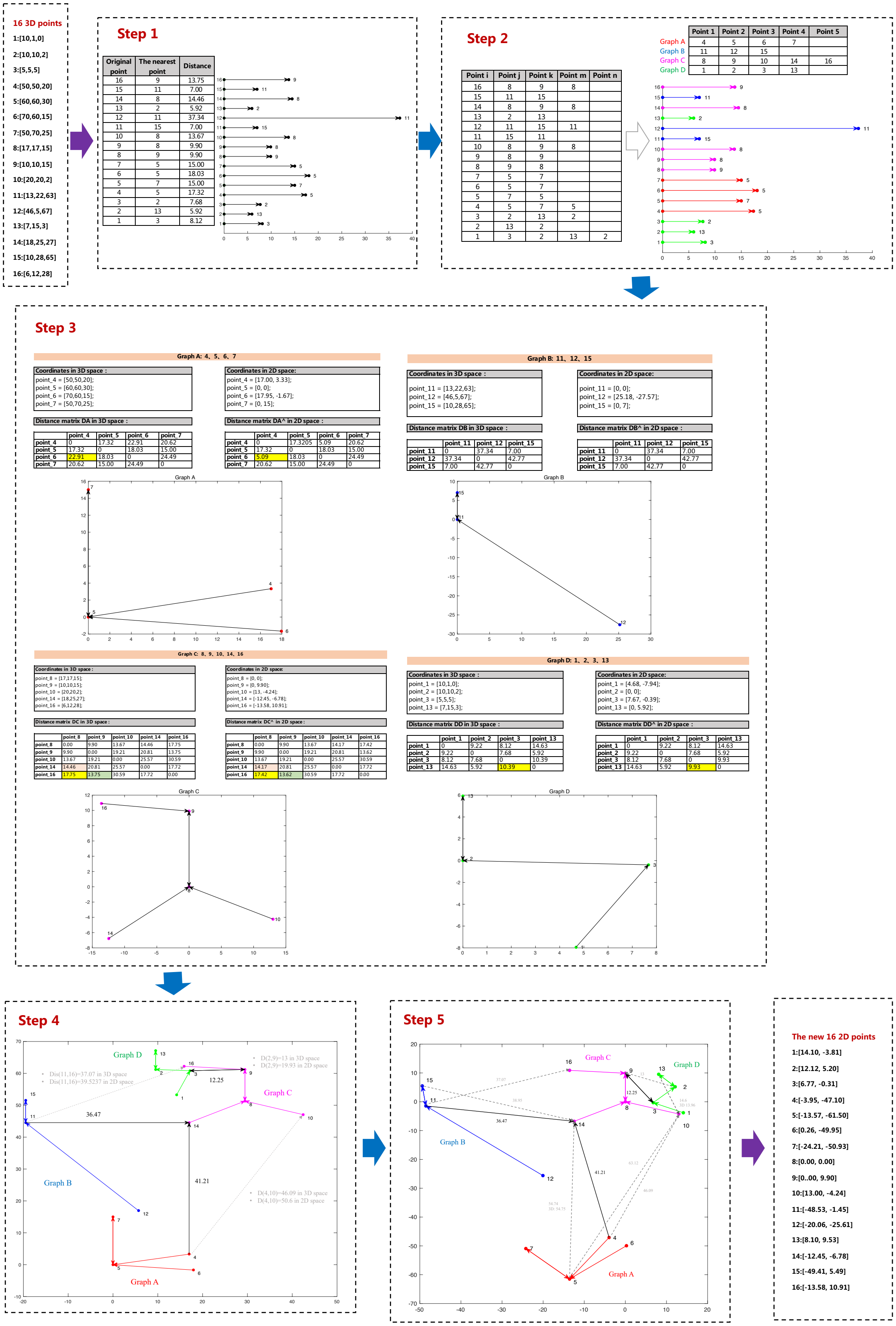
Here, the dataset consists of DNA sequences, which are converted into high-dimensional Natural Vectors. The Euclidean distance is employed to quantify the similarity between sequences. As a result, each dataset corresponds to a distance matrix. The algorithm takes a distance matrix as input. If we aim to visualize the dataset in a low-dimensional space, such as R^2 or R^3, we utilize the aforementioned steps to accomplish the projection task. The distance matrix of the newly projected points approximates the distance matrix of the high-dimensional points.

The graph generation procedure is divided into two parts. The first part involves generating individual subgraphs, while the second part focuses on creating the connected complete graph. The algorithm is unsupervised and operates on a dataset comprising n groups, with each group containing k_i elements (where $k_1 + k_2 + \cdots + k_n = N$, N represents the total number of elements in the dataset). The input to the algorithm is a distance matrix D of size N × N. By identifying the nearest element for each element, the distance matrix D is divided into n submatrices, namely D1, D2, ..., Dn (corresponding to steps 1 and 2 of Algorithm 1). The corresponding subgraphs are denoted as G1, G2, ..., Gn. For each subgraph, a directed line is drawn between the two closest vertices and the second or third nearest vertex is determined through the intersection of two circles (step 3 of Algorithm 1). The next step involves combining the subgraphs into a single graph. We begin by identifying the subgraph with the closest distance to another subgraph. One of the subgraphs is selected as the central component of the drawing, and a rotation is applied toward the other subgraph while considering the shortest distances. The first shortest distance, second shortest distance, and possibly the third shortest distance (depending on the availability of x, y, and rotation θ positions) are determined (steps 4 and 5 of Algorithm 1). It is important to note that the 2-dimensional Natural Graph should consider as many distances as possible, and the consideration of the third closest distance only occurs after the computation of the previous two distances.

3. RESULTS

3.1.The 2D Graphical Representation of Randomly Generated 3D Points

The new Natural Graph algorithm is first built within a randomly generated 3D scenario which made it easier to illustrate the algorithm's principles. We want to visualize the 16 3D points in 2D space. The 16 3D points are displayed in Fig. (**1**), numbered 1, 2, …, 16. The distance matrix of the 16 3D points is computed first, and the nearest point j is determined for each point i (see Step 1 in Fig **1**. $i = 1,2, \ldots, 16, j \in \{1,2, \ldots,16\}$). Then we get four connected subgraphs A, B, C, D (see Step 2 in Fig. **1**). Here Graph A = {Points 4, 5, 6, 7}, Graph B = {Points 11, 12, 15}, Graph C = {Points 8, 9, 10, 14, 16}, Graph D = {Points 1, 2, 3, 13}. Next, the individual distance matrix for each subgraph is calculated (see Step 3 in Fig. **1**). For distance matrix DA, the distance between Points 5 and 7 is the smallest ($d(5,7) = 15$), so there is a double arrow between them. The second largest distance value is $d(4,5) = 17.32$, and there is an arrow from Point 4 to Point 5; $d(4,7) = 20.62$ is also considered, that is, Point 4 is on the circumference $(x - x_7)^2 + (y - y_7)^2 = d(4,7)^2$. Point 6 points to Point 5, and $d(6,4)$ and $d(6,7)$ are considered. The distance matrix DA is similar to DA^, here we ignore $d(4,6)$. For graph B, Point 11 and Point 15 point to each other. Point 12 points to Point 11 and is on the circumference $(x - x_{15})^2 + (y - y_{15})^2 = d(12,15)^2$. For distance matrix DC, the distance between Points 8 and 9 is the smallest, and they point to each other. Point 10 is at the intersection of the circumference $(x - x_8)^2 + (y - y_8)^2 = d(8,10)^2$ and $(x - x_9)^2 + (y - y_9)^2 = d(9,10)^2$. Point 14 points to Point 8, $d(9,14) = 20.81$ and $d(10,14) = 25.57$ are also considered. Point 16 points to Point 9, $d(8,16)$, $d(10,16)$, and $d(14,16)$ are also considered. For distance matrix DD, Points 2 and 13 point to each other. Point 1 is at the intersection of the circumference $(x - x_2)^2 + (y - y_2)^2 = d(1,2)^2$ and $(x - x_{13})^2 + (y - y_{13})^2 = d(1,13)^2$. Point 3 points to Point 2, $d(3,1)$ and $d(13,3)$ are also considered. The four graphs are further connected (see Step 4 in Fig. **1**, Fig. **2**). The distance between two graphs is the minimum distance between any point in one graph and any point in the other graph. We further consider the second or third nearest



16 3D points

1	[10,1,0]
2	[10,10,2]
3	[5,5,5]
4	[50,50,20]
5	[60,60,30]
6	[70,60,15]
7	[50,70,25]
8	[17,17,15]
9	[10,10,15]
10	[20,20,2]
11	[13,22,63]
12	[46,5,67]
13	[7,15,3]
14	[18,25,27]
15	[10,28,65]
16	[6,12,28]

Step 1

Original point	The nearest point	Distance
16	9	13.75
15	11	7.00
14	8	14.46
13	2	5.92
12	11	37.34
11	15	7.00
10	8	13.67
9	8	9.90
8	9	9.90
7	5	15.00
6	5	18.03
5	7	15.00
4	5	17.32
3	2	7.68
2	13	5.92
1	3	8.12

Step 2

Point i	Point j	Point k	Point m	Point n
16	8	9	8	
15	11	15		
14	8	9	8	
13	2	13		
12	11	15	11	
11	15	11		
10	8	9	8	
9	8	9		
8	9	8		
7	5	7		
6	5	7		
5	7	5		
4	5	7	5	
3	2	13	2	
2	13	2		
1	3	2	13	2

	Point 1	Point 2	Point 3	Point 4	Point 5
Graph A	4	5	6	7	
Graph B	11	12	15		
Graph C	8	9	10	14	16
Graph D	1	2	3	13	

Step 3

Graph A: 4、5、6、7

Coordinates in 3D space：

point_4 = [50,50,20];
point_5 = [60,60,30];
point_6 = [70,60,15];
point_7 = [50,70,25];

Coordinates in 2D space:

point_4 = [17.00, 3.33];
point_5 = [0, 0];
point_6 = [17.95, -1.67];
point_7 = [0, 15];

Distance matrix DA in 3D space：

	point_4	point_5	point_6	point_7
point_4	0	17.32	22.91	20.62
point_5	17.32	0	18.03	15.00
point_6	22.91	18.03	0	24.49
point_7	20.62	15.00	24.49	0

Distance matrix DA^ in 2D space：

	point_4	point_5	point_6	point_7
point_4	0	17.3205	5.09	20.62
point_5	17.32	0	18.03	15.00
point_6	5.09	18.03	0	24.49
point_7	20.62	15.00	24.49	0

Graph B: 11、12、15

Coordinates in 3D space：

point_11 = [13,22,63];
point_12 = [46,5,67];
point_15 = [10,28,65];

Coordinates in 2D space:

point_11 = [0, 0];
point_12 = [25.18, -27.57];
point_15 = [0, 7];

Distance matrix DB in 3D space：

	point_11	point_12	point_15
point_11	0	37.34	7.00
point_12	37.34	0	42.77
point_15	7.00	42.77	0

Distance matrix DB^ in 2D space：

	point_11	point_12	point_15
point_11	0	37.34	7.00
point_12	37.34	0	42.77
point_15	7.00	42.77	0

Graph C: 8、9、10、14、16

Coordinates in 3D space :

point_8 = [17,17,15];
point_9 = [10,10,15];
point_10 = [20,20,2];
point_14 = [18,25,27];
point_16 = [6,12,28];

Coordinates in 2D space:

point_8 = [0, 0];
point_9 = [0, 9.90];
point_10 = [13, -4.24];
point_14 = [-12.45, -6.78];
point_16 = [-13.58, 10.91];

Distance matrix DC in 3D space :

	point_8	point_9	point_10	point_14	point_16
point_8	0.00	9.90	13.67	14.46	17.75
point_9	9.90	0.00	19.21	20.81	13.75
point_10	13.67	19.21	0.00	25.57	30.59
point_14	14.46	20.81	25.57	0.00	17.72
point_16	17.75	13.75	30.59	17.72	0.00

Distance matrix DC^ in 2D space :

	point_8	point_9	point_10	point_14	point_16
point_8	0.00	9.90	13.67	14.17	17.42
point_9	9.90	0.00	19.21	20.81	13.62
point_10	13.67	19.21	0.00	25.57	30.59
point_14	14.17	20.81	25.57	0.00	17.72
point_16	17.42	13.62	30.59	17.72	0.00

Graph D: 1、2、3、13

Coordinates in 3D space：

point_1 = [10,1,0];
point_2 = [10,10,2];
point_3 = [5,5,5];
point_13 = [7,15,3];

Coordinates in 2D space:

point_1 = [4.68, -7.94];
point_2 = [0, 0];
point_3 = [7.67, -0.39];
point_13 = [0, 5.92];

Distance matrix DD in 3D space：

	point_1	point_2	point_3	point_13
point_1	0	9.22	8.12	14.63
point_2	9.22	0	7.68	5.92
point_3	8.12	7.68	0	10.39
point_13	14.63	5.92	10.39	0

Distance matrix DD^ in 2D space：

	point_1	point_2	point_3	point_13
point_1	0	9.22	8.12	14.63
point_2	9.22	0	7.68	5.92
point_3	8.12	7.68	0	9.93
point_13	14.63	5.92	9.93	0

Step 4

- Dis(11,16)=37.07 in 3D space
- Dis(11,16)=39.5237 in 2D space
- D(2,9)=13 in 3D space
- D(2,9)=19.93 in 2D space
- D(4,10)=46.09 in 3D space
- D(4,10)=50.6 in 2D space

Step 5

The new 16 2D points

1	[14.10, -3.81]
2	[12.12, 5.20]
3	[6.77, -0.31]
4	[-3.95, -47.10]
5	[-13.57, -61.50]
6	[0.26, -49.95]
7	[-24.21, -50.93]
8	[0.00, 0.00]
9	[0..00, 9.90]
10	[13.00, -4.24]
11	[-48.53, -1.45]
12	[-20.06, -25.61]
13	[8.10, 9.53]
14	[-12.45, -6.78]
15	[-49.41, 5.49]
16	[-13.58, 10.91]

Fig. (1). Example of mapping 16 3D points into 2D space for graphic representation. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

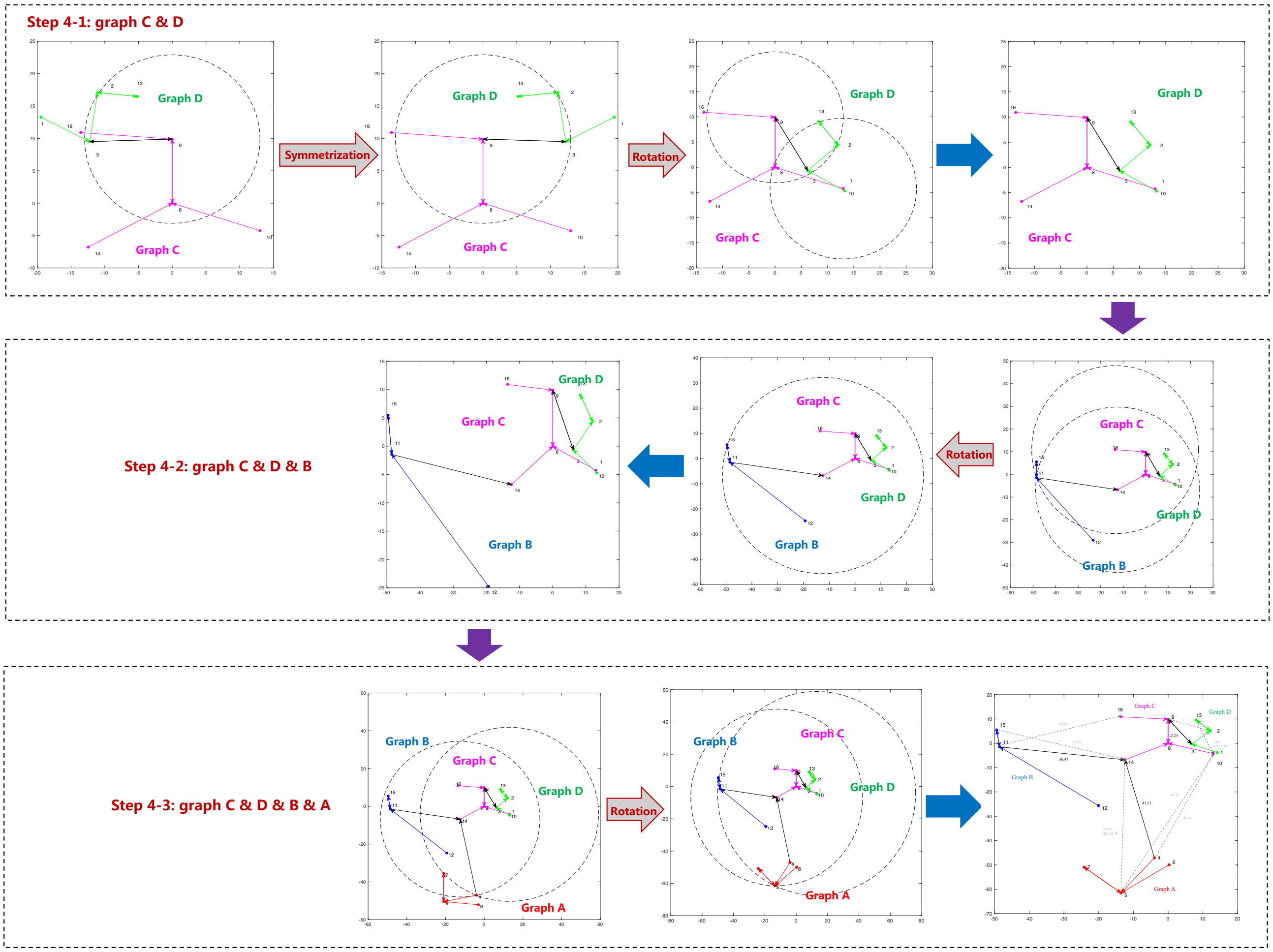


Fig. (2). The implementation details of connected directed graph from step 4 to step 5. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

distance between graphs to obtain the final directed connected graph (see Step 5 in Fig. **1**), these distance values are marked in gray color, and the new 16 2D points are displayed beside. Each subgraph is distinguished in different colors. The dotted lines represent the shortest distances between each subgraph. The graphical representation of the mapping from the m-dimensional space to the n-dimensional space can be similarly implemented (m<n).

The graph in step 4 obtains the direction by considering the second and third closest distances between the subgraphs (Fig. **2**). Specifically, the coordinates are symmetrized or rotated step by step. The symmetry formula is as follows: (x_1, y_1) and (x_2, y_2) are symmetric about line $x = x_0$, if and only if they meet the property: $\begin{cases} x_1 + x_2 = 2x_0 \\ y_1 = y_2 \end{cases}$. The rotation formula is as follows: the original point $\begin{cases} x_1 = Lcos(\phi) \\ y_1 = Lsin(\phi) \end{cases}$ is rotated by θ angle, then the new point is $\begin{cases} x_2 = Lcos(\phi+\theta) = x_1\cos(\theta) - y_1\sin(\theta) \\ y_2 = Lsin(\phi+\theta) = y_1\cos(\theta) + x_1\sin(\theta) \end{cases}$, that is $\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}$; if the rotation center (x_0, y_0) is translated to the origin $(0, 0)$, the rotation formula is $\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \left[\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} - \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \right] + \begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$. Through the above two formulas, we can get the directed connected graph in step 5.

3.2. The 2D Graphical Representation of SARS-CoV-2 Genomes

The COVID-19 pandemic has profoundly affected human health over the past years. Notably, the virus's extensive variability, especially within the S protein, has given rise to the emergence of numerous viral subtypes [52]. According to the classification criteria of GISAID, SARS-CoV-2 can be classified into 12 evolutionary clades: G, GH, GK, GKA, GR, GR, GR, GV, L, O, S, V [53-55]. The G clade is also known as S-D614G, indicating a mutation of the 614th amino acid of the S protein from D (aspartic acid) to G (glycine). The GH clade is referred to as S-D614G+NS3-Q57H, indicating mutations at the 614th amino acid of the S protein

Table 1. The statistical information of the SARS-CoV-2 datasets.

Clade Name	DisMat1		DisMat2		Subgraph label
	Sequence Number	Sequence Label	Sequence Number	Sequence Label	
G	5	1~5	9	1~9	1
GH	5	6~10	7	10~16	2
GR	3	11~13	9	17~25	3
GRA	4	14~17	10	26~35	4
GRY	5	18~22	10	36~45	5
GV	5	23~27	4	46~49	6
L	2	28~29	4	50~53	7
O	5	30~34	10	54~63	8
S	5	35~39	10	64~73	9
GK	3	40~42	10	74~83	10
V	2	43~44	8	84~91	11
GKA	3	45~47	3	92~94	12
Total	47	1~47	94	1~94	12

(D (aspartic acid) to G (glycine)) and the 57th amino acid of the NS3 protein (Q (glutamine) to H (histidine)). Other clades follow a similar naming convention [53, 56]. We obtained all complete genome sequences of SARS-CoV-2 from GISAID as of June 22, 2022 (https://gisaid.org). Only human-derived genomes were included, and low-quality sequences were removed from the dataset, resulting in a total of 115,390 sequences. Among them, the GKA clade had only 3 high-quality sequences. To test the new Natural Graph algorithm, we selected two subsets of the dataset. Specifically, we randomly chose 47 sequences and 94 sequences, and assigned them labels 1 to 47 and 1 to 94, respectively. Their distance matrix was calculated based on k-mer Natural Vectors. The similarity between virus sequences was measured using the distance formula $D_9(x,y) = d_1(x,y) + \frac{1}{2}d_2(x,y) + \frac{1}{2^2}d_3(x,y) + \cdots + \frac{1}{2^8}d_n(x,y)$ [4]. These two distance matrices are referred to as DisMat1 and DisMat2, with sizes of 47 by 47 and 94 by 94, respectively. The statistical information is presented in Table **1**.

Fig. (**3**) illustrates the 2D graph representation of the distance matrix DisMat1. The 47 points are clustered into 12 groups, corresponding to 12 subgraphs as shown in Fig. (**3A**). Each subgraph contains different sequences. For example, subgraph 1 contains sequences labeled 1 to 5, subgraph 2 contains sequences labeled 6 to 10, …, and subgraph 12 contains sequences labeled 45 to 47. According to the sequence label of DisMat1 in Table **1**, each subgraph corresponds exactly to a clade. By considering the second or even third closest distances and applying rotation or translation operations to the subgraphs according to the steps outlined in Section 3.1, the 2D-directed complete graph shown in Fig (**3B**) is obtained. Different clades are distinguished by different colors, and the sequence labels are omitted. Gray dashed lines indicate sequences that have a second or third closest relationship. It can be observed that all 47 sequences are correctly clustered. To facilitate a direct comparison between our method and traditional DNA sequence clustering approaches, we also constructed a phylogenetic tree, as shown in Fig (**4**). The tree was generated using the FastME software [57] with the BioNJ algorithm [58], taking the distance matrix DisMat1 as input. Different colors represent distinct clades. This further demonstrates the validity of our method in producing meaningful results.

To eliminate randomness, we also considered a slightly larger distance matrix, DisMat2, and obtained its 2D graph representation as shown in Supplementary Figs. (**S1** to **S2**). The 94 sequences are also successfully clustered in this representation.

3.3. The 2D Graphical Representation of HIV-1 Genomes

HIV, which stands for Human Immunodeficiency Virus, encompasses two types: HIV-1 and HIV-2 [59]. Among these, HIV-1 is the primary cause of global HIV infections [60]. HIV is a highly mutable virus, with the envelope gene exhibiting the highest mutation rate [61]. Based on the nucleic acid sequence differences in the envelope gene, HIV-1 can be primarily classified into three groups: M, N, and O. The M group further consists of 12 subtypes, namely A, B, C, D, E, F, G, H, I, J, K, and L. The N group only contains the N subtype, while the O group comprises the O subtype [62]. On average, the dissimilarity of the envelope gene sequence among each subtype is approximately 30%. HIV-2, on the other hand, is divided into seven main subtypes: A, B, C, D, E, F, and G. Additionally, there are circulating recombinant forms (CRFs) resulting from recombination events between different subtypes [63], For instance, CRF02_AG is a recombinant form arising from the recombination of subtypes A and G.

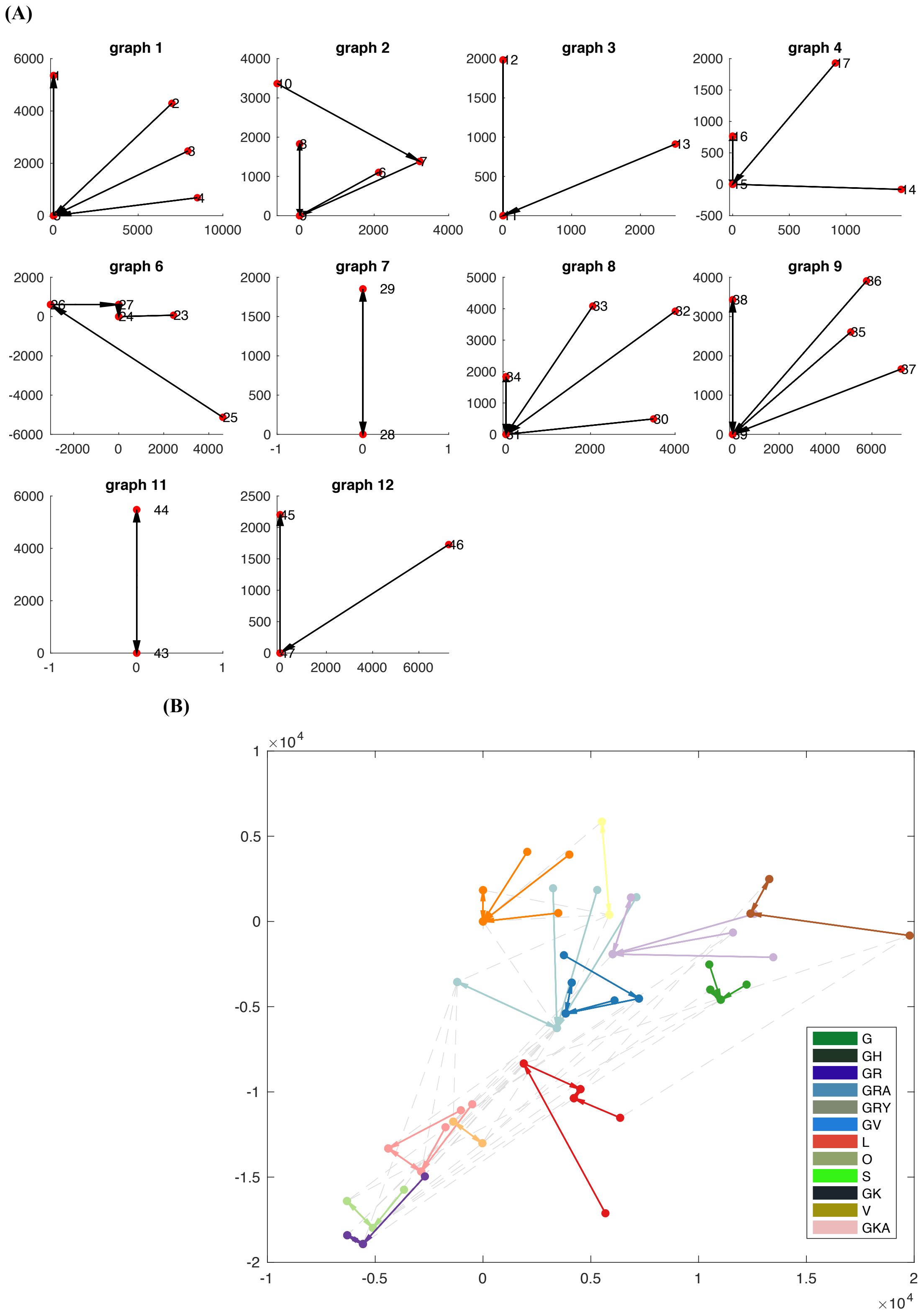


Fig. (3). (**A**) The 47 sequences are divided into 12 subgraphs, with sequence labels indicated. The corresponding clade can be determined based on the sequence labels using Table **1**. The sequence accession numbers are provided in the supplementary data. (**B**) 2D-directed complete graph of the 12 clades. Different clades are distinguished by different colors. The sequence labels are omitted here. Gray dashed lines indicate that the sequences have a second or third closest relationship. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

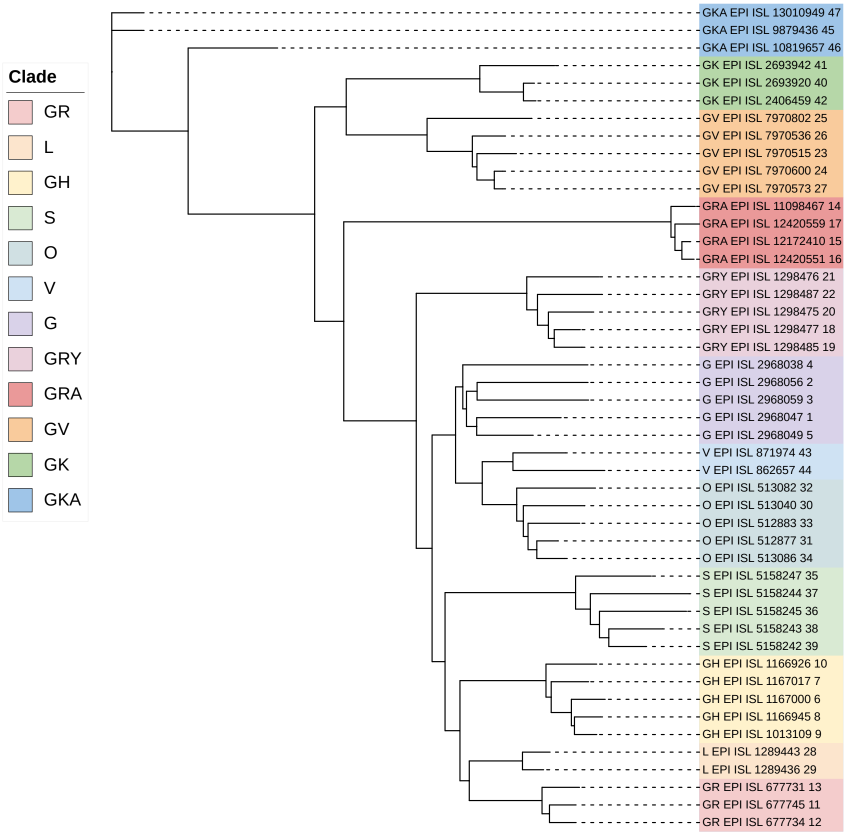


Fig. (4). The phylogenetic tree of the 47 sequences of SARS-CoV-2. The tree is based on the DisMat1 and constructed by FastME software with BioNJ algorithm. Different colors represent different clades. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 2. The statistical information of the HIV-1 datasets.

Subtype Name	DisMat3		DisMat4		Subgraph Label
	Sequence Number	Sequence Label	Sequence Number	Sequence Label	
BC	5	1~5	7	1〜7	1
01B	5	6~10	7	8〜14	2
CD	5	11~15	7	15〜21	3
BF1	5	16~20	7	22〜28	4
D	5	21~25	7	29〜35	5
A1D	5	26~30	7	36〜42	6
02_AG	5	31~35	7	43〜49	7
A1C	5	36~40	3	50〜52	8
A1	5	41~45	7	53〜59	9
01_AE	5	46~50	7	60〜66	10
C	5	51~55	7	67〜73	11
B	2	56~57	3	74〜76	12
Total	57	1〜57	76	1〜76	12

In our study, we focus on HIV-1 and utilize a dataset obtained from a previous paper [64], which was sourced from the HIV sequence Database (https://www.hiv.lanl.gov). This dataset comprises 11,897 high-quality complete genomes, up until April 8, 2022. Among the available subtypes, we specifically consider the top 12 subtypes with a larger number of genomes, namely B, C, 01_AE, A1, A1C, 02_AG, A1D, D, BF1, CD, 01B, and BC. The corresponding sequence numbers for each subtype are provided in Table **2**. For our analysis, we randomly selected 57 sequences and 76 se-

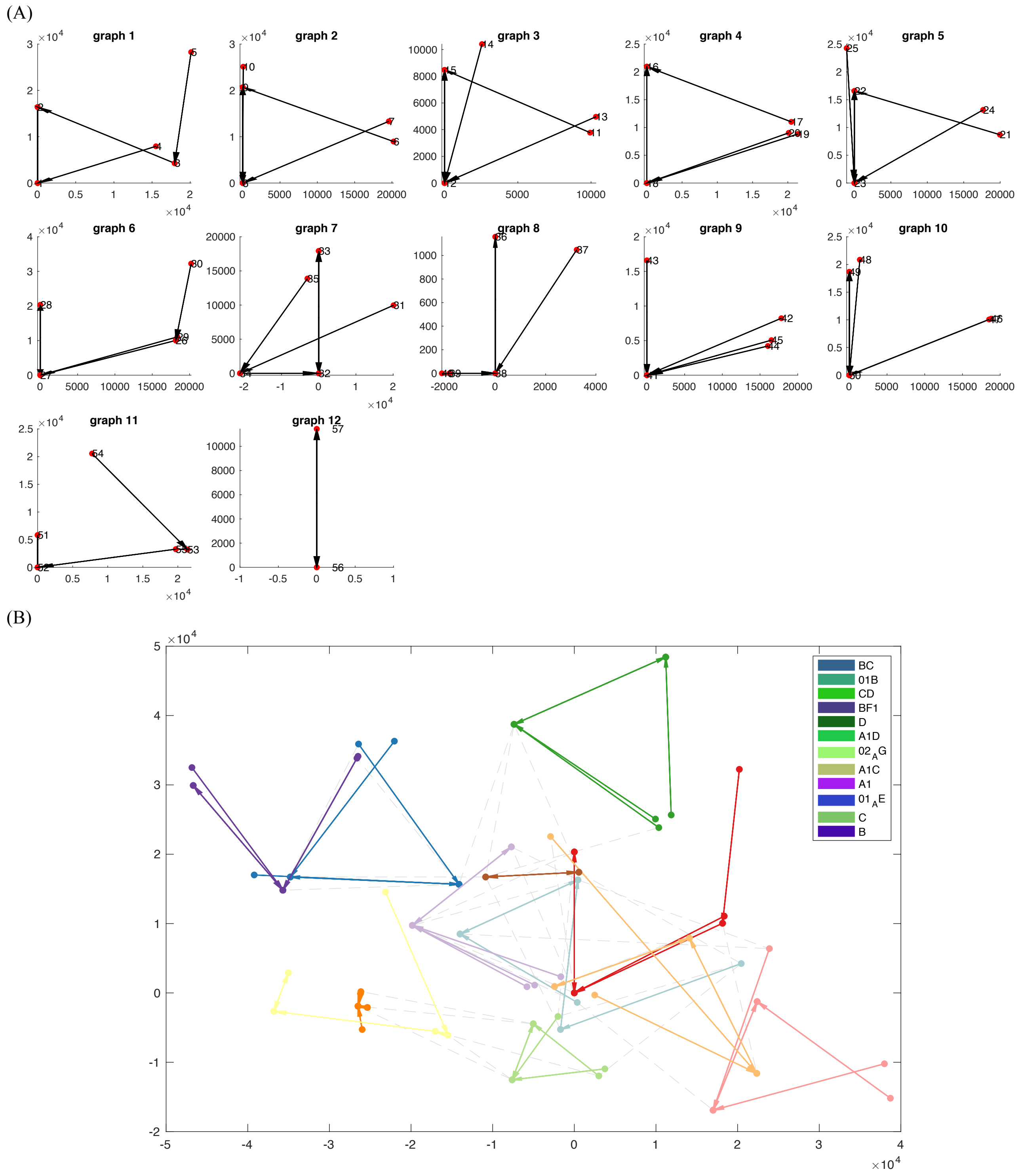


Fig. (5). (**A**) The 57 sequences are divided into 12 subgraphs, with sequence labels indicated. The corresponding subtype can be determined based on the sequence labels using Table **2**. The sequence accession numbers are provided in the supplementary data. (**B**) 2D-directed complete graph of the 12 subtypes. Different subtypes are distinguished by different colors. The sequence labels are omitted here. Gray dashed lines indicate that the sequences have a second or third closest relationship. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

quences, resulting in the calculation of weighted Euclidean distance matrices named DisMat3 (57 by 57) and DisMat4 (76 by 76), respectively.

Fig. (**5**) presents the 2D graph generated from the distance matrix DisMat3. The 57 data points are organized into 12 distinct clusters, which correspond to 12 subgraphs depicted in Fig. (**5A**). Each subgraph contains sequences specific to a particular subtype. Referring to the sequence labels in Table **2**, it is evident that each subgraph corresponds precisely to a subtype. By considering the second or even third

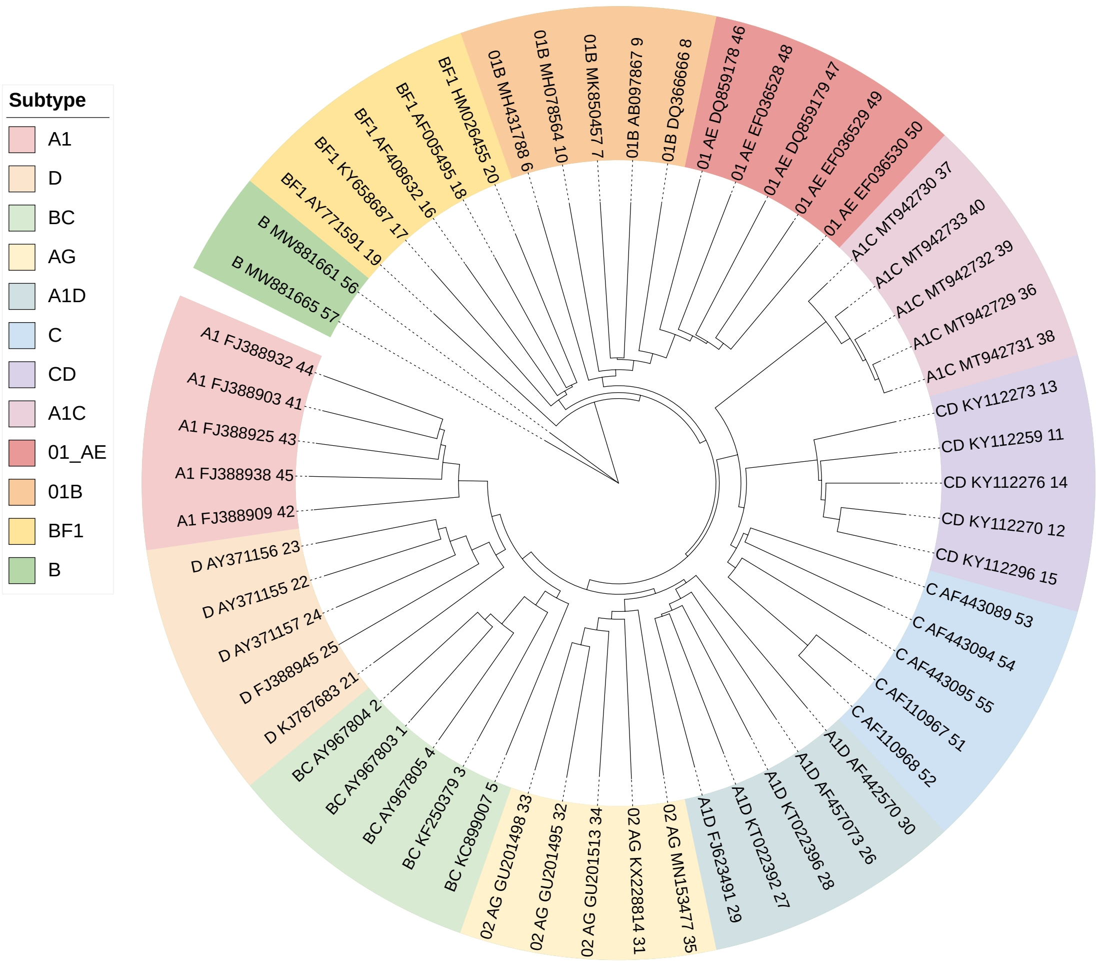


Fig. (6). The phylogenetic tree of the 57 sequences of HIV-1. The tree is based on the DisMat3 and constructed by FastME software with BioNJ algorithm. Different colors represent different subtypes. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

closest distances, we obtain the 2D-directed complete graph displayed in Fig. (**5B**). Notably, all 57 sequences are accurately clustered within the graph. We also constructed a phylogenetic tree to compare with our method, as depicted in Fig. (**6**). Distinct subtypes are represented by different colors, further validating the effectiveness of our approach in producing meaningful results.

To mitigate the effects of randomness, we also examined a slightly expanded distance matrix, DisMat4, and derived its corresponding 2D graph visualization, presented in Supplementary Figs. (**S3** to **S4**). In this representation, the 76 sequences were effectively clustered, further reinforcing the reliability and consistency of our approach.

3.4. The 2D Graphical Representation of H1N1 NS1 Sequence

Influenza, a viral respiratory infection, is categorized into four types: A, B, C, and D [65]. Type A influenza, specifically the H1N1 strain, is characterized by the presence of type 1 hemagglutinin and type 1 neuraminidase proteins [66]. A typical influenza A virus comprises 8 RNA strands [67], one of which encodes the nuclear export protein and non-structural protein 1 (NS1) [65]. NS1 gene mutations are associated with the virulence of H1N1 and can serve as indicators of its evolutionary trajectory [68]. To conduct our analysis, we retrieved a dataset of 18,211 H1N1 NS1 sequences from the NCBI database (https://www.ncbi.nlm.nih.gov) up to April 12, 2022, using the keywords "H1N1" and "NS1". The dataset encompassed sequences from 845 districts. We focused on the top 10 districts with the highest number of genomes, namely California, Singapore, New York, Iowa, Texas, Hong Kong, Minnesota, Florida, Thailand, and Italy. The corresponding sequence counts for each district are presented in Table **3**. We randomly chose 36 sequences totally and calculated the distance matrix, named DisMat5, with dimensions of 36 by 36. This matrix captures the pairwise distances between the sequences, providing insights into their genetic relationships.

Fig. (**7**) illustrates the 2D graph representation of DisMat5, where 36 data points form 10 distinct clusters (Fig. **7A**). Each cluster corresponds to sequences from a specific district, as indicated by Table **3**. Considering the second and third distances, we obtain the 2D-directed complete graph (Fig **7B**) where all 36 sequences are accurately clustered. In addition, we constructed a phylogenetic tree (Fig. **8**) for comparison, highlighting different districts with distinct colors.

Table 3. The statistical information of the H1N1 NS1 dataset.

District Name	DisMat5		Subgraph Label
	Sequence Number	Sequence Label	
California	4	1~4	1
Florida	3	5~7	2
Hong Kong	4	8~11	3
Iowa	3	12~14	4
Italy	4	15~18	5
Minnesota	3	19~21	6
New York	5	22~26	7
Singapore	2	27~28	8
Texas	3	29~31	9
Thailand	5	32~36	10
Total	36	1~36	10

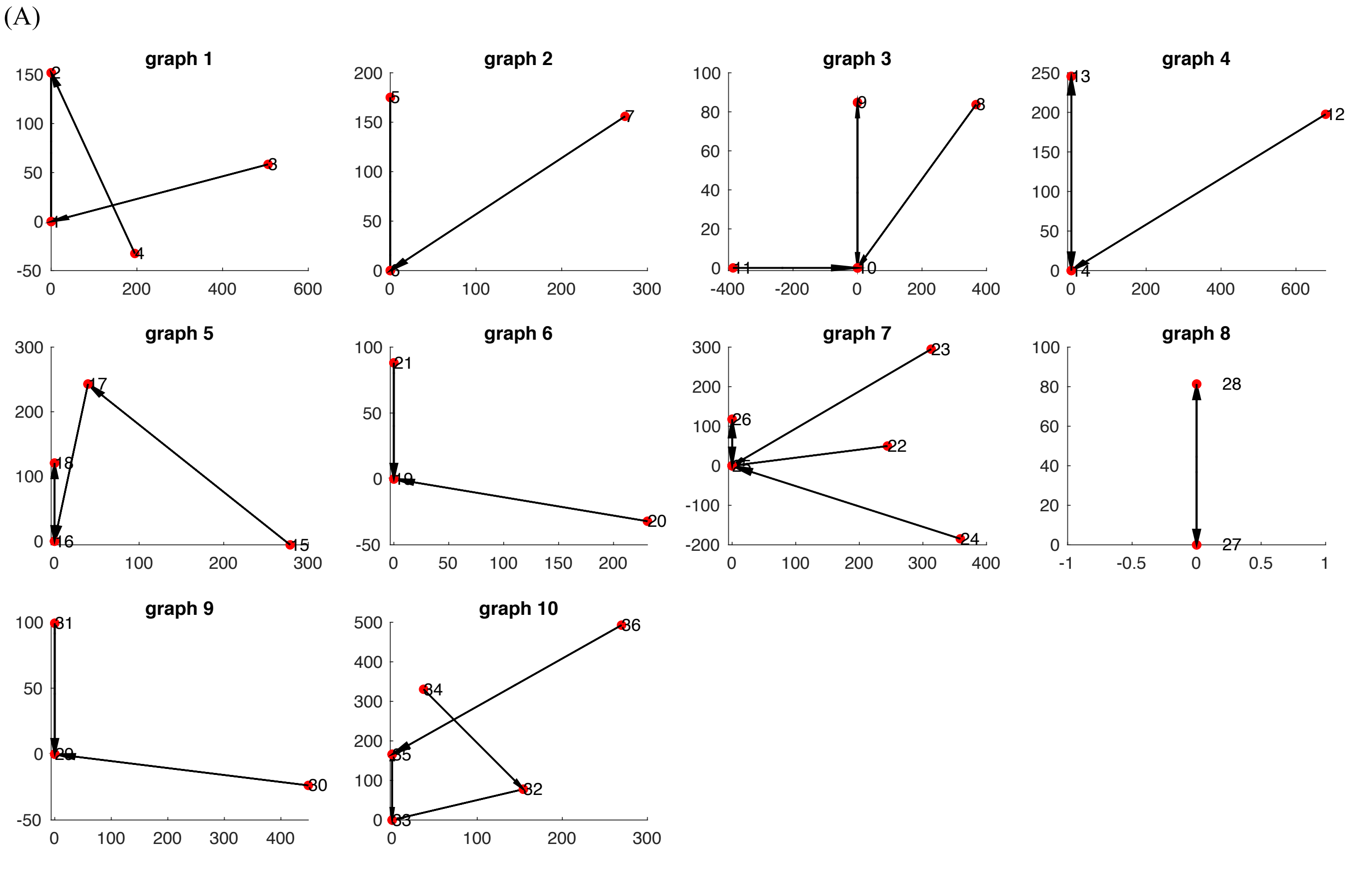


Fig. (7). Contd…

(B)

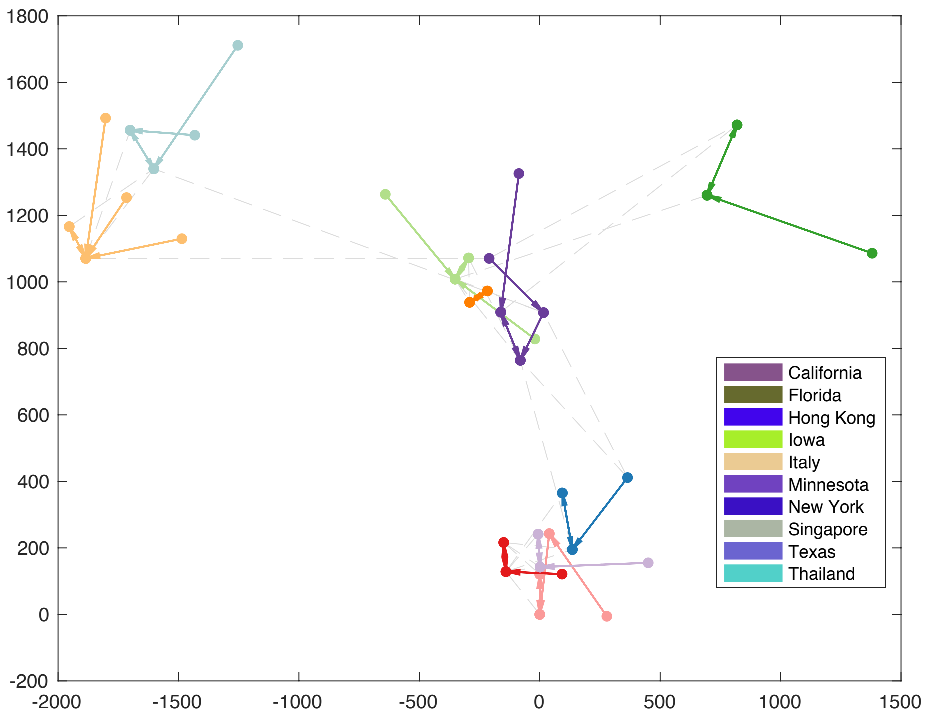


Fig. (7). (**A**) The 36 sequences are divided into 12 subgraphs, with sequence labels indicated. The corresponding district can be determined based on the sequence labels using Table **2**. The sequence accession numbers are provided in the supplementary data. (**B**) 2D-directed complete graph of the 12 districts. Different districts are distinguished by different colors. The sequence labels are omitted here. Gray dashed lines indicate that the sequences have a second or third closest relationship. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

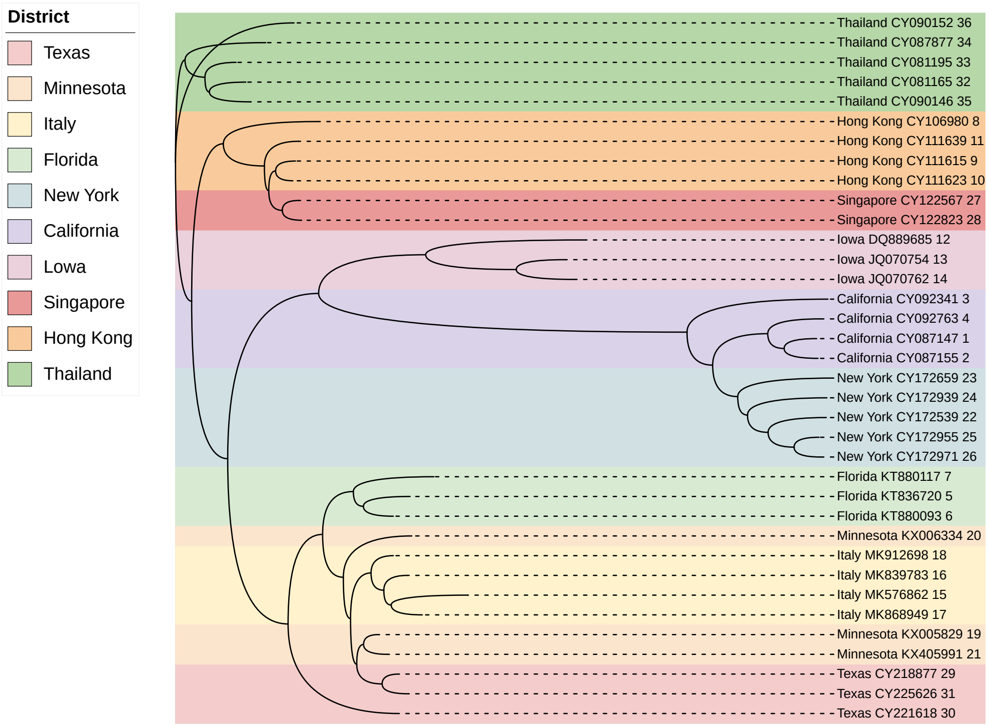


Fig. (8). The phylogenetic tree of the 36 sequences of HIV-1. The tree is based on the DisMat5 and constructed by FastME software with BioNJ algorithm. Different colors represent different districts. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

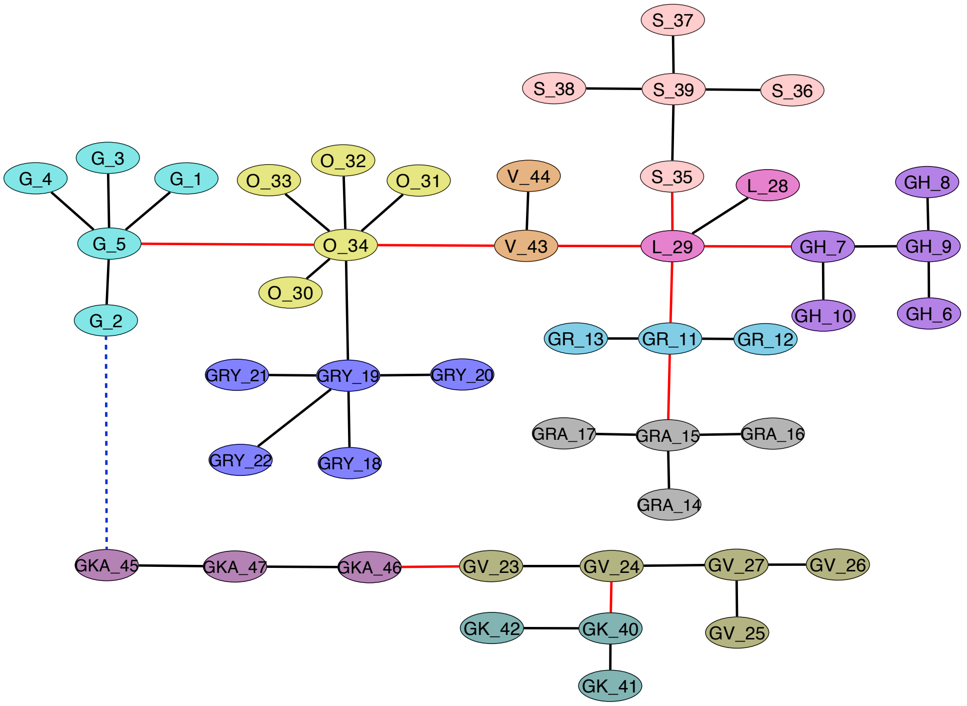


Fig. (9). Genetic network based on DisMat1 with 12 clusters representing distinct clades. Nodes named "Clade name_Sequence label". Solid black lines indicate closest distances between nodes. Solid red lines connect clusters with closest distances. Dashed blue lines link groups with closest distances. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.5. Comparison with an Existing Genetic Network

The previous proposed genetic network depicts viral genome relationships, identifies super-spreaders, and evaluates diagnostics, therapeutics, and vaccines, enabling effective control of pandemic waves and tracing transmission routes with dynamic viral genome variations [69]. It also utilizes a k-mer Natural Vector to characterize the composition and distribution of k-mers in viral sequences. The k-mer Natural Vector method considers the optimal parameter k based on the range [ceil(log4 min(L)), ceil(log4 max(L)) + 1], where L represents the lengths of viral sequences. For SARS-CoV-2 genomes, the chosen value of k is 8 (RefSeq: NC_045512, 29903 bp). For HIV-1 genomes, the chosen value of k is 7 (RefSeq: NC_001802, 9181 bp). For H1N1 NS1 sequences, the chosen value of k is 5 (It is a gene segment of segment 8 of H1N1, and the accession number of segment 8 is NC_026432. The length of the NS1 gene is 660 bp). Each sequence is uniquely represented by a k-mer Natural Vector, and sequence similarity is measured using Spearman distance. The existing genetic network algorithm [69] is as follows:

Algorithm 2: The existing genetic network algorithm

Input: The distance matrix D of sample data.

Step 1: Connect sequences if they have the shortest pairwise distance, resulting in n genetic clusters.

Step 2: Calculate the distance between genetic clusters as the mean of pairwise distances.

Step 3: Link two genetic clusters if their distance is the shortest, connecting the sequences with the shortest pairwise distance. This process creates m groups (G1-Gm) from the n genetic clusters.

Step 4: Update the distances between groups G1-Gm using step (2). Repeat this step until all sequences are connected, completing the genetic network of viral sequences.

Output: A genetic network.

We applied the above algorithm to construct genetic networks based on distance matrices DisMat1 to DisMat5, as shown in Fig. (**9**) and Supplementary Figs. (**S5** to **S8**). The network in Fig. (**9**) demonstrates excellent clustering results, with distinct clades represented by different colors. Each node represents a sequence and follows the naming convention "Clade name_Sequence label". The solid black lines indicate the closest distances between nodes, corresponding to the results obtained in Step 1 of Algorithm 2, resulting in 12 clusters precisely aligned with the 12 clades. The solid red lines represent connections between clusters, where pairs of nodes with the closest distances are linked, corresponding to Step 3. The dashed blue lines indicate connections between groups, with pairs of nodes with the closest distances linked, corresponding to Step 4. By connecting all nodes, a connected graph is obtained. In Fig. (**S5**), the genetic network constructed from DisMat2 reveals 12 clades, with clades "O" and "GRA" having two clusters each. Fig. (**S6**) displays the genetic network constructed from DisMat3, showing 12 subtypes, where subtype "02_AG" is divided into two clusters, while subtypes "01_AE" and "01B" do not exhibit separation. Similarly, in v S7, the genetic network constructed from DisMat4 demonstrates 12 subtypes, with subtypes "BF1" and "C" exhibiting two clusters each, while subtypes "A1D" and "A1" do not exhibit separation. Finally, in Fig. (**S8**), the genetic network constructed from DisMat5 illustrates 10 districts, with districts "Texas" and "Minnesota" showing no separation. The above results demonstrate that our method can achieve improved clustering outcomes. Furthermore, the networks generated by our method exhibit directional connections between nodes, which can provide more interpretable results in certain specific cases.

3.6. Comparison with Multidimensional Scaling

Classical Multidimensional Scaling (MDS) is a kind of unsupervised manifold learning algorithm, which is an approach to non-linear dimensionality reduction [70-72]. The method projects high-dimensional data into low-dimensional space, and the relative distance between each two samples remains unchanged. We compare our method with the MDS method. For sample data $Data = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} \in R^{m\times n}$, where $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$ is the i-th sample data, which has n features. Suppose the data are projected into k-dimensional space $Data' = \begin{bmatrix} y_{11} & \cdots & y_{1\mathrm{k}} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mk} \end{bmatrix}$, where $Y_i = (y_{i1}, y_{i2}, \dots, y_{ik})$ is the i-th new data, the algorithm is as follows:

Algorithm 3: Multidimensional Scaling (MDS)

Input: The distance matrix $D = \left[d_{ij}\right]_{m\times m}$ of m sample data.

Step 1: Compute inner product matrix $B = \left[b_{ij}\right]$, where $b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - \frac{1}{m}\sum_{i=1}^{m} d_{ij} - \frac{1}{m}\sum_{j=1}^{m} d_{ij} + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} d_{ij}\right)$.

Step 2: B is a positive semidefinite matrix, and its eigen-decomposition is $B = V\Lambda V^T$, where $\mathrm{V} = [v_1, v_2, \dots, v_m], \Lambda = diag([\lambda_1, \lambda_2, \dots, \lambda_m])$, the eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ and the corresponding eigenvectors are $v_1, v_2, \dots, v_m$.

Step 3: Choose the first k largest eigenvalues $\Lambda_k = diag([\lambda_1, \lambda_2, \dots, \lambda_k])$ and eigenvectors $\mathrm{V_k} = [v_1, v_2, \dots, v_k]$, reconstruct $\mathrm{V_k}\Lambda_k^{\frac{1}{2}} \triangleq \begin{bmatrix} y_{11} & \cdots & y_{1\mathrm{k}} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mk} \end{bmatrix}$.

Output: The new data $Data' = \begin{bmatrix} y_{11} & \cdots & y_{1\mathrm{k}} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mk} \end{bmatrix} \in R^{m\times k}$.

We applied the MDS method to five datasets in this study for comparison with the results obtained from the novel natural graph approach, as shown in Fig (**S9**). The MDS method has been built in Python, and we use the “sklearn.manifold” module to implement the algorithm. Due to the limitations of the dataset size, we were unable to extract meaningful information from the MDS method. However, our approach successfully clusters nodes using directed lines and provides reasonable clustering results, capable of representing phylogenetic relationships (as illustrated in Figs. **3**-**8**).

4. DISCUSSIONS

Further discussion on the presented approach and findings can shed light on its potential applications and areas for future research. Firstly, the natural graph algorithm showcased its effectiveness in clustering viral genome sequences and representing their genetic relationships. The ability to visualize distinct clades, subtypes, or districts can aid in understanding the transmission dynamics and geographic spread of viruses. Future research could focus on expanding the application of the natural graph algorithm to larger datasets, enabling analysis of broader viral populations and capturing finer-scale genetic variations. As the study's objective is to demonstrate the effectiveness and advantages of the new natural graph approach, and our method has no limit on the input of the number of sequences, expanding its application to larger datasets would necessitate careful consideration of computational resources and optimization strategies. Given the complexity of dealing with large-scale datasets, scalability could be achieved by leveraging distributed computing frameworks and parallel processing techniques. This would allow the method to effectively manage millions of genomes. Additionally, exploring the integration of other genetic features, such as single-nucleotide polymorphisms or structural variations, could enhance the resolution and accuracy of the clustering results. Furthermore, the comparison with existing methods highlighted the advantages of the natural graph approach. Compared to the existing genetic network approach, our method demonstrated improved clustering outcomes and better visualization of the viral genome relationships. When compared to MDS, our approach outperformed in terms of generating meaningful results with the given dataset sizes. By developing the new method, we address challenges that existing methods may encounter when dealing with specific datasets or applications. What sets the new natural graph as superior to others is its utilization of novel algorithms and strategies to enhance data analysis. The results indicate that the new natural graph performs more favorably, and can offer superior performance in certain datasets compared to existing methods. Its ability to generate directed connections between nodes also opens up opportunities for investigating directional transmission pathways. Our method can be scaled to virus sequences linearly (Fig. **S10**). In short, the natural graph algorithm shows promise as a valuable tool in viral genomics research, with potential applications in epidemic control, transmission tracing, and evolutionary studies. Further advancements and refinements in the methodology can contribute to a deeper understanding of viral dynamics and aid in the development of targeted interventions.

CONCLUSION

In conclusion, our study presents a novel approach for the graphical representation and clustering of viral genome sequences using a natural graph algorithm. We applied this method to three different viral datasets, including SARS-CoV-2 genomes, HIV-1 genomes, and H1N1 NS1 sequences, and compared the results with existing genetic network approaches and MDS. The natural graph algorithm successfully clustered the viral sequences into distinct subgraphs or clusters, corresponding to known clades, subtypes, or districts, as validated by the provided sequence labels. The generated 2D graphs accurately represented the genetic relationships among the sequences and the directed connections between nodes provided interpretable results. Overall, the natural graph algorithm proved to be a valuable tool for analyzing and visualizing viral genome sequences, enabling a better understanding of their genetic relationships and aiding in the identification of clades, subtypes, or districts. This approach holds promise for further research in the field of viral genomics and epidemiology.

LIST OF ABBREVIATIONS

COVID-19	=	Coronavirus Disease 2019
CRFs	=	Circulating Recombinant Forms
FM	=	Fitch-Margoliash
H1N1	=	Influenza A Virus Subtype H1N1
HIV	=	Human Immunodeficiency Virus
MDS	=	Multidimensional Scaling
ME	=	Minimum Evolution
NJ	=	Neighbor-Joining
NS1	=	Non-structural Protein 1
SARS-CoV-2	=	Severe Acute Respiratory Syndrome Coronavirus 2
UPGMA	=	Unweighted Pair Group Method with Arithmetic Mean
WPGMA	=	Weighted Pair Group Method with Arithmetic Mean

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data and supportive information are available within the article.

FUNDING

This work is supported by the National Natural Science Foundation of China (NSFC) grant (12171275) and the Tsinghua University Education Foundation fund (042202008).

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Nucleic Acid. Available from:https://www.genome.gov/genetics-glossary/Nucleic-Acids (accessed June, 2023)

[2] What is DNA. Available from:https://whatisdna.net/ (accessed June, 2023)

[3] Watson JD, Crick FHC. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953; 171(4356): 737-8. http://dx.doi.org/10.1038/171737a0 PMID: 13054692

[4] Sun N, Pei S, He L, Yin C, He RL, Yau SST. Geometric construction of viral genome space and its applications. Comput Struct Biotechnol J 2021; 19: 4226-34. http://dx.doi.org/10.1016/j.csbj.2021.07.028 PMID: 34429843

[5] Yu C, 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. http://dx.doi.org/10.1016/j.jtbi.2012.11.005 PMID: 23154188

[6] Deng M, Yu C, Liang Q, He RL, Yau SST. A novel method of characterizing genetic sequences: Genome space with biological distance and applications. PLoS One 2011; 6(3): e17293. http://dx.doi.org/10.1371/journal.pone.0017293 PMID: 21399690

[7] Training E-E. What is genetic variation. Available from:https://www.ebi.ac.uk/training/online/course/human-genetic-variation-i-introduction/what-genetic-variation (accessed June, 2023)

[8] Genetic Variation. Available from: https://www.genome.gov/genetics-glossary/Genomic-Variation (accessed June, 2023)

[9] Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science 2006; 311(5765): 1283-7. http://dx.doi.org/10.1126/science.1123061

[10] Wolf YI, Rogozin IB, Grishin NV, Koonin EV. Genome trees and the tree of life. Trends Genet 2002; 18(9): 472-9. http://dx.doi.org/10.1016/S0168-9525(02)02744-0 PMID: 12175808

[11] Tavassoly I, Goldfarb J, Iyengar R. Systems biology primer: The basic methods and approaches. Essays Biochem 2018; 62(4): 487-500. http://dx.doi.org/10.1042/EBC20180003 PMID: 30287586

[12] Baitaluk M. System biology of gene regulation. Methods Mol Biol 2009; 569: 55-87.

[13] Wen J, Zhang Y, Yau SST. k-mer Sparse matrix model for genetic sequence and its applications in sequence comparison. J Theor Biol 2014; 363: 145-50.
http://dx.doi.org/10.1016/j.jtbi.2014.08.028 PMID: 25158165

[14] Vinje H, Liland KH, Almøy T, Snipen L. Comparing K-mer based methods for improved classification of 16S sequences. BMC Bioinformatics 2015; 16(1): 205.
http://dx.doi.org/10.1186/s12859-015-0647-4 PMID: 26130333

[15] Bohnsack KS, Kaden M, Abel J, Villmann T. Alignment-free sequence comparison: A systematic survey from a machine learning perspective. IEEE/ACM Trans Comput Biol Bioinformatics 2022; 1.
http://dx.doi.org/10.1109/TCBB.2022.3140873

[16] Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method. BMC Evol Biol 2007; 7(1): 41.
http://dx.doi.org/10.1186/1471-2148-7-41 PMID: 17359548

[17] Wang Y, Hill K, Singh S, Kari L. The spectrum of genomic signatures: From dinucleotides to chaos game representation. Gene 2005; 346: 173-85.
http://dx.doi.org/10.1016/j.gene.2004.10.021 PMID: 15716010

[18] Cheng J, Zeng X, Ren G, Liu Z. CGAP: A new comprehensive platform for the comparative analysis of chloroplast genomes. BMC Bioinformatics 2013; 14(1): 95.
http://dx.doi.org/10.1186/1471-2105-14-95 PMID: 23496817

[19] Ondov BD, Treangen TJ, Melsted P, *et al.* Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biol 2016; 17(1): 132.
http://dx.doi.org/10.1186/s13059-016-0997-x PMID: 27323842

[20] Ondov BD, Starrett GJ, Sappington A, *et al.* Mash Screen: High-throughput sequence containment estimation for genome discovery. Genome Biol 2019; 20(1): 232.
http://dx.doi.org/10.1186/s13059-019-1841-x PMID: 31690338

[21] 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(1): 25-34.
http://dx.doi.org/10.1016/j.gene.2014.05.043 PMID: 24858075

[22] Zhang Y, Wen J, Yau SST. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. Genomics 2019; 111(6): 1298-305.
http://dx.doi.org/10.1016/j.ygeno.2018.08.010 PMID: 30195069

[23] Sun N, Yang J, Yau SST. Identification of HIV rapid mutations using differences in nucleotide distribution over time. Genes 2022; 13(2): 170.
http://dx.doi.org/10.3390/genes13020170 PMID: 35205215

[24] Zhao X, Tian K, He RL, Yau SST. Convex hull principle for classification and phylogeny of eukaryotic proteins. Genomics 2019; 111(6): 1777-84.
http://dx.doi.org/10.1016/j.ygeno.2018.11.033 PMID: 30529533

[25] Huang HH, Yu C, Zheng H, *et al.* Global comparison of multiple-segmented viruses in 12-dimensional genome space. Mol Phylogenet Evol 2014; 81: 29-36.
http://dx.doi.org/10.1016/j.ympev.2014.08.003 PMID: 25172357

[26] Yu C, Liang Q, Yin C, He RL, Yau SST. A novel construction of genome space with biological geometry. DNA Res 2010; 17(3): 155-68.
http://dx.doi.org/10.1093/dnares/dsq008 PMID: 20360268

[27] Li Y, Tian K, Yin C, He RL, Yau SST. Virus classification in 60-dimensional protein space. Mol Phylogenet Evol 2016; 99: 53-62.
http://dx.doi.org/10.1016/j.ympev.2016.03.009 PMID: 26988414

[28] Fang M, Xu J, Sun N, Yau SS-T. Generating minimal models of H1N1 NS1 gene sequences using alignment-based and alignment-free algorithms. Genes 2023; 14(1): 186.
http://dx.doi.org/10.3390/genes14010186

[29] Yu C. Real time classification of viruses in 12 Dimensions. Plos one 2013; 8: e64328.
http://dx.doi.org/10.1371/journal.pone.0064328

[30] Tian K, Yang X, Kong Q, Yin C, He RL, Yau SST. Two dimensional Yau-Hausdorff distance with applications on comparison of DNA and protein sequences. PLoS One 2015; 10(9): e0136577.
http://dx.doi.org/10.1371/journal.pone.0136577 PMID: 26384293

[31] Dong R, Zhu Z, Yin C, He RL, Yau SST. A new method to cluster genomes based on cumulative Fourier power spectrum. Gene 2018; 673: 239-50.
http://dx.doi.org/10.1016/j.gene.2018.06.042 PMID: 29935353

[32] Pei S, Dong W, Chen X, He RL, Yau SST. Fast and accurate genome comparison using genome images: The Extended Natural Vector Method. Mol Phylogenet Evol 2019; 141: 106633.
http://dx.doi.org/10.1016/j.ympev.2019.106633 PMID: 31563612

[33] Sun N, Zhao X, Yau SST. An efficient numerical representation of genome sequence: Natural vector with covariance component. PeerJ 2022; 10: e13544.
http://dx.doi.org/10.7717/peerj.13544 PMID: 35729905

[34] Dong R, Pei S, Guan M, *et al.* Full chromosomal relationships between populations and the origin of humans. Front Genet 2022; 12: 828805.
http://dx.doi.org/10.3389/fgene.2021.828805 PMID: 35186019

[35] Sokal M. A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin 1958; 38: pp. 1409-38.

[36] Garcia-Vallvé S, Puigbo P. DendroUPGMA: A dendrogram construction utility. Universitat Rovira i Virgili 2009; pp. 1-14.

[37] Murtagh F. Complexities of hierarchic clustering algorithms: State of the art. Comput Stat Quarterly 1984; 1(2): 101-13.

[38] Olsen GJ. Phylogenetic analysis using ribosomal RNA. In: Methods in enzymology. Elsevier 1988; 164: pp. 793-812.

[39] Erdmann V A, Wolters J. Collection of published 5S, 5.8 S and 4.5 S ribosomal RNA sequences. Nucleic Acids Res 1986; 14(1): 1.

[40] Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 1987; 4(4): 406-25.
PMID: 3447015

[41] Mihaescu R, Levy D, Pachter L. Why neighbor-joining works. Algorithmica 2009; 54(1): 1-24.
http://dx.doi.org/10.1007/s00453-007-9116-4

[42] Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 1994; 11(3): 459-68.
PMID: 8015439

[43] Kidd KK, Sgaramella-Zonta LA. Phylogenetic analysis: Concepts and methods. Am J Hum Genet 1971; 23(3): 235-52.
PMID: 5089842

[44] Catanzaro D. The minimum evolution problem: Overview and classification. Networks 2009; 53(2): 112-25.
http://dx.doi.org/10.1002/net.20280

[45] Rzhetsky A, Nei M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 1993; 10(5): 1073-95.
PMID: 8412650

[46] Fitch WM, Margoliash E. Construction of phylogenetic trees. Science 1967; 155(3760): 279-84.
http://dx.doi.org/10.1126/science.155.3760.279 PMID: 5334057

[47] Saitou N, Imanishi T. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evolu 1989; 6(5): 514.

[48] Leitner T, Escanilla D, Franzén C, Uhlén M, Albert J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc Natl Acad Sci 1996; 93(20): 10864-9.
http://dx.doi.org/10.1073/pnas.93.20.10864 PMID: 8855273

[49] Sullivan J, Joyce P. Model selection in phylogenetics. Annu Rev Ecol Evol Syst 2005; 36(1): 445-66.
http://dx.doi.org/10.1146/annurev.ecolsys.36.102003.152633

[50] Pol D. Empirical problems of the hierarchical likelihood ratio test for model selection. Syst Biol 2004; 53(6): 949-62.
http://dx.doi.org/10.1080/10635150490888868 PMID: 15764562

[51] Abadi S, Azouri D, Pupko T, Mayrose I. Model selection may not be a mandatory step for phylogeny reconstruction. Nat Commun 2019; 10(1): 934.
http://dx.doi.org/10.1038/s41467-019-08822-w PMID: 30804347

[52] Noureddine FY, Chakkour M, El Roz A, *et al.* The emergence of SARS-CoV-2 variant (s) and its impact on the prevalence of COVID-19 cases in the Nabatieh Region, Lebanon. Med Sci 2021; 9(2): 40.
http://dx.doi.org/10.3390/medsci9020040 PMID: 34199617

[53] Alm E, Broberg EK, Connor T, *et al.* Geographical and temporal distribution of SARS-CoV-2 clades in the WHO European Region, January to June 2020. Euro Surveill 2020; 25(32): 2001410.

http://dx.doi.org/10.2807/1560-7917.ES.2020.25.32.2001410 PMID: 32794443

[54] GISAID - hCov19 Variants. Available from:https://gisaid.org/hcov19-variants/ (accessed June, 2023)

[55] GISAID. Clade tree. Available from:https://www.gisaid.org/fileadmin/c/gisaid/files/images/clade_tree.jpg (accessed June, 2023)

[56] Zhukova A, Blassel L, Lemoine F, Morel M, Voznica J, Gascuel O. Origin, evolution and global spread of SARS-CoV-2. C R Biol 2021; 344(1): 57-75. http://dx.doi.org/10.5802/crbiol.29 PMID: 33274614

[57] Lefort V, Desper R, Gascuel O. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol 2015; 32(10): 2798-800. http://dx.doi.org/10.1093/molbev/msv150 PMID: 26130081

[58] Gascuel O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 1997; 14(7): 685-95. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025808 PMID: 9254330

[59] Gilbert PB, McKeague IW, Eisen G, *et al.* Comparison of HIV-1 and HIV-2 infectivity from a prospective cohort study in Senegal. Stat Med 2003; 22(4): 573-93. http://dx.doi.org/10.1002/sim.1342 PMID: 12590415

[60] Douek DC, Roederer M, Koup RA. Emerging concepts in the immunopathogenesis of AIDS. Annu Rev Med 2009; 60(1): 471-84. http://dx.doi.org/10.1146/annurev.med.60.041807.123549 PMID: 18947296

[61] Shankarappa R, Margolick JB, Gange SJ, *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol 1999; 73(12): 10489-502. http://dx.doi.org/10.1128/JVI.73.12.10489-10502.1999 PMID: 10559367

[62] Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. AIDS 2006; 20(16): W13-23. http://dx.doi.org/10.1097/01.aids.0000247564.73009.bc PMID: 17053344

[63] Smith DM, Richman DD, Little SJ. HIV Superinfection. J Infect Dis 2005; 192(3): 438-44. http://dx.doi.org/10.1086/431682 PMID: 15995957

[64] Sun N, Yau SS-T. In-depth investigation of the point mutation pattern of HIV-1. Front Cell Infect Microbiol 2022; 12: 1033481. http://dx.doi.org/10.3389/fcimb.2022.1033481

[65] Krammer F, Smith GJD, Fouchier RAM, *et al.* Influenza. Nat Rev Dis Primers 2018; 4(1): 3. http://dx.doi.org/10.1038/s41572-018-0002-y PMID: 29955068

[66] Sautto GA, Kirchenbaum GA, Ross TM. Towards a universal influenza vaccine: Different approaches for one goal. Virol J 2018; 15(1): 17. http://dx.doi.org/10.1186/s12985-017-0918-y PMID: 29370862

[67] Eisfeld AJ, Neumann G, Kawaoka Y. At the centre: Influenza A virus ribonucleoproteins. Nat Rev Microbiol 2015; 13(1): 28-41. http://dx.doi.org/10.1038/nrmicro3367 PMID: 25417656

[68] Goka E A, Vallely P J, Mutton K J, Klapper P E. Mutations associated with severity of the pandemic influenza A(H1N1)pdm09 in humans: A systematic review and meta-analysis of epidemiological evidence. Arch Virol 2014; 159(12): 3167-83. http://dx.doi.org/10.1007/s00705-014-2179-z

[69] Zhang Y, Wen J, Xi K, Pan Q. Exploring the dynamic variations of viral genomes *via* a novel genetic network. Mol Phylogenet Evol 2022; 175: 107583. http://dx.doi.org/10.1016/j.ympev.2022.107583 PMID: 35810971

[70] Chen C-h, Härdle W, Unwin A, Cox MA, Cox TF. Multidimensional scaling. In: Handbook of data visualization. Berlin, Heidelberg: Springer 2008.

[71] Gordon A. The User's Guide to Multidimensional Scaling, with Special Reference to the Mds (X) Library of Computer Programs. Wiley 1983. http://dx.doi.org/10.2307/2987947

[72] Green PE. Marketing Applications of MDS: Assessment and Outlook: After a decade of development, what have we learned from MDS in marketing? J Mark 1975; 39(1): 24-31. http://dx.doi.org/10.1177/002224297503900105

DISCLAIMER: The above article has been published, as is, ahead-of-print, to provide early visibility but is not the final version. Major publication processes like copyediting, proofing, typesetting and further review are still to be done and may lead to changes in the final published version, if it is eventually published. All legal disclaimers that apply to the final published article also apply to this ahead-of-print version.