JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 33, Number 2, 2026
Mary Ann Liebert, A Part of Sage
© Mary Ann Liebert (NY), LLC
Pp. 255–266
DOI: 10.1177/15578666251391211

Multi-Perspective Natural Vector: A Novel Method for Viral Sequence Feature Extraction

XIANG SHI,[1] JIAYI KANG,[2,3] NAN SUN,[2–4] XIN ZHAO,[2] and STEPHEN S.-T. YAU[1–3]

ABSTRACT

The rapid expansion of biological data in recent decades has highlighted the need for efficient methods in sequence analysis. Traditional pairwise alignment approaches are both time-consuming and memory-intensive. Alignment-free (AF) methods such as natural vector (NV) and k-mer operate on a one-dimensional framework, interpreting DNA primarily as a linear string of nucleotides. To achieve a more comprehensive interpretation of molecular structure, this study incorporates the three-dimensional architectural features of DNA and introduces a novel AF method named Multi-perspective natural vector (MNV). The MNV method maps genome sequences of varying lengths to points within a unified geometric space, facilitating large-size data processing tasks such as variant classification and clustering. Across datasets of different sizes and types, MNV attains a 100% convex hull separation ratio in lower dimensions compared with widely used methods NV and k-mer methods. In neural network classification, MNV achieves better classification accuracy of 99.55% and 98.78% on SARS-CoV-2 and poliovirus datasets respectively, demonstrating its effectiveness in viral genome analysis while maintaining computational efficiency.

Keywords: genome, natural vector, sequence classification, virus.

1. INTRODUCTION

The molecular-level study of pathogenic viruses is essential for deciphering the mechanisms underlying their pathogenicity. Since the 1980s, the advent of gene sequencing technologies has catalyzed an exponential growth in available genome sequences. This surge has spurred the development of alignment-based sequence analysis tools, including BLAST (Altschul et al., 1997), FASTA (Pearson and Lipman, 1988), MUSCLE (Edgar, 2004), and ClustalW (Larkin et al., 2007). Since these methods leverage the sequence collinearity assumption that homologous sequences contain linearly arranged conserved regions, they face critical limitations in practical applications (Zielezinski et al., 2017). Viral genomes, characterized by high mutation rates and extensive variations in base composition and arrangement (Duffy, 2018), frequently violate collinearity assumptions. Moreover, alignment-based approaches suffer from computational inefficiency: the multiple sequence alignment problem is Non-deterministic Polynomial (NP)-hard without approximation (Just,

[1] Department of Mathematical Sciences, Tsinghua University, Beijing, P. R. China.
[2] Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing, P. R. China.
[3] Hetao Institute of Mathematics and Interdisciplinary Sciences (HIMIS), Shenzhen, P. R. China.
[4] Yau Mathematical Sciences Center, Tsinghua University, Beijing, P. R. China.

2001), and alignment complexity increases rapidly with the length of sequence (Zielezinski et al., 2017). Even when alignments are available, the absence of numerical sequence representations still forces pairwise comparisons to classify novel sequences, limiting large-scale genomic analysis (Bohnsack et al., 2023).

In response to these challenges, alignment-free (AF) methods have emerged as powerful alternatives, particularly for sequences that are large in size or exhibit complex relationships between sequence regions (Hu et al., 2019). AF methods convert biological sequences into numerical embeddings, facilitating rapid similarity computations between multiple sequences. A prominent example is the k-mer approach, which maps nucleotide sequences to word sequences over the nucleotide alphabet and then calculates the frequency of all k-length words (Blaisdell, 1991). However, we notice that k-mer methods sometimes fail to distinguish between sequences with minor mutational differences. The natural vector (NV) (Deng et al., 2011) method proposed by Yau et al. considers higher-order statistical information to extract sequence features. NV embeds sequences in a Euclidean space based on nucleotide distribution statistics. The original 12-dimensional NV incorporates nucleotide counts, average position, and second-order moments, with extensions to higher dimensions through additional central moments. NV is effective in sequence comparison and classification. Both k-mer and NV approaches represent DNA sequences solely as ordered linear strings of nucleotides, treating them as one-dimensional symbolic sequences. However, this abstraction fails to capture the structural and topological properties of DNA molecules in their native three-dimensional conformation.

The three-dimensional architecture of DNA is critical to understanding its biological properties. Studies have demonstrated the periodicity of approximately 10–11 base pairs in eukaryotic sequences and prokaryotic coding sequences (Zhurkin, 1981), which is thought to arise from the physical properties of the DNA chain, including the dynamics of helix folding and amphipathic interactions of the α helix in the corresponding protein sequences (Herzel et al., 1999). To better capture such structural periodicities, we propose the Multi-perspective Natural Vector (MNV) method. By integrating trigonometric functions directly into the NV framework, MNV explicitly incorporates the three-dimensional architectural features of DNA, enabling a more semantically informative encoding of viral genome sequences.

In this study, we present MNV as a novel AF method for precise viral classification and clustering at family and subtype levels. Our approach demonstrates superior performance compared to traditional NV and k-mer methods across diverse datasets, establishing a robust framework for large-scale viral genome analysis.

2. METHODS

2.1. Traditional natural vector

A DNA sequence is made up of nucleotides, each consisting of three components: a nitrogenous base, a pentose sugar, and a phosphate group. The nitrogenous bases include adenine (A), guanine (G), cytosine (C), and thymine (T) (Watson and Crick, 1953).

The NV is a $(4+4m)$ dimensional numerical coding of nucleotide sequences defined as follows. For a DNA sequence of length n, S is composed of nucleotides $s_1, s_2, \ldots, s_n$ arranged sequentially, where $s_i \in L = \{A, C, G, T\}, i = 1, 2, \ldots, n$. Define the indicator functions $\omega_k : L \to \{0, 1\}$,

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

where $s_i \in L, i = 1, 2, \ldots, n$ and $k \in L$.

The count of nucleotide k in the sequence S is:

$$D_k^0 = n_k = \sum_{i=1}^{n} \omega_k(s_i).$$

The average location of nucleotide k is:

$$D_k^1 = \mu_k = \sum_{i=1}^{n} i \frac{\omega_k(s_i)}{n_k}.$$

The central moments from the second to the m-th order are:

$$D_k^j = \sum_{i=1}^{n} \frac{(i-\mu_k)^j \omega_k(s_i)}{n_k^{j-1} n^{j-1}}, \quad j = 2, \ldots, m.$$

The denominator is designed to ensure the convergence of central moments as j approaches infinity (Deng et al., 2011).

By concatenating the values calculated above, we obtain the $(4+4m)$ dimensional NV of the sequence S. For example, when $m=2$, the 12-dimensional NV is

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_A^2, D_C^2, D_G^2, D_T^2)$$

In numerical experiments, lower-order expansions of the NV method are commonly used, with second-order moments being the most frequently applied.

2.2. *Multi-perspective natural vector*

The traditional NV method utilizes polynomials to capture sequence features. Although incorporating higher-order moments can enhance the NV method's classification performance, it significantly increases computational complexity. Consequently, two principal research directions have been explored to further advance the capabilities of lower-order NV methods:

- Transforming the sequence into its k-mer representation and calculating the NV of the k-mer representation (Wen et al., 2014).
- Extending the moment-based approach, such as incorporating covariance components into NV (Sun et al., 2022).

Differently, in this study, we introduce a novel conceptual pathway by integrating trigonometric moments based on $\sin(x)$ and $\cos(x)$ functions into the NV framework. This Multi-perspective NV (MNV) method is inspired by the three-dimensional double-helix structure of DNA.

DNA is a double-helix polymer, consisting of two complementary strands wound around each other in a spiral (Watson and Crick, 1953). Geometrically, a single strand can be modeled as a regular cylindrical helix, parameterized as follows:

$$\begin{cases} x(t) = r\cos(t) \\ y(t) = r\sin(t) \\ z(t) = ct \end{cases}$$

where r and c are constants. For illustration, setting r = 1 and c = 1 yields the helix depicted in Figure 1. This figure serves as a conceptual representation of the MNV framework. At key points ($\theta = \pi/2, \pi, 3\pi/2, 2\pi$), the MNV embedding comprises both the z-coordinate projection (encoding sequential positions in traditional NV) and the new XY-projection (incorporating structural context). This geometric analogy underscores the motivation behind our method: just as the helical structure of DNA combines linear and rotational information, our trigonometric moment-based MNV captures both sequential and structural features in an integrated manner.

Although certain viruses utilize RNA as their genetic material, the universality of DNA's role in biological information processing is declared in the central dogma, from the original formulation (Crick, 1958) to its refined framework (Crick, 1970). Therefore, our proposed MNV framework offers a unified mathematical representation for analyzing all types of biological sequences.

MNV is implemented through an expansion of the function space that combines both polynomial and trigonometric components. Crucially, these trigonometric functions exhibit transcendental properties that make their moment representations fundamentally irreducible to finite polynomial expansions.

$$\{1, x, x^2, \ldots, x^m, \sin(x), \cos(x)\}$$

In the new function space, we introduce a trigonometric version of the mean and moments of the NV. To address potential confounding effects from heterogeneous sequence lengths during comparative analyses, we

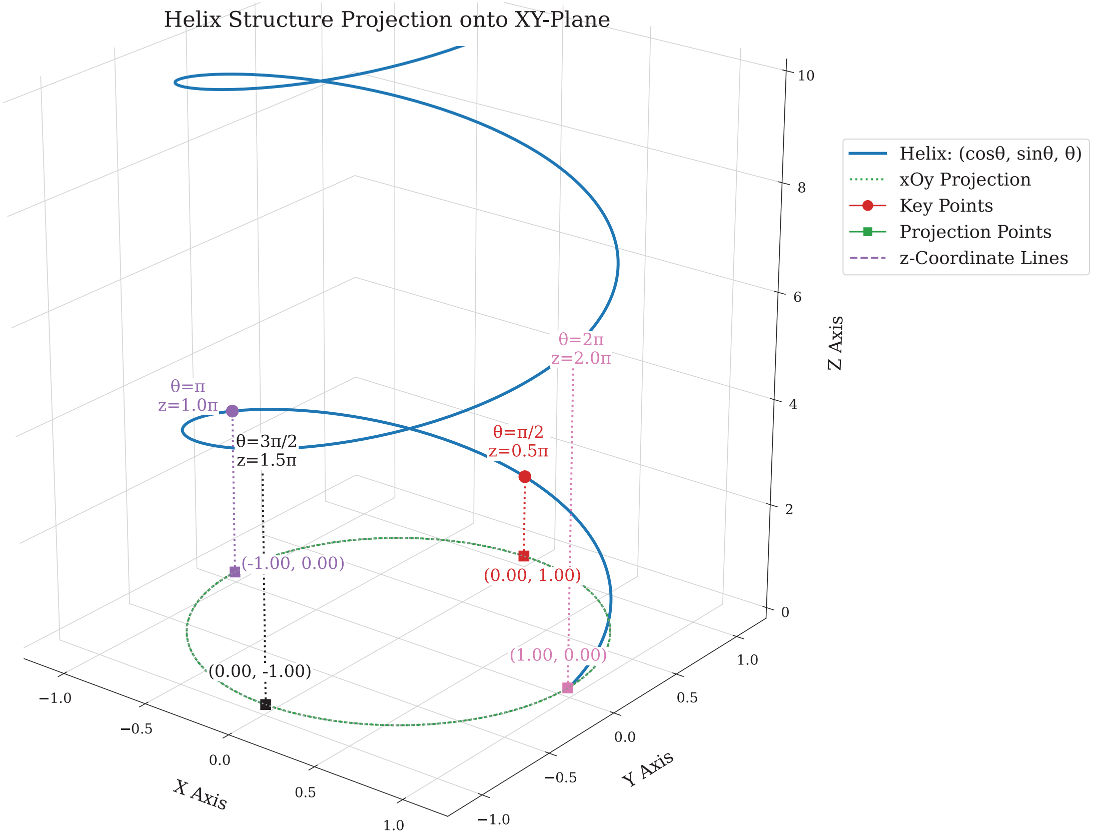

FIG. 1. Conceptual Helix Model for Multi-perspective Embedding.

implement a normalization protocol that projects all genetic sequences onto a standardized trigonometric domain spanning the interval $[0, 2\pi]$ through linear length scaling. For a DNA sequence of length n, the complete MNV embedding is derived by augmenting NV with the following components:

$$Pos_n(i) = \frac{2\pi i}{n}$$

$$D_k^{1\,\cos} = \mu_k^{\cos} = \sum_{i=1}^{n} \cos(Pos_n(i)) \frac{\omega_k(s_i)}{n_k}$$

$$D_k^{1\,\sin} = \mu_k^{\sin} = \sum_{i=1}^{n} \sin(Pos_n(i)) \frac{\omega_k(s_i)}{n_k}$$

$$D_k^{j\,\cos} = \sum_{i=1}^{n} \frac{(\cos(Pos_n(i)) - \mu_k^{\cos})^j \omega_k(s_i)}{n_k}, \; j=2, \ldots, m$$

$$D_k^{j\,\sin} = \sum_{i=1}^{n} \frac{(sin(Pos_n(i)) - \mu_k^{sin})^j \omega_k(s_i)}{n_k}, \; j=2, \ldots, m$$

Finally, we map the original sequence to a $(4+12m)$ dimensional NV. If $m=2$, the 20-dimensional MNV is

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_A^{\cos}, \mu_A^{\sin}, \mu_C, \mu_C^{\cos}, \mu_C^{\sin}, \mu_G, \mu_G^{\cos}, \mu_G^{\sin}, \mu_T, \mu_T^{\cos}, \mu_T^{\sin},$$

$$D_A^2, D_A^{2\,\cos}, D_A^{2\,\sin}, D_C^2, D_C^{2\,\cos}, D_C^{2\,\sin}, D_G^2, D_G^{2\,\cos}, D_G^{2\,\sin}, D_T^2, D_T^{2\,\cos}, D_T^{2\,\sin})$$

TABLE 1. SARS-CoV-2 DATASET

Variants	*Pango lineage*	*Number of sequences*
Alpha	B.1.1.7	54197
Beta	B.1.351	353
Delta	B.1.617.2	4909
Gamma	P.1	5040
Omicron	B.1.1.529	1017

The advantage of the updated version of mean and moments is that they are insensitive to the position of variations. In other words, a single nucleotide variation occurring at the first position or the last position of a sequence does not introduce significant differences. As a result, MNV focuses more on the types and quantities of variations than on the absolute position of mutations. By incorporating trigonometric functions, MNV also introduces periodic components to the embedding.

2.3. *Convex hull classification method*

A convex hull of a given set of points is defined as the smallest convex set that contains all the points. For a finite point set $\mathcal{A} = \{a_1, a_2, \ldots, a_s\}$ in the Euclidean space $\mathbb{R}^d$, the convex hull of $\mathcal{A}$ is the smallest convex polygon $(d=2)$ or polyhedron (in higher dimensions) that encloses all the points in A:

$$\mathrm{Cov}\mathcal{A} = \{\lambda_1 a_1 + \lambda_2 a_2 + \ldots + \lambda_s a_s : a_i \in \mathcal{A}, \lambda_1 + \lambda_2 + \ldots + \lambda_s = 1, \lambda_i \geq 0, i = 1, 2, \ldots, s\}$$

The intersection between the convex hull of $\mathcal{A} = a_1, a_2, \ldots, a_s$ and the convex hull of another set $\mathcal{B} = b_1, b_2, \ldots, b_t$ implies that there exist coefficients λ_i and μ_j such that:

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

The convex hull principle asserts that convex hulls corresponding to different families or variants are pairwise disjoint (Tian et al., 2018). Checking whether two convex hulls intersect is equivalent to solving a feasibility problem to find a set of constants satisfying the constraints in (1).

By computing convex hulls for a dataset of genetic variants, we can establish decision boundaries in the Euclidean space that enable the precise classification of novel sequences. Additionally, convex hull analysis offers a novel method for identifying new sequences that belong to a biological group by examining possible points contained within the group's convex hull.

2.4. *Clustering method*

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) is a widely used clustering technique known for its simplicity and efficiency in identifying clusters in complex datasets. We assessed the quality of DBSCAN clusters through three quantitative indices: ARI, NMI, and FMI.

TABLE 2. POLIOVIRUS DATASET

Serotypes	*Number of sequences*
Poliovirus 1	225
Poliovirus 2	712
Poliovirus 3	210

Table 3. Virus Family Convex Hull Classification Results

Methods	*Dimension*	*Disjoint ratio*
2mer	16	0.9905
3mer	64	0.9999
NV	124	0.9949
MNV	124	1
4mer	256	1

The adjusted Rand Index (ARI) is a normalized version of the Rand Index, designed to measure the degree of agreement between two partitions (Hubert and Arabie, 1985). Given two partitions $A = A_1, A_2, \ldots, A_r$ and $B = B_1, B_2, \ldots, B_s$ of n objects, the calculation formula of ARI is defined as follows:

$$ARI = \frac{\sum_i \binom{n_{ij}}{2} - \frac{\left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right]}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] - \frac{\left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right]}{\binom{n}{2}}}$$

where n_{ij} is the number of observations in $A_i \cap B_j$, $n_i = \sum_j n_{ij}$, and $n_j = \sum_i n_{ij}$. Suppose that A is the ground truth class assignment and B is the clustering result. A higher absolute value of the ARI indicates better clustering performance.

Normalized mutual information (NMI) (Fred and Jain, 2005) is an information-theoretical metric that quantifies the information shared between two data distributions. It is based on entropy, which in information theory measures the amount of information contained in a distribution. For two clusters U and V with label assignments, their corresponding entropy $H(U)$ and $H(V)$ are defined as follows:

$$H(U) = -\sum_{i=1}^{|U|} P(i) \log\left(P(i)\right)$$

$$H(V) = -\sum_{j=1}^{|V|} P'(j) \log\left(P'(j)\right)$$

$$I(U, V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i, j) \log\left(\frac{P(i, j)}{P(i)P'(j)}\right)$$

where $P(i, j) = \mid U_i \cap V_j \mid / N$ represents the probability that an object picked at random falls into both classes U_i and V_j. The NMI is defined as follows. A higher value indicates a better similarity between the two clusters.

$$NMI(U, V) = \frac{2I(U, V)}{H(U) + H(V)}$$

Table 4. SARS-CoV-2 Variant Convex Hull Classification Results

Methods	*Dimension*	*Disjoint ratio*
2mer	16	0.1000
NV	52	0.7000
MNV	52	1
3mer	64	1

Table 5. Poliovirus Subtype Classification Results

Methods	*Dimension*	*Disjoint ratio*
2mer	16	0.3333
NV	28	0.6667
MNV	28	1
3mer	64	1

Fowlkes–Mallows Index (FMI) (Fowlkes and Mallows, 1983) is defined to determine the similarity between two clusterings.

$$FMI = \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$$

FMI value ranges from 0 to 1. A higher value indicates a better similarity between the two clusters.

2.5. Ethical statement

This study did not involve human participants or animal experiments. All viral genome data analyzed in this work were obtained from publicly available databases. The Institutional Review Board (IRB) requirement was waived, as the data are de-identified and open-access.

3. DATASETS

The MNV method was tested on three types of data.

Virus Genome Dataset We downloaded all reference sequences of virus genomes from the National Center for Biotechnology Information (NCBI), specifically from the index of /refseq/release/viral up to 30 May 2024. To ensure data reliability, we filtered out sequences that met any of the following criteria:

(1) sequences lacking family taxonomic information.

(2) sequences that contain undefined nucleotides other than A, C, G, and T.

(3) sequences in the family that have fewer than three sequences.

Following filtration, the final dataset comprises 10,652 viral genome sequences from 169 taxonomic families.

SARS-CoV-2 Genome Dataset Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)-, the causative agent of COVID-19, is a single-stranded positive-sense RNA virus in the Coronaviridae family. A SARS-CoV-2 sequence is approximately 30,000 bases in length. We downloaded all SARS-CoV-2 genome sequences of the five most concerning variants from NCBI up to August 14, 2024, including a total of 65,516 sequences (see Table 1).

Poliovirus Genome Datasets Poliovirus, the causative agent of poliomyelitis (polio), is a single-strand positive-sense RNA virus in the family of Picornaviridae. It comprises three serotypes, numbered 1, 2, and 3. A poliovirus sequence is about 7500 bases in length. We downloaded all available genome sequences of poliovirus from NCBI up to May 30, 2024, including a total of 1147 sequences in the three serotypes (see Table 2).

Table 6. SARS-CoV-2 Variant Clustering and Classification Results

Methods	*ARI*	*NMI*	*FMI*	*Classification accuracy*
2mer	0.0108	0.0078	0.8348	0.9796
52d-NV	0.0291	0.0238	0.8364	0.9898
52d-MNV	**0.6858**	**0.6037**	**0.9182**	**0.9955**
3mer	0.6293	0.5157	0.8625	0.9949

Best results are shown in bold.

Table 7. Poliovirus Subtype Clustering and Classification Results

Methods	*ARI*	*NMI*	*FMI*	*Classification accuracy*
2mer	0.0508	0.0463	0.6758	0.9730
28d-NV	0.3150	0.2637	0.6828	0.9817
28d-MNV	**0.4535**	**0.4300**	**0.7489**	**0.9878**
3mer	0.1009	0.3320	0.5019	0.9869

Best results are shown in bold.

4. RESULTS

Codes for the experiments are available at https://github.com/xiangShi22/Multi-Perspective-Natural-Vector.

4.1. Convex hull classification

For each dataset—where labels correspond to a specific taxonomic level such as family or variant—we calculated the $(4+12m)$ dimensional MNV for each sequence for m from 2 to 4, and then constructed a convex hull for each class. Next, we employed the linear programming method to determine whether the convex hulls of different classes were pairwise disjoint. For comparison, embeddings based on traditional NV and k-mer representations were also generated and evaluated under the same convex hull framework.

We observed that for NV, MNV, and k-mer, the disjoint ratio of convex hull pairs does not decrease as the dimensionality increases; instead, it eventually reaches 100%, which further proves the convex hull principle in taxonomy.

As illustrated in Tables 3–5, MNV achieves pairwise disjoint results for different families and variants at lower dimensions compared with NV and k-mer methods. As a result, the convex hull classification results demonstrate that MNV more effectively distinguishes viral families and subtypes than NV and k-mer methods, further confirming its superior ability to map sequences in a manner that preserves their taxonomic distinctions.

To validate the biological significance of the observed convex hull separation, we performed a randomized label-shuffling experiment on the SARS-Cov-2 and poliovirus dataset. The Virus Genome Dataset was excluded due to an insufficient number of sequences per family. After 10 shuffling iterations, the average disjoint ratio was significantly lower than that under the true labels (shown in Supplementary Table S3), confirming that the observed convex hull separation reflects meaningful biological classification rather than random chance.

4.2. Cluster analysis and neural network classification

To systematically evaluate the performance of MNV, NV, and k-mer methods, we implemented a dual-validation framework including DBSCAN clustering and neural network classification on SARS-CoV-2 and poliovirus datasets. The viral family-level dataset was excluded from model training due to limited sample sizes ($n \leq 10$ in many families), which violates the minimum requirement for robust machine learning applications. DBSCAN enables the assessment of intrinsic pattern separation. The neural network architecture

Table 8. Mutation Sites of Poliovirus Sequences with the Same 2-mer Embedding

Type	*Accession*	*Site 1*	*Site 2*
Subtype 1	KJ170477.1	4515: T	5047: C
	KJ170481.1	4515: C	5047: T
Subtype 2	KJ170548.1	4324: T	5046: C
	KJ170553.1	4324: C	5046: T
Subtype 3	KJ170591.1	2464: T	3611: C
	KJ170615.1	2464: C	3611: T
Subtype 3	KJ170618.1	346: C	2870: T
	KJ170619.1	346: T	2870: C

TABLE 9. NUMBER OF DUPLICATE 2-MER EMBEDDING GROUPS AND CORRESPONDING SEQUENCES IN SARS-COV-2 DATASET

Variant	*Duplicate embeddings*	*Corresponding sequences*
Alpha	1037	2410
Beta	0	0
Gamma	33	70
Delta	54	121
Omicron	2	4

comprises a three-layer fully connected network (FCNN) trained on feature embeddings generated by each method, with classification accuracy (correct predictions/total samples ×100%) as the main metric.

To ensure a reliable neural network evaluation, we performed 5-fold cross-validation on both datasets. To mitigate the pronounced class imbalance in the SARS-CoV-2 Dataset, larger classes were first downsampled randomly to match the size of the smallest class, while the poliovirus dataset was retained in its original form. Subsequently, a stratified K-fold splitting method was applied, ensuring that each fold maintained the same proportional distribution of variants as the dataset before splitting. During cross-validation, fourfolds were used for training and one for validation in each independent iteration. This approach ensured strict separation between training and validation data, providing a robust and unbiased performance estimate for neural network classification tasks.

Table 6 and Table 7 show that MNV achieves the best performance among the three methods on the SARS-Cov-2 and poliovirus datasets. For DBSCAN clustering, we searched for and applied the best hyper-parameters of every experiment, which are illustrated in Supplementary Table S1. The hyper-parameters of FCNN are illustrated in Supplementary Table S2. The FCNN classification and DBSCAN clustering results further demonstrate the enhancement of the MNV method in feature extraction of viral genomes.

4.3. Uniqueness of the embedding

Extended investigation reveals that the MNV method fundamentally overcomes the limitations of the k-mer method in capturing subtle sequence variations. To detect collision, embeddings of each dataset were converted and stored in a hash set. Hashing techniques were applied to quickly compare and identify embeddings that are exact duplicates of previously encountered ones. As a result, no embedding collisions were observed for MNV or NV across our three datasets. Notably, k-mer embeddings fail to achieve this essential uniqueness property. Especially when k is small, certain distinct sequences share identical k-mer representations.

To systematically characterize this degeneracy, we analyzed sequence clusters sharing identical k-mer embeddings. A representative example from the Poliovirus dataset (Table 8) demonstrates that 2-mer embeddings fail to distinguish sequences that harbor single-base antipodal mutations. This phenomenon extends across viral variants: the analysis of SARS-CoV-2 uncovered widespread duplication in both 2-mer and 3-mer embeddings. As shown in Tables 9 and 10, three or more distinct sequences are embedded to identical 2-mer or 3-mer vectors. Notably, Table 11 presents two striking case in which three independent sequences are mapped to the same 3-mer embedding.

In summary, both MNV and NV methods map each genomic sequence to a unique point in Euclidean space, whereas the k-mer method lacks this uniqueness. This duplication problem in the k-mer method demonstrates its limitation in capturing minor variations such as single nucleotide variants. In contrast, MNV and NV ensure

TABLE 10. NUMBER OF DUPLICATE 3-MER EMBEDDINGS AND CORRESPONDING SEQUENCES IN SARS-COV-2 DATASET

Variant	*Duplicate embeddings*	*Corresponding sequences*
Alpha	89	180
Beta	0	0
Gamma	1	2
Delta	7	14
Omicron	0	0

Table 11. Mutation Sites of SARS-Cov-2 Sequences with the Same 3-mer Embedding

Variant	*Accession*	*Site 1*	*Site 2*	*Site 3*
Alpha	MZ970789.1	1465: C	12352: C	14586: T
	MZ454153.1	1465: C	12352: T	14586: C
	MW906170.1	1465: T	12352: C	14586: C
Alpha	MZ131919.1	14742: T	15986: C	24570: C
	MW842371.1	14742: C	15986: C	24570: T
	MW841804.1	14742: C	15986: T	24570: C

the uniqueness of sequence embeddings, preserving the distinctness of each sequence and thereby proving significantly more effective at discriminating between highly similar sequences in rapidly evolving viral genomes.

4.4. Computing time analysis

The proposed MNV method incorporates an optimized algorithmic implementation that achieves significant speed advantages over baseline approaches.

Table 12 compares the processing times of the MNV, NV, and 3-mer methods across the three benchmark datasets. All computations were executed on identical hardware configurations featuring an Intel(R) Xeon(R) Platinum 8352 V CPU and an NVIDIA GeForce RTX 4090 GPU, ensuring fair performance comparisons. All methods were comfortably executed within 0.95 GB RAM.

Benchmarking on three diverse datasets demonstrates significantly faster processing compared to both standard 3-mer enumeration and NV methods, establishing MNV as a computationally efficient solution for large-scale genomic analysis.

5. DISCUSSION

In this study, we present the MNV, a novel AF method for differentiating viral genomes at both the family and subtype levels. MNV enhances the traditional NV method by incorporating trigonometric functions, capturing not only the distribution of nucleotides along the linear sequence but also the spatial and periodic characteristics inherent in the three-dimensional structure of the DNA molecule. This extension enables MNV to provide a richer, more robust representation of genomic sequences. In our experiments, we found that MNV outperforms both NV and k-mer methods, as evidenced by the improved convex hull separation, neural network classification, and clustering results on multiple viral datasets, including virus reference genomes, SARS-CoV-2 and poliovirus. In contrast to the k-mer method, the MNV method maps each sequence to a unique point in Euclidean space, enabling unambiguous identification of single base mutations through quantitative vector analysis.

While this study demonstrates MNV's effectiveness in taxonomic classification and clustering tasks, these represent relatively straightforward discriminative applications. Validation on more complex generative tasks—particularly sequence prediction—remains unexplored. This limitation stems from our initial focus on establishing MNV's representational capacity for viral discrimination. Future work will extend this framework to predictive modeling, for example, predicting immunogenic regions from sequence-structure embeddings and modeling mutation impacts on viral fitness. Such applications would more rigorously test MNV's ability to capture functional sequence determinants beyond taxonomic signatures.

Table 12. Time Comparison of the Three Methods

Dataset/Seconds	*MNV*	*NV*	*3mer*
Virus Genome Dataset	111.15 (124d)	2349.48 (124d)	268.48 (64d)
SARS-CoV-2 Genome Dataset	285.79 (52d)	4825.99 (52d)	1049.52 (64d)
Poliovirus Genome Dataset	3.96 (28d)	19.01 (28d)	7.73 (64d)

In conclusion, MNV offers a powerful approach for genomic sequence analysis, making it a promising tool for viral taxonomy, evolutionary studies, and large-scale data processing. The effectiveness in extracting meaningful features from genomic sequences highlights its potential for broader applications in bioinformatics.

ACKNOWLEDGMENT

The authors thank the associated editor and the anonymous reviewers for their valuable suggestions.

AUTHORS' CONTRIBUTIONS

X.S.: Conceptualization, data curation, software, formal analysis, writing—original draft. J.K.: Methodology, investigation, writing—review and editing. N.S.: Data curation, result validation, writing—review and editing. X.Z.: Visualization, investigation, and formal analysis. S.S.-T.Y.: Funding acquisition, project administration, supervision, writing—review and editing. All authors have read and agreed to the published version of the article.

AUTHOR DISCLOSURE STATEMENT

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

FUNDING INFORMATION

This work is supported by National Natural Science Foundation of China (NSFC) grant (No. 12171275, No. 12201015) and Tsinghua University Education Foundation.

SUPPLEMENTARY MATERIAL

Supplementary Table S1
Supplementary Table S2
Supplementary Table S3

REFERENCES

Altschul SF, Madden TL, Schäffer AA, et al. Gapped blast and psi-blast: A new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389–3402.

Blaisdell BE. Average values of a dissimilarity measure not requiring sequence alignment are twice the averages of conventional mismatch counts requiring sequence alignment for a variety of computer-generated model systems. J Mol Evol 1991;32(6):521–528.

Bohnsack KS, Kaden M, Abel J, et al. Alignment-free sequence comparison: A systematic survey from a machine learning perspective. IEEE/ACM Trans Comput Biol and Bioinf 2023;20(1):119–135.

Crick F. Central dogma of molecular biology. Nature 1970;227(5258):561–563.

Crick FH. On protein synthesis. In Symp Soc Exp Biol 1958;12:8.

Deng M, Yu C, Liang Q, et al. A novel method of characterizing genetic sequences: Genome space with biological distance and applications. PLoS One 2011;6(3):e17293.

Duffy S. Why are rna virus mutation rates so damn high? PLoS Biol 2018;16(8):e3000003.

Edgar RC. Muscle: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32(5):1792–1797.

Ester M, Kriegel H-P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd 1996;96:226–231.

Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc 1983;78(383): 553–569.

Fred AL, Jain AK. Combining multiple clusterings using evidence accumulation. IEEE Trans Pattern Anal Mach Intell 2005;27(6):835–850.

Herzel H, Weiss O, Trifonov EN. 10-11 bp periodicities in complete genomes reflect protein structure and dna folding. Bioinformatics 1999;15(3):187–193.

Hu H, Yuan X, Huang L, et al. Global dynamics of an sirs model with demographics and transfer from infectious to susceptible on heterogeneous networks. Math Biosci Eng 2019;16(5):5729–5749.

Hubert L, Arabie P. Comparing partitions. J Classif 1985;2(1):193–218.

Just W. Computational complexity of multiple sequence alignment with sp-score. J Comput Biol 2001;8(6):615–623.

Larkin MA, Blackshields G, Brown NP, et al. Clustal w and clustal x version 2.0. Bioinformatics 2007;23(21): 2947–2948.

Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8): 2444–2448.

Sun N, Zhao X, Yau SS-T. An efficient numerical representation of genome sequence: Natural vector with covariance component. PeerJ 2022;10:e13544.

Tian K, Zhao X, Yau SS-T. Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. J Theor Biol 2018;456:34–40.

Watson JD, Crick FH. The structure of dna. In: Cold Spring Harbor symposia on quantitative biology, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press; 1953.

Wen J, Chan RH, Yau S-C, et al. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. Gene 2014;546(1):25–34.

Zhurkin VB. Periodicity in dna primary structure is defined by secondary structure of the coded protein. Nucleic Acids Res 1981;9(8):1963–1971.

Zielezinski A, Vinga S, Almeida J, et al. Alignment-free sequence comparison: Benefits, applications, and tools. Genome Biol 2017;18(1):186.

Address correspondence to:
Prof. Stephen S.-T. Yau
Department of Mathematical Sciences
Tsinghua University
Hetao
Beijing 518000
P. R. China

E-mail: yau@uic.edu