Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Entropy-Driven Alignment-Free Phylogenetic Analysis via K-mer Encoding, DNA Physicochemical Properties, and Rough Set Feature Selection

Yong Liu
Meihe Hospital of the First Hospital of Jilin University
Tongliang Xia
Qilu Hospital of Shandong University
Xu Yan
Jinzhou Medical University
He Wang
Jinzhou Medical University
Xinyang Jiang
the Seventh Medical Center of Chinese PLA General Hospital
Shujie Gao
Jinzhou Medical University
Shusheng Li
Jinzhou Medical University
Yan Yang
360226919@qq.com

Jinzhou Medical University

Research Article

Keywords: DNA sequence, Alignment-free method, K-mer, Feature selection, Dataset, Phylogenetic analysis

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10103476/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Entropy-Driven Alignment-Free Phylogenetic Analysis via *K*-mer Encoding, DNA Physicochemical Properties, and Rough Set Feature Selection

Yong Liu[c,*], Tongliang Xia[d,*], Xu Yan[a], He Wang[a], Xinyang Jiang[b], Shujie Gao[a], Shusheng Li[a], Yan Yang[a,b,#]

[a]The College of Health Humanities, Jinzhou Medical University, Jinzhou 121001, China

[b]Department of Pediatrics, the Seventh Medical Center of Chinese PLA General Hospital, Beijing 100010, China

[c]Department of Orthopedics, Meihe Hospital of the First Hospital of Jilin University, Changchun 130000, China

[d]Department of Otorhinolaryngology, Qilu Hospital of Shandong University, Jinan 250063, China

*These authors contributed equally to this work.

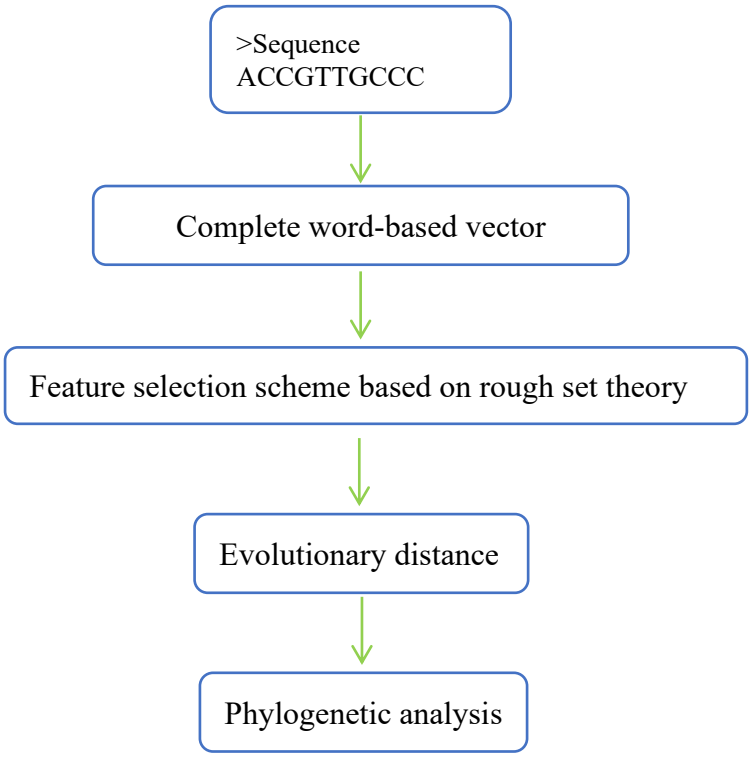
[#]Corresponding authors. Email: 360226919@qq.com (Yang Y), 519804221@qq.com(Gao YL).

Abstract: Introduction: Alignment-free methods based on *k*-mers have been widely used in DNA sequence evolutionary analysis due to their computational efficiency and robustness to genome rearrangements. However, many existing approaches primarily rely on nucleotide composition and neglect both the intrinsic physicochemical properties of DNA and the ordering information of *k*-mers, which may limit their ability to capture biologically meaningful sequence characteristics.

Materials and Methods: In this study, we propose an alignment-free phylogenetic analysis framework that integrates *k*-mer statistics with DNA physicochemical properties through a set of seven characteristic sequences. These characteristic sequences jointly encode nucleotide composition, physicochemical attributes, and strand symmetry, enabling the construction of a *k*-mer word vector representation that preserves both sequence order and biochemical heterogeneity. To reduce redundancy and improve discriminative power, an enhanced rough set–based feature selection method is introduced, incorporating average folding points to identify informative *k*-mers in high-dimensional space. **Results:** To evaluate the performance of our method, we conducted phylogenetic analysis on four datasets. The first dataset served as the training set, where 3005 words were selected from $7 \times 2^{11} = 14336$ *k*-mers to construct the final feature vector. The remaining three datasets served as the testing sets. The results demonstrate that the proposed method effectively captures essential information in DNA sequences. **Discussion:** Experimental outcomes further support the method's reasonable performance in viral molecular phylogenetic inference and its plausibility for such analyses. **Conclusion:** Taken together, this study demonstrates that the proposed framework provides an efficient and interpretable solution for whole-genome phylogenetic inference, with potential applications extending to broader genomic and metagenomic analyses.

Keywords： DNA sequence, Alignment-free method, *K*-mer, Feature selection, Dataset, Phylogenetic analysis

Graphical Abstract



1 Introduction

Advancements in third-generation sequencing technologies have made it more feasible to acquire whole-genome information from various organisms [1–32]. The vast availability of such data presents a significant challenge to traditional nucleotide phylogenetic analysis methods. A key reason is that most existing phylogenetic methods rely on multiple sequence alignment. As noted by Otu and Sayood [33], multiple sequence alignment faces bottlenecks such as computational complexity and the arbitrariness of alignment penalties. Additionally, genetic recombination, substitutions, and inversions at the substring level render it unsuitable for phylogenetic analysis at the whole-genome scale. Thus, the creation of a highly effective and alignment-independent sequence comparison technique is of utmost importance.

In the span of three decades, numerous scientists have investigated diverse alignment-independent techniques to reveal the concealed evolutionary data embedded in biological sequences, approaching the subject from multiple angles.The visual depiction of a DNA sequence acts as a potent instrument for examining extensive DNA sequences [34], This tool enables the monitoring of nucleotide appearances, variations, and repetitions.One of the efficient approaches is utilizing matrix representation for biological sequences, including adjacency matrices, compression matrices, M/M matrices, L/L matrices, as well as "higher-order" matrices [35].Sequence alignment has also been achieved using compression algorithms, including Burrows-Wheeler (BW) transform and Lempel-Ziv (LZ) complexity algorithms [36–42]. Among alignment-free strategies, *k*-mer–based methods are particularly popular due to their simplicity and scalability. These methods represent sequences as vectors of *k*-mer frequencies and infer evolutionary relationships using distance or similarity measures [43–52]. Several extensions have been proposed, including composition vectors, spaced *k*-mers, and sketch-based techniques such as Mash, which approximate *k*-mer distributions using MinHash to achieve remarkable computational efficiency [53–56]. More recently, deep learning–based sequence embeddings have also been explored for phylogenetic inference and sequence classification [47–59].

Despite their success, existing alignment-free methods face notable limitations. Sketch-based approaches trade accuracy for speed and may lose fine-grained biological information, while deep learning models often require large training datasets and suffer from limited interpretability. Moreover, most *k*-mer–based representations focus primarily on nucleotide composition and largely ignore intrinsic physicochemical properties of DNA as well as the ordering relationships among *k*-mers.

To address these issues, we introduce an innovative *k*-mer-based technique aimed at enhancing the extraction of information from biological sequences. Our approach begins with seven binary

sequences derived from a given sequence, and for a specified k, constructs a 7×2^k-dimensional complete word vector. Recognizing that different k-mers vary in their contribution to evolutionary distance computation, we refine a scoring mechanism to prioritize the most impactful k-mers. This allows us to heuristically capture the most informative evolutionary signals. Based on the chosen k-mers, each sequence is represented by a concise, low-dimensional feature vector. This approach is employed in phylogenetic inference. To determine the optimal k value and the number of selected k-mers (denoted as n_k), we used 11 hantavirus sequences as the training set to identify fixed values for k and n_k. With these parameters, we constructed phylogenetic trees using three independent viral datasets from the literature. The results confirmed that our method effectively extracts valuable sequence information.

2 Methods and materials

2.1 Word vector representation of DNA sequences based on 7 characteristic sequences

In a DNA sequence, the four nucleotides can be divided into two categories according to their chemical compositions: purines $R=\{A, G\}$ and pyrimidines $Y=\{C, T\}$, or keto group $M=\{A, C\}$ and amino group $K=\{G, T\}$. Additionally, the four nucleotides can also be divided into two categories based on hydrogen bond strength between nucleotides: weak hydrogen bonds $W=\{A, T\}$ and strong hydrogen bonds $S=\{G, C\}$. By assigning a value of 1 to nucleotides in R, M, and W, and a value of 0 to nucleotides in Y, K, and S in a given DNA sequence, the original DNA sequence can be transformed into three distinct (0, 1)-sequences. These sequences are referred to as the (R, Y)-, (M, K)-, and (W, S)-characteristic sequences of the DNA sequence. Furthermore, four additional binary (0, 1) sequences are constructed to contrast each nucleotide with the remaining three, with each sequence encoding the presence of one nucleotide: for nucleotides A, a value of 0 is assigned, and $\{C, G, T\}$ are assigned the value 1; for C, a value of 0 is assigned, and $\{A, G, T\}$ are assigned the value 1; and similarly for G and T, resulting in four such sequences. Using these seven characteristic sequences, we can obtain seven n-dimensional vectors $V_F^1, V_F^2, V_F^3, V_F^4, V_F^5, V_F^6, V_F^7$, from which a $7n$-dimensional vector $v = (V_F^1, V_F^2, V_F^3, V_F^4, V_F^5, V_F^6, V_F^7)$ can be constructed. This vector serves as a newly constructed feature representation for the corresponding DNA sequence. The format of each binary (0,1) sequence is outlined as follows:

For a (0, 1)-sequence S of length m, the number of occurrences of a k-mer w in S is denoted as $c(w)$. Since there are m-k+1 possible k-mers (including overlapping) in the sequence S, the frequency of the k-mer w is given by $f(w) = c(w)/(m-k+1)$. Once the frequencies of all n possible k-mers in the sequence are obtained, a frequency vector can be constructed as follows:

$$F_k = \left(f(w_{k,1}), f(w_{k,2}), \cdots, f(w_{k,n})\right)$$

Although there are inherent associations between these frequencies, the vector treats them as independent. To reflect the sequential relationship of the *k*-th frequencies, we arrange the components of the frequency vector F_k in ascending order, resulting in a new vector denoted as F_s:

$$F_s = \left(f(w_{k,i_1}), f(w_{k,i_2}), \cdots, f(w_{k,i_n})\right)$$

When the frequencies of two *k*-mers are equal, they are ordered lexicographically. This ensures that the sorted vector F_s satisfies the following condition:

$$f(w_{k,i_1}) \leq f(w_{k,i_2}) \leq \cdots \leq f(w_{k,i_n})$$

Each *k*-mer frequency $f(w)$ has a unique position in F_s, denoted as $g(w)$. By combining the frequencies and their positional information, we construct the *n*-dimensional vector representation V_F for the (0,1)-sequence *S*:

$$V_F = \left(g(w_{k,1}) \times e^{f(w_{k,1})}, g(w_{k,2}) \times e^{f(w_{k,2})} \cdots g(w_{k,n}) \times e^{f(w_{k,n})}\right)$$

The components in the vector are only determined by their positions when $f(w)$=0.

For instance, suppose

S=1001001000111101000000100100110001001001000011010100001110110110110011011110011001100001110010110000001011111001010000010100101111101110010001111000000010001010110000101010000110111110111111011010000001000010011101011000000011111 and *k*=3.

Then there are 8 possible 3-words in total:

000, 001, 010, 011, 100, 101, 110, 111.

It is easy to find that

F_k=(0.1778, 0.1333, 0.1244, 0.1111, 0.1333, 0.1022, 0.1067, 0.1111),

and then we have

F_s=(0.1022, 0.1067, 0.1111, 0.1111, 0.1244, 0.1333, 0.1333, 0.1778).

Therefore,

g=(8, 6, 5, 3, 7, 1, 2, 4)

and

V_F=(9.5568, 6.8556, 5.6625, 3.3525, 7.9982, 1.1076, 2.2252, 4.4700).

We recorded the processing time for sequences of various lengths using the following system configuration: Intel® Core™ i5-10210U CPU @ 1.60 GHz (2.11 GHz), with 8 GB RAM. Processing the longest sequence (31,100 bp) took less than 4 seconds, indicating that the proposed method is

computationally efficient and scalable for long sequences (Table 1).

Table 1. Time taken to conduct numerical characterization for sequences of varying lengths.

Name	Len	*K*	Time
XJP613	10,964 bp	11	1 sec
YN83-Meng83-54	10,965 bp	11	1 sec
Bovine_coronavirus_isolate_BCoV-LUN	31,028 bp	11	4 sec
Bovine_coronavirus_strain_Quebec	31,100 bp	11	4 sec

The resulting word vectors derived from the seven characteristic sequences are concatenated to form a complete *k*-mer representation of the DNA sequence. However, due to the high dimensionality of this representation, not all *k*-mers contribute equally to phylogenetic inference. Therefore, a feature filtering strategy based on rough set theory is introduced to identify a subset of informative *k*-mers for subsequent analysis.

2.2 Feature Filtering Strategy Based on Rough Sets

Rough set theory, originally proposed by the Polish scientist Pawlak in 1982 [60–64], is a powerful mathematical framework for analyzing intelligent systems characterized by ambiguity, uncertainty, and vagueness. It has been widely applied in attribute reduction and feature selection, where the core objective is to identify condition attributes that are most relevant to decision attributes while preserving the classification ability of the original system.

In Section [65], we introduced a rough set–based feature selection method. In this work, we further improved this approach by proposing a novel attribute importance measure based on the concepts of folding points and average folding points. This measure is designed to quantitatively describe the discriminative contribution of each condition attribute in ordered decision tables and to enhance the quality of attribute reduction, especially for DNA sequence data.

An information system S is a quadruple $S = \langle U, A, V, f \rangle$, where $U = \{x_1, x_2, \cdots, x_n\}$ is a non-empty finite set, called the domain, A is a finite set of attributes, V is the set of attribute values, and f is an information function that satisfies certain conditions. $\forall a \in A, \forall x \in U, f(x,a) \in V$. If the set of attributes A is divided into a set of condition attributes C and a set of decision attributes D, then the information system $S = \langle U, C \cup D, V, f \rangle$ is called a decision system, or a decision table.

Given $c_k \in C$, a consistent decision table $S = \langle U, C \cup D, V, f \rangle$, $c_k \in C$, $x, y \in U$, if $f(x,D) \neq f(y,D)$ (this means $f(x,c_k) \neq f(y,c_k)$), we define a "folding point" between x and y, with its value given by:

$$f_{IP}(x,y,c_k) = \frac{1}{2}[f(x,c_k) + f(y,c_k)]$$

Correspondingly, the "average folding point" is defined as follows:

$$A_{IP}(x, y, c_k) = Average[f_{IP}(x, y, c_k)]$$

Furthermore, in the decision table *S*, suppose "≤" is an order relation in $U = \{x_1, x_2, \cdots, x_n\}$. Without loss of generality, assume $x_1 \leq x_2 \leq \cdots \leq x_n$. We can identify the folding points between any two adjacent objects, and consequently determine the corresponding average folding points. We define the smallest value among the folding rates corresponding to c_k as the minimum average folding point, denoted as $A_{mi}(c_k)$.

The larger the value of $A_{mi}(c_k)$, the more important the attribute c_k is. Therefore, we use the value of $A_{mi}(c_k)$ to select important conditions. Later, we will describe in detail how extracting important attributes enables us to obtain the smaller number of features vectors of DNA sequences.

The reduced feature vectors generated by this process are subsequently used for evolutionary distance computation and phylogenetic tree construction. During the training phase with hantavirus sequences, the virus type (HTN-type vs. SEO-type) is used as the decision attribute, while the k-mer–derived features constitute the condition attributes in the decision table.

3 Results

In this section, the effectiveness of our method is assessed by generating phylogenetic trees for a total of four data sets. The four datasets are partitioned into two categories: one is utilized as the training dataset, and the other three function as test datasets.

3.1 Training Phase

The dataset comprises 11 Hantaviruses (refer to Table 2), spread throughout the East Asian region. Among them, Z10, Z5, ZLS6-11, ZLS-12, and Z251 belong to the HTN-type virus, while Gou3, K24-v2, K24-e7, ZT10, ZT71, and Z37 belong to the SEO-type virus.

Table 2. The S segments of 11 HV strains

No	Strain	Type	AC (GenBank)	Region
1	Z10	HTN	AF184987	Shengzhou
2	Z5	HTN	EF103195	Shengzhou
3	Z251	HTN	EF595840	Longquan
4	ZLS6-11	HTN	FJ753396	Lishui
5	ZLS-12	HTN	FJ753398	Lishui
6	Gou3	SEO	AF184988	Jiande
7	K24-v2	SEO	AF288655	Xinchang
8	K24-e7	SEO	AF288653	Xinchang

9	Z37	SEO	AF187082	Wenzhou
10	ZT71	SEO	AY750171	Tiantai
11	ZT10	SEO	AY766368	Tiantai

For a given *k*, following the aforementioned method, we construct a 7×2^k-dimensional feature *k*-mer vector for each Hantavirus and then compute the effective minimum average folding point. Our results show that the training process achieves optimal performance when *k*=11 and $A_{mi}(c_k)$=78.1866 (the average minimum folding rate, see Figure 1). Out of the total $7 \times 2^{11} = 14336$ *k*-mers, 3005 significant *k*-mers are selected as the final feature vectors based on their importance.

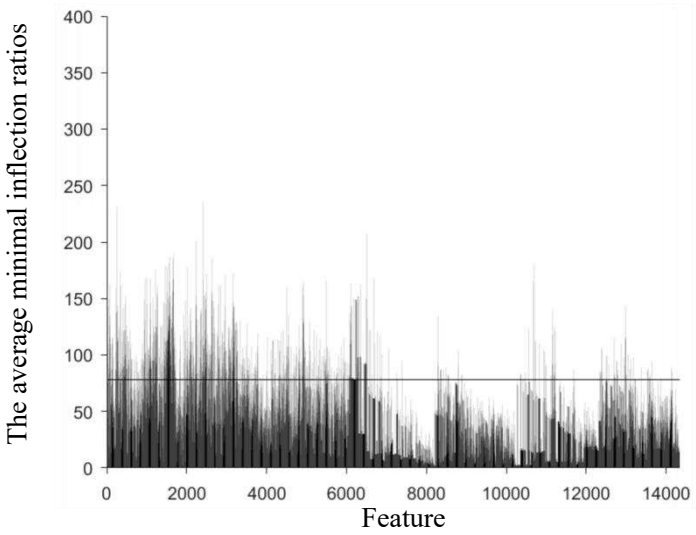


Figure 1. The average minimal inflection ratios (%).

Following the approach reported in multiple studies [35, 66, 67], where the evolutionary distance is calculated as the Euclidean distance divided by the cosine of the angle between vectors, we adopt the same formula for computing evolutionary distances in this study:

$$D(i,j) = d(v_i, v_j) / \cos(v_i, v_j)$$

where v_i and v_j are the feature vectors of sequences *i* and *j*, $\cos(v_i, v_j)$ represents the cosine similarity between v_i and v_j, and $d(v_i, v_j)$ is the Euclidean distance between them.

Using this formula, we calculated the pairwise evolutionary distances among the 11 Hantavirus sequences and constructed a corresponding distance matrix. Based on this matrix, a phylogenetic tree was generated using the UPGMA method implemented in MEGA 11 (Figure 2A). For comparison, multiple sequence alignment and phylogenetic reconstruction were also performed in MEGA, and bootstrap analysis was conducted on the alignment-based reference tree to assess the reliability of the inferred clades (Figure 2B). The resulting phylogenetic tree clearly distinguishes between HTN-type and SEO-type viruses, with strong bootstrap support for the major branches.

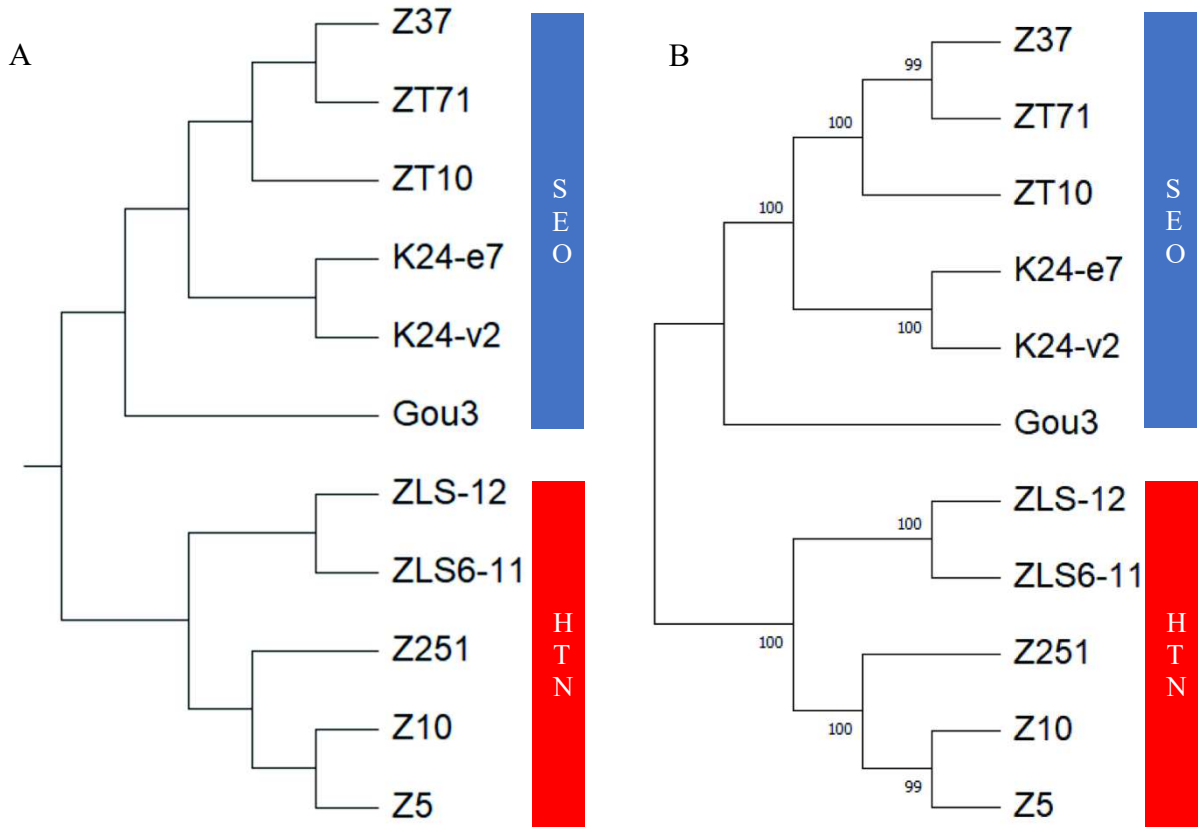


Figure 2. Phylogenetic trees of 11 Hantaviruses. **A:** Phylogenetic tree generated using our method. **B:** Phylogenetic tree generated using the ClustalW method.

3.2 Testing Phase

We use the 3005 *k*-mers selected during the training phase to infer the phylogenetic relationships of 33 complete coronavirus genome sequences (Table 3). Prior to the emergence of SARS and SARS-CoV-2 coronaviruses, classical coronaviruses were classified into three groups based on their genomic and antigenic characteristics: Group I and Group II include mammalian coronaviruses, while Group III contains only avian coronaviruses. Human coronaviruses belong to Group I and Group II, and are important pathogens that cause respiratory infections. About 20% of common colds in humans are caused by coronaviruses. However, since the symptoms of the common cold are mild, they haven't attracted as much attention as SARS and SARS-CoV-2. SARS and SARS-CoV-2 viruses are distinct from previously identified coronaviruses and form an independent fourth branch (Group IV) within the coronavirus family.

Table 3. The 33 Coronavirus

No	Accession	Group	Genome
1	NC_002645	I	Human coronavirus 229E
2	AJ271965	I	Transmissible gastroenteritis virus
3	NC_003436	I	Porcine epidemic diarrhea virus

4	U00735	II	Bovine coronavirus strain Mebus
5	AF391542	II	Bovine coronavirus isolate BCoV-LUN
6	AF220295	II	Bovine coronavirus strain Quebec
7	NC_003045	II	Bovine coronavirus
8	AF208067	II	Murine hepatitis virus strain ML-10
9	AF201929	II	Murine hepatitis virus strain 2
10	AF208066	II	Murine hepatitis virus strain Penn 97-1
11	NC_001846	II	Murine hepatitis virus strain A59
12	NC_001451	III	Avian infectious bronchitis virus
13	DQ001339	III	Avian infectious bronchitis virus IBV-p65
14	AY851295	III	Avian infectious bronchitis virus strain Mass-41
15	AY641576	III	Avian infectious bronchitis virus Peafowl
16	AY283794	IV	SARS coronavirus Sin2500
17	AY283798	IV	SARS coronavirus Sin2774
18	AY291451	IV	SARS coronavirus TW1
19	NC_004718	IV	SARS coronavirus
20	AY278488	IV	SARS coronavirus BJ01
21	AY278741	IV	SARS coronavirus Urbani
22	AY278491	IV	SARS coronavirus HKU-39849
23	AY278554	IV	SARS coronavirus CUHK-W1
24	AY278752	IV	SARS coronavirus CUHK-Su10
25	AY283795	IV	SARS coronavirus SIN2677
26	AY283797	IV	SARS coronavirus SIN2748
27	AY283796	IV	SARS coronavirus SIN2679
28	OU345798	IV	SARS-CoV-2/human/USA/CA-UCSC-508/2020
29	OU345802	IV	SARS-CoV-2/human/USA/CA-UCSC-467/2020
30	OU345800	IV	SARS-CoV-2/human/USA/CA-UCSC-466/2020
31	OU345806	IV	SARS-CoV-2/human/USA/CA-UCSC-445/2020
32	OU345803	IV	SARS-CoV-2/human/USA/CA-UCSC-442/2021
33	OU345799	IV	SARS-CoV-2/human/USA/CA-UCSC-432/2021

We constructed a phylogenetic tree (Figure 3A) in which coronaviruses are divided into four major clades. All SARS-related coronaviruses, including SARS-CoV and SARS-CoV-2, cluster together to form an independent branch. A more detailed examination reveals that SARS-CoV and SARS-CoV-2 each form distinct subclusters within this branch. These results are consistent with previously reported findings [68, 69]. In addition, an alignment-based phylogenetic tree was reconstructed using MEGA, and bootstrap analysis was performed to evaluate the robustness of the

inferred relationships (Figure 3B). The alignment-based tree with bootstrap support further confirms the reliability of the clustering patterns observed using our method.

A



B

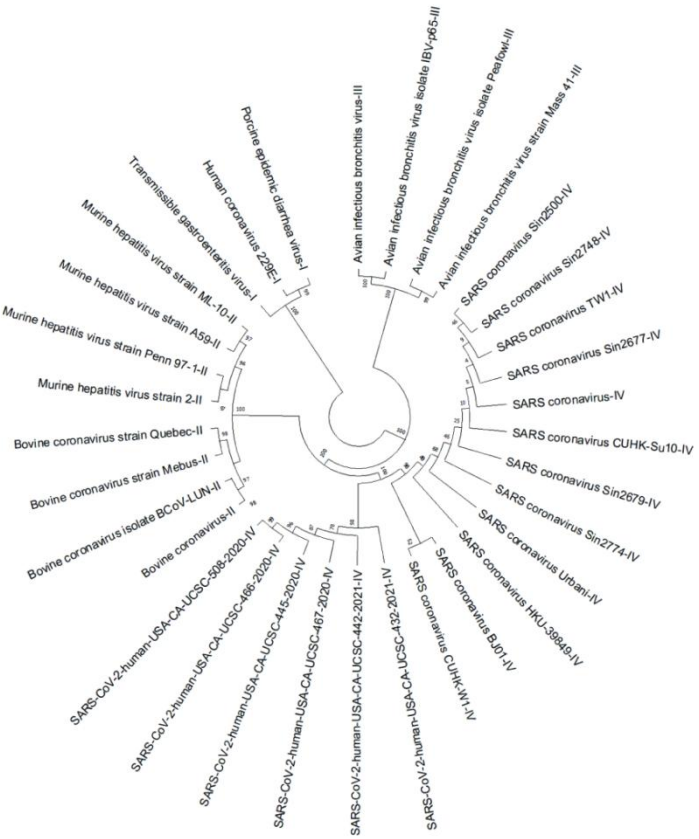


Figure 3. Phylogenetic trees of Coronaviruses. **A:** Phylogenetic tree generated using our method. **B:** Phylogenetic tree generated using the ClustalW method.

The fourth test set consists of 39 complete Japanese encephalitis virus (JEV) genome sequences (Table 4), with MVE-1-51 used as the outgroup. Japanese encephalitis virus (JEV) is the causative agent for epidemic Japanese encephalitis, primarily circulating in Asia but expanding globally, with cases now found in Australia. It is estimated that approximately 3 billion people worldwide live under the threat of JEV. To better analyze the genetic characteristics of JEV, we constructed two phylogenetic trees for the virus (Figure 4).

Table 4. The 39 Japanese encephalitis viruses

No	Strain	Genotype	Location	Source	Genbank (AC)
1	K94P05	I	Korea	Mosquito	AF045551
2	KV1899	I	Korea	Swine	AY316157
3	Ishikawa	I	Japan	Mosquito	AB051292
4	JEV/se/Mie/41/2002	I	Japan	Mosquito	AB241119
5	GZ56	I	China	Homo sapiens	HM366552
6	YN83-Meng83-54	I	China	Midge Culex	JF706282
7	YN0967	I	China	Tritaeniorhynchus	JF706268
8	XJ69	I	China	Mosquito	EU880214
9	XJP613	I	China	mosquito	EU693899
10	HN0421	I	China	IU Culex	JN381841
11	YN05124	I	China	Tritaeniorhynchus	JF706281
12	HN06129	I	China	Armigeres sp.	JF706277
13	SC0415	I	China	IU	JN381838
14	YN05155	I	China	IU	JN381852
15	Clone: JEV-AT31	III	IU	IU	AB196926
16	Clone: JEV-AT31-1	III	IU	IU	AB196923
17	Clone: JEV-rAT	III	IU	IU	AB196925
18	CC27-S6	III	Taiwan	Mosquito	AY303797
19	CC27-L1	III	Taiwan	Mosquito	AY303795
20	CC27-S8	III	Taiwan	Mosquito	AY303798
21	FJ0339	III	China	IU	JN381859
22	FJ0394	III	China	IU	JN381858
23	TlP1-L4	III	IU	IU	AY303792
24	T1P1-S1	III	IU	IU	AY303791
25	CJN-L1	III	Taiwan	IU	AY303794
26	CJN-S1	III	Taiwan	IU	AY303793
27	04940-4	III	India	IU	EF623989

28	057434	III	India	IU	EF623988
29	SA(A)	III	China	IU	D90195
30	SA-14-14-2	III	China	IU	AF315119
31	CH2195LA	III	Taiwan	CH2195 derivative	AF221499
32	CH2195SA	III	Taiwan	CH2195 derivative	AF221500
33	014178	III	India	Human brain	EF623987
34	P20778	III	India	Human brain	AF080251
35	JaOH0566	III	Japan	Human brain	AY508813
36	Nakayama	III	Japan	Human brain	EF571853
37	FU	II	Australia	Human brain	AF217620
38	JKT6468	IV	Indonesia	Mosquito	AY184212
39	MVE-1-51	Out group	Australia	IU	AF161266

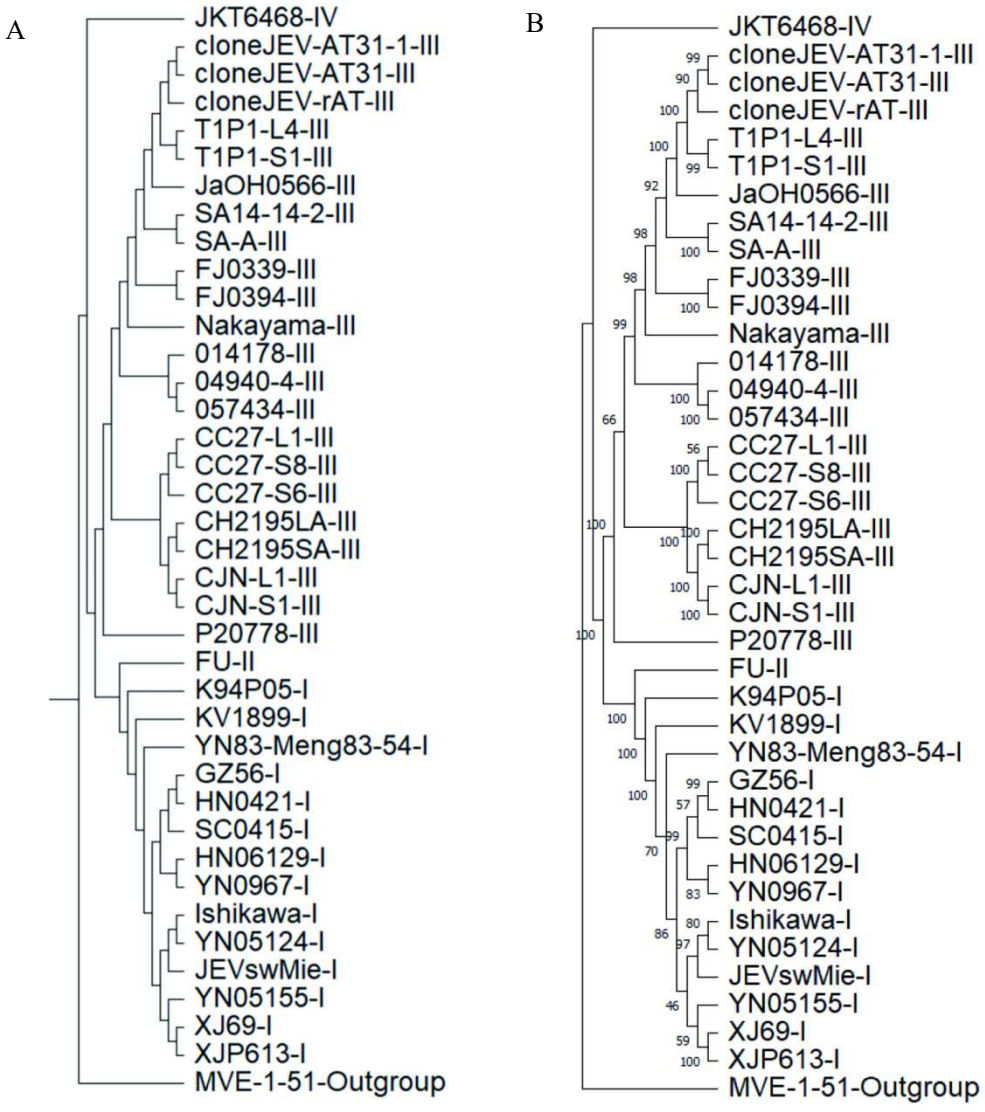


Figure 4. Phylogenetic trees of JEV viruses. **A:** Phylogenetic tree generated using our method. **B:** Phylogenetic tree generated using the ClustalW method.

Figure 4A shows that the JEV genomes are classified into four distinct branches, each corresponding to one of the four identified genotypes of the virus, with MVE-1-51 as the outgroup. For comparison, multiple sequence alignment–based phylogenetic reconstruction was also performed using MEGA, and bootstrap analysis was conducted to evaluate the robustness of the inferred clades. The resulting alignment-based phylogenetic tree with bootstrap support is presented in Figure 4B.

The third test set consists of 55 H7N9 avian influenza virus NA antigen gene sequences (Table 5). H7N9 (Influenza A virus subtype H7N9, abbreviated as A(H7N9) or H7N9) is a subtype of the avian influenza virus. Due to the low mortality rate in birds, it has transferred to humans through genetic reassortment. After infecting humans, it causes a relatively short duration of illness with a higher severity and mortality rate compared to SARS. Currently, there are no specific drugs to treat this disease, and studying the evolutionary relationships of H7N9 is crucial for the development of antiviral drugs.

Table 5. NA gene of H7N9 invluenza virus Nucleotide

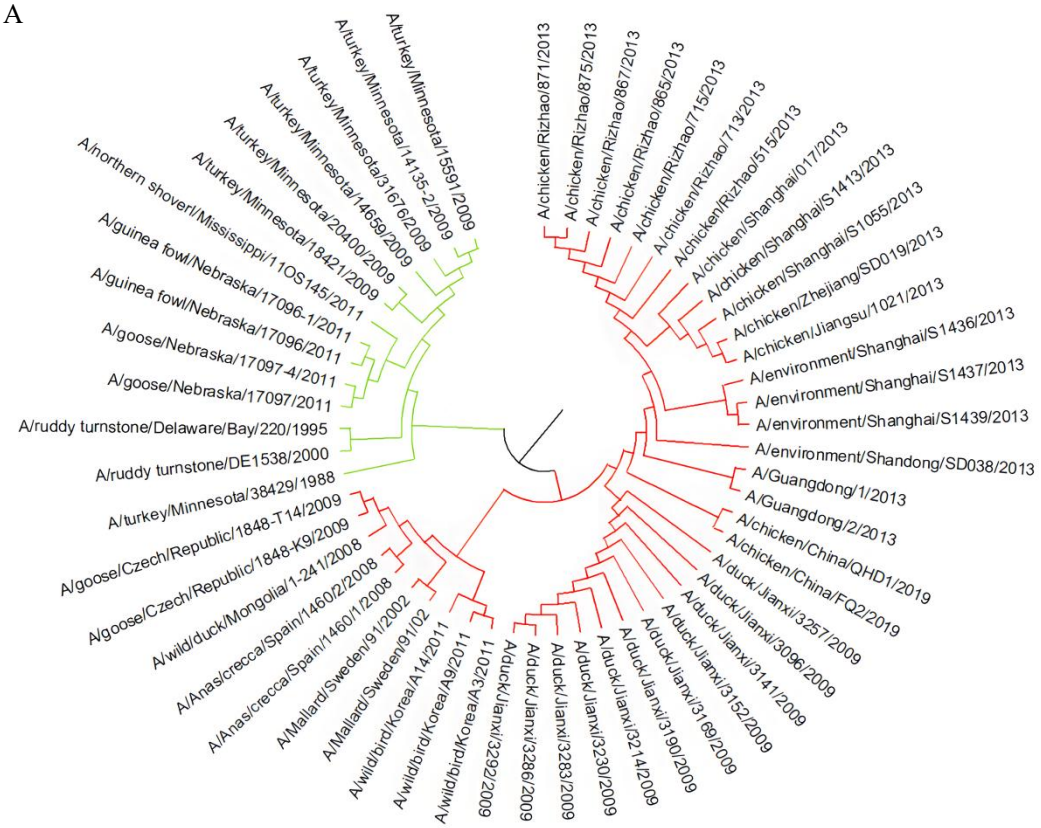
No	Strain	AC (GenBank)
1	A/chicken/Zhejiang/SD019/2013	CY147044
2	A/chicken/Jiangsu/1021/2013	KF938946
3	A/chicken/Shanghai/S1055/2013	CY146956
4	A/chicken/Shanghai/S1413/2013	CY147028
5	A/chicken/Shanghai/017/2013	KF542876
6	A/chicken/Rizhao/515/2013	KF259042
7	A/chicken/Rizhao/713/2013	KF259058
8	A/chicken/Rizhao/715/2013	KF259059
9	A/chicken/Rizhao/865/2013	KF259060
10	A/chicken/Rizhao/867/2013	KF259043
11	A/chicken/Rizhao/871/2013	KF259044
12	A/chicken/Rizhao/875/2013	KF259045
13	A/environment/Shandong/SD038/2013	CY147100
14	A/environment/Shanghai/S1436/2013	CY147132
15	A/environment/Shanghai/S1437/2013	CY147140
16	A/environment/Shanghai/S1439/2013	CY147156
17	A/Guangdong/1/2013	KF662943
18	A/Guangdong/2/2013	KF922738
19	A/duck/Jianxi/3096/2009	KF258951
20	A/duck/Jianxi/3141/2009	KF258952
21	A/duck/Jianxi/3152/2009	KF258987
22	A/duck/Jianxi/3169/2009	KF258953
23	A/duck/Jianxi/3190/2009	KF258954
24	A/duck/Jianxi/3214/2009	KF258955
25	A/duck/Jianxi/3230/2009	KF258956
26	A/duck/Jianxi/3257/2009	KF258988
27	A/duck/Jianxi/3283/2009	KF258985
28	A/duck/Jianxi/3286/2009	KF258957
29	A/duck/Jianxi/3292/2009	KF258992

30	A/wild/bird/Korea/A14/2011	JN244231
31	A/wild/bird/Korea/A3/2011	JN244232
32	A/wild/bird/Korea/A9/2011	JN244234
33	A/Mallard/Sweden/91/02	AY999981
34	A/Mallard/Sweden/91/2002	KF695254
35	A/Anas/crecca/Spain/1460/1/2008	FN386467
36	A/Anas/crecca/Spain/1460/2/2008	HQ244407
37	A/wild/duck/Mongolia/1-241/2008	JN029686
38	A/goose/Czech/Republic/1848-K9/2009	GU060482
39	A/goose/Czech/Republic/1848-T14/2009	HQ244415
40	A/turkey/Minnesota/38429/1988	GU053163
41	A/ruddy turnstone/DE1538/2000	EU684261
42	A/ruddy turnstone/Delaware/Bay/220/1995	CY127253
43	A/turkey/Minnesota/14135-2/2009	KF715076
44	A/turkey/Minnesota/14659/2009	KF771189
45	A/turkey/Minnesota/15591/2009	KF771197
46	A/turkey/Minnesota/18421/2009	KF771205
47	A/turkey/Minnesota/20400/2009	KF771213
48	A/turkey/Minnesota/31676/2009	KF715084
49	A/northern shoverl/Mississippi/11OS145/2011	CY133649
50	A/goose/Nebraska/17097/2011	KF715114
51	A/goose/Nebraska/17097-4/2011	JX899805
52	A/guinea fowl/Nebraska/17096/2011	KF715106
53	A/guinea fowl/Nebraska/17096-1/2011	JX899803
54	A/chicken/China/QHD1/2019	MN700031
55	A/chicken/China/FQ2/2019	MN700030

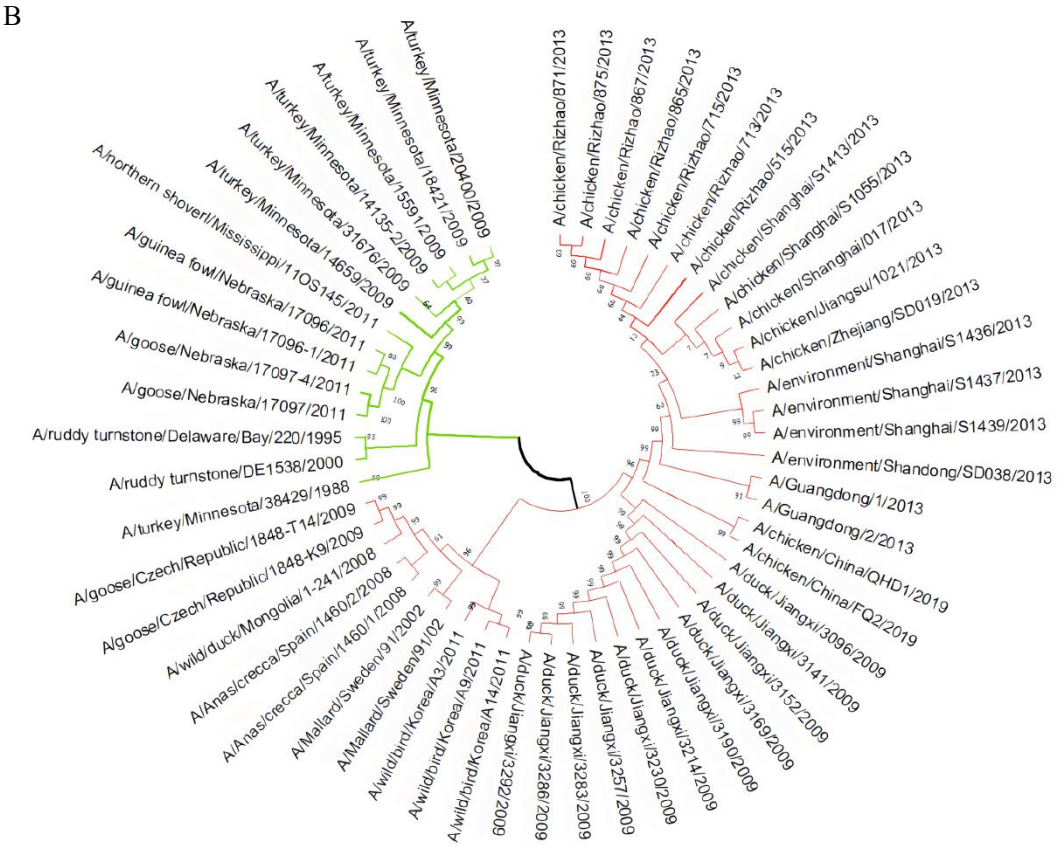
Using the same approach, we constructed a phylogenetic tree using our method (Figure 5A), which is clearly divided into two major lineages: a North American lineage, including viruses from the United States and Guatemala, and a Eurasian lineage, including viruses from Spain, China, South Korea, and Mongolia. In addition, viruses isolated in the same or in temporally proximate years exhibit closer evolutionary relationships, consistent with previous reports [70]. A more detailed examination of the Eurasian lineage further reveals that viruses from geographically proximate regions tend to cluster together, which is also supported by the literature [71, 72].

For comparison, multiple sequence alignment–based phylogenetic reconstruction was also performed using MEGA, and bootstrap analysis was conducted to evaluate the robustness of the inferred clades. The resulting alignment-based phylogenetic tree with bootstrap support is shown in Figure 5B.

A


A/chicken/Rizhao/871/2013
A/chicken/Rizhao/875/2013
A/chicken/Rizhao/867/2013
A/chicken/Rizhao/865/2013
A/chicken/Rizhao/715/2013
A/chicken/Rizhao/713/2013
A/chicken/Rizhao/515/2013
A/chicken/Shanghai/017/2013
A/chicken/Shanghai/S1413/2013
A/chicken/Shanghai/S1055/2013
A/chicken/Zhejiang/SD019/2013
A/chicken/Jiangsu/1021/2013
A/environment/Shanghai/S1436/2013
A/environment/Shanghai/S1437/2013
A/environment/Shanghai/S1439/2013
A/environment/Shandong/SD038/2013
A/Guangdong/1/2013
A/Guangdong/2/2013
A/chicken/China/QHD1/2019
A/chicken/China/FQ2/2019
A/duck/Jianxi/3257/2009
A/duck/Jianxi/3096/2009
A/duck/Jianxi/3141/2009
A/duck/Jianxi/3152/2009
A/duck/Jianxi/3169/2009
A/duck/Jianxi/3190/2009
A/duck/Jianxi/3214/2009
A/duck/Jianxi/3230/2009
A/duck/Jianxi/3283/2009
A/duck/Jianxi/3286/2009
A/duck/Jianxi/3292/2009
A/wild/bird/Korea/A3/2011
A/wild/bird/Korea/A9/2011
A/wild/bird/Korea/A14/2011
A/Mallard/Sweden/91/02
A/Mallard/Sweden/91/2002
A/Anas/crecca/Spain/1460/1/2008
A/Anas/crecca/Spain/1460/2/2008
A/wild/duck/Mongolia/1-241/2008
A/goose/Czech/Republic/1848-K9/2009
A/goose/Czech/Republic/1848-T14/2009
A/turkey/Minnesota/38429/1988
A/ruddy turnstone/DE1538/2000
A/ruddy turnstone/Delaware/Bay/220/1995
A/goose/Nebraska/17097/2011
A/goose/Nebraska/17097-4/2011
A/guinea fowl/Nebraska/17096-1/2011
A/guinea fowl/Nebraska/17096/2011
A/northern shoveI/Mississippi/11OS145/2011
A/turkey/Minnesota/14659/2009
A/turkey/Minnesota/18421/2009
A/turkey/Minnesota/20400/2009
A/turkey/Minnesota/31676/2009
A/turkey/Minnesota/14135-2/2009
A/turkey/Minnesota/15591/2009

B


A/chicken/Rizhao/871/2013
A/chicken/Rizhao/875/2013
A/chicken/Rizhao/867/2013
A/chicken/Rizhao/865/2013
A/chicken/Rizhao/715/2013
A/chicken/Rizhao/713/2013
A/chicken/Rizhao/515/2013
A/chicken/Shanghai/S1413/2013
A/chicken/Shanghai/S1055/2013
A/chicken/Shanghai/017/2013
A/chicken/Jiangsu/1021/2013
A/chicken/Zhejiang/SD019/2013
A/environment/Shanghai/S1436/2013
A/environment/Shanghai/S1437/2013
A/environment/Shanghai/S1439/2013
A/environment/Shandong/SD038/2013
A/Guangdong/1/2013
A/Guangdong/2/2013
A/chicken/China/QHD1/2019
A/chicken/China/FQ2/2019
A/duck/Jiangxi/3096/2009
A/duck/Jiangxi/3141/2009
A/duck/Jiangxi/3152/2009
A/duck/Jiangxi/3169/2009
A/duck/Jiangxi/3190/2009
A/duck/Jiangxi/3214/2009
A/duck/Jiangxi/3230/2009
A/duck/Jiangxi/3257/2009
A/duck/Jiangxi/3283/2009
A/duck/Jiangxi/3286/2009
A/duck/Jiangxi/3292/2009
A/wild/bird/Korea/A3/2011
A/wild/bird/Korea/A9/2011
A/wild/bird/Korea/A14/2011
A/Mallard/Sweden/91/02
A/Mallard/Sweden/91/2002
A/Anas/crecca/Spain/1460/1/2008
A/Anas/crecca/Spain/1460/2/2008
A/wild/duck/Mongolia/1-241/2008
A/goose/Czech/Republic/1848-K9/2009
A/goose/Czech/Republic/1848-T14/2009
A/turkey/Minnesota/38429/1988
A/ruddy turnstone/DE1538/2000
A/ruddy turnstone/Delaware/Bay/220/1995
A/goose/Nebraska/17097/2011
A/goose/Nebraska/17097-4/2011
A/guinea fowl/Nebraska/17096-1/2011
A/guinea fowl/Nebraska/17096/2011
A/northern shoveI/Mississippi/11OS145/2011
A/turkey/Minnesota/14659/2009
A/turkey/Minnesota/31676/2009
A/turkey/Minnesota/14135-2/2009
A/turkey/Minnesota/15591/2009
A/turkey/Minnesota/18421/2009
A/turkey/Minnesota/20400/2009

274

Figure 5. Phylogenetic trees of NA genes of H7N9 influenza virus. **A:** Phylogenetic tree generated using our method. **B:** Phylogenetic tree generated using the ClustalW method.

4 Discussion

In bioinformatics, sequence analysis methods can generally be categorized into two major classes: alignment-based algorithms [73, 74] and alignment-free algorithms [75–77]. Alignment-based approaches encounter substantial difficulties when reconstructing phylogenetic relationships among homologous genes with low sequence similarity or extensive genome rearrangements. In addition, their high computational complexity often restricts their application to short sequences, single genes, or relatively small datasets, limiting their suitability for whole-genome phylogenetic analysis.

Alignment-free methods based on *k*-mers partially overcome these limitations by avoiding sequence alignment altogether. However, most existing *k*-mer–based approaches still neglect several important aspects: (1) the intrinsic sequential relationships among *k*-mers, (2) the physicochemical properties of DNA sequences, and (3) the unequal contributions of different *k*-mer features to evolutionary distance estimation.

In this study, we propose a novel alignment-free framework that integrates DNA physicochemical characteristics with *k*-mer representations, together with an attribute reduction strategy to address feature redundancy and weighting issues. By combining sequence-order–aware *k*-mer word vectors with rough set–based feature selection, the proposed method captures biologically meaningful evolutionary signals while maintaining computational efficiency. For example, for sequences longer than 10,000 bp, graphical representation methods reported in the literature require more than 5 minutes [78], whereas our method completes the task in only 4 seconds.

Despite these advantages, the proposed method has several limitations. First, compared with graphical representation methods, it does not provide an intuitive visual depiction of sequence differences, which may limit direct qualitative interpretation. Second, although the rough set–based feature selection effectively reduces dimensionality, the choice of parameters such as *k* and feature size may still influence performance and requires empirical tuning. Finally, like most alignment-free methods, the approach does not explicitly model recombination events, which may affect phylogenetic inference in highly recombinant genomes.

Future work will focus on integrating graphical visualization techniques with the proposed framework to enhance interpretability, as well as extending the method to explicitly account for recombination and other complex evolutionary processes. By combining the strengths of physicochemical modeling, *k*-mer statistics, and visualization-based approaches, we aim to further improve the robustness and applicability of alignment-free phylogenetic analysis.

5 Conclusions

In non-alignment models for sequence comparison, the *k*-mer frequency models are well established. Nevertheless, the majority of current word-based models fail to account for the correlations among *k*-mer frequencies, and those that do take these correlations into account often overlook the *k*-mer frequencies themselves.In this study, we present a new *k*-mer model that effectively tackles both of these challenges.

We construct a 7×2^k-dimensional complete *k*-mer vector for DNA sequences and apply rough set theory to design a feature selection process that extracts the most informative *k*-mers for DNA sequence representation. To assess the performance of our approach, we performed phylogenetic analysis on four datasets. One dataset served as the training set, from which 3005 key features were chosen based on their significance, while the other datasets were used as test sets. The results demonstrate that our method effectively captures essential information and provides a feasible framework for molecular phylogenetic analysis.

Conflicts of interest

The authors report no conflicts of interest relevant to this article.

Acknowledgements

We would like to extend our appreciation to Dr. Xianchao Guo and Dr. Meng Li from Beijing Mengya Runze Biotechnology Co., Ltd.for their insightful feedback in enhancing our approach.

Contributors

All authors participated in the study design and data analysis. They also contributed to data interpretation by reviewing later versions of the manuscript. All authors have read and approved the final manuscript.

Funding

The research presented in this paper was financially supported by the Liaoning Provincial Natural Science Foundation for Doctoral Research Start-up Projects (2025-BS-0646), the Fundamental Research Funds for the Central University (SCU2025D005), the 2022 General Project of Liaoning Province General Higher Education Undergraduate Teaching Reform Research (SJJGYB20220503), the Liaoning Provincial Economic and Social Development Research Project (20261s1ybkt-005), the Education and Teaching Research and Reform Project of Jinzhou Medical University (YD2025046), Key Project of the Social Committee of “Liao San Hui” (LSHXHL2025-002) and the Key Project of the Liaoning Provincial Department of Education (LJ112410160013).

Data availability statement

The entire dataset pertinent to the research is presented within the paper, with additional data available as supporting material.Given that this analysis is based on publicly accessible data and does not involve any direct interaction with human subjects, it was determined that ethical approval was not required.

References

1. Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu S., et al. Pan-Genome of Wild and Cultivated Soybeans. *Cell.*, **2020,** 182(1), 162-176.
2. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., et al. The complete sequence of a human genome. *Science.,* **2022,** 376(6588)*,* 44-53.
3. Jian, C., Zijian, W., Kaiwen, T., Wei, H., Junpeng, Shi., Tong, L., et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.,* **2023,** 55(7), 1221-1231.
4. Shen, F., Xu, S., Shen, Q., Bi, C., Lysak, M.A. The allotetraploid horseradish genome provides insights into subgenome diversification and formation of critical traits. *Nat. Commun.,* **2023,** 14(1), 4102.
5. Zhang, C., Xie, L., Yu, H., Wang, J.H., Chen, Q.S., Wang, H. The T2T genome assembly of soybean cultivar ZH13 and its epigenetic landscapes. *Mol. Plant.,* **2023,** 16(11)*,* 1715-1718.
6. Wang, T., Wang, B., Hua, X., Tang, H.B., Zhang, Z.Y., Gao, R.T., et al. A complete gap-free diploid genome in *Saccharum* complex and the genomic footprints of evolution in the highly polyploid *Saccharum* genus. *Nat. Plants.,* **2023,** 9(4), 554-571.
7. He, Y., Chu, Y., Guo, S., Hu, J., Li, R., Zheng, Y., et al. T2T-YAO: A telomere-to-telomere assembled diploid reference genome for Han Chinese. *Genom. Proteom. Bioinf.,* **2023,** 21(6), 1085-1100.
8. Shi, X., Cao, S., Wang, X., Huang, S., Wang, Y., Liu, Z., et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Hortic. Res.,* **2023,** 10(5), uhad061.
9. Shi, J., Tian, Z., Lai, J., Huang, X. Plant pan-genomics and its applications. *Mol. Plant.,* **2023,** 16(1), 168-186.
10. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., et al. The complete sequence of a human Y chromosome. *Nature.,* **2023,** 621(7978), 344-354.
11. Bai, M., Jiang, S., Chu, S., Yu, Y., Shan, D., Liu, C., et al. The telomere-to-telomere (T2T) genome of Peucedanum praeruptorum Dunn provides insights into the genome evolution and coumarin biosynthesis. *Gigascience.,* **2024,** 13, giae025.
12. Li, M., Chen, C., Wang, H., Qin, H., Hou, S., Yang, X., et al. Telomere-to-telomere genome assembly of sorghum. *Sci. Data.,* **2024,** 11(1), 835.
13. Nguinkal, J.A., Zoclancloun, Y.A.B., Brunner, R.M., Chen, Y., Goldammer, T. Haplotype-resolved and near-T2T genome assembly of the African catfish (*Clarias gariepinus*). *Sci. Data.,* **2024,** 11(1), 1095.
14. Yang, C., Zhou, Y., Song, Y., Wu, D., Zeng, Y., Nie, L., et al. The complete and fully-phased diploid genome of a male Han Chinese. *Cell. Res.,* **2023,** 33(10), 745-761.
15. Li, Q., Qiao, X., Li, L., Gu, C., Yin, H., Qi, K., et al. Haplotype-resolved T2T genome assemblies and pangenome graph of pear reveal diverse patterns of allele-specific expression and the genomic basis of fruit quality traits. *Plant Commun.*, **2024,** 5(10), 101000.

16. Garg, V., Bohra, A., Mascher, M., Spannagl, M., Xu, X., Bevan, M.W., Bennetzen, J.L., Varshney, R.K. Unlocking plant genetics with telomere-to-telomere genome assemblies. *Nat. Genet.,* **2024,** 56(9), 1788-1799.
17. Wang, Y., Dong, M., Wu, Y., Zhang, F., Ren, W., Lin, Y., et al. Telomere-to-telomere and haplotype-resolved genome of the kiwifruit Actinidia eriantha. *Mol. Hortic.,* **2023,** 3(1), 4.
18. Tao, Y., He, C., Lin, D., Gu, Z., Pu, W. Comprehensive identification of mitochondrial pseudogenes (NUMTs) in the human telomere-to-telomere reference genome. *Genes (Basel).,* **2023,** 14(11), 2092.
19. Feng, Y., Zhou, J., Li, D., Wang, Z., Peng, C., Zhu, G. The haplotype-resolved T2T genome assembly of the wild potato species Solanum commersonii provides molecular insights into its freezing tolerance. *Plant Commun.,* **2024,** 5(10), 100980.
20. Tong, Z., Huang, Y., Zhu, Q.H., Fan, L., Xiao, B., Shen, E. Retrospect and prospect of *Nicotiana tabacum* genome sequencing. *Front. Plant. Sci.,* **2024,** 15, 1474658.
21. Chao, K.H., Zimin, A.V., Pertea, M., Salzberg, S.L. The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual. *G3 (Bethesda).,* **2023,** 13(3), 321.
22. Wang, Y.H., Liu, P.Z., Liu, H., Zhang, R.R., Liang, Y., Xu, Z.S., et al. Telomere-to-telomere carrot (*Daucus carota*) genome assembly reveals carotenoid characteristics. *Hortic. Res.,* **2023,** 10(7), 103.
23. Yang, H., Lian, C., Liu, J., Yu, H., Zhao, L., He, N., et al. High-quality assembly of the T2T genome for Isodon rubescens *f. lushanensis* reveals genomic structure variations between 2 typical forms of Isodon rubescens. *Gigascience.,* **2024,** 13*,* giae075.
24. Kalbfleisch ,T.S., McKay, S.D., Murdoch, B.M., Adelson, D.L., Almansa-Villa, D., Becker, G., et al. The Ruminant Telomere-to-Telomere (RT2T) Consortium. *Nat. Genet.,* **2024,** 56(8), 1566-1573.
25. Wang, L., Li, L.L., Chen, L., Zhang, R.G., Zhao, S.W., Yan, H., et al. Telomere-to-telomere and haplotype-resolved genome assembly of the Chinese cork oak (*Quercus variabilis*). *Front. Plant. Sci.,* **2023,** 14*,* 1290913.
26. Mo, C., Wang, H., Wei, M., Zeng, Q., Zhang, X., Fei, Z., et al. Complete genome assembly provides a high-quality skeleton for pan-NLRome construction in melon. *Plant J.,* **2024,** 118(6), 2249-2268.
27. Gao, Z., Lu, Y., Chong, Y., Li, M., Hong, J., Wu, J., et al. Beef cattle genome project: advances in genome sequencing, assembly, and functional genes discovery. *Int. J. Mol. Sci.,* **2024,** 25(13), 7147.
28. Zhao, H., Zhou, H., Sun, G., Dong, B., Zhu, W., Mu, X., et al. Telomere-to-telomere genome assembly of the goose Anser cygnoides. *Sci. Data.,* **2024,** 11(1), 741.
29. Manley, B.F., Lotharukpong, J.S., Barrera-Redondo, J., Llewellyn, T., Yildirir, G., Sperschneider, J., et al. A highly contiguous genome assembly reveals sources of genomic novelty in the symbiotic fungus Rhizophagus irregularis. *G3 (Bethesda).,* **2023,** 13(6), jkad077.
30. Sun, M., Yao, C., Shu, Q., He, Y., Chen, G., Yang, G., et al. Telomere-to-telomere pear (*Pyrus pyrifolia*) reference genome reveals segmental and whole genome duplication driving genome evolution. *Hortic. Res.,* **2023,** 10(11), uhad201.
31. Wei, M., Huang, Y., Mo, C., Wang, H., Zeng, Q., Yang, W., et al. Telomere-to-telomere genome assembly of melon (*Cucumis melo* L. var. *inodorus*) provides a high-quality reference for meta-QTL analysis of important traits. *Hortic. Res.,* **2023,** 10(10), 189.

32. Zheng, Y., Yang, D., Yin, X., Yang, X., Chen, M., Li, X., et al. The chromosome-level genome assembly of *Cananga odorata* provides insights into its evolution and terpenoid biosynthesis. *New Phytol.,* **2024,** 243(6), 2279-2294.
33. Otu H.H., Sayood K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*., **2003,** 19(16), 2122-30.
34. Randić, M.; Balaban, A.T. On A Four-Dimensional Representation of DNA Primary Sequences. *J. Chem. Inf. Comptu. Sci.,* **2003,** 43(2), 532-539.
35. Yang, Y.; Zhang, Y.; Jia, M.; Li, C.; Meng, L. Non-degenerate graphical representation of DNA sequences and its applications to phylogenetic analysis. *Comb. Chem. High Throughput Screen.,* **2013,** 16(8), 585-9.
36. Li, C., Li, Z.X., Zheng, X.Q., Ma, H., Yu, X.Q., A generalization of Lempel-Ziv complexity and its application to the comparison of protein sequences. *J. Math. Chem.,* **2010,** 48, 330-338.
37. Yu, C., He, R.L., Yau, S.S. Viral genome phylogeny based on Lempel-Ziv complexity and Hausdorff distance. *J Theor Biol.,* **2014,** 348, 12-20.
38. Munagala, N.V.T.S., Amanchi, P.K., Balasubramanian, K., Panicker, A., Nagaraj, N. Compression-complexity measures for analysis and classification of coronaviruses. *Entropy (Basel),* **2022,** 25(1), 81.
39. Li, H., Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* **2009,** 25(14), 1754-60.
40. Naseri, A., Zhi, D., Zhang, S. Multi-allelic positional Burrows-Wheeler transform. *BMC Bioinformatics,* **2019,** 20(Suppl 11), 279.
41. Wertenbroek, R., Xenarios, I., Thoma, Y., Delaneau, O. Exploiting parallelization in positional Burrows-Wheeler transform (PBWT) algorithms for efficient haplotype matching and compression. *Bioinform Adv.,* **2023,** 3(1), vbad021.
42. Orlov, Y.L., Orlova, N.G. Bioinformatics tools for the sequence complexity estimates. *Biophys Rev.,* **2023,** 15(5), 1367-1378.
43. Yang, X., Wang, T. A novel statistical measure for sequence comparison on the basis of *k*-word counts. *J Theor Biol.,* **2013,** 318, 91-100.
44. Yu, X., Zheng, X., Meng, L., Li, C., Wang, J. A support vector machine based method to predict success for polymerase chain reactions. *Comb. Chem. High Throughput Screen.,* **2012,** 15(6), 486-91.
45. Tang, R., Yu, Z., Li, J. KINN: An alignment-free accurate phylogeny reconstruction method based on inner distance distributions of *k*-mer pairs in biological sequences. *Mol. Phylogenet. Evol.,* **2023,** 179, 107662.
46. Chow, C.F.W., Ghosh, S., Hadarovich, A., Toth-Petroczy, A. SHARK enables sensitive detection of evolutionary homologs and functional analogs in unalignable and disordered sequences. *Proc. Natl. Acad. Sci. USA.,* **2024,** 121(42), e2401622121.
47. Shi, G., Dai, Y., Zhou, D., Chen, M., Zhang, J., Bi, Y., Liu, S., Wu Q. An alignment- and reference-free strategy using *k*-mer present pattern for population genomic analyses. *Mycology.,* **2024,** 16(1), 309-323.
48. Islam, R., Rahman, A. An alignment-free method for detection of missing regions for phylogenetic analysis. *Heliyon*, **2024,** 10(11), e32227.

49. Van, Etten, J., Stephens, T.G., Bhattacharya, D. A *k*-mer-Based Approach for Phylogenetic Classification of Taxa in Environmental Genomic Data. *Syst. Biol.,* **2023,** 72(5), 1101-1118.
50. Moore, M.P., Laager, M., Ribeca, P., Didelot, X. KmerAperture: Retaining *k*-mer synteny for alignment-free extraction of core and accessory differences between bacterial genomes. *PLoS Genet.,* **2024,** 20(4), e1011184.
51. Rachtman, E., Jiang, Y., Mirarab, S. Machine learning enables alignment-free distance calculation and phylogenetic placement using *k*-mer frequencies. *Mol. Ecol. Resour.,* **2025,** 25(8), e70055.
52. Yan, L., Liu, Y., Yang, Y. Predicting polymerase chain reaction success: integrating the *k*-word order model, physicochemical properties modeling of double bases, and support vector machine. *Comb. Chem. High Throughput Screen.,* **2025,** doi: 10.2174/0113862073351071250102100221.
53. Qi, J., Wang, B., Hao, B.I. Whole proteome prokaryote phylogeny without sequence alignment: a *k*-string composition approach. *J. Mol. Evol.,* **2004,** 58(1), 1-11.
54. Xu, Z., Hao, B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res.,* **2009,** 37, W174-8.
55. Leimeister, C.A., Morgenstern, B. Kmacs: the *k*-mismatch average common substring approach to alignment-free sequence comparison. **2014,** *Bioinformatics.,* 30(14), 2000-8.
56. Ondov, B.D., Starrett, G.J., Sappington, A., Kostic, A., Koren, S., Buck, C.B., Phillippy, A.M. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.,* **2019,** 20(1), 232.
57. Zhang, X., Ding, S., Yu, C., Zhao, J., Bu, D. Accurate and efficient phylogenetic inference through end-to-end deep learning. *Mol. Biol. Evol.,* **2025,** 42(11), msaf260.
58. Jiang, Y., Balaban, M., Zhu, Q., Mirarab, S. DEPP: deep learning enables extending species trees using single genes. *Syst. Biol.,* **2023,** 72(1), 17-34.
59. Braichenko, S., Borges, R., Kosiol, C. Phylogenetic methods meet deep learning. *Genome Biol. Evol.,* **2025,** 17(10), evaf177.
60. Skowron, A., Dutta, S. Rough sets: past, present, and future. *Nat. Comput.,* **2018,** 17(4), 855-876.
61. Hu, Q.H., Zhao, H., Xie, Z.X., Yu, D.R. Consistency based attribute reduction, in: PAKDD, in: Lecture Notes in Computer Science, vol. 4426, **2007,** pp. 96–107.
62. Luo, S., Shi, L., Chen, L., Cao, X. Accelerated feature selection via discernibility hashing: a rough set approach. *Entropy (Basel)*, **2025,** 27(12):1222.
63. Wang, C., Chen, D., Hu, Q. Some invariant properties of ordered information systems under homomorphism. *Sci. China Inf. Sci.,* **2010,** 53, 1816–1825.
64. Wang, P., Zhao, Z., Wang, Z.L.Z. Fuzzy set-valued information systems and their homomorphisms based on data compression. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, **2022,** 43(3), 3311-3321.
65. Li, C., Yang, Y., Jia, M.D., Zhang, Y.Y., Yu, X.Q., Wang, C.Z. Phylogenetic analysis of DNA sequences based on *k*-word and rough set theory, *Physica A,* **2014,** 398, 162-171.
66. Yu, Z.G., Anh, V., Lau, K.S. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J. Theor. Biol.,* **2004,** 226(3), 341-8.
67. Vinga, S., Almeida, J. Alignment-free sequence comparison-a review. *Bioinformatics.,* **2003,** 19(4), 513-23.

68. Zheng, W.X., Chen, L.L., Qu, H.Y., Gao, F., Zhang, C.T.. Coronaviruses phylogeny based on a geometric approach. *Mol. Phylogenet. Evol.,* **2005,** 2, 224-232.
69. Luk, H.K.H, Li, X., Fung, J., Lau, S.K.P., Woo, P.C.Y. Molecular epidemiology, evolution and phylogeny of SARS coronavirus. *Infect. Genet. Evol.,* **2019,** 71, 21-30.
70. Sun, Y.T., Jin, N.Z., Zhang, H.R., Shen, Y.L., Ma, Y., Peng, T.S. Evolutionary analysis of neuraminidase gene of A/H7N9 influenza virus. *Chinese journal of virology,* **2014,** 30, 44-50.
71. Butler, D. Novel bird flu kills two in China, *Nature,* **2013,** 2, 2
72. Liu, Y., Chen, Y., Yang, Z., Lin, Y., Fu, S., Chen, J., Xu, L., Liu, T., Niu, B., Huang, Q., Liu, H., Zheng, C., Liao, M., Jia, W. Evolution and antigenic differentiation of avian influenza A(H7N9) virus, China. *Emerg Infect Dis.,* **2024,** 30(6), 1218-1222.
73. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Research*, **1988,** 16(22), 10881-10890.
74. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research,* **2004,** 32(5), 17921797.
75. Vinga, S., Almeida, J. Alignment free sequence comparison—a review. *Bioinformatics,* **2003,** 19(4), 513-523.
76. Ramanathan, N., Ramamurthy, J., Natarajan, G. Numerical characterization of DNA sequences for alignment-free sequence comparison - A review. *Comb. Chem. High. Throughput. Screen.,* **2022,** 25(3), 365-380.
77. Anjum, N., Nabil, R.L., Rafi, R.I., Bayzid, M.S., Rahman, M.S. CD-MAWS: an alignment-free phylogeny estimation method using cosine distance on minimal absent word sets. *IEEE/ACM Trans Comput Biol Bioinform.,* **2023,** 20(1), 196-205.
78. Dong, L., Jiang, X., Liu, Y., Gao, Y., Yang, Y. Improved visualization method of DNA sequences and its application in phylogenetic analysis. *Comb. Chem. High. Throughput Screen.,* **2025,** doi: 10.2174/0113862073379972250612103433.