A new alignment-free method: Subsequence Correlation Coefficient Vector(SCCFV) for influenza A comparison using virus genomes

Lily He[a,*], Zhenglong Yu[a,**], Xinrui Wu[a,**], Yi Zhu[a,**]

[a]*School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, P.R. China*

Abstract

Influenza viruses remain a formidable threat to global public health due to their high mutability and infectivity. Accurate prediction of influenza virus subtypes is crucial for clinical treatment and disease prevention. In recent years, machine learning methods have played an important role in studying influenza viruses. This study proposes a new alignment-free method based on the correlation of k-grams called Subsequence Correlation Coefficient Vector (SCCFV) to subtype hemagglutinin (HA) and neuraminidase (NA) of influenza virus. In the method, each influenza virus sequence is converted to four time series and the correlation coefficients of time series are utilized to extract the features of sequences. Then the supervised learning methods are used for the subtype classification of influenza viruses. We compare the effectiveness of the random forest, decision tree and support vector machine classifiers. Experimental results show that the random forest method achieves the best performance with an accuracy of 0.99979, an precision of 0.99996 and a recall of 0.99997. All prediction indicators of our method are significantly higher than traditional methods.

Keywords: Alignment-free, SCCFV, Influenza A virus, Machine learning

1. Introduction

One of the key respiratory infectious diseases affecting humans is influenza[1]. Influenza viruses have acute and high infectivity and belong to the Orthomyxoviridae family[2]. The virus

*Corresponding author

**These authors contributed equally to this work.

Email address: lilyhe6@163.com (Lily He)

was first identified in 1932 and has been a major focus of virus research in recent years [3].According to the differences in nucleocapside protein (NP) and matrix protein (MP), it can be divided into three types : type A, type B and type C [4]. Influenza C virus contains 7 RNA segments [5, 6], whereas influenza A and B [7] viruses both have 8 RNA segments. Influenza A virus has attracted widespread attention due to its capacity for cross-species transmission and high variability. Additionally, it exhibits a broad range of host adaptability, whose hosts include humans, wild birds, domestic poultry, and a variety of mammals [8, 9].The virus poses a serious threat to human and animal health mainly due to the antigenic drift and antigenic transfer of its surface glycoprotein, which makes it easy to mutate across different hosts [10, 11].

Hemolysin (HA) and neuraminidase (NA) are glycoproteins on the surface of influenza virus. The subtype of influenza virus is determined by the types of these two type of proteins. For influenza A viruses, 16 HA subtypes [H1-H16] and 9 NA subtypes [N1-N9] have been identified based on antigenic differences [12] . For determining the subtype classification of influenza virus HA and NA fragments, the most commonly used methods by WHO are the hemolysis inhibition (HI) test and the neuraminidase inhibition (NI) test. Although this subtype identification method is relatively economical compared to molecular detection, it still requires a lot of experiments when dealing with unknown virus subtype, which will consume a lot of time and materials[13]. Machine learning methods can provide a feasible and resource efficient option for the subtype classification problem, which may have higher efficiency and significantly lower cost compared to traditional methods.

In the past decades of research on influenza virus, machine learning methods have been used to infer certain pathogenic markers in the hemagglutinin (HA) gene [14] . To classify the pathogenicity classification of H5Nx avian influenza strains, Akshay Chadha et al. [15]compared the performance of different machine learning classifiers such as logistic regression (LR) with lasso and ridge regularization, random forest (RF), K-nearest neighbor (KNN), naive Bayes (NB), support vector machine (SVM) and convolutional neural networks (CNN). Edyta Swiezton et al. [16] used Bayesian methods for phylogenetic and molecular analysis of H5N8 and H5N5 viruses.

In the research of Fahad Humayun et al.[17], a statistical method for predicting avian influenza virus subtypes was proposed, which made a great contribution to predicting avian influenza subtypes through feature extraction and machine learning methods. However, there are few methods for classifying the subtype of Avian Influenza A virus. Even if there is a methodology, the prediction effect needs to be improved due to the way it is encoded.

To improve the accuracy of influenza virus subtype prediction, we adopted a novel feature extraction method inspired by the correlation coefficients of time series of DNA sequences. Subsequently, based on the features extracted, we performed influenza virus subtype classification by machine learning methods. After comparing four different methods: random forest, decision tree, SVM and KNN, we finally found that random forest showed the highest accuracy in predicting influenza virus subtypes.

2. Materials and methods

2.1. Hemagglutinin and Neuraminidase Subtype sequences

All Hemagglutinin(HA) subtype sequences from 2002 to 2022 years stored in NCBI were downloaded by Influenza Virus Tool (http:/www.ncbi.nlm.nih.gov/
genomes/FLU). We retained nucleotide sequences with a sequence length of at least 1600 bp. The virus subtypes containing only a few sequences (less than 10 sequences) were also eliminated. The resulted data set included 95,585 influenza virus sequences with subtypes from H1 to H14, and H16. The distribution of sequence length was listed in Table 6. Among these sequences, the 70% of them were utilized for model training, while the remaining 30% were utilized for model test.

Table 1: Summary of the HA sequence data.

Subtype	Number	Min Length (bp)	Max Length (bp)
H1	33044	1683	1938
H2	428	1689	1805
H3	40563	1671	1867
H4	2020	1695	1789
H5	7210	1683	1870
H6	1823	1695	1780
H7	2773	1659	1763
H8	158	1701	1745
H9	4456	1671	1805
H10	1227	1686	1730
H11	683	1698	1772
H12	312	1695	1737
H13	562	1698	1770
H14	26	1707	1749
H16	299	1698	1776

All Neuraminidase(NA) subtype sequences from 2002 to 2022 years in NCBI were also downloaded. The DNA sequences with at least 1300 bp were kept. The virus subtypes containing only a few sequences were also deleted. Then the dataset we obtained included N1 to N9 subtypes of avian influenza virus, with a total of 88,638 sequences. The distribution of sequence length was listed in Table 7. Among these sequences, the 70% of them were utilized for model training, while the remaining 30% were utilized for model test.

Table 2: Summary of the NA sequence data.

Subtype	Number	Min Length (bp)	Max Length (bp)
N1	30256	1335	2349
N2	46153	1332	1498
N3	1786	1329	1469
N4	335	1413	1463
N5	575	1308	1472
N6	3452	1335	1580
N7	959	1356	1466
N8	3427	1356	1701
N9	1694	1389	1471

2.2. Subsequence Correlation Coefficient Feature Vector (SCCFV)

To intricately examine the distribution of nucleotides within the DNA sequence, we initially transform the DNA sequence into a corresponding feature vector in multi-dimensional space. For a virus DNA sequence $X = (x_1, x_2, \ldots, x_n)$, with each x_i belonging to the set $\Phi = \{G, T, A, C\}$,$i = 1, 2 \cdots, n$, and n denoting the length of the sequence, we transform the sequence into four time series as follows: Let $V_t = v_t(1)v_t(2) \cdots v_t(N)$, $t \in \{G, T, A, C\}$

$$v_t(i) = \begin{cases} 1, & x_i = t \\ 0, & x_i \neq t \end{cases} \quad i = 1, 2, \cdots, n. \tag{1}$$

We define the M-step delay V_{t+M} of V_t as: $V_{t+M} = v_t(M+1)v_t(M+2) \cdots v_t(M+n)$,where $v_t(n+1) = \cdots = v_t(M+n) = 0$, especially, $V_t = V_{t+0}$. M is a preset integer ($M \ll N$) which can be adjusted from datasets.

Thus the average value, auto-correlation of one time series and cross-correlation among time series can be computed for the four time series. First, we define the average occurrence number of

$t \in \{G, T, A, C\}$ as:

$$f_t = \frac{1}{N}(v_t(1) + v_t(2) + \ldots + v_t(n)) \tag{2}$$

We illustrate the above formula with a sequence "AGCTGTACCTG":

$V_A = 1\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0$; $V_G = 0\,1\,0\,0\,1\,0\,0\,0\,0\,0\,1$;

$V_{A+2} = 0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0$; $V_{G+2} = 0\,0\,1\,0\,0\,0\,0\,0\,1\,0\,0$;

$f_A = \frac{1}{11}(v_A(1) + v_A(2) + \ldots + v_A(11)) = \frac{1}{11}(1+1) = \frac{2}{11}$;

$f_G = \frac{1}{11}(v_G(1) + v_G(2) + \ldots + v_G(11)) = \frac{1}{11}(1+1+1) = \frac{3}{11}$;

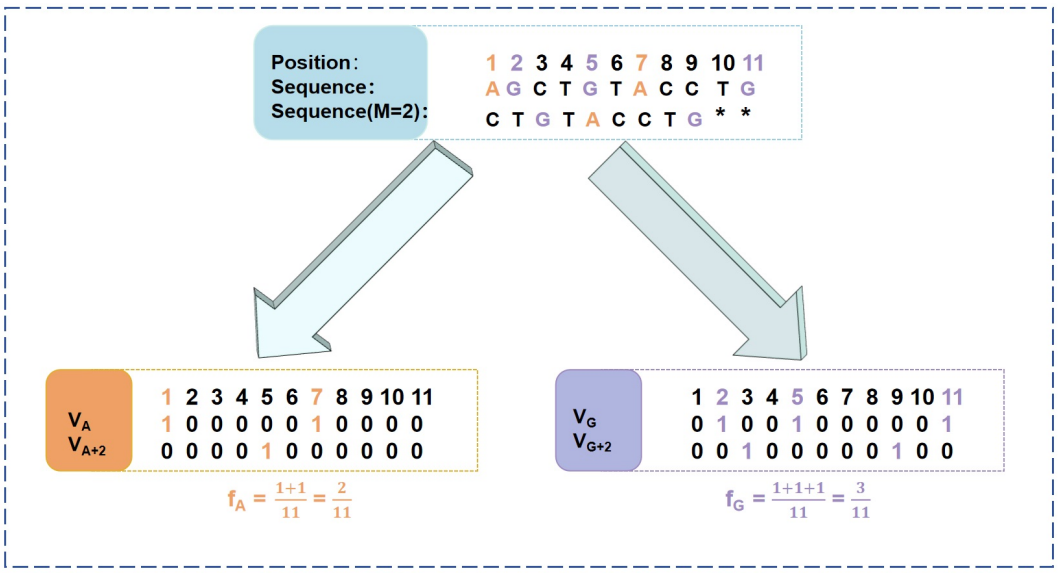

Fig. 1: The computational methodology for determining V_G, V_{G+2}, f_A and f_T.

Please refer to the detailed computation in Figure 6. In order to assess the correlation of nucleotides with a sequence and between sequences, this study conducted computation for the autocorrelation coefficient and cross-correlation coefficient. The M lag autocorrelation $\rho_{cc}(M)$ in a sequence is defined as:

$$\rho_{cc}(M) = \frac{1}{n}\sum_{i=1}^{n}(v_c(i) - f_c)(v_{c+M}(i) - f_c) \tag{3}$$

The M-step delay correlation $\rho_{xc}(M)$ of nucleotides x and c in a sequence can be defined using the

cross-correlation between the two time series v_x and v_c :

$$\rho_{xc}(M) = \frac{1}{n}\sum_{i=1}^{n}(v_x(i) - f_x)(v_{c+M}(i) - f_c) \tag{4}$$

The standardized cross-correlation also called cross-correlation coefficient $\tau_{xc}(M)$ between nucleotides x and c is defined as:

$$\tau_{xc}(M) = \frac{\rho_{xc}(M)}{\sqrt{\rho_{xx}(0) \cdot \rho_{cc}(0)}} \tag{5}$$

Specifically, the autocorrelation coefficient $\tau_{cc}(M)$ of nucleotide c itself can be defined as:

$$\tau_{cc}(M) = \frac{\rho_{cc}(M)}{\rho_{cc}(0)} \tag{6}$$

Subsequently, we obtain a feature vector of nucleotides named the correlation coefficient feature vector [18] with dimensions $16 \times M$:

$$\begin{aligned}
&(\tau_{AA}(1), \tau_{AA}(2), \ldots, \tau_{AA}(M), \tau_{AC}(1), \tau_{AC}(2), \ldots, \tau_{AC}(M),\\
&\tau_{AG}(1), \tau_{AG}(2), \ldots, \tau_{AG}(M), \tau_{AT}(1), \tau_{AT}(2), \ldots, \tau_{AT}(M),\\
&\ldots\ldots\ldots\ldots, \tau_{TT}(1), \tau_{TT}(2), \ldots, \tau_{TT}(M)).
\end{aligned} \tag{7}$$

Drawing upon the methodology introduced by He et al.[19], we systematically apply a procedure to divide sequences into segments with nearly equal length. Consider the virus DNA sequence $X = (x_1, x_2, \ldots, x_n)$, the sequence is segmented into k non-overlapping subsequences, where the initial r segments $(Substr_1, Substr_2, \cdots, Substr_k)$ are comprised of $Z + 1$ nucleotides, and the remaining $k - r$ segments $(Substr_r + 1, Substr_r + 2, \ldots, Substr_L)$ containing z nucleotides. The define of z is the quotient and r is the remainder when dividing n by k.

$$z = [\frac{n}{k}],\ r = n - k * z.\ (0 \leq r < k) \tag{8}$$

For each segment of the sequence, some correlation coefficient feature vectors can be evaluated. Thus the feature vectors can be integrated to a new vector denoted as Subsequence Correlation Coefficient Feature Vector (SCCFV) characterized by a length of $16 \times M \times k$:

$$
\begin{aligned}
&(\tau_{AA}(11), \tau_{AA}(12), \ldots, \tau_{AA}(1M), \tau_{AC}(11), \tau_{AC}(12), \ldots, \tau_{AC}(1M),\\
&\tau_{AG}(11), \tau_{AG}(12), \ldots, \tau_{AG}(1M), \tau_{AT}(11), \tau_{AT}(12), \ldots, \tau_{AT}(1M),\\
&\ldots\ldots\ldots\ldots, \ldots\ldots\ldots\ldots\ldots\ldots \tau_{TT}(11), \tau_{TT}(12), \ldots, \tau_{TT}(1M),\\
&\tau_{AA}(21), \tau_{AA}(22), \ldots, \tau_{AA}(2M), \tau_{AC}(21), \tau_{AC}(22), \ldots, \tau_{AC}(2M),\\
&\tau_{AG}(21), \tau_{AG}(22), \ldots, \tau_{AG}(2M), \tau_{AT}(21), \tau_{AT}(22), \ldots, \tau_{AT}(2M),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \tau_{TT}(21), \tau_{TT}(22), \ldots, \tau_{TT}(2M), \ldots\ldots,\\
&\tau_{AA}(k1), \tau_{AA}(k2), \ldots, \tau_{AA}(kM), \tau_{AC}(K1), \tau_{AC}(k2), \ldots, \tau_{AC}(kM),\\
&\tau_{AG}(k1), \tau_{AG}(k2), \ldots, \tau_{AG}(kM), \tau_{AT}(k1), \tau_{AT}(k2), \ldots, \tau_{AT}(kM),\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \tau_{TT}(k1), \tau_{TT}(k2), \ldots, \tau_{TT}(kM)).
\end{aligned}
$$

2.3. Model Definition and Evaluation

We used three machine learning methods decision tree, random forest and SVM to train our model based on the above feature vectors. The decision tree is a machine learning model based on a tree structure. It constructs a tree by recursively partitioning a dataset, where each node represents a feature, each branch represents a possible value for that feature, and leaf nodes represent the final classification result. The classification of a decision tree is based on a set of rules learned from the features of the input data [20]. Random forest uses multiple decision trees to perform classification or regression, and combines the results of these trees through voting or averaging [21]. The goal of SVM is to find a hyperplane that can divide different classes in the feature space, while maximizing the separation between the sample points (support vectors) closest to the hyperplane for classification [22]. Fig 2 illustrates the pipeline of our method used in this study. We first used

the SCCFV method to convert sequence information into feature vectors, then used three classifiers i.e. random forest, decision tree, and support vector machine to train our model. Once our models were well trained, the prediction of influenza virus subtypes were conducted by the models.

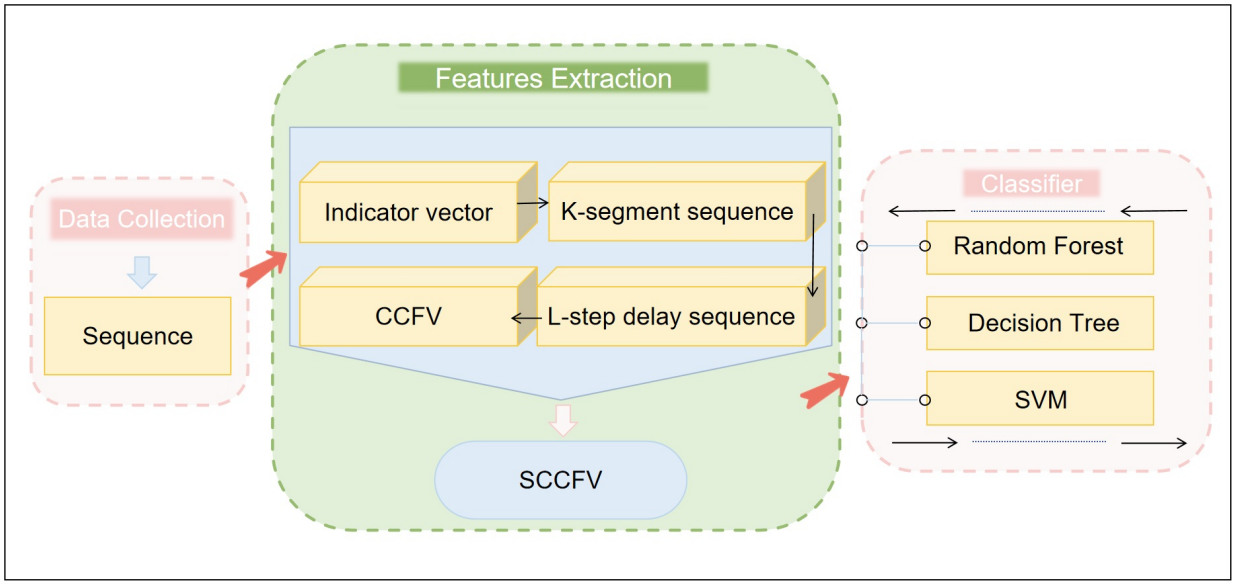

Figure 2: The pipeline of the methodology followed in this study.

In order to evaluate the effectiveness of three machine learning methods in classifying influenza A viruses, this study used four evaluation indicators: Accuracy, Precision, Recall and F1-score. Accuracy represents the proportion of actual positive predicted and negative predicted actual samples to the total sample, Precision represents the proportion of positive predicted actual positive to the predicted positive, Recall represents the proportion of positive predicted actual positive samples to the total number of positive samples, and F1- Score can be regarded as a comprehensive measure of Precision and Recall, which is their harmonic mean. These indicators are selected to comprehensively evaluate the performance of the classification model. Accuracy provides the overall classification accuracy, Precision focuses on the accuracy of positive examples, Recall focuses on the model's coverage ability of positive examples, and F1-score balances the accuracy of Precision and Recall.The calculation formula for these four variables is as follows :

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|} \tag{9}$$

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{10}$$

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{11}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Presion + Recall} \tag{12}$$

The Python source code in this paper is freely available to the public upon request. All calculations in this article were performed on a Dell Inspiron 15 3511 with an Intel Core i5-1135G7 processor, running Microsoft Windows 11 Chinese version, with 16GB RAM.

3. Results

In our work, we classified the HA and NA fragments of influenza A viruses into subtypes respectively. We first extracted features of virus DNA sequences based on the SCCFV method, and then applied three classifiers to obtain our training results. After many classified data tests, we selected k as 5 and M as 2. The specific test results can be referred to Fig 7 .

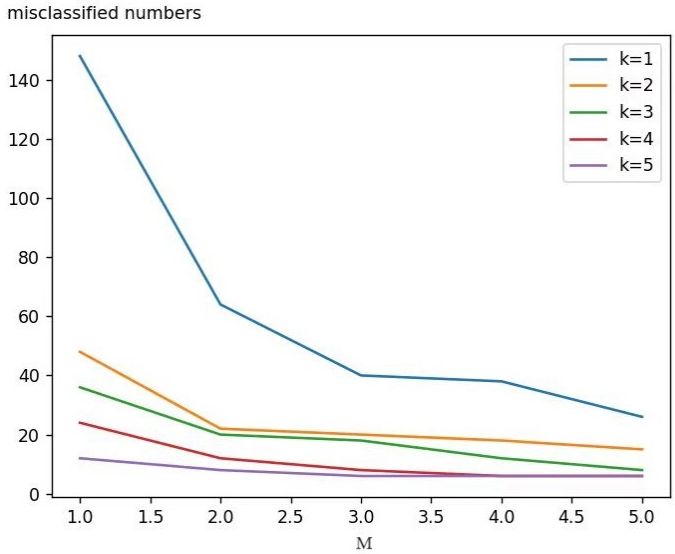

Fig. 3: The number of misclassified samples under different combinations of k and M.

3.1. HA subtype classification

The classifiers were applied to classify the subtype of HA and the results were listed in Table 8. Among the results, Random Forest (referred to as RF) achieves the highest performance with

an accuracy of 0.99979, a precision of 0.99996, a recall of 0.99997, and an F1 score of 0.99996. The decision tree and SVM methods obtain almost same performance. All three methods achieve almost 100% accuracy, precision, recall and F1 score, which ultimately leads to the conclusion that our SCCFV algorithm comprehensively captures the sequence information of HA subtypes.

Table 3: The classification performance of different classifiers on HA subtypes.

	Accuracy	Precision	Recall	F1 score
RF	0.99979	0.99996	0.99997	0.99996
DT	0.99745	0.99756	0.99745	0.99749
SVM	0.99965	0.99994	0.99965	0.99972

To demonstrate the accuracy and reliability of our SCCFV approch, we also constructed a phylogenetic tree of 16 subtypes of HA viruses with 160 HA DNA sequence data. The resulted tree was shown in the following Fig 8 and viruses from the same subtype were labeled with the same color. Notably, in this tree, viruses from the same subtype cluster together, indicating that our method can accurately reconstruct the phylogeny of HA subtype sequences.

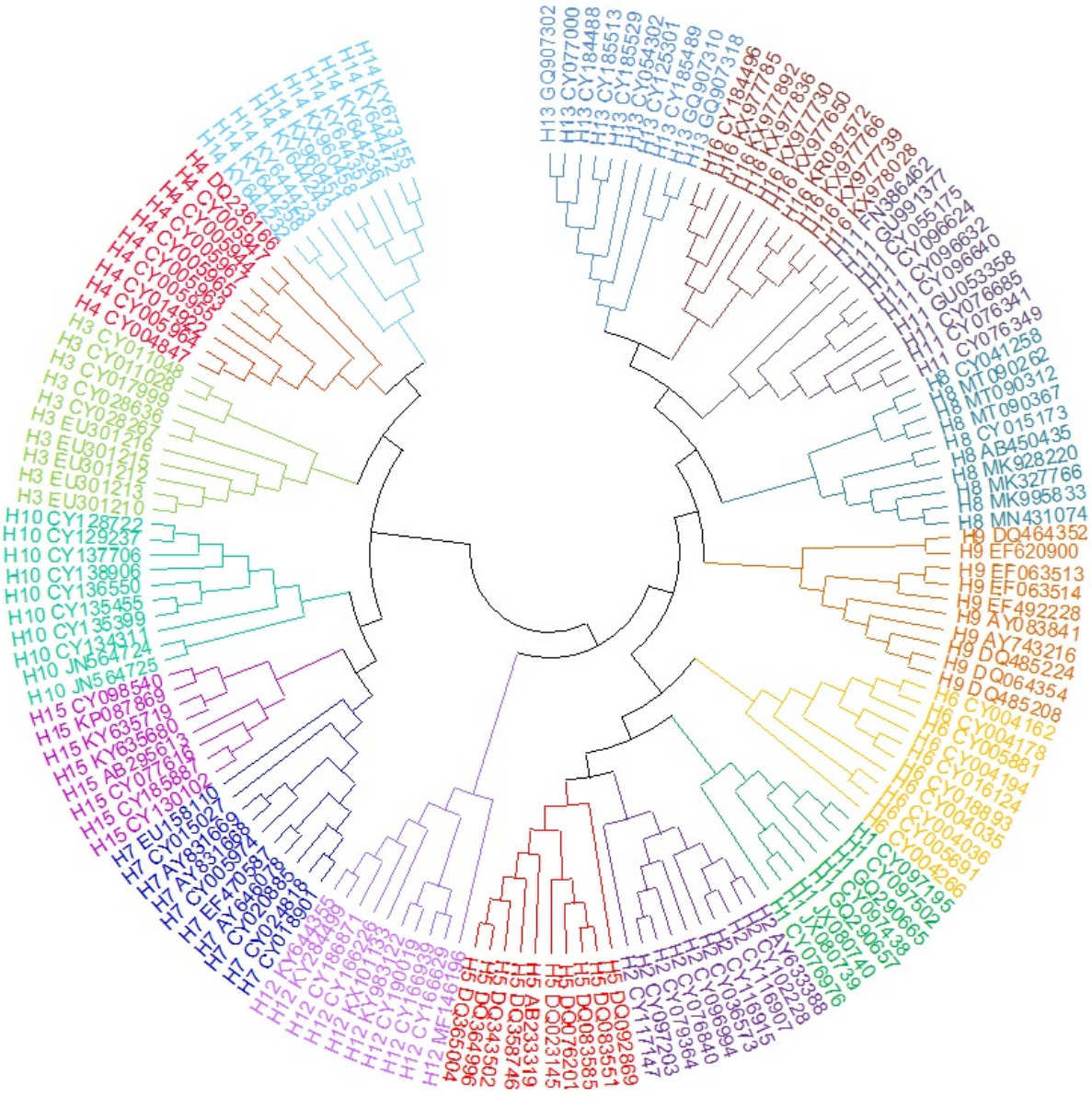

Fig. 4: Phylogenetic tree of 160 HA genes, with 16 subtypes by SCCFV (k=5, M=2).

3.2. NA subtype classification

For the classification of NA subtypes, the Random Forest method still yields the most favorable result, attaining an impressive accuracy of 0.99979, a precision of 0.99996, a recall of 0.99997, and an F1 score of 0.99996. The accuracy, precision, recall and F1 score of the decision tree and SVM both achieve almost 100%. The detailed results are listed in Table 9. From the results of HA and NA, we conclude that the three method can almost 100% accurately classify the HA and NA

subtypes and the RF method has the optimal performance.

Table 4: The classification performance of different classifiers on NA subtypes.

	Accuracy	Precision	Recall	F1 score
RF	0.99979	0.99996	0.99997	0.99996
DT	0.99777	0.99778	0.99777	0.99777
SVM	0.99932	0.99954	0.99932	0.99939

We also used 90 NA data of 9 subtypes to construct a phylogenetic tree and the resulted tree was shown in the following Fig 6 .

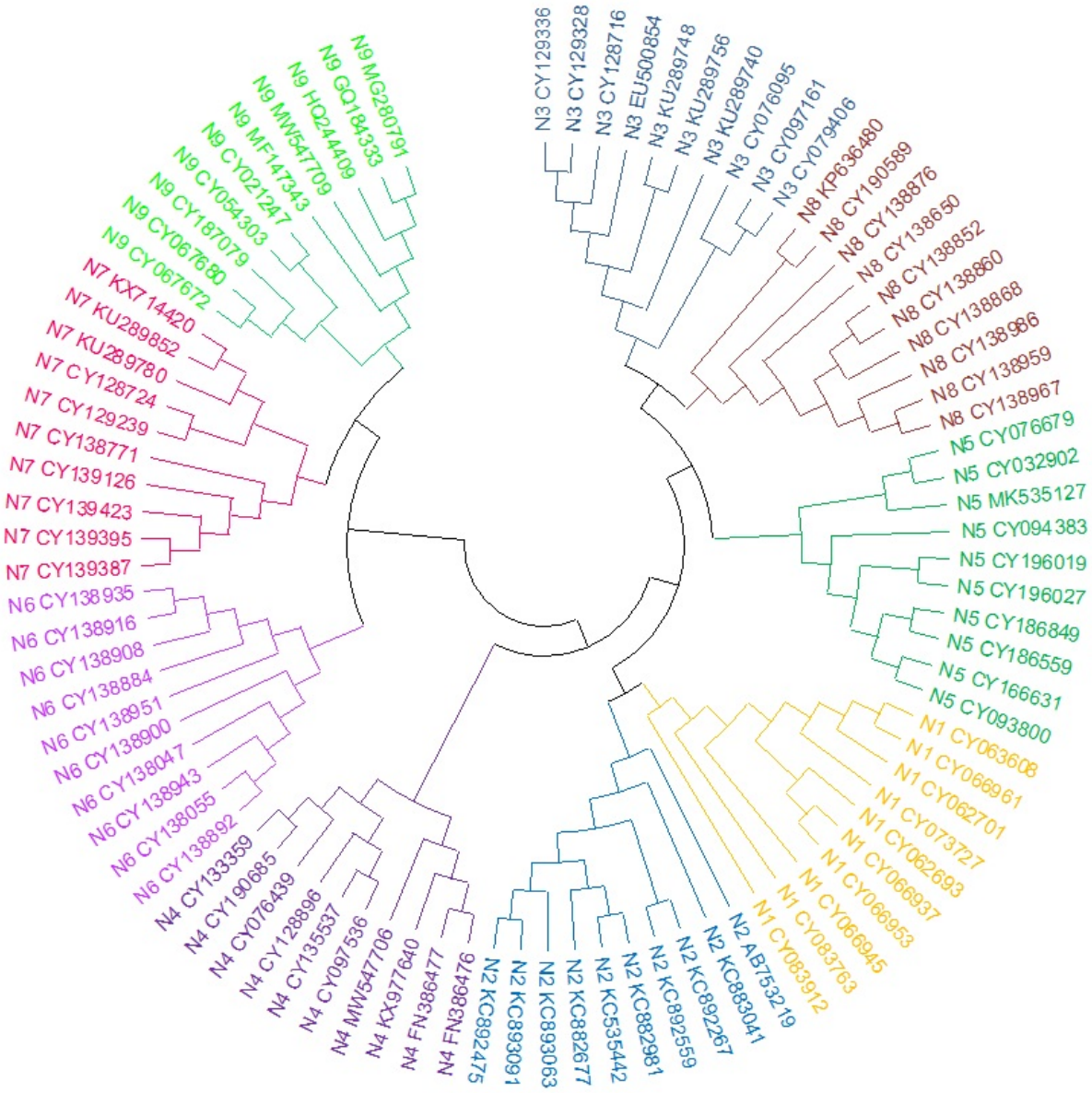

Fig. 5: Phylogenetic tree of 90 NA genes, with 9 subtypes by SCCFV (k=5, M=2).

By observing the phylogenetic tree, we note that taxa of the same subtype tend to cluster together, suggesting that our method may reconstruct the phylogeny of NA subtypes.

In the past, Fahad Humayun used three methods: k-gram, discrete wavelet transform (DWT) and multivariate mutual information (MMI) to extract features of sequences, and then utilized four classifiers to predict the subtype of Type A influenza virus. In his research, the decision tree has the best prediction effect. Compared with Fahad Humayun's experiment, we utilized the same dataset

for the classification of Type A avian influenza. Based on the RF classifier, we classified the Type A avian influenza virus. The classification results were detailed in Table 5. The results indicated that our prediction accuracy was significantly higher than Fahad Humayun's best classification results.

Table 5: Comparison of classification results between our method and F. Humayun's method .

Result \ Method	F's Method	Our Method
Accuracy	0.9571	0.99973
Precision	0.9514	0.99987
Recall	0.9535	0.99973
F1 score	0.9524	0.99973

4. Discussion and conclusions

Since its initial identification, the Type A influenza virus has undergone continuous evolution and widespread transmission, maintaining a significant impact on both human and animal health. Based on the different biological, physical and chemical properties of different Type A influenza virus subtypes, more targeted prevention, control strategies and immune vaccines can be developed. In this work, we studied the subtype classification of Type A influenza virus using 95,585 HA and 88,638 NA DNA sequences. A novel alignment-free method SSCFV was proposed to extract the information contained in the sequences. The approach converted each DNA sequence to four numerical time series and utilized the average, correlation coefficient of them to extract features. To capture the local correlation, the DNA sequence was divided into k segments and the features of each segment were computed. Thus a $16 \times M \times k$ dimensional numerical vector was formed to represent each Type A influenza virus.

Based on the feature vectors of viruses, three machine learning methods random forest, decision tree and SVM were used for the subtype classification of HA and NA. All these methods achieve extremely high prediction effect with almost 100% accuracy, precision, recall, and F1 score. Among them, the random forest perform best. To validate the effect of our SCCFV approach, the phylogenetic trees of HA and NA subtypes were also constructed. In the trees, the viruses

from same subtype cluster together, indicating the method has the potential to reconstruct the evolutionary history of HA and NA subtypes.

In addition, our method was also compared with Fahad Humayun's work for the subtype classification of Type A influenza virus. The prediction results using the random forest classifier based on the SCCFV approach were all superior to Fahad Humayun's experiment. Our method is able to achieve higher classification accuracy in a shorter time.

Besides, our method may have certain limitations. During the preliminary sequence processing, we exclude subtypes containing less than 10 sequences, which may subsequently lead to a certain extent of prediction inaccuracies when it comes to classifying less common subtypes of the Type A influenza virus. However, subtypes with less than 10 sequences can be clustered by phylogenetic trees. In the future, few shot learning methods can be studied to deal with the classification of these less frequently occurring subtypes.

5. Conflict of interest statement

The authors declare no competing financial interests.

6. Acknowledgements

This study is supported by the National Natural Science Foundation of China (Z23146), and Beijing Municipal Education Commission(Z22053 and Z22055) for providing excellent research environment while part of this research was done.

References

[1] F. R. Krammer F., Smith G.J.D, Influenza, Nature Reviews Disease Primers 4 (1) (2018) 3.

[2] K. V. Nuwarda R.F., Alharbi A.A., An overview of influenza viruses and vaccines, Vaccines (Basel) 9 (9) (2021) 1032.

[3] C. W. Potter, A history of influenza, Journal of applied microbiology 91 (4) (2001) 572–579.

[4] G.-K. E. Tyrrell C.S., Allen J.L.Y., Ainfluenza: epidemiology and hospital management, Medicine (Abingdon) 49 (12) (2021) 797–804.

[5] T. V. A. Fodor E., Structure and function of the influenza virus transcription and replication machinery, Cold Spring Harbor Perspectives in Medicine 10 (9) (2020) a038398.

[6] S. Nakada, P. Graves, U. Desselberger, R. Creager, M. Krystal, P. Palese, Influenza c virus rna 7 codes for a nonstructural protein, Journal of virology 56 (1) (1985) 221–226.

[7] K. M. Dadonaite B., Gilbertson B., The structure of the influenza a virus genome, Nature Microbiology 4 (11) (2019) 1781–1789.

[8] B. S. Zhang M., Liu M., Influenza a virus-host specificity: An ongoing cross-talk between viral and host factors, Front Microbiol 5 (12) (2021) 777885.

[9] D. L. Suarez, Influenza a virus, Animal influenza (2016) 1–30.

[10] W. R. Harrington W.N., Kackos C.M., The evolution and future of influenza pandemic preparedness, Experimental And Molecular Medicine 53 (5) (2021) 737–749.

[11] B. D.F., Structural consequences of antigenic variants of human a/h3n2 influenza viruses, Viruses 19 (4) (2023) 1008.

[12] Y. J. Kosik I., Influenza hemagglutinin and neuraminidase: Yin-yang proteins coevolving to thwart immunity, Viruses 11 (4) (2019) 346.

[13] J. C. Pedersen, Hemagglutination-inhibition test for avian influenza virus subtype identification and the detection and quantitation of serum antibodies to the avian influenza virus, Avian influenza virus (2008) 53–66.

[14] A. L. Tarca, V. J. Carey, X.-w. Chen, R. Romero, S. Drăghici, Machine learning and its applications to biology, PLoS computational biology 3 (6) (2007) e116.

[15] A. Chadha, R. Dara, D. L. Pearl, S. Sharif, Z. Poljak, Predictive analysis for pathogenicity classification of h5nx avian influenza strains using machine learning techniques, Preventive Veterinary Medicine 216 (2023) 105924.

[16] E. Świętoń, K. Śmietanka, Phylogenetic and molecular analysis of highly pathogenic avian influenza h5n8 and h5n5 viruses detected in poland in 2016–2017, Transboundary and emerging diseases 65 (6) (2018) 1664–1670.

[17] F. Humayun, F. Khan, N. Fawad, S. Shamas, S. Fazal, A. Khan, A. Ali, A. Farhan, D.-Q. Wei, Computational method for classification of avian influenza a virus using dna sequence information and physicochemical properties, Frontiers in Genetics 12 (2021) 599321.

[18] L. He, S. Sun, Q. Zhang, X. Bao, P. K. Li, Alignment-free sequence comparison for virus genomes based on location correlation coefficient, Infection, Genetics and Evolution 96 (2021) 105106.

[19] L. He, R. Dong, R. L. He, S. S.-T. Yau, A novel alignment-free method for hiv-1 subtype classification, Infection, Genetics and Evolution 77 (2020) 104080.

[20] Y.-Y. Song, L. Ying, Decision tree methods: applications for classification and prediction, Shanghai archives of psychiatry 27 (2) (2015) 130.

[21] M. Belgiu, L. Drăguţ, Random forest in remote sensing: A review of applications and future directions, ISPRS journal of photogrammetry and remote sensing 114 (2016) 24–31.

[22] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, W. Xu, Applications of support vector machine (svm) learning in cancer genomics, Cancer genomics & proteomics 15 (1) (2018) 41–51.

Table 6: Summary of the HA sequence data.

Subtype	Number	Min Length (bp)	Max Length (bp)
H1	33044	1683	1938
H2	428	1689	1805
H3	40563	1671	1867
H4	2020	1695	1789
H5	7210	1683	1870
H6	1823	1695	1780
H7	2773	1659	1763
H8	158	1701	1745
H9	4456	1671	1805
H10	1227	1686	1730
H11	683	1698	1772
H12	312	1695	1737
H13	562	1698	1770
H14	26	1707	1749
H16	299	1698	1776

Table 7: Summary of the NA sequence data.

Subtype	Number	Min Length (bp)	Max Length (bp)
N1	30256	1335	2349
N2	46153	1332	1498
N3	1786	1329	1469
N4	335	1413	1463
N5	575	1308	1472
N6	3452	1335	1580
N7	959	1356	1466
N8	3427	1356	1701
N9	1694	1389	1471

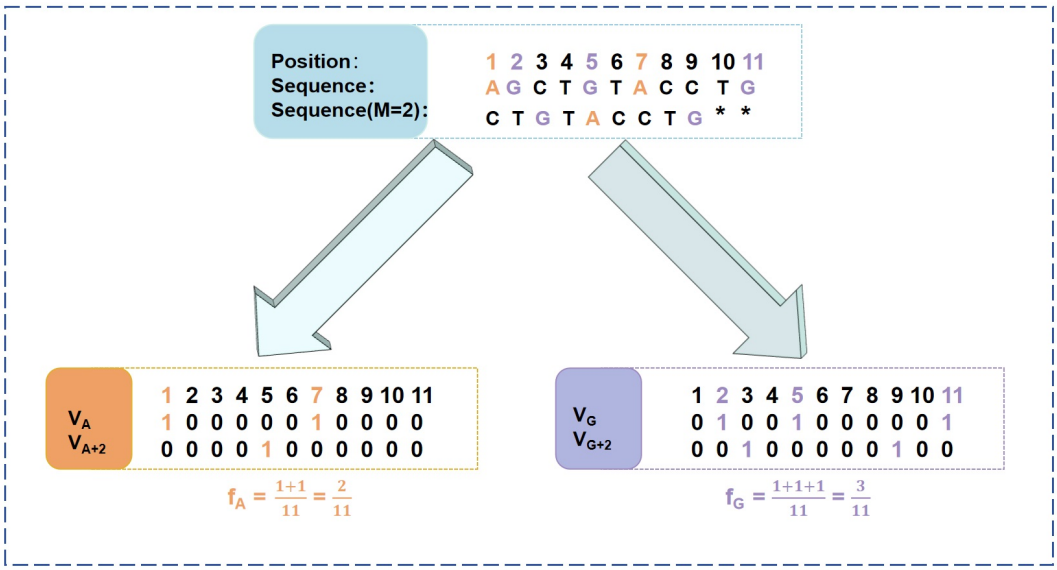

Fig. 6: The computational methodology for determining V_G, V_{G+2}, f_A and f_T.

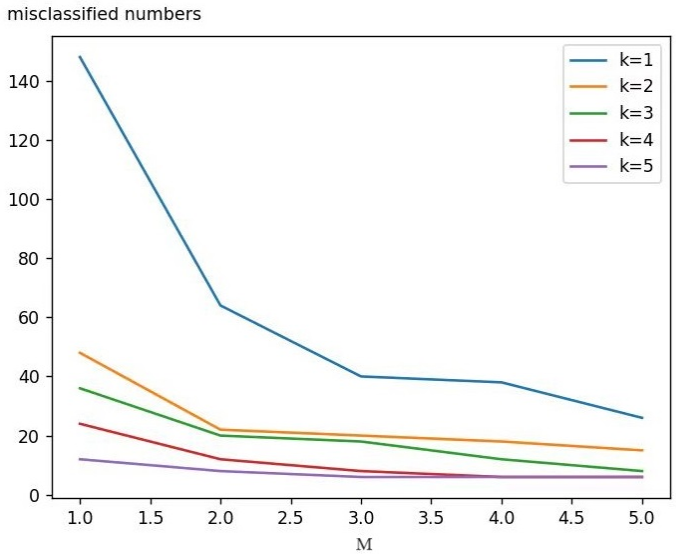

Fig. 7: The number of misclassified samples under different combinations of k and M.

Table 8: The classification performance of different classifiers on HA subtypes.

	Accuracy	Precision	Recall	F1 score
RF	0.99979	0.99996	0.99997	0.99996
DT	0.99745	0.99756	0.99745	0.99749
SVM	0.99965	0.99994	0.99965	0.99972

Fig. 8: Phylogenetic tree of 160 HA genes, with 16 subtypes by SCCFV (k=5, M=2).

Table 9: The classification performance of different classifiers on NA subtypes.

	Accuracy	Precision	Recall	F1 score
RF	0.99979	0.99996	0.99997	0.99996
DT	0.99777	0.99778	0.99777	0.99777
SVM	0.99932	0.99954	0.99932	0.99939