Classification of Host Origin in Influenza A virus by Transferring Protein Sequences into Numerical Feature Vectors

FAYROZ F.SHERIF*[1], NOURHAN ZAYED[1], MAHMOUD FAKHR[1]

***Abstract*—** Global outbreaks of human influenza occur from influenza A viruses with novel Hemagglutinin (HA) molecules to which humans have no immunity. So accurate detection of influenza viral origin is of particular importance to improve influenza surveillance and vaccine development. Here, a total of 1500 and 2349 protein sequences for Hemagglutinin (HA) and Neuraminidase (NA) respectively were selected to be involved in our study. We used two techniques to transfer the protein sequences into feature vectors firstly, the feature vector constructed from the composition of amino acids (AAC) and secondly the feature vector constructed from the Composition, Transition, Distribution (CTD). Both used separately for the training of machine learning algorithms. Host of origin classification models constructed using KNN and random forest based on AAC and CTD feature vectors. The results guarantee that the classification performance using AAC feature vector achieves slightly better performance than using CDT feature vector. Furthermore host classification using HA protein segment achieved higher accuracy results than NA. The highest host classification model was HA-human using random forest with accuracy 96.6% and 95.3% for AAC and CDT respectively.

Keywords— Influenza A virus, machine learning, host classification, KNN, random forest, composition of amino acids, Composition, Transition, Distribution (CTD).

I. Introduction

THIS Influenza A viruses belong to the Orthomyxoviridae family of negative sense, single-stranded, segmented RNA viruses. The RNA core consists of 8 gene segments. Immunologically, the most significant surface proteins include Hemagglutinin HA (16 subtypes) and Neuraminidase NA (9 subtypes). Influenza A subtypes are usually identified by their HA and NA proteins [1]. The HA and NA proteins are integral membrane proteins and consider as the major surface antigen of the influenza virus virion. The Hemagglutinin (HA) of influenza A viruses is a major surface glycoprotein that is responsible for attachment of the virus to the cell surface of host receptors. The role of NA is to free virus particles from host cell receptors, to allow progeny virions to escape from the cell in which they arose, and so facilitate virus spread [2, 3]. All known subtypes of influenza A viruses are found among avian species that serve as main reservoirs for these agents [4]. In general, an influenza virus infects only a single species; however, whole viruses may occasionally be transmitted from one species to another, and genetic reassortment between viruses from two different hosts can produce a new virus capable of infecting a third host. Avian influenza viruses are not readily introduced into humans [5], possibly because humans do not possess the a(2,3)-sialyllactose (NeuAc-2,3Gal) receptors required for attachment of the viruses to epithelial cells. However, individual viral genes can be transmitted between humans and avian species, as demonstrated by avian human reassortant viruses that caused the 1957 and 1968 influenza pandemics [6]. This finding suggested that an middle host may be needed for genetic reassortment of human and avian viruses. Pigs are considered a logical candidate for this role because they can be infected by either avian or human viruses and because they possess both NeuAc-2,3Gal and NeuAc-2,6Gal receptors. In addition, there is good evidence that pigs are more frequently involved in interspecies transmission of influenza A viruses than are other animals [7, 8].

Previous studies have also defined host specificity markers. For example, [9] predicted positions in the genome associated with human host specificity. However, the host markers that these workers identified in the surface glycoproteins HA and NA and in the polymerase protein PB1, as well as the alternate transcripts NS2, M2, and PB1-F2, were poor-quality host discriminators. In a previous study, Host-specific signatures were identified using class associative rule mining to identify and confirm significant variations between different influenza hosts with lower accuracy[10]. Another study [11] used random forest for the prediction of host tropism from both avian and human samples only. Another previous computational prediction model in [12] could successfully classify avian and human strains only using support vector machine (SVM). They had another drawback that, is the use of only inner proteins of influenza. This method ignores the importance of HA and NA in determining host tropism. Because of the important functional role of HA and NA in cell-receptor attachment, entry, and infectivity, our focus in this study was specifically on the host markers that were found only in HA and NA.

[1]*Computers and Systems Department, Electronics Research Institute, Giza 12622, Egypt*
Correspondence:
Fayroz F. Sherif
Computers and Systems Department
Electronics Research Institute (ERI)
Giza 12622, Egypt
E-mail: Fayroz_Farouk@eri.sci.eg

The aim of the present study was to establish accurate host of origin classifiers that are capable of indicating signatures in human, avian, and swine influenza viral genomes for HA and NA proteins, using two different feature vectors; Amino Acid Composition (AAC) and Composition/Transition / Distribution (CTD). KNN and Random forest classification models were used to classify viral sequences by host species.
The paper is organized as follows. Section 2 describes the dataset used in this study, how to transfer the protein sequences into numerical feature vectors, describe in details the AAC and CTD methods and introduces KNN and random forest learning algorithms. Section 3 combines the results of HA and NA host models and compares them. Finally, Section 4 presents our conclusions.

II. Materials and Methods

All protein sequences that isolated from human, avian and swine hosts were downloaded from the NCBI's Influenza Virus Resources (http://www.ncbi.nlm.nih.gov/genomes/ FLU /FLU.html). The downloaded sequences were forced to be non-redundant and complete isolation of HA and NA segments. A total of 1500 and 2345 HA and NA protein sequences respectively were selected to be involved in our study. 70% of the data is used for training and the remaining part is used for testing. We used amino acid sequences (20 letter alphabet) because they are known to give more reliable results than nucleotide sequences when the sequence divergence is high.

To compare the genomic patterns of avian, swine and human influenza viruses with each other, we downloaded protein sequences of HA and NA from NCBI's Influenza Virus Resources, isolated from various host species. The detailed count of sequences used in this study for each host is indicated in table 1. The sequences were grouped according to host type, and cover all the viral subtypes found in that host. Downloaded FASTA format sequences were parsed into each category such as accession number, subtype, gene, host, occurring year, and other parameters.

A. Transforming protein sequence into feature vectors

A protein or peptide sequence with N amino acid residues could be generally represented as (R_1, R_2, . . . , R_n), where R_i represents the residue at the i-th position in the sequence. The labels i and j are used to index amino acid position in a sequence, and r, s, t are used to represent the amino acid type. Amino acids composition and amino acid physicochemical properties (Composition / Transition / Distribution) were extracted from protein sequences as numerical feature vectors to train the machine learning algorithms [13]. The following subsections present the details of the two methods.

Table 1 The count of sequences used for each host of the HA and NA

Protein Segment	Human	Avian	Swine	Total
HA	500	500	500	1500
NA	1213	757	379	2349

1. Amino Acid Composition (AAC)

Composition of amino acids were extracted from protein sequences as feature vectors for the training of machine learning algorithms. These feature vectors represent the composition of each individual amino acid in the protein sequence. The Amino Acid Composition (AAC) is the fraction of each amino acid type within a protein. Composition of each of the 20 standard amino acids was first computed, yielding 20 feature vectors. This was performed by calculating the frequency of each amino acid along the length of the entire protein sequence.
The fractions of all 20 natural amino acids are calculated as:

$$f(r) = \frac{N_r}{N} \qquad r = 1, 2, \ldots, 20.$$

where N_r is the number of the amino acid type r and N is the length of the sequence.

2. Composition/ Transition/ Distribution (CTD)

We generate another feature vectors based on the overall Composition, Transition and Distribution (CTD) of amino acid into three groups, for each attribute of these seven attributes: hydrophobicity, polarizability, normalized van der Waals volume, secondary structure, polarizability, charge, and solvent accessibility of the protein sequences. The amino acids are divided in three classes according to its attribute and each amino acid is encoded by one of the indices 1, 2, 3 according to which class it belonged. The corresponding division is shown in the table 2. The detailed computational procedures are illustrated as follows.

Composition (C)
Composition is the total percent for each encoded class in the sequence. It can be defined as the number of amino acids of a specific property divided by the whole number of amino acids. Number of vectors = 21 (3 groups * 7 attributes)

$$C_r = \frac{n_r}{n} \quad r = 1, 2, 3$$

Transition (T)
Transition descriptor characterizes the percent frequency with which amino acids of a specific property is followed by amino acids of a different property. Transition descriptor can be calculated as

$$T_{rs} = \frac{n_{rs} + n_{sr}}{N-1}$$

$rs = '12', '13', '23'$

Table 2 The amino acid attributes and division of the amino acids to groups

	Group 1	Group 2	Group 3

Attribute			
Hydrophobicity	Polar {Q, E, R, K, D, N}	Neutral {G, P, H, A, S, T,Y}	Hydrophobic {C, V, F, L, I, M,W}
Polarizability	(0-1.08) {S, D, G, A, T}	(0.128-0.186) {C, Q, I, P, N, V, E L}	(0.219-0.409) {Y, M, K, R,H, F,W}
Normalized van der Waals volume	(0-2.78) {S, C, G, A,T, P, D}	(2.95-4.0) {E, Q, N, V, I, L}	(2.95-4.0) {K, F, M, H, R, Y, W}
Polarity	(4.9-6.2) {W, C, L, I, F,M, V, Y}	(8.0-9.2) {T, G, P, A, S}	(10.4-13.0) {K, N, H, Q, R,E, D}
Solvent accessibility	Buried {A, L, F, C, G, I, V, W }	Exposed {R, K, Q, E, N, D }	Intermediate {M, S, P, T, H, Y }
Secondary structure	Helix {E, A, L, M, Q, K, R, H}	Strand {V, I, Y, C, W, F, T }	Coil {G, N, P, S, D }
Charge	Positive {K, R}	Neutral {A, N, C, Q, G, H, I, L, M, F, P, S, T, W, Y, V}	Negative {D, E}

where n_{rs} and n_{sr} *are* the numbers of dipeptide encoded as 'rs' and 'sr' respectively in the sequence. *N* is the length of the sequence. Number of vectors = 21 (3 groups * 7 attributes)

Distribution (D)

The distribution descriptor describes the distribution of each attribute in the sequence. It measures the chain length within which the first, 25, 50, 75 and 100% of the amino acids of a specific property is located respectively. Number of vectors = 105 (3 groups * 7 attributes * 5 distribution positions). The complete parameter vector for these three descriptors contains 21(C)+21(T)+105(D)=147 scalar components. This means that each protein sequence was represented by 147 biochemical and physicochemical features.

B. Training Machine Learning classifiers

There is imperfect classifications and there is no perfect classifier for all dataset. It is helpful to compare the performance of various classifiers to determine which one works better on a given data. The performance of the classifier is often evaluated using test set to predict the class labels of unknown samples. This section reviews some of the classifiers commonly used for building host prediction models. Two popular classification techniques including K-nearest neighbors (KNN) and random forest (RF) were applied to identify the model that best fit the dataset and correctly predict the host of test set.

1. K-nearest neighbor (KNN)

KNN is a simple method for classifying objects based on closest training points in the feature space. KNN assumes that objects, which are close together, are probable to have the same classification. The chance that a point x belongs to a class can be estimated by the majority voting for the training data sets. in a specified neighborhood of x that belong to that class. The Euclidean distance that calculate the distances from x to all points in the training set is the most common distance metric used in K-nearest neighbor [14].

2. Random Forest (RF)

RF is a classification method based on a collection of decision trees CART classifiers. RF uses bootstrap samples from the dataset to build a set of trees. To classify a new sample, a majority vote method is utilized to make a decision about class label. RF has better performance over the single (CART) [15, 16].

C. Model Evaluations

Performance of prediction models were evaluated from a number of measures including prediction accuracy, sensitivity, specificity. Prediction accuracy measures of the overall accuracy of the classifier by calculating the number of correctly classified human, avian or swine samples over the total number of samples in the dataset. Sensitivity and specificity summarize the accuracies of positive and negative predictions respectively where sensitivity calculates the ratio of samples correctly predicted among all positive samples in the dataset and specificity describes the ratio of samples correctly predicted among all negative samples in the dataset.

III. Results and Discussion

Our results confirm that the applied machine learning algorithms; KNN and random forest can successfully be used for classifying all Influenza A strains through identifying Hemaglutinin (HA) and Neuraminidase (NA) segments with high accuracy.

Host classification of any viral sequence as human, avian or swine, varied according to HA and NA segments. In general host classification using HA achieved higher accuracy than NA. Figures 1, and 2 compare the performance of the two classifiers; KNN and random forest in terms of accuracy, sensitivity and specificity in HA and NA models respectively using amino acid composition (AAC). However Figures 3, and 4 compare the performance of the same classifiers using Composition ,Transition, Distribution (CTD) in HA and NA models respectively.

The classification models constructed from Amino Acid Composition (AAC) feature vectors and Composition, Transition and Distribution (CTD), all achieved high prediction performance that indicate clear difference in human, avian and swine proteins.

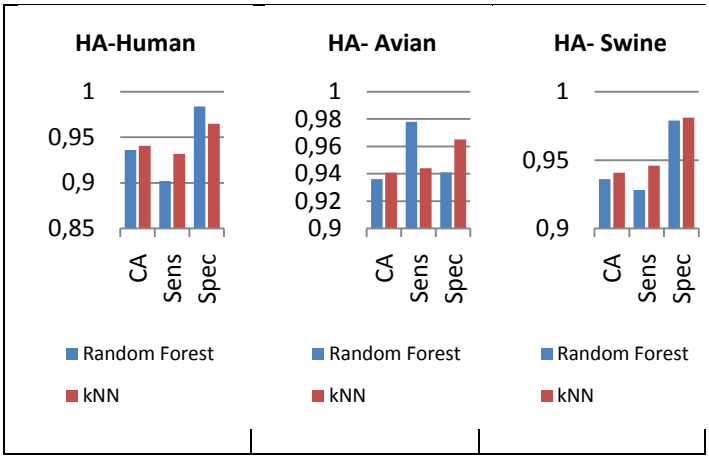

Figure 1. HA classification results using AAC

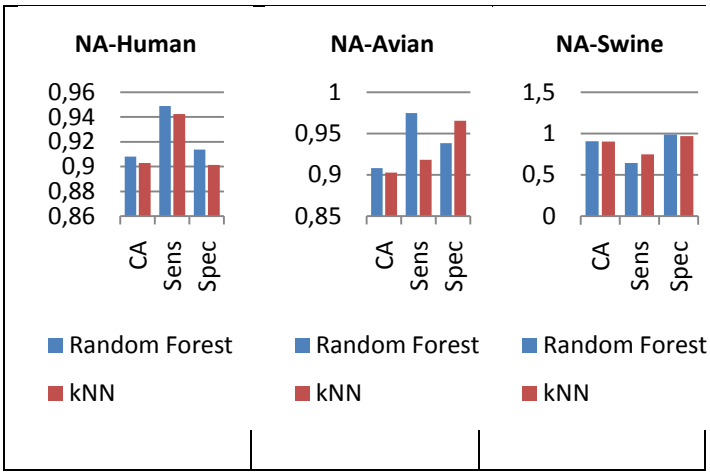

Figure 2. NA classification results using AAC

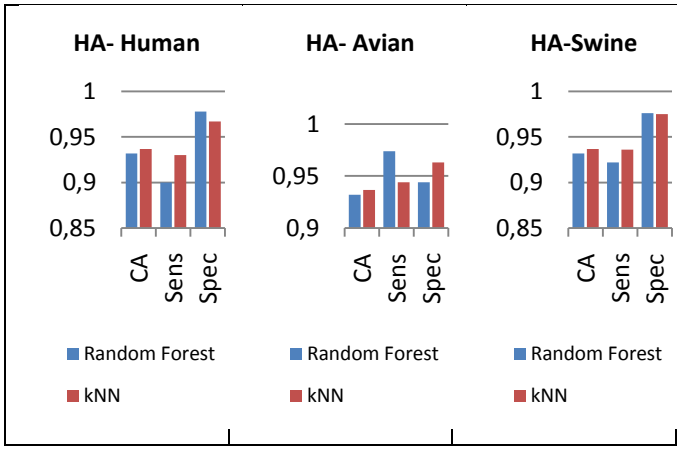

Figure 3. HA classification results using CDT

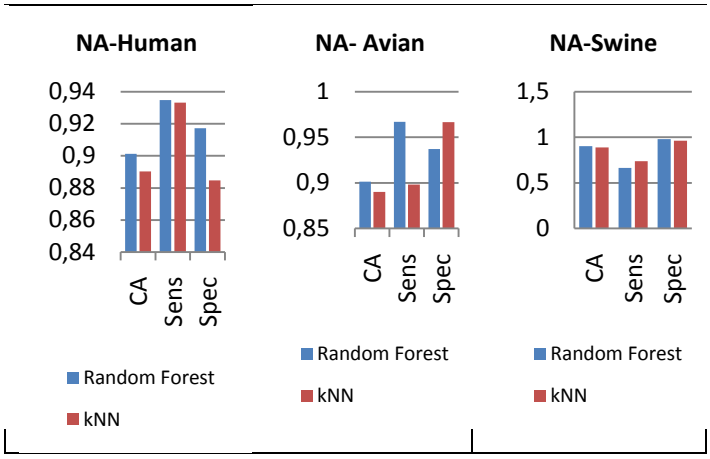

Figure 4. NA classification results using CDT

In general, the classifications results using AAC exceed the CDT method with small differences. By comparing the results of AAC method we found that, HA-human model achieved the highest accuracy in host classification (96.6%) over HA-avian and HA-swine (89.2% and 95.7%) respectively using random forest method. And NA-avian model achieved the highest accuracy (92.7%) in host classification over NA-human and NA-swine (91.1% and 92.3%) respectively using KNN method as shown in table 3.

For CTD method we found that, HA-human model also achieved a higher accuracy in host classification (95.3%) over HA-avian and HA-swine (89.7% and 95.1%) resistively using random forest method. And NA-avian model also achieved a higher accuracy (92.8%) in host classification over NA-human and NA-swine (89.6% and 79%) respectively using KNN method as shown in table 4. These results seem sensible so that cross species infections are usually taken place in these segments of different hosts. This study revealed that AAC achieved higher performance than CTD although it had little feature vectors (only 20) compared with CTD (147 features). So it is recommended for computational time efficiency and simplicity. The power of this influenza host prediction method lies not in its almost high prediction accuracy, but rather when it makes a mistake in classifying human, avian or swine proteins. The rate of misclassified hosts that range from (0.9 % to 6.4%) in HA and (1% to 11.9 %) in NA.

Our results achieved a higher accuracy than the accuracy reported in a related work used the protein sequences and nucleotide sequences in their research without transformation. The research in [7] yielded accuracies ranging from 50 % to 95% for host classification, depending on HA subtype. Decision tree (DT) in [9] gave higher host classification accuracies, ranging from 91.2 % to 94.6 %, as opposed to HMMs. However this was done for some HA subtypes only as (H1, H2, H3, H5 and H9).

Table 3 Comparative performance of HA and NA using AAC

Hemaglutinin (HA)			
	Human	**Avian**	**Swine**
KNN	93.0 %	93.1 %	96.1%
Random forest	96.6 %	89.2 %	95.7 %
Neuraminidase (NA)			
	Human	**Avian**	**Swine**
KNN	91.1 %	92.7 %	82.3 %
Random forest	92.2 %	88.3 %	92.4 %

Table 4 Comparative performance of HA and NA using CTD.

Hemaglutinin (HA)			
	Human	**Avian**	**Swine**
KNN	93.4 %	92.7 %	95.1 %
Random forest	95.3 %	89.7 %	95.1 %
Neuraminidase (NA)			
	Human	**Avian**	**Swine**
KNN	89.6%	92.8%	79%
Random forest	92.3%	88%	86.9%

IV. Conclusion

Accurate detection of influenza viral origin can significantly improve influenza surveillance and vaccine development. This study provides prediction models for HA and NA influenza proteins determining host of origin classification for virus strains. Here we introduced many varieties in our study, first through using Hemaglutinin (HA) and Neuraminidase (NA) segments, second by using two feature extraction methods; Amino Acid Composition (AAC) and Composition/Transition / Distribution (CTD) and finally by utilizing two different machine learning techniques; random forest and KNN. These varieties confirm that we can successfully classify all Influenza host strains as human, avian or swine using AAC or CTD with high accuracy. In general host classification using HA achieved higher accuracy than NA. The highest host-origin model was HA-human model using random forest with accuracy 96.6%. Interestingly, This study revealed that the simple AAC achieved higher performance than CTD with HA and NA proteins despite its little parameters, consequently it is preferable to use it in modeling the evolution of the influenza A through different hosts and in understanding its specificity. We hope that our work will facilitate reliable detection of influenza viral origin to improve influenza surveillance and vaccine development.

References

[1] T. Horimoto and Y. Kawaoka, "Influenza: lessons from past pandemics, warnings from current incidents," *Nat Rev Microbiol*, vol. 3, pp. 591-600, Aug 2005.

[2] Y. Zhang, X. Lin, F. Zhang, J. Wu, W. Tan, S. Bi, J. Zhou, Y. Shu, and Y. Wang, "Hemagglutinin and neuraminidase matching patterns of two influenza A virus strains related to the 1918 and 2009 global pandemics," *Biochem Biophys Res Commun*, vol. 387, pp. 405-8, Sep 18 2009.

[3] C. L. Eng, J. C. Tong, and T. W. Tan, "Distinct Host Tropism Protein Signatures to Identify Possible Zoonotic Influenza A Viruses," *PLoS One*, vol. 11, p. e0150173, 2016.

[4] S. S. Wong and K. Y. Yuen, "Avian influenza virus infections in humans," *Chest*, vol. 129, pp. 156-68, Jan 2006.

[5] J. J. Skehel and D. C. Wiley, "Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin," *Annu Rev Biochem*, vol. 69, pp. 531-69, 2000.

[6] G. W. Chen and S. R. Shih, "Genomic signatures of influenza A pandemic (H1N1) 2009 virus," *Emerg Infect Dis*, vol. 15, pp. 1897-1903, Dec 2009.

[7] V. Shinde, C. B. Bridges, T. M. Uyeki, B. Shu, A. Balish, X. Xu, S. Lindstrom, L. V. Gubareva, V. Deyde, R. J. Garten, M. Harris, S. Gerber, S. Vagasky, F. Smith, N. Pascoe, K. Martin, D. Dufficy, K. Ritger, C. Conover, P. Quinlisk, A. Klimov, J. S. Bresee, and L. Finelli, "Triple-reassortant swine influenza A (H1) in humans in the United States, 2005-2009," *N Engl J Med*, vol. 360, pp. 2616-25, Jun 18 2009.

[8] F. S. Fayroz, Y. M. Kadah, and M. El-Hefnawi, "Genomic signatures and associative classification of the Hemagglutinin protein for Human versus Avian versus Swine Influenza A viruses," in *Proc. 28th National Radio Science Conference*, Cairo, Egypt, 2011.

[9] J. E. Allen, S. N. Gardner, E. A. Vitalis, and T. R. Slezak, "Conserved amino acid markers from past influenza pandemic strains," *BMC Microbiol*, vol. 9, p. 77, 2009.

[10] M. ElHefnawi and F. F. Sherif, "Accurate classification and hemagglutinin amino acid signatures for influenza A virus host-origin association and subtyping," *Virology*, vol. 449, pp. 328-38, Jan 20 2014.

[11] C. L. Eng, J. C. Tong, and T. W. Tan, "Predicting host tropism of influenza A virus proteins using random forest," *BMC Med Genomics*, vol. 7 Suppl 3, p. S1, 2014.

[12] J. Wang, C. Ma, Z. Kou, Y. H. Zhou, and H. L. Liu, "Predicting transmission of avian influenza A viruses from avian to human by using informative physicochemical properties," *Int J Data Min Bioinform*, vol. 7, pp. 166-79, 2013.

[13] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu, "protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, pp. 1857-9, Jun 01 2015.

[14] D. T. Larose and C. D. Larose, "k-Nearest Neighbor Algorithm," in *Discovering Knowledge in Data*: John Wiley & Sons, Inc., 2014 pp. 149-164.

[15] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5 - 32, 2001.

[16] B. rard, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063-1095, 2012.