Phylogenetic Analysis of HA Protein of Influenza A Virus Based on a Novel Alignment-Free Method

Ping-an He\textsuperscript{1,*}, Daoli Yu\textsuperscript{1}, Tingting Ma\textsuperscript{1}, Zhixin Tie\textsuperscript{2}

\textsuperscript{1}School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, P. R. China
\textsuperscript{2}School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, P. R. China

(Received February 21, 2015)

Abstract

When a new subtype of HA gene is transmitted from aquatic birds to human, the influenza pandemics occur sometimes. Research of the origin and evolution of HA gene is an important task for biologists. In this study, based on an efficient mathematical method, a novel 3D graphical representation of protein sequence has been proposed to analyze HA protein of 351 strains of different viral subtype of influenza A virus. Then a mathematical descriptor was suggested to compute the distance between two different proteins which is very convenient for computation. And phylogenic tree without requiring multiple alignment was constructed based on the single linkage method. The evolutionary tree obtained is consistent with the previous studies. The results show that species virus host and geographical distribution possess vital effect on evolution routes of HA protein of influenza A virus.

Introduction

Influenza is an acute respiratory infections caused by influenza viruses. Influenza viruses, which belong to the Orthomyxoviridae family, are classified as A, B, and C based on antigenic differences in their nucleoprotein (NP) and matrix (M1) protein \cite{1,2}. All avian influenza viruses are classified as type A. Influenza A virus is a negative sense single stranded RNA virus with eight genomic segments encoding hemagglutinin (HA), matrix protein (MP), neuraminidase (NA), nucleoprotein (NP), non-structural protein (NS) and RNA polymerases

* Correspondence to: P. A. He; E-mail address: pinganhe@zstu.edu.cn
Based on two surface antigens of the viral envelope HA, of which there are 18 subtypes (H1–H18), and NA, of which there are 11 subtypes (N1–N11), influenza A viruses are categorized into 198 subtypes. All of which are maintained in wild avian populations, except for the new H17N10 subtype which has been found in bats [4,5]. It is believed that all subtypes of HA are perpetuated in the aquatic bird population and reassorted with each other with high frequency [6-8]. New influenza viruses and genotypes continuously emerge due to the frequent evolutionary events including genetic reassortment and mutation [9-10]. Influenza pandemics, defined as global outbreaks of the disease due to these emerging influenza viruses, have exacted a high death toll from human populations [1] and caused several pandemics, including H1N1 in 1918, H2N2 in 1957, and H3N2 in 1968 [11-15]. Furthermore, each year seasonal influenza results in a considerable death toll worldwide [16].

Influenza genotype analysis reflects influenza viral evolutionary footprints, and thus is critical for preparing a strategy to prevent and control influenza epidemics and pandemics [10]. When a new subtype of HA gene is transmitted from aquatic birds to human, the influenza pandemics occur sometimes. Research of the origin and evolution of HA gene is an important task for biologists [2]. Based on the graphical representation of protein, the evolutionary relationships of influenza A virus HA proteins are researched in the paper.

Recently, the graphical representation of protein sequences has been widely used in the analysis of the evolution of many species, such as ND5 proteins of nine species [17-20], ND6 proteins of eight species [21,22,25], H5N1 and H7N9 influenza virus [9,23-24]. Graphical representation of protein sequence provides a simple way of viewing, sorting and comparing various protein structures [25]. For example, Liao et al. proposed a 2D graphical representation of DNA sequences to construct phylogenetic tree of H5N1 avian influenza virus [23]. Nandy et al., have identified some distinct structural motifs of the neuraminidase RNA sequences of H5N1 using a 2D graphical representation [13]. Liu and Zhang analyzed the phylogenetic relationship among H5N1 avian influenza virus using the graphical representation of protein sequences [9]. Applying a graphical representation of protein sequences, Bai et al. infer the evolution route of H7N9 avian influenza virus [24].

In this study, a novel graphical representation of protein is proposed to analyze HA protein of influenza A virus. A mathematical descriptor is extracted from the graphical representation to
compare the similarities and dissimilarities of HA protein of 351 strains of different viral subtypes. And phylogenetic tree without requiring multiple alignment is constructed based on the single linkage method. The results show that species virus host and geographical distribution possess vital effect on evolution routes of HA protein of influenza A virus.

**Materials**

For influenza A virus, there are eight segments, each of which codes for one or more proteins. The HA protein is an integral membrane protein and the major surface antigen of the influenza virus virion [2]. Owing to error-prone viral RNA polymerase activity, influenza virus HA is subject to a very high rate of mutation [8]. The HA protein is considered for deeper analysis due to its key role in the molecular evolution of avian influenza virus.

The influenza datasets are downloaded from influenza virus resource database at NCBI. To avoid sampling bias, 318 strains of different viral subtypes of HA proteins isolated from avian species for the period 1997 to the present are selected randomly, whose lengths are all from 555 to 565. In addition, 21 strains of H5N1 virus, 9 strains of H7N9 virus and 3 strains of H9N2 virus are selected from human. Thus, the information of 351 influenza virus HA proteins is listed in Table 1.

**3D graphical representation of protein sequences**

Proteins are linear polymers composed of twenty different amino acids, linked by covalent bonds. Various physicochemical properties of amino acids, such as the relative molecular mass, solubility limit, specific rotation, isoelectric point, hydrophobicity, melting point, and $pK_a$ values for terminal amino acid groups of $COOH$ and $NH_3^+$, can be used to study protein

### Table 1   Information of 351 HA proteins of influenza A virus

<table>
<thead>
<tr>
<th>subtype</th>
<th>Category</th>
<th>Serial number, ID(Virus Name)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-1(2)</td>
<td></td>
<td>2, AC88463 (A/redheaded duck/Minnesota/Sg-00123/2007(H1N1)); 3, ACT84668 (A/mallard/Minnesota/Sg-00627/2008(H1N1))</td>
</tr>
<tr>
<td>H1-2(2)</td>
<td></td>
<td>5, AF77481 (A/mallard/Republic of Georgia/4/2010(H1N1)); 12, ACT84002 (A/mallard/Minnesota/Sg-00108/2007(H1N3))</td>
</tr>
<tr>
<td>H1-3(7)</td>
<td></td>
<td>4, ACT84661 (A/mallard/Minnesota/Sg-00579/2008(H1N1))</td>
</tr>
</tbody>
</table>
sequence structure and function [18]. Three parameters of these physicochemical properties, $p_k$ values for terminal amino acid groups of $NH_3^+$ ($p_k$), isoelectric point ($pI$) at 25°C and hydrophobicity ($h$) are adopted to construct the 3D vectors for 20 amino acids, whose coordinates are shown in Table 2.

<table>
<thead>
<tr>
<th>Table2 Amino acid side chain properties and their coordinates in 3D space.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Alanine</td>
</tr>
<tr>
<td>Phenylalanine</td>
</tr>
<tr>
<td>Isoleucine</td>
</tr>
<tr>
<td>Leucine</td>
</tr>
<tr>
<td>Methionine</td>
</tr>
<tr>
<td>Amino Acid</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Proline</td>
</tr>
<tr>
<td>Valine</td>
</tr>
<tr>
<td>Tryptophan</td>
</tr>
<tr>
<td>Cysteine</td>
</tr>
<tr>
<td>Glycine</td>
</tr>
<tr>
<td>Aspargine</td>
</tr>
<tr>
<td>Glutamine</td>
</tr>
<tr>
<td>Serine</td>
</tr>
<tr>
<td>Threonine</td>
</tr>
<tr>
<td>Tyrosine</td>
</tr>
<tr>
<td>Histidine</td>
</tr>
<tr>
<td>Lysine</td>
</tr>
<tr>
<td>Arginine</td>
</tr>
<tr>
<td>Glutamic acid</td>
</tr>
<tr>
<td>Aspartic acid</td>
</tr>
</tbody>
</table>

As shown in Table 2, all the values of the \( p_{k_1} \) and \( p_{I} \) parameters are more than zero. In order to save the space, the data can be standardized by the Eq.(1),

\[
\begin{align*}
    x_i &= \frac{p_{k_1} - \min(p_{k_1})}{\max(p_{k_1}) - \min(p_{k_1})} \\
    y_i &= \frac{p_{I} - \min(p_{I})}{\max(p_{I}) - \min(p_{I})} \\
    z_i &= h_i - \bar{h}
\end{align*}
\]

where \( \bar{h} = \frac{1}{20} \sum_{i=1}^{20} h_i \), \( p_{k_1} \), \( p_{I} \) and \( h_i \) are the values of \( p_{k_1} \), \( p_{I} \) and hydrophobicity for each amino acid, \( \max \) and \( \min \) are the maximum value and minimum value of the corresponding parameters among 20 amino acids. Based on the above transformation, the new coordinates of all amino acids are listed at the last three columns in Table 2. Each amino acid is denoted numerically by a special vector in 3D space.

Then a protein sequence can be transformed into a set of points in 3D space. For a protein sequence with \( N \) amino acids \( \text{SS}_1 \text{SS}_2 .. \text{SS}_N \), scan it by searching one amino acid step by step at a time. For the step \( i \) (\( i=1,2, ..., N \)), the point \( P_i(x_i,y_i,z_i) \) can be constructed based on the Eq.(2) as follows:
where \( s_j^1 (j = 1, 2, 3) \) represents the \( j \)-th component of the vector corresponding to \( S \), and we set \( P_0 = (0, 0, 0) \). Thus, \( N \) points \( P_1, P_2, \cdots, P_N \) can be obtained when \( i \) count from 1 to \( N \). Connecting the adjacent points, a curve can be plotted in 3D space for protein sequence \( S \).

Because the first two components of vector corresponding to each amino acid are nonnegative, the 3D graphical representation of a protein based on the Eq. (2) should be acyclic and non-degenerate. Therefore, there is a one-to-one correspondence between protein sequences and 3D curves.

The numerical characterization of protein sequences

The similarities/dissimilarities of different protein sequences can be compared quantitatively from their 3D graphical representations with mathematical descriptor.

Firstly, the quotient of the Euclidean distances and cosine values between two continuous vertices in their 3D graphical representations is proposed as the mathematical descriptor to compare the similarities/dissimilarities of different sequences, denoted \( d \). For example, given two sequences \( S \) and \( S' \), \( d \) between two vectors \( \overrightarrow{E}_i = (x_i, y_i, z_i) \) and \( \overrightarrow{E'}_i = (x'_i, y'_i, z'_i) \) is defined as follows:

\[
d(i) = \frac{d(\overrightarrow{E}_i, \overrightarrow{E'}_i)}{\cos(\overrightarrow{E}_i, \overrightarrow{E'}_i)}, \quad \text{for } i = 1, 2, \cdots, n \quad (3),
\]

where \( d(\overrightarrow{E}_i, \overrightarrow{E'}_i) \) represents the Euclidean distance between the vector \( \overrightarrow{E}_i \) and \( \overrightarrow{E'}_i \). Secondly, the cumulative distances from the first amino acid to all the following amino acids are calculated. When we compute the cumulative distances between two proteins with different lengths \( m > n \), we consider \( \max(d(i)) \) as distances between last \( m - n \) points, in which \( d(i)(i = 1, 2, \ldots, n) \) denotes the distance between the \( i^{th} \) points of graphical representation of the two proteins. In other words, if the two protein sequences have the same amino acids, the value of \( d_i \) is 0 between last \( m - n \) points, otherwise the value of \( d_i \) is larger than 0. Thus, the cumulative distance between two sequences is defined by Eq. (4) as follows:
Thus, the similarities/dissimilarities between two protein sequences can be compared based on the cumulative distances. In general, the smaller the corresponding cumulative distance of two proteins is, the more similar they are.

\[ D = \sum_{i=1}^{n} d(i) + (m - n) \times \max\{ d(i) \} \ (4). \]

Similarities/Dissimilarities of 9 ND6 proteins

As discussed above, the similarity of sequences can be compared based on their distance. To illustrate our method, we consider the numerical characterization of mutations and analyze the similarities among sequences belonging to eight ND6 proteins: human (P03923), gorilla (Q34573), common chimpanzee (Q9T9V6), wallaroo (P92670), harbor seal (Q00543), gray seal (P38603), rat (P03926), and mouse (P03925), whose sequence data were all downloaded from UniProtKB. The distances among proteins were calculated using the equation (4). If the total number of all proteins is \( N \), a real symmetric \( N \times N \) matrix \( D \) is constructed, whose element \( D_{ij} \) is used to reveal the evolutionary distance between the proteins \( i \) and \( j \). Using the single linkage method, the phylogenetic tree is obtained based on the distance matrix \( D \), shown in Fig. 1.

Observing Fig. 1, the ND6 proteins are more similar for group \{human, gorilla, common
chimpanzee}, group {harbor seal, gray seal} and the group {rat, mouse} respectively. On the other hand, ND6 protein of wallaroo is very dissimilar to others among the eight species, which is consistent with the actual evolutionary evidences. The same results have been obtained with ClustalW methods and some other methods recently [21, 22 and 25]. The result shows that our approach is effective to construct the phylogenetic tree of protein.

The phylogenetic tree of HA proteins of influenza A virus

In this section, the 351HA protein of 14 subtypes of influenza A virus in Table 1 are compared based on above method. Using the single linkage method, the phylogenetic tree is obtained based on the cumulative distances, shown in Fig. 2. In Fig. 2, the name of each clade is denoted by subtype-the i-th branch of this subtype (the number of virus strains in this clade). The ID number and names of virus strains are listed Table 1. For example, H3-1(9) represents the first cluster of H3 subtype virus including 9 strains.

All the viruses are sampled from different virus hosts, different time periods and different areas. Analyzing the phylogenetic relationship of HA protein of influence viruses in Fig.2, some evolution information among those influence viruses can be obtained.

Observing Fig.2, we believe that it is not accidental that most of HA proteins of same subtype virus is located at the same clade. Furthermore, some different virus subtype is clustered into a clade. For example, H2 subtype and H10 subtype form a cluster. It implies that the HA protein of H2 and H10 are very closer to each other. Similarly, the three subtypes H13, H16, H12, and two subtypes H5, H11 are in the same group respectively.

Additionally, some subtypes are divided into many small clades, such as H1, H3 and H6 subtypes, leading us to analyze their evolution path hardly. For instance, all H3 viruses are divided into the eight clades: four clusters, H3-1(9), H3-2(5), H3-3(2), H3-4(5), and four H3 viruses ACI89631 (H3N8), AEI29907 (H3N2), AEI29899 (H3N1), and ACT84086 (H3N6). In the H3-3(2), two virus strains are extracted from duck. Again, five strains of H3 subtype included in the H3-4(5) clade are from USA. These categories share the common feature, which has a great relationship with the species virus host and geographical distribution in each clade.
Fig. 2 Phylogenetic tree of 351 strains of different viral subtypes of HA proteins (The name of each clade is denoted with the subtype-the i-th branch of this subtype(the number of virus strains in this clade). The ID number and names of virus strains are listed Table 1. For example, H3-1(9) represents the first branch of H3 virus including 9 strains.)

Observing the clades of H1 subtype in Fig.2, we can find that the distance among H1-1(2), H1-2(2), H1-3(7), ACT84661 (H1N1) is far from each other, although all of them belong to the same subtype H1. This implies the HA proteins of H1 subtype virus come into different degrees of variation. We may infer that with the increasing of the immune pressure for these subtypes, the amino acids of HA protein sequences are more or less changed.

In Fig 2, 43 strains of H7 subtype and 45 strains of H5 subtype, which are taken the two
clades in Fig.3 and Fig.4, cluster together in one clade, respectively.

In Fig 3, 9 HA proteins of H7 subtype, the strains 340-348, are grouped together which are all isolated from the same host, human. It is the same with four clusters, {170, 175}, {174, 176 and 185}, {171-173 and 184}, and {162, 163}. The hosts of viruses in four clusters are chicken, mallard, ruddy turnstone and quail, respectively.

Moreover, the virus strains of cluster of 7 HA proteins, {166,168, 186, 190, 193-195}, are from same location, USA. The same case applies to the three clusters, {164,167}, {179, 182}, and {180, 181}, respectively. Hence, we could deduce that the host and geographical
distributions factor play very important roles in the HA protein of the subtype H7 evolution process.

The H7 phylogenetic tree also shows that H7N9 viruses infected human are genetically close to ducks, AEK84770 (H7N7), BAH22785 (H7N9) and BAN16711 (H7N9). It indicates that the HA proteins of the novel avian influenza H7N9 virus might have originated from avian influenza viruses of duck origin. The results are all in accordance with the conclusion of the Refs.[26-29].

Similar to the H7 phylogeny, the phylogenetic tree of H5 can be divided into two main branches roughly in Fig.4 according to their host, avian and human. Furthermore, four
subclades are from four different areas, {111, 112, 116-119, 130,136} in USA, {331, 332, 335, 336, 338and339} in Vietnam, {319, 321} in Cambodia, and {322-324} in Hong Kong, respectively. So the phylogenetic relationship of HA protein of H5 is mainly related to species virus host. If the HA proteins of H5 have the same host, the location would become important for their phylogeny.

From the phylogenetic tree of H5 and H7, species virus host and geographical distribution are the most important two factors for HA protein evolution.

**Conclusions**

The 3D graphical representation has been proposed for studying the origin and evolution of influenza A virus HA protein. And there is a one-to-one correspondence between protein sequences and their 3D graphical representation. To show the utility of the approach, the phylogenetic tree of eight ND6 proteins was constructed.

The phylogenetic tree is constructed for different subtypes of influenza A virus HA protein sequences. From the phylogenetic tree of HA of H5 and H7, we could infer that this phylogenetic relationship is mainly related to species virus host and geographical distribution. The same results can be found in other subtypes. So we can infer that species virus host and geographical distribution possess vital effect on evolution routes of HA protein of avian influenza A. The result is also identified with the previous study[10, 30].

In addition, the HA proteins of the novel avian influenza A virus H7N9 might have originated from avian influenza viruses of duck origin based on the H7 phylogenetic tree.

**Acknowledgments:** We thank the referees for many valuable comments that have improved this manuscript. This work was supported by the National Natural Science Foundation of China under Grant(61170110, 11171042), the Zhejiang Provincial Natural Science Foundation of China (LY14F020049,LY13F020043), and Zhejiang Sci-Tech University 521 talent project.
References


