

Research Article

Phylogenetic Analysis of H7N9 Avian Influenza Virus Based on a Novel Mathematical Descriptor

Yusheng Bai,¹ Tingting Ma,¹ Yuhua Yao,² Qi Dai,² and Ping-an He¹

¹ School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China

² School of Life Science, Zhejiang Sci-Tech University, Hangzhou 310018, China

Correspondence should be addressed to Ping-an He; pinganhe@zstu.edu.cn

Received 28 February 2014; Revised 13 May 2014; Accepted 23 May 2014; Published 16 June 2014

Academic Editor: Tatsuya Akutsu

Copyright © 2014 Yusheng Bai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A new mathematical descriptor was proposed based on 3D graphical representation. Using the method, we construct the phylogenetic trees of nine proteins of H7N9 influenza virus to analyze the originated source of H7N9. The results show that the evolution route of H7N9 avian influenza is from America through Europe to Asia. Furthermore, two samples collected from environment in Nanjing and Zhejiang and one sample collected from chicken are the sources of H7N9 influenza virus that infected human in China.

1. Introduction

In February 2013, two patients with severe pneumonia were admitted to Shanghai Fifth Hospital affiliated with Fudan University [1]. Both patients died in one week, respectively. Chinese experts concluded that the patients were infected with novel avian influenza A H7N9 [2, 3]. 131 people were diagnosed with novel avian influenza A H7N9 in China according to National Population and Family Planning Commission of China, updated to May 7, 2013. This particular A (H7N9) virus had not previously been seen in either animals or people until 2013 in China [4].

The influenza virus consists of eight negative-strand RNA molecules surrounded by an envelope. The envelope contains the HA and NA proteins. There are 17 known HA subtypes and 10 known NA subtypes. Many different combinations of HA and NA proteins are possible. Six of the eight mRNAs code for single proteins, while the remaining two code for two proteins by differential splicing of the RNA. Each mRNA segment is associated with multiple copies of the nucleocapsid protein (NP) and an RNA polymerase (made from the viral proteins PB1, PB2, and PA). H7N9 avian influenza virus is previously only isolated from birds. No case of human infection has been reported before. The Lancet published a

research report, in which the H7N9 avian influenza virus has some types of mutations associated with circulation in humans [5].

Graphical representation of biological sequences is one of the most commonly used models to analyze protein sequences [6–26]. Since Hamori and Ruskin proposed H-curve for studying genomic data in 1983, more and more improved graphical representations of DNA sequences were introduced to analyze gene data [6–9]. Then, a number of the graphical representations of DNA sequences were expanded into protein sequences to describe and analyze protein sequences [9–26]. For instance, analogously to the scheme of Jeffrey for graphical representation of DNA, Randić et al. [11, 12], Yang et al. [13], Rasouli et al. [14], and He et al. [15, 16] have suggested several 2D graphical representations of protein sequences. According to the indices of some physicochemical properties of the twenty amino acids, some graphical representations of protein sequences have been proposed to compare the similarities/dissimilarities of proteins [17–26]. Among these graphical representations, 20 amino acids are usually first represented by 20 pre-given vectors. Based on these vectors, an iterated function system is given to generate a curve representing protein sequences. The numerical characterizations of the curves are used to describe

corresponding protein sequences. The representation above provides a simple way of viewing, sorting, and comparing various protein sequences.

In this paper, we divided 20 amino acids into five groups based on the two kinds of physicochemical properties of amino acids: charge and polarity. According to the classification, we proposed a novel 3D graphical representation to describe protein sequences. A novel numerical characterization method was suggested, in which each protein sequence can be transformed into eleven vectors. The phylogenetic tree of nine ND5 proteins was constructed to illustrate our approach. The result shows that our method has better representation accuracy and improvement on generation of phylogenetic tree for protein. And then, the phylogenetic trees of nine kinds of protein sequence of H7N9 avian influenza virus were constructed based on this method. The results show that the evolution route of H7N9 avian influenza virus is from America through Europe to Asia. Moreover, the H7N9 influenza virus that infected humans is originated from three strains of which one extracted from chicken in Zhejiang and the others extracted from environment in Nanjing and Zhejiang. Furthermore, all these three strains are originated from the wild bird in Korea which is the earlier source.

2. Materials and Methods

2.1. Materials. For our analysis, we collected complete protein sequences (HA, M1, M2, NP, NS1, NS2, PA, PB1, and PB2) of the H7N9 influenza virus from NCBI-Flu database with release date updates to June 14, 2013. The number of NA, PA-X, NEP, PB1-N40, and PB1-F2 protein sequences of the H7N9 influenza virus is too small to analyze phylogeny of H7N9 influenza virus. Thus, we took 26 HA protein sequences, 27 M1 protein sequences, 25 M2 protein sequences, 27 NP protein sequences, 28 NS1 protein sequences, 25 NS2 protein sequences, 24 PA protein sequences, 21 PB1 protein sequences, and 21 PB2 protein sequences to construct the phylogenetic tree for H7N9 avian influenza virus. These proteins come from 30 H7N9 avian influenza virus strains around the world.

2.2. Methods

2.2.1. 3D Graphical Representation. Amino acids are the structural units that make up proteins. The physicochemical properties of amino acids in a protein are very important factors for three-dimensional structure and chemical reactivity of protein. In the section, we consider two physicochemical properties of amino acids: polarity and charge.

In the work of Huang et al. [23], 20 amino acids were divided into five groups according to polarity and charge: I = {A, V, L, I}, II = {F, W, M, P}, III = {G, S, T, C}, IV = {Y, N, Q, H}, and V = {K, R, D, E}. Both the first sort and second sort of amino acids are nonpolar amino acids, in which the first sort of amino acids is nonpolar aliphatic amino acids and the second sort is nonpolar aromatic amino acids. The third sort and the fourth sort of amino acids belong to uncharged amino acids. The fifth sort of amino

acids is composed of positive charged amino acids (lysine and arginine) and negative charged amino acids (aspartic and glutamic).

In this paper, we also adopt the above classification for 20 amino acids. Given a protein sequence $S = S_{(1)}S_{(2)} \cdots S_{(n)}$ with the length n , $S_{(i)}$ means the i th amino acids of the protein sequence; inspect it by stepping one amino acid at a time. Using the classification above, the protein sequence is transformed into n points in 2D space by (1)

$$\begin{aligned} x_k &= x_{k-1} + a, & k = 2, 3, \dots, n, & x_1 = 1, \\ y_k &= b \cdot j + c, & S_{(k)} \text{ belong to } j\text{th set.} \end{aligned} \quad (1)$$

For (1), the parameters a , b , and c are arbitrary constants. In the paper, we take $a = 3$, $b = 10$, and $c = -30$. We call the graph as the PC graph of S , which is denoted as C_S .

For C_S , we take the points on the line $y = 20$ as a subsequence $S^{(1)} = S_{(1)}^{(1)}S_{(2)}^{(1)} \cdots S_{(k_1)}^{(1)}$ of the protein. For the sequence $S^{(1)}$, a map $\phi_1: A \rightarrow (1, 1), V \rightarrow (-1, 1), I \rightarrow (-1, -1), L \rightarrow (1, -1)$ is adopted. Inspecting $S^{(1)}$ by stepping one amino acid at a time, for step k ($k = 1, 2, \dots, k_1$), the vertex P_i can be constructed according to iterated function system $\vec{X}_i = (1/2)\vec{X}_{i-1} + (1/2)S_{(k)}^{(1)}$, where \vec{X}_i denotes the vector corresponding to the i th point. Connecting the adjacent points, a curve can be obtained in 2D space for protein sequence $S^{(1)}$, called $C_{S^{(1)}}$. Then, we construct the L/L matrix to describe numerically the curve $C_{S^{(1)}}$. The elements of the L/L matrix are defined as the quotient of the Euclidean distance between a pair of vertexes of the curve $C_{S^{(1)}}$ and the sum of distances between the same pair of vertexes measured along the curve $C_{S^{(1)}}$. The leading eigenvalue of the L/L matrix is computed, which is denoted as $z^{(1)}$.

Again, when the points are located at line $y = 10, 0, -10$, and -20 , operating the same steps, we can also obtain other four parameters $z^{(2)}, z^{(3)}, z^{(4)}$, and $z^{(5)}$, respectively.

Finally, we construct a 3D graphical representation of protein sequence S based on the graph C_S and $z^{(i)}$ ($i = 1, 2, 3, 4$ and 5) as follows:

$$\begin{aligned} x_k &= x_{k-1} + 3, & k = 2, 3, \dots, n, & x_1 = 1, \\ y_k &= 10j - 30, & S_{(k)} \text{ belong to } j\text{th set,} \end{aligned} \quad (2)$$

$$z_k = \sum_{j=1}^5 p_k^{(j)} z_k^{(j)},$$

where $p_k^{(j)}$ is the quotient of the length of subsequence $S^{(j)}$ and protein sequence S with the length k . Connecting the adjacent points, we can obtain a curve G_S in 3D space for protein sequence S . The 3D graphical representation can be proved to be acyclic and nondegenerate in mathematics. Therefore, the correspondence between protein sequences and graphical curves is one to one.

3. Numerical Characterization

In order to numerically characterize a protein, a novel sequence descriptor is introduced as follows.

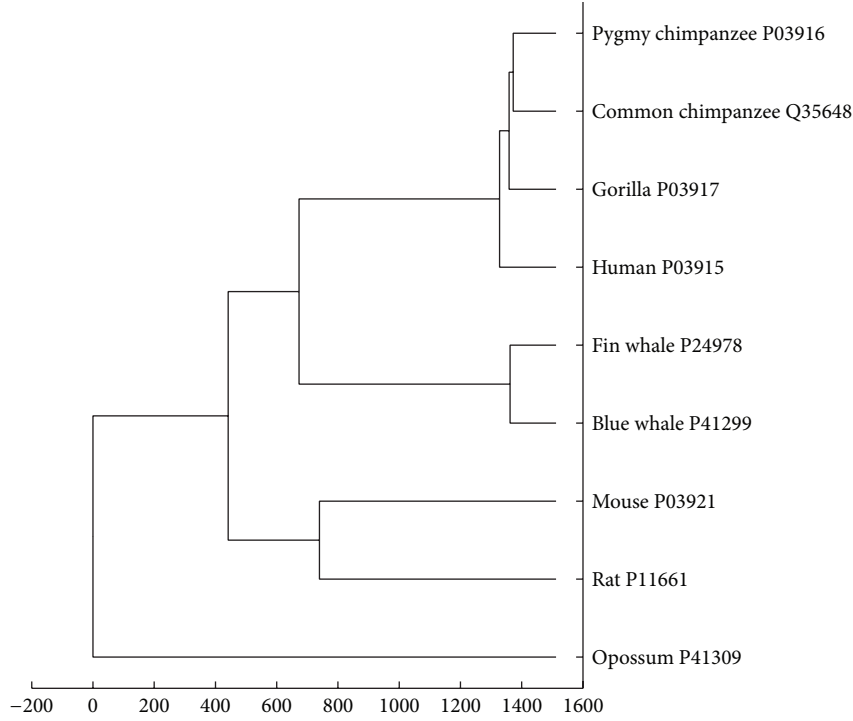


FIGURE 1: Phylogenetic tree for ND5 proteins of nine species.

For G_S , we calculate the coordinate of the center point of all points on each line $y = 20(l_1)$, $y = 10(l_2)$, $y = 0(l_3)$, $y = -10(l_4)$, and $y = -20(l_5)$, respectively, which is denoted as $(\bar{x}_k, \bar{y}_k, \bar{z}_k)$ ($k = 1, 2, 3, 4, 5$) correspondingly to five lines. For the coordinate of the center point of all points on five lines l_1-l_5 in G_S , we denote it as $(\bar{x}_6, \bar{y}_6, \bar{z}_6)$. Similarly, $(\bar{x}_7, \bar{y}_7, \bar{z}_7)$ denotes the center point of all points on four lines l_1-l_4 , $(\bar{x}_8, \bar{y}_8, \bar{z}_8)$ the center point of all points on three lines l_3-l_5 , and $(\bar{x}_9, \bar{y}_9, \bar{z}_9)$ the center point of all points on two lines l_1-l_2 , respectively. For the points on the line l_5 , we divide them into two parts according to the charge, positive or negative. One contains {K, R}; the other contains {D, E}. Using the same calculation above, two center points are obtained, $(\bar{x}_{10}, \bar{y}_{10}, \bar{z}_{10})$ and $(\bar{x}_{11}, \bar{y}_{11}, \bar{z}_{11})$.

Thus, protein sequences can be described by 11 vectors, $(\bar{x}_i, \bar{y}_i, \bar{z}_i)$ ($i = 1-11$). For two protein sequences S^i and S^j , we define their distance as the following equation:

$$D_{ij} = \sum_{k=1}^{11} \sqrt{\left((\bar{x}_k^i - \bar{x}_k^j)^2 + (\bar{y}_k^i - \bar{y}_k^j)^2 + (\bar{z}_k^i - \bar{z}_k^j)^2 \right)}. \quad (3)$$

The smaller the distance is, the closer the two protein sequences are.

3.1. Similarities/Dissimilarities of 9 ND5 Proteins. As discussed above, the similarity of sequences can be compared with the distance among them. To illustrate our method, we consider the numerical characterization of mutations and analyze the similarities among sequences belonging to nine ND5 proteins: human (*Homo sapiens*, P03915), gorilla (*Gorilla gorilla*, P03917), common chimpanzee (*Pan*

trogodytes, Q35648), pygmy chimpanzee (*Pan paniscus*, P03916), fin whale (*Balaenoptera physalus*, P24978), blue whale (*Balaenoptera musculus*, P41299), rat (*Rattus norvegicus*, P11661), mouse (*Mus musculus*, P03921), and opossum (*Didelphis virginiana*, P41309), whose sequence data were all downloaded from UniProtKB. The distances among proteins were calculated using (3). If the total number of proteins is N , a real symmetric $N \times N$ distance matrix D is constructed, whose element D_{ij} is used to reveal the evolutionary distance between protein sequences S^i and S^j .

Using the UPGMA method [27], the phylogenetic tree is obtained based on the distances between each pair of ND5 proteins, shown in Figure 1, which is consistent with the results obtained with ClustalW methods and some other methods proposed recently [19–23]. The results show that our approach has better representation accuracy and improvement on generation of phylogenetic tree for protein.

4. Result and Discussion

The avian influenza A virus genome is composed of eight single (nonpaired) RNA strands that can code for up to 14 proteins. For H7N9 influenza virus, we can compare the similarities of nine kinds of protein sequences based on their distances in the section. Using UPGMA method, the phylogenetic tree was obtained, shown in Figures 2–10. The phylogenetic trees of H7N9 genes are consistent with other published studies [4, 5, 28, 29].

Figure 2 shows that the virus strains isolated from the humans are grouped together in one clade. The virus strains, A/Nanjing/1/2013, A/Zhejiang/HZ1/2013, and

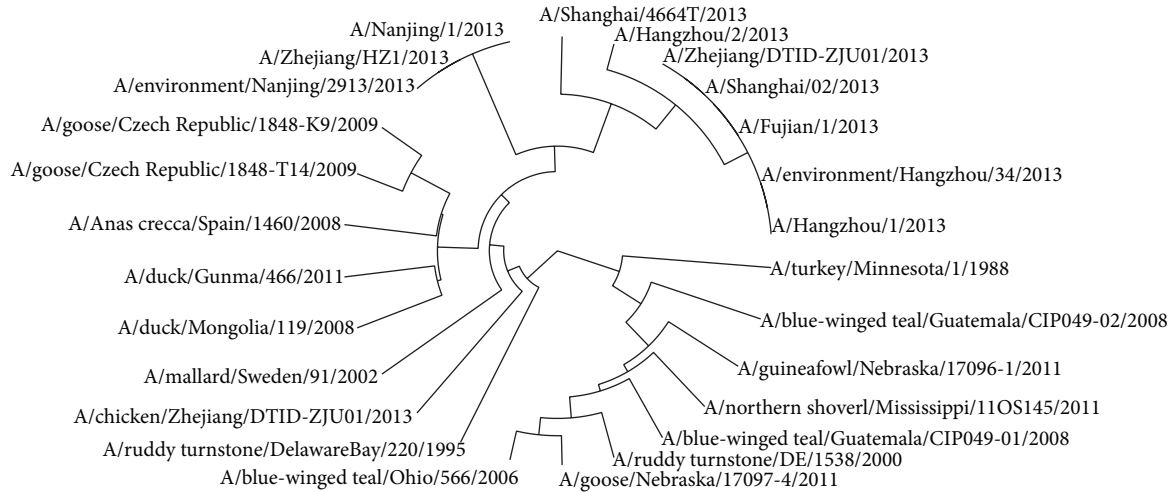


FIGURE 2: Phylogenetic trees of HA protein of H7N9.

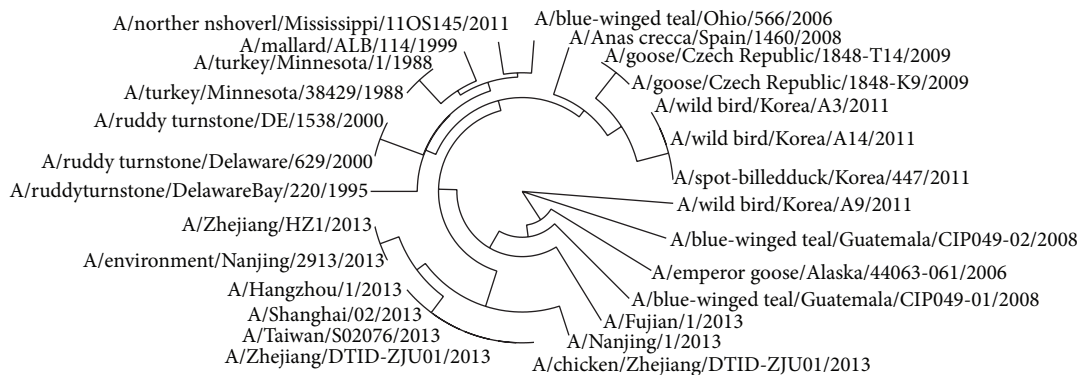


FIGURE 3: Phylogenetic trees of NP protein of H7N9.

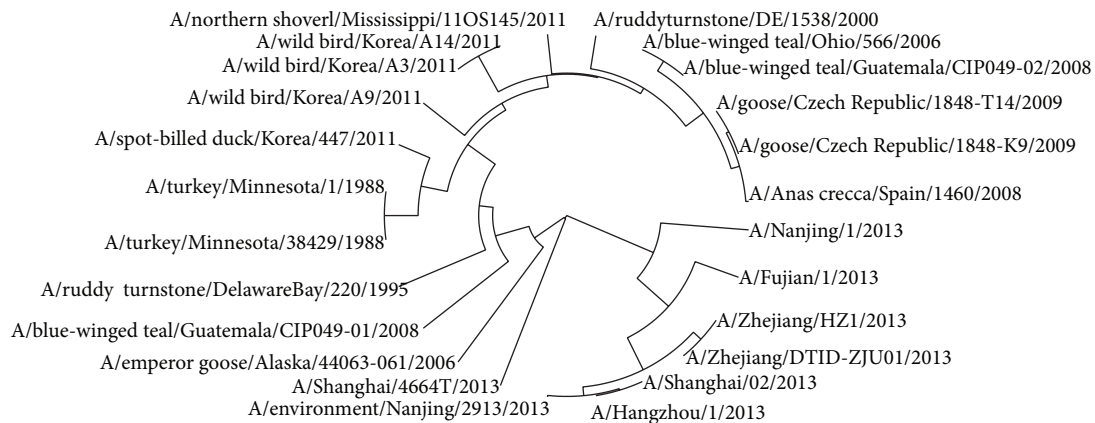


FIGURE 4: Phylogenetic trees of PA protein of H7N9.

A/environment/Nanjing/2913/2013, cluster together, which implies the HA protein of virus strains A/Nanjing/1/2013 and A/Zhejiang/HZ1/2013 probably originates from the virus strain A/environment/Nanjing/2913/2013. It is the same with all virus strains in the other cluster and the strain A/environment/Hangzhou/34/2013 is probably the source of the HA

protein of all virus strains in the other cluster. Moreover, the evolution route of 26 strains of H7N9 influenza virus suggests the HA protein of virus strain A/chicken/Zhejiang/DTID-ZJU01/2013 most probably originates from America.

In Figure 3, 9 virus strains isolated from the humans are also grouped together in a clade which contains

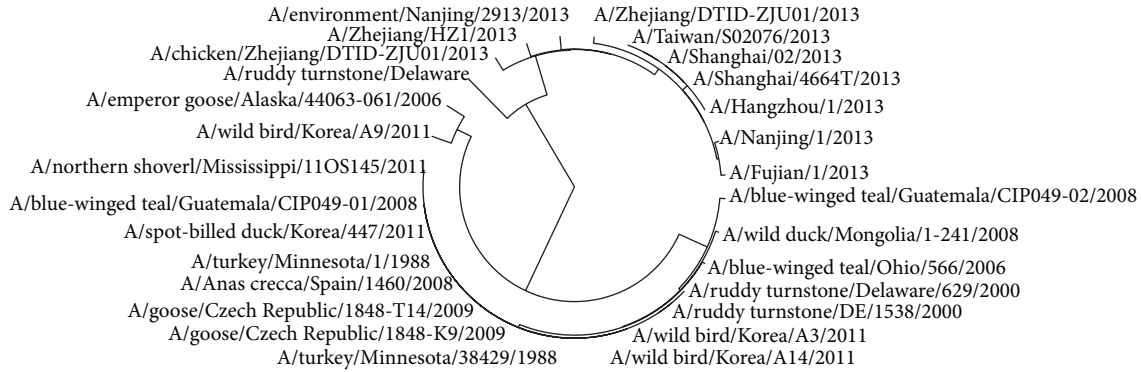


FIGURE 5: Phylogenetic trees of NS1 protein of H7N9.

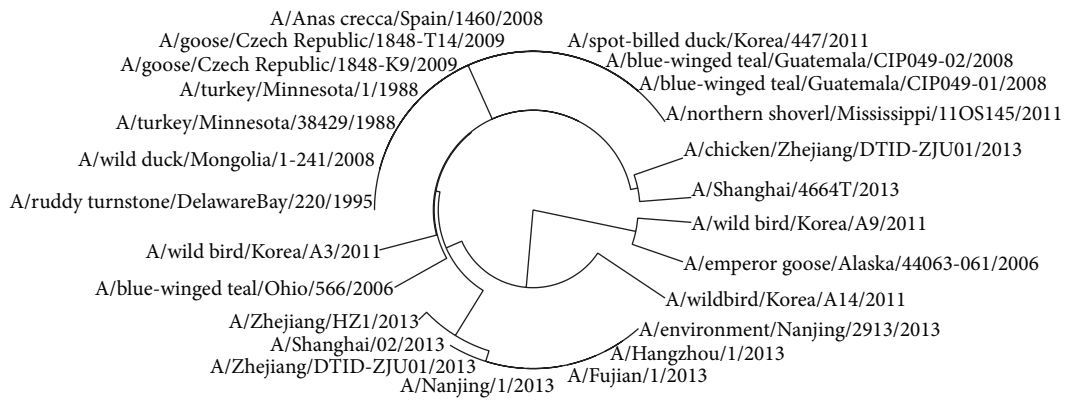


FIGURE 6: Phylogenetic trees of NS2 protein of H7N9.

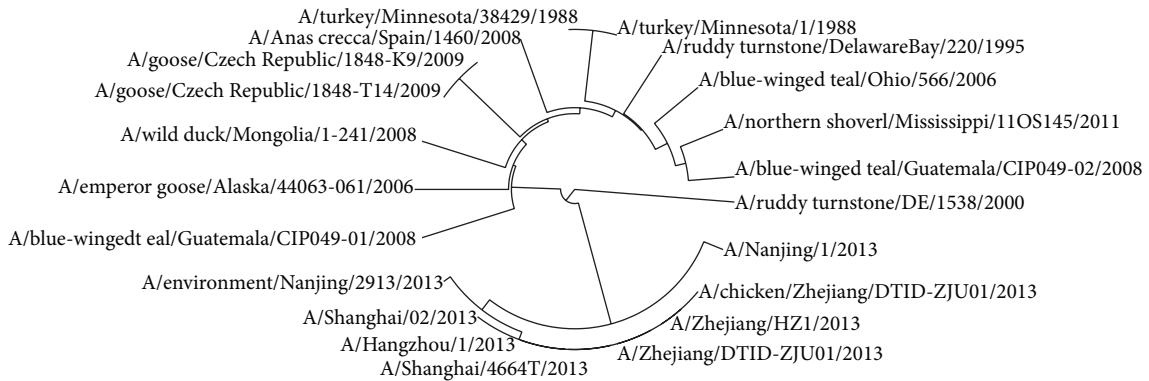


FIGURE 7: Phylogenetic trees of PB1 protein of H7N9.

two clusters. One is the virus strains A/Nanjing/1/2013, A/Zhejiang/HZ1/2013, and A/environment/Nanjing/2913/2013, which may show that virus strain A/environment/Nanjing/2913/2013 is the root of others. The other clade contains 5 virus strains, A/Hangzhou/1/2013, A/Shanghai/02/2013, A/Taiwan/so2076/2013, A/Zhejiang/DTID-ZJU01/2013, and A/chicken/Zhejiang/DTID-ZJU01/2013, which indicates that the strain A/chicken/Zhejiang/DTID-ZJU01/2013 is the source of NP protein in this clade.

Similar to Figure 3, the PA proteins of 7 virus strains that infected humans are grouped together in a clade in

Figure 4, which suggests that they originate from the strain A/environment/Nanjing/2913/2013. The evolution route of all 24 strains of H7N9 influenza virus suggests America is the earliest source. The complete PA sequences of A/chicken/Zhejiang/DTID-ZJU01/2013 and A/environment/Hangzhou/34/2013 are not provided in the NCBI-Flu database.

In Figure 5, the evolution route of 28 strains of viruses suggests that the strain A/environment/Nanjing/2913/2013 is the source of NS1 protein isolated from humans, whose earlier source is the strain A/emperor goose/Alaska/44063-061/2006

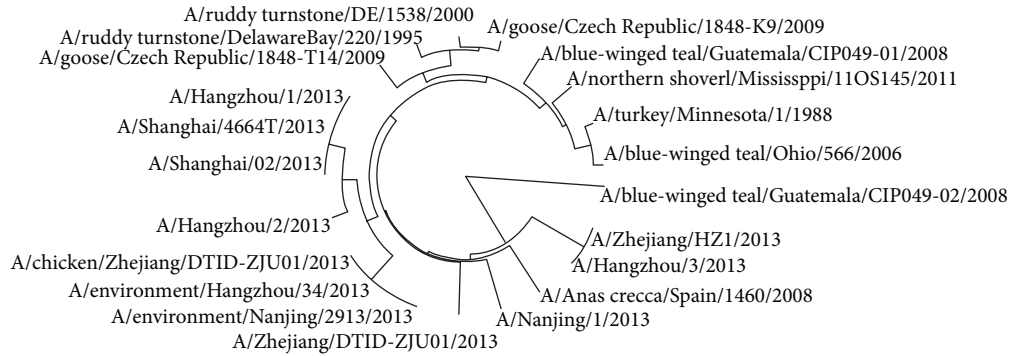


FIGURE 8: Phylogenetic trees of PB2 protein of H7N9.

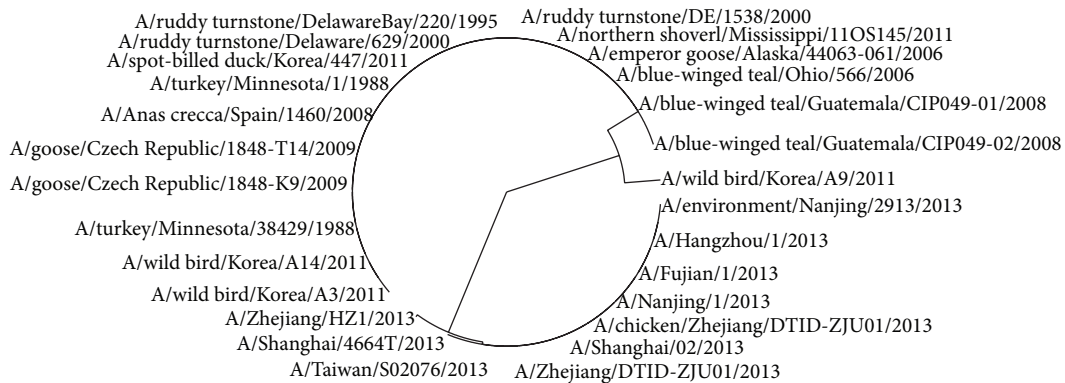


FIGURE 9: Phylogenetic trees of M1 protein of H7N9.

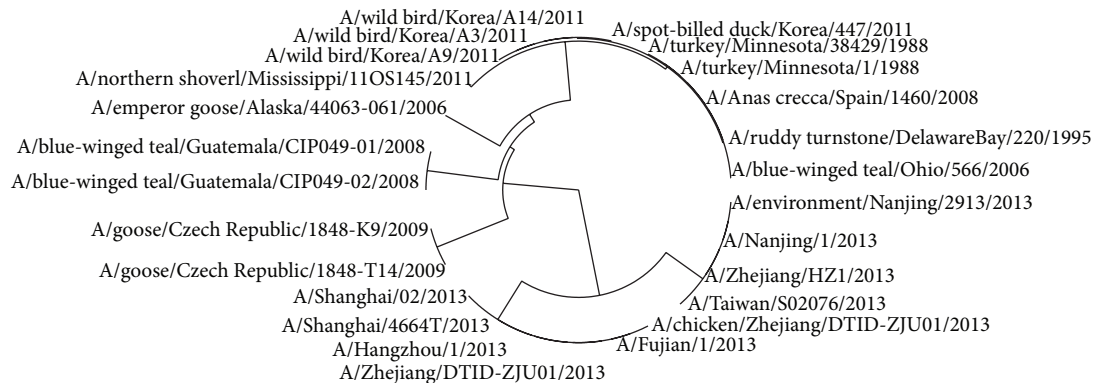


FIGURE 10: Phylogenetic trees of M2 protein of H7N9.

or A/wild bird/Korea/A9/2011. China, Korea, and Alaska are about thousands of kilometers of distance away from each other and human beings could be infected and some of patients contacted poultry before [28]. It may indicate that the strains are sourced probably through poultry trade between the above places.

In Figure 6, the phylogenetic tree of NS2 protein is similar to that of NS1. The strain A/Shanghai/4664T/2013 originates from A/chicken/Zhejiang/DTID-ZJU01/2013 and the strain A/environment/Nanjing/2913/2013 is the source of another NS2 protein of H7N9 virus extracted from human.

Observing the group of virus strains extracted from human in Figure 7, we can see that the virus strain A/chicken/Zhejiang/DTID-ZJU01/2013 is most probably the source of PB1 protein of H7N9 virus extracted from human.

In Figure 8, the PB2 protein sequences of three H7N9 virus strains A/chicken/Zhejiang/DTID-ZJU01/2013, A/environment/Nanjing/2913/2013, and A/environment/Hangzhou/34/2013 are identical, which implies their source should be the same virus. Observing Figure 8, we can see that A/Anas crecca/Spain/1400/2008 is the source of PB2 protein of H7N9 virus extracted from human.

TABLE 1: Statistics information of identical protein sequences of H7N9 available in the NCBI-Flu database.

HA	AGK82158-AGL44438-AGJ51953; AGJ73503-AGM16242-AGJ73510;
NP	ACX53684-ADN34728; AGK84853-AGL44439-AGL95089-AGJ51954-AGJ72862; ABI84698-ACZ46730; AEK84689-AEK84687;
PA	ACX53682-ADN34726; AGK84852-AGL44436; AGJ51952-AGM16237; ABI84701-ACZ46350; AEK84707-AEK84708;
NS1	ACX53686-ADN34732-ADN34743-ADK71142-AEK84817-ABI84699-ACZ47493-AEK84809-AEK84811; AGL44443-AGI60296-AGL95093; AAY96591-AAY96593;
NS2	ACX53687-ADN34733-ADN34744-ADK71143-ADK71154-AEK84818-ABI84700-ACZ47494-AEM98235; AGK82163-AGK84854-AGJ73509-AGL44444-AGJ51958-AGJ73516;
PB1	ACX53680-ADN34724; AGK84851-AGI60294-AGJ51960-AGM16239-AGJ72859; ABI84702-ACZ45969;
PB2	AGK84859-AGM16245; AGJ72858-AGK84862-AGJ73521;
M1	AGK82160-AGI60298-AGJ73506-AGL44441-AGI60290-AGL95091-AGJ51957-AGJ72864-AGJ73511; ACU44782-ADN34730-ADN34741-ADK71140-ADK71151-ABS89410-AFM09448-AGE08099-AAY96476- AFX85261-AAY96478-AEK84670-ABI84695-ACZ48319-AEK84654-AEK84656;
M2	ACU44783-ADN34731; ADN34742-AFX85262-ABI84696-ACZ48320; AGK82161-AGI60299-AGL44442-AGL44442-AGI60291-AGJ51956-AGJ72865; AGJ73507-AGL95092-AGM16244-AGJ73512; ADK71141-ADK71152; AEK84671-AEK84655-AEK84657-AEK84659;

All M1 protein sequences of H7N9 virus extracted from China are identical except A/Zhejiang/HZ1/2013, while the other avian H7N9 viruses also possess identical M1 protein sequences except A/Wild bird/Korea/A9/2011 as shown in Figure 9.

There are also two main branches roughly for Figure 10 according to their host. The strains A/environment/Nanjing/2913/2013 and A/chicken/Zhejiang/DTID-ZJU01/2013 are two resources of M2 protein of H7N9 virus extracted from human.

In summary, three strains, A/environment/Nanjing/2913/2013, A/chicken/Zhejiang/DTID-ZJU01/2013, and A/environment/Hangzhou/34/2013, are three main sources of H7N9 virus that infected humans. The phylogenetic trees of PB2 and M1 show that they have identical amino acid sequences. We conclude that three strains originated from the same source.

In addition, the H7N9 outbreak at the years 1988, 1995, 1999, 2000, 2006, 2008, 2009, 2011, and 2013 is shown in Figure 11 according to records available in the NCBI-Flu database. The time interval between two outbreaks is 7, 4, 1, and 6 years before the year 2006 and every two years afterwards. The outbreak of H7N9 is more and more frequent, which implies that H7N9 evolution is speeding up.

Observing each distance matrix *D*, we can see the same protein sequences are identical. For instance, we find that the strains A/Fujian/1/2013, A/Shanghai/02/2013, and A/Zhejiang/DTID-ZJU01/2013 have identical HA protein sequences, and the HA protein of strain A/Nanjing/1/2013 duplicated in strains A/Zhejiang/HZ1/2013 and A/environment/Nanjing/2913/2013. All identical protein sequences are listed in Table 1.

From the perspective of geography [30], we show identical sequences (NS1, NS2, M1, and M2) occurring in different locations connected by arrow lines in Figure 12. Observing the date and location of occurrence for each identical

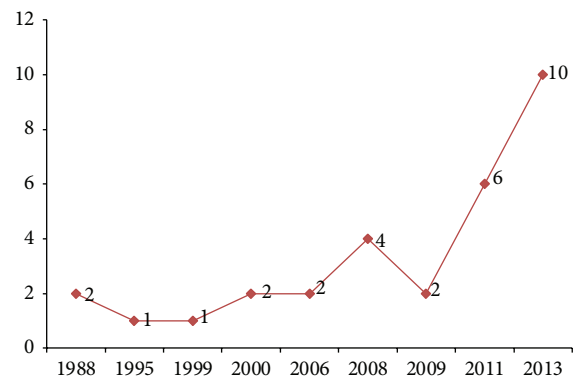


FIGURE 11: H7N9 avian influenza virus outbreak timeline.

protein sequence above, we find that the earliest H7N9 avian influenza virus comes from America, spreading to Europe and finally arriving in Asia. For example, M1 protein sequence of H7N9 avian influenza started from Minnesota (1988), passing Delaware Bay (1995), Delaware (2000), Alaska (2006), and Ohio (2006), then Guatemala (2008), Spain (2008), and Czech Republic (2009), and finally arrived at Korea in Figure 11 (2011). In another instance, the evolution route of NS1 protein sequences of H7N9 avian influenza virus is also from America (Minnesota, 1988; Guatemala, 2008) through Europe (Czech Republic, 2009) to Asia (Korea, 2011).

Comparing the results of phylogenetic trees of nine kinds of H7N9 protein, we can see that the evolution route of H7N9 avian influenza virus is from America through Europe to Asia.

5. Conclusion

The physicochemical properties of amino acids determine three-dimensional structure and the biological activity of the

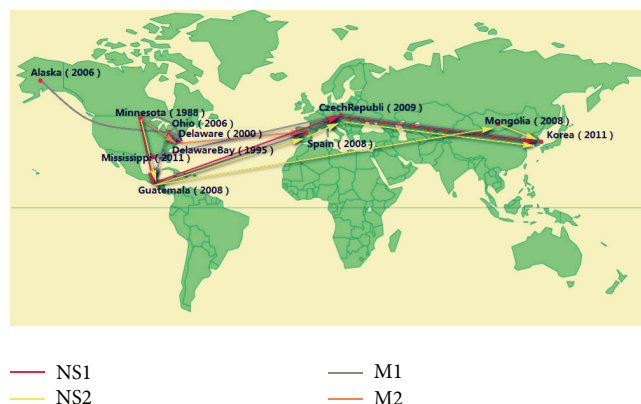


FIGURE 12: Location of identical M1, M2, NS1, and NS2 sequences found widely separated in space and time for H7N9 avian influenza virus.

protein. In the paper, according to charge and polarity of 20 amino acids, we divided them into five sorts. Then a novel 3D graphical representation was proposed, called G_S Curve, for a protein sequence. Using the defined distance, we obtain similarity matrix D by computing the distance of nine kinds of H7N9 protein sequences. By the single linkage method, their phylogenetic trees are constructed based on similarity matrix D . To show the utility of the approach, the phylogenetic tree of nine ND5 proteins was constructed.

Furthermore, the phylogenetic trees of nine kinds of protein sequences of H7N9 avian influenza virus were constructed based on this method. From the phylogenetic trees of nine kinds of H7N9 protein sequences, we deduce that two samples collected from environment in Zhejiang and Nanjing and one sample collected from chicken in Zhejiang are sources of strains that infected humans in 2013. The phylogenetic tree of PB2 and M1 shows that the three strains may originate from the same source. The result is supported by the geographic analysis of virus outbreak record.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank the referees for many valuable comments that have improved this paper. This work was supported by the National Natural Science Foundation of China under Grant nos. (61170110, 61272312), the Zhejiang Provincial Natural Science Foundation (LY14F020049), Zhejiang Qianjiang talent plan (2013R10061) and Program of Graduate Education and Teaching Reform in Zhejiang Sci-Tech University.

References

[1] T. Kageyama, S. Fujisaki, E. Takashita et al., "Genetic analysis of novel avian A(H7N9) influenza viruses isolated from patients

in China, February to April 2013," *Eurosurveillance*, vol. 18, no. 15, pp. 18–21, 2013.

- [2] J. Zhou, D. Wang, R. Gao et al., "Biological features of novel avian influenza A (H7N9) virus," *Nature*, vol. 499, no. 7459, pp. 500–503, 2013.
- [3] D. Holmes, "The world waits for H7N9 to yield up its secrets," *The Lancet Infectious Diseases*, vol. 13, pp. 477–478, 2013.
- [4] D. Liu, W. Shi, Y. Shi et al., "Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses," *The Lancet*, vol. 381, no. 9881, pp. 1926–1932, 2013.
- [5] Y. Chen, W. Liang, S. Yang et al., "Human infections with the emerging avian influenza A H7N9 virus from wet market poultry: clinical analysis and characterisation of viral genome," *The Lancet*, vol. 381, no. 9881, pp. 1916–1925, 2013.
- [6] D. Bielińska-Waz, "Graphical and numerical representations of DNA sequences: statistical aspects of similarity," *Journal of Mathematical Chemistry*, vol. 49, no. 10, pp. 2345–2407, 2011.
- [7] D. Li, "Similarity analysis of DNA sequences based on CLZ complexity," *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 2, pp. 481–487, 2013.
- [8] Y. W. Liu and Y. Peng, "A novel 2-D graphical representation for DNA sequences based on the chemical properties," *Journal of Computational and Theoretical Nanoscience*, vol. 10, pp. 2102–2105, 2013.
- [9] M. Randić, J. Zupan, A. T. Balaban, D. Vikić-Topić, and D. Plavšić, "Graphical representation of proteins," *Chemical Reviews*, vol. 111, pp. 790–862, 2011.
- [10] M. Randić, M. Nović, and D. Plavšić, "Milestones in graphical bioinformatics," *International Journal of Quantum Chemistry*, vol. 113, pp. 2413–2446, 2013.
- [11] M. Randić, M. Novič, A. R. Choudhury, and D. Plavšić, "On graphical representation of trans-membrane proteins," *SAR and QSAR in Environmental Research*, vol. 23, no. 3-4, pp. 327–343, 2012.
- [12] M. Randić, J. Zupan, and D. Vikić-Topić, "On representation of proteins by star-like graphs," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 1, pp. 290–305, 2007.
- [13] J.-Y. Yang, Z.-G. Yu, and V. Anh, "Protein structure classification based on chaos game representation and multifractal analysis," in *Proceedings of the 4th International Conference on Natural Computation (ICNC '08)*, pp. 665–669, October 2008.
- [14] M. Rasouli, G. Rasouli, F. A. Lenz, D. S. Borrett, L. Verhagen, and H. C. Kwan, "Chaos game representation of human pallid spike trains," *Journal of Biological Physics*, vol. 36, no. 2, pp. 197–205, 2010.
- [15] P. He, "A new graphical representation of similarity/dissimilarity studies of protein sequences," *SAR and QSAR in Environmental Research*, vol. 21, no. 5-6, pp. 571–580, 2010.
- [16] P.-A. He, D. Li, Y. Zhang, X. Wang, and Y. Yao, "A 3D graphical representation of protein sequences based on the Gray code," *Journal of Theoretical Biology*, vol. 304, pp. 81–87, 2012.
- [17] M. Randić, "2-D graphical representation of proteins based on physico-chemical properties of amino acids," *Chemical Physics Letters*, vol. 444, no. 1-3, pp. 176–180, 2007.
- [18] M. Randić, K. Mehulic, D. Vukicevic, T. Pisanski, D. Vikić-Topić, and D. Plavšić, "Graphical representation of proteins as four-color maps and their numerical characterization," *Journal of Molecular Graphics and Modelling*, vol. 27, pp. 637–641, 2009.
- [19] M. I. Abo El Maaty, M. M. Abo-Elkhier, and M. A. Abd Elwahaab, "3D graphical representation of protein sequences

- and their statistical characterization,” *Physica A*, vol. 389, no. 21, pp. 4668–4676, 2010.
- [20] A. Ghosh and A. Nandy, “Graphical representation and mathematical characterization of protein sequences and applications to viral proteins,” *Advances in Protein Chemistry and Structural Biology*, vol. 83, pp. 1–42, 2011.
- [21] B. Liao, B. Liao, X. Lu, and Z. Cao, “A novel graphical representation of protein sequences and its application,” *Journal of Computational Chemistry*, vol. 32, no. 12, pp. 2539–2544, 2011.
- [22] P.-A. He, J. Wei, Y. Yao, and Z. Tie, “A novel graphical representation of proteins and its application,” *Physica A*, vol. 391, no. 1-2, pp. 93–99, 2012.
- [23] L. Huang, H. Tan, and B. Liao, “HR-Curve: a novel 2D graphical representation of protein sequence and its multi-application,” *Journal of Computational and Theoretical Nanoscience*, vol. 10, no. 1, pp. 257–264, 2013.
- [24] C. Yu, R. L. He, and S. S.-T. Yau, “Protein sequence comparison based on K-string dictionary,” *Gene*, vol. 529, pp. 250–256, 2013.
- [25] Y. X. Liu, D. Li, K. B. Lu, Y. D. Jiao, and P. A. He, “P-H curve, a graphical representation of protein sequences for similarities analysis,” *MATCH Communications in Mathematical and in Computer Chemistry*, vol. 70, no. 1, pp. 451–466, 2013.
- [26] T. T. Ma, Y. X. Liu, Q. Dai, Y. H. Yao, and P. A. He, “A graphical representation of protein based on a novel iterated function system,” *Physica A*, vol. 403, pp. 21–28, 2014.
- [27] H. H. Otu and K. Sayood, “A new sequence distance measure for phylogenetic tree construction,” *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.
- [28] R. Gao, B. Cao, Y. Hu et al., “Human infection with a novel avian-origin influenza A (H7N9) virus,” *The New England Journal of Medicine*, vol. 368, no. 20, pp. 1888–1897, 2013.
- [29] Y. Wang, Z. Dai, H. Cheng et al., “Towards a better understanding of the novel avian-origin H7N9 influenza A virus in China,” *Scientific Reports*, vol. 3, article 2318, 2013.
- [30] A. Ghosh, A. Nandy, P. Nandy, B. D. Gute, and S. C. Basak, “Computational study of dispersion and extent of mutated and duplicated sequences of the H5N1 influenza neuraminidase over the period 1997–2008,” *Journal of Chemical Information and Modeling*, vol. 49, no. 11, pp. 2627–2638, 2009.