Evolution Analysis for HA Gene of Human Influenza A H3N2 Virus (1990 - 2013)

Su-Li Li, Meng-Zhe Jin, Zhao-Hui Qi*
College of Information Science and Technology, Shijiazhuang Tiedao University
Shijiazhuang, Hebei, 050043, People's Republic of China
*E-mail: zhqi_wy2013@163.com

***Abstract*—Human influenza A virus is an important pathogen which threatens the health of human in a long time. The mutation study of HA gene is the most important. Here we investigate the evolution characteristics of HA gene of H3N2 influenza virus from 1990 to 2013. Numerical mapping and PCA clustering analysis are applied to the gene evolution analysis. The clustering diagram by MATLAB represents the mapping of HA gene in 2D space. The first two principal components account for 78.48% by PCA analysis. And the points are clustered into three parts, 1990~1999, 2000~2005 and 2006~2013. However, there is no obvious interval among them. Then we show the graphical representation of HA gene sequences according to the emerging time of isolates and different continents. Results show that during 1990 to 2013 human influenza A H3N2 virus has been evoluting gradually. There was not large genetic recombination. Even so, it is necessary to continuously monitor the human influenza A (H3N2) viruses.**

Keywords—HA gene; H3N2; evolution; human influenza

I. INTRODUCTION

In a long time, influenza A virus remains an important pathogen which threatens human health. The viruses can spread with acute onset, contagious and high incidence of local high incidence. There are three kinds of antigens of influenza virus. They are hemagglutinin (HA), neuraminidase (NA) and membrane protein (MZ). Their protrusions on the surface of the virus cause the host's immune system to produce the immune response. Hemagglutinin (HA) is the most important for the human influenza A virus, since it can induce the production of antibodies that can stop the virus continues to spread.

Human influenza A (H3N2) appeared in 1968. Then it broke out to the global and eventually killed about one million people [1]. It is a disease of the respiratory system caused by influenza A H3N2 virus. The influenza virus has continuous antigenic mutations. The HA gene mutation is the most important mutation [2]and has the highest mutation frequency [3] among the three antigens. The genetic mutation can make it difficult for the host immune system to identify and remove viruses. Therefore, the virus can break the body's immune barrier repeatedly and effectivelythat leads to the spread of influenza. Thus the research on HA gene evolution has an important practical significance. In order to have a good analysis on the evolution characteristics of HA gene, we chose the HA gene samples (1990 - 2013) downloaded from NCBI as the research object, using the method of numerical mapping and PCA clustering. The results would help to predict the evolution trend of human influenza A H3N2 virus.

II. MATERIALS AND DATA PREPROCESSING

A. Acquisition of HA Gene Sequence

We can query and download the HA gene sequences from influenza virus database NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html). NCBI is one of the largest three bioinformatics database. The database contains the most global influenza virus gene sequence and is free to download. We obtained 7629 worldwide full-length HA gene sequences of human influenza A (H3N2) viruses from 1968 to 2013.

B. Distribution of HA Gene Sequence Samples

We first give a simple statistical analysis on HA gene sequence samples, so that we can find out the change rule by different continents and different time. Table 1 shows the distribution of sequence samples.

TABLE I. RELEASE FULL-LENGTH SEQUENCES BY DIFFERENT REGION AND TIME

Continents	*1968-1979*	*1980-1989*	*1990-1999*	*2000-2005*	*2006-2013*	*Sum*
Africa	0	0	5	0	46	51
Asia	41	36	139	365	884	1465
Europe	49	28	131	282	148	638
N. America	52	26	413	400	3435	4326
Oceania	11	5	33	641	28	718
S. America	0	0	0	0	431	431
Sum	153	95	721	1688	4972	7629

From Table 1 we see that the number of sequences shows a rapid growth according to the published data in recent years. Data samples from 1968 to 1989 are limited because of the difficulty of gene collection. The data can not cover the global due to the different national conditions and the difficulty of collecting the gene. We choose the data of 1990 - 2013 as the research object. From the overall view, our choices about the sequences have stochastic characteristics. The stochastic

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

characteristics can avoid the biases for the evolution history of influenza.

C. Preprocessing of HA Gene Sequence

HA gene samples contain their emerging date, but the downloaded sequences are not in chronological order. However, we need the chronological gene sequences as the sample data of PCA method. To eliminate data redundancy, we must merge the same sequences with the same date into one. So we need to do some preprocessing on the sequences. The sequences are sorted by their time record using Perl language. After merging the same sequences with the same date into one, the total number of sequences changes into 7381 from the original 7629. Here we give a simple example to show the HA gene sequences ordered by time.

>cds:AHL89054 A/New York/1005/2006 2006/01/04 HA
ATCATGTGGGCCTGCCAAAAAGGCAACATTAGGTGC
AACATTTGCATTTGA
>cds:AHL89066 A/New York/1006/2006 2006/01/08 HA
ATCATGTGGGCCTGCCAAAAAGGCAACATTAGGTGC
AACATTTGCATTTGA
>cds:AEG65208 A/Georgia/NHRC0001/2006 2006/01/09 HA
ATCATGTGGGCCTGCCAAAAAGGCAACATTAGGTGC
AACATTTGCATTTGA

III. Digital Mapping of HA Gene Sequence

HA gene sequences are alphabetic strings composed of four bases A, G, C and T with some rules. Exploration of the arrangement rules is one of the most important research topics of bioinformatics, which seems to be hidden deeply in the sequences. In general, it is difficult to come to a conclusion by analyzing DNA sequence with great deals of characters, especially when the sequence is long. In this fields, statistical method has become an important means in biological information processing, such as frequency, correlation analysis, Markov chain, neural network, Bayesian statistics, information entropy [4-6].

Statistical method used in this paper is to calculate the frequencies of the four bases and map the four nucleotides to four frequencies. Then the gene sequence is converted to a four-dimensional frequency vector.

The frequency vector is calculated as follows,

$$P(x) = \frac{N(x)}{N}, \quad x = \{A, G, C, T\}$$

where x is one of the four nucleotides and $N(x)$ is the repeat count of x in the sequence. The N denotes the total number of bases in a sequence.

$$N = N(A) + N(G) + N(C) + N(T)$$

Here we give an example to show the proposed frequency vector. There are five randomly chosen HA gene samples of H3N2 from Table 1, ACI26560 (A/Siena/4/1990), ACI26571 (A/Siena/10/1990), AFG99512 (A/England/260/1991), AFG72151 (A/Geneva/6447/1991) and AFG99855 (A/Paris/417/1991). We write a Perl program to calculate the base frequencies of HA gene sequences. The results are a set of high-dimensional vectors as shown in Table 2.

TABLE II. Frequency Vectors of Five Randomly Chosen HA Gene Samples

Accession	*A*	*G*	*C*	*T*
ACI26560	0.22923976	0.20467836	0.23450292	0.32631578
ACI26571	0.23040935	0.20350877	0.23450292	0.32631578
AFG99512	0.23040935	0.20116959	0.23742690	0.32573099
AFG72151	0.23157894	0.20233918	0.23567251	0.32514619
AFG99855	0.23216374	0.20350877	0.23450292	0.32456140

From Table 2, we can see there is no significant change of the four frequencies of bases A, G, C and T in different sequences. In addition, it is complicated to analyze the four-dimensional data. Therefore, some specific data analysis method is needed to process the high dimension data with redundancy. In this paper, we adopt PCA (Principal Component Analysis) method, which will be introduced in Section 4 in detail.

IV. Evolution Analysis for HA Gene Based on PCA Method

A. PCA Method

About feature selection in pattern recognition, we should choose the data that have the same dimensions as the original data. This, however, is only in theory. High dimensional data are correlated to each other, and have certain data redundancy. At the same time, the high-dimensional correlated data increase the complexity of analyzing. Therefore, we hope to find a transformation that the data set could be efficiently transformed to lower dimension. PCA is such a solution in multivariate statistical method.

Now, Principal components analysis (PCA) and its many expanded methods have been successfully applied to the resolution of many biological problems [7-9]. PCA is a projection method to analyze data set [10]. It reduces data set from high-dimensional space to few hidden variables while keeping important information on its variability. This method can extract the most important features from the correlation variables. Thus we can eliminate data redundancy and transform the high-dimensional data to be lower dimensions. Then we will get a more simplified data model. The data will be compressed indirectly to release the simple structure behind the complex data. At the same time, we can largely retain the original data information.

Compared with other data dimensionality reduction method, PCA is a universal applicable method. It has many advantages such as perfect in theory, simple in concept, convenient in calculation. It has optimal linear reconstruction error. PCA method is a linear dimensionality reduction method. Its dealing with the real data is simple and effective, comparing with nonlinear dimensionality reduction methods. The method is suitable for eliminating secondary characteristics. It can extract the main factors from multi-dimensional vectors and expresses the main features of the vectors.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

B. PCA Cluster of HA gene sequences

The HA gene sequences are converted into a 4D vector set based on the time-ordered sequences in Section II and the digital mapping in Section III. However, it is difficult for us to understand the high-dimensional data directly. In this paper, we use PCA method to reduce the dimensions of the digital vectors dataset from 4D to 2D. Then graphical representations of the data in a 2D space is obtained in this section.

We use the princomp function of Matlab to process the 4D data set. The command format shows as follows,

$$PC = princomp(M);$$

$$[PC, SCORE, lantent, tsquare] = princomp(M);$$

The input matrix M is 4D digital base frequency representation of HA gene samples. The output PC is the matrix comprised of principal components. The SCORE is the sample principal component score. The tsquare is the Hotelling T^2 statistic data for each data point.

Table I. shows the distribution of 7629 worldwide full-length HA gene sequences of H3N2 viruses from 1968 to 2013. We first get the corresponding 4D base frequency vector set of 7629 HA gene sequences. Then the 4D data set is reduced to a 2D data set with the PCA method. Table III. shows some parameters of principal component analysis on the initial variables.

TABLE III. SEVERAL PARAMETERS OF PRINCIPAL COMPONENT ANALYSIS ON THE 4D FREQUENCY VECTOR SET

Principal axes	*Eigenvalues (10^{-6})*	*Variance contribution (%)*	*Cumulative variance contribution (%)*
The first	14.6310683	57.88	57.88
The second	5.20693331	20.60	78.48
The third	4.37947098	17.33	95.81
The fourth	1.06079334	4.20	100.00

From Table III. we can see that the first and the second principal axes account for 57.88% and 20.60% of the total inertia of the 4D space, respectively. The first two principal axes contain the most information of the sample. So it is feasible to reduce the four-dimensional data into the two-dimensional data by the first two principal axes.

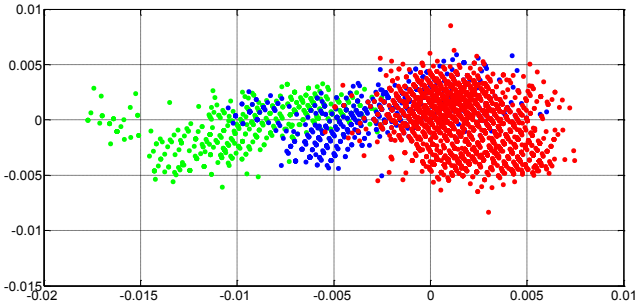
C. Graphic representation of isolate clusters from 1990 to 2013

The clustering characteristics of influenza A H3N2 virus can be obtained directly by the graphical representation method. So we can easily analyze the trend in the evolution of influenza virus. Using image processing function of MATLAB software, the data after dimension reduction can be represented in the Fig. 1 and Fig. 2. In this paper, we take the first principal axis as the X axis and the second principal axis as the Y axis. Then each HA gene sequence can be plotted in a 2D clustering figure. In order to get the evolution characteristics of sequences in time dimension, we add the time-ordered serial number of the sequences from the data set as the Z coordinate. Then we get the 3D graphical representation of the data set in Fig. 2.

In Fig. 1, the data are divided into 3 classes. We take the data from 1990 to 1999 as the first class and mark it in green, and mark the data from 2000 to 2005 as the second class in blue color and the data from 2006 to 2013 as the third in red.

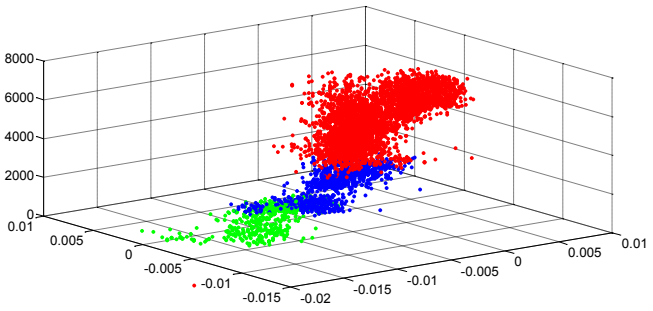
The results in Fig. 1 illustrate that the evolution of the human influenza A H3N2 virus from 1990 to 2013 seems to be a gradual and irregular process. However, there is still some clear information of evolution. The evolution trend shows in Fig.1 is from left to right and points to the opposite direction of early virus genes. The recent virus has more significant difference with the samples from 1990 to 1999 than other years. This evolution trend suggests that in the following years the new clusters should be located on the right side of samples from 2006 to 2013. Besides, the non-obvious aggregation phenomenon in Fig. 1 clearly indicates that there is no frequent genetic recombination of the human influenza A H3N2 virus from 1990 to 2013.

FIG I. 2D SPACE CLUSTERING MAP



As seen in Fig. 2, the overall 3D figure of H3N2 virus can be regarded as an approximately continuous curve. There is no distinct gap between the different classes. H3N2 influenza virus is constantly mutating during these years, evolving and arousing rounds of flu outbreak.

FIG II. 3D SPACE CLUSTERING MAP



D. Continent-difference evolution characteristics by Graphic representation from 1990 to 2013

In order to further describe the evolution characteristics of H3N2 influenza virus, the regional distribution of H3N2 influenza virus is considered in this section. We classify the viruses into 6 types according to their collection continents and then analyze their continent-difference evolution characteristics. The graphical representation of HA genes of the viruses collected in North America, Asia, Europe, Oceania, South

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

America and Africa are illustrated in Fig. 3 respectively by using the same method in Fig. 1.

FIG III. 2D SPACE CLUSTERING MAP OF SIX REGIONS

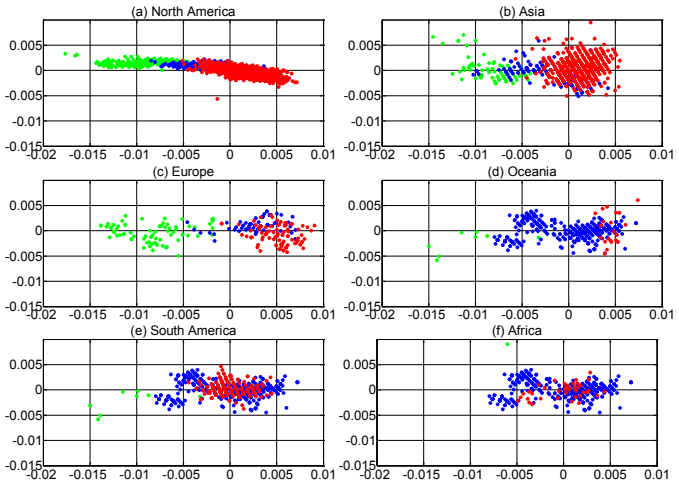


Fig. 3 shows that the two pictures representing Asia and Europe are quite similar to each other. That is to say, the evolution characteristics of H3N2 influenza virus in Asia and Europe are approximately identical during the time from 1990 to 2013. Further, it suggests that viral populations in the two regions are the same. However, we can also find that the evolution characteristics in North America are quite different from the others. It can be inferred the special evolution characteristics are the reason why H3N2 influenza virus displays more widespread property in North America than in other regions. From the point of view of northern and southern hemispheres, there are also significant differences between the pictures representing North America, Asia and Europe and those representing Oceania and South America.

V. CONCLUSIONS

In this paper, we analyze the evolution characteristics of human influenza A (H3N2) viruses base on the HA gene of the H3N2 isolates from 1990 to 2013. In order to get a quantitative analysis, we first build up a digital mapping of DNA sequence according to the frequency of bases A, G, C and T in the sequence. We get a 4D vector set consisting of the frequencies. Then PCA method is used to reduce the dimensionalities of the 4D vector set from 4D to 2D. The graphical representation of the 2D vector set shows the moderate evolution characteristics of human influenza A (H3N2) viruses from 1990 to 2013. Even so, it is necessary to monitor continuously the human influenza A (H3N2) viruses. Then the continent-difference 2D graphical representation is illustrated in 6 pictures. And we compare the different evolution characteristics of the H3N2 viruses from the 6 different continents.

ACKNOWLEDGMENT

This work supported by the National Natural Science Foundation of China (Grant No. 61272254), and by the Natural Science Foundation of Hebei Province, China (Project No. F2012210017), and by the Humanities and Social Sciences Research of Ministry of Education of China (Project name, The Origin, Propagation and Migration of Human Influenza Epidemic (1918-2010) from Space-time Perspective; Project No. 11YJCZH132).

REFERENCES

[1] M. I. Nelson, E. C. Holmes, The evolution of epidemic influenza. Nature, 2007, 8(3), 196-205.

[2] S. E. Lindstrom, N. J. Cox, A. Klimov, Genetic analysis of human H2N2 and early H3N2 influenza viruses, 1957-1972: evidence for genetic divergence and multiple reassortment events, Virology, 2004, 328(1), 101-119.

[3] Y. C. Liao, M. S. Lee, C. Y. Ko, A. H. Chao, Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus, Bioinformatics, 2008, 24(4), 505-512.

[4] Z. H. Qi, X. Q. Qi, Numerical characterization of DNA sequences based on digital signal method, Computers in Biology and Medicine, 2009, 39, 388-391.

[5] Z. H. Qi, J. M. Wang, X. Q. Qi, Classification analysis of dual nucleotides using dimension reduction, Journal of Theoretical Biology, 2009, 260, 104-109.

[6] O. A. Schmitt, H. Herzel, Estimating the Entropy of DNA Sequences, Journal of Theoretical Biology, 1997, 188(3), 369-377.

[7] J.C. Costa, M. M. Alves, E. C. Ferreira, Principal component analysis and quantitative image analysis to predict effects of toxics in anaerobic granular sludge, Bioresource Technology, 2009, 100, 1180-1185.

[8] F. Xiao, S. W. Wang, X. H. Xu, G. M. Ge, An isolation enhanced PCA method with expert-based multivariate decoupling for sensor FDD in air-conditioning systems, Applied Thermal Engineering, 2009, 29(4), 712-722.

[9] Z. H. Qi, J. Feng, C. C. Liu, Evolution trends of the 2009 pandemic influenza A (H1N1) viruses in different continents from March 2009 to April 2012, Biologia, 2014, 69(4), 407-418.

[10] J. E. Jackson, J. W. Wiley, User's Guide to Principle Components. Wiley-Interscience, 1991, New York.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE