Development of Neural Network for BLSOM Clustering of HA Genes of Avian Influenza Viruses Isolated in Guangdong Province

Chunjian TIAN*, Qiong LUO, Jiahui GAO, Zhixiong LIN, Haiqiong YU, Zhiling LIU, Ru CHEN, Xiaowei WU

Guangdong Inspection and Quarantine Technology Center/Guangdong Provincial Key Laboratory of Animal and Plant and Food Import and Export Technology/AQSIQ State Key Laboratory of Avian Influenza, Guangzhou 510623, China

Abstract A neural network classification method, and a batch-learning self-organizing map (BLSOM), was established using trinucleotide and tetranucleotide in the hemagglutinin gene sequences of 25 avian influenza viruses isolated in Guangdong Province. Statistical analysis and normalization of the fragment number were done and MATLAB function was used to simulate the human brain thinking for self-organizing learning. When the number of training steps was 100 and above, the strains could be successfully clustered. H_1, H_3, H_5, H_7 and H_9 subtype strains fell within different classes, respectively, and the HA gene cluster map of H_3N_2 and H_7N_9 strains was quite similar, suggesting that these strains shared the same origin; H_5N_1 strain was quite different in different years; H_1N_1 and H_9N_2 strains could be clustered into one group, indicating the natural recombinant variation in the two kinds of viruses, thereby providing a reference for high-risk strain screening and traceability.

Key words Avian influenza virus, BLSOM, HA gene, Neural network, Classification

1 Introduction

Avian influenza can cause great economic losses to poultry industry, and according to statistics of OIE, H_5N_1 avian influenza infected 668 humans, and the mortality rate reached 58.83%. Currently, the emerging H_7N_9 avian influenza continues to be prevalent, and there are new human cases, making it more difficult to prevent and control avian influenza. In fact, the zoonotic influenza virus is spread to humans ultimately, and brings a great threat to public health[1]. In the avian influenza virus monitoring, with the rapid progress of high-throughput sequencing technology, the gene pool data increase sharply, making the classic phylogenetic tree analysis difficult, so there is a need to conduct technological innovation. Pearl River Delta is located in the international migratory routes of migratory birds, and has a mild and humid climate. It is the world's avian influenza outbreak center. Survey shows that the avian influenza infection is increasingly serious in the live poultry markets of Guangzhou, Jiangmen, Zhaoqing and other places, and these markets become important repository of virus[2-4]. The positive rate reaches 32.73%, and it is even up to 75% on the chopping board for the slaughter[5]. It is also found that it is dominated by H_9 subtype, and there are also H_5 and H_7 subtypes, greatly different from the vaccine strains[6]. At the same time, the majority of residents are often in contact with live poultry, and humans will be in critical condition or even die after infection[7-8]. Therefore, it is of great significance to study the classification methods for avian influenza virus.

Received: September 18, 2016 Accepted: November 20, 2016

Supported by National Entry-Exit Inspection and Quarantine Research Project (2015IK054); Special Project of Ministry of Science and Technology for Development of Major Scientific Instruments and Equipments (2012YQ0901-9705).

* Corresponding author. E-mail: gzvettian@163.com

2 Materials and methods

2.1 Data sources All strains used in this study came from Guangdong Province, providing all of the HA gene sequences. Among them, H_1N_1 influenza virus was isolated from the pig strains by South China Agricultural University and from the human strains by the medical unit; H_3N_2 strains were viral samples isolated by Guangdong Provincial Center for Disease Control from the patient; H_5N_1 viruses were the poultry strains isolated by Harbin Veterinary Research Institute and South China Agricultural University, and human infections isolated by National Influenza Center; H_7N_9 viruses were the recent epidemic strains, provided by Guangdong Provincial Center for Disease Control, 3 isolated from humans and 2 from chickens; H_9N_2 was isolated from chickens by Harbin Veterinary Research Institute and South China Agricultural University.

2.2 Research methods

2.2.1 BLSOM algorithm. We established BLSOM (Batch-Learning Self-Organizing Map) artificial neural network to receive the external input to produce different response regions and simulate human brain's self-organizing learning process[9]. The Euclidean distance was calculated as follows:

$$d_j = Px - w_jP = \sqrt{\sum_{i=1}^{N}[x_i(t) - W_{ij}(t)]^2}.$$

2.2.2 Data normalization. We measured the number of HA gene fragments for each strain[10], and the normalization formula was as follows:

$$\overline{X_i} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

2.2.3 MATLAB realization. The programming was based on Reference [11], and MATLAB (2014) was run. Part of the program codes are as follows:

```
fx > >
%% clear environment variable
clc
clear
%% input data
% load data
load ('c:\data.mat');
P = data;
……
```

The function *newsom* was used to establish SOM network, and the competition layer consisted of $6 \times 6 = 36$ neurons. The functions *train* and *sim* were used for training and simulation; the function *plotsom* was used to draw the variable relation diagram; the function *vec2ind* was used to convert data.

Table 1 BLSOM characteristic genetic fragment statistics about the avian influenza virus prevalent in Guangdong Province

Strains	Gene bank codes	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ
H_1N_1(1)	A/GuangzhouSB/01/2009(H_1N_1)	7	59	6	0	48	4
H_1N_1(2)	A/Guangzhou/506/2006(H_1N_1)	7	58	7	0	49	4
H_1N_1(3)	A/swine/Guangdong/1/2010(H_1N_1)	5	52	5	2	40	1
H_1N_1(4)	A/swine/Guangdong/11/2009(H_1N_1)	6	55	3	4	44	3
H_1N_1(5)	A/Guangdong/1513/2012(H_1N_1)	7	57	5	0	46	4
H_3N_2(1)	A/Guangdong/522/2009(H_3N_2)	5	58	8	3	34	6
H_3N_2(2)	A/Guangdong/472/2009(H_3N_2)	5	56	6	2	35	6
H_3N_2(3)	A/Guangdong/93/2008(H_3N_2)	5	53	7	3	34	6
H_3N_2(4)	A/Guangdong/578/2008(H_3N_2)	5	54	7	3	34	6
H_3N_2(5)	A/Guangdong/560/2009(H_3N_2)	5	58	8	3	34	6
H_5N_1(1)	A/Goose/Guangdong/1/96(H_5N_1)	8	50	6	6	39	3
H_5N_1(2)	A/duck/Guangdong/23/2004(H_5N_1)	5	52	5	6	48	5
H_5N_1(3)	A/Guangdong/1/2006(H_5N_1)	6	48	6	6	45	5
H_5N_1(4)	A/Goose/Guangdong/3/97(H_5N_1)	8	53	7	7	41	3
H_5N_1(5)	A/parrot/Guangdong/C99/2005(H_5N_1)	8	52	8	7	45	4
H_7N_9(1)	A/Guangdong/05/2013(H_7N_9)	5	49	11	4	39	1
H_7N_9(2)	A/Guangdong/04/2013(H_7N_9)	5	49	11	4	39	1
H_7N_9(3)	A/Guangdong/03/2013(H_7N_9)	5	48	12	4	40	1
H_7N_9(4)	A/environment/Guangdong/25/2013(H_7N_9)	5	44	10	4	38	1
H_7N_9(5)	A/environment/Guangdong/30/2013(H_7N_9)	5	42	10	4	38	1
H_9N_2(1)	A/chicken/Guangdong/LY/2010(H_9N_2)	5	49	4	11	40	5
H_9N_2(2)	A/chicken/Guangdong/BL/2010(H_9N_2)	5	44	5	10	36	1
H_9N_2(3)	A/chicken/Guangdong/6/97(H_9N_2)	6	47	5	9	43	2
H_9N_2(4)	A/chicken/Guangdong/56/01(H_9N_2)	7	49	7	5	41	2
H_9N_2(5)	A/chicken/Guangdong/5/97(H_9N_2)	6	45	6	9	41	3

Note: The characteristic genetic fragments were Ⅰ. GGGG, Ⅱ. AAA, Ⅲ. TTTC, Ⅳ. TCTT, Ⅴ. AAG, Ⅵ. ACGG, respectively.

3 Results and analysis

3.1 Microorganism BLSOM classification At present, microbial genomic information has increased significantly, and there is a need for new technical means to conduct a comprehensive analysis. The common GC value analysis method for microbial genome is simple, not suitable for processing large amounts of genomic information, and the results can not reflect the essential characteristics of microbial genetic variation. The non-sequence-alignment self-organizing map (SOM) and the improved BLSOM method, can be used to classify the 1 kb genetic fragment and predict the direction of variation. BLSOM uses the visual classification tool of advanced computer software which can reveal virus host-dependence and codon bias caused by natural selection, to identify high-risk types in millions of microbial gene data and monitor high-risk strains. It is of great significance to biomedicine and preventive veterinary medicine. The new MATLAB software toolbox provides neural network function, and can simulate the human brain to complete BLSOM competitive learning and training, pattern recognition, classification and identification. It is widely applied in engineering, finance, agriculture, environmental protection, education, public security and a variety of scientific studies[11-12]. In this paper, it was used to study the avian influenza virus and achieved initial success, worthy of further study.

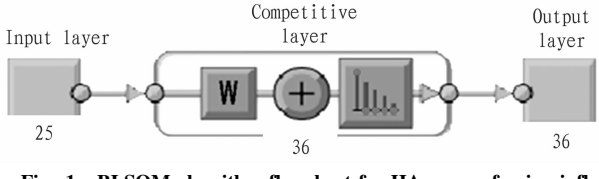

Fig. 1 BLSOM algorithm flowchart for HA genes of avian influenza virus isolated in Guangdong Province

3.2 Avian influenza classification standard and BLSOM classification According to statistics, the current nucleic acid sequences about avian influenza virus in the gene pool have been as many as 730000, and the number for H_1N_1, H_3N_2, H_5N_1 and H_9N_2 is 110000, 83000, 27000 and 14000, respectively. The con-

ventional evolutionary tree and other analysis tools fail to see the whole picture[13]. BLSOM method can handle more than one million gene sequences at the same time, and the analysis results are consistent with the evolutionary tree. At the genetic level and in the oligonucleotide (2 – 4 bp) fragment composition, avian influenza shows significant host dependence, namely the self-organization and classification characteristics based on host, which is the biological basis of BLSOM classification. The influenza virus growth depends on many host factors such as nucleotides, amino acids and tRNA, while avoiding the antiviral mechanism of host such as the action of antibodies, cytotoxic T cells, interferon and RNA interference, thereby forming unique host dependence of gene structure. However, we often can not draw the conclusion of host dependence from the single nucleotide BLSOM, and the tetranucleotide BLSOM classification effect based on host is good. The results from Table 2, Fig. 3, Table 3 show that except HA genes, all 8 genetic fragments can be used for BLSOM analysis.

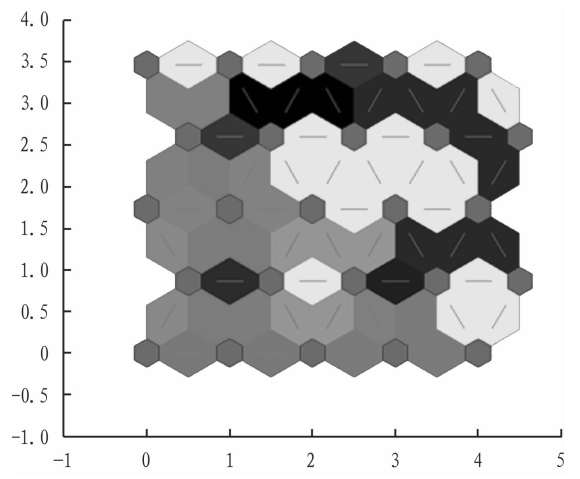

Fig. 2 Distance between adjacent BLSOM neurons for some avian influenza virus strains prevalent in Guangdong

Table 2 BLSOM training steps and clustering results for some avian influenza virus strains prevalent in Guangdong

Number of training steps	Clustering results					
10	36	1	4	36	36	1
30	31	1	12	36	25	6
50	36	1	23	35	24	1
100	13	35	6	5	19	36
200	21	12	1	31	4	36
500	33	9	6	24	25	1
1000	1	12	28	31	9	36

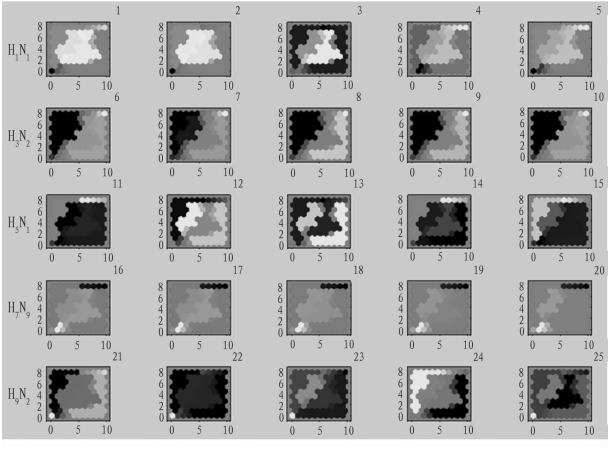

Fig. 3 BLSOM clustering map for some avian influenza virus strains prevalent in Guangdong

Table 3 BLSOM clustering results for some avian influenza virus strains prevalent in Guangdong

Type	Name of the strain
1	$H_1N_1(1)$, $H_1N_1(2)$, $H_1N_1(4)$, $H_1N_1(5)$
2	$H_3N_2(1)$, $H_3N_2(2)$, $H_3N_2(3)$, $H_3N_2(4)$, $H_3N_2(5)$
3	$H_5N_1(1)$, $H_5N_1(2)$, $H_5N_1(3)$, $H_5N_1(4)$, $H_5N_1(5)$
4	$H_7N_9(1)$, $H_7N_9(2)$, $H_7N_9(3)$, $H_7N_9(4)$, $H_7N_9(5)$
5	$H_9N_2(1)$, $H_9N_2(3)$, $H_9N_2(4)$, $H_9N_2(5)$
6	$H_1N_1(3)$, $H_9N_2(2)$

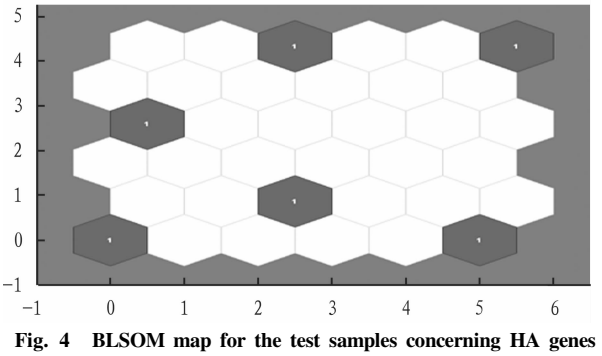

Fig. 4 BLSOM map for the test samples concerning HA genes of avian influenza virus isolated in Guangdong Province

3.3 BLSOM software tool, parameter and short nucleotide fragment selection The natural selection of avian influenza virus gene codon has pressure on all 8 fragments which can be used for BLSOM analysis to obtain genetic characteristics and host preference information. Tetra-BLSOM can clearly divide the human influenza virus from swine influenza virus according to region and color, and the high-risk strain located in the border between two regions due to the variation can be identified. From Fig. 3, we can clearly see that the strain structure of H_3N_2 and H_7N_9 is consistent, and there is strain variation for H_9N_2, H_5N_1 and H_1N_1. In this study, 6 kinds of oligonucleotide fragments are selected for BLSOM analysis, and the main avian influenza strains can be identified. When the number of training steps is more than 100, the effective clustering effect is generated (Table 2). The neurons in the upper right corner are distant (Fig. 2), and the test samples

are evenly distributed (Fig. 4). The methods and softwares that can handle large data sequences simultaneously are for further study. BLSOM is used to analyze the host dependence on oligonucleotide, identify sequence variation direction in huge gene data, and monitor high-risk strains of animals and humans, thus it is an important issue on viral molecular evolution study[9]. In this study, there is large variation in H_5N_1, and BLSOM map is significantly different (Fig. 3). There is a need to conduct further studies on the use of BLSOM as virus warning and traceability tool.

References

[1] SONG JD, ZHU DG, YUAN LP, *et al.* Etiology analysis of avian influenza pathogen from the live poultry market in Jiangmen Region during 2011 -2013[J]. Guangdong Journal of Animal and Veterinary Science,2014, 31(6):6 -9. (in Chinese).

[2] LU QF, CAO JW, FENG XH, *et al.* Etiology analysis of avian influenza pathogen from the live poultry market in Jiangmen Region during 2011 -2013[J]. Guangdong Journal of Animal and Veterinary Science,2014,39 (4):18 -19. (in Chinese).

[3] LU EJ, CHEN YY, LIU JW, *et al.* Surveillance analysis on the occupational population exposed to avian influenza virus and contamination in market environment of Guangzhou city in 2013[J]. Chinese Journal of Pest Control,2014,30(9):980 -984. (in Chinese).

[4] LU JY, LU EJ, LI KB, *et al.* Analysis on avian influenza surveillance among occupational population in poultry environment in Guangzhou from 2011 to 2012[J]. Chinese Journal of Pest Control,2013,29(6):591 -593. (in Chinese).

[5] ZHU BL, HUANG GH, MAI W, *et al.* Surveillance analysis on the high risk population and environment of avian influenza in Zhaoqing, 2011 -2012[J]. Journal Of Tropical Medicine,2014,14(1):115 -117. (in Chinese).

[6] LI GW, YAN ZQ, LIAO CT, *et al.* Phylogenetic analysis of hemagglutinin genes of H_9N_2 subtype avian influenza viruses isolated in Guangdong and Guangxi provinces during 2011 -2012[J]. Chinese Journal of Veterinary Science,2014,34(3):461 -464. (in Chinese).

[7] CHEN B, MA ZC, RAO DP, *et al.* Epidemiological investigation on the cases of human infection With H_7N_9 avian influenza in Shenzhen[J]. The Journal of Medical Theory and Practice,2014,27(21):2924 -2925. (in Chinese).

[8] KONG DF, QIN YM, MEI SJ, *et al.* Epidemiological analysis on 2 cases infected with highly pathogenic human avain influenza in Shenzhen[J]. Chinese Journal of Pest Control,2013,29(12):1390 -1392. (in Chinese).

[9] IWASAKI Y, ABE T, WADA K, *et al.* Prediction of directional changes of Influenza A virus genome sequences with emphasis on pandemic H_1N_1/09 as a model case[J]. DNA Research, 2011, 18(2): 125 -136.

[10] IWASAKI Y, ABE T, WADA Y, *et al.* Novel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains[J]. BMC Infectious Diseases,2013,13:386.

[11] WANG XC, SHI F, YU L, *et al.* Analysis on 43 cases about MATLAB neural network[M]. Beijing:Beihang University Press,2013. (in Chinese).

[12] ZHANG XR, ZHANG YL, LIU LS, *et al.* Zoning by land types based on SOFM network:A case study on transect of eastern Tibetan Plateau [J]. Geographical Research,2013,32(5):839 -847. (in Chinese).

[13] SONG QQ, CHAI ZX, ZHONG JC, *et al.* Codon usage bias and cluster analysis on genes of avian influenza virus[J]. Biotechnology,2014,24(2):48 -53. (in Chinese).

(From page 100)

skill training and improving agricultural labor forces' skills. We can organize agricultural technical personnel to carry out service in rural areas, enhance rural labor skill training and establish entrepreneurship training base, actively mobilize agricultural vocational schools, training institutions, science and technology volunteers and other social forces to carry out the training on agricultural machinery and equipment operating skills and agricultural management knowledge. There is a need to enhance farmers' self-confidence, abandon the traditional conservative ideas, and further stimulate farmers' willingness to transfer land.

References

[1] XU MY. An empirical study on the farmers wishes about transfer of rural land in developed areas [J]. Journal of Nanjing Agricultural University (Social Science Edition),2014,11(6):97 -105. (in Chinese).

[2] QIAN WR, ZHANG ZM. The farmers' desired scale of land management-An analysis based on a survey in the Middle-Lower Reaches of Yangtze River: An empirical study [J]. Problems of Agricultural Economy, 2007, 28(5):28 -34. (in Chinese).

[3] ZHAO XQ, LI HJ. An empiricial analysis on the influencing factors of farmers' land transfer willingness in the western region [J]. Chinese Rural Economy, 2009(8): 70 -78. (in Chinese).

[4] QIAO HQ, CHENG WS, XU B. Farmer's desire characteristics and influence factors of farmland circulation in Hexi corridor [J]. Research of Soil and Water Conservation, 2016(6):209 -213. (in Chinese).