Avens Publishing Group
J Gene Ther
January 2014 Vol.:1, Issue:1

Comparative Analysis of H1N1 Avian Influenza Virus by Multiple Sequence Alignment and Support Vector Machine

Open Access Research Article

Journal of Gene Therapy

Yufei Liu, Libin Zhang* and Yanhong Zhou*

Department of Biomedical Engineering, School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

Address for Correspondence

Libin Zhang, Department of Biomedical Engineering, School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China; E-mail: libinzhang@hust.edu.cn

Yanhong Zhou, Department of Biomedical Engineering, School of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China; E-mail: yhzhou@hust.edu.cn

Submission: 11 September 2013
Accepted: 03 January 2014
Published: 10 January 2014

Reviewed & Approved by: Dr. Yujing Li
Senior Scientist, Emory University School of Medicine, USA

Keywords: Avian influenza virus; Bioinformatics; H1N1; Hemagglutinin; Machine learning; Support vector machine; Sequence alignment

Abstract

The H1N1 is a subtype of avian influenza virus (AIV) that is able to break the host barrier to seriously endanger human health. Investigating the molecular mechanisms of AIV interspecies transmission is important for preventing the influenza epidemics. In this study, we used bioinformatics approaches to identify factors that may cause the avian-to-human transmission in hemagglutinin (HA) sequences of H1N1. First, the multiple sequence alignment analysis reveals 10 signature regions of HA that are highly conserved in intra-species, but largely divergent between interspecies. Then, the avian-to-human transformation was modeled as a binary classification problem in a machine learning (ML) context. A computational prediction model was developed to predict the avian-to-human transmission of H1N1 with advanced ML techniques by characterizing amino acid residues in these signatures regions. The evaluation results suggested that these amino acid residues have a discrimination ability to distinguish H1N1 strains isolated from human to those from avian. The proposed bioinformatic framework would be helpful for further understanding the transmission mechanisms of H1N1 and other AIV viruses.

Introduction

Influenza is a paramount epidemic in the world because of the continuing evolution of virus *via* antigenic drift and genetic shift [1]. The avian influenza virus is infectious for birds, pigs, horses and human and lead to the lesions of avian body or respiratory tract, which does great harm to the breeding of poultry such as chickens, turkeys and ducks etc [2-6]. Therefore, the study of avian influenza virus not only has great significance to the poultry industry but also to human health.

H1N1 is the subtype of influenza A virus (AIV) that can break the host barrier to seriously endanger human health, exemplified by the 2009 swine-origin H1N1 influenza A epidemic [7]. Although the origins and evolutionary of human-isolated H1N1 virus can be easily inferred from the phylogenic analysis, the determinants of cross-species transmission are still not fully understood [8,9]. The H1N1 genome codes six internal proteins (NP, M1, M2, PB1, PB2 and PA), two non-structural proteins (NS1 and NS2), and two coat proteins (HA and NA) [10]. Among these ten proteins, HA (hegagglutinin) has been demonstrated to be particularly important for virus infection against the host, by mediating the attachment of the virus to the host cell surface and the entry of viral RNA into the host cell [11,12]. Therefore, the properties of the HA protein in H1N1 virus are very worthwhile to be studied, which will provide a clue to better understand the infection mechanism of influenza viruses and monitor the interspecies transmission of influenza virus.

In this study, we used bioinformatic approaches to comparatively analyze the HA protein sequences of H1N1 viruses isolated from avian and human hosts, and identified several signature regions in which amino acid segments specifically are conserved in avian- and human-isolated H1N1 viruses. Further statistical analysis revealed that there are different patterns of amino acid content in the HA sequences of H1N1 viruses isolated before and after the 2009 H1N1 pandemic. We also applied machine learning techniques to effectively distinguish human-isolated H1N1 from avian-isolated H1N1.

Methods

Multiple sequence alignment

In bioinformatics, multiple sequence alignment is used for arranging the sequences of DNA, RNA, or protein to identify similar regions that may conclude the important conclusions of function, structure, or evolution of species [13]. Multiple sequence alignment aligned all of the sequences to a unified format for analyzing the functional or structural regions in samples. Multiple sequence alignment is also the necessary steps to construct phylogenetic trees for aiding in identifying evolutionary relationships [14,15]. In this study, the multiple sequence alignment was performed with online alignment software integrated into the NCBI Influenza virus resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html).

Support vector machine

The Support Vector Machine (SVM) is a well-formulated machine learning technology first proposed by Corinna Cortes and Vapnik in 1995 [16]. The SVM mapped the features in the original space into a high-dimensional feature space with a kernel function to perform the classification problems. Due to the advantages of solving small samples, nonlinear and high-dimensional pattern recognition, it has been applied for various classification and prediction problems, including the classification of translation initiation starts [17], protein subcellular localization [18] and protein function [19]. In this study, we performed the SVM algorithm by considering the amino

ISSN: 2381-3326

acid content in signature regions through implementing the 'e1071' package with the default parameters in R programming language (http://cran.r-project.org).

Results

Signature regions identified in HA genes of H1N1 by multiple sequence alignment analysis

All HA sequences of the H1N1 avian influenza virus isolated from human (3377) and avian (127) analyzed in this study were downloaded from the Influenza Virus Resource of NCBI. Examining the host specificity of amino acid (AA) residues is helpful to identify important regions that may play roles in the cross-transmission of H1N1 in the HA sequence. The AA residues in a successive positions were consider to be host specific if they were highly conserved in the same species of the host, but evidently divergent between two species. It should be noted that this definition of host specificity of AA residues can also be applied on other virus proteins and at the structure level. We found that several regions (denoted as signature regions) in HA sequences were host specific (Table 1). At the regions of 489~490 and 144~152, the most frequency of AA residues in avian-isolated H1N1 viruses were "DDE" (94.6%) and "ETTKGVTAA" (93%), respectively. While in human-isolated H1N1 viruses, the corresponding frequency were about 61%. This result indicated that amino acids of this signature region in human-isolated H1N1 viruses evolved more rapidly than those in avian-isolated H1N1 viruses. We also observed that at the region of 9~16, the frequencies of "FCTFTVLK" residues were comparable between H1N1 isolated from avian and human hosts (65% vs 63%). Further statistical analysis revealed that nearly 60% contained all the identified amino acid residues (Figure 1A, red amino acids, allowed for 2 amino acids mismatch) among 3377 H1N1 strains in human. On the other hand, the majority of H1N1 strains in avian (total number, 127) were found to contain all the identified amino acids (Figure 1C, blue amino acids, allowed for 2 amino acids mismatch). Finally, NA protein was selected for same analysis as HA protein. We picked out a total of 904 H1N1 strains from human to perform multiple protein sequence alignment and the amino acid frequencies in 470 sites were counted. As shown in Supplementary Table 1, the result showed that the majority of amino acid frequencies in 470 sites are more than 80% except for sites 241, 248 and 369. Moreover, the number of the sites in which the frequencies of amino acids are more than 85% is 451, which indicated that the multiple sequence alignment analysis of NA protein displayed a single pattern in human-isolated H1N1 viruses.

Table 1: The different frequency of amino acid residues in HA sequences of avian- and human-isolated H1N1 viruses. "AA Residues" denotes the amino acid residues with the highest frequency in the signature regions.

Signature regions	Avian-isolated H1N1 viruses		Human-isolated H1N1 viruses	
	AA Residues	Frequency	AA Residues	Frequency
9-16	FCTFTVLK	65%	LYTFATAN	63%
85-91	DLLLTAN	85%	ESLSTAS	64.6%
97-103	IETSNSE	91.5%	VETSSSD	63.3%
144-152	ETTKGVTAA	93%	DSNKGVTAA	60.9%
154-159	SYSGAS	73.6%	PHAGAK	64.9%
163-170	RNLLWITK	70.5%	KNLIWLVK	66.2%
200-204	PTTSE	72.9%	STSAD	59.4%
274-279	LNKGSD	84.5%	MERNAG	64.3%
284-294	TSDAPVHNCDT	69.8%	ISDTPVHDCNT	61.9%
489-490	DDE	94.6%	DNT	66.4%

Figure 1: Pattern classification of the H1N1 avian influenza virus. (A). Pattern I of the H1N1 avian influenza virus in human (containing all the amino acids marked in red, allow for 2 amino acids mismatch). (B). Pattern II of the H1N1 avian influenza virus in human (containing all the amino acids marked in green, allow for 2 amino acids mismatch). (C).The major pattern of H1N1 avian influenza virus in avian (containing all the amino acids marked in blue, allow for 2 amino acids mismatch).

Support vector machine verification

By drawing the identified signature regions as the characteristics and setting H1N1 strains in human as the negative samples and H1N1 strains in avian as the positive samples, we built up a classification mode between avian-isolated and human-isolated H1N1 with SVM. Although the classification of AIV from avian and human hosts has been recently investigated [20,21], we further performed the classification problems on H1N1 strains isolated in different years. As shown in Table 2, the result revealed that the HA fragments we picked up can be used to identify the differences between isolated- and human-isolated H1N1 viruses, and suggested that SVM is an effective classifier for distinguishing the H1N1 influenza viruses isolated from avian and human hosts. In addition the SVM analysis among different H1N1 species in human was shown in Table 3. On the basis of Table 3, the avian influenza virus in human was obviously different every year. This result not only indicated influenza viruses evolved very fast, but also confirmed that H1N1 influenza virus with different species and same species at different times were both distinguished by HA fragments.

Discussion

In this study, we performed the multiple sequence alignment analysis to identify several signature regions of HA in which the

ISSN: 2381-3326

frequencies of amino acid fragments in human-isolated H1N1 were different from those in avian-isolated H1N1. Using machine learning technique, the amino acids in these signature regions helped us to build prediction model for classifying the avian- and human-isolated H1N1 viruses.

The multiple sequence alignment analysis also helped us to identify three patterns of amino acid content of HA protein in human- and avian-isolated H1N1 viruses. As shown in Figure 1A, statistical analysis demonstrated that nearly 60% contained all the amino acids (marked in red, allow for 2 amino acids mismatch) among 3377 H1N1 strains in human. We described these human H1N1 strains as pattern I H1N1 virus in human (Figure 1A). On the other hand, nearly 40% H1N1 strains in human contained all the amino acids (marked in green, allow for 2 amino acids mismatch) and were described as pattern II H1N1 virus in human (Figure 1B). Also, statistical analysis demonstrated that majority of H1N1 strains in avian (total number, 127) contained all the amino acids (marked in blue, allow for 2 amino acids mismatch) and were described as major H1N1 pattern in avian (Figure 1C). Most of pattern I H1N1 influenza viruses in human appeared recently after the 2009 AIV pandemic. Moreover, more than 90% of human H1N1 viruses appeared in Europe and Asia after 2009 were classified into pattern I. On the contrary, Most of pattern II H1N1 influenza viruses in human appeared before 2009 and the number of viruses decreased very quickly after 2009. Further investigation of these patterns would be helpful for understanding the evolution of H1N1, and forecasting the potential AIV pandemic. Different from HA protein, the multiple sequence alignment analysis of NA protein (Supplementary Table 1) did not display the two patterns (pattern I and II, please discussion) appearing in HA protein analysis, which suggested that NA is more conserved than HA and HA plays a more important role in the interspecies transmission process of avian influenza virus.

Table 2: The SVM results of human H1N1 vs avian H1N1 (2007-2011).

		Strains in avian	
		-1	1
2007 strains in human	-1	235	0
	1	0	127
2008 strains in human	-1	257	0
	1	0	127
2009 strains in human	-1	2057	0
	1	0	127
2010 strains in human	-1	445	0
	1	0	127
2011 strains in human	-1	226	0
	1	0	127

*H1N1 strains in human were set as the negative samples and H1N1 strains in avian were set as the positive samples.

Table 3: The SVM results of H1N1 in human according to the year.

		2007 strains in human		2008 strains in human		2009 strains in human		2010 strains in human	
		-1	1	-1	1	-1	1	-1	1
2008 strains in human	-1	167	18						
	1	68	239						
2009 strains in human	-1	229	4	167	7				
	1	6	127	90	2068				
2010 strains in human	-1	235	0	257	0	2024	337		
	1	0	445	0	445	33	108		
2011 strains in human	-1	235	0	257	0	2030	128	403	43
	1	0	226	0	226	27	98	42	183

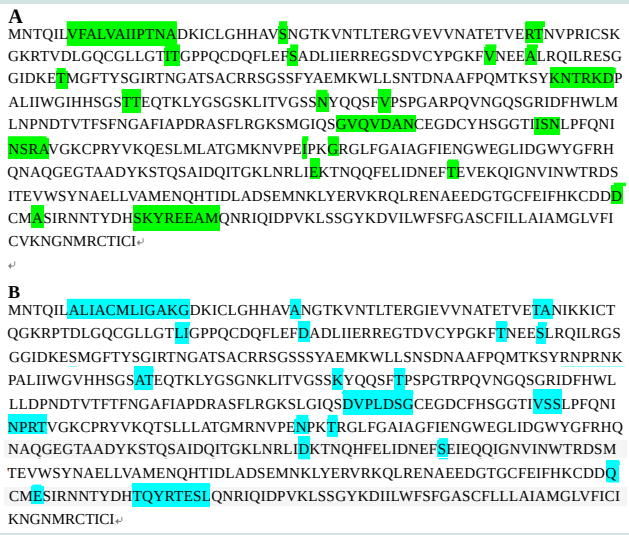

Figure 2: Pattern classification of the H7N9 avian influenza virus. (A). Pattern I of the H7N9 avian influenza virus in avian (containing all the amino acids marked in green, allow for 2 amino acids mismatch). (B). Pattern II of H7N9 avian influenza virus in avian (containing all the amino acids marked in blue, allow for 2 amino acids mismatch).

In the meanwhile, H7N9 sequences deposited in NCBI database including 9 strains isolated from human and 39 strains isolated from poultry were downloaded for multiple sequence alignment analysis. The result showed that the strains from poultry also displayed two patterns (pattern I and pattern II). As shown in Figure 2, pattern I has different conserved amino acid sites from pattern II and the pattern I strains have high sequence similarity with strains from human. Furthermore, the similarity of amino acids between pattern I strains and strains from human added up to 89.8%. Therefore, we speculated that pattern I H7N9 strains from poultry are easier to infect human and need to be monitored. Although the number of collected H7N9 strains is limited, the obtained result can be used as a reference for further study. Collectively, the bioinformatics framework applied in this study will facilitate in-depth knowledge discovery of transmission mechanisms from the protein sequences of H1N1 viruses.

References

1. Subbarao K, Joseph T (2007) Scientific barriers to developing vaccines against avian influenza viruses. Nat Rev Immunol 7: 267-278.
2. Webster RG, Bean WJ, Gorlnan OT, Chambers TM, Kawaoka Y (1992) Evolution and ecology of influenza A viruses. Microbiol Rev 56: 152-179.
3. Ferguson NM, Fraser C, Donnelly CA, Ghani AC, Anderson RM (2004) Public health risk from the Avian H5N1 influenza epidemic. Science 304: 968-969.
4. Zhou N, Senne D, Landgraf JS, Swenson SL, Erickson G, et al. (1999) Genetic reassortment of avian, swine, and human influenza viruses in American pigs. J Virol 73: 8851-8856.
5. Bean WJ, Schell M, Katz J, Kawaoka Y, Naeve C, et al. (1992) Evolution of

ISSN: 2381-3326

the H3 influenza virus hemagglutinin from human and nonhuman hosts. J Virol 66: 1129-1138.

6. Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, et al. (2004) The structure and receptor binding properties of the 1918 iInfluenza Hemagglutinin. Science 303: 1838-1842.
7. Smith GJ, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, et al. (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459: 1122-1125.
8. Lin T, Wang G, Li A, Zhang Q, Wu C, et al. (2009) The hemagglutinin structure of an avian H1N1 influenza A virus, Virology 392: 73-81.
9. Landolt GA, Olsen CW (2007) Up to new tricks-A review of cross-species transmission of inluenza A viruses. Anim Health Res Rev 8: 1-21.
10. Knipe DM, Howley PM (2001) In: Fields Virology (4thedn), Philadelphia: Lippincott Williams & Wilkins Publishers.
11. Blixt O, Head S, Mondala T, Scanlan C, Huflejt ME, et al. (2004) Printed covalent glycan array for ligand profiling of diverse glycan binding proteins. Proc Natl Acad Sci, 101: 17033–17038.
12. Skehel JJ, Wiley DC (2002) Influenza haemagglutinin. Vaccine 20: S51-S54.
13. Mount D (2004) Bioinformatics: sequence and genome analysis (2ndedn), Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.
14. Wang L, Jiang T (1994) On the complexity of multiple sequence alignment. J Comput Biol 1: 337–348.
15. Elias I (2006) Settling the intractability of multiple alignment. J Comput Biol 13: 1323–1339.
16. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20: 273-297.
17. Ma C, Zhou D, Zhou Y (2006) Feature mining and integration for improving the prediction accuracy of translation initiation sites in eukaryotic mRNAs. Grid and Cooperative Computing Workshops 349-356.
18. Wan S, Mak MW, Kung SY (2012) mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines. BMC Bioinformatics 13: 290.
19. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ, et al. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. 31: 3692-3697.
20. Wang J, Kou Z, Duan M, Ma C, Zhou Y (2013) Using amino acid factor scores to predict avian-to-human transmission of avian influenza viruses: a machine learning study. Protein Pept Lett 20:1115-1121.
21. Wang J, Ma C, Kou Z, Zhou YH, Liu HL (2013) Predicting transmission of avian influenza A viruses from avian to human by using informative physicochemical properties. Int J Data Min Bioinform 7:166-179.