Bioinformation Volume 5 open access
www.bioinformation.net Issue 1 Hypothesis

Evolutionary analyses on the HA gene of pandemic H1N1/09: early findings

Yuki Furuse, Akira Suzuki, Hitoshi Oshitani*

Department of Virology, Tohoku University Graduate School of Medicine, 2-1 Seiryou-machi Aoba-ku, Sendai, Japan; Hitoshi Oshitani - E-mail: oshitanih@mail.tains.tohoku.ac.jp *corresponding author

Received March 3, 2010; accepted April 9, 2010; published June 15, 2010

Abstract:
The HA protein is responsible for influenza virus attachment and the subsequent fusion of viral and cellular membranes. Antigenic drift is driven by an accumulation of point mutations in the HA. And, the receptor-binding specificity of HA is responsible for the host range restriction of the virus. In April 2009, large outbreaks of novel H1N1 influenza in human population were reported from North America. The pandemic H1N1 virus originated from swine influenza virus. Evolutionary process of the pandemic virus after its introduction to human population remains to be clarified. We conducted phylogenetic analyses constructing a phylogenetic tree for and calculating site-by-site selective pressures in the HA gene. Phylogenetic tree showed that pandemic viruses were not clustered clearly by their geographical location or isolation time in the phylogenetic tree. The virus has been circulating the globe extensively with multiple introductions into most geographical areas. We found 3 sites positively selected in the HA gene for pandemic H1N1 virus. Among them, position 206 is located in an antigenic site. We did not find significant negative selection on any of the receptor binding sites. The virus has been evolving under unique selective pressure.

Background:

The influenza virus is a common cause of respiratory infection all over the world. The subtype of influenza A virus is determined by the antigenicity of two surface glycoproteins, hemaglutinin (HA) and neuraminidase (NA). The subtypes circulating in the human population are H1N1 and H3N2. The HA protein is responsible for virus attachment and the subsequent fusion of viral and cellular membranes **[1, 2].** Besides, there may also be a structural role for the HA in budding and particle formation **[1, 2]**. The HA is a rod-shaped molecule with parts inserted into the viral membrane and projecting as spike away from the viral surface. It is synthesized as a single polypeptide chain (HA0), and the HA0 precursor cleaves into HA1 and HA2 subunits. A long fibrous stem is comprised of HA2 part and globular head is comprised of HA1 part. Receptor-binding site lies within the globular head of the molecule [1]. The receptor-binding specificity of HA, which is determined by the nature of the amino acids that form the receptor-binding pocked, is responsible for the host range restriction of the virus **[3].**

Antigenic drift is driven by an accumulation of point mutations in the HA or NA proteins, leading minor and gradual antigenic changes **[4].** Drift variants emerge due to a positive selection of spontaneous mutants that can escape from existing host’s antibodies. The HA protein is the major antigenic component of the virus. For H1 viruses, the antigenic sites are designated Ca1, Ca2, Cb, Sa, and Sb **[5].** The antigenic sites surround the sialic acid receptor-binding site. Single point mutations in one HA antigenic site can be sufficient for antigenic variation. Operating natural selection theory on HA, positively selected sites could be found **[6].** They may be useful for identifying the epitopes involved in the elimination of viruses from infected patients.

In April 2009, large outbreaks of novel H1N1 influenza in human population were reported from Mexico and the U.S. The H1N1 virus has spread worldwide and the World Health Organization (WHO) raised its official pandemic alert level phase 6 on its six-phased scale on June 11, 2009. According to the WHO, more than 208 countries and overseas territories or communities have reported laboratory-confirmed cases of pandemic influenza, including at least 12,799 deaths as of January 2010. The pandemic H1N1 virus originated from swine influenza virus **[7]**. Reassortment between swine influenza viruses in two distinct lineages (‘triple-reassortant’ and ‘Eurasian avian-like swine’) led to generation of the pandemic virus. The source of a swine lineage (triple-reassortant) itself comprised genes derived from avian, human and swine lineages.

Mechanisms the pandemic virus broke the host range restriction is a major topic to be studied. Besides, evolutionary process of the pandemic virus after its introduction to human population remains to be clarified. For Spanish Flu in 1918 to 1919, it is believed that the virus evolved becoming much more lethal after its emergence, leading the second wave with high impact **[8].** Thus, genetic analysis on current pandemic virus should provide critical insights into evolutionary course and evolutionary mechanisms of the influenza virus. Here, we conducted phylogenetic analyses constructing a phylogenetic tree for and calculating site-by-site selective pressures in the HA gene.

Methodology:

Sequence data

All data were obtained from the influenza sequence database [Influenza Virus Resource on: http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html, accessed on January 14, 2010]. All sequencing data for the strains with a full-length HA gene of pandemic H1N1 human influenza A virus were included. Sequences derived from laboratory strains were excluded and a total of 1,332 sequences were obtained. Sequences containing minor insertions, minor deletions or untranslatable codons were excluded. A multiple alignment of the nucleotide sequences was constructed using ClustalW.

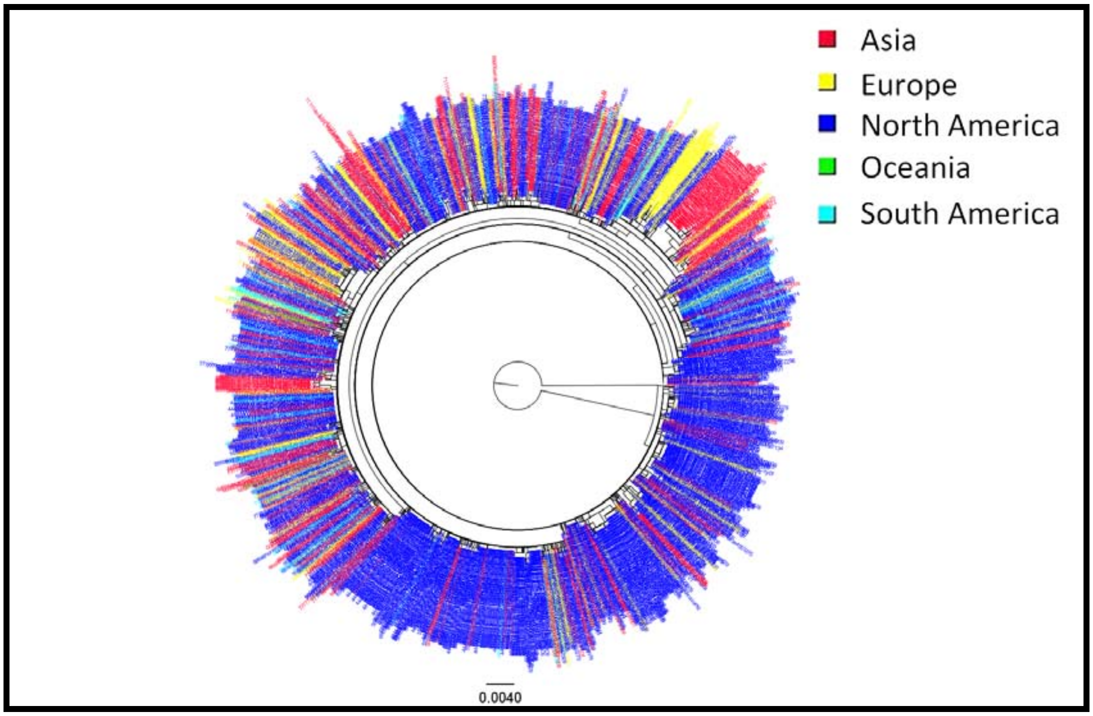

Figure 1: Phylogenetic tree for the HA gene. Phylogenetic trees constructed using RAxML. Scale bar shows evolutionary distance inferred by RAxML algorithm. Trees are shaded in colors according to geographical information.

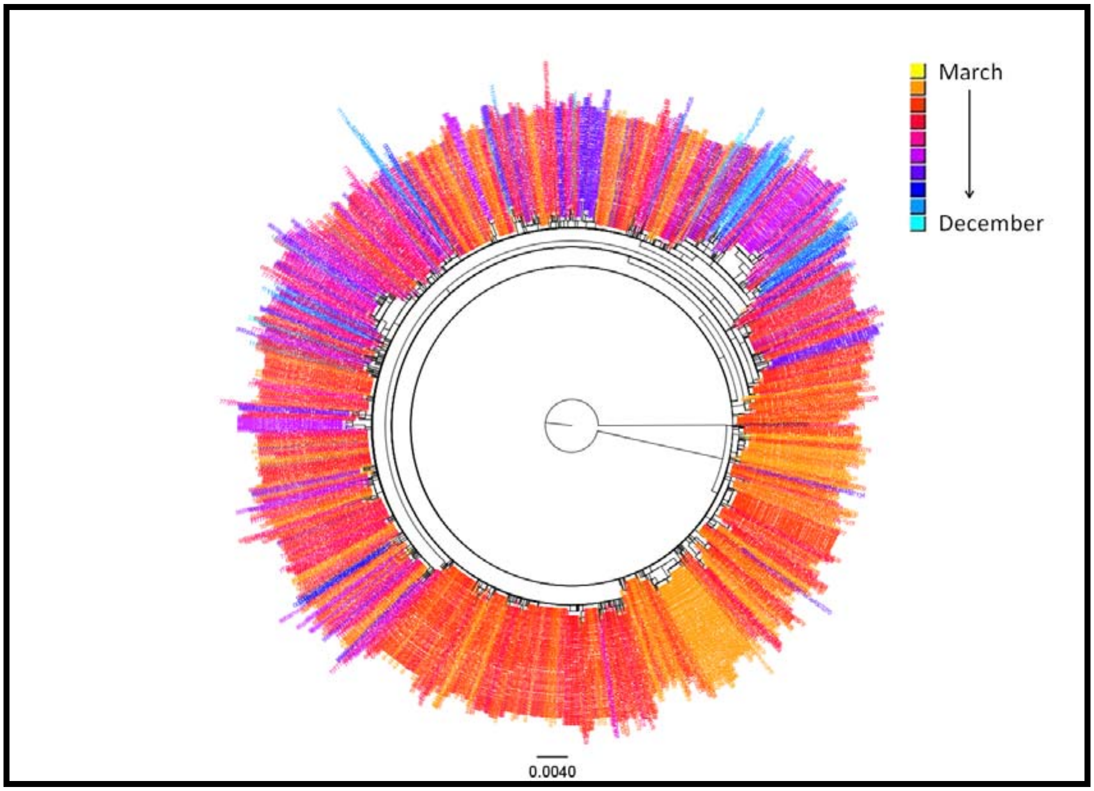

Figure 2: Phylogenetic tree for the HA gene. Phylogenetic trees constructed using RAxML. Scale bar shows evolutionary distance inferred by RAxML algorithm. Trees are shaded in colors according to temporal information.

Phylogenetic tree

A phylogenetic tree was inferred by RAxML [9]. The sequences data only for the coding region were used. The basic sequential algorithm of RAxML is described elsewhere [10]. RAxML is one of the fastest and most accurate sequential phylogeny programs. In this method, a rapid bootstrap search is combined with a rapid maximum likelihood search on the original alignment. The tree was constructed using Web-servers, RAxML BlackBox (mirror at CIPRES): "http://8ball.sdsc.edu:8889/cipres-web/Home.do". The tree is color-coded according to geographical or temporal information using FigTree (ver.1.3.1).

Selective pressure

Identical nucleotide sequences in the dataset were removed before calculating selective pressures. A phylogenetic tree was constructed using the maximum-likelihood method implemented in PhyML-aLRT **[11]** with the GTR model (four rate categories, all parameters estimated from the data). Positive selection sites were detected using fixed-effects likelihood (FEL) method, which is based on maximum-likelihood estimates. The relative rates of non-synonymous and synonymous substitutions (dN/dS) in each codon were compared. Sites where dN/dS >1 and dN/dS <1 were inferred as positively and negatively selected, respectively. The details of the method are described elsewhere **[12].** H1 numbering was counted from initiation codon to termination codon based on A/South Carolina/1/18 (H1N1). H3 numbering was counted by HA1 and HA2 regions seperately **[13].**

Discussion:

We constructed a phylogenetic tree for the HA gene of pandemic influenza viruses collected worldwide in 2009. The tree is shaded in colors by geographical or temporal information **(Figures 1 and 2).** Pandemic viruses were not clustered clearly by their geographical location or isolation time in the phylogenetic tree. It should be noted that North American strains collected in April and May look to be clustering in the lower right part of the tree because intensive surveillance was done in countries in North America at that time. Basically, the tree shows scattered and mixed strains from different countries at different time periods. The results suggest that multiple importations from outside occurred frequently. Nelson et al. showed genetic multiple lineages of pandemic viruses and co-circulation of these lineages **[14].** The virus has been circulating the globe extensively with multiple introductions into most geographical areas **[14, 15].** In future, it is intriguing to know whether a specific strain (and its progeny) will predominate globally or not. We should keep monitoring the spatial movements and evolution of the pandemic virus.

Then, we calculated site-by-site selective pressures. We found 3 sites positively selected in the HA gene for pandemic H1N1 virus: 1 in signal domain, 2 in HA1 and 0 in HA2 **(Table 1 in supplementary material).** Reason for positive selection at position 8 (H1 numbering) is unknown. The site is located in signal domain and is not involved in the structure of HA. Position 206 (H3 numbering) are located in an antigenic site. Mutation on the site could be due to positive selection imposed by host's immune response. It is surprising that selective pressure has been already imposed on pandemic virus, although only a year has passed since its introduction into human population. The virus must be circulating in human population very rapidly and extensively, evolving under positive selection by host's immune response. It is believed that most people do not possess immunity to the pandemic virus, except for cross-reactive immunity observed in some elderly people **[16, 17].** Therefore, re-infection in persons who had been infected once may lead antigenic drift. Actually, re-infection with the pandemic virus was reported **[18].** Otherwise, mutation on the site can be advantageous for the virus such as increasing receptor binding affinity and/or replication ability.

Other sites in antigenic sites are not under significant positive selection although selective pressures on some of them are higher than 1 **(Table in supplementary material).** Non-synonymous mutations in at least one strain were observed in 49.0% of antigenic sites (25/51). In contrast, non-synonymous mutations in at least one strain were observed in only 33.6% of all codons in the HA gene (190/566). We should keep monitoring whether mutation on the antigenic sites will occur leading antigenic drift. Further, no sites which are involved in receptor binding specificity are under significant positive selection (table). Pandemic virus is considered to recognize human-type receptors **[19].** Receptor binding sites do not have to be mutated to increase adaptation in humans. However, we also did not find significant negative selection on any of the receptor binding sites. We should keep monitoring whether mutation on the receptor binding sites will occur, leading change in receptor binding specificity.

We identified a site at position 264 (H3 numbering) as positively selected. However, its specific role is unknown. Experimental studies are needed to clarify the significance. The site may possess unknown characteristics such as a part of unidentified antigenic sites. Besides, mutation at position 225 was reported to be associated with fatal cases **[20]**. We did not find positive selection at the site. Yet, the site possess various residues (D, E, G, and N) and selective pressure on the site is higher than 1 (1.7, P value = 0.45).

Conclusion

This study precisely intrigue to investigate the evolutionary way of the HA gene of pandemic H1N1 influenza virus in the first year of its emergence. Our analyses revealed that the virus has been circulating extensively and globally, evolving under unique selective pressure. Furthermore, the study predicts sites which may be involved in significant roles.

Acknowledgement

YF is a recipient of a scholarship through the Honjo International Scholarship Foundation. This work was supported by the Health and Labor Science Research Grants from the Ministry of Health, Labor, and Welfare, Japan.

References:

[1] JJ Skehel *et al. Annu Rev Biochem*. **69**: 531 (2000) [PMID: 10966468]
[2] LJ Earp *et al. Curr Top Microbiol Immunol.* **285**: 25 (2005) [PMID: 15609500]
[3] GN Rogers *et al. Virology* **127**: 361 (1983) [PMID: 6868370]
[4] RG Webster *et al. Microbiological Reviews* **56**: 152 (1992) [PMID: 1579108]
[5] W Gerhard *et al. Nature* **290**: 713 (1981) [PMID: 6163993]
[6] Y Suzuki, *Gene* **427**: 111 (2008) [PMID: 18848975]
[7] GJD Smith *et al. Nature* **459**: 1122 (2009) [PMID: 19516283]
[8] AH Reid *et al. Emerg Infect Dis*. **9**: 1249 (2003) [PMID: 14609459]
[9] A Stamatakis *et al. Systematic Biology* **57**: 758 (2008) [PMID: 18853362]
[10] J Felsenstein, *J Mol. Evol*. **17**: 368 (1981) [PMID: 7288891]
[11] M Anisimova *et al. Syst Biol*. **55**: 539 (2006) [PMID: 16785212]
[12] www.hyphy.org/pubs/hyphybook2007.pdf
[13] E Nobusawa *et al. Virology* **182**: 475 (1991) [PMID: 2024485]
[14] M Nelson *et al. PLoS Curr Influenza.* RRN1126 (2009)
[15] A Rambaut *et al. PLoS Curr Influenza*. RRN1003 (2009) [PMID: 20025195]
[16] Y Itoh *et al. Nature.* **460**: 1021 (2009) [PMID: 19672242]
[17] RJ Garten *et al. Science* **325**: 197 (2009) [PMID: 19465683]
[18] http://www.foxnews.com/story/0,2933,576966,00.html
[19] G Neumann *et al. Nature* **459**: 931 (2009) [PMID: 19525932]
[20] http://www.who.int/csr/disease/swineflu/notes/briefing_20091120/en/index.html

Edited by P. Kangueane
Citation: Furuse *et al.* Bioinformation 5(1): 7-10 (2010)

Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 1 **Hypothesis**

Supplementary material :

Table 1 : Selective pressure on each codon

Comment	Position	dN/dS	P value	Comment	Position	dN/dS	P value
	(H3 [H1] numbering)						
Sites selected positively	- [8]	inf	**0.0063**		192 [206]	*0*	***0.033***
	206 [220]	inf	**0.00026**		193 [207]	*0.47*	*0.56*
	264 [278]	inf	**0.028**		194 [208]	1.1	0.93
Receptor binding specificity	98 [108]	*0*	*0.14*		195 [209]	*0*	*1*
	136 [150]	*0*	*0.16*		196 [210]	*0*	*0.23*
	153 [167]	*0*	*1*		197 [211]	*0*	*1*
	183 [197]	*0*	*1*		192 [206]	*0*	***0.033***
	190 [204]	*0.18*	*0.14*				
	194 [208]	1.1	0.93	Ca1	169 [183]	inf	0.26
	225 [239]	1.7	0.45		170 [184]	*0*	*1*
	226 [240]	inf	0.078		171 [185]	*0*	*1*
	228 [242]	*0*	*0.14*		172 [186]	inf	0.48
Antigenicity					173 [187]	*0*	*1*
Sa	128 [141]	*0*	*0.15*		206 [220]	inf	**0.00026**
	129 [142]	*0*	*1*		207 [221]	*0*	*1*
	156 [170]	*0*	*0.084*		208 [222]	inf	0.23
	157 [171]	inf	0.48		238 [252]	*0.53*	*0.45*
	158 [172]	*0.088*	***0.012***		239 [253]	*0*	*0.056*
	159 [173]	*0.098*	***0.032***		240 [254]	*0*	*1*
	160 [174]	*0*	*1*	Ca2	140 [154]	inf	0.37
	161 [175]	*0*	*0.14*		141 [155]	*0*	*0.092*
	162 [176]	*0*	*1*		142 [156]	*0*	*1*
	163 [177]	inf	0.35		143 [157]	inf	0.44
	164 [178]	*0*	***0.042***		144 [158]	*0.37*	*0.5*
	165 [179]	*0.93*	*0.95*		145 [159]	*0*	*1*
	166 [180]	*0*	*1*		224 [238]	inf	0.38
	167 [181]	*0*	*1*		225 [239]	1.7	0.45
Sb	187 [201]	*0*	***0.028***	Cb	78 [86]	*0*	*1*
	188 [202]	inf	0.15		79 [87]	*0.55*	*0.68*
	189 [203]	*0*	*1*		80 [88]	inf	0.18
	190 [204]	*0.18*	*0.14*		81 [89]	*0*	*0.11*
	191 [205]	*0.47*	*0.6*		82 [90]	inf	0.17
					- [91]	inf	0.3
					83 [92]	inf	0.36

Significance of the FEL test for positive selection levels is given as P values, and italic values indicate P-values for negative selection. Bold values are those deemed to indicate significantly positive or negative selection ($P < 0.05$). Positions 8 and 91 (H1 numbering) does not have H3 numbering. "inf" means infinity as denominator is 0.