Published in final edited form as:
J Mol Biol. 2012 September 7; 422(1): 145–155. doi:10.1016/j.jmb.2012.05.011.

Identifying antigenicity associated sites in highly pathogenic H5N1 influenza virus hemagglutinin by using sparse learning

Zhipeng Cai[1,†], **Mariette F. Ducatez**[2,†,&], **Jialiang Yang**[1], **Tong Zhang**[3], **Li-Ping Long**[1], **Adrianus C. Boon**[2,#], **Richard J. Webby**[2], and **Xiu-Feng Wan**[1,*]

[1]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, 240 Wise Center Drive, P.O. Box 6100, Mississippi State, MS 39762, USA

[2]Department of Infectious Diseases, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA

[3]Department of Statistics and Biostatistics, Rutgers University, 501 Hill Center, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA

Abstract

Since the isolation of A/goose/Guangdong/1/1996 (H5N1) in farmed geese in southern China, highly pathogenic H5N1 avian influenza viruses have posed a continuous threat to both public and animal health. The non-synonymous mutation of the H5 hemagglutinin gene has resulted in antigenic drift, leading to difficulties in both clinical diagnosis and vaccine strain selection. Characterizing H5N1's antigenic profiles would help resolve these problems. In this study, a novel sparse learning method was developed to identify antigenicity associated sites in influenza A viruses on the basis of immunologic datasets (i.e., from hemagglutination inhibition and microneutralization assays) and HA protein sequences. Twenty-one potential antigenicity associated sites were identified. A total of seventeen H5N1 mutants were used to validate the effects of eleven of these predicted sites on H5N1's antigenicity, including seven newly identified sites not located in reported antibody binding sites. The experimental data confirmed that mutations of these tested sites lead to changes in viral antigenicity, validating our method.

INTRODUCTION

Since the isolation of A/goose/Guangdong/1/1996 (H5N1) in farmed geese in southern China [1; 2], H5N1 highly pathogenic avian influenza viruses (HPAIV) have spread to more than 30 countries throughout Asia, Europe, and Africa. This virus has caused more than 600 laboratory-documented cases in humans (with a fatality rate of approximately 60%) and millions of deaths in birds, and these numbers are still increasing (http://www.who.int/influenza/human_animal_interface/avian_influenza/archive/en/index.html).

[*]Correspondence: Dr. Xiu-Feng Wan, Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762. wan@cvm.msstate.edu or wanhenry@yahoo.com. Tel: +1(662)325-3559; Fax: +1(662)325-3884.

[†]Equally contributed to this study

[#]Current affiliation: Washington University School of Medicine, Department of Internal Medicine, Division of Infectious Diseases, 660 South Euclid Avenue, St. Louis, MO 63110, USA

[&]Current affiliation: INRA UMR 1255, Interactions hôtes pathogènes, ENVT, 23 chemin des Capelles, Toulouse, France

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Over the past decades, H5N1 viruses have undergone considerable evolution, including both frequent reassortment events, leading to the emergence of novel genotypes, and mutations in the surface glycoproteins hemagglutinin (HA) and neuraminidase (NA), resulting in antigenic drifts. More than 40 genotypes of H5N1 viruses were identified between 1996 and 2006 [3; 4; 5], and novel genotypes are still emerging [6; 7]. The HA of H5N1 virus has been classified into 10 genetic clades (namely clade 0 to clade 9) and many more subclades [8; 9] with different antigenic properties [10]. The cross reactions between different clusters can be very limited; for instance, the viruses from clades 1, 2.1, 2.2, and 2.3 can be separated into distinct antigenic clusters [6]. On the basis of hemagglutination inhibition (HI) assay results from a panel of 17 monoclonal antibodies, H5N1 viruses were separated into 4 antigenic groups: group A (clades 1, 2.1, 2.4, 8), group B (clades 1, 2.1, 4, 5, 7, 9), group C (clades 2.1, 2.2, 2.3), and group D (clades 2.3, 5) [11].

Understanding the mutation patterns on the antigenicity associated residues that lead to H5N1 HPAIVs' antigenic drifts will facilitate not only the detection of H5N1 antigenic variants but also the selection of influenza strains for subsequent vaccine design, including prepandemic vaccine preparation. Identifying influenza's antigenicity associated sites is not trivial, and the commonly used methods can be grouped into 3 categories: (1) escaped mutant strategy [12; 13; 14; 15; 16]; (2) genetic polymorphism strategy [11; 17]; and (3) positive selection strategy [18; 19]. Although these strategies have led to many advances in the field, they each have drawbacks, which ultimately led us to search for a new method.

In this study, we developed a novel sparse learning algorithm to identify antigenicity associated sites. Our method is different from those just described because it derives the antigenicity associated sites by correlating the changes in the antigenic distance of immunologic data (as measured by antigenic cartography) with the genetic profiles represented by mutations in HA's protein sequence (FIG. 1). This method was applied to H5N1 HPAIVs and effectively identified antigenicity associated sites in influenza viruses.

RESULTS

Relationship between antigenic clusters and genetic clades

Antigenic cartography based on the HI assay dataset showed that 23 of the 27 H5N1 HPAIVs used in this study (TABLE 1) can be loosely grouped into at least 3 antigenic clusters (FIG. 2A). Most groupings are consistent with those of previous reports based on monoclonal antibody studies in which 4 antigenic groups were assigned: group A (clades 1, 2.1, 2.4, 8), group B (clades 1, 2.1, 4, 5, 7, 9), group C (clades 2.1, 2.2, 2.3), and group D (clades 2.3, 5) [11]. However, many clade 1 viruses were surprisingly grouped with clade 2.2 viruses (FIG. 2A).

The antigenic cartography of the H5 HA showed that the interclade antigenic relationships gleaned from our HI and MN assay results were similar. In our study, 02-CHN (clade 2.1) is grouped with 04-VNM (clade 1) and 06-CHN (clade 4); 05-SAU (clade 2.2) is grouped with 04-CHN (clade 2.3.1) and 06-HKG (clade 2.3.4); and 97-HKG (clade 0) is grouped with most clade 2.2, clade 1, and clade 2.3.4 viruses (FIG. 2B). These results were consistent to those previously described [20].

To correlate the genetic data with the antigenic data, we made phylogenetic trees of the selected 27 viruses (FIG. 2C). Our results demonstrated that the antigenic properties of viruses in the same genetic cluster are not necessarily the same. For example, in both HI assay–based and MN assay–based cartographies, 03-HKG (clade 1) and 05-MNG (clade 2.2) are outliers in the antigenic cartography, while both viruses are genetically closer to being associated with other clade 1 or 2.2 viruses in the phylogenetic tree, respectively (FIG. 2C).

Antigenicity associated sites identified by using the sparse learning method

By applying our novel sparse learning method, we identified 21 antigenicity associated sites having bootstrap values greater than 60% in both HI and MN datasets (TABLE 2), including eight sites not located in reported H3 antibody binding sites A-E (at amino acid positions 94, 120, 124, 162, 227, 252, 263, and 282, H5 numbering). The reported antibody binding sites were annotated based on previous studies [21; 22], and the position correlations between H1, H3, and H5 were based on the alignments of their protein structures [23; 24; 25]. The genetic polymorphisms of these positions are shown in TABLE 3. The receptor-binding site 124 is predicted to be an antigenicity associated site as are thirteen positions that correspond to reported antibody binding sites (FIG. 3). The remaining seven predicted sites do not correspond to reported antibody binding sites.

Experimental validation of the identified antigenicity associated sites

To validate the effects of the predicted antigenicity associated sites on influenza antigenicity, we used 3 rg H5N1 viruses (04-VNM, 05-MNG, and 07-HKG), 13 mutants generated by using site-directed mutagenesis and reverse genetics, and four in-house available single mutants previously generated by mixing murine monoclonal antibodies with 04-VNM. The mutated sites in these mutants include 5 of 13 residues (at amino acid positions 45, 83, 129, 140, and 141, H5 numbering, TABLE 4) corresponding to the reported antibody binding site A to E in H3N2 influenza A virus, and seven newly identified residues (at amino acid positions 94, 124, 162, 227, 252, 263, and 282, H5 numbering, TABLE 4) not corresponding to known antibody binding sites in H3N2 influenza A virus. Besides the 11 predicted sites from sparse learning, these 12 mutated sites included residue 129 (H5 numbering) not shown in the results from sparse learning. The site 129 had a bootstrap value of 95 in HI assay data and 55 in MN assay data (data not shown).

Our results in HI assay showed that the antigenic changes of all 17 mutants but 07-HKG-N45D, 04-VNM-D94N, 04-VNM-S124D, and 07-HKG-K162R had at least 1 unit change, which corresponds to a 2-fold change of the HI titer from the corresponding parent strain in the antigenic cartography (FIG. 4).

The results from MN experiments are consistent with those from HI experiments: all 10 mutants tested in MN experiments, except for 07-HKG-N45D, had at least 1 unit antigenic distance from the associated parent strain in the antigenic cartography (FIG. S1 and TABLE S1–3).

DISCUSSION

In this study we developed a novel computational method using a sparse learning algorithm, LASSO, to identify antigenicity associated sites in influenza viruses. Moreover, we experimentally confirmed the effectiveness of our new method: 13 of the 17 mutants we tested had at least one antigenicity unit change, which corresponds to a 2-fold change of the HI titer from their parent strain in the antigenic cartography.

The sparse learning algorithm proposed in the present study identified 21 antigenicity associated sites using HI and MN data, and some of these predicted sites are consistent with those reported from other studies (TABLE 2): for instance, residues 86 and 124 were reported in [11]; residues 140, 141, 156, and 162 in [26]; residues 124 and 263 in [27]; residues 86 and 154 in [28].

Current knowledge about antibody binding sites in influenza HAs mainly concerns H1 (site Sa, Sb, Ca1, Ca2, Cb)[22] and H3 (site A–E)[21; 24] subtypes. Although the structure of H5N1 HA has been resolved[25], the correlation between antigenic binding sites of H5N1, H1N1, and H3N2 HAs remains unclear. Nine of the 21 antigenicity associated sites derived from our studies are located in the antibody binding sites A, B or D characterized for H3N2 influenza A virus, and four in antibody binding sites C and E. Our results suggested that, similar to H3N2 seasonal influenza A virus, antigenic drift events in H5N1 favor more the residues in sites A, B or D than those in C and E [29].

Although both HI assay data and MN assay data were used to predict the antigenicity associated sites, four of the predicted positions appear in only one type of assay’s dataset (data not shown). Such a discrepancy is very likely to be generated from the location shift of some strains between the antigenic cartographies based on the HI assay data and those based on the MN assay data. For instance, the clade 2.1 virus 02-CHN is grouped with the clade 1 virus 04-VNM and the clade 4 virus 06-CHN in the MN assay dataset but not in the HI assay dataset. Upon examining the virus’ sequences, 02-CHN has an S residue in position 129, yet both 04-VNM and 06-CHN have an L in this position. Because these three viruses have a very close antigenic distance according to MN assay data, this position will not be recognized strongly as an antigenicity associated site (i.e., one with a low bootstrap value) when using this dataset; however, because the viruses have a greater antigenic distance according to HI assay data, the position will have a high enough bootstrap value to be recognized as an antigenicity associated site (TABLE 2). Such a discrepancy could potentially be solved if more viruses from clade 2.1 were used in HI and MN assays. Nevertheless, our study suggests that antigenicity associated site predictions based on both HI and MN datasets are more robust than those based on a single type of dataset.

Our proposed sparse learning method identifies antigenicity associated sites directly from immunologic datasets. This differs from the 3 most commonly used methods: (1) escape mutant strategy; (2) polymorphism strategy; and (3) positive selection strategy. Among these methods, the escape mutant strategy may be the most straightforward but is very labor intensive and limited by the availability of monoclonal antibody. Although less reliant on monoclonal antibodies than the escaped mutant strategy, the information derived from using the genetic polymorphism method is very likely to be incomplete because the antibody binding information may vary among HA subtypes. Additionally, this strategy is greatly affected by the clustering results from the phylogenetic analysis or antigenic clustering that must be performed as its first step; thus, it will be a guessing process after identifying genetic polymorphisms among these sequences. Finally, the positive selection strategy makes predictions derived directly from the sequence and could result in false-positive predictions, especially if the positively selected mutations are generated by selective pressure other than herd immunity, as they likely are in avian influenza viruses including H5N1 HPAIV [30].

To reduce the false-positive rate, most studies using the polymorphism strategy and positive selection strategy limit the sites examined to the ones corresponding to the reported antibody binding sites in H3 [11; 17]. The method proposed in this study can avoid the limitations of the 3 common strategies just mentioned, making it be useful for identifying antigenicity associated sites when the immunologic data and the sequences for the antigens are available. One potential pitfall of this method is that it requires a relatively large training dataset. However, this issue does not create a major bottleneck for influenza virus because HI and MN assay datasets are easily generated. To reduce false-positive results, the antigen and antisera should be carefully selected to avoid sampling biases; in general, multiple antigens from the same genetic or antigenic groups should be chosen. Using multiple types of immunologic datasets (e.g., HI- and MN-assay datasets) can minimize the errors from these sampling biases.

To overcome the potential pitfall of machine learning on the H5N1 study, we carefully selected 23 antigens covering the majorly circulating strains. Four predicted ancestral strains were also included. If multiple strains were selected from the same genetic clusters, genetically diverse strains were attempted to be included in the training data. In addition, both HI and MN data were included in this study. Experimental validation on novel binding sites demonstrated the sparse learning method is effective. To further evaluate our method, we applied this sparse learning method on another independent HI dataset by Wu et al. (2008)[11], which includes 41 antigens and 17 monoclonal antibodies. All of these 8 newly identified sites, and eight of 13 sites in reported antibody binding sites (except five conserved sites 45, 133, 154, 175, and 210, H5 numbering, in Wu data), were also identified (data not shown). In conclusion, our new method can be used to successfully identify antigenicity associated sites in influenza virus HAs and it should contribute to solve some of the difficulties caused by the virus antigenic drift in both clinical diagnosis and vaccine strain selection.

MATERIALS AND METHODS

H5N1 viruses and sera

We used 27 H5N1 strains for the present study (TABLE 1). Virus names were abbreviated as follows: year of first isolation–code of country where the virus was first isolated. The mutants were generated by using reverse genetics of HA, NA, and 6 internal genes from A/ Puerto Rico/8/1934 as described previously [31]. The HA connecting peptide (site for the proteolytic cleavage of the HA enabling the fusion of the protein to the endosome and the virus entry into the host cell cytoplasm[32]) was modified to match that of low-pathogenicity viruses so that the modified strains could be used in biosafety level (BSL) 2+ laboratories; the H5N1 HPAIVs were manipulated in BSL3+ laboratories. The ancestral strains A, B, C, and D have been described [20]. The reverse genetics (rg) 6+2 BSL2 strain 07-HKG was used to generate 07-HKG-N45D and 07-HKG-K162R; rg 6+2 BSL2 strain 05-MNG to generate 05-MNG-I83A, 05-MNG-N94D, 05-MNG-E227D, 05-MNG-N252Y, 05-MNG-T263A, and 05-MNG-I282M; and rg 6+2 BSL2 strain 04-VNM to generate the remaining 11 mutants. The ferret antisera against each of these parent wild-type or 6+2 reassortant viruses were generated as described previously [33].

Sparse learning algorithm

Sparse learning is a machine learning technique that can be used to find a small number of important genetic positions in HA's protein sequence such that genetic changes at these positions lead to significant changes of antigenic properties of the virus. In our approach, the genetic change between two viruses is encoded as a vector x; where each element i of vector x corresponds to a position in the HA1 protein sequence (322 amino acids), and its value x_i indicates the mutation strength at position i measured by a scoring function (e.g. the number of amino acid differences among the compared sequences). The importance of the positions can be measured by a vector w, which we refer to as "weight vector" using standard terminology in machine learning. Each element w_i of w is a numerical value that indicates the importance of position i. For every pair of viruses, we denote by y the change of the antigenic profile, which is the distance between the two viruses measured by antigenic cartography [34; 35; 36] based on a serological dataset such as HI or microneutralisation (MN) data.

Mathematically, we can correlate the antigenic profile change and genetic profile change by modeling the distance y between two viruses (antigenic profile change) as w (position importance) and x (genetic profile change):

$$y \approx w \cdot x \quad (1)$$

Here

$$w \cdot x = \sum_i w_i x_i \quad (2)$$

Our goal in sparse learning is to find a vector w with a small number of nonzero coefficients (which is also referred to as features) w_i that indicate the important genetic positions highly correlated to the antigenic property. This paper employs the LASSO (Least Absolute Shrinkage and Selection Operator) method for sparse learning, which is effective for selecting a small to moderate number of features [37]. The LASSO method finds w by solving the following optimization problem:

$$\widehat{w} = \arg\min_{\mathrm{w} \in \mathrm{R}^{\mathrm{d}}} \sum_{j=1}^{m} (y_j - w \cdot x_j)^2 \quad \text{subject to } \|\mathrm{w}\|_1 \le L \quad (3)$$

where m is the number of virus pairs; j denotes the j-th pair; d is the number of sequence positions; and

$$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_i| \quad (4)$$

Here L is related to the number of predicted antigenicity associated sites. The size of L is tuned by cross-validation and bootstrapping. The higher the absolute value of w_i is, the more impact the residue at position i will have on the antigenicity. It is known that the LASSO formulation leads to a sparse solution vector w corresponding to the most important genetic positions in the HA1 sequence.

Bootstrapping methods

To minimize the likelihood of overfitting and false-positive rates, we used the entire dataset to measure the antigenic distance matrix and then randomly selected 70% of the data entries from this distance matrix. A total of 100 runs were performed, and the detection rates on each site being one of the top 25 sites from each run were used as the confidence level for this site being the antigenicity associated site. All the newly identified antigenicity associated sites not located at reported antibody binding sites will be further validated through experiments.

Antigenic cartography construction

The HI or MN assay data were normalized by using a reference value between that of a reference antigen and reference antiserum as described elsewhere [20]. The antigenic cartography was constructed by using matrix completion-multidimensional scaling described earlier in [34; 35; 36]. A low-rank matrix completion procedure helped to remove noise that is present in HI and MN experiments. Multi-dimensional scaling projected the antigens into a two-dimensional space for visualization.

HI and MN assays

Ferret sera were treated with receptor-destroying enzyme (Denka Seiken Co., Japan) overnight at 37°C, heat-inactivated at 56°C for 30 min, diluted 1:10 with phosphate-buffered saline (PBS), and tested by performing an HI assay with 0.5% packed chicken red blood cells (cRBC) as described in the WHO Manual on Animal Influenza Diagnosis and Surveillance (http://www.who.int/vaccine_research/diseases/influenza/WHO_manual_on_animal-diagnosis_and_surveillance_2002_5.pdf). HI titers were expressed as the reciprocal of the highest dilution at which virus binding to red blood cells is blocked.

Filtered sera were tested by performing an MN assay in MDCK cells. Neutralizing titers were expressed as the reciprocal of the serum dilution that inhibited 50% of viral growth of 100 tissue culture infectious dose 50 ($TCID_{50}$) of virus. The HI and MN data are available in TABLE S1 and S2, respectively.

Because both HI and MN titers may be affected by the respective types of cells used, both assay types were used to ensure the robustness of the prediction from our sparse learning method. Only the residues appearing in the results from both HI and MN assay datasets and having bootstrap values more than 60% are considered to be antigenicity associated sites.

Molecular cloning and site-directed mutagenesis

Full-length HA and NA genes were amplified by reverse transcriptase PCR. The PCR fragments were cloned into the vector pHW2000. The site-directed mutagenesis was performed by using QuikChange Site-Directed Mutagenesis kit (Agilent). The identities of all clones were confirmed by full-length sequencing.

Influenza virus mutant generation

The H5N1 parent wild-type viruses were grown in embryonated chickens eggs or cultured cells. Each mutant was generated to contain 6 gene segments (PB2, PB1, PA, NP, M, NS) of A/Puerto Rico/8/34 and the NA gene and mutated HA gene of the target virus (so-called 6+2 viruses) by using pHW2000 plasmid–based reverse genetics (rg) [38]. The virus was rescued in 293T and MDCK cell lines as described elsewhere [33]. The mutagenesis targeted seven newly identified sites not located in reported antibody binding sites (except 120, H5 numbering) and amino acid position 45 (H5 numbering), which corresponds to site C (54, H3 numbering). The mutants at amino acid positions 129, 140 and 141 (H5 numbering) were available from other studies at the Webby Laboratory, and they were generated by mixing murine monoclonal antibodies with 04-VNM.

Phylogenetic tree construction and sequence analysis

The multiple-sequence alignments were conducted by using the MUSCLE software package [39], and the protein sequence alignments were examined manually to ensure that they were correct. The phylogenetic analyses and bootstrap resampling analyses were performed using PAUP* 4.0 Beta [40] to apply a maximum parsimony method, as described earlier [41].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Cherise Guess (St. Jude Children's Research Hospital) for editorial assistance, and Adam Rubrum and Daniel Darnell (Department of Infectious Diseases, St. Jude Children's Research Hospital) for excellent technical assistance. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, the National Institute of Allergy and Infectious Diseases or the National Institutes of Health.

Funding

This work was supported by the National Institute of Allergy and Infectious Diseases, RC1AI086830 to C.Z., T.Z., L.P. Long, and X.F.W., HHSN266200700005C to M.D., A.C.B., and R.J.W.; and by the American Lebanese Syrian Associated Charities (ALSAC) to M.D., A.C.B., and R.J.W.

References

1. Wan, X-F. Master thesis. South China Agricultural University; 1998. Isolation and characterization of avian influenza viruses in China.
2. Guo Y, Xu X, Wan X. [Genetic characterization of an avian influenza A (H5N1) virus isolated from a sick goose in China]. Zhonghua Shi Yan He Lin Chuang Bing Du Xue Za Zhi. 1998; 12:322–325. [PubMed: 12526344]
3. Zhao ZM, Shortridge KF, Garcia M, Guan Y, Wan XF. Genotypic diversity of H5N1 highly pathogenic avian influenza viruses. J Gen Virol. 2008; 89:2182–2193. [PubMed: 18753228]
4. Duan L, Bahl J, Smith GJ, Wang J, Vijaykrishna D, Zhang LJ, Zhang JX, Li KS, Fan XH, Cheung CL, Huang K, Poon LL, Shortridge KF, Webster RG, Peiris JS, Chen H, Guan Y. The development and genetic diversity of H5N1 influenza virus in China, 1996–2006. Virology. 2008; 380:243–254. [PubMed: 18774155]
5. Chen H, Deng G, Li Z, Tian G, Li Y, Jiao P, Zhang L, Liu Z, Webster RG, Yu K. The evolution of H5N1 influenza viruses in ducks in southern China. Proc Natl Acad Sci U S A. 2004; 101:10452–10457. [PubMed: 15235128]
6. Wan XF, Nguyen T, Davis CT, Smith CB, Zhao ZM, Carrel M, Inui K, Do HT, Mai DT, Jadhao S, Balish A, Shu B, Luo F, Emch M, Matsuoka Y, Lindstrom SE, Cox NJ, Nguyen CV, Klimov A, Donis RO. Evolution of highly pathogenic H5N1 avian influenza viruses in Vietnam between 2001 and 2008. PLoS One. 2007; 3:e3462. [PubMed: 18941631]
7. Ducatez MF, Olinger CM, Owoade AA, De Landtsheer S, Ammerlaan W, Niesters HG, Osterhaus AD, Fouchier RA, Muller CP. Avian flu: multiple introductions of H5N1 in Nigeria. Nature. 2006; 442:37. [PubMed: 16823443]
8. WHO. Toward a unified nomenclature system for highly pathogenic avian influenza virus (H5N1). Emerg Infect Dis. 2008; 14:e1.
9. WHO. Continuing progress towards a unified nomenclature for the highly pathogenic H5N1 avian influenza viruses: divergence of clade 2.2 viruses. Influenza Other Respi Viruses. 2009; 3:59–62. [PubMed: 19496842]
10. Guan Y, Poon LL, Cheung CY, Ellis TM, Lim W, Lipatov AS, Chan KH, Sturm-Ramirez KM, Cheung CL, Leung YH, Yuen KY, Webster RG, Peiris JS. H5N1 influenza: a protean pandemic threat. Proc Natl Acad Sci U S A. 2004; 101:8156–8161. [PubMed: 15148370]
11. Wu WL, Chen Y, Wang P, Song W, Lau SY, Rayner JM, Smith GJ, Webster RG, Peiris JS, Lin T, Xia N, Guan Y, Chen H. Antigenic profile of avian H5N1 viruses in Asia from 2002 to 2008. J Virol. 2007; 82:1798–1807. [PubMed: 18077726]
12. Knossow M, Daniels RS, Douglas AR, Skehel JJ, Wiley DC. Three-dimensional structure of an antigenic mutant of the influenza virus haemagglutinin. Nature. 1984; 311:678–680. [PubMed: 6207440]
13. Natali A, Oxford JS, Schild GC. Frequency of naturally occurring antibody to influenza virus antigenic variants selected in vitro with monoclonal antibody. J Hyg (Lond). 1981; 87:185–190. [PubMed: 7288173]

14. Varghese JN, Webster RG, Laver WG, Colman PM. Structure of an escape mutant of glycoprotein N2 neuraminidase of influenza virus A/Tokyo/3/67 at 3 A. J Mol Biol. 1988; 200:201–203. [PubMed: 3379640]

15. Air GM, Laver WG, Webster RG. Mechanism of antigenic variation in an individual epitope on influenza virus N9 neuraminidase. J Virol. 1990; 64:5797–5803. [PubMed: 1700825]

16. Bizebard T, Mauguen Y, Petek F, Rigolet P, Skehel JJ, Knossow M. Crystallization and preliminary X-ray diffraction studies of a monoclonal antibody Fab fragment specific for an influenza virus haemagglutinin and of an escape mutant of that haemagglutinin. J Mol Biol. 1990; 216:513–514. [PubMed: 2258927]

17. Hoffmann E, Lipatov AS, Webby RJ, Govorkova EA, Webster RG. Role of specific hemagglutinin amino acids in the immunogenicity and protection of H5N1 influenza virus vaccines. Proc Natl Acad Sci U S A. 2005; 102:12915–12920. [PubMed: 16118277]

18. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. Science. 1999; 286:1921–1925. [PubMed: 10583948]

19. Bush RM, Fitch WM, Bender CA, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol. 1999; 16:1457–1465. [PubMed: 10555276]

20. Ducatez MF, Cai Z, Peiris M, Guan Y, Ye Z, Wan XF, Webby RJ. Extent of antigenic cross-reactivity among highly pathogenic H5N1 influenza viruses. J Clin Microbiol. 2011; 49:3531–3536. [PubMed: 21832017]

21. Wilson IA, Cox NJ. Structural basis of immune recognition of influenza virus hemagglutinin. Annu Rev Immunol. 1990; 8:737–771. [PubMed: 2188678]

22. Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. The antigenic structure of the influenza virus A/PR/8/34 hemagglutinin (H1 subtype). Cell. 1982; 31:417–427. [PubMed: 6186384]

23. Ha Y, Stevens DJ, Skehel JJ, Wiley DC. H5 avian and H9 swine influenza virus haemagglutinin structures: possible origin of influenza subtypes. EMBO J. 2002; 21:865–875. [PubMed: 11867515]

24. Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE Jr, Wilson IA. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science. 2010; 328:357–360. [PubMed: 20339031]

25. Stevens J, Blixt O, Tumpey TM, Taubenberger JK, Paulson JC, Wilson IA. Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. Science. 2006; 312:404–410. [PubMed: 16543414]

26. Kaverin NV, Rudneva IA, Govorkova EA, Timofeeva TA, Shilov AA, Kochergin-Nikitsky KS, Krylov PS, Webster RG. Epitope mapping of the hemagglutinin molecule of a highly pathogenic H5N1 influenza virus by using monoclonal antibodies. J Virol. 2007; 81:12911–12917. [PubMed: 17881439]

27. Li J, Wang Y, Liang Y, Ni B, Wan Y, Liao Z, Chan KH, Yuen KY, Fu X, Shang X, Wang S, Yi D, Guo B, Di B, Wang M, Che X, Wu Y. Fine antigenic variation within H5N1 influenza virus hemagglutinin's antigenic sites defined by yeast cell surface display. Eur J Immunol. 2009; 39:3498–3510. [PubMed: 19798682]

28. Mulyanto CC, Saleh R. Prediction of a neutralizing epitope of a H5N1 virus hemagglutinin complexed with an antibody variable fragment using molecular dynamics simulation. Journal of Biophysical Chemistry. 2011; 2:258–267.

29. Ndifon W, Wingreen NS, Levin SA. Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. Proc Natl Acad Sci U S A. 2009; 106:8701–8706. [PubMed: 19439657]

30. Pereira HG, Rinaldi A, Nardelli L. Antigenic variation among avian influenza A viruses. Bull World Health Organ. 1967; 37:553–558. [PubMed: 5301736]

31. Hoffmann E, Neumann G, Kawaoka Y, Hobom G, Webster RG. A DNA transfection system for generation of influenza A virus from eight plasmids. Proc Natl Acad Sci U S A. 2000; 97:6108–6113. [PubMed: 10801978]

32. Skehel JJ, Wiley DC. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem. 2000; 69:531–569. [PubMed: 10966468]

33. Ducatez MF, Bahl J, Griffin Y, Stigger-Rosser E, Franks J, Barman S, Vijaykrishna D, Webb A, Guan Y, Webster RG, Smith GJ, Webby RJ. Feasibility of reconstructed ancestral H5N1 influenza viruses for cross-clade protective vaccine development. Proc Natl Acad Sci U S A. 2011; 108:349–354. [PubMed: 21173241]
34. Cai Z, Zhang T, Wan X-F. Concepts and applications for influenza antigenic cartography. Influenza and Other Respiratory Viruses. 2011; 5:204–207. [PubMed: 21761589]
35. Cai Z, Zhang T, Wan XF. A computational framework for influenza antigenic cartography. PLoS Comput Biol. 2010; 6:e1000949. [PubMed: 20949097]
36. Barnett JL, Yang J, Cai Z, Zhang T, Wan X-F. AntigenMap 3D: an online antigenic cartography resource. Bioinformatics. 2012 **In Press**.
37. Tibshirani R. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society (Series B). 1996; 58:267–288.
38. Hoffmann E, Krauss S, Perez D, Webby R, Webster RG. Eight-plasmid system for rapid generation of influenza virus vaccines. Vaccine. 2002; 20:3165–3170. [PubMed: 12163268]
39. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]
40. Swofford, DL. PAUP*: Phylogenic analysis using Parsimony. Sunderland, Massachusetts: Sinauer; 1998.
41. Wan XF, Ren T, Luo KJ, Liao M, Zhang GH, Chen JD, Cao WS, Li Y, Jin NY, Xu D, Xin CA. Genetic characterization of H5N1 avian influenza viruses isolated in southern China during the 2003-04 avian influenza outbreaks. Arch Virol. 2005; 150:1257–1266. [PubMed: 15717120]

Highlights

- A novel method for influenza antigenic site identification was developed
- This method integrates protein sequences and immunological datasets
- 21 antigenic sites were identified in HA gene of H5N1 virus
- Wet lab experiments validated predicting results

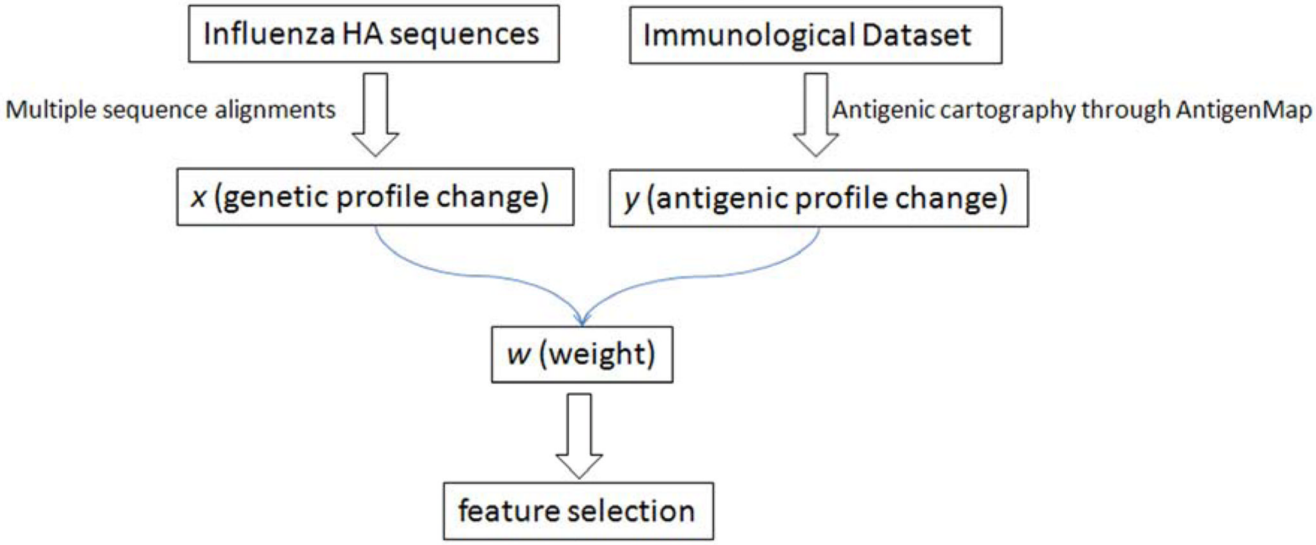
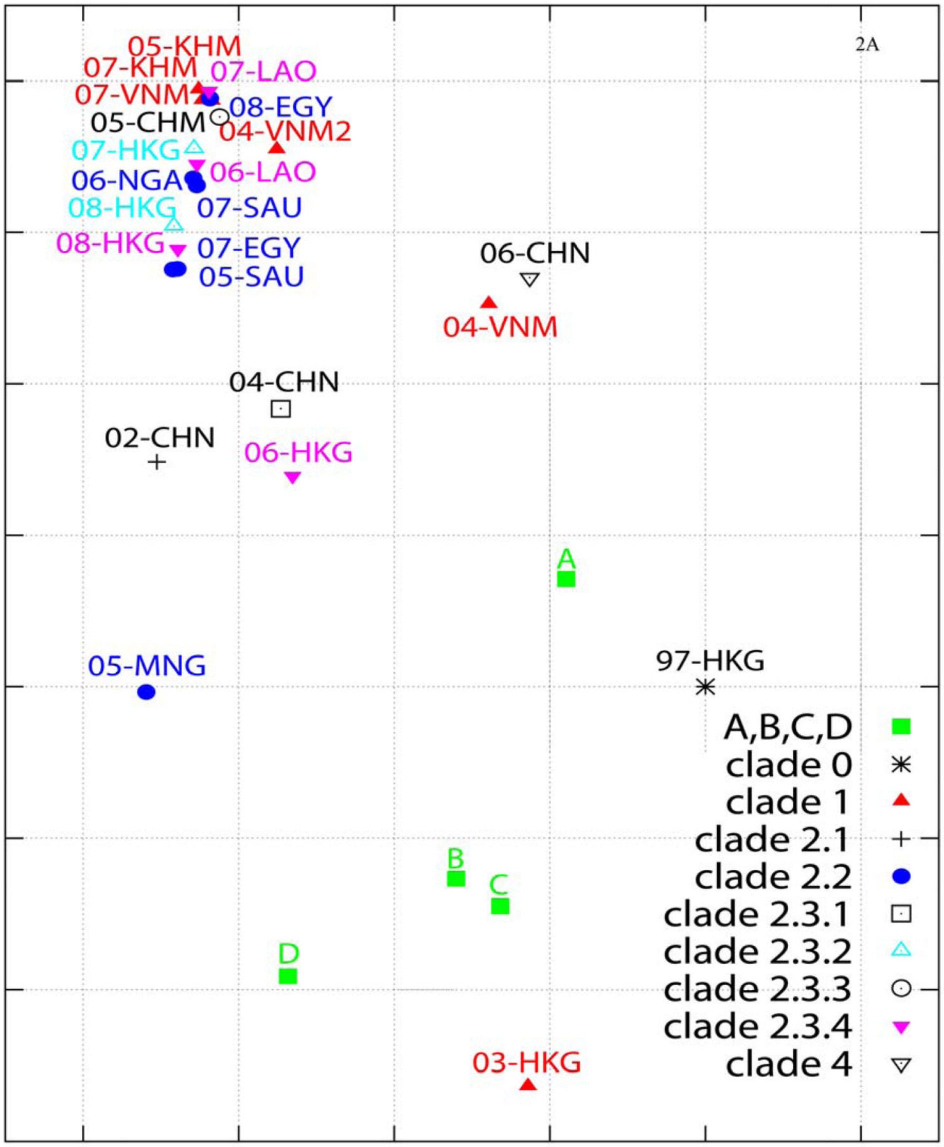
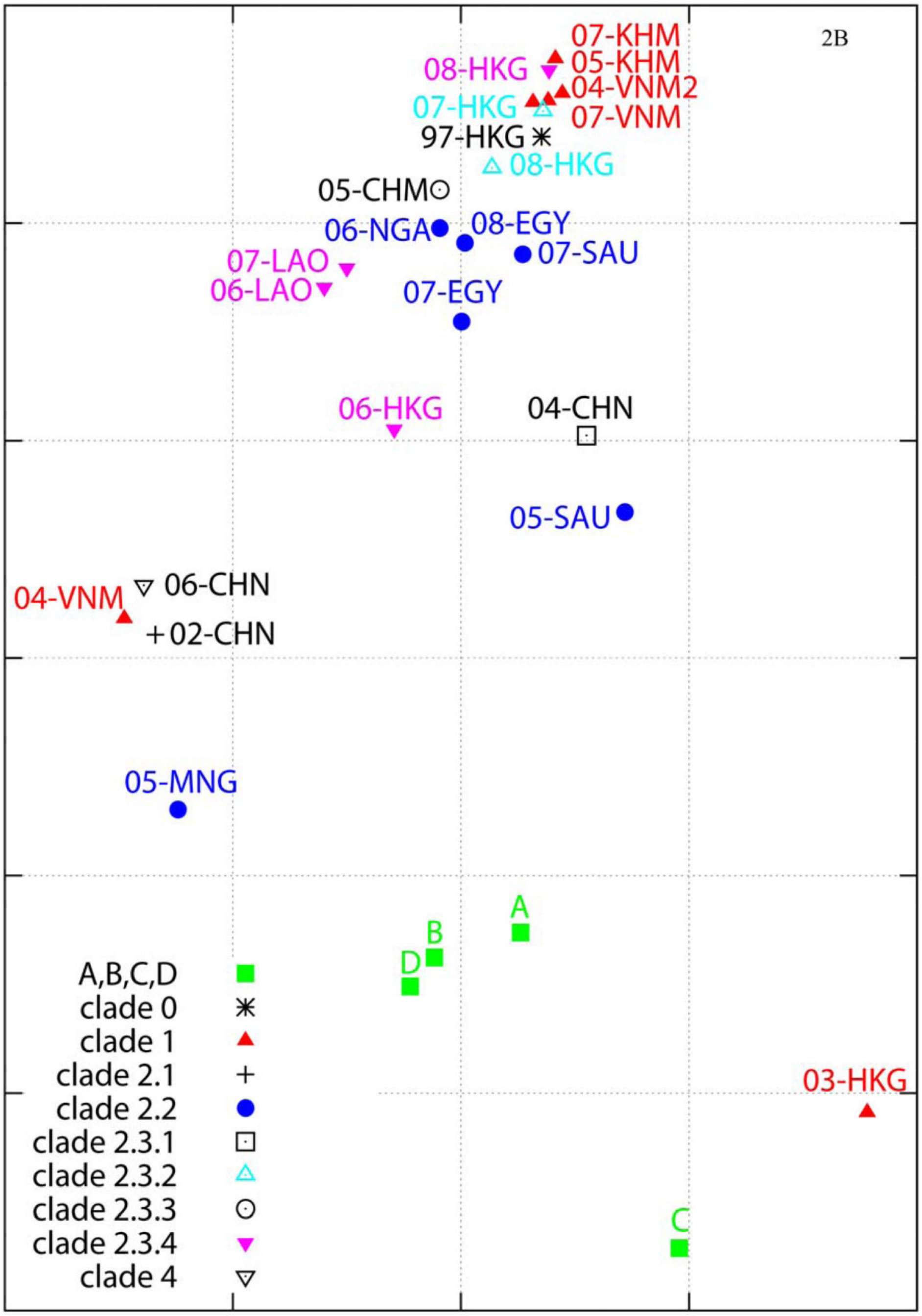

Influenza HA sequences
Immunological Dataset
Multiple sequence alignments
Antigenic cartography through AntigenMap
x (genetic profile change)
y (antigenic profile change)
w (weight)
feature selection

FIG. 1.

The workflow for the different steps involved in the integrative sparse learning method.


2A
05-KHM
07-KHM
07-LAO
07-VNM
08-EGY
05-CHM
04-VNM2
07-HKG
06-LAO
06-NGA
08-HKG
07-SAU
08-HKG
07-EGY
05-SAU
06-CHN
04-VNM
04-CHN
02-CHN
06-HKG
A
05-MNG
97-HKG
A,B,C,D
clade 0
clade 1
clade 2.1
clade 2.2
clade 2.3.1
clade 2.3.2
clade 2.3.3
clade 2.3.4
clade 4
B
C
D
03-HKG


2B
07-KHM
05-KHM
08-HKG
04-VNM2
07-HKG
07-VNM
97-HKG
08-HKG
05-CHM
06-NGA
08-EGY
07-SAU
07-LAO
06-LAO
07-EGY
06-HKG
04-CHN
05-SAU
04-VNM
06-CHN
02-CHN
05-MNG
A
B
D
C
03-HKG
A,B,C,D
clade 0
clade 1
clade 2.1
clade 2.2
clade 2.3.1
clade 2.3.2
clade 2.3.3
clade 2.3.4
clade 4

2C

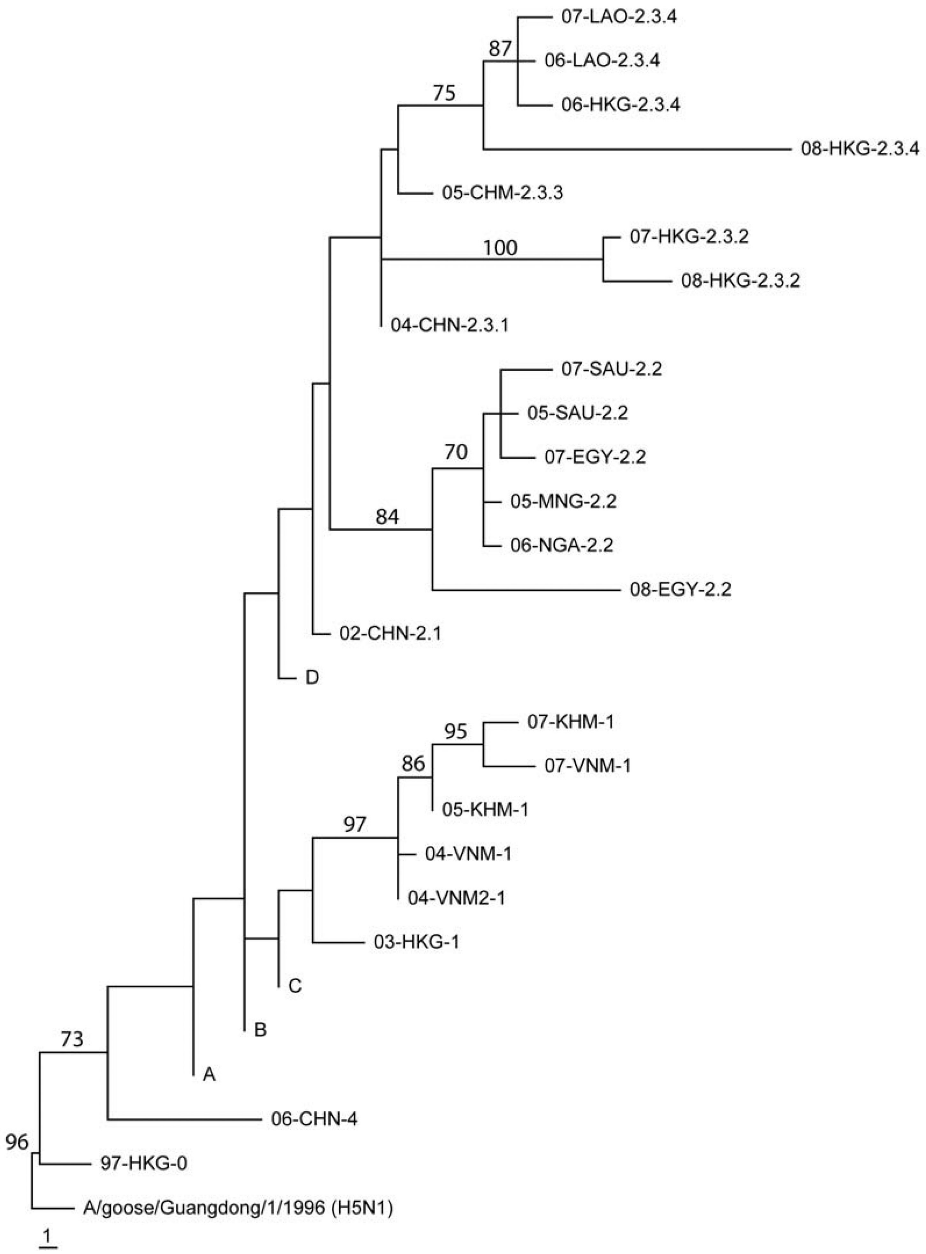


FIG. 2.

The antigenic cartography of H5N1 highly pathogenic avian influenza virus (HPAIV) made by using AntigenMap (http://sysbio.cvm.msstate.edu/AntigenMap)[34; 35; 36]. Cartography based on the results from hemagglutination inhibition assays in chicken red blood cells (**A**) or microneutralization assays in MDCK cells (**B**). One unit (grid) represents a 2-fold change in HI assay results. The cartography includes 27 influenza viruses (listed in TABLE 1), which includes viruses from 9 clades or subclades (**C**) [8; 9]. The antigenic clusters are marked with a large circle for visualization purpose.

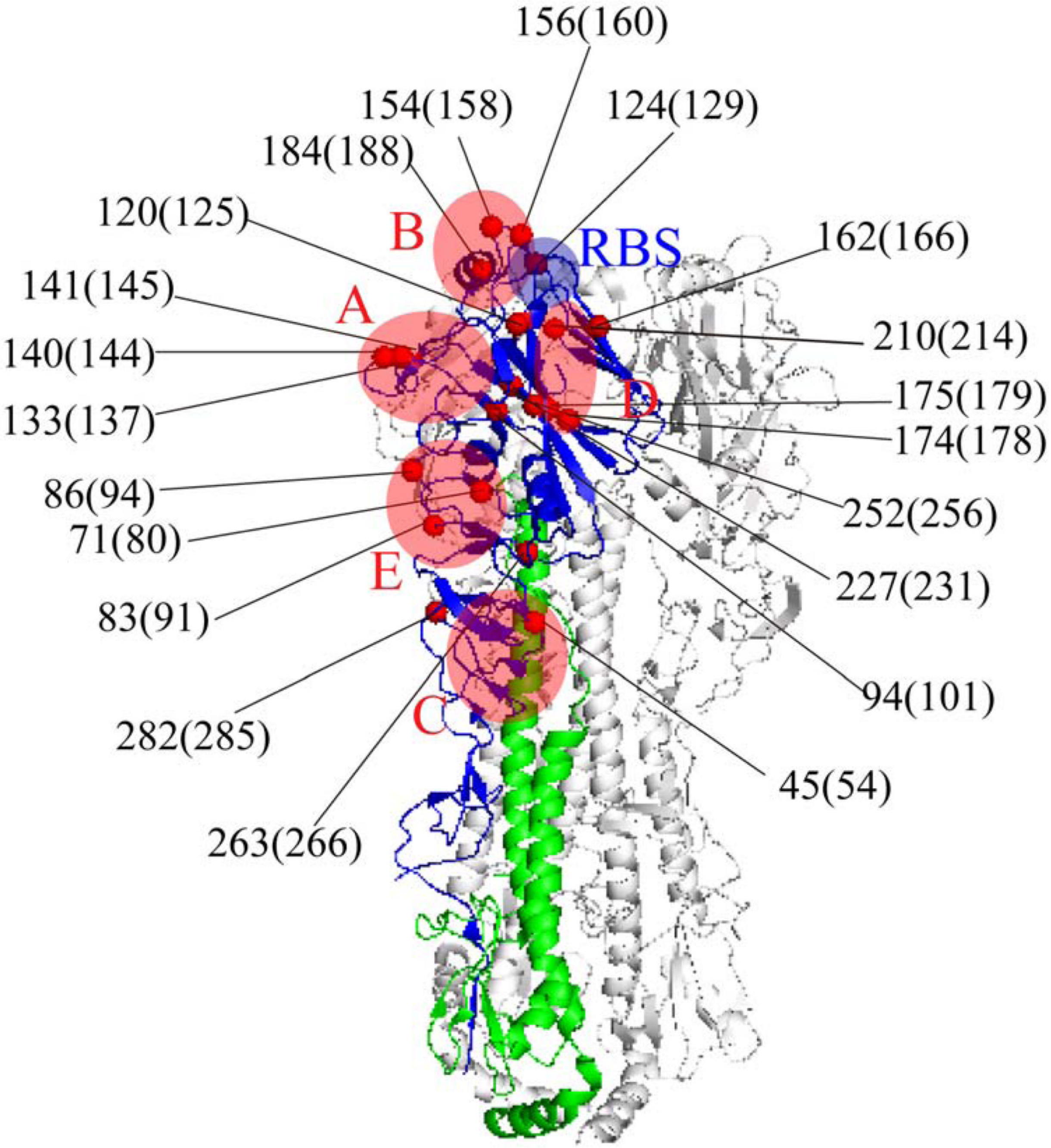


FIG. 3.
Ribbon diagram of the trimeric hemagglutinin (HA) molecule with the 21 identified antigenicity associated sites (all sites listed in TABLE 2) shown in red spheres. The mutant positions were numbered based on the corresponding residue number in HA of H5N1 HPAIVs, and the corresponding numbers in H3 HA are listed in parentheses. The antibody binding sites A–E are shown in red, and the receptor-binding site is shown in blue.

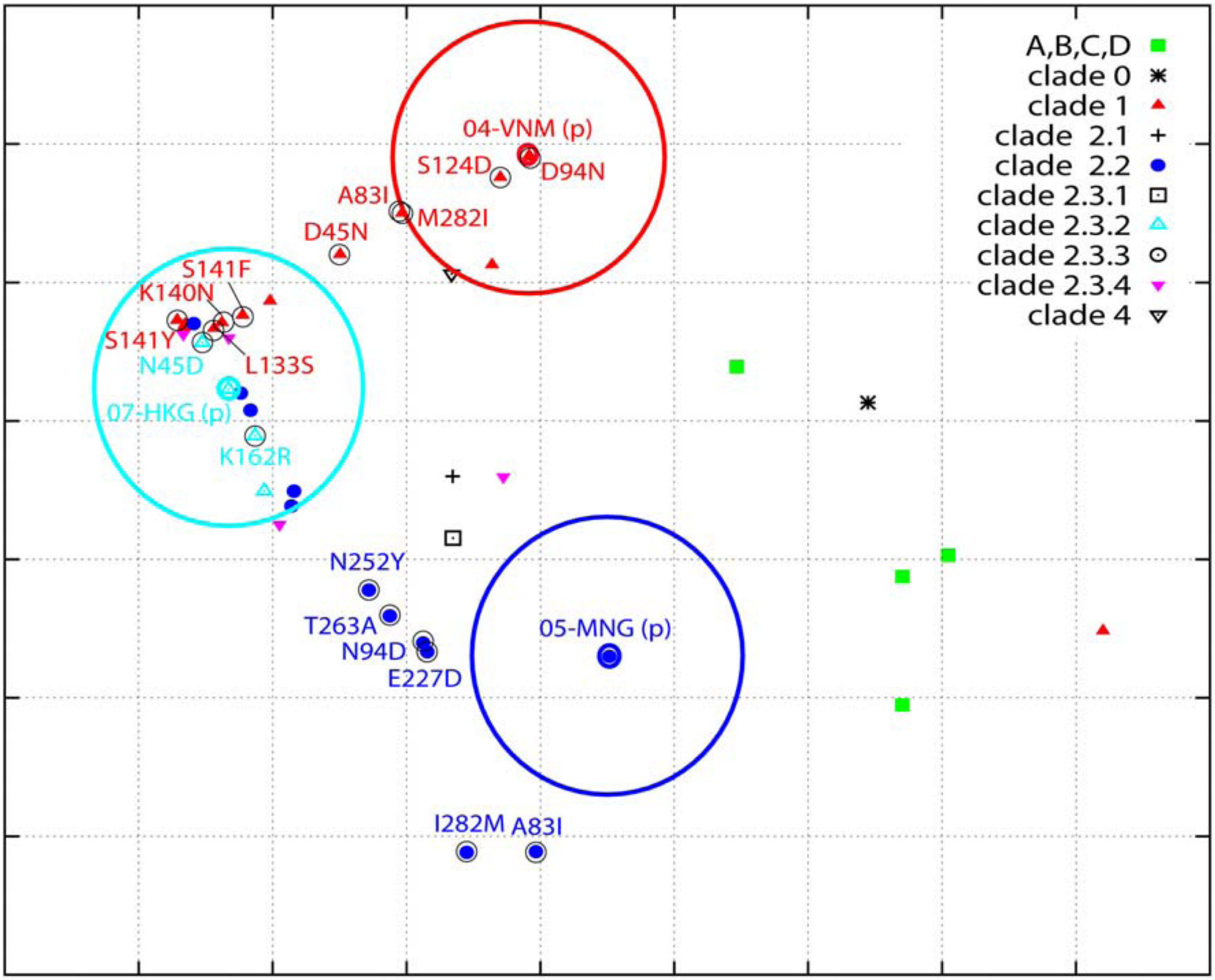


FIG. 4.

The antigenic cartography of H5N1 HPAIVs and their mutants (listed in TABLE 4) made by using AntigenMap (http://sysbio.cvm.msstate.edu/AntigenMap)[34; 35; 36]. The mutant positions were numbered based on the corresponding residue number in HA of H5N1 HPAIVs. The parental viruses are labeled with heavily colored circles, and the mutants with lightly colored circles. The parental viruses are also encircled (circle = 2-unit diameter). A symbol (p) is appended after the name of each parent strain. One unit (grid) corresponds to a 2-fold change in HI titer.

TABLE 1

The H5N1 viruses used in this study.

Clade	Virus	Abbreviation
	A	A
	B	B
	C	C
	D	D
0	A/Hong Kong/156/97	97-HKG
1	A/Hong Kong/213/03	03-HKG
1	A/Vietnam/1203/04	04-VNM
1	A/Vietnam/1194/04	04-VNM2
1	A/chicken/Cambodia/13LC1/05	05-KHM
1	A/muscovy duck/Vietnam/33/07	07-VNM
1	A/Cambodia/R0405050/07	07-KHM
2.1	A/duck/Hunan/795/02	02-CHN
2.2	A/whooper swan/Mongolia/244/05	05-MNG
2.2	A/falcon/Saudi Arabia/D1795/05	05-SAU
2.2	A/chicken/Nigeria/42/06	06-NGA
2.2	A/falcon/Saudi Arabia/D1936/07	07-SAU
2.2	A/turkey/Egypt/7/07	07-EGY
2.2	A/chicken/Egypt/1/08	08-EGY
2.3.1	A/duck/Hunan/101/04	04-CHN
2.3.2	A/common magpie/Hong Kong/5052/07	07-HKG
2.3.2	A/grey heron/Hong Kong/1046/08	08-HKG
2.3.3	A/chicken/Guiyang/3570/05	05-CHN
2.3.4	A/duck/Laos/3295/06	06-LAO
2.3.4	A/Japanese white-eye/Hong Kong/1038/06	06-HKG
2.3.4	A/duck/Laos/A0301/07	07-LAO
2.3.4	A/chicken/Hong Kong/AP156/08	08-HKG
4	A/goose/Guiyang/337/06	06-CHN

TABLE 2

Antigenicity associated sites identified in H5N1's HA protein by using a sparse learning method.

position[a]			mutation[c]	weight (SD) using HI data[d]	weight (SD) using MN data[d]	*bootstrap value*[e]		Reference
H5	H3	H1[b]				HI	MN	
45	54(C)	48	D/N	0.163(0.067)	0.246(0.079)	65	89	
71	80(E)	76	I/L	0.385(0.196)	0.628(0.195)	93	100	
83	91(E)	87	A/I	0.251(0.166)	0.200(0.116)	78	78	
86	94(E)	90	A/V	0.412(0.093)	0.252(0.077)	100	100	11; 28
94	101	98	N/D/V	0.195(0.099)	0.256(0.090)	73	96	
120	125	122	S/N/D	0.357(0.110)	0.231(0.081)	100	98	
124	129(RBS)	129(Sa)	D/N/S	0.640(0.165)	0.165(0.091)	100	81	11; 27
133	137(A)	136	S/A	0.192(0.137)	0.404(0.136)	74	98	
140	144(A)	144	K/R/T/G/M/E	0.249(0.101)	0.136(0.064)	91	64	26
141	145(A)	145(Ca2)	S/P	0.169(0.119)	0.181(0.074)	64	68	26
154	158(B)	158(Sb)	N/D/G	0.158(0.067)	0.135(0.052)	74	61	28
156	160(B)	160(Sa)	A/T/K	0.293(0.094)	0.155(0.056)	98	92	26
162	166	166(Sa)	R/I/K/S	0.321(0.095)	0.243(0.088)	94	96	26
174	178(D)	178	V/I	0.297(0.213)	0.177(0.125)	89	72	
175	179(D)	179	L/M	0.224(0.113)	0.399(0.114)	80	91	
184	188(B)	188	A/E	0.222(0.118)	0.238(0.107)	81	90	
210	214(D)	214	V/T/I	0.210(0.072)	0.185(0.052)	82	87	
227	231	231	E/D	0.202(0.128)	0.216(0.089)	60	84	
252	256	256	Y/N	0.213(0.116)	0.186(0.074)	72	72	
263	266	267	A/T	0.304(0.074)	0.358(0.056)	89	89	27
282	285	286	M/I	0.305(0.095)	0.248(0.074)	89	87	

[a] The position is numbered using the corresponding residue number in the H5 HA sequence or the corresponding one in the H3 HA protein;

[b] The position in H1 was based on the amino acid number in HA of PR8 (PDB ID 1RU7);

[c] The mutations are predicted to affect the titers in immunologic datasets, and the corresponding antibody binding sites in H3 HA are shown in parentheses. RBS denotes receptor-binding site;

[d] The weight was an mean of absolute weight values calculated from a total of 100 runs using HI or MN data, and the numbers in the parenthesis denote the corresponding standard deviations;

[e] The bootstrap values were calculated based on 100 runs.

TABLE 3

Genetic polymorphism of H5N1 virus' *HA* gene among the antigenicity associated sites identified by using a sparse learning method.

		Position																			
	H5	**45**	**71**	**83**	**86**	**94**	**120**	**124**	**133**	**140**	**141**	**154**	**156**	**162**	**174**	**175**	**184**	**210**	**227**	**252**	**263**	**282**
Group	**H3**	**54**	**80**	**91**	**94**	**101**	**125**	**129**	**137**	**144**	**145**	**158**	**160**	**166**	**178**	**179**	**188**	**214**	**231**	**256**	**266**	**285**
	H1	**48**	**76**	**87**	**90**	**98**	**122**	**129**	**136**	**144**	**145**	**158**	**160**	**166**	**178**	**179**	**188**	**214**	**231**	**256**	**267**	**286**
ABCD		D	I	A	A	D/N	S	N/S/D	S	K	S	N	A	R	V	L	A	V	E/D	Y	A	M
Clade 0		N	I	A	A	N	S	N	S	R	S	N	A	R	V	L	A	V	E	Y	T	M
Clade 1		D	I	A	A/V	N/V	S/N	S	S/A	K	S	N	T/A	R	V	M/L	A	V/T	E	Y	T/A	M
Clade 2.1		D	I	A	A	N	S	D	S	K	S	N	A	R	V	L	A	V	E	Y	A	M
Clade 2.2		D	L	I	A/V	N	S	D	S	R/G	S/P	D/N	A	R/I/K	V	L	A/E	V	E	N	T/A	I
Clade 2.3.1		D	I	A	A	N	S	D	S	K	S	N	A	R	V	L	A	V	D	Y	A	M
Clade 2.3.2		N	I	A	A	N	D	D	S	N	S	G/D	A	K	V	L	E	V	D	Y	T	M
Clade 2.3.3		D	I	A	A	N	S	D	S	K	S	N	A	R	V	L	A	I	D	Y	A	M
Clade 2.3.4		D	I/N	A	A	N	S/N	D	S/A	T/M	P	N	T/K	R/S	I	L	A/E	V	D	Y	A	I
Clade 4		D	I	A	A	D	S	N	S	E	S	N	S	R	V	L	A	V	E	Y	A	M

TABLE 4

Antigenic distance changes between mutant influenza viruses and their respective parent influenza viruses.

Position [a]			Mutant [b]	Antigenic Distance Change [c]	Position [a]			Mutant [b]	Antigenic Distance Change [c]
H5	H3	H1			H5	H3	H1		
45	54(C)	48	04-VNM-D45N	1.5866	141	145(A)	145(Ca2)	04-VNM-S141F	2.4349
			07-HKG-N45D	0.3897				04-VNM-S141Y	2.8910
83	91(E)	87	04-VNM-A83I	1.0408	162	166	166(Sa)	07-HKG-K162R	0.3901
			05-MNG-I83A	1.5136	227	231	231	05-MNG-E227D	1.3612
94	101	98	04-VNM-D94N	0.0176	252	256	256	05-MNG-N252Y	1.8631
			05-MNG-N94D	1.3984	263	266	267	05-MNG-T263A	1.6685
124	129(RBS)	129(Sa)	04-VNM-S124D	0.2685	282	285	286	04-VNM-M282I	1.0400
129	133(A)	133	04-VNM-L129S	2.6660				05-MNG-I282M	1.7717
140	144(A)	144	04-VNM-K140N	2.5963					

[a] The positions are numbered based on the amino-acid position in the HA of H5N1 HPAIV and corresponding positions in HA of H3N2 influenza A virus, respectively;

[b] The 6+2 mutants generated by using site-directed mutagenesis and reverse genetics were named according to HA/NA donor-mutation, 04-VNM denotes A/Vietnam/1203/04(H5N1), 05-MNG A/whooper swan/Mongolia/244/05(H5N1), and 07-HKG A/common magpie/Hong Kong/5052/07(H5N1);

[c] These antigenic distances were measured by using antigenic cartography [34; 35] (see FIG. 2). One grid unit corresponds to a 2-fold change in HI titer.