ON LINE SUPPORTING INFORMATION for

Using Sequence Data To Infer the Antigenicity of Influenza Virus

Running title: Identify influenza antigenic variant via sequences

Hailiang Sun[a,1], Jialiang Yang[a,1], Tong Zhang[b], Li-Ping Long[a], Kun Jia[a], Guohua Yang[a], Richard J. Webby[c], and Xiu-Feng Wan[a,2]

[a]Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA; [b]Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA; [c]Department of Infectious Diseases, St. Jude Children's Research Hospital, 262 Danny Thomas Place, Memphis, TN 38105, USA.

[1] These authors contributed equally to this work.

[2] To whom correspondence should be addressed:

Xiu-Feng Wan

Department of Basic Sciences

College of Veterinary Medicine

Mississippi State University

240 Wise Center Dr.

Mississippi State, MS 39762, USA

Phone: (662)325-3559; Fax: (662)325-3884; E-mail: wan@cvm.msstate.edu

Keywords: influenza A virus, antigenicity, antibody binding site, genomic sequence, antigenic variant, vaccine strain selection, machine learning

This folder contains the following supporting documents.

Supporting Files

1. H3N26807_HI.tab: the HI table contains 512 viruses and 133 serum used for training.
2. H3N26807_512_seq_HA1.aln: the 512 HA1 protein sequence alignment.
3. H3N26812_3332_seq_HA1.aln: the alignment of 3332 non-overlapping H3N2 HA1 protein sequences from 1968 to 2012.
4. H3N26812_3332_seq_mutationpattern.xlsx: the mutation pattern of 39 identified key antigenicity associated sites over the 3332 sequences.
5. H3N2Site.surface: the surface sites used in the study.

Supporting Tables

TABLE W1. Accuracy of binary and PIMA scoring functions

TABLE W2. GetArea-calculated surface sites

Supporting Figures

FIG. W1. Three-dimensional structure of H3 influenza's hemagglutinin showing the mutations that drive designated antigenic drift events (A-I). The residues and their antigenic binding domains are marked. Seven of nine sets of multiple mutations leading to antigenic drift are in the same antibody-binding sites. Seven of the nine sets of mutations (TABLE 1) that lead to the same antigenic drift are in antibody-binding sites that are the same (e.g., Y155H and K189R, which drive antigenic drift from BK79 to SI87 [*C*]) or neighboring (e.g., E135K, K145N, and E156K, which all drive antigenic drift from BE89 to BE92 [*E*]); the other 2 sets are in distant antibody-binding sites (e.g., N145K and G172D, which drive antigenic drift from BE92 toWU95 [F]; and K62E, K156Q, and, E158K, which drive antigenic drift from WU95 to SY97 [*G*]). The protein structure was visualized by using PyMOL with the H3 HA structure (pdb: 2VIU).

FIG. W2. The amino acid hierarchical scoring schema adapted from pattern-induced multi-sequence alignment (PIMA). In PIMA, each leaf represents an amino

acid, each ancestral node represents an amino acid class, and the number in the cardinality column denotes the level of each node. The similarity score of amino acids x and y is calculated as six minus the level of the amino acids' most recent ancestral node.

FIG. W3. Banded structure for H3N2 influenza HI data. The high reactors are predominantly in the diagonal zone, and the low reactors and the missing values appear more approaching the two corners. The sparse-learning method used in this study derived the long-distance matrix by using a temporal model and the short-distance matrix without using a temporal model. The long-distance matrix was used to learn the long-weight matrix, and the short-distance matrix was used to learn the short-weight matrix.

FIG. W4. The root mean square error (RMSE) and Pearson correlation coefficient (CC) curves used to determine the number of sites to select from the H3N2 data (1968 to 2007 viruses). The triangle curve plots the trend of RMSE against the number of sites selected, and the dot curve plots the trend of CC against the number of sites.

FIG. W5. Maps showing the effects of predicted residues in antibody-binding sites in driving designated antigenic drifts *(A-G)*. Each colored ball on the map represents a virus: gray balls denote the viruses in the previous cluster of an antigenic drift event; blue balls denote the viruses in the drifted cluster; and the yellow ball represents the wild-type strain from which we generated simulated drift variants by mutating key residues in its HA sequence. Each mutant is marked with its mutation type on the wild-type strain. For example, N145K denotes the viruses obtained by mutating the amino acid asparagine (N) to lysine (K) in position 145 of the wild-type strain, and G135K-N145K represents a double mutation derived similarly. The mutations not driving antigenic drift are marked in red, and those driving antigenic drift are marked in light blue. One unit on the map corresponds to a 2-fold change in HI titer