Modeling the Antigenic Evolution of Influenza Viruses from Sequences

Taijiao Jiang

Center of Systems Medicine, Chinese Academy of Medical Sciences
Suzhou Institute of Systems Medicine

October 8-10, 2015. Sino-German Symposium, Beijing
Overview of the talk

I. Backgrounds:
   Influenza virus, antigenic evolution and vaccination.

II. Modeling influenza virus antigenic evolution
   - Methods for predicting antigenic variants:
     1). EADpred
     2). Feature-based Naive Bayes model
   - Methods for predicting antigenic evolution patterns
     1). Genomic co-occurrence network

III. Application of PREDAC to recommend vaccine strains for seasonal influenza viruses.
Overview of the talk

I. Backgrounds:
   Influenza virus, antigenic evolution and vaccination.

II. Modeling influenza virus antigenic evolution
   - Methods for predicting antigenic variants:
     1). EADpred
     2). Feature-based Naive Bayes model
   - Methods for predicting antigenic evolution patterns
     1). Genomic co-occurrence network

III. Application of PREDAC to recommend vaccine strains for seasonal influenza viruses.
The impact of influenza virus on public health

Pandemic flu viruses: up to tens of millions deaths
Seasonal flu viruses: infecting 10-15% human population each year, up to ¼ million deaths (H1N1, H3N2 and B).
The newly emerging flu viruses: occasional infections with high rate of death, causing social panic.
What does a flu virus look like?

Influenza genome and proteome

Genetic composition of a flu virus

Negative single-stranded RNA virus.

8 RNA segments encoding at least 12 proteins:
- **NP, PA, PB1 and PB2** code for viral replication machinery
- **NS** for non-structural proteins 1 and 2 (NS-1, 2),
- **MP** for matrix proteins 1 and 2 (M-1 and 2).
- **HA** for hemagglutinin
- **NA** for neuraminidase

The surface glycoprotein, HA, controls viral entry into the cells and is the major antigenic target of the host immune responses.
HA structure

Antigenic regions

Receptor-binding region

Vaccination: the most effective way to control influenza infection in humans

Vaccination

Flu seasons:

North Hemisphere: October - April next year.

In mid-Feb of every year, WHO will recommend trivalent vaccine composition: A/H1N1, A/H3N2, and B.

Grown in eggs

Inactivated virus or live attenuated virus

Injection (Flu shot)
Effective vaccination requires that the antigenic property of vaccine strain should match that of circulating strains.

The critical step in the vaccine strategy is to select the vaccine strains that match future circulating strains in antigenicity.
WHO has set up a Global Influenza Surveillance Network (GISN)

- 135 National influenza centers from 105 countries
- 6 WHO Collaborating centers (WHOCC: USA, UK, Japan, Australia and China)
- 12 reference labs for H5 viruses

- Perform antigenic analysis and HA sequencing
- Assess epidemiologic behavior
- Identify potential circulating strains
- Recommend vaccine strains
Current situation: the frequent vaccine mismatches

Our analysis on the vaccine matches from 2002 to 2009.

Influenza virus is highly mutable

Flu virus mainly uses two strategies to change its antigenicity to escape host immune surveillance.

1. Gene mutation
2. Gene reassortment

Co-infection

Genetic drift to antigenic drift

Genetic shift to antigenic shift

Reassorted viruses
The experimental method for flu virus antigenicity analysis: Hemagglutination inhibition (HI) assay

Disadvantages of HI assay:
- Labor-intensive & time-consuming
- Not consistent & low sensitivity

Antigenic variants detection
Vaccine strain recommendation
Sequence-based approaches have become important alternatives to assist vaccine strain recommendation.

At present, sequencing of HA or even whole viral genome has been a routine work in influenza surveillance.

Sequence

Identify influenza antigenic variants and recommend vaccine strains.
Overview of the talk

I. Backgrounds:
   Influenza virus, antigenic evolution and vaccination.

II. Modeling influenza virus antigenic evolution
   - Methods for predicting antigenic variants:
     1). EADpred
     2). Feature-based Naive Bayes model
   - Methods for predicting antigenic evolution patterns
     1). Genomic co-occurrence network

III. Application of PREDAC to recommend vaccine strains for seasonal influenza viruses.
WHO relies on the two critical steps in vaccine strain recommendation

**Step 1.** Monitor antigenic variants across the world. Can we use only sequence information for the mission of the two steps?

**Step 2.** Assess their potential dominance in the coming season.
The key question 1: how to accurately predict antigenic variants from HA sequences?
Previous models to predict antigenic relationships

Either counting amino acid changes on HA, particularly those in the epitope regions.

Wilson et al. 1990, Annu Rev Immunol 8, 737
Lee et al. 2004, Emerg Infect Dis 10, 1385
Munoz et al. 2005, Vaccine 23, 1114
Gupta et al. 2006, Vaccine 24, 3881
Lee et al. 2007, Vaccine 25, 8133
Liao et al. 2008, Bioinformatics 24(4), 505
Huang et al. 2009, BMC Bioinformatics 10 (Sup 1), S41
Lees et al. 2010, Bioinformatics, Epub ahead of print
Previous models to predict antigenic relationships

<table>
<thead>
<tr>
<th>Position</th>
<th>Antigenic domain</th>
<th>Residue frequency among the 45 viruses in the training dataset</th>
<th>GM4</th>
<th>GM5</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA82</td>
<td>E</td>
<td>17 (E), 28 (K)</td>
<td>0.998</td>
<td>1.037</td>
</tr>
<tr>
<td>AA92</td>
<td>E</td>
<td>1 (E), 44 (K)</td>
<td>0.941</td>
<td>0.920</td>
</tr>
<tr>
<td>AA121</td>
<td>D</td>
<td>28 (I), 10 (N), 7 (T)</td>
<td>0.495</td>
<td>0.546</td>
</tr>
<tr>
<td>AA124</td>
<td>A</td>
<td>13 (D), 20 (G), 2 (N), 10 (S)</td>
<td>0.298</td>
<td></td>
</tr>
<tr>
<td>AA129</td>
<td>B</td>
<td>1 (E), 44 (G)</td>
<td>1.748</td>
<td></td>
</tr>
<tr>
<td>AA135</td>
<td>A</td>
<td>1 (E), 23 (G), 6 (K), 15 (T)</td>
<td>0.954</td>
<td>1.021</td>
</tr>
<tr>
<td>AA144</td>
<td>A</td>
<td>13 (D), 3 (I), 5 (N), 24 (V)</td>
<td>0.716</td>
<td>0.683</td>
</tr>
<tr>
<td>AA145</td>
<td>A</td>
<td>1 (E), 18 (K), 22 (N), 1 (R), 3 (S)</td>
<td>1.209</td>
<td>1.282</td>
</tr>
<tr>
<td>AA155</td>
<td>B</td>
<td>30 (H), 2 (T), 13 (Y)</td>
<td>1.202</td>
<td>1.582</td>
</tr>
<tr>
<td>AA156</td>
<td>B</td>
<td>8 (E), 1 (H), 27 (K), 9 (Q)</td>
<td>0.400</td>
<td>0.294</td>
</tr>
<tr>
<td>AA157</td>
<td>B</td>
<td>26 (L), 19 (S)</td>
<td>0.423</td>
<td>0.448</td>
</tr>
<tr>
<td>AA158</td>
<td>B</td>
<td>29 (E), 7 (G), 9 (K)</td>
<td>0.761</td>
<td>0.715</td>
</tr>
<tr>
<td>AA160</td>
<td>B</td>
<td>1 (A), 35 (K), 1 (R), 1 (S), 7 (T)</td>
<td>1.072</td>
<td>1.073</td>
</tr>
<tr>
<td>AA173</td>
<td>D</td>
<td>34 (K), 11 (N)</td>
<td>1.285</td>
<td>1.301</td>
</tr>
<tr>
<td>AA174</td>
<td>D</td>
<td>40 (F), 4 (S), 1 (V)</td>
<td>0.613</td>
<td>0.633</td>
</tr>
<tr>
<td>AA188</td>
<td>B</td>
<td>42 (D), 1 (E), 1 (N), 1 (Y)</td>
<td>1.087</td>
<td>1.234</td>
</tr>
<tr>
<td>AA189</td>
<td>B</td>
<td>8 (K), 5 (Q), 8 (R), 24 (S)</td>
<td>0.721</td>
<td>0.684</td>
</tr>
<tr>
<td>AA240</td>
<td>D</td>
<td>44 (G), 1 (R)</td>
<td>0.690</td>
<td>0.708</td>
</tr>
<tr>
<td>AA273</td>
<td>C</td>
<td>44 (P), 1 (S)</td>
<td>0.779</td>
<td>0.738</td>
</tr>
<tr>
<td>AA276</td>
<td>C</td>
<td>9 (K), 14 (N), 22 (T)</td>
<td>1.830</td>
<td>2.287</td>
</tr>
</tbody>
</table>

Agreement rate in the training dataset (N=181): 93.37% 92.82%

Agreement rate in the validation dataset (N=96): 91.67% 91.67%

Or considering the specific amino acid changes at certain positions on HA observed during influenza evolution (site-dependent models).

\[
D_{HI} \text{ distance} = e^{\sum_{j} w_j s_j}, \quad s_j = \begin{cases} 
1, & \text{if } X_j = 1 \\
0, & \text{if } X_j = 0 
\end{cases}
\]

Liao et al. 2008, Bioinformatics 24(4), 505
Huang et al. 2009, BMC Bioinformatics 10 (Sup 1), S41
Lees et al. 2010, Bioinformatics, Epub ahead of print
The motivation of our modeling work

AA Change $\neq$ Antigenic Change
Influenza virus may explore new amino acid changes even different amino-acid positions during influenza evolution!

So,
we are interested in developing a more general computational model to predict influenza antigenic relationships.
We view the antigenic changes as the changes of the interactions between the HA and host antibodies.

Based on HA sequence-structure modeling, we developed two models:

1. **EADpred (Epitope-based Antigenic Distance Prediction)**

2. **Machine learning-based model**
EADpred: Epitope-based Antigenic Distance Prediction

Antigenic epitopes as basic structural units to mediate HA-antibody interactions

Physicochemical rules underlying HA-antibody interactions
Four steps of the EADpred

Step 1: Identify the antigenic epitopes

Step 2: Transform amino acid changes in each antigenic epitope into the changes of physicochemical properties

\[
f(E) = N_{\omega_{\mu}} + a_1 N_{\omega_{\mu \sigma}} + a_2 N_{\pi_{\mu}} + a_3 N_{\gamma_{\mu}} + a_4 f_{\lambda_{\mu \sigma}}
\]

Step 3: Integrate the contributions of all derived antigenic epitopes into a single predictor.

\[
d = f(E) + b_1 f(E_2) + b_2 f(E_3) + b_3 f(E_4) + b_4 f(E_5) + b_5 f(E_6) + b_6
\]

Step 4: Parameterize and assess the model

Correlation of Influenza Virus Excess Mortality with Antigenic Variation: Application to Rapid Estimation of Influenza Mortality Burden

Aiping Wu\textsuperscript{1,9}, Yousong Peng\textsuperscript{1,2,9}, Xiangjun Du\textsuperscript{1,2}, Yuelong Shu\textsuperscript{3}, Taijiao Jiang\textsuperscript{1*}
EADpred achieves a good accuracy in quantifying antigenic distances between different viral strains

<table>
<thead>
<tr>
<th>Method</th>
<th>Agreement</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site-dependent*</td>
<td>83.5</td>
<td>80.8</td>
<td>85.4</td>
<td>0.75</td>
</tr>
<tr>
<td>Epitope-Based (EADpred)</td>
<td>86.9</td>
<td>80.8</td>
<td>91.3</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Feature-based Naive Bayes model is a more generalized model to predict antigenic variation.

**Attributes:**
- Groups
  - Epitope A
  - Epitope B
  - Epitope C
  - Epitope D
  - Epitope E
- Hydrophobicity
- Volume
- Charge
- Accessible surface area
- Polarity
- Receptor Binding Inhibition
- Glycosylation

**Training Dataset**
- Smith Dataset (each pair)
- Pairs with mutation <= 9
- Discretion Attributes: Make the positive and negative set most different

**Naive Bayesian**

$$
P(Y_{new} = 1|X_{new}, X, y) = \frac{P(Y_{new} = 1, X_{new}, X, y)}{P(Y_{new} = 1, X, y)}$$

$$P(Y_{new} = 1, X_{new}, X, y) = \prod_{j=1}^{m} P(Y_{new} = 1, X_{new}, X, y)$$

**Prediction**

Retrospective Testing shows that feature-based NB model can model antigenic relationship with high accuracy.
WHO relies on the two critical steps in vaccine strain recommendation

Step 1. Monitor antigenic variants across the world.

Step 2. Assess their potential dominance in the coming season.
The key question 2: how to accurately model influenza antigenic patterns for vaccine recommendation?
Cluster-wise evolution of influenza antigenicity

253 H3N2 viruses during 1968-2002

HA sequencing

HI assay + computational clustering

Antigenic cluster

Continuous genetic changes

Discontinuous antigenic drift

Conventional method to analyze the patterns of influenza evolution: Phylogenetic tree analysis

Trunk-like pattern
Side lineages
Phylogenetic clades
Phylogenetic clades ≠ Antigenic clusters

New computational models are needed to capture the antigenic evolutionary patterns!!!
Based on whole genome information, we have developed a network model to understand influenza evolution.

✓ We developed a network model to capture the co-evolutionary signals across the whole viral genomes.
✓ We used the network to describe human influenza evolutionary patterns.

Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution

Xiangjun Du, Zhuo Wang, Aiping Wu, Lin Song, Yang Cao, Haiying Hang and Taijiao Jiang

Genome Res. 2008 18: 178-187; originally published online Nov 21, 2007;
Access the most recent version at doi:10.1101/gr.6969007
Construction of genomic co-occurrence networks

Step 1. Align each of the eight gene segments

Step 2. Compute the co-occurring nucleotide pairs

Step 3. Connect the co-occurring nucleotide pairs for the viruses involved

Each virus was represented as a nucleotide co-occuring network
The structural changes in co-occurrence network rather than the sequence changes capture the antigenic changes of human H3N2 virus.

Continuous AA changes cannot reflect the clusterwise antigenic changes.

The topology changes of the network are not continuous, capturing the clusterwise antigenic changes of H3N2 viruses.

(Du et al., Genome Research 2008)
Cooperative mutations drive flu antigenic variation.

Antigenic structures of the major antigen HA

Red: Network-captured amino acid changes.
Yellow: Other mutations

Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution
Xiangjun Du, Zhuo Wang, Aiping Wu, Lin Song, Yang Cao, Hailin Hang and Taijiao Jiang
Genome Res. 2008 18: 178-187; originally published online Nov 21, 2007;
Access the most recent version at doi:10.1101/gr.7969007
Predicting antigenic clusters from HA sequences only
PREDAC: PREDictor of Antigenic Clusters

HA sequences of a group of influenza viruses → Predict antigenic relationship → Construct antigenic correlation network → Identify antigenic clusters

PREDAC vividly delineates the antigenic evolution of H3N2 viruses as the replacement of antigenic clusters

1. ~1100 influenza A (H3N2) viruses isolated between 1968 and 2009.

2. 17/20 dominant antigenic clusters, and each persisted for one to five seasons.
Transmission of H3N2 virus in North and South China

A 0.5
B 0.5

Rate of influenza A (H3N2) viruses

Northern China

Southern China
Overview of the talk

I. Backgrounds:

Influenza virus, antigenic evolution and vaccination.

II. Modeling influenza virus antigenic evolution

- Methods for predicting antigenic variants:
  1). EADpred
  2). Feature-based Naive Bayes model

- Methods for predicting antigenic evolution patterns
  1). Genomic co-occurrence network

III. Application of PREDAC to recommend vaccine strains for seasonal flu.
The automatic vaccine strain selection

- New Antigenic cluster
- Percentage >=10%

Patent applied.
Comparison with WHO-recommendations

Our accuracy: 6/7 for China; 6/7 for North America and Europe

Based on 7 seasons during 2002-2009.

WHO: 2/7 for China; 4/7 for North America and Europe

1, Patent obtained.
2, China CDC is using our method in selecting candidate vaccine strains
Predicting H3N2 strains for vaccination

Nature Communications
February 20, 2012

A new computational method that predicts antigenic clusters of the influenza A type virus H3N2 is reported this week in Nature Communications. These findings may aid in the identification of viral strains for use in vaccine production. Influenza vaccines need to be updated every 2-5 years to keep pace with the circulating viral strains that continually change their features. Previous methods have analysed the evolutionary trees of the haemagglutinin antigen gene, which is found in influenza strains. Taijiro Jiao and
Internet-based flu activity predictor

Drawbacks:
1. Can not determine which type or subtype of flu virus in circulation?
2. Can not determine whether the virus has been mutated?

Therefore, the internet-based approach can not be used to recommend vaccine strains and to select drugs for treating the flu.

Integrated approach: combine internet big data and gene big data for fighting flu

Targeted sampling & sequencing

Big data modeling
- Epidemic dynamics
- Pathogen genetic & functional evolution
- Rapid pathogen identification

Acknowledgements

Thank you for attention

Project “973” ; National Key Project for Infectious diseases
NSFC (Outstanding Youth).