Automating Genetic Classification for Hemagglutinin and Neuraminidase Genes from Influenza A Viruses through Machine Learning Methods

M. Zeller¹, T. Anderson², A. Vincent², P. Gauger³

INTRODUCTION
To help control the transmission of influenza A virus (IAV) in swine, surveillance of circulating strains and rapid genetic classification is necessary to monitor viral evolution, detect emerging IAV, and facilitate vaccine antigen selection. Identifying the genetic clade is the initial step required to match circulating field strains with vaccine strains. Consequently, a method that quickly and accurately classifies IAV into genetic clades will allow veterinarians to monitor current IAV circulating in their production systems, rapidly detect new or emerging IAV, and inform strain updates in farm-specific or commercially available vaccines that help protect swine against infection with genetically similar IAV. The objective of this study was to develop an automated method for assigning IAV phylogenetic clade classifications using machine learning methods.

OBJECTIVES
• Automatic genetic clade classification with IAV sequences circulating in swine
• Develop a notification method that detects unique IAV sequences in swine
• Perform quality control methods to ensure accurate genetic classifications

METHODS
• Obtain all IAV swine sequences from ISU FLUture on a daily (nightly) basis
• Separate HA1/HA and NA sequences by subtype and align with MAFFT
• Remove uninformative sites from aligned sequence
• One-hot encode the remaining nucleotide positions
• Train multiclass logistic regression classifier on sequences that have prior clade classification, using L2 regularization (C = 1.0) with scikit-learn library for python (Eq. 1)
• Predict phylogenetic clades with logistic regression classifier (Eq. 3), selecting classification with maximum probability
• Reject classification if no probability ≥ 85%, identify sequences for follow-up
• Update clade designations in the ISU FLUture SQL database
• Validate phylogenetic clade designations by creating phylogenetic trees with prespecified reference sequences using FastTree2
• Confirm clade designations with IRD Swine H1 Clade Classification tool (https://www.fludb.org)

IMPLEMENTING AS A PUBLIC WEB INTERFACE
• Align previously classified set of sequences
• Sparingly select most important genetic features for accurate classification
• Train multiclass classifier, export resultant equations to JavaScript
• Allow user to input sequences, use heuristic alignment to select same features
• Calculate the scores and select max, return to user

RESULTS validated using phylogenetic methods
Consequently, Eq. (1) is the logistic hypothesis function fitted for each clade designation to calculate the log-odds. Eq. (2) The maximum indicator is selected, given that the log-odds probability is > 0.85.

Calculating the scores for IAV swine genetic clade designations. Eq. 1) The binary features xᵢ are nucleic acids and their weights θ, as determined using the liblinear solver with regularization. Eq. 2) The logistic hypothesis function fitted for each clade designation to calculate the log-odds. Eq. 3) The maximum indicator is selected, given that the log-odds probability is > 0.85.

RESULTS
A. Clade Designations
B. Classified + Unknown
C. Classification Results

Fig 1 Using prior classified IAV sequences (n=58) to define the clade of unclassified sequences using logistic regression (n=450). A) Maximum likelihood tree of H1 swine hemagglutinin (HA) sequences with known genetic clades. B) Maximum likelihood tree of H1 HA sequences with known genetic clades combined with HA sequences with unknown genetic clades. C) Maximum likelihood tree of combined known and predicted genetic clades. A single run of logistic regression was used described in the methods was used for clade predictions. Black branches represent where maximum prediction was less than 85%. Additional runs or manual classification are used to follow up branches without a prediction.

Acknowledgments
The Iowa State University Presidential Interdisciplinary Research Initiative for funding.

Neuraminidase Classification Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>N2.1998</th>
<th>N2.2002</th>
<th>N1.classical</th>
<th>N2.hu-2016</th>
<th>N1.pandemic</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2.1998</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N2.2002</td>
<td>2</td>
<td>317</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N1.classical</td>
<td>0</td>
<td>4</td>
<td>210</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>N2.hu-2016</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>N1.pandemic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
</tr>
</tbody>
</table>

Accuracy: 97.9%
Precision: 97.9%
Recall: 97.9%

BLAST TOOL FOR CROSS CHECKING

Fig 4 A beta implementation of BLAST to find HA sequences with identity greater than or equal to 98% on the ISU FLUture website. Currently under development for future use to inform ISU VDL clients, stakeholders and animal health industry.

CONCLUSIONS
• Multiclass logistic regression was able to accurately and quickly classify IAV clades
• Logistic regression was straightforward to implement as a client script
• Quick cross checking was conducted using BLAST tool
• Results validated using phylogenetic methods

ISU FLUture
http://influenza.cvm.iastate.edu