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

M. Zeller[1], T. Anderson[2], A. Vincent [2], P. Gauger[3]

[1]Veterinary Microbiology and Preventive Medicine, Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa;
[2]Virus and Prion Research Unit, National Animal Disease Center, USDA, Agricultural Research Service, Ames, Iowa;
[3]Department of Veterinary Diagnostic and Production Animal Medicine, College of Veterinary Medicine, Iowa State University, Ames, Iowa;

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

- Automate 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 *FLU*ture** on a daily (nightly) basis
- Separate HA1/H3 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. 2-3), selecting classification with maximum probability
- Reject classification if no probability ≥ 85%, identify sequences for follow-up
- Update clade designations in the **ISU *FLU*ture** SQL database
- Validate phylogenetic clade designations by creating phylogenetic trees with predesignated reference sequences using FastTree2
- Confirm clade designations with IRD Swine H1 Clade Classification tool (https://www.fludb.org)

$$\theta^T x = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n \qquad \textit{(Eq. 1)}$$

$$h_\theta^{(i)}(x) = \frac{1}{1 - e^{-\theta^T x}} \qquad \textit{(Eq. 2)}$$

$$y = \begin{cases} \max\limits_i h_\theta^{(i)}(x) & \text{if } \max\limits_i h_\theta^{(i)}(x) \geq 0.85 \\ \text{"other"} & \text{if } \max\limits_i h_\theta^{(i)}(x) < 0.85 \end{cases} \qquad \textit{(Eq. 3)}$$

Equations 1-3. Calculating the scores for IAV swine genetic clade designations. Eq. 1) The binary features x_n as nucleic acids and their weights θ_n 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

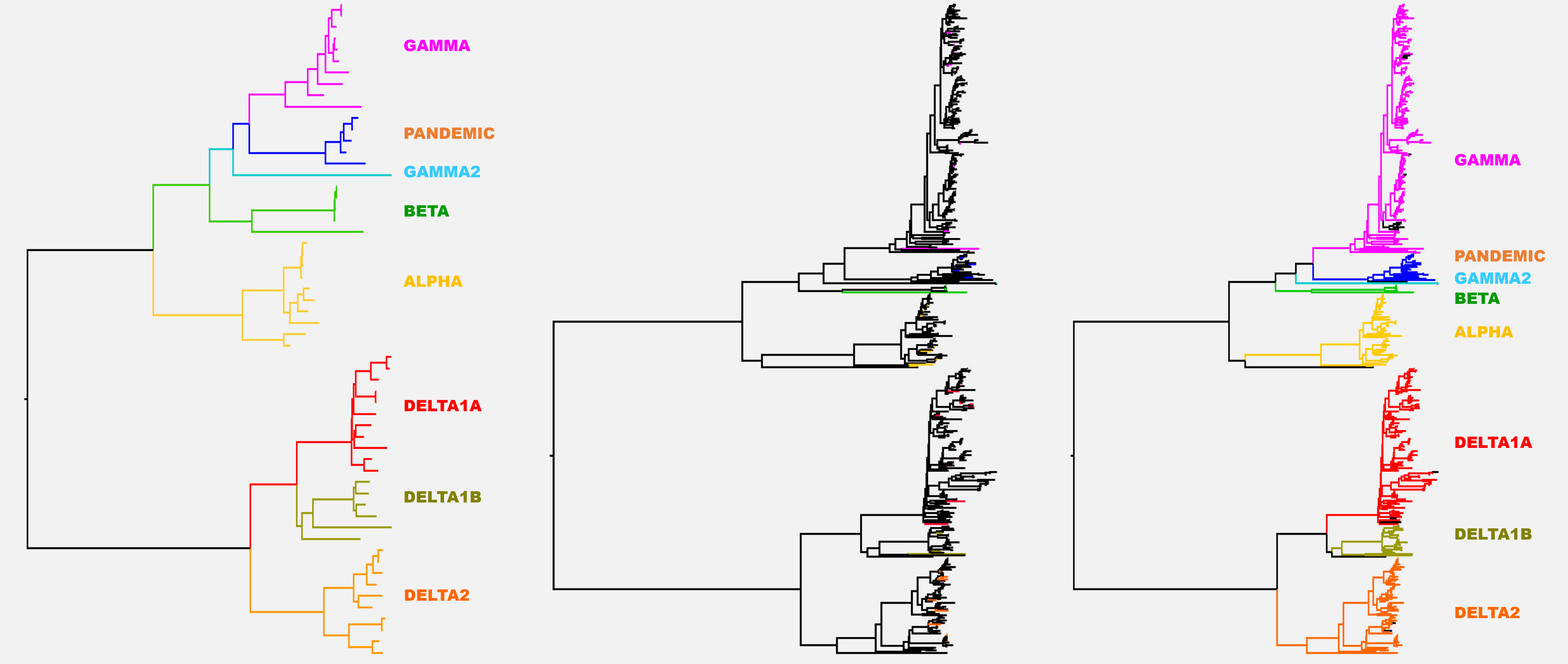

A. Clade Designations **B. Classified + Unknown** **C. Classification Results**

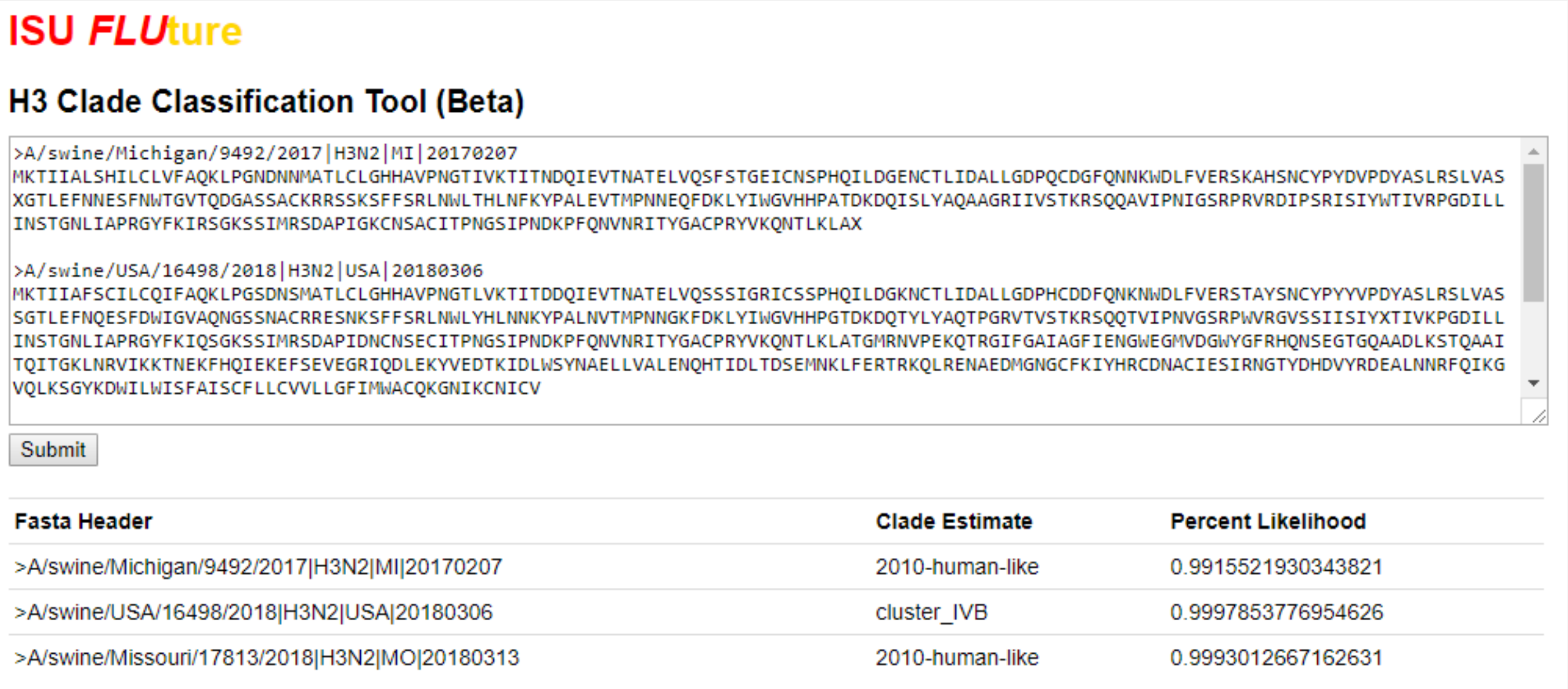
Fig 1 Using prior classified IAV sequences (n=58) to define the clade of unclassified sequences using logistic regression (n=450). A) Maximum likelihood tree of H1 swine hemagglutinin (HA) sequences with known genetic clades. B) Maximum likelihood tree of H1 HA sequences with known genetic clades combined with HA sequences with unknown genetic clades. C) Maximum likelihood tree of combined known and predicted genetic clades. A single run of logistic regression as 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.

Implementing as a Public Web Interface

- Align previously classified set of sequences
- Sparsely 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


ISU *FLU*ture

H3 Clade Classification Tool (Beta)

```
>A/swine/Michigan/9492/2017|H3N2|MI|20170207
MKTIIALSHILCLVFAQKLPGNDNNMATLCLGHHAVPNGTIVKTITNDQIEVTNATELVQSFSTGEICNSPHQILDGENCTLIDALLGDPQCDGFQNNKWDLFVERSKAHSNCYPYDVPDYASLRSLVAS
XGTLEFNNESFNWTGVTQDGASSACKRRSSKSFFSRLNWLTHLNFKYPALEVTMPNNEQFDKLYIWGVHHPATDKDQISLYAQAAGRIIVSTKRSQQAVIPNIGSRPRVRDIPSRISIYWTIVRPGDILL
INSTGNLIAPRGYFKIRSGKSSIMRSDAPIGKCNSACITPNGSIPNDKPFQNVNRITYGACPRYVKQNTLKLAX

>A/swine/USA/16498/2018|H3N2|USA|20180306
MKTIIAFSCILCQIFAQKLPGSDNSMATLCLGHHAVPNGTLVKTITDDQIEVTNATELVQSSSIGRICSSPHQILDGKNCTLIDALLGDPHCDDFQNKNWDLFVERSTAYSNCYPYYVPDYASLRSLVAS
SGTLEFNQESFDWIGVAQNGSSNACRRESNKSFFSRLNWLYHLNNKYPALNVTMPNNGKFDKLYIWGVHHPGTDKDQTYLYAQTPGRVTVSTKRSQQTVIPNVGSRPWVRGVSSIISIYXTIVKPGDILL
INSTGNLIAPRGYFKIQSGKSSIMRSDAPIDNCNSECITPNGSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGTGQAADLKSTQAAI
TQITGKLNRVIKKTNEKFHQIEKEFSEVEGRIQDLEKYVEDTKIDLWSYNAELLVALENQHTIDLTDSEMNKLFERTRKQLRENAEDMGNGCFKIYHRCDNACIESIRNGTYDHDVYRDEALNNRFQIKG
VQLKSGYKDWILWISFAISCFLLCVVLLGFIMWACQKGNIKCNICV
```

Fasta Header	Clade Estimate	Percent Likelihood
>A/swine/Michigan/9492/2017|H3N2|MI|20170207	2010-human-like	0.9915521930343821
>A/swine/USA/16498/2018|H3N2|USA|20180306	cluster_IVB	0.9997853776954626
>A/swine/Missouri/17813/2018|H3N2|MO|20180313	2010-human-like	0.9993012667162631

Fig 3 A beta implementation of exposing the logistic regression classifier on the **ISU *FLU*ture** website.

Neuraminidase Classification Confusion Matrix

	N2.1998	N2.2002	N1.classical	N2.hu-2016	N1.pandemic
N2.1998	67	0	0	0	0
N2.2002	2	317	6	0	0
N1.classical	0	4	210	0	0
N2.hu-2016	0	0	0	1	0
N1.pandemic	0	0	0	0	23

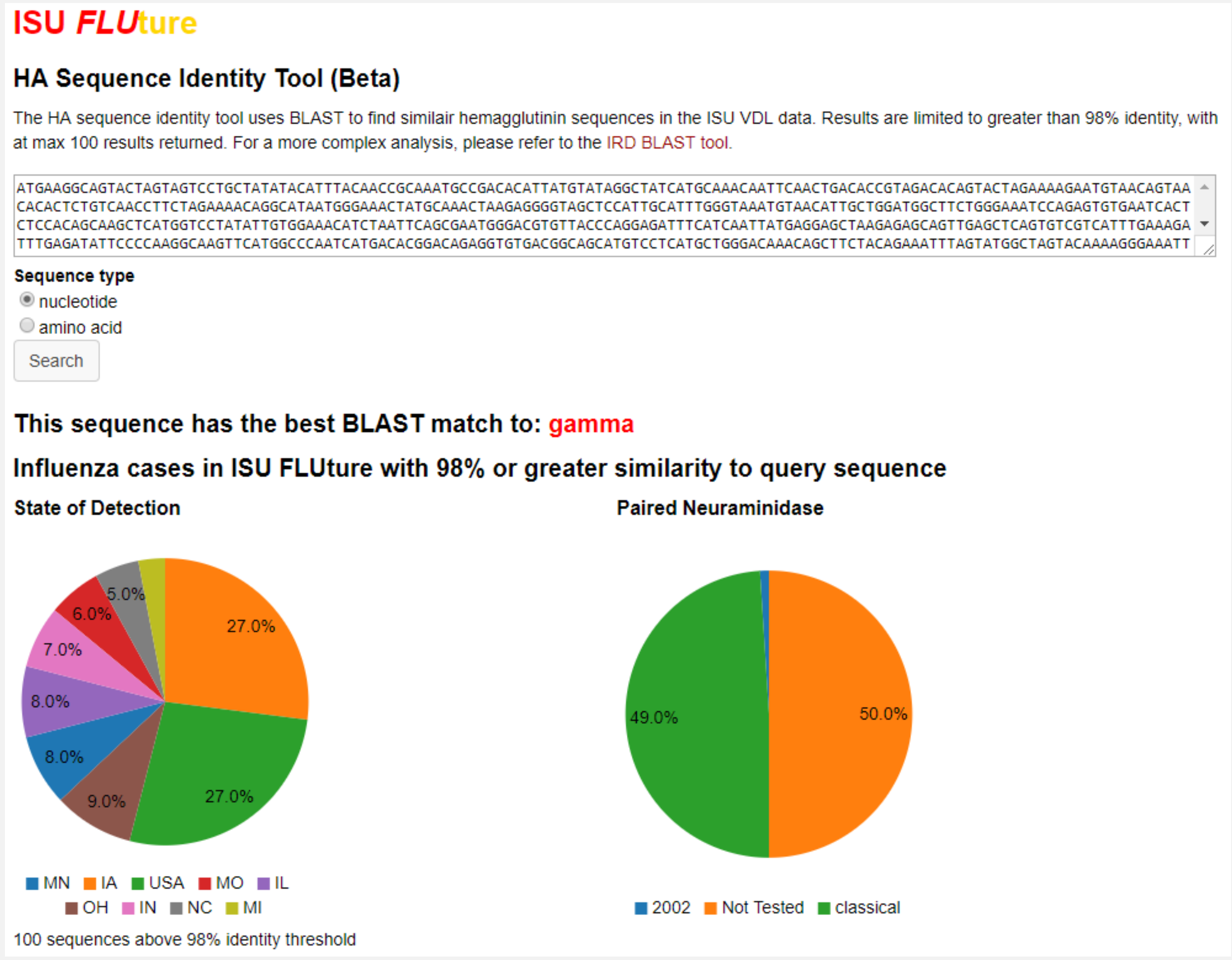
Accuracy: 97.9%

Precision: 97.9%

Recall: 97.9%

Fig 2 A confusion matrix generated from splitting classified neuraminidase sequences into 80% training and 20% testing sets. Columns represent predicted, while rows represent actual classification. Numbers along the diagonal represent correct classification, while numbers outside of the diagonal represent incorrect classification. A total of 12 out of 630 samples were misclassified, highlighted in red.

BLAST Tool for Cross Checking


ISU *FLU*ture

HA Sequence Identity Tool (Beta)

The HA sequence identity tool uses BLAST to find similair hemagglutinin sequences in the ISU VDL data. Results are limited to greater than 98% identity, with at max 100 results returned. For a more complex analysis, please refer to the IRD BLAST tool.

```
ATGAAGGCAGTACTAGTAGTCCTGCTATATACATTTACAACCGCAAATGCCGACACATTATGTATAGGCTATCATGCAAACAATTCAACTGACACCGTAGACACAGTACTAGAAAAGAATGTAACAGTAA
CACACTCTGTCAACCTTCTAGAAAACAGGCATAATGGGAAACTATGCAAACTAAGAGGGGTAGCTCCATTGCATTTGGGTAAATGTAACATTGCTGGATGGCTTCTGGGAAATCCAGAGTGTGAATCACT
CTCCACAGCAAGCTCATGGTCCTATATTGTGGAAACATCTAATTCAGCGAATGGGACGTGTTACCCAGGAGATTTCATCAATTATGAGGAGCTAAGAGAGCAGTTGAGCTCAGTGTCGTCATTTGAAAGA
TTTGAGATATTCCCCAAGGCAAGTTCATGGCCCAATCATGACACGGACAGAGGTGTGACGGCAGCATGTCCTCATGCTGGGACAAACAGCTTCTACAGAAATTTAGTATGGCTAGTACAAAAGGGAAATT
```

Sequence type
● nucleotide
○ amino acid

This sequence has the best BLAST match to: gamma

Influenza cases in ISU FLUture with 98% or greater similarity to query sequence

100 sequences above 98% identity threshold

Fig 4 A beta implementation of BLAST to find HA sequences with identity greater than or equal to 98% on the **ISU *FLU*ture** 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 *FLU*ture

http://influenza.cvm.iastate.edu

Iowa State University

USDA

Agricultural Research Service

NADC National Animal Disease Center