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

Authors: Michael Zeller, Tavis Anderson, Amy Vincent, Phillip Gauger, Iowa State University, Ames, Iowa;

Keywords: influenza A virus, swine, machine-learning, classification

Abstract:

Increased surveillance efforts have expanded the number of swine related Influenza A virus (IAV) genetic sequences collected each year. The volume of sequences has created a problem that requires methodology to classify large amounts of genetic data quickly and accurately into phylogenetic clades. This capability allows veterinarians to monitor current or newly emerging IAV circulating in production systems, and inform strain updates in farm-specific or commercially available vaccines. The objective of this study was to develop an automated method for assigning IAV phylogenetic clade classifications using machine learning methods.

Machine learning methods were applied to classify unknown hemagglutinin (HA) and neuraminidase (NA) sequence data from IAV detected in United States (US) swine to known genetic clades circulating in North America. Training (70%) and cross validation (30%) sets of HA (n = 450) and NA (n=630) were assigned genetic clade labels using maximum-likelihood phylogenetic methods. Labels for each gene segment were assigned based on nearest neighbor identity. A multiclass one-versus-all logistic regression classifier with regularization (C=1.0) was developed with the scikit-learn library. The classifier was fitted using aligned nucleotides as binary features and was used to classify a test dataset. Probabilities < 0.85 were assigned an ‘other’ classification (HA=20, NA=12). The model was evaluated through precision and recall (>0.95). The machine learning classifications were validated against phylogenetically-informed classification with no disagreement in sequences not given an ‘other’ designation.

This automated classifier implemented a machine-learning algorithm and provided rapid and accurate genetic classification of unknown HA and NA sequence data for North American swine IAV. This classifier requires little computational power and it can be further developed into a web interface for public use (http://influenza.cvm.iastate.edu/). This classifier has been integrated into automated pipelines used by the Iowa State Veterinary Diagnostic Lab to ease the burden of classifying large amounts of genetic data.