Machine learning approaches for estimating cross-neutralization potential among FMD serotype O viruses

Running title: *In silico* prediction of r_1 values in FMDVs

Dennis N Makau [a, c#], Jonathan Arzt [b], Kimberly VanderWaal [c]

[a] Department of Biomedical and Diagnostic Sciences, College of Veterinary Medicine, University of Tennessee, Knoxville, USA.

[b] Foreign Animal Disease Research Unit, USDA-ARS, Plum Island Animal Disease Center, Orient Pt., NY 11957, USA

[c] Department of Veterinary Population Medicine, College of Veterinary Medicine, University of Minnesota, USA

#Address correspondence to Dennis N Makau, dmakau@utk.edu

Abstract

In this study, we aimed to develop an algorithm that uses sequence data to estimate cross-neutralization between serotype O foot-and-mouth disease viruses (FMDV) based on r1 values, while identifying key genomic sites associated with high or low r1 values. The ability to estimate cross-neutralization potential among co-circulating FMDVs in silico is significant for vaccine developers, animal health agencies making herd immunization decisions, and disease preparedness. Using published data on virus neutralization titer (VNT) assays and associated VP1 sequences from GenBank, we applied machine learning algorithms (BORUTA and random forest) to predict potential cross-reaction between serum/vaccine-virus pairs for 73 distinct serotype O FMDV strains. Model optimization involved tenfold cross-validation and sub-sampling to address data imbalance and improve performance. Model predictors included amino acid distances, site-wise amino acid polymorphisms, and differences in potential N-glycosylation sites.

The dataset comprised 108 observations (serum-virus pairs) from 73 distinct viruses with r1 values. Observations were dichotomized using a 0.3 threshold, yielding putative non-cross-neutralizing (< 0.3 r1 values) and cross-neutralizing groups (≥ 0.3 r1 values). The best model had a training accuracy, sensitivity, and specificity of 0.96 (95% CI: 0.88-0.99), 0.93, and 0.96, respectively, and an accuracy of 0.94 (95% CI: 0.71-1.00), sensitivity of 1.00, and specificity of 0.93, positive, and negative predictive values of 0.60 and 1.00, respectively, on one testing dataset and an accuracy, AUC, sensitivity, specificity, and predictive values all approaching 1.00 on a second testing dataset. Additionally, amino acid positions 48, 100, 135, 150, and 151 in the VP1 region alongside amino acid distance were found to be important predictors of cross-neutralization.

Our study highlights the value of genetic/genomic data for informing immunization strategies in disease management and understanding potential immune-mediated competition amongst related

endemic strains of serotype O FMDVs in the field. We also showcase leveraging routinely generated sequence data and applying a parsimonious machine learning model to expedite decision-making in selection of vaccine candidates and application of vaccines for controlling FMD, particularly serotype O. A similar approach can be applied to other serotypes.

Key words: cross-protection; r_1 values, cross-reactivity; viral neutralization; machine learning; bioinformatics; computational immunology

Introduction

Foot and mouth disease (FMD) is a viral disease caused by the FMD virus (FMDV), a member of the *Piconaviridae* family (1) that affects cloven-hoofed ungulates. Though typically not fatal, the impact of FMD on food security and livelihoods in endemic countries (particularly low-and middle-income countries (LMICs)) cannot be overemphasized (2). FMD also continues to be a major stumbling block to livestock production and global trade in many parts of the world, especially with the continued threat of emergence and introduction of new FMDV lineages/strains into FMD-free countries. As such, numerous efforts and pathways to achieve and maintain disease free status in most countries has required, among other things, the ability and capacity to identify viruses and vaccinate animals with appropriate vaccines for countries where elimination is yet to be achieved. Although there are 7 documented antigenically distinct serotypes (O, A, C, SAT1, SAT2, SAT3 and Asia1), serotype A and O have been reported to be the most common causes of FMD globally (3–6).

Continued efforts for disease management have included the identification of suitable vaccine candidates and formulation of vaccines applied to the animal populations at risk. To identify these vaccine candidates, a process known as vaccine matching is done, which entails the use of viral neutralization titer (VNT) assays to identify effective cross-reaction and cross-neutralization of viruses by candidate vaccines. The measure of effectiveness of the cross-neutralization and subsequent cross-protection is expressed as the r_1 value and is briefly defined as the ratio between the neutralizing serum titer against the heterologous virus (usually a field strain) to the neutralizing serum titer against the homologous virus (usually the vaccine strain). Subsequently, r_1 values between virus-serum pairs of ≥ 0.3 are considered evidence of cross-neutralization between the two involved viruses, while r_1 values < 0.3 are considered to be non-neutralizing (7).

Generating r_1 data for comparison demands significant effort, precision, and laboratory resources. This involves producing sera in live animals and running *in vitro* assays, to evaluate the cross-reactivity of target viruses with various sera samples. Advances in computational biology empower us to harness machine learning and artificial intelligence to build computation tools aimed at estimating the cross-neutralization potential between different viruses. While previous studies have explored predictive models of FMDV antigenicity, they have varied objectives. Some focused on predicting FMDV antigenicity and identifying antigenicity descriptors (8), identifying and predicting specific epitopes and genomic regions of antigenic importance (9, 10), or developing intricate models to forecast antigenic changes among SAT 1 and 2 FMDV serotypes (11–13). There is hardly any literature specific to serotype O despite it being one of two widely occurring serotypes, nor has research on the application of machine learning to estimate r_1 values been published. Although there have been some divergent views about the adequacy of using r_1

values in identifying potential cross-protection (14), it is still one of the metrics relied upon when interpreting data generated from *in vitro* VNT assays and selecting candidate vaccine strains. Successful application of machine learning to aid in decision making when selecting vaccine candidates and immunization programs has been demonstrated in diseases such as influenza and dengue virus (15–17). Our intention in this study was to develop a simple, yet robust, predictive tool that can be used to estimate potential cross-neutralization between viruses by leveraging machine learning algorithms and genetic characteristics of FMDV.

Therefore, the objective of this study was to develop an algorithm to estimate potential cross-neutralization between strains of serotype O FMDV using r_1 values and identify important genomic sites that influence high or low r_1 values between viruses. The ability to distinguish and estimate the potential for cross-protection between different cocirculating FMD viruses *in silico* will support decision-making by vaccine developers in vaccine candidate matching, animal health agencies/institutions in decision-making for herd immunization practices, and overall disease preparedness and response in cases of emergent strains especially for serotype O about which there is a dearth of information.

Materials and methods

Data preparation

We obtained data from published manuscripts on r_1 values for FMDV serotype O. Viruses utilized in these papers came from 14 countries. These manuscripts were obtained from a search from google and PubMed databases using the search terms 'FMDV and r_1 values' and 'FMDV and vaccine matching' yielding 4 manuscripts (18–21). To be included, the study must have followed standard WHOA (22) methodology for generation of r_1 values for serotype O viruses using virus neutralization tests (VNTs), as well as have associated viral sequence data. The r_1 values indicate the serological compatibility between the vaccine strain and the field isolate, determined by comparing the reactivity of the heterologous versus homologous virus to the antisera produced against the specific viruses. The ratio of the heterologous to homologous neutralization titers between field isolates and vaccine sera serve as a measure of cross-protection.

Since majority of the published studies had uploaded only VP1 data to GenBank, we downloaded sequence data for the VP1 gene of 73 distinct viruses used in the cross-reactivity and generation of r_1 data summarized in the manuscripts described above and any whole genomes available were trimmed to the VP1 region using Muscle in Aliview 1.26 (23). Accession numbers to the sequences used in this analysis are available in supplemental material 1. Using Muscle in Aliview 1.26 (23) we aligned and translated sequences into amino acid sequences. Using MEGA X (24), we calculated Poisson corrected amino acid distance between all pairs (24, 25) and used R packages *stringr* (26), *bioseq* (27), *ape* (28) and *tidysq* (29) to code site-wise differences at any polymorphic amino acid sites using R v4.2 software (30). For each polymorphic amino acid site, serum-virus pairs were tabulated as 0 if they shared the same amino acid and 1 if they had a different amino acid at that site. We also scanned the genome for sites of potential N-glycosylation (31, 32) in the VP1 region using a custom-built R script. Pairs with identical sets of inferred glycosylated sites were coded as 0 and non-identical sets as 1. The r_1 values, as reported in the manuscripts, amino acid distance calculated in MEGA X, and amino acid site-wise differences were concatenated into one data frame, resulting in 108 observations (serum-virus pairs). Although some studies on influenza have explored the pros and cons of weighting of specific amino acid sites based on *a*

priori knowledge of their evolutionary history and influence on antigenicity (33, 34), our modeling approach was naïve to *a priori* assumptions about the relative importance of specific site-wise amino acid differences and thus no weighting was used (e.g., I→N ≡ N→I) (15).

Initial analysis involved descriptive statistics of the data and correlation analysis. We performed a phylogenetic analysis of the sequence data by constructing a bootstrapped maximum likelihood tree in RAxMl with 500 bootstraps to depict the distribution/representation of the different topotypes in our data and to visualize relatedness of the serum/vaccine candidates cross-reacted in VNT assays to generate r_1 values. Additionally, we performed a Spearman's correlation analysis between p-amino acid distance and r1 scores. The r_1 values were then dichotomized at 0.3, which is the threshold recommended by WHOA (22).

Training and testing machine learning models

Using a stepwise approach, we developed a random forest classification model with sub-sampling steps and tenfold cross validation to achieve the best model performance. Upon dichotomization of the r_1 values into binomial data, there was an imbalance in the 0 (non-cross-reacting) vs 1 (cross-reacting) classes with more than twice as many observations in the 0 category as those in the 1 category (0 = 87 observations, 1 = 21 observations). As such, there was need to subsample the data to adjust for this imbalance of the outcome in the training dataset; we thus used the synthetic minority oversampling technique (SMOTE) which statistically increased the number of cases in the dataset to balance the distribution of cases vs non-cases (1 vs 0) in the outcome to achieve better model performance. Feature-reduction was then implemented to reduce the number of model features (i.e., predictors) using the Boruta algorithm to optimize model fitting. Upon concatenating all parts of the data into a single data frame, the final data frame had 216 columns (amino acid distance and 214 site-wise amino acid differences and potential sites for N-glycosylation). With highly dimensional data, machine learning models may struggle to achieve good model accuracy and prediction due to noise introduced as the model tries to optimize the estimated contribution of each model feature to the variation in the outcome (r_1 class in our study). As such, the Boruta algorithm has been proposed as a way of eliminating correlated and redundant features from the model, thus reducing the number of model features needed in the final model and optimizing model performance. After parsing the data through Boruta, we included all model features classified as important or tentatively important, resulting in a data frame with 35 model features. These features were VP1 amino acid distance and site-wise differences in amino acid positions (4, 13, 24, 32, 33, 43, 48, 49, 57, 69, 97, 100, 124, 135, 139-141, 143, 145, 149-151, 154, 156, 159, 166, 173, 175, 195, 198, 199, 210, 213, 214) and potential N-glycosylation profile.

Subsequently we used the 80:20 split rule to randomly split the data into three subsets to generate one training and two test datasets with 70, 17 and 21 observations each. Using *caret* (35) and *randomForest* (36) packages in R for model building, training and evaluation, including hyper-parameter tuning and 10-fold cross-validation, we trained the model on500 trees, tested it, and made predictions as summarized in the results section. We based model performance on accuracy (overall percent of observations correctly classified as cross-reacting/non-cross-reacting), sensitivity (percent of high observations correctly classified), specificity (percent of low observations correctly classified), and predictive values (proportion (%) of times the classification (non-cross-protecting-0 vs cross-protecting-1) is the true r_1 group class.

Additionally, from the random forest model, model features were ranked in their importance to the performance of our model using mean decrease in model accuracy when a features' data were

randomized relative to the outcome (the relative prediction strength of a variable) and improvement in Gini index (measure of node impurity associated with a variable) when data were split on a variable (37, 38). This allowed us to highlight highly ranked amino acids in our VP1data, and considered the relative importance of different amino acid sites based on their role in improving model accuracy and node purity in outcome classification (15, 39).

To test the performance of the machine learning algorithm's predictions on a completely external set of sequences that have been reported to be antigenically novel, we used data from Bachanek-Bankowska et al., (40). In this study, they reported the isolation of three antigenically distinct viruses isolated from outbreaks in Pakistan in 2016 and 2017. Bachanek-Bankowska et al., described three isolates (GenBank accession numbers MH784403, MH784404, and MH784405) from a single genetic sublineage displaying distinct antigenic phenotypes against three commonly used vaccines in the region (O 3039 and O Manisa [Boehringer Ingelheim] and O TUR 5/2009 [MSD]).

Results

Descriptive analysis

In this dataset, there were108 observations (serum-virus pairs) from 73 distinct viruses for which r_1 values had been reported. In summary, the mean r_1 value was 0.22 ± 0.23 (SD), with a range of (0.0-0.94) and a median of 0.16. Upon dichotomizing the data using the 0.3 threshold as cutoff, the two resulting groups used in subsequent modelling comprised of 87 pairs in the non-cross-neutralizing group (< 0.3 r_1 values) and 21 in the cross-neutralizing group (≥ 0.3 r_1 values).

Additionally, upon phylogenetic assessment, our dataset included viruses belonging to five topotypes for FMDV serotype O and the serum/vaccines used in the assays were evenly distributed among the topotype groupings (Figure 1). The mean VP1 p-amino acid distance for the pairs was 0.12 ± 0.03 with a range of 0.04 – 0.15 and median of 0.13. There was a significant (p=0.0001) moderate negative correlation (Spearman rho = -0.4) between amino acid distance and r_1 values between the pairs (Figure 2).

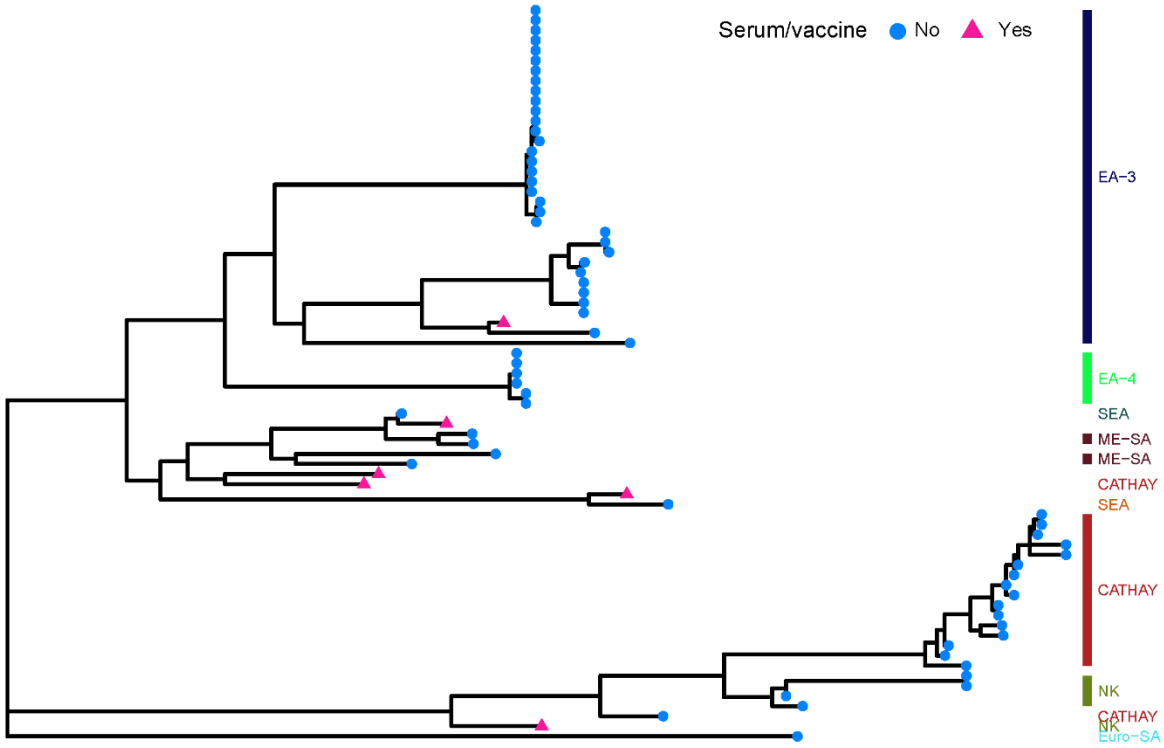

Figure 1: Maximum likelihood tree depicting the phylogenetic distribution of 73 foot and mouth viruses (serum, vaccine (pink triangles) and field strains (blue circles)) representing serotype O topotypes used in the development of an r_1 predictive model. The colored bars on the right of the tree represent the different topotypes included in the data (*NK=Not Known*).

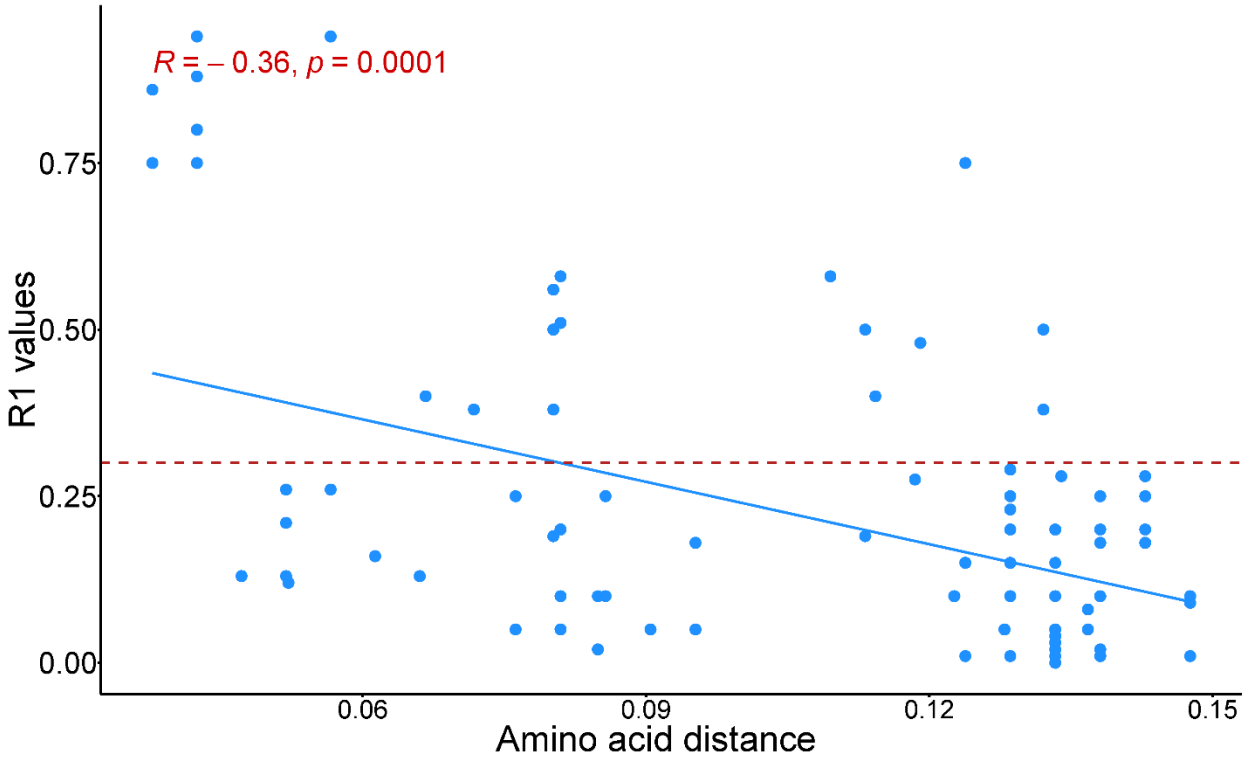

Figure 2: Correlation plot depicting an inverse moderate Spearman's correlation (blue) between amino acid distance of the VP1 region for serotype O foot and mouth disease viruses and r_1 values obtained from *in vitro* virus neutralization assays. The red-dotted line represents the cut-off of 0.3 threshold for potentials cross-neutralization between serum-virus pairs.

Random forest model performance and predictions

The best performing model had an accuracy of 0.96 (95% CI 0.88-0.99), AUC of 0.95, sensitivity and specificity of 0.93 and 0.96, and negative and positive predictive values of 0.98 and 0.87, respectively on training data (n=70: 0=56, 1=14). On one testing dataset (n=17: 0=14,1=3), the model accuracy was 0.88 (95% CI 0.64-0.99), AUC of 0.93, sensitivity of 1.00 and specificity of 0.86, and positive and negative predictive values of 0.60 and 1.00, respectively. The model performance on the second test/validation data (n=21: 0=17, 1=4) was 1.00 accuracy (95% CI 0.84-1.00), AUC of 1.00, sensitivity and specificity of 1.00 and 1.00, and positive and negative predictive values of 1.00 and 1.00, respectively.

From the variable importance analysis that extracted which predictors had the most influence in the models' predictive ability and accuracy, several sites were ranked highly besides the amino acid distance. These include amino acid positions 48, 100, 135, 150, and 151 in the VP1 region (Figure 3).

When the trained model was applied to the antigenically novel sequences described by Bachanek-Bankowska (40), the model correctly predicted the antigenic relationship of one out of three antigenically distinct viruses, i.e., between MH784405.1 (PAK/14/2017) and commonly used vaccines in the region. The model however incorrectly predicted that commonly used vaccines in

the regions were good matches for both MH784403.1 (PAK/10/2016) and MH784404.1 (PAK/4/2017) ($r_1 \geq 0.3$). The *in vitro* vaccine matching experiment had indicated non-cross-neutralization between the two outbreak viruses and commonly used vaccines in this region (Supplementary table 1).

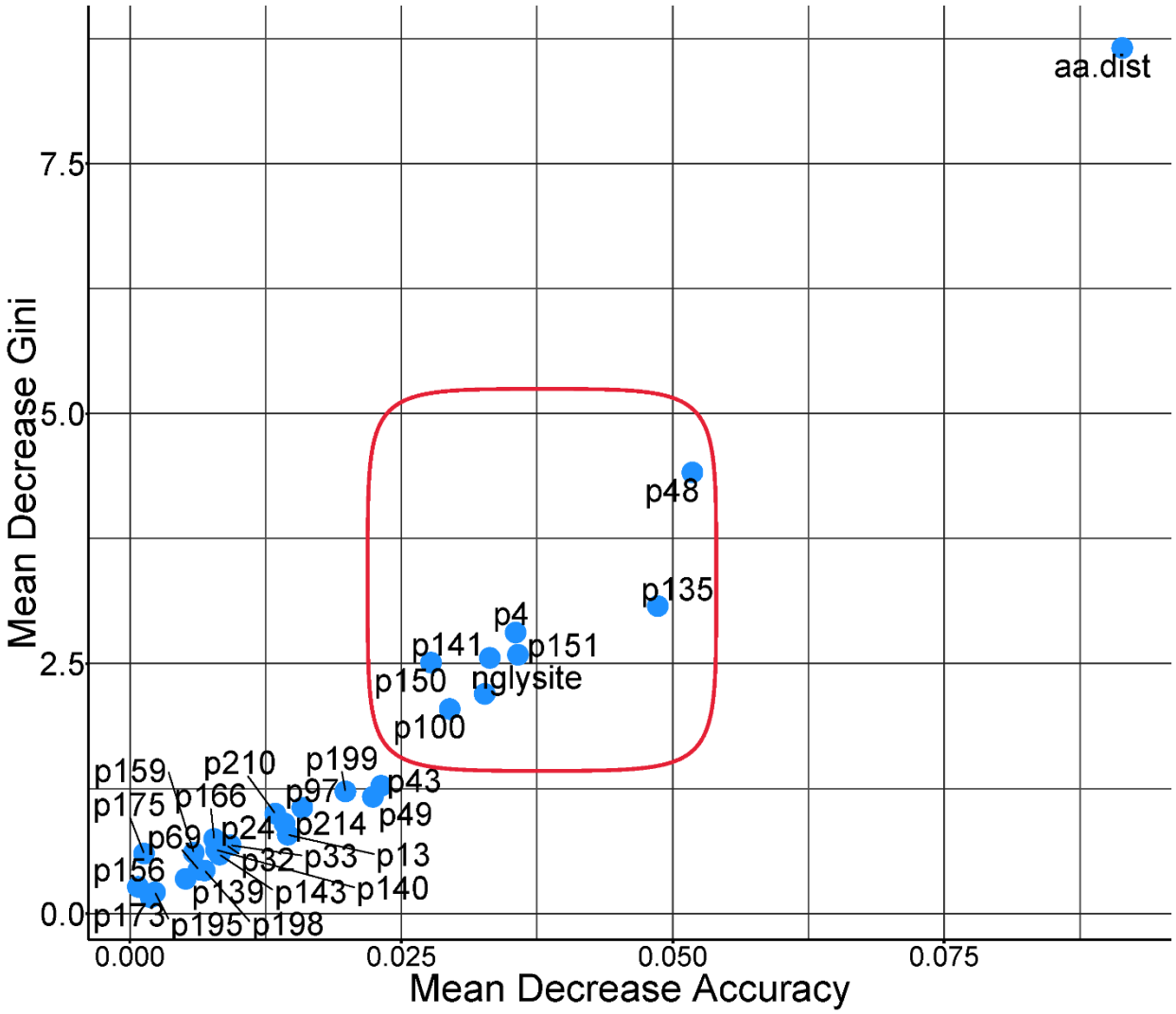

Figure 3: Multiway importance plot for model covariates included in the final random forest model predicting r_1 values in VP1 region for serotype O foot and mouth disease virus. Feature labels in the graph indicate the specific amino acid site (p) while 'aa.dist' indicates the amino acid distance between serum-virus pairs, 'nglysite' similarity or difference in positions for potential N-glycosylation.

Discussion

We developed a parsimonious random forest classification model that estimates potential cross-neutralization between serum and viruses i.e., r_1 values for viruses belonging to serotype O FMDV with an accuracy of more than 85%. This model adds to a growing body of research and efforts towards leveraging bioinformatic data to streamline and enhance our understanding on viral antigenicity and host immune response as exemplified in diseases like influenza (16, 17, 41–44). diseases like influenza. Optimized and robust predictive models can be valuable in decision making and implementing strategic interventions towards the control, and ultimately eradication, of FMDV. This model requires further training and refinement with diverse datasets because when applied to the antigenically novel sequences described by Bachanek-Bankowska (40), the model correctly predicted the antigenic relationship of only one of three novel viruses.

The suitability of using *in vitro* methods to estimate *in vivo* cross-protection in FMD (45), r_1 values and 0.3 thresholds compared to other measures of immune response and cross reactivity has been argued by several studies (14, 45–48). However, though this model is based on r_1 values it can be modified to accommodate other measures of antigenic variability like raw VNT results. Since r_1 values are still the most-used option when conducting vaccine matching and selection of candidates, whether achieved through simple ELISA or a modified combination of techniques to estimate cross-protection, our model is well trained to aid in the identification of potential cross-neutralizing and non-cross-neutralizing sera thus streamlining the selection of potential vaccine candidates and facilitating immediate comparisons and of field strains where necessary. Moreover, when applied in non-research settings, animal health agencies/institutions can easily upload sequences to the model from isolates or clinical samples and promptly identify potential options for immunization. The continued improvement and accessibility of sequencing technology such as Minion Nanopore sequencing, which can be deployed in LMIC settings, may also increase the frequency and capacity of in-country generation of sequences (especially in countries where access to reference labs is limited) complementing the utility of predictive tools supported by this model.

By highlighting the highly ranked/important amino acids in the VP1 region, we posit that those amino acids influence host immune induction and response as differences in those sites appear to be important for estimating the r_1 values between serum-virus pairs. This supports findings by other studies (10, 12) that certain immunodominant sites exist in the VP1 region. To better understand challenges with vaccine effectiveness, causes of vaccine failure and breakthrough outbreaks in vaccinated populations as well as the mechanisms of emergence and spread of new viruses, dynamics of mutations in these sites can be investigated further (49). In these data, seven amino acid positions were most influential in the ability of our model to accurately distinguish between cross-neutralizing ($\geq$0.3) vs non-cross-neutralizing (<0.3). These were amino acids in positions 4, 48, 100, 135, 141, 150 and 151. Except position 4 and 100, all these amino acids are part of immunogenically important regions described as the B-C and G-H loops in the VP1 protein of serotype O (6) in which certain mutations in combination or independently have been thought to confer serological heterogeneity among FMDVs (50). In sequences used for this study, N-glycosylation was predicted to occur most often at sites 86, then 101, and least often at site 132.

Even with improved predictive ability to accurately classify cross-neutralizing sera/vaccine as possible, there are drawbacks of *in silico* modelling for biological processes. The immune response is a composite process which commonly involves the interplay of multiple compartments of the host immune response system and different factors (e.g., environment and host genetics) that may influence host-virus interactions. Such factors contribute to why measures of cross-reactivity *in vitro* often do not translate well to *in vivo* cross-protection in the field. Our model does not account

for these factors. Nevertheless, with current data, we were able to accurately categorize 8 to 9 out of every 10 serum-virus pairs for serotype O as cross-neutralizing or non-cross-neutralizing (based on r_1 values). However, the wider confidence interval suggests that there is some misclassification and misidentification of sera/vaccine candidates.

As such, continued improvement of *in silico* models is necessary both with diverse datasets and potential validation with *in vivo*/field application data as attempted in this study. That limitation notwithstanding, with the limited data used here, the model can at least aid in filtering potential candidates to be considered for in-depth *in vitro*/*in vivo* assays or blends in vaccination programs which would save resources and increase efficiency of the decision-making process and FMD management.

Lastly, other VP regions in the P1 portion of the genome also likely play a role in host-virus interaction and immune response. However, since we only relied on publicly available secondary data, we did not have access to full genome sequences for all 73 viruses included in the study hence the need to restrict the analysis to the role of VP1 in estimating r_1 values. Whole genomic data could improve our model. Future efforts will explore the benefits of a broader genomic analysis and more sensitive threshold of cross-protection.

Conclusion

Ultimately, this study adds to the growing body of literature that machine learning can be applied to genetic/genomic data to achieve a more nuanced analysis of the relationship between genetic variability and cross-recognition of viruses by host immune systems. Such models can support immunization as a pathway to disease management. In this study, we demonstrate the opportunities and potential for leveraging routinely generated sequence data and the application of a parsimonious machine learning model to streamline the process of decision-making in vaccine development and application to control FMD, especially serotype O (but also applicable to other serotypes). Specifically, by obtaining an accurate model with high sensitivity and specificity, appropriate vaccine candidates may be able to be selected more quickly, although *in vivo* experiments would still ultimately be necessary to assess cross-protection. Also, this capability would enable tailoring immunization protocols for use in the field with a faster turnaround time for decision-making. Lastly, outputs from such a model can be combined with other mathematical models to understand drivers and trends of viral emergence, especially the role of immune pressure in driving the evolution and spread of FMDV. The latest version of the r_1 predictive model is available for access via a Shiny dashboard (https://dmakau.shinyapps.io/PredImmune-FMD/).

Data availability. The nucleotide sequences of the FMDV used in this analysis are already publicly available on GenBank and the list provided in Supplemental material 1 can help in identification and downloading of the sequences if needed.

Supplemental material

All supplementary materials have been provided in one file.

Acknowledgements

This project was supported by the USDA Agricultural Research Service, award 58-8064-2-006.

The authors declare there are no conflicts of interest.

References

1. Zell R, Delwart E, Gorbalenya AE, Hovi T, King AMQ, Knowles NJ, Lindberg AM, Pallansch MA, Palmenberg AC, Reuter G, Simmonds P, Skern T, Stanway G, Yamashita T. 2017. ICTV virus taxonomy profile: Picornaviridae. J Gen Virol 98:2421–2422.

2. Knight-Jones TJD, McLaws M, Rushton J. 2017. Foot-and-Mouth Disease Impact on Smallholders - What Do We Know, What Don't We Know and How Can We Find Out More? Transbound Emerg Dis 64:1079.

3. He Y, Li K, Cao Y, Sun Z, Li P, Bao H, Wang S, Zhu G, Bai X, Sun P, Liu X, Yang C, Liu Z, Lu Z, Rao Z, Lou Z. 2021. Structures of Foot-and-mouth disease virus with neutralizing antibodies derived from recovered natural host reveal a mechanism for cross-serotype neutralization. PLoS Pathog 17:1–20.

4. Brito BP, Rodriguez LL, Hammond JM, Pinto J, Perez AM. 2017. Review of the Global Distribution of Foot-and-Mouth Disease Virus from 2007 to 2014. Transbound Emerg Dis 64:316–332.

5. Grubman MJ, Baxt B. 2004. Foot-and-mouth disease. Clin Microbiol Rev 17:465–493.

6. Ranaweera LT, Wijesundara UK, Jayarathne HSM, Knowles N, Wadsworth J, Mioulet V, Adikari J, Weebadde C, Sooriyapathirana SS. 2019. Characterization of the FMDV-serotype-O isolates collected during 1962 and 1997 discloses new topotypes, CEY-1 and WCSA-1, and six new lineages. Sci Reports 2019 91 9:1–10.

7. OIE/FAO Foot-and-Mouth Disease Reference Laboratories Network. 2022. Foot-and-Mouth Disease European July-September 2022 Quarterly report.

8. Qiu J, Qiu T, Dong Q, Xu D, Wang X, Zhang Q, Pan J, Liu Q. 2021. Predicting the Antigenic Relationship of Foot-and-Mouth Disease Virus for Vaccine Selection through a Computational Model. IEEE/ACM Trans Comput Biol Bioinforma 18:677–685.

9. Borley DW, Mahapatra M, Paton DJ, Esnouf RM, Stuart DI, Fry EE. 2013. Evaluation and Use of In-Silico Structure-Based Epitope Prediction with Foot-and-Mouth Disease Virus. PLoS One 8.

10. Bari FD, Parida S, Asfor AS, Haydon DT, Reeve R, Paton DJ, Mahapatra M. 2015. Prediction and characterization of novel epitopes of serotype A foot-and-mouth disease viruses circulating in East Africa using site-directed mutagenesis. J Gen Virol 96:1033–1041.

11. Rahman T, Mahapatra M, Laing E, Jin Y. 2015. Evolutionary non-linear modelling for selecting vaccines against antigenically variable viruses. Bioinformatics 31:834–840.

12. Davies V, Reeve R, Harvey WT, Maree FF, Husmeier D. 2017. A sparse hierarchical Bayesian model for detecting relevant antigenic sites in virus evolution. Comput Stat 32:803–843.

13. Reeve R, Blignaut B, Esterhuysen JJ, Opperman P, Matthews L, Fry EE, de Beer TAP, Theron J, Rieder E, Vosloo W, O'Neill HG, Haydon DT, Maree FF. 2010. Sequence-

Based Prediction for Vaccine Strain Selection and Identification of Antigenic Variability in Foot-and-Mouth Disease Virus. PLOS Comput Biol 6:e1001027.

14. Brito BP, Perez AM, Capozzo AV. 2014. Accuracy of traditional and novel serology tests for predicting cross-protection in foot-and-mouth disease vaccinated cattle. Vaccine 32:433–436.

15. Zeller MA, Gauger PC, Arendsee ZW, Souza CK, Vincent AL, Anderson TK. 2021. Machine Learning Prediction and Experimental Validation of Antigenic Drift in H3 Influenza A Viruses in Swine. mSphere 6.

16. Bell SM, Katzelnick L, Bedford T. 2019. Dengue genetic divergence generates within-serotype antigenic variation, but serotypes dominate evolutionary dynamics. Elife 8.

17. Mansfield KL, Horton DL, Johnson N, Li L, Barrett ADT, Smith DJ, Galbraith SE, Solomon T, Fooks AR. 2011. Flavivirus-induced antibody cross-reactivity. J Gen Virol 92:2821–2829.

18. Maree FF, Blignaut B, Esterhuysen JJ, de Beer TAP, Theron J, O'Neill HG, Rieder E. 2011. Predicting antigenic sites on the foot-and-mouth disease virus capsid of the South African Territories types using virus neutralization data. J Gen Virol 92:2297–2309.

19. Tesfaye Y, Khan F, Yami M, Wadsworth J, Knowles NJ, King DP, Gelaye E. 2020. A vaccine-matching assessment of different genetic variants of serotype O foot-and-mouth disease virus isolated in Ethiopia between 2011 and 2014. Arch Virol 165:1749–1757.

20. Upadhyaya S, Mahapatra M, Mioulet V, Parida S. 2021. Molecular Basis of Antigenic Drift in Serotype O Foot-and-Mouth Disease Viruses (2013-2018) from Southeast Asia. Viruses 13.

21. Yang M, Xu W, Goolia M, Zhang Z. 2014. Characterization of monoclonal antibodies against foot-and-mouth disease virus serotype O and application in identification of antigenic variation in relation to vaccine strain selection. Virol J 11:136.

22. World Organization for Animal Health. 2022. Foot and Mouth Disease (Infection with Foot and Mouth Disease Virus), p. 1–34. *In* OIE Terrestrial Manual.

23. Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276–3278.

24. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35:1547–1549.

25. Zuckerkandl E, Pauling L. 1965. Evolutionary Divergence and Convergence in Proteins, p. 97–166. *In* Evolving Genes and Proteins. Elsevier.

26. CRAN - Package stringr. https://stringr.tidyverse.org/authors.html. Retrieved 17 June 2022.

27. Keck F. 2020. Handling biological sequences in R with the bioseq package. Methods Ecol Evol 11:1728–1732.

28. Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and

397 evolutionary analyses in R. Bioinformatics 35:526–528.

398 29. CRAN - Package tidysq. https://cran.rstudio.com/web/packages/tidysq/index.html.
399 Retrieved 17 June 2022.

400 30. R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna,
401 Austria.

402 31. Chirkova Z V., Filimonov SI, Prituzhalov I V., Vasanov EA, Abramov IG. 2017.
403 Synthesis of chalcones from 3-formyl-substituted pyrrolo[3,4-f]indole-5,7-diones. Russ
404 Chem Bull 66:882–885.

405 32. Paploski IAD, Makau DN, Pamornchainavakul N, Baker JP, Schroeder D, Rovira A,
406 VanderWaal K. 2022. Potential Novel N-Glycosylation Patterns Associated with the
407 Emergence of New Genetic Variants of PRRSV-2 in the U.S. Vaccines 10:2021.

408 33. Yao Y, Li X, Liao B, Huang L, He P, Wang F, Yang J, Sun H, Zhao Y, Yang J. 2017.
409 Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint
410 random forest method. Sci Rep 7:1545.

411 34. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell
412 CA, Smith DJ, Rambaut A. 2014. Integrating influenza antigenic dynamics with
413 molecular evolution. Elife 2014.

414 35. Kuhn M. 2008. Building Predictive Models in R Using the caret Package. J Stat Softw
415 28:1–26.

416 36. Liaw A, Wiener M. 2002. Classification and Regression by randomForest. R News 2:18–
417 22.

418 37. Hastie T, Tibshirani R, Friedman J. 2009. Random Forests, p. 587–604. *In* Springer (ed.),
419 The Elements of Statistical LearningSecond. Springer Series in Statistics, New York, NY.

420 38. Machado G, Mendoza MR, Corbellini LG. 2015. What variables are important in
421 predicting bovine viral diarrhea virus? A random forest approach. Vet Res 46:1–15.

422 39. Makau DN, Prieto C, Martínez-Lobo FJ, Paploski IAD, VanderWaal K. 2023. Predicting
423 Antigenic Distance from Genetic Data for PRRSV-Type 1: Applications of Machine
424 Learning. Microbiol Spectr 11.

425 40. Bachanek-Bankowska K, Wadsworth J, Henry E, Ludi AB, Bin-Tarif A, Statham B, King
426 DP, Afzal M, Hussain M, Manzoor S, Abubakar M, Knowles NJ. 2019. Genome
427 Sequences of Antigenically Distinct Serotype O Foot-and-Mouth Disease Viruses from
428 Pakistan. Microbiol Resour Announc 8.

429 41. Neher RA, Russell CA, Shraiman BI. 2014. Predicting evolution from the shape of
430 genealogical trees. Elife 3.

431 42. Xia Y-L, Li W, Li Y, Ji X-L, Fu Y-X, Liu S-Q. 2021. A Deep Learning Approach for
432 Predicting Antigenic Variation of Influenza A H3N2. Comput Math Methods Med
433 2021:1–10.

434 43. Forghani M, Khachay M. 2020. Convolutional Neural Network Based Approach to In

435 Silico Non-Anticipating Prediction of Antigenic Distance for Influenza Virus. Viruses
436 12:1019.

437 44. Lee EK, Tian H, Nakaya HI. 2020. Antigenicity prediction and vaccine recommendation
438 of human influenza virus A (H3N2) using convolutional neural networks. Hum Vaccin
439 Immunother 16:2690–2708.

440 45. Paton DJ, Reeve R, Capozzo A V., Ludi A. 2019. Estimating the protection afforded by
441 foot-and-mouth disease vaccines in the laboratory. Vaccine 37:5515–5524.

442 46. Lavoria M ángeles, Di-Giacomo S, Bucafusco D, Franco-Mahecha OL, Pérez-Filgueira
443 DM, Capozzo AV. 2012. Avidity and subtyping of specific antibodies applied to the
444 indirect assessment of heterologous protection against Foot-and-Mouth Disease Virus in
445 cattle. Vaccine 30:6845–6850.

446 47. Robiolo B, La Torre J, Maradei E, Beascoechea CP, Perez A, Seki C, Smitsaart E,
447 Fondevila N, Palma E, Goris N, De Clercq K, Mattion N. 2010. Confidence in indirect
448 assessment of foot-and-mouth disease vaccine potency and vaccine matching carried out
449 by liquid phase ELISA and virus neutralization tests. Vaccine 28:6235–6241.

450 48. Mattion N, Goris N, Willems T, Robiolo B, Maradei E, Beascoechea CP, Perez A,
451 Smitsaart E, Fondevila N, Palma E, De Clercq K, La Torre J. 2009. Some guidelines for
452 determining foot-and-mouth disease vaccine strain matching by serology. Vaccine
453 27:741–747.

454 49. Makau DN, Lycett S, Michalska-Smith M, Paploski IAD, Cheeran MC-J, Craft ME, Kao
455 RR, Schroeder DC, Doeschl-Wilson A, VanderWaal K. 2022. Ecological and evolutionary
456 dynamics of multi-strain RNA viruses. Nat Ecol Evol 1–9.

457 50. Islam MR, Rahman MS, Amin M Al, Alam ASMRU, Siddique MA, Sultana M, Hossain
458 MA. 2021. Evidence of combined effect of amino acid substitutions within G-H and B-C
459 loops of VP1 conferring serological heterogeneity in foot-and-mouth disease virus
460 serotype A. Transbound Emerg Dis 68:375–384.

461

462