

AMERICAN UNIVERSITY OF BEIRUT

ESTIMATING COMPONENTS OF CROSS-IMMUNITY
BETWEEN INFLUENZA H3N2 STRAINS BASED ON
HEMAGGLUTININ SEQUENCE DATA

by
AYBAK SAMIR HASSAN

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Arts
to the Department of Biology
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
June 2016

AMERICAN UNIVERSITY OF BEIRUT

ESTIMATING COMPONENTS OF CROSS-IMMUNITY
BETWEEN INFLUENZA H3N2 STRAINS BASED ON
HEMAGGLUTININ SEQUENCE DATA

by
AYBAK SAMIR HASSAN

Approved by:




Dr. Heinrich zu Dohna, Assistant Professor
Department of Biology

Advisor



Dr. Mike Osta, Associate Professor
Department of Biology

Member of Committee



Dr. Hassan Zaraket, Assistant Professor
Department of Experimental Pathology, Immunology
& Microbiology

Member of Committee

Date of thesis defense: June 13th, 2016

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DESSERTATION, PROJECT RELEASE FORM

Student Name: _____

Last

First

Middle

Master's Thesis

Master's Project

Doctoral Dissertation

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, **three years after the date of submitting my thesis, dissertation, or project**, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

Signature

Date

AN ABSTRACT OF THE THESIS OF

Aybak Samir Hassan for Master of Science
Major: Biology

Title: Estimating components of cross-immunity between influenza H3N2 strains based on hemagglutinin sequence data

Global influenza epidemics cause thousands of deaths annually. Vaccination campaigns are an important tool to mitigate effects of influenza epidemics. Because of the fast evolution of the influenza virus, selecting a vaccine strain that confers protection against the main circulating strain remains a key challenge. The WHO uses hemagglutination inhibition (HI) assays, where anti-sera against one strain (serum strain) is set against another strain (virus strain), to determine which strain from a set of vaccine-candidate strains confers best protection against the dominant circulating strains. Even though several studies have shown that cross-immunity can be predicted from hemagglutinin (HA) sequences, current techniques still use HI assays. One shortcoming of sequence-based predictions is that they rely only on the differences between two HA sequences, and thus assume symmetry in cross immunity. Our study introduces a method for sequence-based prediction of cross-immunity that relaxes the symmetry assumption. In our method, each amino acid of the virus strain HA, each amino acid of the serum strain HA and each amino acid difference between the virus and serum strain HA were included as a potential predictor variable and log-transformed HI titers were used as response variable. Regression coefficients were estimated via elastic net regression with cross-validation. The data was split in a training and validation set. The training set was used to estimate regression coefficients and the validation set was used to predict HI titers based on estimated coefficients and compare them to the actual HI titer values. The coefficients for the correlations between estimated and actual HI titers were 0.72 and 0.67 for training and validation sets, respectively. Most amino acid positions that received non-zero regression coefficients fell within the epitope regions or were in close proximity to those regions on the 3D structure of the HA protein. Our results suggest that the proposed model can predict HI titers and find antigenically important positions on the HA protein.

CONTENTS

ABSTRACT	v
LIST OF ILLUSTRATIONS	viii
LIST OF TABLES	ix
Chapter	
I. INTRODUCTION	1
II. MATERIALS AND METHODS	5
A. Data collection	5
B. Regression Model	5
C. Response variable transformation	7
D. Baseline model and analysis	7
E. Decade dependent model	9
F. Other modifications	10
G. Model validation	11
III. RESULTS	12
A. HI prediction	12
B. Positions with significant coefficients	13
C. Time plots	16
IV. DISCUSSION	19
REFERENCES	23

APPENDIX 1.....25

LIST OF ILLUSTRATIONS

Figure	Page
1. HI values estimated based on sequence data plotted against observed HI values..	13
2. 3D structure of the hemagglutinin protein trimer (PDB: 2HMG) highlighting the positions selected by the model	14
3. Positions selected by the model within the canonical epitope	15
4. Histogram of regression coefficients.....	16
5. Estimated serum and virus effect over time	17
6. Effect of the time range between the virus and the serum strain difference effects (A), and residuals of the model (B).	17
7. Frequency of numbers of decades.....	18

LIST OF TABLES

Table.....	Page
1. Predictor model combinations tested	11
2. Reported R^2 values of the baseline model and modifications	13

CHAPTER I

INTRODUCTION

The influenza virus of types A and B cause seasonal global epidemics in humans leading to hundreds of thousands of deaths annually. Vaccinations are currently the main defense strategy against the virus (WHO, 2014). However, antigenic drift, caused by amino acid substitutions in the virus' surface proteins requires regular vaccine updates in order to match the constantly changing circulating strains (WHO, 2014).

Hemagglutinin (HA) and neuraminidase (NA) are two virus surface proteins that are recognized by the immune system. HA is the dominant target for immune recognition and the HA inhibition (HI) assay is the most commonly used technique to determine antibody neutralizing capacity against influenza virus (Russell et al., 2008). The HI assay relies on the HA protein's ability to bind to red blood cells. Because of this property, red blood cells agglutinate when mixed with influenza virus particles. However, when antibodies specific to the agglutinating influenza virus strain are added to the blood/virus mixture, they neutralise the HA protein and prevent blood cell agglutination. The minimum amount of antibody necessary to prevent agglutination of red blood cells is determined by a series of twofold dilutions. The highest antibody dilution level that still prevents agglutination is called the HI titer and is an indication of how well particular antibodies neutralize a given influenza virus.

An HI titer is called *homologous* if the virus strain to be neutralized is the same as the strain used to raise the antibodies. If these two virus strains differ, the HI titer is called *heterologous*. Heterologous HI titers provide some information about cross-immunity between virus strains. However, heterologous titers alone do not measure cross-immunity between strains because they are influenced by factors that are unrelated to the antigen-antibody binding. One such factor is the virus avidity to host cell receptors. Viruses with strong avidity to blood-cell receptors require more antibodies to neutralize agglutination and the HI values would erroneously suggest that the antibodies provide a low protection against such virus. This effect is asymmetric, i.e. if virus i has a higher avidity than virus j , the heterologous HI obtained by using serum raised against virus j to neutralize virus i is lower than the heterologous HI obtained by using serum raised against virus i to neutralize virus j .

To account for this problem of varying receptor-binding avidity, Archetti and Horsfall (1950) suggested converting HI titers of any two strains into a symmetric antigenic distance (AD). The AD is calculated as the geometric mean of the two ratios of homologous to heterologous titer. Specifically, the AD between two viruses i and j , is calculated based on the homologous titers for virus i and j (H_{ii} and H_{jj} , respectively), the heterologous titer obtained by neutralizing virus j with serum raised on virus i (H_{ij}), and the titer obtained by neutralizing virus i with serum raised on virus j (H_{ji}) and is given by the following equation:

$$AD = \sqrt{\frac{H_{ii} * H_{jj}}{H_{ij} * H_{ji}}}$$

The AD removes asymmetric effects. This can easily be seen from the fact that exchanging i and j in the above formula does not change the AD. The AD has been shown to predict between-strain vaccine efficacies better than HI titers (Ndifon, Dushoff, & Levin, 2009) and is now widely used as a measure of cross-immunity between influenza strains.

Setting up HI assays however is relatively expensive and time consuming. Monoclonal antibodies specific to one influenza strain are required for the assay. Recent advances in sequencing technology triggered increased interest in computational approaches that can predict cross-immunity based on HA protein sequence data. The purpose of such computational approaches is twofold. On the one hand, predicting cross-immunity from sequence data could allow replacing of HI assays by cheap and fast sequencing techniques. In addition, predicting cross-immunity from sequence data can reveal amino acid residues that can alter the antigenic properties of a given virus and thereby improves our understanding of antigen-antibody interactions. Most computational approaches that estimate antigenic similarity from protein sequences use AD as measure for antigenic distance between strains (Bedford et al., 2014; Lee & Chen, 2004; Lees, Moss, & Shepherd, 2010; Liao, Lee, Ko, & Hsiung, 2008; Sun et al., 2013).

While AD has been proven to be a very useful measure for antigenic similarity between influenza strains (Ndifon et al., 2009), it has some limitations. The utility of AD as a predictor of vaccine efficacy has been called into question by a later study (Pan & Deem, 2009). A more important limitation of AD is that it removes all

asymmetric effects on HI titers. While the AD calculation removes non-antigenic effects such as virus avidity to host receptor, it also removes asymmetric effects that are antigenically relevant, such as general serum neutralizing capacity. If the antibodies raised against virus i have stronger neutralizing capability than antibodies raised against virus j , the heterologous titer H_{ij} obtained by neutralizing virus j with serum raised on virus i , is larger than H_{ji} , the titer obtained by neutralizing virus i with serum raised on virus j . Since the AD calculation removes serum neutralizing capacity effects, computational studies that relate AD to protein sequences fail to identify amino acid residues that influence serum neutralizing capacity.

The purpose of the present study is to develop a novel framework for predicting cross-immunity from HA sequence data. According to this framework the response variable are HI titers instead of AD and predictor variables do not only include HA sequence differences but also virus receptor binding avidity and serum potency effects. Choosing HI titers rather than AD as response variable allows identifying amino acid residues with asymmetric effects, e.g. residues that influence virus receptor binding avidity or serum potency. The two principal aims of this study are to determine how well HI titers can be predicted from sequence data and to identify amino acid residues that influence different factors affecting HI titers.

CHAPTER II

MATERIALS AND METHODS

A. Data collection

HI information collated by Bedford et al. (2014) was downloaded from an online repository (doi:10.5061/dryad.rc515). This data set contains over 10,000 HI assay results from 465 different H3N2 virus strains that were collected between 1968 and 2011. HA amino acid sequences of the strains found in this data set were downloaded from NCBI (<http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/>) and GISAID (<http://platform.gisaid.org/epi3/frontend#1136b8>). The downloaded sequences were aligned using MUSCLE (Edgar, 2004). The virus and serum strain of each HI titer was matched to its HA sequence from the alignment using strain names.

B. Regression Model

Regression models were fitted to predict HI titers from the HA protein sequence of the virus strain, the HA protein sequence of the serum strain, and the differences between the two HA protein sequences. The core of our model is described by the standard multiple regression equation:

$$y_j = \beta_0 + \sum_i^n \beta_i x_{ij} + \varepsilon_j$$

Where y_j is the j^{th} value of the response variable, in our case a log-transformed HI titer, x_{ij} is the value of the i^{th} predictor variable in the j^{th} measurement, n is the total number of predictor variables, β_0 and β_i are the estimated regression coefficients, and ε_j is an error term, describing the difference between prediction and observation.

Our baseline model generated three predictor variables per amino acid position on the HA protein, namely one for the amino acid residue on the virus strain HA, one for amino acid residue on the serum strain HA and one for the amino acid residue difference between the virus and serum strain HA. We included all 322 amino acid residues of the H3 protein as potential predictors, and hence our baseline model contained $n = 322 * 3 = 966$ predictor variables per HI value.

To avoid overfitting due to the large number of predictor variables we estimated regression coefficients using penalized regression approaches. Regression coefficients were estimated via elastic net regression. In elastic net (Zou & Hastie, 2005), each coefficient is calculated through minimizing the following equation:

$$\frac{1}{2N} \sum_j \sum_i (y_j - \beta_0 - \beta_i x_{ij})^2 + \lambda \left[(1 - \alpha) \frac{1}{2} \beta_i^2 + \alpha |\beta_i| \right]$$

The first term in in the above equation is the squared deviation between prediction and observation and the second term is an additional penalty term. A characteristic of elastic net is that its penalty term has two components (β_i^2 and $|\beta_i|$) and a parameter (α) that determines the weight of each penalty component. The first component of the penalty term is called ridge penalty. It tends to give equal coefficients to highly correlated predictors. The second component of the penalty term is called the lasso penalty. It tends to minimize the number of non-zero coefficients. We obtained

the optimal values for λ and α in a stepwise procedure. First a grid of α values was set up (ranging from 0 to 1 with a step width of 0.1, followed by a fine tuning set up with range from 0.5 to 1 with a step width of 0.05) and for each α the λ that minimize the mean prediction error was determined through five-fold internal cross validation using the R-package *glmnet* (Friedman, Hastie, & Tibshirani, 2010). Next the combination of α and λ was selected that lead the lowest overall mean prediction error (for more details check script in appendix). The estimated regression coefficients (β) were used to predict HI titers of any two strains based on their HA sequences.

Since the regression model required numeric predictor variables, virus and serum amino acid sequences were converted to numeric scores as described below. To simplify the descriptions we will refer to the HA amino acid sequences of the serum and virus strain as the *serum sequence* and *virus sequence*, respectively.

C. Response variable transformation

The response variable of our regression models were log transformed HI values (base 10). For example, if HI = 320, or response variable is $\log_{10}(320) = 2.5$. Whenever HI values were specified by an upper limit (e.g. < 20) we chose that upper bound.

D. Baseline model and analysis

Values for sequence difference predictor variables were calculated for each pair of HA sequences by assigning to each amino acid position a value of zero if it

contained the same amino acid in the virus and serum sequence and a value of one otherwise. In addition, separate predictor variables were generated for the virus and serum sequences. Each amino acid position received a value of zero if it contained the amino acid that is most common at the respective position across all strains and a value of one otherwise.

After running the model, positions that received non-zero coefficients were compared with the positions of the five canonical epitopes on the HA protein (Wiley, Wilson, & Skehel, 1981) and with positions that have been shown to influence HA affinity to human receptors (Lin et al., 2012). The positions selected by the model were highlighted on the 3D structure of the H3 trimer (PDB: 2HMG) in comparison to the canonical epitope regions using the visualization tool provided by the Influenza Research Database (IRD) website (direct link to the online rendering tool: <http://www.fludb.org/brc/structureOperation.spg?accession=2HMG&decorator=influenza&context=1463220628681#>).

A histogram of virus, serum and sequence difference coefficients was plotted to assess the estimated coefficients' signs and magnitude. We also investigated temporal trends in virus and serum effects. Virus and serum effects were calculated by multiplying each position-specific coefficient estimated by the model with its corresponding predictor variable value in a particular strain, and then summing this product over all positions. This sum was calculated for each strain and plotted against the year of its isolation. We also explored whether the time between the isolation dates of two strains that form an HI titer modifies difference effect. We calculated the time range by subtracting serum strain year from virus strain year, and calculated difference

effect for each HI pair by summing the product of difference predictor variable values and estimated coefficients over all amino acid positions. We plotted these difference effects against the time range. Last, we plotted the residuals of the model, i.e. actual HI titers subtracted from estimated titers, against the time between the isolation dates of two strains that form an HI value. Such a plot indicates whether the model systematically over or underestimates HI at certain time ranges.

E. Decade dependent model

The decade dependent scoring model allowed for sequence effects to vary by time. Each position in the virus and serum sequences was split into six dummy variables, one per decade from the 1960's to the 2010's. If the amino acid at a particular position in a particular strain differed from the most common amino acid at this position, the value on this position was set to one in the dummy variable corresponding to the decade in which the respective strain was collected and to zero in all other dummy variables. If the amino acid at a particular position in a particular strain was the most common amino acid on that position, the value was set to zero in all dummy variables. If the virus and sequence effects were consistent over time, the same amino acid position should produce the same regression coefficient across different decades. To summarize the temporal consistency of regression coefficients we therefore counted for each amino acid position with at least one non-zero coefficient the number decades in which its regression coefficient was non-zero. If effects were decade-specific, this number should be equal to one for most amino acid positions. If on the other hand, effects were consistent through time, this number should equal to six

for most positions, because then any amino acid position that showed an effect in one decade showed this effect in all other decades as well. We summarized the number of decades in which a coefficient is non-zero separately for virus and serum coefficients.

F. Other Modifications

Other predictor variable modifications were tested in an attempt to improve model prediction. The first modification factored in specific amino acids seen in each position of the HA of the virus and serum strain; every position in the strain sequence received a dummy variable for each amino acid seen at that specific position. Similar to the time variable, on each position, a value of one was given to the dummy variable corresponding to the amino acid seen in that specific position, and a zero to all other dummy variables.

The second modification incorporated interactions between all possible position pairs in the difference variables. Every position in the difference string was multiplied by every other possible position. Coefficients estimated of this modification were analysed for sign and magnitude. The different predictor models were tested in different combinations (Table 1).

Table 1: Predictor model combinations tested

Experiment	Virus/Serum scoring model	Difference scoring model
1	Baseline	Baseline
2	Decade dependent	Baseline
3	Amino acid specific	Baseline
4	Baseline	Interaction model

G. Model validation

Validation experiments were done, where one third of the data was removed from the estimation procedure, to later test the performance of the model on data not used to estimate regression coefficients. The data used to estimate regression coefficients is called the training set, the data used to test the model predictions is called the validation set. The first experiment selected the all strains after 2005 as the validation set. This was done to simulate real life situations where the model would be used to estimate HI for new strains, based on data of existing strains. However, in standard model validation, the training and validation sets are supposed to be random draws from the same distribution. Therefore, a second experiment was performed where the validation set was selected randomly. Correlations between actual HI titers and titers estimated by the model were used to evaluate strength of the model.

CHAPTER III

RESULTS

A. HI prediction

The calibrations found that an α of 0.7 and a λ around 0.8 minimized the prediction error. The correlation between predicted and observed HI values was similar across all versions of our model for the training data ($R^2 > 0.7$ in all models, Fig. 1, Table 2). For the validation data on the other hand, the correlation between predicted and observed HI depended on how the validation data were chosen. When the validation data were selected based on a time threshold the R^2 value for the correlation between predicted and observed HI among validation data was 0.096, whereas selecting validation data at random produced an R^2 value of 0.67 (Fig. 1, Table 2). Modifications on scoring method show little improvement over the baseline scoring method (Table 2).

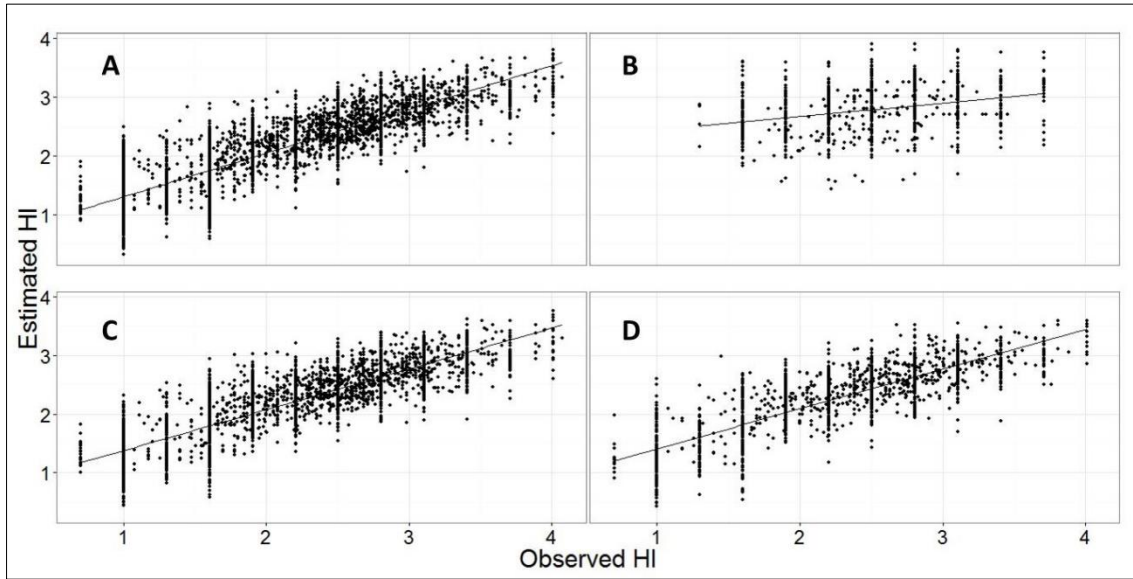


Figure 1: Log-transformed HI titers estimated based on sequence data plotted against observed HI values. (A) Comparison of predicted and observed HI titers in training set; training and validation set were separated by time. (B) Comparison of predicted and observed HI titers in validation set; training and validation set were separated by time. (C) Comparison of predicted and observed HI titers in training set; training and validation set were separated randomly. (D) Comparison of predicted and observed HI titers in validation set; training and validation set were separated randomly.

Table 2: Reported R^2 values of the baseline model and modifications

Model	Time Based Validation		Random Validation	
	Training R^2	Validation R^2	Training R^2	Validation R^2
Baseline model	0.76	0.069	0.72	0.67
Decade dependent	0.78	0.095	0.75	0.72
Amino acid specific	0.79	0.085	0.75	0.73
Interaction model	0.81	0.074	0.79	0.71

B. Positions with significant coefficients

Estimated coefficients are an indicator of position importance and predictors with zero coefficients are positions estimated to have no effect on virus immune recognition. The baseline model selected non-zero coefficients for 132 positions in at

least one of the three predictor variable groups, 93 of which are within the epitope regions (Fig. 2). Forty five of these 132 positions have non-zero coefficients for all three predictor variable groups (virus, serum and difference effects), 37 of these 45 positions fall within the epitope regions (Fig. 3). Most of the 39 positions with at least one non-zero coefficient that fall outside the epitope regions resided in the vicinity of these regions (Fig. 2). Eight of these positions have an effect in all three predictor groups (the positions are: 31, 199, 202, 222, 223, 225, 233, and 269). All positions reported to have a strong effect on host cell receptor binding by Lin et al. (2012) were selected by the model (positions: 190, 193, 222, 225, 226, and 227).

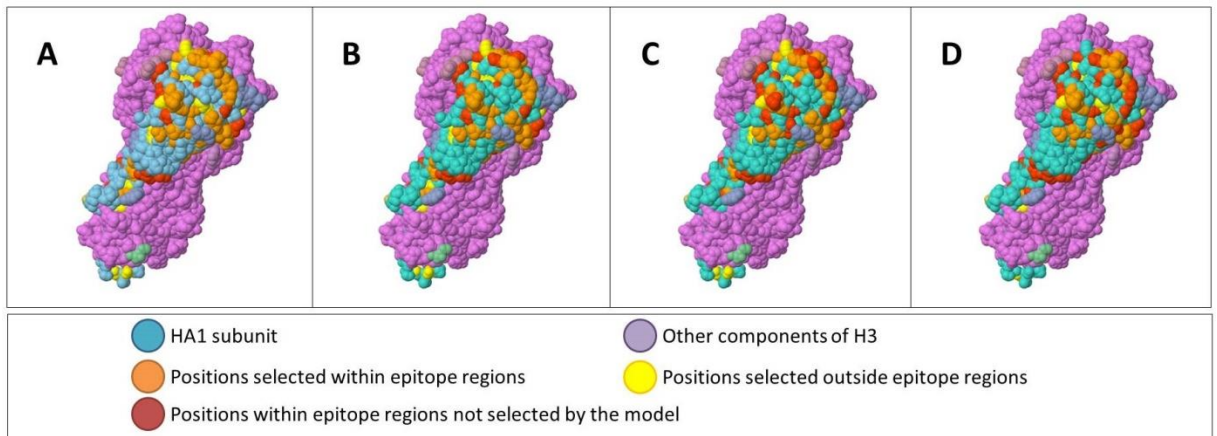


Figure 2: 3D structure of the hemagglutinin protein trimer (PDB: 2HMG) highlighting the positions selected by the model. Total positions selected (A), difference positions (B), virus positions (C) and serum positions (D) are highlighted above

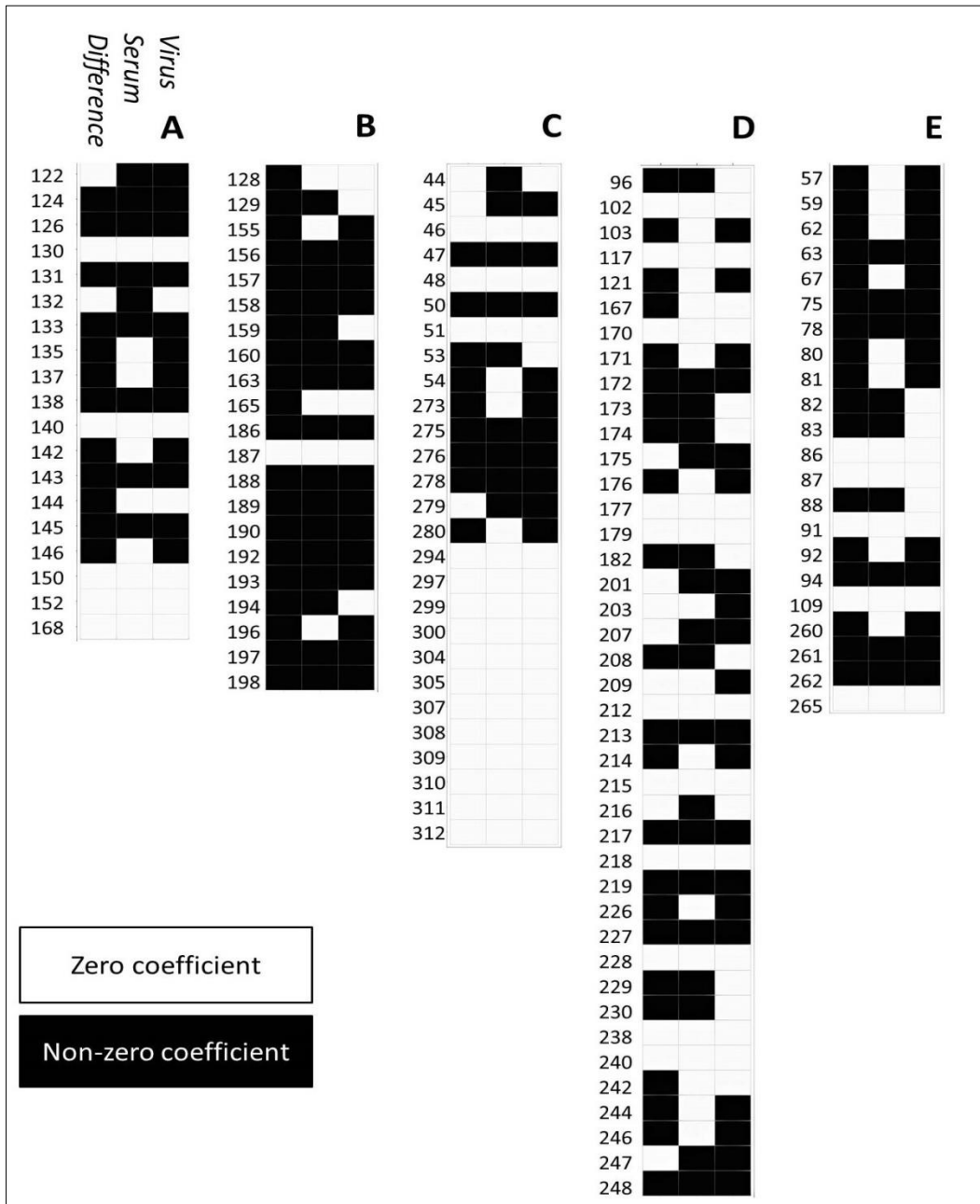


Figure 3: Positions that received zero and non-zero coefficients within known HA epitopes. Each box shows a canonical epitope and the name of the epitope is indicated by the bold letter on the left. The three different rows correspond to coefficients corresponding to effects of the virus strain HA, serum strain HA and the HA differences.

Coefficients of the sequence difference part of the estimation are more negative than positive, while the coefficients of the virus and serum components are

evenly distributed (Fig. 4). The interaction modification selected a total of 780 interactions from 90 different positions. Coefficients of this modification were mostly negative (about 60%).

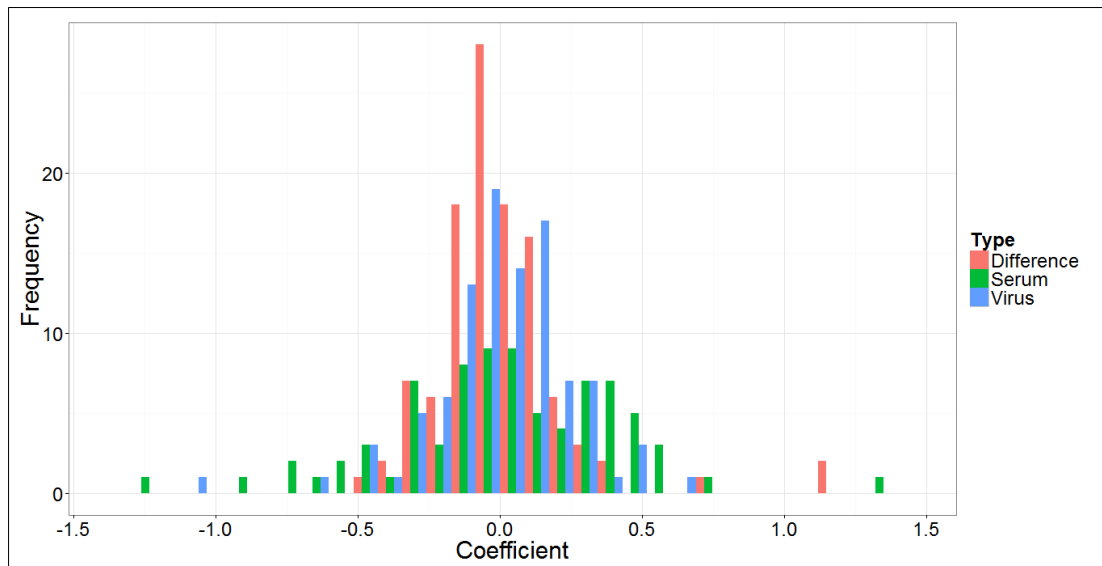


Figure 4: histogram of coefficients of the estimation

C. Time plots

Virus and serum effects do not follow a consistent temporal trend (Fig. 5). Estimated effects of sequence difference become more negative with the increase of time between strains (Fig. 6A). The residuals of the estimation tend to be negative when virus and serum are about 15 to 20 years apart (Fig. 6B).

Positions selected by the decade dependent model were examined for the number of decades they appear in. Most of the positions selected by the model appear in only one or two decades (Fig. 7).

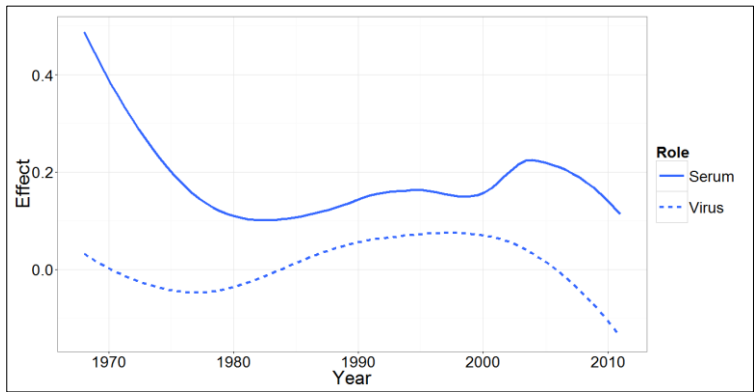


Figure 5: Estimated serum and virus effect over time. The effects are calculated using the coefficients estimated for each role by the model and the strain sequences. Effects are plotted against the year that specific strain was seen.

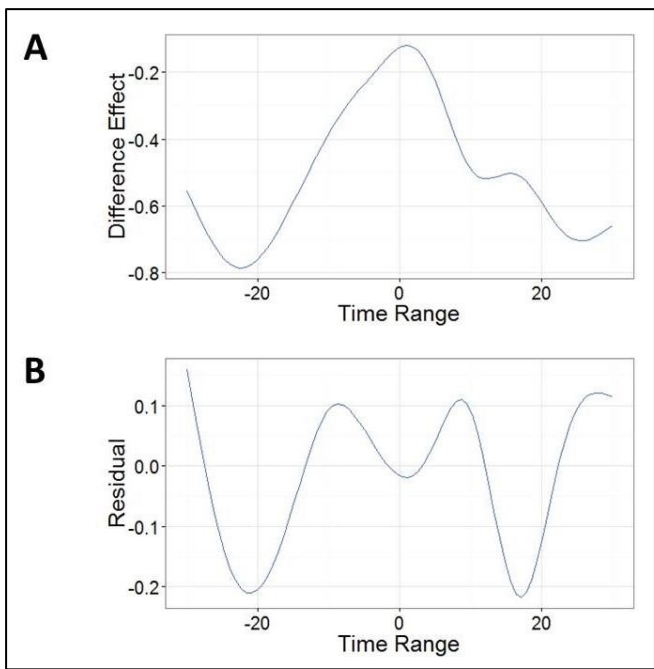


Figure 6: Effect of the time range between the virus and the serum strain difference effects (A), and residuals of the model (B).

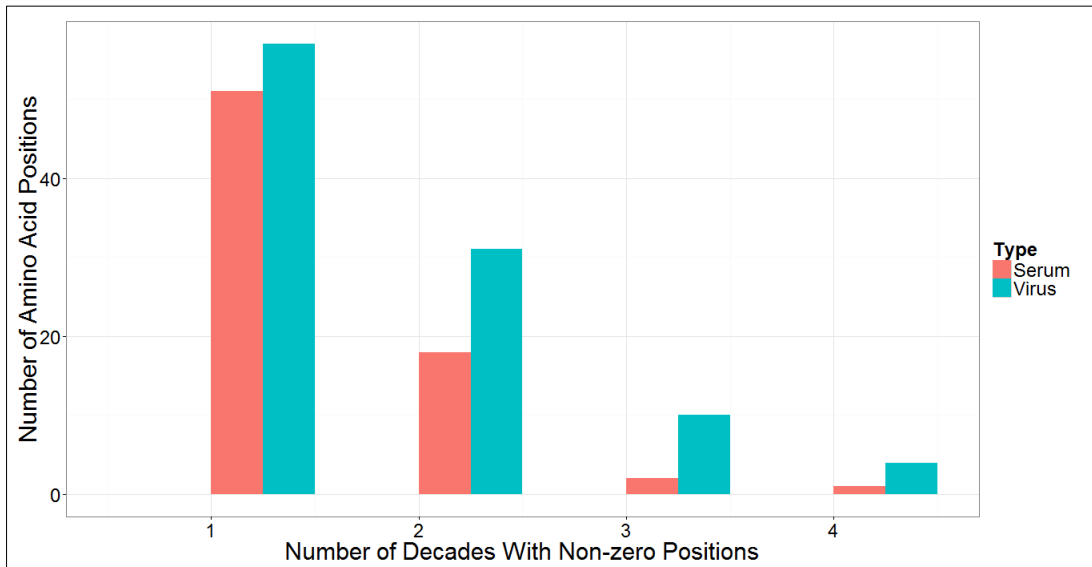


Figure 7: Frequency distribution of numbers of decades with non-zero coefficients per amino acid position.

CHAPTER IV

DISCUSSION

In this project we showed that it is possible to predict HI titers between any two influenza strains based on their HA protein sequences. However, there is some evidence that the coefficients estimated by our model from data in a particular time period cannot be extrapolated to a different time period. Nevertheless, the model does find antigenically important positions in or around epitopes and positions that have been shown to influence receptor-binding avidity. In contrast to previous AD-based approaches, our model can identify amino acids with effects that are specific to the virus or serum strain and are therefore potentially related to virus receptor binding avidity or serum potency.

The correlations between predicted and observed HI values in the training data of our study were comparable to the results of a previous study that predicted AD from HA sequence data without a validation step (Lee & Chen, 2004). Correlations between predicted and observed HI values in the validation data of our study were not far below the correlations in the training data if the validation data were selected at random. If, however, training and validation data were chosen according to a time threshold, the correlation between predicted and observed HI titers in the validation data dropped dramatically (Fig. 1B). This drop in correlation indicates that the coefficients estimated from data in one time period cannot be used to make predictions in another time period. The result that estimated coefficients are time-specific is corroborated by the model that estimated decade-specific coefficients. Only a small

minority of amino acid positions showed effects that were consistent across multiple decades (Fig. 7). Furthermore, serum strain effects were even more ephemeral than virus strain effects (Fig. 7). According to the residual plots, the model tends to overpredict HI titers for strains that are separated by 15-20 years (Fig. 5B). In other words, if the virus and serum strain that form an HI titer are separated by 15-20 years, these strains are antigenically more different from each other than what our model predicts based on their HA sequences. One possible interpretation of this pattern is that multiple amino acid changes have synergistic antigenic effects such that the combined result of many amino acid differences is stronger than the sum of individual effects. The fact that the majority estimated coefficients for interaction terms are negative confirms this interpretation since a negative interaction coefficient means that differences between serum and virus HA sequence at two amino acid positions lower the HI titer even more than the sum of effects of the two individual differences.

Amino acid positions that received non-zero coefficients were identified as important by the model. Most of these selected positions fell within the epitope regions (Fig. 3). Selected positions outside the epitope ranges tend to be around those regions (Fig. 2). All the positions that were previously found to be important for virus receptor binding avidity (Lin et al. 2012) were also selected by our model, even the positions that fall outside the epitope regions (positions: 222 and 225). The model also selects some positions that are under positive selection reported by other methods in all three predictor groups, such as positions: 63, 126, and 133 (Zaraket et al., 2009) (more positions are selected in at least one predictor group), positions: 124, 133, 138, 145, 156, 158, 186, 190, 193, 197, 262, and 275 (Bush, Bender, Subbarao, Cox, & Fitch, 2007) (all positions by Bush et al. are seen in at least one predictor group). This

suggests that the model can find antigenically important positions only using sequences.

While the spatial locations of amino acid positions with non-zero coefficients are largely consistent with prior biological knowledge of their functions, the sign of these coefficients did not always confirm prior expectation. Virus and serum effects could either increase or decrease HI titers, and indeed for each of these effects the number of positions with positive and negative coefficients are roughly equal. However, for difference predictors one would expect only negative coefficients since amino acid differences between HA proteins of two strains should lower cross-immunity between strains and thereby decreased HI titer values. In contrast to these expectations about 40% of the coefficients for difference predictors are positive. This is most likely a statistical artefact. Sun et al. (2013) observed a similar phenomenon in their analysis of AD data. While the positions selected by our model appear to be biologically meaningful, the magnitude of individual coefficients should be interpreted with caution.

Our approach allows estimating total virus and serum effects per strain by summing the products of fitted coefficients and predictor variable values. These strain-specific effects can give indications of trends in virus properties that affect HI titers. It is unclear what biological mechanisms influence the serum and virus effects estimated by our model. One potential driver of virus effects is virus avidity to host cell receptors. Lin et al. (2012) reported for H3N2 influenza a general decrease of receptor binding avidity over time from 1968 - 2010. Our model does not show the same trend for virus effects. It is therefore unlikely that the virus effects estimated by our model are due to receptor-binding avidity.

The analyses presented here could be extended in various ways. Phylogenetic comparative techniques (Pagel & Meade, 2006) could be used to test whether residue changes on positions that our model identified are correlated, indicating potential interactions between different amino acid residues. Furthermore, the decade-specific coefficients could be analysed to determine whether at specific times some regions on the H3 protein were most antigenically active. The model also provides a potential avenue to explore aspects of serum potency by analysing the positions selected from the serum predictor group.

This study has shown that predicting HI from HA sequences can produce novel insights into different aspects of virus-antigen binding. Our results revealed antigenically important amino acid positions and trends in receptor binding avidity that are consistent with and expand previous results. The approach presented here provides an important new avenue for studying the influenza virus antigenicity.

REFERENCES

- Archetti, I., & Horsfall, F. L. (1950). Persistent antigenic variation of influenza A viruses after incomplete neutralization in ovo with heterologous immune serum. *The Journal of Experimental Medicine*, 92(5), 441–62.
<http://doi.org/10.1084/jem.92.5.441>
- Bedford, T., Suchard, M. a., Lemey, P., Dudas, G., Gregory, V., Hay, A. J., ... Rambaut, A. (2014). Integrating influenza antigenic dynamics with molecular evolution. *eLife*, 2014(3), 1–26. <http://doi.org/10.7554/eLife.01914>
- Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J., & Fitch, W. M. (2007). Predicting the evolution of human influenza A. *Science*, 1921(1999), 10–15.
<http://doi.org/10.1126/science.286.5446.1921>
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.
<http://doi.org/10.1093/nar/gkh340>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22.
<http://doi.org/10.1359/JBMR.0301229>
- Gupta, V., Earl, D. J., & Deem, M. W. (2006). Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, 24(18), 3881–3888.
<http://doi.org/10.1016/j.vaccine.2006.01.010>
- Lee, M. S., & Chen, J. S. E. (2004). Predicting antigenic variants of influenza A/H3N2 viruses. *Emerging Infectious Diseases*, 10(8), 1385–1390.
<http://doi.org/10.3201/eid1008.040107>
- Lees, W. D., Moss, D. S., & Shepherd, A. J. (2010). A computational analysis of the antigenic properties of haemagglutinin in influenza A H3N2. *Bioinformatics*, 26(11), 1403–1408. <http://doi.org/10.1093/bioinformatics/btq160>
- Liao, Y. C., Lee, M. S., Ko, C. Y., & Hsiung, C. a. (2008). Bioinformatics models for predicting antigenic variants of influenza A/ H3N2 virus. *Bioinformatics*, 24(4), 505–512. <http://doi.org/10.1093/bioinformatics/btm638>
- Lin, Y. P., Xiong, X., Wharton, S. A., Martin, S. R., Coombs, P. J., Vachieri, S. G., ... McCauley, J. W. (2012). Evolution of the receptor binding properties of the influenza A(H3N2) hemagglutinin. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52), 21474–21479.
<http://doi.org/10.1073/pnas.1218841110>
- Ndifon, W., Dushoff, J., & Levin, S. a. (2009). On the use of hemagglutination-inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness. *Vaccine*, 27(18), 2447–52.
<http://doi.org/10.1016/j.vaccine.2009.02.047>

- Pagel, M., & Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov Chain Monte Carlo. *The American Naturalist*, 167(6), 808–825. <http://doi.org/10.1086/503444>
- Pan, K., & Deem, M. W. (2009). Comment on Ndifon et al., “On the use of hemagglutination-inhibition for influenza surveillance: Surveillance data are predictive of influenza vaccine effectiveness.” *Vaccine*, 27(18), 5033–5034. <http://doi.org/10.1016/j.vaccine.2009.02.047>
- Russell, C. a., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., ... Smith, D. J. (2008). Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. *Vaccine*, 26(SUPPL. 4), 31–34. <http://doi.org/10.1016/j.vaccine.2008.07.078>
- Sun, H., Yang, J., Zhang, T., Long, L. P., Jia, K., Yang, G., ... Wan, X. F. (2013). Using sequence data to infer the antigenicity of influenza virus. *mBio*, 4(4), 1–9. <http://doi.org/10.1128/mBio.00230-13>
- WHO. (2014). No Title. Retrieved November 18, 2015, from <http://www.who.int/mediacentre/factsheets/fs211/en/>
- Wiley, D. C., Wilson, I. A., & Skehel, J. J. (1981). Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*, 289(5796), 373–378. Retrieved from <http://dx.doi.org/10.1038/289373a0>
- Zaraket, H., Saito, R., Sato, I., Suzuki, Y., Li, D., Dapat, C., ... Suzuki, H. (2009). Molecular evolution of human influenza A viruses in a local area during eight influenza epidemics from 2000 to 2007. *Archives of Virology*, 154(2), 285–295. <http://doi.org/10.1007/s00705-009-0309-9>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2), 301–320. <http://doi.org/10.1111/j.1467-9868.2005.00503.x>

APPENDIX

Below is the script that runs the baseline prediction model and calculates the correlations between predicted and actual HI values.

```
library(ape)
library(seqinr)
library(glmnet)

# Read in alignment and convert to uppercase
Alignment <- read.fasta("Data/ProcessedData/AASeq_HI_alignednogap.fas")
Alignment <- t(sapply(Alignment, function(x) (x)))
Alignment <- toupper(Alignment)

#read in HIData
HIData <- read.delim("Data/RawData/H3N2_HI_data.txt", stringsAsFactors=FALSE)

#remove homologous titers
HIData <- HIData[HIData$virusStrain != HIData$serumStrain,]

# Create a variable that indicates whether the titer value is exact (TRUE)
# or upper bound (FALSE)
HIData$TiterExact <- !c(1:nrow(HIData)) %in% grep("<", HIData$titer)

# remove all "<" from titer column and convert to numeric vector
HIData$titer <- gsub("<", "", HIData$titer)
HIData$titer <- as.numeric(HIData$titer)

# log all titers
HIData$logT <- log(HIData$titer, base = 10)

# Calculate scoremat (difference between every strain combination found in
# HI table)
ScoreMat <- CalcScoreMat(HIData$virusStrain, HIData$serumStrain, Alignment,
                          1:ncol(Alignment), "binary")

# remove non-existing sequences from HIData
HIData <- HIData[!is.na(ScoreMat[,1]),]

## OPTIONAL
# create training and validation sets by date (commented out)
#AVGyear <- (HIData$virusYear + HIData$serumYear)/2
#THIData <- HIData[AVGyear<2005,]
#VHIData <- HIData[AVGyear>=2005,]
```

```

# create random training and validation set
random <- sample(1:3, length(HIData$titer), replace = TRUE)
THIData <- HIData[random>1,]
VHIData <- HIData[random==1,]

# recalculate scoremat (for validation set only)
ScoreMat <- CalcScoreMat(THIData$virusStrain, THIData$serumStrain, Alignment,
  1:ncol(Alignment), "binary")

# Fit model for Virus, serum and difference coefficients
# based on sequences

# Create a vector that gives for each AA position the number of unique AAs
NrAA <- apply(Alignment, 2, function(x) length(unique(x)))

# Turn alignment into a score by assigning 0 to the most common AA and 1
# to all others.
AlignScore <- apply(Alignment, 2, function(x) {
  Counts <- table(x)
  as.numeric(x != names(Counts)[which.max(Counts)])
})

# Match names from coefficients to alignment row names
NameMatchVirusAlign <- match(THIData$virusStrain, row.names(Alignment))
NameMatchSerumAlign <- match(THIData$serumStrain, row.names(Alignment))

# Subset alignment columns to get only columns with more than one AA
PredictCols <- which(NrAA > 1)

## create predictor alignment data frame
# create virus and serum position scoring matrix
VirusPos <- AlignScore[NameMatchVirusAlign, PredictCols]
SerumPos <- AlignScore[NameMatchSerumAlign, PredictCols]

# create Validation Predictor Matrix
Tpredictor <- cbind(ScoreMat, VirusPos, SerumPos)
colnames(Tpredictor) <- c(paste("ScoreMat", 1:ncol(ScoreMat), sep = ""),
  paste("Virus", 1:length(PredictCols), sep = ""),
  paste("Serum", 1:length(PredictCols), sep = ""))

## to select best alpha we run the following commented out section
## create k-fold vector
#FoldID <- sample(1:5, length(THIData$titer), replace = TRUE)

## create list of alpha values to be tested
#alphaslist <- seq(0.5,1,by=0.05)

## run elasticnet testing all possible alpha values

```

```

#elasticnet <- lapply(alphaslist, function(a){
# cv.glmnet(Tpredictor, THIDData$logT, alpha=a, foldid = FoldID)})

## print prediction errors of all alpha values tested and select the alpha
## with the lowest error
#for (i in 1:11) {print(min(elasticnet[[i]]$svm))}

# run Elasticnet (several runs show that 0.7 is the best alpha to be used)
ElasticFit <- cv.glmnet(Tpredictor, THIDData$logT, alpha=0.7)

# plot coefficients of estimation across different values of lambda
plot(ElasticFit$glmnet.fit, "lambda", label = TRUE)

# check fitted prediction
FittedT <- as.vector(predict(ElasticFit, Tpredictor))

# Run validation for Virus, serum and difference coefficients
# based on sequences

# Calculate V scorematrix
VScoreMat <- CalcScoreMat(VHIDData$virusStrain, VHIDData$serumStrain, Alignment,
1:ncol(Alignment), "binary")

# recreate virus/serum alignment match vectors
NameMatchVirusAlign <- match(VHIDData$virusStrain, row.names(Alignment))
NameMatchSerumAlign <- match(VHIDData$serumStrain, row.names(Alignment))

# create virus and serum position scoring matrix
VirusPos <- AlignScore[NameMatchVirusAlign, PredictCols]
SerumPos <- AlignScore[NameMatchSerumAlign, PredictCols]

# predict Elastic fit
Vpredictor <- cbind(VScoreMat, VirusPos, SerumPos)
colnames(Vpredictor) <- c(paste("VScoreMat", 1:ncol(VScoreMat), sep = ""),
paste("Virus", 1:length(PredictCols), sep = ""),
paste("Serum", 1:length(PredictCols), sep = ""))
FittedV <- as.vector(predict(ElasticFit, as.matrix(Vpredictor)))

# Diagnostic Plots
plot(THIDData$logT, FittedT, xlab = "Actual", ylab = "Fitted",
main = "Elastic training")
abline(0,1)
abline(fit <- lm(FittedT ~THIDData$logT), col='red')
legend(x = max(THIDData$logT) - 1, y = min(FittedT) + 2, bty="n",
legend=paste("R2 is",format(summary(fit)$adj.r.squared, digits=3)))

plot(VHIDData$logT, FittedV, xlab = "Actual", ylab = "Fitted",
main = "Elastic Validation")

```

```
abline(0,1)
abline(fit <- lm(FittedV ~VHIData$logT), col='red')
legend(x = max(FittedV) - 1, y = min(VHIData$logT) + 1.5, bty="n",
       legend=paste("R2 is",format(summary(fit)$adj.r.squared
```

