Bio-Medical Materials and Engineering 24 (2014) 3729–3735
DOI 10.3233/BME-141201
IOS Press

Using multiple linear regression and physicochemical changes of amino acid mutations to predict antigenic variants of influenza A/H3N2 viruses

Haibo Cui[a], Xiaomei Wei[b], Yu Huang[b], Bin Hu[b], Yaping Fang[b] and Jia Wang[b,*]
[a]*School of Computer Science and Information Engineering, Hubei University, Wuhan 430070, China*
[b]*College of Informatics, Huazhong Agricultural University, Wuhan 430070, China*

Abstract. Among human influenza viruses, strain A/H3N2 accounts for over a quarter of a million deaths annually. Antigenic variants of these viruses often render current vaccinations ineffective and lead to repeated infections. In this study, a computational model was developed to predict antigenic variants of the A/H3N2 strain. First, 18 critical antigenic amino acids in the hemagglutinin (HA) protein were recognized using a scoring method combining phi (ϕ) coefficient and information entropy. Next, a prediction model was developed by integrating multiple linear regression method with eight types of physicochemical changes in critical amino acid positions. When compared to other three known models, our prediction model achieved the best performance not only on the training dataset but also on the commonly-used testing dataset composed of 31878 antigenic relationships of the H3N2 influenza virus.

Keywords: Influenza A virus, H3N2, antigenic variant, multiple linear regression, physicochemical properties

1. Introduction

Influenza A viruses circulate in the human population every year and cause enormous losses for global economics, social medicine and human health. According to statistics from the WHO, influenza A viruses cause three to five million severe illnesses annually and ~500,000 deaths all over the world [1]. Currently, the influenza A viruses that have spread widely in the human population consists mainly of H3N2, H1N1 and H2N2 subtypes, among which the H3N2 subtype causes the largest fraction of influenza illness [1,2]. At present, vaccination is the principal method for preventing influenza infection [3]. However, point mutations in viral proteins allows the virus to change their antigenic properties, thereby escaping the human immune system [1,4]. This also poses challenges to 1) the effectiveness of vaccination, 2) global influenza virus surveillance, and 3) vaccine selection.

*Corresponding author: Jia Wang, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China. Tel.: +86-27-87282492; Fax: +86-27-87282019; E-mail: wang.jia@mail.hzau.edu.cn.

0959-2989/14/$27.50 © 2014 – IOS Press and the authors.

This article is published with Open Access and distributed under the terms of the Creative Commons Attribution and Non-Commercial License.

Therefore, it is of great value to identify critical mutations and make predictions for future possible antigenic variants of the influenza A viruses.

Currently, determining antigenic variants of influenza A viruses mainly depends on ferret serum hemagglutinin-inhibition (HI) assays [5]. However, this technique is expensive and time-consuming. Computational methods not only can predict antigenic variants of influenza A viruses, it can also identify potential key amino acid positions for antigenic changes. In a recent study, Lee and Chen constructed five computational models to predict antigenic relationships of influenza A/H3N2 viruses, among which the model using hamming distance of the five epitopes in the hemagglutinin (HA) glycoprotein achieved the best performance [5]. In 2008, Liao and his colleagues employed and compared six grouping strategies and four bioinformatics models to predict antigenic variants [6]. Most recently, Huang's group identified 19 key positions and constructed a decision tree composed of six nodes to predict antigenic variants [7]. Although the above models achieve ~90% accuracy on the same training dataset composed of 181 pairs of antigenic relationship records, these models attained a high false positive rate on the testing dataset containing 31,878 HI records proposed by Smith's group [8]. As an example, Huang's model made 1131 wrong predictions for 4780 antigenic similarities, a false positive rate of 23.66%. In addition, their training sample size is relatively small (181 pairs). Therefore, new strategies and approaches are needed to 1) increase the predictive effect, 2) reduce the false positive rate, and 3) increase the sample size to reflect HI data from studies and surveillance reports.

In this study, we first increased the sample size and then attempted to use a new scoring method combing phi (ϕ) coefficient and information entropy to identify potential key antigenic amino acid positions. Based on scoring method and multiple linear regression, 18 key amino acid positions were determined. Additionally, the grouping method in Liao's study [6] inspires us that some amino acid mutations itself may not cause antigenic variants, and that antigenic variants may only develop upon some physicochemical changes. Therefore, physicochemical changes of the 18 amino acid positions were applied to predict antigenic variants. Compared to other models, our model attained a lower false positive rate and achieved optimal performance on both training and testing datasets. In addition, the identified 18 potential key amino acid positions would be valuable for research in related fields of influenza A viruses.

2. Materials and methods

2.1. Dataset

We first manually collected 394 HI assay records of influenza A/H3N2 viruses from related publications, the Weekly Epidemiological Records (WER) from the WHO and the influenza surveillance reports from the American Center for Disease Control and Prevention (CDC), among which there were a total of 94 influenza A/H3N2 viral strains. Based on the Archetti-Horsfall method [9,10], the antigenic distance between two influenza viruses can be calculated with the formula $d_{ij} = \ln\sqrt{(H_{ii}H_{jj}/(H_{ij}H_{ji})}$, where H_{ii} and H_{jj} refers to homologous antibody titers, H_{ij} and H_{ji} refers to heterologous antibody titers. If $d_{ij} \geqslant \ln 4$, the two viruses can be considered as antigenic variants. Otherwise, the relationship of antigenic similarity should be assigned. Thus, the training dataset composed of 394 pairs of influenza A/H3N2 viruses were split into 208 antigenic variants and 186 antigenic similarities. Additionally, an independent testing dataset composed of 27098 antigenic variants and 4780 antigenic similarities of H3N2 influenza viruses was used in this study to validate

the prediction model. This testing dataset was generated from 11 antigenic clusters of H3N2 influenza viruses in Smith's study [8], which was also used to validate Liao's [6] and Huang's models [7].

2.2. Recognition of critical amino acid positions in antigenic variants

In order to identify critical amino acid positions for antigenic variants of H3N2 influenza viruses, a scoring method combining phi (ϕ) coefficient and information entropy was employed. Considering pairwise mutation of H3N2 influenza viruses at each of the 329 amino acid positions in the HA1 subunit as a random variable X, it has two states: 1 representing mutation and 0 representing non-mutation. Meanwhile, the antigenic relationship of pair-wise viruses can be considered as another variable Y: 1 representing variant and 0 representing similarity. Therefore, the ϕ coefficients of all 329 amino acid positions can be computed as follows:

$$\phi_i = (N_{11}N_{00} - N_{10}N_{01}) \Big/ \sqrt{N_{X1}N_{X0}N_{1Y}N_{0Y}}\,, i = 1,2,...,329 \tag{1}$$

where N_{mn} (m=1,0 and n=1,0) is the number of virus pairs with X=m and Y=n, N_{Xn} is the number of virus pairs with Y=n and N_{mY} is the number of virus pairs with X=m. The phi coefficients range from −1 to +1, where ±1 demonstrates perfect agreement or disagreement and 0 indicates no correlation. We further added a weight factor for each amino acid position according to the entropy method that has been successfully used in different bioinformatics fields [11–12]. An entropy value is defined at an aligned amino acid position according to the formula E_i=-ΣP_j*log(P_j), where P_j is the observed probability for each of the 20 amino acids. Taken together, a scoring formula for the amino acid positions was defined as $S_i = \left|\varphi_i\right| * E_i, i = 1,2,...,329$. Thus, the significance of all 329 HA1 amino acid positions was calculated. We further set different thresholds including 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35 and 0.4 to generate 9 filtered position sets. Combining multiple linear regression and these position sets, prediction models were constructed. Finally, the 18 amino acid positions used in the best model were identified as potential key amino acid positions for antigenic variations.

2.3. Clusters of 531 physicochemical properties

In order to reduce the false positive rate of the prediction model, we attempted to use physicochemical properties in AAIndex [13] to improve the prediction model. However, the AAIndex database has too many physicochemical properties and redundancy needs to be eliminated. So we first used mutual information to cluster 531 physicochemical properties in AAIndex. Actually, mutual information can be used to measure the mutual dependence of the two random variables. Each physicochemical feature can be characterized with 18 mutual information values calculated with the following equation:

$$MI(X,i) = \sum_R \sum_T P(X = R, A_i = T) \log(P(X = R \mid A_i = T) \big/ P(X = R)) \tag{2}$$

where i is 18 key amino acid positions, $P(X=R)$ is the observed probability of antigenic relationship is R and R includes two states: variant and similarity, $P(X=R,A_i=T)$ is the joint probability of antigenic relationship and physicochemical property in given position i. T represents different AAIndex values and $P(X=R|A_i=T)$ is the conditional probability. Subsequently, according to calculated mutual

information features, 531 physicochemical properties were clustered to 12 groups by using hierarchical cluster algorithm in Matlab. Further, the 'find(T==k)' function was used to find out which physicochemical properties are contained in group *k* and the first one in each group was chosen as the representative. Thus, 12 representatives such as positive charge, alpha-CH chemical shifts, number of hydrogen bond donors etc, were chosen to encode 18 key amino acid positions and construct the prediction model.

2.4. Encoding the key amino acid positions to construct the prediction model

After 12 representative physicochemical properties were selected from the AAIndex database, we employed these physicochemical changes of 18 key amino acid positions to fit antigenic distances of H3N2 influenza virus pairs. Firstly, 18 key positions were respectively encoded by each of 12 physicochemical properties. Taking position 145 and the physicochemical property of "alpha-CH chemical shifts" as an example, if the amino acid at position 145 is mutated from isoleucine to serine, it shall be encoded with 0.55(4.5-3.95) according to the fact that alpha-CH chemical shifts of serine and isoleucine in AAIndex are 4.5 and 3.95 respectively. Thus, each of 18 key positions can be encoded to 12 physicochemical change values. We then input the 216 (18*12) features to execute multiple stepwise regression to fit antigenic distances. Finally, according to the rule of antigenic distance$\geqslant$ln4, the prediction model was constructed.

3. Results and discussion

3.1. The significance scores of 329 amino acid positions

Based on the significance scoring method mentioned in Section 2.2, 329 amino acid positions in HA1 subunit were ranked, among which 206 scored zero. Thus, these 206 positions were excluded from the prediction model. The remaining 123 positions achieved positive scores among which the highest is ~0.7. Most of the scores are below 0.15 and distribute on the interval from 0 to 0.05, only nine positions have scores higher than 0.4. Considering most of the virus pairs with more than nine mutations in the testing dataset are antigenic variant pairs, the nine positions were included in the prediction model and different thresholds below 0.4 were set to generate more antigenic critical amino acid positions.

3.2. Antigenic critical amino acid positions

By setting different thresholds from 0 to 0.4, 9 candidate position sets were generated. Combining each candidate position set with multiple linear regression on the training dataset, 9 prediction models were then constructed. Moreover, these models were also validated on the testing dataset containing 31878 virus pairs. The experiment results are shown in Figure 1. When all 123 positions with non-zero scores are used to construct the prediction model, the highest accuracy on the training dataset and the lowest accuracy on the testing dataset was attained. This suggests over-fitting had occurred, and that some critical positions are specific for the training dataset and are not representative of the larger sample. The candidate position sets selected by the thresholds of 0.05, 0.1, 0.35 and 0.4 also resulted in model over-fitting and under-fitting. Therefore, these position sets were excluded. Interestingly, models constructed using the remaining four position sets all achieved relatively high accuracies on

both the training and testing datasets. Furthermore, within these four models, the one constructed by 18 positions selected with a cutoff of 0.25 achieved the best performance. Therefore, these 18 positions were considered as critical antigenic amino acid positions.

Table 1 shows the significance scores and other information of the 18 critical amino acid positions. Based on the table, the importance of the 18 positions can be easily observed. For example, at position 145, there were a total 123 virus pairs containing amino acid mutations, among which 99 pairs became antigenic variants. Some of the 18 critical positions are consistent with previous studies. Huang's group determined six key amino acid positions, among which four positions were also identified in this study. Based on our significance measure, positions 213 and 214 only achieved a score below 0.15. This suggests that the importance of position 213 and 214 for antigenic variation may have diminished with the evolution of influenza A/H3N2 viruses. Statistical analyses including Mann-Whitney U-test and F-score were further implemented to validate statistical differences between the positive and the negative samples in the 18 positions. All the F-scores of these 18 positions are larger than 0.178, whereas the P-values are lower than 4.86e-4. This highlights a clear distinction between the antigenic

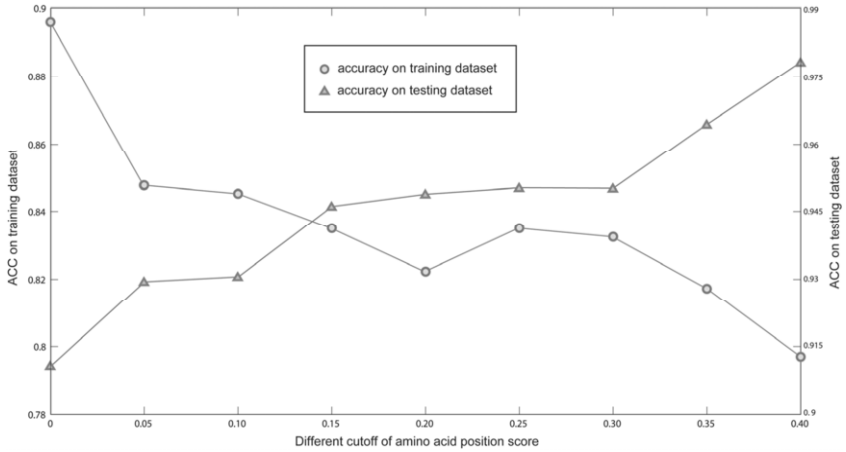

Fig. 1. Prediction accuracy (ACC) changes of the models with different score cutoffs on training and testing dataset.

Table 1

The significance scores and other information of 18 critical amino acid positions

Position	N_{11}	N_{10}	N_{01}	N_{00}	$\|\phi\|$	Entropy	Significance	Position	N_{11}	N_{10}	N_{01}	N_{00}	$\|\phi\|$	Entropy	Significance
50	57	11	151	175	0.28	1.66	0.47	158	63	7	145	179	0.35	1.6	0.55
62	51	5	157	181	0.31	1.49	0.47	160	49	6	159	180	0.29	0.89	0.26
83	52	9	156	177	0.28	1.26	0.35	172	47	10	161	176	0.24	1.55	0.38
133	52	8	156	178	0.29	1.38	0.4	186	44	69	164	117	0.18	1.66	0.29
135	47	11	161	175	0.24	1.44	0.34	189	89	35	119	151	0.26	2.17	0.56
137	82	18	126	168	0.34	1.53	0.52	193	86	43	122	143	0.19	1.89	0.37
145	99	24	109	162	0.37	1.57	0.59	197	40	2	168	184	0.29	0.94	0.27
155	53	10	155	176	0.27	1.56	0.43	276	30	5	178	181	0.21	1.46	0.3
156	97	28	111	158	0.34	1.98	0.67	278	46	2	162	184	0.32	1.4	0.45

Note: N_{mn} represents the number of virus pairs with X=m and Y=n, where X is mutation state and Y is antigenic variant state.

variants versus the antigenic similarities. Taken together, the above data indicate that the 18 positions identified in this study are of great importance to the antigenic variation of influenza A/H3N2 viruses.

3.3. The final prediction model and performance evaluation

Combining multiple linear regression and physicochemical changes of 18 key amino acid positions, the final prediction model was constructed. Table 2 shows physicochemical properties and regression coefficients of the critical mutation positions used in the model. In order to accurately evaluate the performance of the prediction model, 10-fold cross validation and four common-used measures including accuracy (ACC), sensitivity (SN), specificity (SP) and Matthews correlation coefficient (MCC) were used. Moreover, our model was also compared with three existing models including the Hamming distance model in Lee's study, the multiple regression model in Liao's study and the decision tree model in Huang's study. Table 3 shows the performances and comparison results of the four models on the training and testing datasets.

As shown in Table 3, the model constructed in this study achieved the best performance in terms of ACC and the comprehensive measurement of MCC on both the training and testing datasets. The ACC on the training dataset was higher than 86% and the MCC was larger than 0.72. With regard to the SP, although Liao's model achieved a performance of 83.33% that is slightly higher than our model on the testing dataset, this may be attributed to the imbalance of 27098 antigenic variants against 4780 antigenic similarities. Apart from this, our model has made observable improvement compared with the other three models. Especially against Huang's decision tree model, the SP is improved from 66.67% to 86.46% on the training dataset, indicating the false positive rate has been greatly reduced. Therefore, the model constructed in this study is superior to the other three models and will be more robust and effective for predicting new antigenic variants of influenza A/H3N2 viruses.

Table 2

The physicochemical properties and critical amino acid positions in the final model

Position	Physicochemical property	Coeff	Position	Physicochemical property	Coeff
135	alpha-CH chemical shifts	0.99	145	Number of hydrogen bond donors	0.35
155	alpha-CH chemical shifts	7.78	155	Loss of Side chain hydropathy by helix form	-1.18
193	alpha-CH chemical shifts	0.84	158	Loss of Side chain hydropathy by helix form	0.44
278	alpha-CH chemical shifts	0.82	160	Loss of Side chain hydropathy by helix form	-7.26
62	The number of atoms in the side chain	0.62	193	Bitterness	0.45
137	The number of atoms in the side chain	0.65	145	Amphiphilicity index	0.11
145	A parameter of charge transfer capability	0.65	156	Amphiphilicity index	0.18
189	Positive charge	-1.65	276	Amphiphilicity index	0.36
135	Positive charge	0.69	189	Amphiphilicity index	1.03

Table 3

The performance comparison of the models respectively constructed in Lee's, Liao's, Huang's and this study

Model	Training dataset				Testing dataset			
	ACC(%)	SN(%)	SP(%)	MCC	ACC(%)	SN(%)	SP(%)	MCC
This study	86.06	85.63	86.46	0.726	96.96	99.55	82.30	0.877
Liao's model	83.25	81.73	84.95	0.666	96.54	98.87	83.33	0.860
Huang's model	78.68	89.42	66.67	0.560	96.23	99.73	76.34	0.846
Lee's model	77.66	85.10	69.35	0.554	92.44	96.34	70.33	0.693

4. Conclusion

In this study, we focused on the prediction of antigenic variants of influenza A/H3N2 viruses and the identification of critical antigenic amino acid positions. A multiple regression prediction model combined with physicochemical changes of key amino acid positions was developed. By incorporating physicochemical changes, our model can predict with a low false positive rate whether two H3N2 virus strains are antigenic variants of one another. Moreover, the identified 18 key positions were demonstrated to be of great importance for antigenic variations. Together, we anticipate our work will be valuable for public health and future research on antigenic variants of influenza A/H3N2 viruses.

Acknowledgment

This work is supported by the Fundamental Research Funds for the Central Universities (2014JC003), Doctoral Research Funding of Huazhong Agricultural University (52902-0900206172) and the National Natural Science Foundation of China (Grant No: 61202304).

References

[1] P. Keyao and W.D. Michael, Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza, Journal of The Royal Society Interface **8** (2011), 1644–1653.
[2] C.A. Russell, T.C. Jones, I.G. Barr, N.J. Cox, R.J. Garten, V. Gregory, I.D Gust, A.W. Hampson, A.J. Hay, A.C. Hurt, J.C. de Jong, A. Kelso, A.I. Klimov, T. Kageyama, N. Komadina, A.S. Lapedes, Y.P. Lin, A. Mosterin, M. Obuchi, T. Odagiri, A.D. Osterhaus, G.F. Rimmelzwaan, M.W. Shaw, E. Skepner, K. Stohr, M. Tashiro, R.A. Fouchier and D.J. Smith, The global circulation of seasonal influenza A (H3N2) viruses, Science **320** (2008), 340–346.
[3] K. Florian and P. Peter, Universal influenza virus vaccines: need for clinical trials, Nature Immunology **15** (2014), 3–5.
[4] G. Vishal, J.E. David and W.D. Michael, Quantifying influenza vaccine efficacy and antigenic distance, Vaccine **24** (2006), 3881–3888.
[5] M.S. Lee and J.S. Chen, Predicting antigenic variants of influenza A/H3N2 viruses, Emerging Infectious Diseases **10** (2004), 1385–1390.
[6] Y.C. Liao, M.S. Lee, C.Y. Ko and C.A. Hsiung, Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus, Bioinformatics **24** (2008), 505–512.
[7] J.W. Huang, C.C. King and J.M. Yang, Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses, BMC Bioinformatics **10** (2009), S41.
[8] D.J. Smith, A.S. Lapedes, J.C. de Jong, T.M. Bestebroer, G.F. Rimmelzwaan, A.D. Osterhaus and R.A. Fouchier, Mapping the antigenic and genetic evolution of influenza virus, Science **305** (2004), 371–376.
[9] W. Ndifon, J. Dushoff and S.A. Levin, On the use of hemagglutination inhibition for influenza surveillance: surveillance data are predictive of influenza vaccine effectiveness, Vaccine **27** (2009), 2447–2452.
[10] I. Archetti and F.L. Horsfall Jr., Persistent antigenic variation of influenza A viruses after incomplete neutralization in ovo with heterologous immune serum, The Journal of Experimental Medicine **92** (1950), 441–462.
[11] J. Wang, Z. Kou, M. Duan, C. Ma and Y. Zhou, Using amino acid factor scores to predict avian-to-human transmission of avian influenza viruses: a machine learning study, Protein & Peptide Letters **20** (2013), 1115–1121.
[12] K. Ye, E.M. Lameijer, M.W. Beukers and A.P. Ijzerman, A two-entropy analysis to identify functional positions in the transmembrane region of class AG protein-coupled receptors, Proteins **63** (2006), 1018–1030.
[13] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Research **36** (2008), 202–205.