BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 4 2008, pages 505–512
doi:10.1093/bioinformatics/btm638

Sequence analysis

Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus

Yu-Chieh Liao[1,†], Min-Shi Lee[2,†], Chin-Yu Ko[1] and Chao A. Hsiung[1,*]

[1]Division of Biostatistics and Bioinformatics and [2]Vaccine R&D Center, National Health Research Institutes, Zhunan 350, Taiwan

Received on July 30, 2007; revised on November 15, 2007; accepted on December 17, 2007

Advance Access publication January 10, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Continual and accumulated mutations in hemagglutinin (HA) protein of influenza A virus generate novel antigenic strains that cause annual epidemics.

Results: We propose a model by incorporating scoring and regression methods to predict antigenic variants. Based on collected sequences of influenza A/H3N2 viruses isolated between 1971 and 2002, our model can be used to accurately predict the antigenic variants in 1999–2004 (agreement rate = 91.67%). Twenty amino acid positions identified in our model contribute significantly to antigenic difference and are potential immunodominant positions.

Contact: hsiung@nhri.org.tw

Supplementary information: The supplementary information includes 62 amino acid sequences of H3N2 viruses and 277 pair-wise antigenic distances.

1 INTRODUCTION

Influenza A virus, a viral respiratory pathogen, can cause human mortality and morbidity. The virus is classified into different subtypes, e.g. H1N1 and H3N2, based on differences in its surface proteins, hemagglutinin (HA) and neuraminidase (NA). HA is the most important antigen for inducing protective antibody responses. Accumulation of mutations on the HA cause antigenic drift, i.e. immunologically distinct strains (also named as antigenic variants), which are produced continuously. This antigenic drift requires new vaccine strains to be updated frequently. The HA protein is a trimer of identical subunits, each of which consists of two chains, HA1 and HA2, which are 329 and 175 residues long, respectively. The HA1 mutates more frequently than HA2 and experiences a strong selection for novel variants (Bush *et al*., 1999a, b; Plotkin and Dushoff, 2003). Although the NA gene also acquires substitution rapidly, NA is not considered an antigenic determinant due to its relatively small prevalence on surfaces and weak binding with antibodies (Plotkin and Dushoff, 2003). Influenza A/H1N1, A/H3N2 and B viruses have been circulating throughout the world since 1977, and thus current vaccines are usually trivalent, containing these three strains. The antigenic variants are determined currently by using the ferret serum hemagglutinin-inhibition (HI) assay. The serology assay, however, is labor-intensive and time-consuming. Alternatively, molecular sequencing is high-throughput and yields a more accessible and precise sequence of virus. Various studies have revealed the feasibility of HA1 sequence data of influenza A for prediction of antigenic drift (Bush *et al*., 1999a; Lee and Chen, 2004; Smith *et al*., 2004). In a previous study, five models were employed to predict antigenic variants of influenza A/H3N2 viruses, with a model based on 131 amino acid positions on the five antigenic sites shown to have a good agreement rate of 83% to serological data (Lee and Chen, 2004). To refine the predictive power, in this study we established bioinformatics models that incorporated scoring methods in the analyses for predicting antigenic variants of influenza A H3. In addition, potential immunodominant positions could be identified and can be used to accelerate the process for selection of vaccine strain.

2 METHODS

2.1 Data

Forty-five HA1 amino acid sequences of H3N2 viruses isolated between 1971 and 2002 and 181 pair-wise antigenic distances were available among these 45 viruses collected as a training dataset (Lee and Chen, 2004). A validation dataset consists of 96 pair-wise antigenic distances among 19 viruses isolated between 1999 and 2004. The antigenicity of influenza viruses is characterized based on ferret serum HI antibody cross-reactivity. The antigenic distance between two viruses is defined as the reciprocal of the geometric mean of two ratios between the heterologous and homologous antibody titers (Kilbourne *et al*., 1990). The amino acid sequences of viruses and their pair-wise antigenic distances are shown in Supplementary Data.

2.2 Methods

2.2.1 Scoring methods The previous models are based solely on the number of amino acid changes between two aligned sequences. In the current study, we aim to improve the models by considering the polarity, charge and structure of amino acids. In order to quantitatively evaluate the amino acid difference of pair-wise comparisons, different similarity classes for grouping amino acids and scoring matrices were employed. The non-grouping method used in the previous models is the

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

simplest scoring method that assigns zero or one depending on whether two amino acid residues are identical or not. As for grouping methods (GM1–GM6), the score is assigned one as the amino acid residues change between different classes; however, change of the amino acid residue within the same class is ignored, and thus zero is assigned. As a consequence, the scores of pair-wise amino acid comparisons obtained from non-grouping and grouping methods are binary. The similarity classes in each grouping method are listed below:

GM1: {non-polar: A, F, G, I, L, M, P, V, W}, {polar: C, N, Q, S, T, Y}, {charged: D, E, H, K, R}
GM2: {non-polar aliphatic: A, G, I, L, M, V}, {non-polar aromatic: F, P, W}, {polar: C, N, Q, S, T, Y}, {charged: D, E, H, K, R}
GM3: {non-polar: A, F, G, I, L, M, P, V, W}, {polar: C, N, Q, S, T, Y}, {positively charged: H, K, R}, {negatively charged: D, E}
GM4: {non-polar aliphatic: A, G, I, L, M, V}, {non-polar aromatic: F, P, W}, {polar: C, N, Q, S, T, Y}, {positively charged: H, K, R}, {negatively charged: D, E}
GM5: {non-polar aliphatic: A, I, L, M, P, V}, {non-polar aromatic: F, W, Y}, {polar: N, Q, S, T}, {positively charged: H, K, R}, {negatively charged: D, E}, {C}, {G} (Espadaler *et al*., 2005)
GM6: {non-polar aliphatic: A, I, L, M, P, V}, {non-polar aromatic: F, W, Y}, {polar: N, Q, S, T}, {charged: D, E, H, K, R}, {C}, {G}

GM1 divided the 20 amino acid residues into the three most common classes: non-polar, polar and charged. GM2–GM4 further divided the non-polar class into aliphatic and aromatic classes, and the charged class into positively and negatively charged classes. GM5 and GM6 are based on side chains of amino acid residues referred to in Espadaler's study (Espadaler *et al*., 2005). Cysteine (C) was placed in its own class because of its propensity to form disulfide bridges.

Alternatively, in order to provide quantitatively continuous scores for amino acid differences, this study employed substitution matrices and some amino acid pair distances. The most common substitution matrices, PAM250 and BLOSUM62—abbreviated as PAM and BLO, respectively—were modified to PAM.A and BLO.A by replacing the diagonal values with 0. The original diagonal of the matrix represents the identities, which means there are no differences between the two amino acid residues. Furthermore, the negative values in PAM and BLO represent the seldom-observed substitutions or dissimilarity of amino acids; but the positive values represent the frequently observed substitutions or similarity of amino acids. Therefore, PAM.B and BLO.B were produced by replacing the positive values in PAM and BLO with 0 and making the negative values become positive. In addition, three matrices associated with amino acid differences are used in this study. In the Miyata matrix, the amino acid pair distance was made according to hydrophobicity and volume of amino acid residues, etc. (Miyata *et al*., 1979). The Collins matrix was obtained from taking the logarithm of relative mutability of the 20 amino acid residues (Collins and Jukes, 1994). In the Grantham matrix, the chemical distance for each amino acid pair was constructed based on the composition, polarity and molecular volume of amino acid residues (Grantham, 1974). Each pair-wise amino acid sequence alignment can be converted to a 1×329 vector based on the matrices described earlier and then can be used for further analysis.

2.2.2 Analysis methods The pair-wise sequence comparisons were converted to binary and continuous scores according to the scoring methods described in Section 2.2.1. Their corresponding antigenic distances were considered as the continuous dependent variables in regression analyses on amino acid changes according to different scoring and grouping methods. One can also assign two categories—S (similar viruses) or V (antigenic variants)—according to the antigenic distance with the cutoff value 4 (<4 or $\geqq 4$) (Lee and Chen, 2004). The purpose of the analysis is to find potential immunodominant positions that can predict antigenicity with a high agreement rate. In this study, the potential immunodominant positions are defined as amino acid substitutions that are positively associated with antigenic distances. Four statistical and machine-learning methods were used in this study for analyzing the pair-wise scoring vectors.

Iterative filtering algorithm For each amino acid position in the sequences, the count of 1 in antigenic variants minus the corresponding count of 1 in similar viruses will be used to rank the amino acid positions in decreasing order. The amino acid position with rank one is put into a model, with the remaining amino acid positions tested for their contribution to the model. Each position is put into the model depending on whether it can improve the performance. Whenever one position has entered the model, other positions in the model will be examined again and remain in the model as long as their existence contributes to the performance. After this iterative procedure, some amino acid positions with the maximum agreement rate were kept in the model.

Multiple regression The pair-wise comparisons of amino acid sequences obtained from scoring methods for each amino acid position and logarithmic transformation of their corresponding antigenic distances (lnAD) could be deemed independent variables and dependent variables, respectively, in stepwise multiple regression. Linear combinations of the amino acid positions were, therefore, obtained for prediction of antigenic distances. SPSS 13.0 (SPSS Inc., Chicago, Illinois) was used to carry out stepwise multiple regressions. The predicted value is then classified as S or V using ln(4) as the cutoff.

Logistic regression The pair-wise comparisons of amino acid sequences obtained from scoring methods for each amino acid position and their corresponding probability of antigenic variants (0 for similar viruses and 1 for antigenic variants) were used in logistic regression. The backward conditional method was employed with SPSS.

Support vector machine (SVM) The pair-wise comparisons of amino acid sequences obtained from scoring methods were used as inputs for SVM. Classification problem of predicting S or V was solved by SVM. The SVM package in R language is in a package called e1071.

2.2.3 Testing strategy and performance comparison The models combining the scoring and analysis methods were built based on the training dataset and then applied to test the validation dataset. Antigenic variants and similar viruses are predicted based on proposed models. Agreement rate, sensitivity and specificity were calculated for comparisons of different models, using the serologically antigenic distances as the gold standard. The agreement rate is defined as the ratio of all truly predicted pairs to the number of all virus pairs. The ratio of predicted variants to true variants and the ratio of predicted similar viruses to true similar viruses multiplied by 100% are defined as sensitivity and specificity, respectively.

2.2.4 Applying the model to flu sequences without serology data We applied our model to human influenza A/H3N2 sequence data downloaded from influenza sequence database [ISD, (Macken *et al*., 2001)]. After aligning sequences by MUSCLE (Edgar, 2004) and removing those with a length shorter than 329 residues, 3098 sequences were used for further analysis. Virus sequences were compared with WHO-recommended vaccine strain sequences based on their years of isolation and then predicted as antigenic variants or similar viruses by our model. Percentage of predicted antigenic variants was obtained in each year. Fisher's exact test was used to test whether there is any relationship between vaccine strain update and the percentage of predicted antigenic variants in the previous year.

3 RESULTS

In the previous study, the non-grouping method was used to transform the pair-wise amino acid sequence comparisons into

Table 1. Results of analysis of binary data from different grouping methods

Analysis method	Grouping method	No. of selected positions	Training dataset ($N = 181$)			Validation dataset ($N = 96$)		
			Agreement	Sensitivity	Specificity	Agreement	Sensitivity	Specificity
Iterative filtering	Non-grouping	9	89.50	95.20	76.79	54.17	100.0	42.86
	GM1	11	91.71	95.20	83.93	76.04	100.0	70.13
	GM2	9	91.71	94.40	85.71	76.04	100.0	70.13
	GM3	9	89.50	93.60	80.36	76.04	100.0	70.13
	GM4	9	89.50	93.60	80.36	76.04	100.0	70.13
	GM5	10	90.06	94.40	80.36	76.04	100.0	70.13
	GM6	9	91.71	94.40	85.71	76.04	100.0	70.13
Multiple regression	Non-grouping	19	91.16	90.40	92.86	82.29	94.74	79.22
	GM1	19	86.74	85.60	89.29	89.58	68.42	94.81
	GM2	19	86.74	84.80	91.07	89.58	84.21	90.91
	GM3	23	90.06	88.00	94.64	91.67	84.21	93.51
	GM4	23	91.16	89.60	94.64	91.67	84.21	93.51
	GM5	20	91.71	90.40	94.64	91.67	84.21	93.51
	GM6	19	90.61	88.80	94.64	82.29	94.74	79.22
Logistic regression	Non-grouping	25	93.92	96.00	89.29	78.13	78.95	77.92
	GM1	19	93.37	96.80	85.71	71.88	84.21	68.83
	GM2	19	92.82	96.80	83.93	73.96	84.21	71.43
	GM3	19	93.37	96.80	85.71	68.75	89.47	63.63
	GM4	23	94.48	96.00	91.07	68.75	94.74	62.34
	GM5	21	93.92	95.20	91.07	60.42	100.0	50.65
	GM6	21	94.48	94.40	94.64	79.17	84.21	77.92
Support vector machine	Non-grouping	45	93.92	96.80	87.50	78.13	100.0	72.73
	GM1	35	93.37	96.00	87.50	80.21	94.74	76.62
	GM2	35	92.82	96.00	85.71	80.21	94.74	76.62
	GM3	25	93.92	96.00	89.29	80.21	94.74	76.62
	GM4	35	93.37	95.20	89.29	79.17	84.21	77.92
	GM5	40	93.37	95.20	89.29	80.21	94.74	76.62
	GM6	28	93.92	95.20	91.07	80.21	94.74	76.62

0 and 1 (Lee and Chen, 2004). In this study, we combined the concept of similarity classes of amino acids with the models proposed by Lee and Chen, but the agreement rate was raised only from 82.9 to 84.0% for their best model. We would like to further establish bioinformatics models to refine the predictive power.

The reaction of virus strains is compared in a HI test with that of a panel of reference virus strains using specific antisera raised against these viruses by infection of ferrets. These serological studies were undertaken in collaborative laboratories within the network of the WHO. In this study, we collected the 181 pair-wise antigenic distances to build the models and then intended to predict whether the new virus strains in the validation dataset are drift variants to the reference strains. Two reference strains, A/Panama/2007/99 and A/Fujian/411/2002, were both included in the training and validation datasets. In the validation dataset, 29 pairs were associated with these two reference strains and 67 pairs were pair-wise comparisons among the other 17 virus strains isolated in 2002–2004.

In order to identify the potential immunodominant positions, the four analysis methods described in Section 2.2.2 were combined with different grouping methods. The results are shown in Table 1. Using the non-grouping and the iterative filtering algorithm, the agreement rates were 89.50 and 54.17% in the training and validation datasets, respectively. Although the grouping methods GM1, GM2 and GM6 followed by iterative filtering had the same performances in the training and validation datasets, GM6 was found to be the best scoring method prior to iterative filtering by using 10-fold cross-validation that merges the training and validation datasets together. The procedure was repeated a thousand times to evaluate agreement rates for various grouping methods. A *t*-statistic was used to test whether the mean agreement rate obtained from GM6 is higher than that of other methods. The *P*-values of the *t*-tests for the comparisons of GM6 with GM1 and GM2 were $4.8 \times 10^{-7}$and 1.1×10^{-6}, respectively. Therefore, GM6 is an appropriate scoring method prior to iterative filtering.

Using the regression method, some amino acid positions with negative regression coefficients do not meet the definition of potential immunodominant positions. The coefficients of AA88 and AA262 were -1.040 and –0.598, respectively, in the model that combines the non-grouping method and multiple regression analysis. Nevertheless, the grouping methods could reduce the noise and all coefficients of selected positions from multiple regressions were positive. The model combining the regression method and the GM5 has the best performance (agreement rates are 91.71 and 91.67% in the training and validation datasets, respectively). Apparently, the grouping methods could make the pair-wise amino acid comparisons clearer and more appropriately be applied to further analysis. In addition, we divided the training dataset (181 pair-wise comparisons) into 5 sub-datasets representing different degrees of sequence similarity (as shown in the Supplementary Data). The five sub-datasets have similar numbers of pair-wise comparisons (range 29–43). The agreement rates of the five sub-datasets in the training and validation datasets were 84–100% and 70–91%, respectively. Proportion of antigenic variants in the sub-datasets is highly correlated with average number of amino acid differences. The high sequence similarity (1 ~ 5 amino acid differences) seems to have lower agreement rates in the training and validation datasets. However, caution needs to be taken interpreting the results due to the small numbers of amino acid differences in pair-wise comparisons and the low proportion of antigenic variants.

Although the logistic regression had the good agreement rates of 93–94% in the training dataset, it had poor predictive power for the validation dataset—probably because of overfitting. SVM following grouping methods could be a good predictor and the linear kernel made SVM a simple model that is one kind of linear combination of amino acid positions. The number of amino acid positions selected by SVM, however, exceeded 25, many more than in the other methods. The amino acid positions selected by each method are shown in the Supplementary Data.

In addition to the binary data obtained from different grouping methods, the continuous data transformed from various matrices—e.g. PAM, BLO, Miyata, Collins and Grantham—were further employed for analyses. Multiple regression, logistic regression and SVM were used to analyze the continuous data transformed from these matrices. However, there were serious overfittings in logistic regression and SVM (data not shown), so we present only the results of multiple regression (Fig. 1). Unexpectedly, the predictive power of the modified substitution matrices is worse than that of the original substitution matrices. In spite of the meaningful modification, the positive scores in the original substitution matrices were replaced to zero and thus less information was provided in multiple regressions and led to poor results.

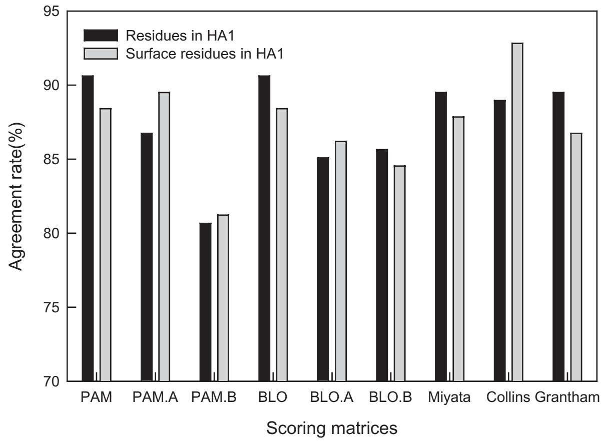

Fig. 1. Agreement rates of multiple regression models using different scoring matrices.

Table 2. Regression coefficients of amino acid positions using the grouping method 4 (GM4) and 5 (GM5) combined with the multiple regression of surface positions

Position	Antigenic domain	Residue frequency among the 45 viruses in the training dataset	GM4	GM5
AA82	E	17 (E), 28 (K)	0.998	1.037
AA92	E	1 (E), 44 (K)	0.941	0.920
AA121	D	28 (I), 10 (N), 7 (T)	0.495	0.546
AA124	A	13 (D), 20 (G), 2 (N), 10 (S)	0.298	–
AA129	B	1 (E), 44 (G)	1.748	–
AA135	A	1 (E), 23 (G), 6 (K), 15 (T)	0.954	1.021
AA144	A	13 (D), 3 (I), 5 (N), 24 (V)	0.716	0.683
AA145	A	1 (I), 18 (K), 22 (N), 1 (R), 3 (S)	1.209	1.282
AA155	B	30 (H), 2 (T), 13 (Y)	1.202	1.582
AA156	B	8 (E), 1 (H), 27 (K), 9 (Q)	0.400	0.294
AA157	B	26 (L), 19 (S)	0.423	0.448
AA158	B	29 (E), 7 (G), 9 (K)	0.761	0.715
AA160	B	1 (A), 35 (K), 1 (R), 1 (S), 7 (T)	1.072	1.073
AA173	D	34 (K), 11 (N)	1.285	1.301
AA174	D	40 (F), 4 (S), 1 (V)	0.613	0.633
AA188	B	42 (D), 1 (E), 1 (N), 1 (Y)	1.087	1.234
AA189	B	8 (K), 5 (Q), 8 (R), 24 (S)	0.721	0.684
AA240	D	44 (G), 1 (R)	0.690	0.708
AA273	C	44 (P), 1 (S)	0.779	0.738
AA276	C	9 (K), 14 (N), 22 (T)	1.830	2.287
Agreement rate in the training dataset ($N = 181$)			93.37%	92.82%
Agreement rate in the validation dataset ($N = 96$)			91.67%	91.67%

Because the amino acid residues involved in antibody binding are likely to form the exposed surface of the protein, the surface positions of HA1 were selected for further analysis. There are 148 amino acid positions on surface of the trimer form of HA1. These 148 positions were selected for further analysis; the results are shown in Figure 1. The agreement rates for the training dataset in the Miyata, Collins and Grantham scoring matrices are similar; the introduction of surface positions, however, could improve the agreement from 88.95 to 92.82% in the Collins matrix. Despite the good performance obtained from combining the multiple regression analysis and the continuous scoring matrices, the grouping method was determinedly used in our model, since it integrates meaningful biological information and outperforms substitution matrices. After thorough testing, a feasible strategy had the best agreement rates: 93.37 and 91.67% in the training and validation datasets, respectively (Table 2). This strategy is described as follows: the GM4 was used to convert pair-wise amino acid comparisons into binary data, and surface positions were then selected for stepwise multiple regression. The GM5 was also used for the same process, producing a 92.82% agreement rate in the training dataset. The regression coefficients of amino

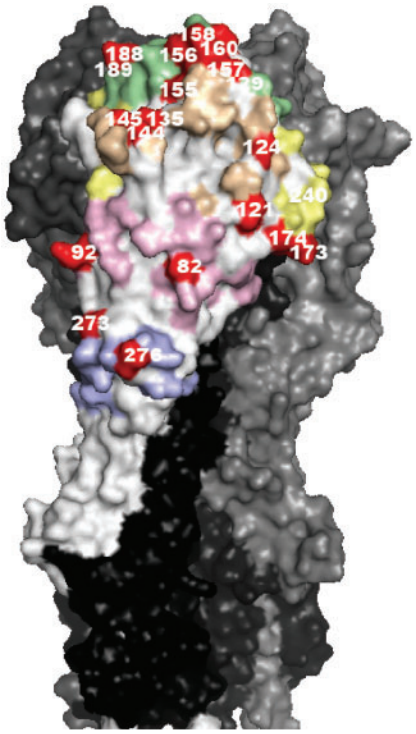

Fig. 2. Influenza virus hemagglutinin in its trimer form (pdb: 2HMG). Twenty potential immunodominant positions identified by our models are highlighted in red and all of them are in the five antigenic domains that are shown in different colors. Antigenic site A (orange), site B (green), site C (blue), site D (yellow) and site E (pink).

acid positions obtained by the models are shown in Table 2. These 18–20 positions are all in the antigenic sites, as shown in 3D structure of hemagglutinin (Fig. 2), and are believed to be potential immunodominant positions.

We applied our model to the sequence data downloaded from ISD, and predicted whether the viruses are antigenic variants of the WHO-recommended vaccine strain by making use of the pair-wise sequence comparisons. The percentage of predicted antigenic variants was thus obtained in each year. The results, shown in the Supplementary Data, may reflect the association between the percentage of antigenic variants and the vaccine strain update in the following year (P-value = 0.07, Fisher's exact test). However, these samples may not be good representations of total populations and some years have relative small sample sizes (<50 sequences), which need to be taken into consideration for interpreting the results.

4 DISCUSSION

Currently, the serum HI assays is the most common way to detect antigenic variants, and global influenza surveillance relies heavily on this technique. It is, however, labor-intensive and time-consuming. Meanwhile, phylogenetic analysis is widely used to elucidate genetic relatedness. A neighbor-joining phylogenetic tree of the 62 HA1 amino acid sequences in the training and validation datasets, as shown in Figure 3, was constructed by using the Web-based phylogenetic analysis tool named POWER (Lin *et al.*, 2005). Phylogenetic tree analysis is a useful tool to construct genetic relatedness; neighboring strains in the phylogenetic tree are usually regarded as genetically similar viruses. For example, the antigenic distances between A/HongKong/23/92 (HK92) and A/Beijing/32/92 (BEI92), A/Alaska/10/95 (ALA95) and A/Fujian/47/96 (FUJ96) and A/Nanchang/933/95 (NCH95) and A/Wuhan/359/95 (WHN95) were 2.00, 2.00 and 1.00, respectively, which are successfully indicated by the phylogenetic analysis. Two pairs, however—A/Bangkok/1/79 (BAN179) and A/Bangkok/2/79 (BAN279), and A/Shanghai/11/87 (SHA87) and A/Victoria/7/87 (VIC87)—had antigenic distances of 9.24 and 5.66, respectively, making them antigenic variants; yet these were grouped together in the phylogenetic tree. Moreover, although A/Sydney/05/97 (SYD97) and A/Panama/2007/99 (PAN99) are antigenically similar viruses (antigenic distance 1.41), they are separated widely in the phylogenetic tree (12 amino acid residues difference). These examples reveal

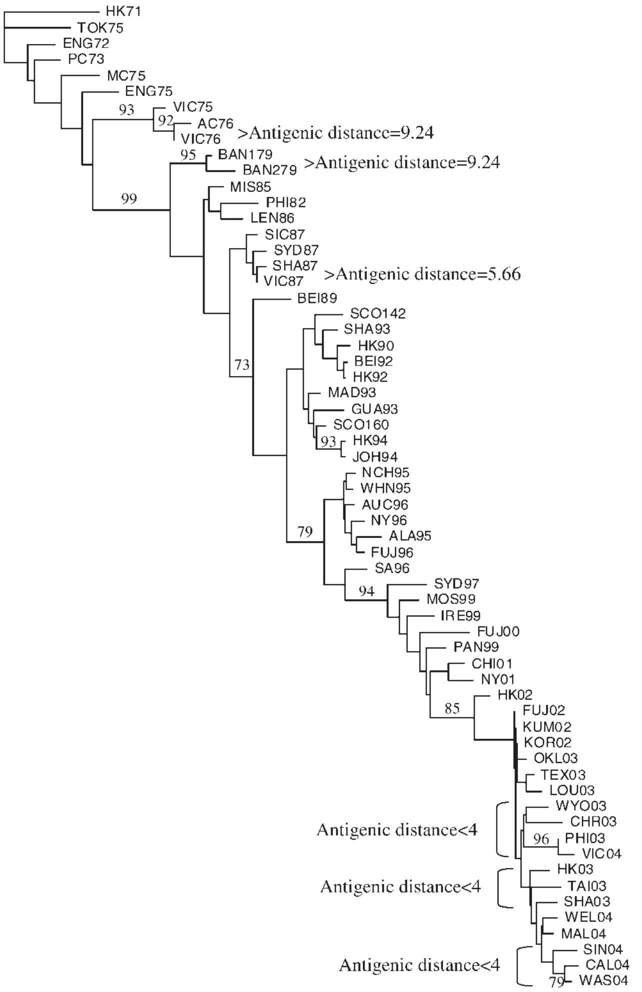

Fig. 3. Neighbor-joining phylogenetic tree of the 62 influenza H3N2 viruses (HA1 protein with 329 amino acid residues). The full names of viruses are in the Supplementary Data. Numbers at branch nodes refer to the percentage of 1000 bootstrap repetitions, and only those occurring at a percentage >70% are shown.

that some amino acid positions are more important than others; thus, potential immunodominant positions need to be identified. Although molecular sequencing is high-throughput and robust, the phylogenetic tree based on sequence analysis still cannot confidently predict antigenic variants because it lacks a good criterion to determine how the distance between two viruses can be related to antigenic variants.

In this study, we combine the sequence comparisons and antigenic distances to develop the bioinformatics models for prediction of antigenic variants. In sequence comparisons, we use various grouping and scoring methods. The amino acid positions identified in the iterative filtering method were weighted equally; thus, antigenic variant can easily be predicted according to whether their amino acid residues mutate in these positions. In the model based on the iterative filtering algorithm, around 10 amino acid positions were required to determine the antigenic variants with around a 90% agreement rate (Table 1). It is quite apparent that the non-grouping method combined with iterative filtering resulted in the worst agreement in the validation dataset. Looking into the false prediction of the validation dataset, we found that it was due mainly to the amino acid position AA189. The viruses with the amino acid residue serine (S) in AA189 mutated to asparagine (N) and thus were predicted falsely to be antigenic variants. In contrast, the grouping methods placed serine and asparagine into the same class based on their polarity and thus could resolve this misclassification problem and had better agreement rates in the validation dataset. Although the iterative filtering had only 76% agreement rates in the validation dataset, the extremely high sensitivity still helps us identify antigenic variants more easily.

In addition to the substitution matrices, three matrices associated with amino acid differences (Miyata, Collins and Grantham) were employed in this study. The three matrices have different scores. The identical amino acid residues were assigned as 0; but the most dissimilar pairs of amino acids—glycine (G) to tryptophan (W), cysteine (C) to threonine (T) and cysteine (C) to tryptophan (W)—were assigned 5.13, 0.60 and 215 in Miyata, Collins and Grantham matrices, respectively. The agreement rates of multiple regression analyses of continuous data obtained from substitution matrices PAM and BLO were around 90% (Fig. 1); but it is difficult to explain the coefficients of the selected positions. In addition, Blosum90 and PAM6.4–8.7 were also applied here because the hemagglutinins of influenza A have a 90% sequence identity within subtypes (Knossow and Skehel, 2006). The agreement rates of multiple regressions of these continuous data were <90%. As for the other three matrices associated with amino acid differences, their agreement rates are nearly the same with the grouping methods. As a consequence, the grouping methods were determinedly used in our model.

Stepwise multiple regression and backward conditional logistic regression were employed to select the amino acid positions and to weight differently on each position. Although influenza viruses change continuously to escape from immunity, the mutations of amino acid residues on HA1 are functional and structural constraints for the stability of proteins. The net charge of protein critically affects on protein structure and thus the adjacent residues mutate together. For example, the amino acid residue changed from glutamic acid (E) to lysine (K) in AA82 but from K to E in AA83. GM1, GM2 and GM6 grouped the charged amino acids in the same class and had the better agreements than GM3, GM4 and GM5 in use of the iterative filtering method; this is because iterative filtering following GM3–GM5 picked the position AA2 that lies far away from the regions of antibody combination and is buried inside the folded trimer. However, the agreement rates in multiple regressions had the contrary trends. To divide the charged amino acids into positively charged class and negatively charged class could provide more information to multiple regression analysis for achieving a better agreement rate in prediction. The results imply that different grouping methods may suit different analysis methods. Cross-validation was repeated a 100 times for evaluating the difference in predictive performance of the two different models: the model combining iterative filtering and GM6, and the model combining the regression method and GM4. Their testing agreement rates were 84.46 ± 6.37 and 88.39 ± 6.12%, respectively. In bioinformatics views, one can make use of the grouping methods combined with iterative filtering, multiple regression or SVM to predict antigenic variants; however, we mainly focus on multiple regression to explore the biological meaning of identified positions.

We intended to explore the antigenic drifts over the last 35 years; but the short supply of serological data may limit the predictability of our models. Nevertheless, the high predictive performance of testing the validation dataset demonstrated that our models could predict antigenic variants prospectively. In addition, we merged the validation dataset with the training dataset to go through the same process: GM4 was used to convert pair-wise amino acid comparisons into binary data, and surface positions were then selected for stepwise multiple regression. Twenty-two amino acid positions were identified with a 92.06% agreement rate; and only two positions (AA193 and AA227) were newly identified, which indicates this model is stable. Finkenstadt and colleagues have stated that antigenic variant strains usually appear first in the southern hemisphere and that this information would be available to predict the following epidemic year of northern hemisphere (Finkenstadt *et al.*, 2005). In addition, mutations at new positions occur frequently in the HA of influenza A/H3N2 viruses (Lee *et al.*, 2007). Therefore, in order to increase the predictive power for facilitating vaccine-strain selection, new cross-reactivity data based on recent viruses need to be added to the training dataset continuously.

Although the model was built based on the 181 pair-wise sequence comparisons among 45 viruses isolated during 1971–2002, it is useful for prediction of antigenic variants. In addition to A/Fujian/411/2002 that was included in the training dataset, the circulating viruses, e.g. A/Wyoming/3/2003 and A/Wellington/1/2004, were antigenic variants in comparison with A/Panama/2007/99, which widely used as the vaccine strain during 2000–2003. Our model successfully predicted these events.

According to the documentation of vaccine-strain selection, vaccines containing A/Panama/2007/99 antigens stimulated anti-HA antibodies, which were lower titer to A/Wyoming/3/2003 (A/Fujian/411/2002-like) and A/Wellington/1/2004 viruses than to the vaccine virus (http://www.flu.lanl.gov/vaccine/). Therefore, A/Fujian/411/2002-like and A/Wellington/1/2004 had been considered as southern hemisphere vaccine strains in 2004 and 2005, respectively (Macken *et al*., 2001). A recent experimental study found that mutations at amino acid positions AA155 and AA156 are the molecular determinants of the antigenic drift from A/Panama/2007/99 to A/Fujian/411/2002-like viruses (Jin *et al*., 2005). The amino acid residues in these two positions of HA1 protein of A/Fujian-like viruses had complementary changes of charged property, H155T and Q156H, and thus their structure are stabilized. In our model, the regression coefficients of AA155 and AA156 are 1.202 and 0.400, respectively, which indicates simultaneous changes of these two amino acid residues could cause antigenic drift ($\mathrm{ln}AD = 1.602 > \mathrm{ln}4$). The threshold used to identify antigenic variants as four is not arbitrary, which has some biological reasons. In the literatures, a virus that shows 4-fold or greater difference to reference antisera is considered an antigenic variant (Besselaar *et al*., 2004; Daum *et al*., 2005; Ellis *et al*., 1997). In order to evaluate whether our models are dependent on the threshold, square roots of 8 and 32 (i.e. 2.828 and 5.657) were used as thresholds to define antigenic variants. The agreement rates of using thresholds at 2.828 and 5.657 were 89.50 and 90.06%, respectively, which are not significantly different from that using threshold at four (*P*-values = 0.121 and 0.211, McNemar test). Therefore, this result revealed that our models are not significantly dependent on the threshold.

In examination of the prediction results from 96 pair-wise comparisons in the validation dataset, our different models (iterative filtering, multiple regression and support vector machine) all predicted five pairs variants, which were inconsistent with the collected validation dataset. A/Texas/40/2003 and A/Oklahoma/8/2003 were similar to A/Panama/2007/99 according the serological data (as shown in Supplementary Data); however, our models predicted A/Texas/40/2003 and A/Oklahoma/8/2003 antigenic variants of A/Panama/2007/99. Daum *et al*. performed antigenic analyses and the result revealed that the antigenic distances between A/Texas and A/Panama as well as between A/Oklahoma and A/Panama were both 5.66 (Daum *et al*., 2005), which agreed with our model prediction. In addition, our models predicted A/California/7/2004 as an antigenic variant of A/Korea/770/2002, A/Wyoming/3/2003 and A/Wellington/1/2004. The prediction results suggested that A/Wyoming/3/2003 and A/Wellington/1/2004 were no longer suitable vaccine strains when A/California/7/2004 emerged predominantly. In February 2005, WHO recommended inclusion of A/California/7/2004 in the trivalent influenza vaccine. These examples indicated that the HI assay might not be sensitive because of laboratory variability. Since neutralization arrays for detecting influenza virus antibody responses are more sensitive than the HI assay, they may be the better data source.

Gupta *et al*. reported that their new definition of antigenic distance—pepitope, based on the pair-wise sequence comparison—is highly correlated with vaccine efficacy (Gupta *et al*., 2006). However, they only used 19 pair-wise sequence comparisons among 18 viruses to demonstrate the correlation between pepitope and vaccine efficacies. Although they made use of the pair-wise sequence comparisons to estimate the vaccine efficacy by their model, the evidence is insufficient due to the limited data.

Smith *et al*. provided an antigenic and genetic map of influenza H3 viruses and found 45 amino acid positions related to cluster-difference substitutions (Smith *et al*., 2004). Out of the 20 amino acid positions listed in Table 2, we identified 16 amino acid positions contributing to cluster transition (Table 3). The four amino acid positions—AA92, AA129, AA240 and AA273—were excluded because only one amino acid residue was substituted in these positions among 45 viruses in the training dataset in our study. Based on regression coefficients of these 16 amino acid positions obtained from multiple regression, 31878 pair-wise comparisons among 253 viral isolates belonging to 11 clusters [data extracted from the Supplementary Data (Smith *et al*., 2004)] were tested exhaustively. We assumed that a pair of viruses sampled from different clusters or from the same cluster were antigenic variants or similar viruses. The result showed that a 97.08% (30973/31878) agreement rate was reached by using 16 positions. The highly

Table 3. Amino acid positions related to cluster transition

Cluster transition	Cluster-difference substitution	
	Smith *et al*., 2004	This study
HK68-EN72	T122N, G144D, T155Y, N188D, R207K	G144D, N188D
EN72-VI75	N53D, R102K, N137S, S145N, L164Q, F174S, Q189K, S193D, I213V, I217V, I230V, I278S	F174S, Q189K
VI75-TX77	K50R, D53N, E82K, S137Y, G158E, Q164L, S174F, D193N, K201R, V213I, V230I, M260I	E82K, G158E, S174F
TX77-BK79	N53D, N54S, I62K, K82E, N133S, P143S, G146S, K156E, T160K, D172G, Q197R, V217I, V244L	K82E, K156E, T160K, N173K
BK79-SI87	G124D, Y155H, K189R	E82K, G124D, Y155H
SI87-BE89	N145K	G135E, N145K
BE89-BE92	S133D, K145N, E156K, E190D, T262N	K145N, E156K, S157L, R189S
BE92-WU95	N145K	I121T, D124G, N145K
WU95-SY97	K62E, K156Q, E158K, V196A, N276K	G124S, K156Q, E158K, N276K
SY97-FU02	V25I, R50G, H75Q, E83K, A131T, H155T, Q156H, V202I, W222R, G225D	H155T, Q156H

consistent result revealed that our model indeed reduced the positions related to cluster transitions to a small number and could still stand comparison with the antigenic map. We also used the 20 positions as shown in Table 2 for an exhaustive sampling study and got 89.89% (28654/31878) of agreement rate. It is hard to compare the performances of two studies because Smith *et al.* used a *k*-means clustering algorithm to determine the clusters in the antigenic map, which are not always consistent with antigenic distances. Ideally, the real performances should be compared by using the same serological data. Our models provide a feasible strategy for incorporating bioinformatics scoring in statistical methods and further identifying the potential immunodominant positions for predicting antigenic variants, which could be readily integrated to the global influenza surveillance system. Moreover, the methods could also be applied to other highly mutable viruses (Lipsitch and O'Hagan J, 2007).

ACKNOWLEDGEMENTS

This work is supported by the Centers for Disease Control, Taiwan (DOH95-DC-1401) and the National Research Program for Genomic Medicine, National Science Council, Taiwan (NSC95-3112-B-400-005).

Conflict of Interest: none declared.

REFERENCES

Besselaar,T.G. *et al.* (2004) Antigenic and molecular analysis of influenza A (H3N2) virus strains isolated from a localised influenza outbreak in South Africa in 2003. *J. Med. Virol.*, **73**, 71–78.

Bush,R.M. *et al.* (1999a) Predicting the evolution of human influenza A. *Science*, **286**, 1921–1925.

Bush,R.M. *et al.* (1999b) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.*, **16**, 1457–1465.

Collins,D.W. and Jukes,T.H. (1994) Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics*, **20**, 386–396.

Daum,L.T. *et al.* (2005) Influenza A (H3N2) outbreak, Nepal. *Emerg. Infect. Dis.*, **11**, 1186–1191.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Ellis,J.S. *et al.* (1997) Analysis of influenza A H3N2 strains isolated in England during 1995–1996 using polymerase chain reaction restriction. *J. Med. Virol.*, **51**, 234–241.

Espadaler,J. *et al.* (2005) Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, **21**, 3360–3368.

Finkenstadt,B.F. *et al.* (2005) Modelling antigenic drift in weekly flu incidence. *Stat. Med.*, **24**, 3447–3461.

Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

Gupta,V. *et al.* (2006) Quantifying influenza vaccine efficacy and antigenic distance. *Vaccine*, **24**, 3881–3888.

Jin,H. *et al.* (2005) Two residues in the hemagglutinin of A/Fujian/411/02-like influenza viruses are responsible for antigenic drift from A/Panama/2007/99. *Virology*, **336**, 113–119.

Kilbourne,E.D. *et al.* (1990) Independent and disparate evolution in nature of influenza A virus hemagglutinin and neuraminidase glycoproteins. *Proc. Natl Acad. Sci.*, **87**, 786–790.

Knossow,M. and Skehel,J.J. (2006) Variation and infectivity neutralization in influenza. *Immunology*, **119**, 1–7.

Lee,M.S. and Chen,J.S. (2004) Predicting antigenic variants of influenza A/H3N2 viruses. *Emerg. Infect. Dis.*, **10**, 1385–1390.

Lee,M.S. *et al.* (2007) Identifying potential immunodominant positions and predicting antigenic variants of influenza A/H3N2 viruses. *Vaccine*, **25**, 8133–8139.

Lin,C.Y. *et al.* (2005) POWER: PhylOgenetic WEb Repeater–an integrated and user-optimized framework for biomolecular phylogenetic analysis. *Nucleic Acids Res.*, **33**, W553–W556.

Lipsitch,M. and O'Hagan,J.J. (2007) Patterns of antigenic diversity and the mechanisms that maintain them. *J. R. Soc. Interface*, **4**, 787–802.

Macken,C. *et al.* (2001) The value of a database in surveillance and vaccine selection. In Osterhaus,A.D.M.E. *et al.* (eds) *Options for the Control of Influenza IV*. Elsevier Science, Amsterdam.

Miyata,T. *et al.* (1979) Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–236.

Plotkin,J.B. and Dushoff,J. (2003) Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. *Proc. Natl Acad. Sci. USA*, **100**, 7152–7157.

Smith,D.J. *et al.* (2004) Mapping the antigenic and genetic evolution of influenza virus. *Science*, **305**, 371–376.