Identifying Discriminative Amino Acids Within the Hemagglutinin of Human Influenza A H5N1 Virus Using a Decision Tree

Li-Ching Wu, Jorng-Tzong Horng, *Member, IEEE*, Hsien-Da Huang, and Wei-Long Chen

***Abstract*—Recently, the H5N1 virus has had an increasingly important impact on human life. This is because more and more people are becoming infected with this virus, and the possibility of a serious pandemic with human to human transmission is looming. This might occur if the genome of this influenza virus mutates either by antigenic drift or by antigenic shift, especially if there is a mutation of the hemagglutinin (HA) glycoprotein. The HA is the surface glycoprotein, and it binds to sialic acid of the host cell surface receptor. Thus, the combination of HA and sialic acid are central to whether influenza virus infects humans. In this study, we selected 497 HA protein sequences from the National Center for Biotechnology Information (NCBI) Influenza Resource database, and used a decision tree method to identify discriminative amino acids in the HA protein sequences that may possibly influence the binding of HA to sialic acid. Four such amino acid positions at 54, 55, 241, and 281 were identified and these may play an important role in infection by H5N1 influenza virus.**

***Index Terms*—Decision tree, hemagglutinin (HA), influenza A H5N1 virus.**

I. Introduction

RECENTLY, between 1997 and the present day, outbreaks of the highly pathogenic H5N1 influenza A virus have caused a significant number of human deaths. The infected region in Asia stretches from Japan in the north to Indonesia in the south [1]–[3]. Current evidence suggests that the fatal human cases resulted from direct transmission of virus from birds to humans [1], [2]. Evolution and mutation of the H5N1 virus is high likely by either antigenic drift or antigenic shift and this has the potential, if H5N1 viruses become transmissible from human to human, to cause a dreadful pandemic disaster. This would be because humans are not resistant to this virus and it is likely that rapid spread would occur.

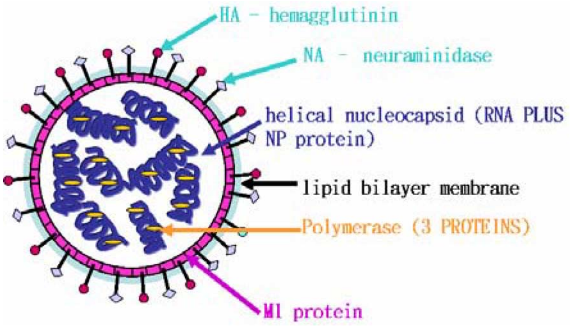


Fig. 1. Diagram of an influenza virus. There are eight RNA segments found within this virus and HA and NA are used to divide the viruses into different subtypes.

Influenza A virus belongs to the orthomyxoviruses, which has eight RNA segments. The virus is divided into different subtypes based on the fourth segment hemagglutinin (HA), and the sixth segment neuraminidase (NA) [4], [5]. The HA forms the spikes at the surface of virions (Fig. 1) [2], [4]. Furthermore, this protein is responsible for virus absorption onto the sialoside receptors, which allows the virus to enter the cell by cell membrane fusion. A number of cases where avian influenza viruses (AIVs) have infected mammals have been reported, including two separate cases of conjunctivitis in humans and epidemic outbreaks in pigs, horses, and seals [6], [7]. With time, any AIV will mutate by either antigenic drift or antigenic shift. Antigenic drift is caused by point mutations and results in a new strain within the original subtype. Antigenic drift occurs in both type A and type B influenza viruses and is caused by genetic recombination that results in a new strain of virus with a new host range perhaps including humans. Antigenic shift occurs only within the same influenza virus type.

In 1997, in Hong Kong, 18 people were infected with the H5N1 influenza virus and six of these people died [5], [8]–[10]. Recently, well after this initial outbreak, from January 2004 to September 2005, H5N1 caused 12 human deaths in Thailand, 38 in Vietnam, and 4 in Cambodia [11]. The evolution of the HA of H5N1 shows that the viruses isolated from birds and humans, between 1997 and 2003 in Hong Kong, were clade 3 and clade 1, respectively. However, the viruses isolated from birds and humans in Vietnam, Thailand, and Cambodia were clade 1 and clade 2, and these are the same viruses that are found in birds from China, Indonesia, and South Korea [2].

Antigenic drift due to amino acid mutation within the HA protein sequence may be able to cause changes in sialic acid

Manuscript received January 17, 2007; revised January 25, 2007. Current version published November 5, 2008. This work was supported in part by the National Science Council, Taiwan, under Contract NSC95–2221-E-008–129 and NSC96–2311-B-008–004-MY2.

L.-C. Wu is with the Institute of Systems Biology and Bioinformatics, National Central University, Jhongli City 320, Taiwan (e-mail: Richard@db.csie.ncu.edu.tw).

J.-T. Horng is with the Department of Computer Science and Information Engineering, Institute of Systems Biology and Bioinformatics, National Central University, Jhongli City 320, Taiwan, and also with the Department of Bioinformatics, Asia University, Taichung County, Taiwan (e-mail: horng@db.csie.ncu.edu.tw).

H.-D. Huang is with the Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan (e-mail: bryan@mail.nctu.edu.tw).

W.-L. Chen is with the Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan (e-mail: sdragon9@msn.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2008.896871

affinity activity with respect to the HA protein [12]. As a result, this might influence infection among humans. Therefore, we have developed an effective method to analyze changes in each amino acid of the HA protein, and see if these might influence affinity to sialic acid. Up to present, most relevant studies have only used a few selected representative H5 protein sequences and used these to identify by projection some important positions like the receptor-binding site or potential glycosylation sites. Using this information, it is possible to identify whether changes at these sites might affect H5 absorption ability. This approach is both costly and inefficient.

Previous research has shown that some amino acids within the HA protein may affect binding with sialic acid and researchers have often used the hemagglutination inhibition (HI) assay and real-time polymerase chain reaction (RT-PCR) to carry out these experiments [1], [10]. Antigenicity was analyzed by the HI assay, and then it was analyzed what the effect of these point mutations inserted into the HA protein sequence by PCR and RT-PCR would be [1], [8], [11]. This has been successful and the results suggested that some amino acid positions are more important than others. Examples are positions 153 and 242 [1], [10], which are within the receptor-binding site, and positions 146 and 175 [8], [10], which are potential glycosylation sites; these sites are projected to influence receptor binding. Using a decision tree approach on a very wide range of HA sequences, we hope to identify more such positions that may influence receptor binding by the H5 protein.

II. Material and Methods

A. Materials

1) Influenza Virus Resource Database: The Influenza Virus Resource presents data obtained from the National Institute of Allergy and Infectious Disease (NIAID) Influenza Genome Sequencing Project as well as from GenBank and these are combined with tools for flu sequence analysis. In addition, it provides links to other resources that contain flu sequences, publications, and general information about flu viruses [13]. In this database, it is possible to obtain a great deal of information on influenza virus, such as protein sequence, RNA sequence, and influenza virus origin. In this study, the HA protein sequences of influenza H5N1 virus were downloaded.

2) NetNGlyc 1.0 Server: The NetNglyc server predicts N-glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn–Xaa–Ser/Thr sequons [14]. Glycosylation is important in the infection of influenza virus, and therefore, the NetNglyc server was used to predict N-glycosylation sites in H5N1 virus, and these potential glycosylation sites were added to our selected features.

3) Weka: Weka is a collection of machine learning algorithms for data mining tasks that contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes [15]. Here, we use the classification tool to analyze H5N1 influenza virus HA protein sequences and select the discriminative amino acids.

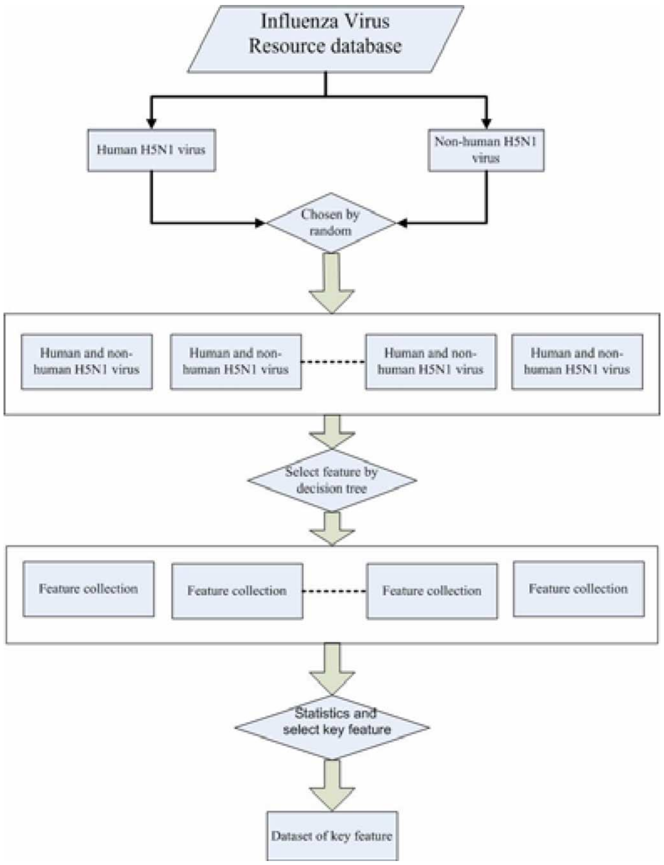
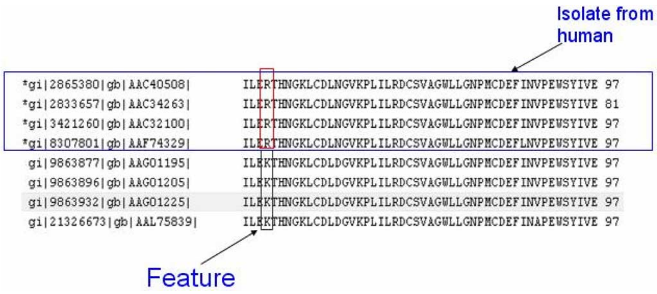

Influenza Virus Resource database
Human H5N1 virus
Non-human H5N1 virus
Chosen by random
Human and non-human H5N1 virus
Human and non-human H5N1 virus
Human and non-human H5N1 virus
Human and non-human H5N1 virus
Select feature by decision tree
Feature collection
Feature collection
Feature collection
Feature collection
Statistics and select key feature
Dataset of key feature

Fig. 2. Analysis flow of selecting discriminative features.

B. Methods

1) Selection of Discriminative Features and the System Flow: First, it was necessary to separate the H5 protein sequence data into two groups, those isolated from human sources and those isolated from nonhuman sources. It is obvious that the result of the decision tree would be deflected toward the nonhuman data group because the number of sequences from this source are many more than those isolated from humans. For instance, if it is supposed 20% of the nonhuman HA protein sequences have arginine at the position 80 and 80% of the human HA protein sequences also have arginine at this position, the ratio at this position favors arginine for the human HA, but if human sequences only make up 10% of the whole database, then the overall ratio will still favor arginine in nonhuman HA. The result will be that arginine at the position 80 is classified as a nonhuman HA trait. Therefore, it was necessary to access these two groups of sequence data separately and randomly in order to balance the dataset in order to allow a correct classification to be made. In order to obtain enough information, the dataset was accessed repeatedly and J48 was used to classify these data and build a decision tree based on the random data. Each of the tree nodes in the decision tree are the features that need to be picked out, and then, the repeated tree nodes from the multiple decision trees are computed. The tree nodes with the highest frequency are finally chosen as the discriminative features. Fig. 2 shows the system flow.



```
*gi|2865380|gb|AAC40508|        ILERTHNGKLCDLNGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE 97
*gi|2833657|gb|AAC34263|        ILERTHNGKLCDLNGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE 81
*gi|3421260|gb|AAC32100|        ILERTHNGKLCDLNGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE 97
*gi|8307801|gb|AAF74329|        ILERTHNGKLCDLNGVKPLILRDCSVAGWLLGNPMCDEFLNVPEWSYIVE 97
 gi|9863877|gb|AAG01195|        ILEKTHNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE 97
 gi|9863896|gb|AAG01205|        ILEKTHNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE 97
 gi|9863932|gb|AAG01225|        ILEKTHNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE 97
 gi|21326673|gb|AAL75839|       ILEKTHNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINAPEWSYIVE 97
```

Fig. 3. Amino acid position selected as a feature when the amino acid is different in different strains.

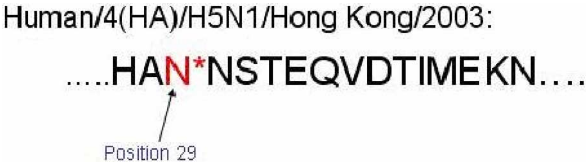
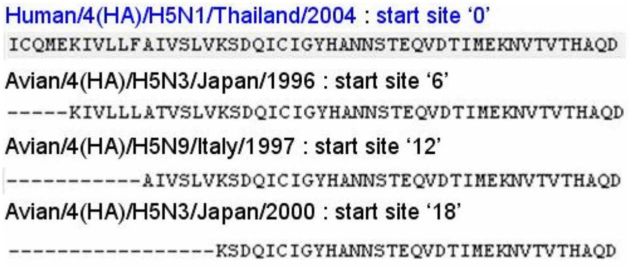


Fig. 4. Marking the potential glycosylation site and inserting it into the feature set.

2) Data Source: We downloaded the HA protein sequence from the Influenza Virus Resource database at NCBI. There were 458 sequences isolated from nonhuman and 39 sequences isolated from humans. We used ClustalW of European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI) to align all of the sequences, and we then selected the sequence Human/4(HA)/H5N1/Thailand/2004 (accession number is AAS65615 in NCBI) and the sequence/Human/4(HA)/H5N1/Hong Kong/1997 (accession number is AAC32098 in NCBI) as templates. The first three amino acids in AAS65615 are not present in AAC32098, and there are seven amino acids in AAC32098 that extend beyond the end of the AAS65615 sequence (see Appendix A).

3) Extract Features: There were 578 amino acid positions in the template after we used the ClusterW [16] to create the alignment. We took the amino acids at each different position as a feature (Fig. 3) and took the different amino acid at each feature as an attribute, for example, in Fig. 3, arginine is one attribute and lysine is another attribute. These results were then filtered to remove features where the amino acid was identical at the same position in order to decrease the number of features analyzed. Furthermore, the letter "X" was used to specify any deletion.

Glycosylation is an important element for the HA binding to sialic acid; we therefore added potential glycosylation sites (Fig. 4) to our feature set and hoped that the potential glycosylation site would be linked to a discriminative feature that we had selected. We used the NetNGlyc 1.0 Server to predict N-link glycosylation sites and these were marked by "$*$" at the potential glycosylation site. Furthermore, we also add the sequence start site (Fig. 5) to the feature collection. Finally, we used the created features set (Fig. 6) to analyze the amino acid structure of the H5N1 virus HA protein.



```
Human/4(HA)/H5N1/Thailand/2004 : start site '0'
ICQMEKIVLLFAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQD
Avian/4(HA)/H5N3/Japan/1996 : start site '6'
-----KIVLLLATVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQD
Avian/4(HA)/H5N9/Italy/1997 : start site '12'
-----------AIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQD
Avian/4(HA)/H5N3/Japan/2000 : start site '18'
-----------------KSDQICIGYHANNSTEQVDTIMEKNVTVTHAQD
```

Fig. 5. Computing the start site of protein sequence and inserting it into the feature set.

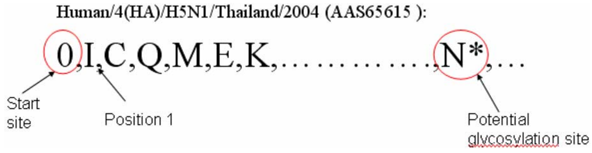

Human/4(HA)/H5N1/Thailand/2004 (AAS65615):
0,I,C,Q,M,E,K,.............,N*,...
Start site
Position 1
Potential glycosylation site

Fig. 6. Example of the feature set extracted.

III. Results

A. *Influenza Virus Worldwide*

We downloaded 539 HA sequences of H5N1 using "any" as the parameter for region/country in order to select strains from worldwide, and then, filtered 42 sequences as they were too short to align with the remaining sequences. This left 497 sequences, which were divisible into 458 HA protein sequences that where isolated from nonhuman sources and 39 HA protein sequences that were isolated from human sources. All 497 HA protein sequences were used in this analysis. We randomly accessed 35 HA sequences from the two pools of 458 non-human HA and the 39 human HA and combined these into individual datasets of sequences. The reason for accessing 35 HA sequences from each pool was first to balance the number of sequences in each dataset and second to avoid over fitting. This process was repeated 1000 times and the datasets was used to build the decision trees using J48. Therefore, in total, 1000 different decision trees were created and the repeated nodes in the decision trees were used as the discriminative features for this study.

In order to confirm whether these discriminative features are stable or not, we accessed twice the number of H5 protein sequences isolated from nonhuman than those isolated from human. Next, the three repeating most highly discriminative features, namely positions 54, 242, and 272, were chosen (Fig. 7). We then analyzed the amino acids at these three positions (Fig. 8). From these results, it was possible to detect obvious differentiation in the proportion of amino acids at the three positions when H5 protein sequences from humans were compared to H5 protein sequences from nonhumans sources.

B. *Hong Kong Influenza Virus*

Next, we reduced the regional range and analyzed the Hong Kong H5N1 influenza virus independently. This dataset included a total of 25 HA protein sequences isolated from humans and 96 HA protein sequences isolated from nonhuman. In contrast to the previous approach, the analysis only identified highly

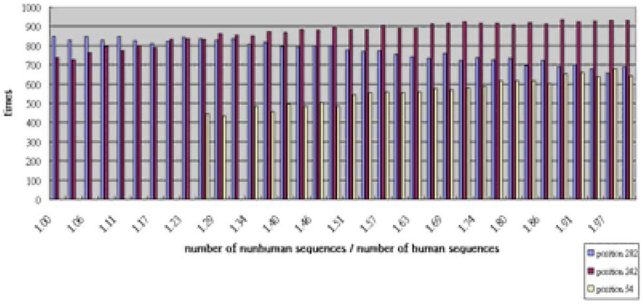


Fig. 7. Selection of three amino acid positions at 54, 242, and 282 in H5. In this figure, it is shown that the stability when accessing twice the number of H5 protein sequences from nonhumans compared to the same number of H5 protein sequences from human sources.

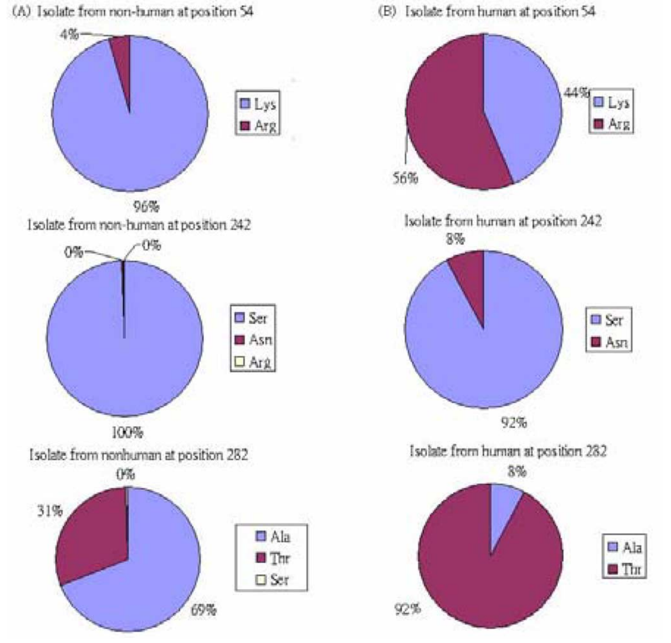


Fig. 8. From the worldwide dataset. (A) Amino acid proportion for H5 sequences isolated from nonhumans at positions 54, 242, and 282. (B) Amino acid proportion for H5 sequences isolated from humans at positions 54, 242, and 282.

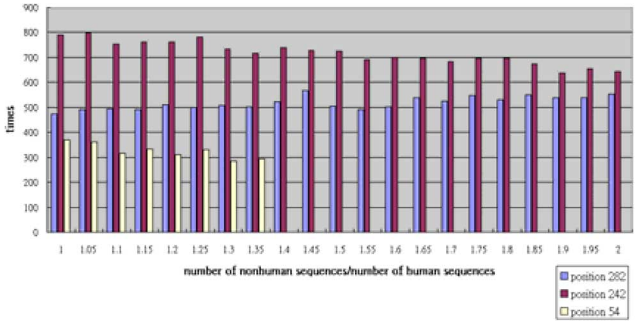


Fig. 9. Position 54 does not show a high enough discrimination for selection. Thus, only positions 242 and 282 were selected as discriminative features from the Hong Kong dataset.

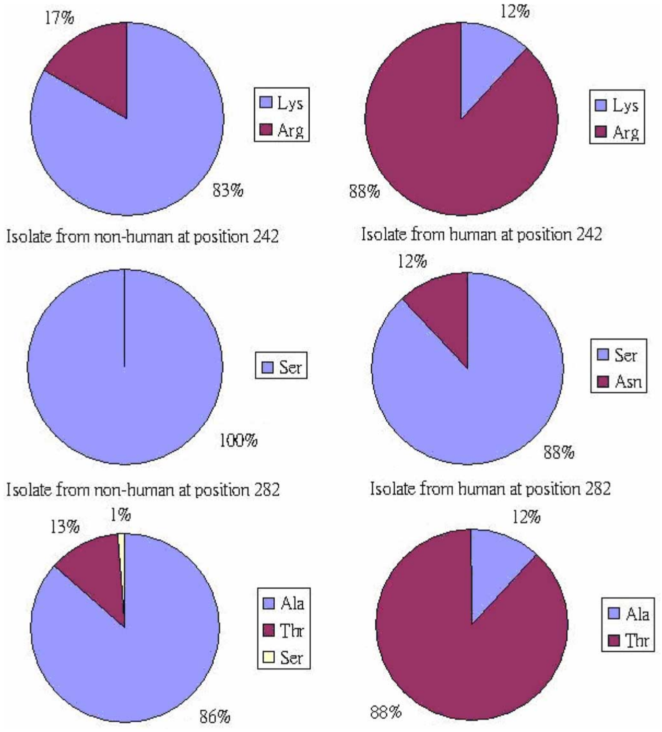


Fig. 10. From the Hong Kong dataset. (A) Amino acid proportion for H5 sequences isolated from nonhumans at positions 54, 242, and 282. (B) Amino acid proportion for H5 sequences isolated from humans at positions 54, 242, and 282.

repeated positions at 242 and 282 (Fig. 9). This was because, unlike the previous results, position 54 failed to show a significant difference between the human and nonhuman samples. However, we still analyzed the proportion of the amino acids at all three positions, and it can bee seen that position 54 is still quite dissimilar when human and nonhuman isolates are compared (Fig. 10). The reason that the position 54 is not selected is because the sequences of position 54 with Lys are mostly from year 2003. In the mean time, these sequences have the same rule that their positions 242 and 282 are all Asn and Ala. Thus, two rules (54 is K) and (242282 are N, A) are equivalent. Now look at the sequences isolated from the nonhuman part, there are lots of sequence having (54 is K). Thus, the rule of (54 is K) will not be selected when the number of sequences of nonhuman source increase as the nonhuman sample increased (Fig. 9). Thus, position 54 is not as significant as positions 242 and 282 in Hong Kong dataset. Based on the fact that a virus isolated in Hong Kong can only be distinguished by positions 242 and 282 when the J48 analysis is used for the classification, it can be suggested that the time and geographic region may influence the discriminative features identified by analysis.

C. Evolution of H5N1

Based on the regional analysis of H5 sequences isolated from humans (Fig. 11) and using the three discriminative features described earlier, it was found that the H5 strains in Hong Kong, Thailand, and Vietnam were significantly different for the features chosen, and we suggest that this is a consequence of virus evolution.

Next, we compared the timelines of the H5 sequences isolated from human at position 54 (Fig. 12). This analysis showed

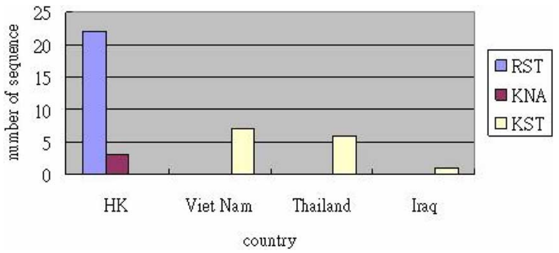


Fig. 11. Comparison of amino acids at positions 54, 242, and 282 is able to discriminate the various clades of the H5. Here, K represents lysine, S represents serine, T represents threonine, R represents arginine, N represents asparagine, and A represents alanine.

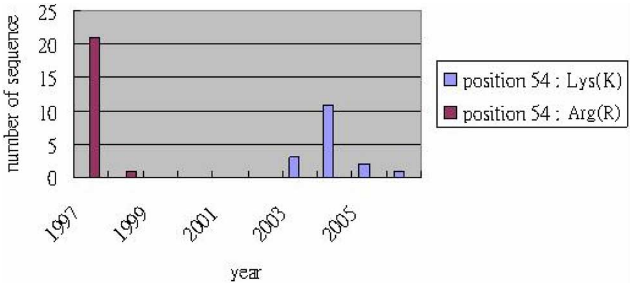


Fig. 12. At position 54, there is a change in the amino acid from arginine (R) in 1997–1998 to lysine (K) in 2003–2006.

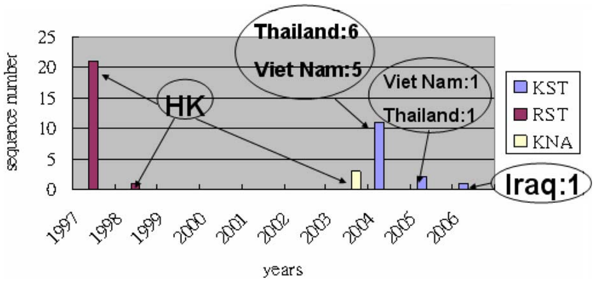


Fig. 13. Evolution of H5N1 virus. Viruses from Hong Kong in 1997–1998 belong to clade 3, viruses from HK in 2003 belong to clade 1 and viruses from Thailand and Vietnam in 2004–2006 belong to clade 1. This figure shows that the different combination of amino acid at positions 54, 242, and 282 are able to classify the viruses into the different clades.

that there was a difference in the amino acids present at position 54 for 1997–1998 compared to 2003–2006. The amino acid at position 54 was arginine during 1997–1998 and lysine during 2003–2006. This suggests that position 54 may be very important in terms of evolution. Based on this, we carried out a full regional and timeline analysis (Fig. 13), and this supports the hypothesis that the three discriminative features play an important role in H5 evolution (Figs. 14 and 15).

Analysis of the amino acid groups: Next, the distribution of amino acids in terms of a Venn diagram was analyzed (Fig. 16). In a similar way, we used a decision tree to implement this analysis. Based on this, we selected position 282 as showing differentiation. At position 282, in the nonhuman group, 69% are alanine, which belongs to the nonpolar group, while on the human group, only 8% are alanine. Furthermore, this compares to 31% being threonine, a polar amino acid, in nonhuman group versus 92% being threonine in human group. We added various characteristics of amino acids, such as hydrophobicity, charge, etc., to our analysis. We observed that a small group was able to discriminate between serine and asparagine at position 242. Therefore, we infer two reasons that may influence binding between HA and sialic acid. One is the amino acid position in the protein and the other is whether the amino acid is polar or not.

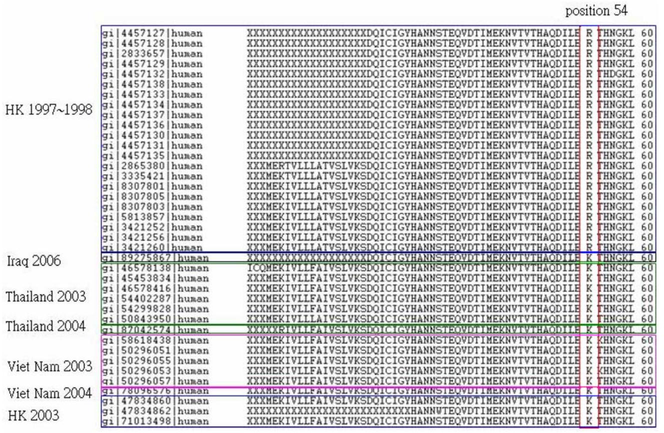


Fig. 14. Alignment showing that the amino acids differ from country to country and from year to year at position 54.

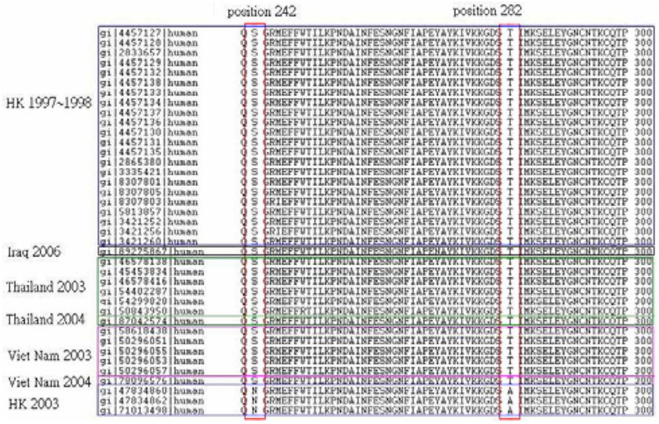
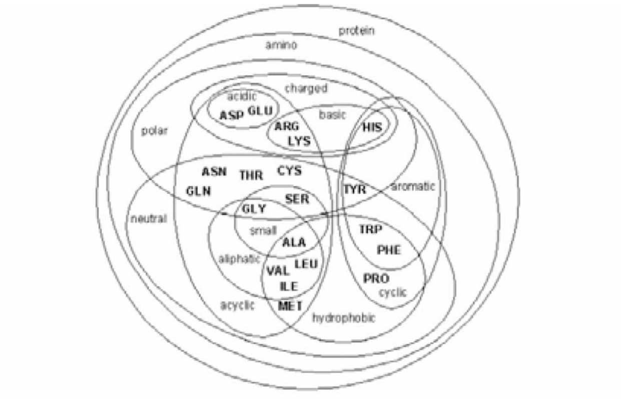


Fig. 15. Alignment showing that the amino acids are differ form country to country and from year to year at position 242 and position 282.



RasMol/Chime Venn diagram v1.2
Drawn by Kurt Giles (kurt@inn-prot.weizmann.ac.il) for the Israeli National Node of EMBnet

Fig. 16. Venn diagram of amino acid groups.

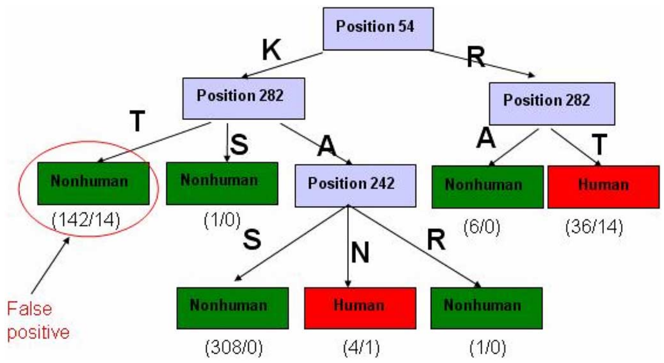


Fig. 17. Decision tree built using three discriminative feature positions 54, 242, and 282. When position 54 is K and position 282 is T, we used 14 H5 proteins isolated from humans and compared them to 45 isolates from nonhumans.

TABLE I
COMPARISON OF THE ACCURACY WITH DIFFERENT CLASSIFIED METHODS USING POSITIONS 64, 242, AND 282

Method	Sensitivity	Specificity
J48	0.625	0.969
ADTree	0.625	0.969
SMO	0.625	0.969

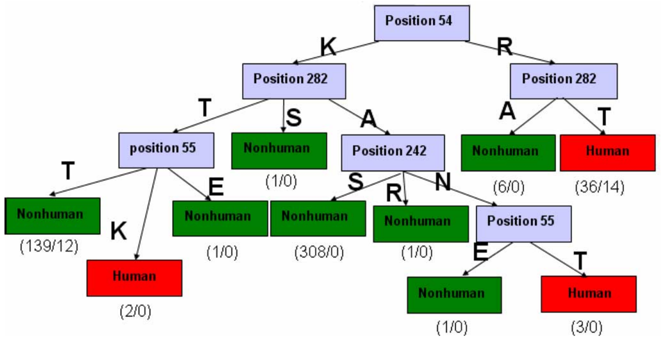


Fig. 18. Position 55 was added to the decision tree and this raises the accuracy when the position 54 is K and 282 is T.

Building the decision tree: First, we have only used the J48 approach to build the decision trees and used the results to identify three discriminative features that are able to classify the virus groups as either human or nonhuman (Fig. 17). This was a limitation, and therefore, we next compared accuracy using three different classification tools (Table I).

We realized that it was not possible to distinguish H5 virus isolated from humans and nonhumans when position 54 is lysine and position 282 is threonine. Therefore, we selected all the sequences that fulfilled these criteria at positions 54 and 282 and used these as a new dataset. Using the same approach as previously, we attempted to identify more discriminative features. This approach gave us positions 55, 90, 194, and 560. We then added these positions to our decision tree, respectively. However, accuracy was only raised when we added position 55 to the decision tree (Fig. 18). The improvement of adding position 55 in sensitivity/specificity is not large since it only separates three special cases. But adding this feature may show the unique amino acid used in position 55 when other features failed to classify.Therefore, in the end, four features were identified and we compared the accuracy using three classification tools (Table II).

TABLE II
COMPARISON OF THE ACCURACY WITH DIFFERENT CLASSIFIED METHOD USING POSITIONS 54, 55, 242, AND 282

Method	Sensitivity	Specificity
J48	0.625	0.969
ADTree	0.659	0.974
SMO	0.643	0.974

IV. DISCUSSION

It is well known that the binding of HA to sialic acid is important for infection. In this study, based on recently isolated H5N1 sequences, we used a decision tree approach to search for possible amino acids in H5 that might influence H5 binding to sialic acid. One of the positions identified is 242, which has been discussed in a number of recent publications [1], [2], and [10]. However, it is likely that single amino acid changes will not strongly influence absorption [10]. Therefore, we have identified four positions that give rise to three combinations, namely (K54, K55, X242, T282), (K54, T55, N242, A282), and (R54, X55, X242, T282) that seem to influence HA binding to sialic acid. In the aforementioned combinations, A is a nonpolar amino acid and T is a polar amino acid at position 282.

Although in the final decision tree, position 54 lay in the tree root, which shows that position 54 is critical; in our Hong Kong dataset, analysis shows that position 54 may be less important than others. The reason position 54 is in the root of the decision tree is that there is a high proportion of human source H5N1 sequence is from Hong Kong between 1997 and 1998. These sequences have the feature that position 54 is Arg. This leads to a mathematical paradox, that is, most human source H5N1 sequences have Arg at position 54 (statistically). But sequences after year 2000 all have Lys at position 54 (shown in Fig. 12). Since there are 17% nonhuman source H5N1 sequences that have Arg at position 54, this might be just a case of oversampling from Hong Kong from 1997 to 1998.

Potential glycosylation sites have always been considered as important discriminative positions for binding of HA and sialic acid, especially potential glycosylation at position N172. Based on this, we perhaps ought to have identified this potential glycosylation site as a feature, especially because it is close to the receptor-binding site. However, this analysis did not identify position 172, and when checked, we found that there was no discriminatory relationship for this position between human and nonhuman viruses. This is supported by previous publications that have shown that the presence or absence of the glycosylation at position 172 is not important for H5 to cross the species barrier [1], [6].

Using the approach described here, we have identified four discriminative features that may decide whether a red blood cell will absorb H5 or not. However, the accuracy rate is not much higher than 70% after classification by the decision tree. The main reason for this is because there are not enough H5 sequences available. Although our approach tried to balance the uneven sequence number between human and nonhuman isolates in order to avoid deflecting the decision tree, we still were unable to choose enough discriminative features to raise the accuracy rate. If the sample number is increased, we believe that it will be possible to identify more discriminative features, and thereby, increase the accuracy rate.

APPENDIX

```
Human/4(HA)/H5N1/Thailand/2004 (accession number : AAS65615) :
   ICQMEKIVLLFAIVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQD   50
   ILEKTHNGKLCDLDGVKPLILRDCSVAGWLLGNPMCDEFINVPEWSYIVE   100
   KANPVNDLCYPGDFNDYEELKHLLSRINHFEKIQIIPKSSWSSHEASLGV   150
   SSACPYQRKSSFFRNVVWLIKKNSTYPTIKRSYNNTNQEDLLVLWGIHHP   200
   NDAAEQTKLYQNPTTYISVGTSTLNQRLVPRIATRSKVNGQSGRMEFFWT   250
   ILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMKSELEYGNCNTKCQTP   300
   MGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNSPQRERRRKKRG   350
   LFGAIAGFIEGGWQGMVDGWYGYHHSNEQGSGYAADKESTQKAIDGVTNK   400
   VNSIIDKMNTQFEAVGREFNNLERRIENLNKKMEDGFLDVWTYNAELLVL   450
   MENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKCDNECMES   500
   VRNGTYDYPQYSEEARLKREEISGVKLESIGIYQILSIYSTVASSLALAI   550
   MVAGLSLWMCSNGSLQCRICI-------                         571

/Human/4(HA)/H5N1/ Hong Kong/1997 (accession number : AAC32098) :
   ---MEKIVLLLATVSLVKSDQICIGYHANNSTEQVDTIMEKNVTVTHAQD   47
   ILERTHNGKLCDLNGVKPLILRDCSVAGWXLGNPMCDEFLNVPEWSYIVE   97
   KTSPANDLCYPGHFNDYEELKHLLSRINHFEKIQIIPKSSWSNHDASSGV   147
   SSACPYLGRSSFFRNVVWLIKKNSAYPTIKRSYNNTNQEDLLVLWGIHHP   197
   NDAAEQIKLYQNPTSYISVGTSTLNQRLVPEIATRPKVNGQSGRMEFFWT   247
   ILKPNDAINFESNGNFIAPEYAYKIVKKGDSTIMKSELEYGNCNTKCQTP   297
   MGAINSSMPFHNIHPLTIGECPKYVKSNRLVLATGLRNTPQRERRRKKRG   347
   LFGAIAGFIEGGWQGMVDGWYGYHHSNEQGSGYAADKESTQKAIDGVTNK   397
   VNSIINKMNTQFEAVGREFNNLERRIENLNKKMEDGFLDVWTYNAELLVL   447
   MENERTLDFHDSNVKNLYDKVRLQLRDNAKELGNGCFEFYHKCDNECMES   497
   VKNGTYDYPQYSEEARLNREEISGVKLESMGTYQILSIYSTVASSLALAI   547
   MVAGLSLWMCSNGSLQCRICIYFCEFRL                         575
```

REFERENCES

[1] E. Hoffmann, A. S. Lipatov, R. J. Webby, E. A. Govorkova, and R. G. Webster, "Role of specific hemagglutinin amino acids in the immunogenicity and protection of H5N1 influenza virus vaccines," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 12915–12920, 2005.

[2] "Evolution of H5N1 avian influenza viruses in Asia," *Emerg. Infect. Dis.*, vol. 11, pp. 1515–1521, 2005.

[3] K. S. Li, Y. Guan, J. Wang, G. J. Smith, K. M. Xu, L. Duan, A. P. Rahardjo, P. Puthavathana, C. Buranathai, T. D. Nguyen, A. T. Estoepangestie, A. Chaisingh, P. Auewarakul, H. T. Long, N. T. Hanh, R. J. Webby, L. L. Poon, H. Chen, K. F. Shortridge, K. Y. Yuen, R. G. Webster, and J. S. Peiris, "Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia," *Nature*, vol. 430, pp. 209–213, 2004.

[4] J. B. Plotkin, J. Dushoff, and S. A. Levin, "Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus," *Proc. Natl. Acad. Sci. USA*, vol. 99, pp. 6263–6268, 2002.

[5] E. C. Claas, A. D. Osterhaus, R. van Beek, J. C. De Jong, G. F. Rimmelzwaan, D. A. Senne, S. Krauss, K. F. Shortridge, and R. G. Webster, "Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus," *Lancet*, vol. 351, pp. 472–477, 1998.

[6] D. L. Suarez, M. L. Perdue, N. Cox, T. Rowe, C. Bender, J. Huang, and D. E. Swayne, "Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong," *J. Virol.*, vol. 72, pp. 6678–6688, 1998.

[7] Y. Guan, K. F. Shortridge, S. Krauss, P. H. Li, Y. Kawaoka, and R. G. Webster, "Emergence of avian H1N1 influenza viruses in pigs in China," *J. Virol.*, vol. 70, pp. 8041–8046, 1996.

[8] C. Bender, H. Hall, J. Huang, A. Klimov, N. Cox, A. Hay, V. Gregory, K. Cameron, W. Lim, and K. Subbarao, "Characterization of the surface proteins of influenza A (H5N1) viruses isolated from humans in 1997–1998," *Virology*, vol. 254, pp. 115–123, 1999.

[9] E. K. Ng, P. K. Cheng, A. Y. Ng, T. L. Hoang, and W. W. Lim, "Influenza A H5N1 detection," *Emerg. Infect. Dis.*, vol. 11, pp. 1303–1305, 2005.

[10] K. Iwatsuki-Horimoto, R. Kanazawa, S. Sugii, Y. Kawaoka, and T. Horimoto, "The index influenza A virus subtype H5N1 isolated from a human in 1997 differs in its receptor-binding properties from a virulent avian influenza virus," *J. Gen. Virol.*, vol. 85, pp. 1001–1005, 2004.

[11] D. J. Hulse-Post, K. M. Sturm-Ramirez, J. Humberd, P. Seiler, E. A. Govorkova, S. Krauss, C. Scholtissek, P. Puthavathana, C. Buranathai, T. D. Nguyen, H. T. Long, T. S. Naipospos, H. Chen, T. M. Ellis, Y. Guan, J. S. Peiris, and R. G. Webster, "Role of domestic ducks in the propagation and biological evolution of highly pathogenic H5N1 influenza viruses in Asia," *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 10682–10687, 2005.

[12] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch, "Predicting the evolution of human influenza A," *Science*, vol. 286, pp. 1921–1925, 1999.

[13] J. P. Jenuth, "The NCBI. Publicly available tools and resources on the Web," *Methods Mol. Biol.*, vol. 132, pp. 301–312, 2000.

[14] E. Jung, S. Brunak, and R. Gupta, "Prediction of N-glycosylation sites in human proteins," submitted for publication.

[15] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*. San Mateo, CA: Morgan Kaufmann, 2000.

[16] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res.*, vol. 22, pp. 4673–4680, 1994.

Li-Ching Wu was born in Taipei, Taiwan, in 1973. He received the Ph.D. degree in computer science and information engineering from the National Central University, Jhongli City, Taiwan, in June 2004.

He is currently with the Institute of Systems Biology and Bioinformatics, National Central University. His current research interests include bioinformatics, database systems, and data mining.

Jorng-Tzong Horng (M'02) was born in Nantou, Taiwan, on April 10, 1960. He received the Ph.D. degree in computer science and information engineering from the National Taiwan University, Taipei, Taiwan, in April 1993.

Since 1993, he has been with the Department of Computer Science and Information Engineering, Institute of Systems Biology and Bioinformatics, National Central University, Jhongli City, Taiwan, where he became a Professor in 2002. His current research interests include database systems, data mining, genetic algorithms, and bioinformatics.

Hsien-Da Huang was born in Taoyuan, Taiwan, in 1975. He received the Ph.D. degree in computer science and information engineering from the National Central University, Jhongli City, Taiwan, in June 2003.

Since 2003, he has been with the Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu, Taiwan. His current research interests include bioinformatics, database systems, and data mining.

Wei-Long Chen received the M.S. degree in computer science and information engineering from the National Central University, Jhongli City, Taiwan, in 2006.

He is now in his duty of military service.