Virology 449 (2014) 328–338

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Accurate classification and hemagglutinin amino acid signatures for influenza A virus host-origin association and subtyping

Mahmoud ElHefnawi [a,*], Fayroz F. Sherif [b,c]

[a] Informatics and Systems Department and Biomedical Informatics and Chemoinformatics group, Division of Engineering Research and Centre of Excellence for Advanced Sciences, National Research Centre, Tahrir Street, 12311 Cairo, Egypt
[b] Biomedical Engineering Department, Cairo University, 12613 Giza, Egypt
[c] Bioelectronics Department, Modern University for Technology and Information, Katameya, Egypt

ARTICLE INFO

Article history:
Received 24 July 2013
Returned to author for revisions
3 September 2013
Accepted 6 November 2013
Available online 22 December 2013

Keywords:
Influenza
HA protein
Virus signature
Hemagglutinin subtyping

ABSTRACT

Host-origin classification and signatures of influenza A viruses were investigated based on the HA protein for tracking of the HA host of origin. Hidden Markov models (HMMs), decision trees and associative classification for each influenza A virus subtype and its major hosts (human, avian, swine) were generated. Features of the HA protein signatures that were host-and subtype-specific were sought. Host-associated signatures that occurred in different subtypes of the virus were identified. Evaluation of the classification models based on ROC curves and support and confidence ratings for the amino acid class-association rules was performed. Host classification based on the HA subtype achieved accuracies between 91.2% and 100% using decision trees after feature selection. Host-specific class association rules for avian-host origins gave better support and confidence ratings, followed by human and finally swine origin. This finding indicated the lower specificity of the swine host, perhaps pointing to its ability to mix different strains.

Introduction

Influenza A viruses

Influenza is one of the most important emerging and re-emerging infectious diseases, causing high morbidity and mortality (Allen et al., 2009). Global outbreaks of human influenza arise from influenza A viruses with novel hemagglutinin (HA) proteins, to which humans have no immunity (Finkelstein et al., 2007). The influenza A viral genome is segmented into eight parts which allows the exchange of entire genes between the different viral strains producing new viruses (Horimoto and Kawaoka, 2005; Triki, 1997). The HA protein is responsible for the binding of virions to host cell receptors and for fusion between the virion envelope and the host cell (Wiley and Skehel, 1987). The role of NA is to free virus particles from host cell receptors, to permit progeny virions to escape from the cell in which they arose, and so facilitate the spread of the virus (Chander et al., 2010). There are at least 16 different HA and nine different NA influenza A subtypes (Zhang et al., 2009) classified according to the immunological nature of the strains.

Influenza subtyping and pattern discovery

Rapid virus subtype and evolutionary host of origin identification is critical for accurate diagnosis of human infections, effective response to epidemic outbreaks, and global-scale surveillance of highly pathogenic subtypes (Garten et al., 2009). The hemagglutination inhibition assay is a classical subtyping method but it requires extensive laboratory support for reagent libraries (Pedersen, 2008). Another way of subtyping the HA genes is by reverse transcriptase PCR, or real-time PCR, which is highly specific (Starick et al., 2000). Sequencing methods can also be used for viral characterization by BLAST searches against known viral sequences (Altschul et al., 1997); however, the BLAST results cannot reveal host-origin, or host-related signatures which are important mutations that may be related to the structure and function of HA proteins. Further, the BLAST scores would not reliably reveal the host-origin because a few mutations could separate two hosts of the same subtype.

Discriminative pattern recognition and identification of conserved regions for the influenza A virus proteins are important for capturing the signatures associated with seasonal changes (ElHefnawi et al., 2011a,b; Gendoo et al., 2008) because these changes can provide functional insights into the roles of the influenza proteins and the HA segment of the viral genome. In addition, antigenic drifts and antigenic shifts have been related to pandemics occurrences; especially to the high-infectivity 2009 H1N1 pandemic (manuscript in press).

* Corresponding author.
E-mail addresses: mahef@aucegypt.edu (M. ElHefnawi), FFS@K-space.org, fayroz_farouk@yahoo.com (F.F. Sherif).

0042-6822/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.virol.2013.11.010

Influenza host-origins and the HA protein

To elucidate the mechanism by which pandemic influenza virus strains are generated, we must first understand the host range restrictions at a molecular level, and the mechanisms and processes behind such restrictions. All 16 subtypes of HA and nine subtypes of NA are found in the avian influenza virus (Munch et al., 2001). Four viral subtypes, H1, H2, H3 and H5 (Scholtissek et al., 1978) are known human influenza viral strains, although, recently, the H7 subtype was found in the Netherlands. Swine-origin influenza virus is limited to subtypes H1N1, H1N2, H3N1, and H3N2. Influenza virus infection is initiated by interactions between the viral HA and sialic acid (SA)-containing carbohydrates on the surface of the target cells. Avian influenza viruses are not readily introduced into humans (Beare and Webster, 1991), possibly because humans do not possess the α(2,3)-sialyllactose (NeuAc-2,3Gal) receptors required for the attachment of the viruses to epithelial cells. However, individual viral genes can be transmitted between human and avian species, as demonstrated by the avian human reassortant viruses that caused the 1957 and 1968 influenza pandemics (Scholtissek et al., 1978; Kawaoka et al., 1989). This finding suggested that an intermediate host might be needed for the genetic reassortment of human and avian viruses. Swine are considered likely candidates for this role because they can be infected by either avian or human viruses (Kida et al., 1994; Schultz et al., 1991), and because they possess both NeuAc-α2,3Gal and NeuAc-α2,6Gal receptors (Kida et al., 1994; Rota et al., 1989; Scholtissek et al., 1983).

HA protein analysis, data mining and influenza host of origin

The amino acid residues of HA that make up the receptor-binding site (RBS) are highly conserved among the HAs of different subtypes of the avian influenza virus; however the amino acids in the RBS of the human influenza viruses display distinct variability. In particular, the residues at positions 138, 190, 194, 225, 226, and 228 are highly conserved in the avian RBSs, whereas in the human RBSs there are substitutions at these positions (Matrosovich et al., 1997). In the H2 and H3 influenza virus strains, residues at positions 226 and 228 in the HA sequence correlate with the preferential recognition of the SA-Gal (referred to above as NeuAc-Gal) linkage by HA and the host species from which the virus was isolated. HAs with Leu at position 226 (Leu-226) and Ser-228 (human viruses) preferentially recognize SA-α2,6Gal, whereas those with Gln-226 and Gly-228 (avian and equine viruses) recognize SA-α2,3Gal (Connor et al., 1994). Highly pathogenic avian influenza H5N1 virus strains can transmit directly from avian species to humans and a cause severe form of the disease. The change of one amino acid in the RBS of the H5 HA protein could be sufficient to change the receptor-binding specificity of H5N1 viruses, easily overcoming barriers between interspecies transmission (Wong and Yuen, 2006). This process will be elaborated upon in the Discussion section.

Previously, we used hidden Markov models for subtyping of influenza a virus using the Hemagglutanine protein with 100% accuracy, and for host of origin classification with accuracies ranging between 50–90% (Fayroz et al., 2012). An integrated approach, using both decision trees (DTs) and profile hidden Markov models (HMMs) for the subtype prediction of human influenza A virus was presented by Attaluri et al. (2009a) and was reported to have achieved a subtype prediction accuracy of 88% for the human subtypes. In another study, these workers applied two machine learning techniques (DTs and support vector machines) to identify and discriminate the origin of the pandemic (H1N1) 2009 viral strains with 95% accuracy and concluded that the H1N1 strain was of swine origin (Attaluri et al., 2009b). In their most recent study, Attaluri et al. (2010) applied a feed-forward backpropagation neural network to predict important influenza virus antigenic types and hosts and found that the highest accuracy was achieved when using HAs and NAs for human host classification.

Large-scale sequence analyses have revealed ‘signature’ amino acids at specific positions in the viral proteins that distinguish human influenza viruses from avian influenza viruses (Finkelstein et al., 2007; Chen et al., 2006). These host lineage-specific amino acids were found mainly in the components of the viral RNA polymerase complex, such as the PB2 and PA polymerases and the nucleocapsid protein, that is essential for viral genome replication (Deng et al., 2006; Klumpp et al., 1997). It is likely that these amino acids contribute to the host-range restriction of influenza viruses (Gabriel et al., 2005; Scholtissek et al., 1985); although, with the exception of the amino acids at positions 627 and 701 of PB2 whose importance in virulence has been demonstrated in a rodent model (Shinya et al., 2004; Steel et al., 2009), their biological significance remains to be established. In another study of the pandemic (H1N1) 2009 virus, the human–swine signatures and amino acid sequences at the host species-specific positions of the proteins were analyzed to elucidate the adaptive mutation of the strain in these host species (Chen and Shih, 2009). Signatures that distinguish swine viruses from human viruses were also present.

Aim of the present work

The aim of the present study was to identify molecular signatures and establish accurate host of origin classifiers for the influenza HA protein. A profile HMM-subtype classification for the influenza A virus was performed. Then, HMM and DT classification models were used for host-origin classification of a particular subtype. Host-origin classification independent of subtype was conducted next. Finally, host-origin signatures and classification rules were inferred for the three major hosts (human, avian, and swine). Here, we generalized and extended on previous studies that analyzed one subtype or host, to include all influenza A viral subtypes and major host origins using better classification models. We also identified host-specific genomic signatures in the HA proteins for human vs. swine vs. avian origin influenza viruses. Amino acid residues that were specific to either human, swine or avian influenza viruses were selected as potential host-associated signatures. Class association rules were generated for the sites with statistically significant variations between different hosts in both the comparative sequence logo and the viral epidemiology signature pattern analysis (VESPA) as detailed in the Methods section and flow chart (Fig. 1). We subsequently validated the robustness of the candidate signatures against human, avian and

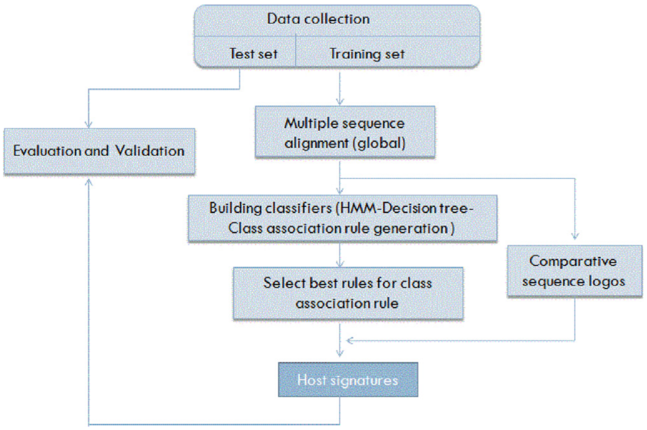

Data collection
Test set
Training set
Evaluation and Validation
Multiple sequence alignment (global)
Building classifiers (HMM-Decision tree-Class association rule generation)
Comparative sequence logos
Select best rules for class association rule
Host signatures

Fig. 1. Protocol used to extract host-origin classification signatures for the influenza A virus subtypes. The classification methods incorporated amino acid signatures into class-association rules for the HA protein for human vs. avian vs. swine influenza A viruses.

swine sequences downloaded from the National Center for Biotechnology Information (NCBI) Influenza Virus Resource database. Using DTs, we found that these signature classifiers achieved host-origin prediction accuracies based on HA subtypes between 91.2% and 100%. Thus, analysis of a near-complete collection of species-specific influenza A viral sequences comprising the long-evolving avian, and recent ancestral swine and human viruses, as well as the pandemic (H1N1) 2009 viruses was performed, and host-specific signatures and class-association rules were generated that would help in tracking and understanding of the influenza virus.

Results

Subtyping and hidden Markov models and decision trees host of origin classification

Multiple sequence alignments were carried out separately for 16 HA subtypes and nine NA subtypes using the ClustalX program. Then, using the HMMER suite for each group, profile HMM models were built and calibrated to produce a database for each group as previously performed (Sherif et al., 2012). The accuracies of the classification results using the HMMs was 100% for all the HA and NA subtypes. The same method was then used for subtype classification of the host of origin.

For each host-specific group, sequences belonging to each HA subtype were aligned using the ClustalW program. Each of the 12 groups of data (H1 – human, H1 – avian, H1 – swine, H2 – human, H2 – avian, and so on) were aligned and analyzed. Host classification was done by applying the two comparative techniques (profile HMM and DTs) to identify the origin of the viral strains. The pre-identified HA subtype has scored with the corresponding 'HA-host' HMM models for better matching. The host classification using the profile HMMs had accuracies between 53% and 100% (Table 2).

Host-specific signatures of the human, swine, and avian viruses were extracted using VESPA with a minimum threshold of 90%. A total of 23, 22, 12, 7 and 8 most informative amino acid positions were used to collectively identify the different H1, H2, H3, H5 and H9 hosts respectively. Comparative sequence logos were used to graphically represent the most informative positions for each of the HA subtypes for the different hosts (Fig. 2 and Additional file 1 in Appendix A). To identify more precisely the host of origin of the viruses, these important positions were used to generate DTs for the H1, H2, H3, H5, and H9 HAs (Fig. 3 and Additional file 2 in Appendix A). The DT models had accuracies between 91.2% and 100% (Table 2), outperforming the profile HMM models. The performance of the DTs was comparable to the performance of DTs reported in previous studies, as elaborated upon in the Discussion section.

General host of origin signatures identification (non-subtype-specific)

A multiple sequence alignment of all 1500 HA protein sequences from the three hosts was performed and then the alignment was separated into three sets of alignments (human, avian and swine) for comparison. Two groups of sequences (one from each of two sets) were compared, generating six different comparisons between the hosts: human vs. avian, human vs. swine, avian vs. swine, and their permutations (Fig. 4 and Additional file 4 in Appendix A). Positional variations in the HA sequences between different hosts were compared using VESPA and comparative sequence logos (Fig. 4) (see Methods section for details). The most informative positions are those that were relatively variable in one group compared with the other and

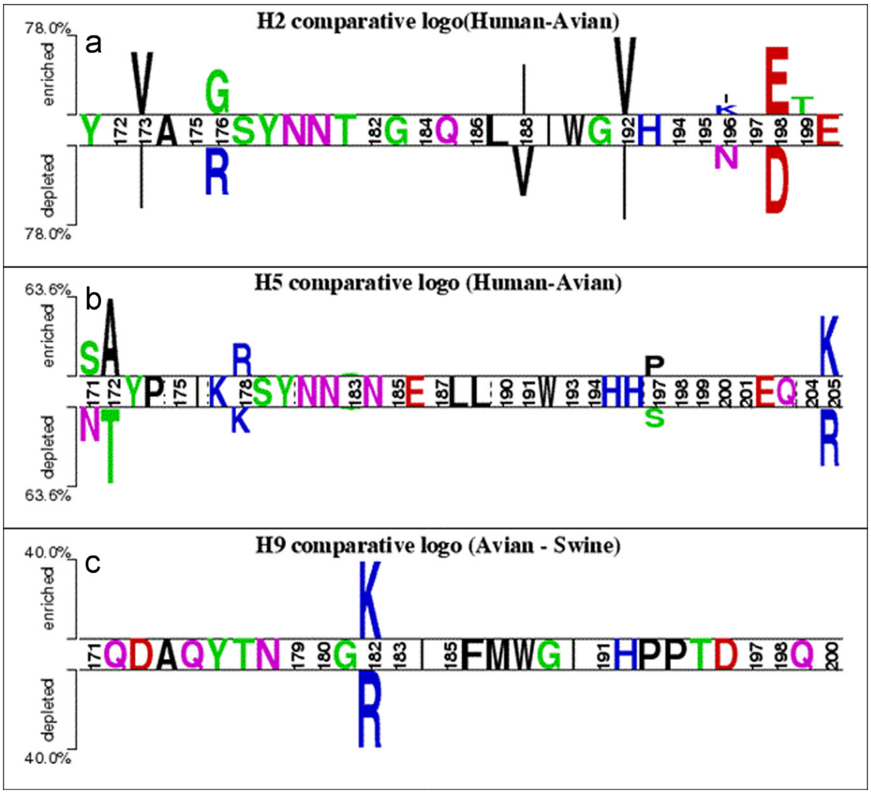


Fig. 2. Representative comparative sequence logos for subtype-specific HA signatures in different hosts. Comparative sequence logos representing the most informative positions for the H2, H5 and H9 HA proteins from different hosts are shown. (A) H2-avian sequences were set as the negative sample and H2-human sequences were set as the positive sample. (B) H5-avian sequences were set as the negative sample and H2-human sequences were set as the positive sample. (C) H9-avian sequences were set as the negative sample and H9-swine sequences were set as the positive sample. The letters in the middle bar represent conserved positions. The empty positions represent variations within each group but no significant variations between the two groups.

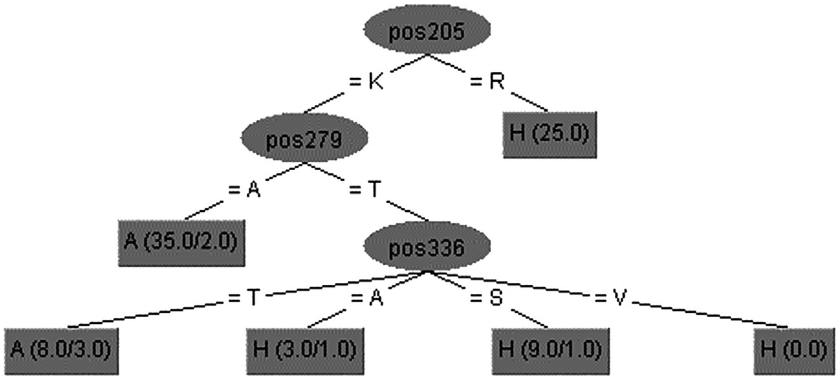


Fig. 3. DT classifier to classify H5 influenza A virus host of origin as human or avian. The DT for the H5 HA that was used to predict the H5 host of origin as human (H) or avian (A). The numbers in brackets indicate.

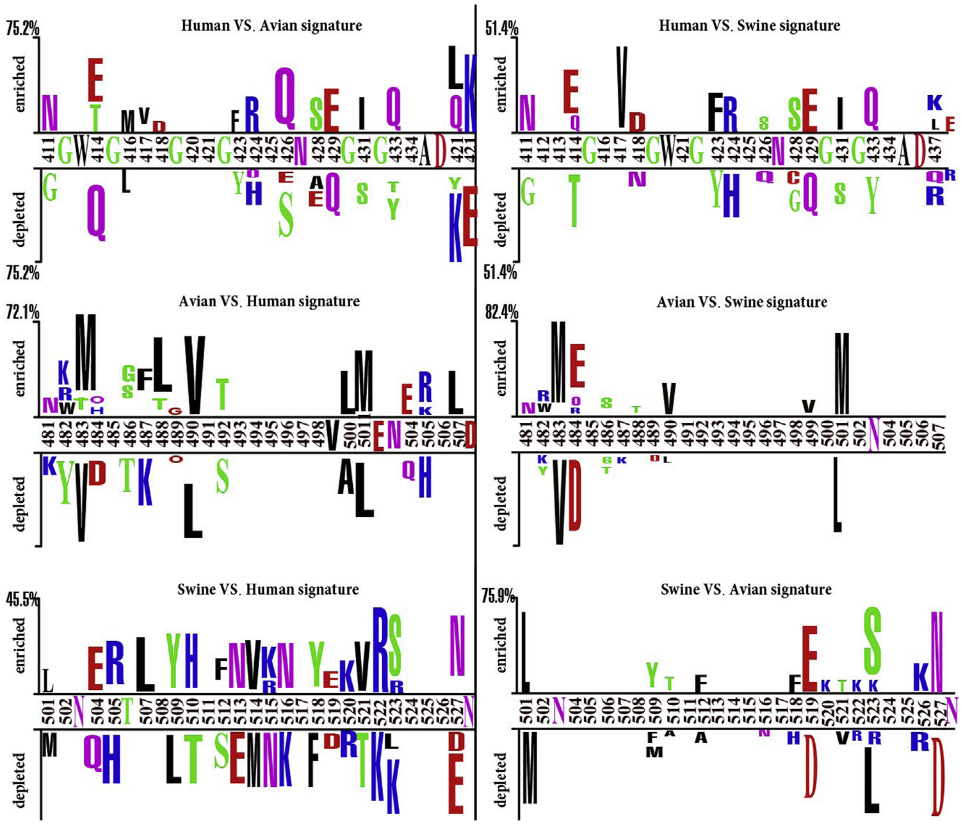


Fig. 4. Comparative sequence logos for HA host signatures across different subtypes. Logos for human vs. avian, human vs. swine, avian vs. swine and vice versa are shown. The letters in the middle bar represent conserved positions. The empty positions represent variations within each group but no significant variations between the two groups. The variable positions in the different hosts that were also found by VESPA are shown [see the tables in Additional file 4 in Appendix A for details].

their relative frequencies [see the tables in Additional file 4 in Appendix A for details]. Informative class association rules with a certain threshold of support and confidence ratings were generated using the VESPA and comparative sequence logos results (Tables 3 and 4). The complete set of rules with their support and confidence ratings is available in additional tables [see Additional file 4 in Appendix A]. The most accurate sets of class association rules extracted from the most informative positions are shown in Table 3 and the top-ranked rules are listed in Table 4. The amino acid positions shown in Table 3 refer to the H1N1 reference sequence [GenBank:NP_040980].

A total of 9, 31, 11, 6, 22, and 31 most informative amino acid positions among the 630 aligned residues in the HAs, revealed significant differences between the avian vs. human, human vs. avian, human vs. swine, swine vs. human, avian vs. swine, and swine vs. avian HAs respectively (Table 3 and the tables in Additional file 4 in Appendix A). Two signature residues at positions 291 (K in human, N in avian, D in swine) and 408 (H in human, E in avian, and T in swine) of the HAs (highlighted in yellow in Table 3) exhibited dominant changes between the three hosts with strong support in the avian and moderate support in the human and swine viruses. Positions 244 (M–I) and 312 (I–V) (356) (highlighted in green in Table 3) were both informative for finding signatures for human vs. avian and human vs. swine. These two positions could separate the HAs into two classes (human and avian/swine) with support ratings of 82.8% and 82.4% respectively.

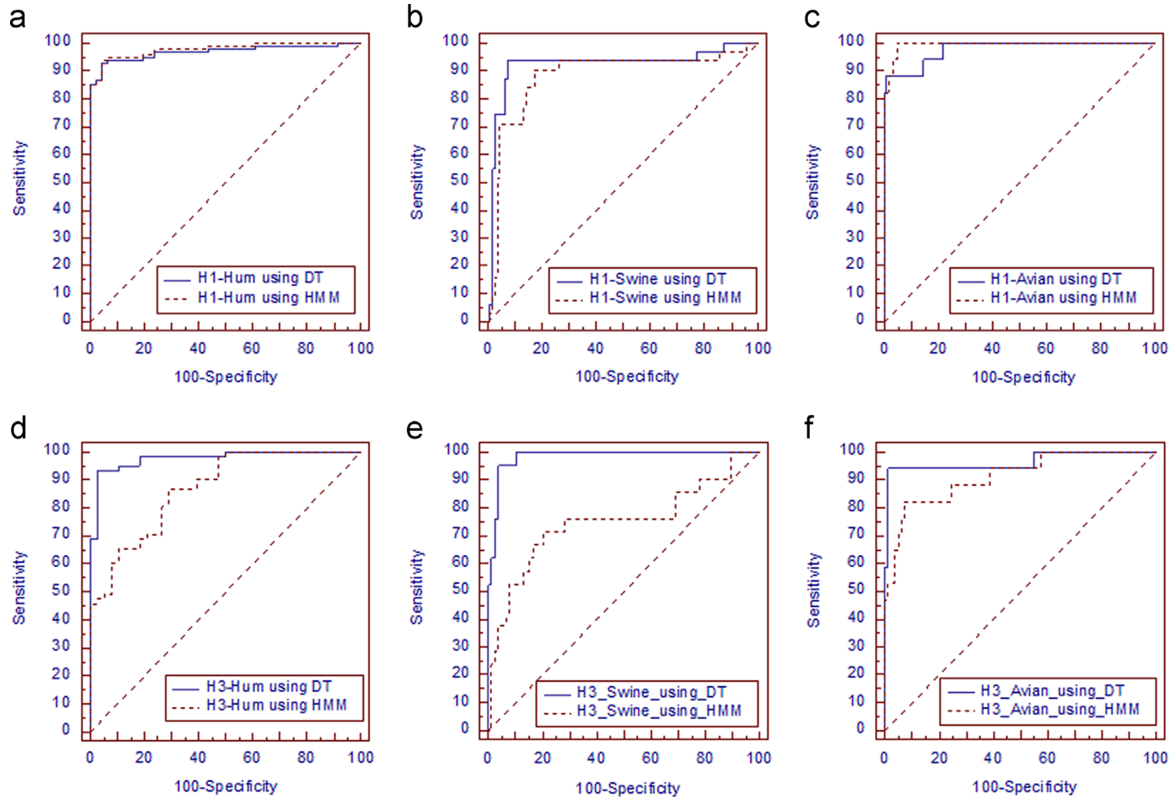
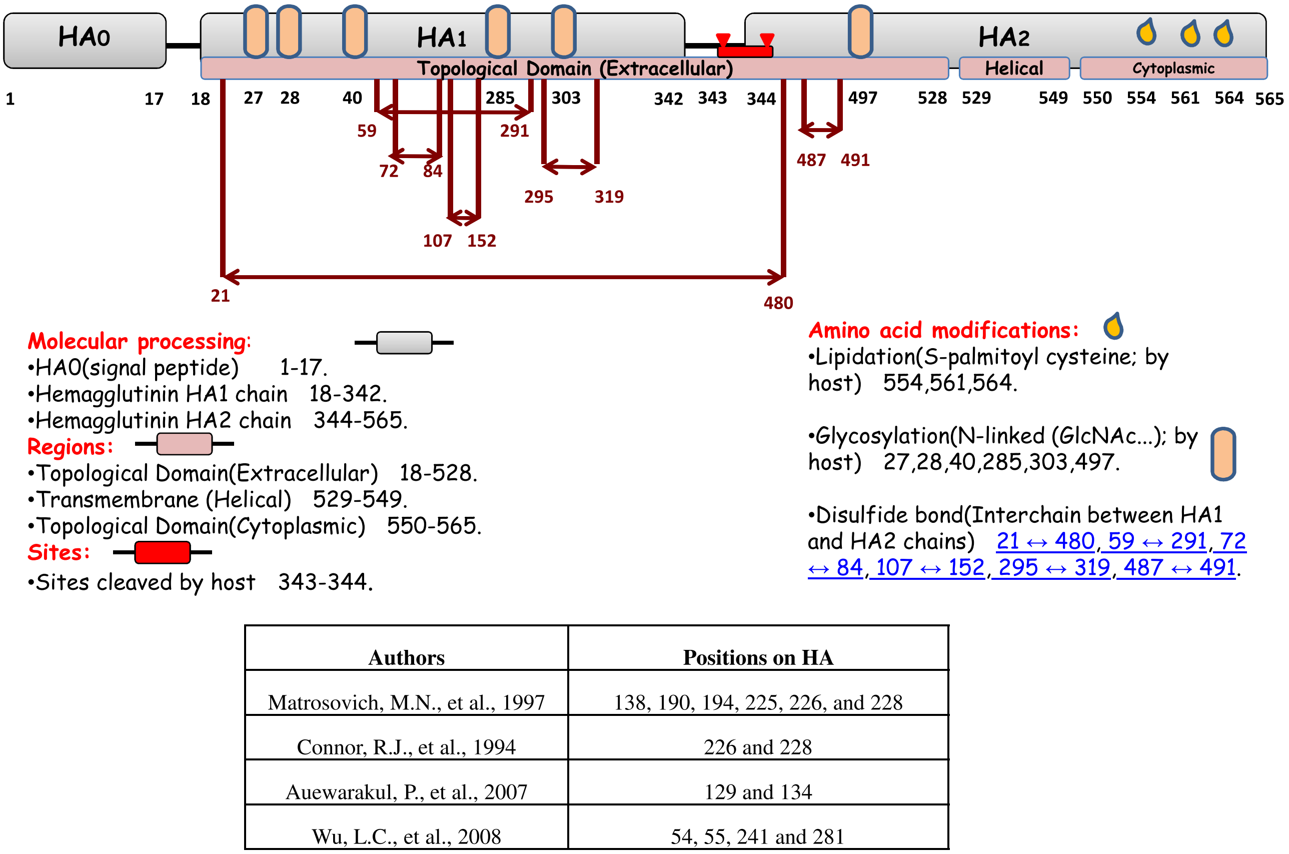


Fig. 5. Comparative ROC curves between the HMMs and DTs for host identification of different HA subtypes. The ROC curves for (A) H1 – human, (B) H1 – swine, (C) H1 – avian, (D) H3 – human, (E) H3 – swine, and (F) H3 – avian models are shown.



Molecular processing:
- •HA0(signal peptide) 1-17.
- •Hemagglutinin HA1 chain 18-342.
- •Hemagglutinin HA2 chain 344-565.

Regions:
- •Topological Domain(Extracellular) 18-528.
- •Transmembrane (Helical) 529-549.
- •Topological Domain(Cytoplasmic) 550-565.

Sites:
- •Sites cleaved by host 343-344.

Amino acid modifications:
- •Lipidation(S-palmitoyl cysteine; by host) 554,561,564.
- •Glycosylation(N-linked (GlcNAc...); by host) 27,28,40,285,303,497.
- •Disulfide bond(Interchain between HA1 and HA2 chains) 21 ↔ 480, 59 ↔ 291, 72 ↔ 84, 107 ↔ 152, 295 ↔ 319, 487 ↔ 491.

Authors	Positions on HA
Matrosovich, M.N., et al., 1997	138, 190, 194, 225, 226, and 228
Connor, R.J., et al., 1994	226 and 228
Auewarakul, P., et al., 2007	129 and 134
Wu, L.C., et al., 2008	54, 55, 241 and 281

Fig. 6. Features of the HA protein mapped to the H1N1 reference sequence [GenBank:NP_040980]. The positions of the HA subdomains, post-modification sites, transmembrane region, and receptor-binding sites are indicated. Specific strain and host-associated positions are shown in the accompanying table.

Positions 266 (I–V)), 267 (A–V), and 526 (G–R) (highlighted in blue in Table 3) were informative for finding signatures for both swine vs. human and swine vs. avian. These positions could separate the HAs into two classes (swine and avian/human) but their support rating was about 55.5%. This poor result is plausible because the swine host is a known reservoir for both the avian and human strains. Notably, except for the few positions mentioned above, none of the positions listed in Table 3 could simultaneously differentiate between the three host classes.

The avian signature showed the largest number of informative positions (the 21 non-highlighted positions in Table 3). These positions were informative for finding signatures for both human vs. avian and swine vs. avian and separated the hosts into two classes (human/swine and avian) with support ratings between 59% and 99.3% (Table 3).

Some of the positional markers that produced the most informative rules with the highest support and confidence ratings (Table 4) could not distinguish between the functional classes by themselves; however, when combined with other markers they improved the class predictions. These findings demonstrate the importance of the informative residues for receptor specificity and for host range restriction of the virus.

The support and confidence ratings are measures of a rule's interestingness and reflect the usefulness and certainty, respectively, of the postulated rules. For example, we found a support rating of 90% for an association rule at position 408 (H in human, E in avian, and T in swine), indicating that in 90% of the sequences analyzed “E” and “avian” occurred together. A confidence rating of 79.8% indicates that 79.8% of the sequences that contain the residue “E” at position 408 were found in avian. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. These thresholds can be set by users or domain experts.

Model evaluation: ROC curves

The classification models were evaluated using ROC curves which chart the number of true positives vs. the number of false positives. The ROC curves were generated using the MedCalc program. True positives are homologous pairs and false positives are non-homologous pairs with scores above a certain threshold. By varying the threshold score, the curve of true positives vs. false positives can be traced (Nomura, 1979). We used comparative ROC curves to test the statistical significance of the difference between the areas under two or more ROC curves. The ROC curves for comparisons between HMMs and DTs for host identification of the different HA subtypes are shown in Fig. 5 and Additional file 3 in Appendix A. The curves indicate the superiority of the DT models.

A diagram of the HA reference protein from H1N1 indicating some of its important features such as subdomains, motifs, receptor binding sites, and post-translational modifications, is shown in Fig. 6. Mapping the markers to the annotated HA reference sequence may reveal their host restriction roles.

Discussion

Recent large-scale sequence analyses revealed ‘signature’ amino acids at specific positions in viral proteins that distinguish human influenza viruses from avian or swine viruses. Therefore, it is likely that, because of immune pressure and the receptor specificity of the HA receptor binding site, there are markers in the HA glycoprotein on the surface of the host cells.

Subtype classification

Our results confirm that protein profile HMMs can be used successfully to subtype influenza A strains hosted in all three species. The HA and NA subtypes were identified with 100% accuracy and the 16 HA and nine NA models all had a sensitivity of 100% and specificity of 100% as previously performed (Sherif et al., 2012). Our results achieved a higher accuracy than the accuracy reported by Attaluri et al. (2010) in a previous study. In their study, the accuracy for subtype classification was over 91% when the frequencies of k-mer nucleotide strings were used as input to a neural network and a higher value of k was reported to achieve relatively better classification results. In the present study, we used protein sequences rather than nucleotide sequences because protein sequences tend to be more discriminative as reflected in the higher accuracy levels that we achieved.

Subtype-specific host-origin classification

Host classification of any viral sequence depends on the HA subtype. Some of the HA subtypes that can infect more than one species vary greatly between the human, swine, and avian viruses, while others varied so little that it was difficult to identify their host of origin.

Here, because of its particular importance, we focused on the HA protein and compared the results of the HMM and DT analyses to classify the influenza host. For the HMM models, the detailed results are also discussed elsewhere (Altschul et al., 1997; Fayroz et al., 2012).

Interestingly, we found that using DTs improved the accuracy of the influenza host classification by identifying the most informative positions that differentiate different hosts within the same HA subtype. For the major HA subtypes, DTs achieved a higher accuracy in host classification than the HMMs for the same HA subtype. For example, the DT for the H5 HA that was built to identify the H5 host of origin (human or avian), predicted the H5-human and H5-avian models with accuracies of 91.3% and 91.2% respectively, much higher than the accuracies for the equivalent HMM models. The DT analysis focuses on some specific discriminative positions, making it is possible to identify accurately the host of origin. In contrast, profile HMMs often model complete protein sequences and then search the query sequence for global matching using this model. Thus, despite the genetic similarities that exist in the human, swine, and avian viruses for the same HA subtype, we were able to successfully identify specific signatures using our combined approach.

Similarly, Attaluri et al. (2009a, 2010) found that the accuracy of virus classification varies from host to host and from gene segment to gene segment and the highest host classification accuracies were achieved for the HA and NA genes. In the present study and in the Attaluri et al. studies, compared with avian and swine hosts, the human host was predicted with the highest accuracy, no matter which method was used.

General host-origin specific signatures

Here, we proposed a computational approach that is capable of indicating species-associated signatures in human, avian, and swine influenza viral genomes. Because of the important functional role of HA in cell-receptor attachment, entry, and infectivity, our focus in this study was specifically on the persistent host-origin conserved markers and host markers that were found only in the surface HA glycoproteins. Variations in the HA protein are caused by immune pressures, and host restrictions occur because of the receptor specificity of the HA receptor binding sites. Nevertheless, our results could still be applied to study evolutionary processes in different hosts and to investigate host adaptation.

These class association rules which are position- and amino acid-specific, proved to be more appropriate for host identification than the profile HMMs that we reported previously (Fayroz et al., 2012). Some of the rules gave support and confidence ratings that were high enough either to separate between the three hosts or to separate one of the hosts from the other two (Table 3). The amino acids at positions 185 (K–T), 238 (R–N), 266 (I–V), 383 (K–E), 408 (T–E) and 547 (L–V) gave the best support and confidence ratings for influenza HA protein host discrimination. Thus, the class association rules that were extracted from the VESPA results and confirmed by comparative sequence logos can be used to increase the sensitivity and specificity of genetic biomarker discovery in general (ElHefnawi et al., 2010). Comparative sequence logos confirm our signature analysis for every pair of host groups because almost all the positional markers coincided with the comparative sequence logo and signature pattern analysis [see the tables in Additional file 4 in Appendix A]. However, the comparative logos detected some additional signature positions, such as positions 433 I–L (488) and 435 I–V (490), which were statistically significant in the avian vs. human signature.

Previous studies have also defined host specificity markers. For example, Allen et al. (2009) predicted positions in the genome associated with human host specificity. However, the host markers that these workers identified in the surface glycoproteins HA and NA and in the polymerase protein PB1, as well as the alternate transcripts NS2, M2, and PB1-F2, were poor-quality host discriminators. Chen et al. (2006) looked for human markers beyond the pandemic conserved regions. However, their approach of identifying species-associated signatures by entropy are less useful for the HA and NA genes because the genetic diversity that exists in these two gene segments in human or avian viruses can boost their respective entropy to more negative values, making it difficult to find residues that were conserved sufficiently to identify such signatures (Chen et al., 2006). Similarly, using genetic distance or phylogenic analysis for host-origin discrimination may not be applicable because the different hosts would not cluster appropriately. Some phylogenetic trees that clustered HA subtypes but not hosts of origins are available in an additional file [see Additional file 6 in Appendix A].

For the inaccurately classified swine host HA sequences H1N1, H1N2, H3N1, and H3N2, the classification errors appeared to be due to recent reassortment events, suggesting that some influenza genomes are a mix of both human and avian strains (Sherif et al., 2011). Matrosovich et al. (1997) reported six amino positions (138, 190, 194, 225, 226, and 228) that distinguish human and avian influenza viral sequences. In a virus isolated from a fatal human influenza case, Auewarakul et al. (2007) showed that substitutions at positions L129V and A134V in the HA protein could change the receptor-binding preference of the HA of the H5N1 virus from SA-2,3Gal to both SA-2,3Gal and SA-2,6Gal. Likewise, Wu et al. (2008) identified four discriminative amino acid positions (54, 55, 241 and 281) in the HA protein sequences within H5N1 using a DT. None of these positions were identified in the non-specific host-origin subtype signatures mainly because, for general host-origin signature identification, we pooled all the HA subtypes into one “host” class which would have substantially altered the reported set of persistent markers for a specific subtype.

Our study could also have a number of limitations. In addition to the data limitations, accurate HA sequence alignments are difficult to generate because of the high sequence variability in the HA proteins, and this is despite the care taken in manual editing. Thus, false-negative errors may occur as the result of alignment errors.

Conclusion

Accurate detection of the viral origin of influenza can significantly improve influenza surveillance and vaccine development. Here, subtyping and host identification of influenza A virus was performed based on profile HMMs for the HA and NA subtypes and DTs for the major host of origin. Critical amino acid positions and identities inside the HA proteins were identified to act as host-specific signatures. HA host-origin comparisons revealed host-specific sites and amino acids that could help in modeling the evolution of the influenza A HA protein through different hosts and in understanding its specificity. Informative class association rules with a certain minimum threshold of support and confidence were generated to improve host-origin determination.

Hence, the power of extracting conserved and discriminative positions from integrating the multiple sequence alignments, and DTs approaches in classifying influenza A viral strains and their host of origin was demonstrated. We found that the subtyping of the HA and NA proteins using profile HMMs was an accurate and easy to apply method. When the DT and HMM approaches for host-origin classification was compared, we found that the DT method was superior.

Finally, to extract the most important motifs, discriminative pattern analysis on a very wide range of complete HA sequences was performed. The host markers that were identified were confirmed and validated using human, avian, and swine test data sets. The host-specific class association rules that we built gave higher support and confidence ratings for avian compared with for human or swine. The protein sequences from the different hosts were numbered based on the H1N1 influenza A virus Puerto Rico strain HA protein [GenBank:NP_040980.1] which was used as the reference sequence. Some of the highest host-origin class association rules were located at amino acid positions 238 (N–R), 383 (K–E) and 266 (I–V) in the avian, human and swine HAs respectively. Also, two HA signatures, one at position 291 (K in human, N in avian, D in swine) and another at position 408 (H in human, E in avian, T in swine), exhibited dominant changes in the three hosts, suggesting that these signatures may be useful as host-specific markers as described in the Results section and in Tables 3 and 4. Thus, the residues at these positions are potential markers for the prediction of influenza host origin.

Methods

The workflow (Fig. 1) that we followed to classify the influenza A viral subtypes and hosts origins and their associated signatures consists of sequence collection and sorting, and multiple sequence alignments. This was followed by the training and testing of profile HMM influenza subtype models. The profile HMMs were then used for host-origin classification. For host-origin classification using DTs, informative site identification and feature selection was performed using comparative sequence logos (for quick identification of positions that are statistically different) and VESPA (for positional enumeration of amino acids variations between two hosts) after global multiple sequence alignments for each HA protein host group were generated. For host identification, comparisons between the HMMs and DTs for the multi-host of origin subtypes H1, H2, H3, H5 and H9 were conducted.

To find host-origin signatures or markers in the HA proteins, we used informative site identification and feature selection for the positional enumeration of amino acids in each host group regardless of HA subtype. Then, class association rules were generated and the best set of rules was selected.

Table 1
List of influenza A virus sequences used in this study. The count of sequences used for each of the HA and NA subtypes is shown. The total number of each subtype, and the number in each group based on the host are listed. For example, 'H1 – human' is the number of H1 viral sequences isolated from human. The number of sequences in the training and test sets used for building and testing the HMMs are indicated.

HA segment	Group	# Of training sequences	# Of test sequences
H1	Total H1	1154	178
	H1 – human	749	100
	H1 – avian	105	10
	H1 – swine	300	68
H2	Total H2	174	43
	H2 – human	50	13
	H2 – avian	124	30
H3	Total H3	913	128
	H3 – human	550	69
	H3 – avian	263	30
	H3 – swine	100	29
H4	Total (avian)	200	64
H5	Total H5	1310	217
	H5 – human	110	33
	H5 – avian	1200	184
H6	Total (avian)	150	40
H7	Total (avian)	200	64
H8	Total (avian	15	4
H9	Total H9	413	44
	H9 – avian	400	42
	H9 – swine	13	2
H10	Total (avian)	40	7
H11	Total (avian)	40	11
H12	Total (avian)	15	4
H13	Total (avian)	25	5
H14	Total (avian)	10	2
H15	Total (avian)	10	2
H16	Total (avian)	12	4
NA segment			
N1	Total N1	1530	146
	N1 – human	600	56
	N1 – avian	830	70
	N1 – swine	100	20
N2	Total N2	1652	264
	N2 – human	761	100
	N2 – avian	700	124
	N2 – swine	191	40
N3	Total N3	102	40
N4	Total N4	40	10
N5	Total N5	65	20
N6	Total N1	261	15
N7	Total N7	100	20
N8	Total N8	300	30
N9	Total N9	80	20

Sequence collection and data analysis

Viral protein sequences were downloaded from the NCBI Influenza Virus Resource (Bao et al., 2008), including sequences from laboratory-adapted viruses and pandemic (H1N1) 2009 sequences from within the human host. Of the downloaded sequences, only non-redundant HA and NA segments were selected, giving a total of 3850 and 1220 HA and NA protein sequences respectively. Non-redundant HA and NA segments for each HA and NA subtype were selected to ensure that the different subtypes and host origins within each subtype were represented. Random uniform sampling was carried out for sequences from within each subtype and from each of the three host species (human, avian and swine). The sequences were grouped according to subtype or host, and covered all the viral subtypes found in that host. The sequences were downloaded in FASTA format (including accession number, subtype, gene, host, occurring year, and other parameters) and were parsed into each category. We used the amino acid sequences (20 letter alphabet) because they are known to give more reliable results than nucleotide sequences (4 letters alphabet), whose divergence is high (Suzuki and Nei, 2002).

In this study, along with the five most important subtypes, H1, H2, H3, H5 and H9, we also chose three host groups, human, avian, and swine, because birds appear to be the reservoir of the influenza A virus and swine act as an intermediary between avian and human viruses. The HA segment alone was used for host classification modeling and signature identification; part of the data was used for training and the remaining part was used for testing (Table 1).

To compare the genomic patterns of the avian, swine and human influenza viruses, we downloaded 1500 HA protein sequences that had been isolated from the three host species (500 for each host, applying random uniform sampling) from the NCBI Influenza Virus Resource. The signatures that we obtained by analyzing the primary dataset were validated and tested using human, avian, or swine test sets. The signatures and positions in the sequences from the different hosts were numbered in accordance with the H1N1 influenza A virus Puerto Rico strain HA reference protein [GenBank:NP_040980.1]. The training sets from each group (Table 1) were aligned using ClustalW 2.0.

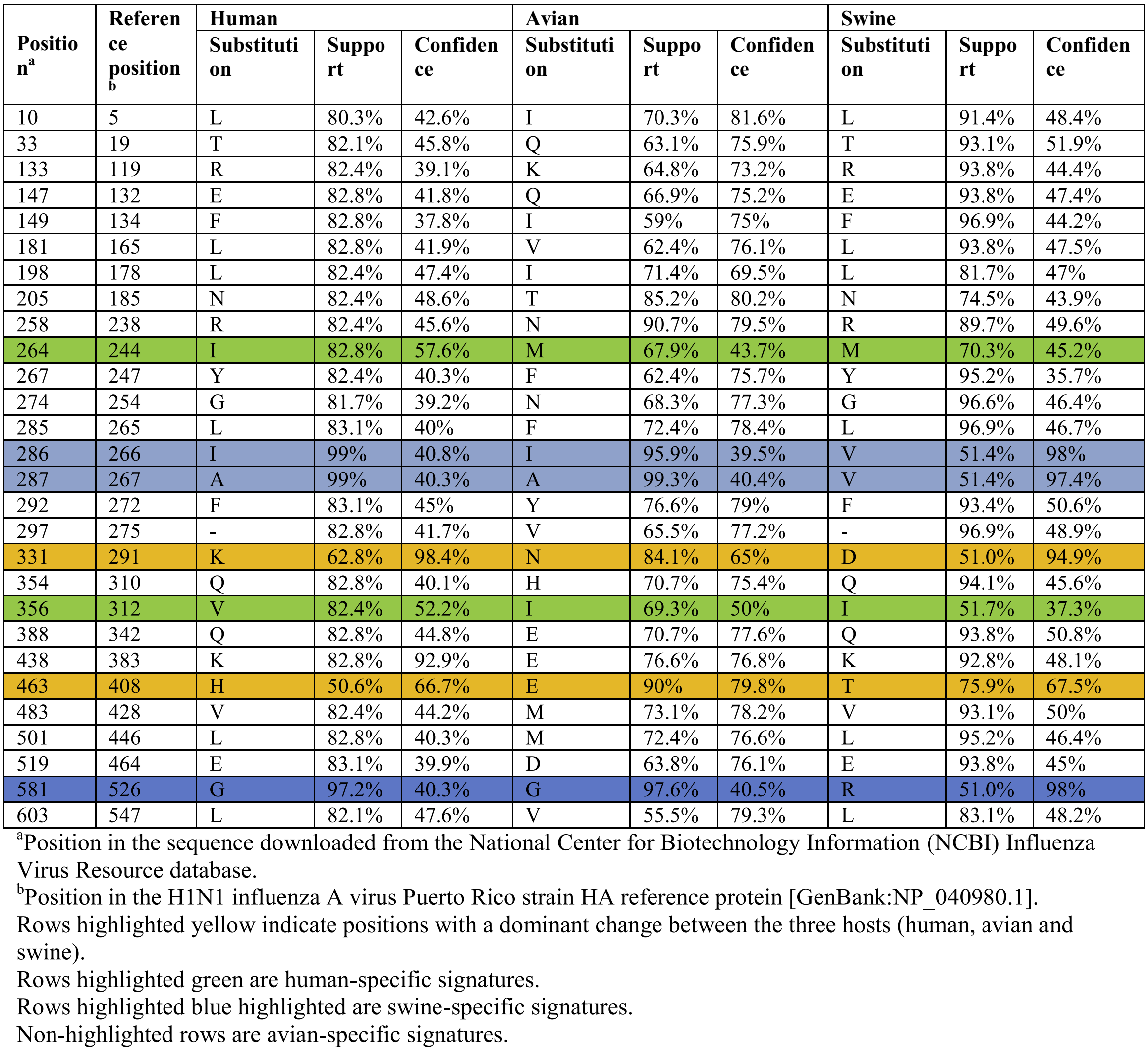
Pattern discovery and feature selection

The VESPA program (available from http://www.hcv.lanl.gov) can be used to quickly detect amino acids that characterize

Table 2
Summary of host classification results for influenza A virus using HMMs and DTs.

HA subtype	Host	Using HMM			Using DT		
		Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
H1	Human	94.4	93.7	95.7	91.3	92.9	95.5
	Avian	89.5	100	95.3	94.6	90	99
	Swine	84.5	90.3	82.9	92	91.1	92.5
H2	Human	100	94.1	100	100	100	100
	Avian	90	91.7	87.5	100	100	100
H3	Human	80.8	86.9	71.1	94.3	94	98
	Avian	90.9	82.4	92.7	93	95	97.5
	Swine	78.7	71.4	78.8	94	94	96
H5	Human	53	95.8	39.5	91.3	95	87.5
	Avian	55	44.7	87.5	91.2	87.5	95
H9	Avian	55	46.7	80	100	100	100
	Swine	90	80	93.3	100	100	100

Table 3
Host-origin class association positional signatures in HA and their support and confidence ratings. Positional signatures for human vs. avian vs. swine are shown.



Position[a]	Reference position[b]	Human			Avian			Swine		
		Substitution	Support	Confidence	Substitution	Support	Confidence	Substitution	Support	Confidence
10	5	L	80.3%	42.6%	I	70.3%	81.6%	L	91.4%	48.4%
33	19	T	82.1%	45.8%	Q	63.1%	75.9%	T	93.1%	51.9%
133	119	R	82.4%	39.1%	K	64.8%	73.2%	R	93.8%	44.4%
147	132	E	82.8%	41.8%	Q	66.9%	75.2%	E	93.8%	47.4%
149	134	F	82.8%	37.8%	I	59%	75%	F	96.9%	44.2%
181	165	L	82.8%	41.9%	V	62.4%	76.1%	L	93.8%	47.5%
198	178	L	82.4%	47.4%	I	71.4%	69.5%	L	81.7%	47%
205	185	N	82.4%	48.6%	T	85.2%	80.2%	N	74.5%	43.9%
258	238	R	82.4%	45.6%	N	90.7%	79.5%	R	89.7%	49.6%
264	244	I	82.8%	57.6%	M	67.9%	43.7%	M	70.3%	45.2%
267	247	Y	82.4%	40.3%	F	62.4%	75.7%	Y	95.2%	35.7%
274	254	G	81.7%	39.2%	N	68.3%	77.3%	G	96.6%	46.4%
285	265	L	83.1%	40%	F	72.4%	78.4%	L	96.9%	46.7%
286	266	I	99%	40.8%	I	95.9%	39.5%	V	51.4%	98%
287	267	A	99%	40.3%	A	99.3%	40.4%	V	51.4%	97.4%
292	272	F	83.1%	45%	Y	76.6%	79%	F	93.4%	50.6%
297	275	-	82.8%	41.7%	V	65.5%	77.2%	-	96.9%	48.9%
331	291	K	62.8%	98.4%	N	84.1%	65%	D	51.0%	94.9%
354	310	Q	82.8%	40.1%	H	70.7%	75.4%	Q	94.1%	45.6%
356	312	V	82.4%	52.2%	I	69.3%	50%	I	51.7%	37.3%
388	342	Q	82.8%	44.8%	E	70.7%	77.6%	Q	93.8%	50.8%
438	383	K	82.8%	92.9%	E	76.6%	76.8%	K	92.8%	48.1%
463	408	H	50.6%	66.7%	E	90%	79.8%	T	75.9%	67.5%
483	428	V	82.4%	44.2%	M	73.1%	78.2%	V	93.1%	50%
501	446	L	82.8%	40.3%	M	72.4%	76.6%	L	95.2%	46.4%
519	464	E	83.1%	39.9%	D	63.8%	76.1%	E	93.8%	45%
581	526	G	97.2%	40.3%	G	97.6%	40.5%	R	51.0%	98%
603	547	L	82.1%	47.6%	V	55.5%	79.3%	L	83.1%	48.2%

[a] Position in the sequence downloaded from the National Center for Biotechnology Information (NCBI) Influenza Virus Resource database.
[b] Position in the H1N1 influenza A virus Puerto Rico strain HA reference protein [GenBank:NP_040980.1].
Rows highlighted yellow indicate positions with a dominant change between the three hosts (human, avian and swine).
Rows highlighted green are human-specific signatures.
Rows highlighted blue highlighted are swine-specific signatures.
Non-highlighted rows are avian-specific signatures.

differences between two groups of sequences. It enumerates a set of amino acids that are conserved in one group, and differ in another group (Korber and Myers, 1992). VESPA will pick out the differentiated amino acids and calculate their frequencies in each set. The sequences should all be of the same length (the total length of the aligned protein sequences was 630 amino acids), and the threshold should be adjusted for the minimum degree of conservation of signature amino acids in the query set.

Two Sample Logo is a tool that can be used to calculate the statistical significance of the relative position-specific symbol frequencies between two sets of aligned sequences (Vacic and Radivojac, 2006). We used it to generate host-origin graphical logos that are subtype specific (Fig. 2 and Additional file 1), and non-subtype specific (Fig. 4 and Additional file 4). These graphical logos have been used previously to discriminate between different classes of sequences such as viral sequences from responders to treatment vs. non-responders (ElHefnawi et al., 2010).

Table 4
Top-ranked host-origin class association rules extracted from the most informative positions (markers). The most informative rules with the highest support and confidence ratings are listed.

Position[a]	Reference position[b]	Rule	Support rating (%)	Confidence rating (%)
438K	383K	K→ Human	82.8	92.9
258N	238R	N→ Avian	90.7	79.5
463E	408T	E→Avian	90	79.8
205T	185K	T→Avian	85.2	80.2
603L	547L	V→Avian	55.5	79.3
286V	266I	V→ Swine	51.4	98

[a] Position in the sequence downloaded from the National Center for Biotechnology Information (NCBI) Influenza Virus Resource database.
[b] Position in the H1N1 influenza A virus Puerto Rico strain HA reference protein [GenBank:NP_040980.1].

Subtyping and host-typing classifier models

The first classifier models that we implemented using the HMMER package version 2.3.2 were the profile HMMs for subtyping and host-typing (Eddy, 1998). Profile HMMs are statistical models of multiple sequence alignments that can be used for protein homology detection (Schuster-Bockler and Bateman, 2007; Eddy, 1998). They capture position-specific information about each column of the multiple sequence alignment. This makes the HMMs more sensitive for remote homology database searches than those based on pairwise alignments. HMMER is the engine that is used in other databases, including TIGRFAM (Haft et al., 2003) and SMART (; Schultz et al., 2000).

Profile HMM models were first built for each subtype; followed by models for each major host in each subtype. Next, multiple alignments for each host, irrespective of subtype, were carried out and models for each host, irrespective of subtype, were used to find signatures as described above. Database searches to score a sequence against the model followed. Model building involved converting the multiple alignment of each group of sequences into a probabilistic model, while database searches involved scoring a sequence to the profile HMM (Eddy, 1998).

A DT model and associative classification were also used to classify host origins. These are standard data mining techniques that have been used for a wide range of applications in classification problems (C4.5). DT algorithms such as C4.5, CART and regression trees can also be used to classify and identify important features for classification. A DT is a supervised approach to classification. Each node in a DT represents a feature in the instance to be classified, and each branch represents a value that the node can assume (Murthy, 1998). The WEKA classifier package was used to implement this classifier. The package is a collection of machine learning algorithms for data mining tasks such as DT classification (Gewehr et al., 2007). The most informative positions (found from both the VESPA and comparative logo analyses) that differed between the human, swine and avian influenza viruses were used as features (attribute) for DT generation. The DTs were generated using the C4.5 algorithm as implemented in the WEKA 3.7.5 program known as J48. Five separate DTs for the HA proteins from H1, H2, H3, H5 and H9 were generated, to identify more precisely the host of origin for each subtype. The remaining HA subtypes were found in avian hosts only, so identifying their subtypes was enough and here was no need for further host classification.

The third classifier, associative classification, is easy to implement and interpret. Class association rules were generated for the sites with statistically significant variations between the host groups in both the VESPA and the comparative sequence logo analysis; only sites whose support and confidence ratings were above 40% were retained.

Evaluation of classification models

Evaluation of the different classification models was performed using a 10-fold cross validation for the DT models in the WEKA tool. Specificity, sensitivity, accuracy, and area under the curve for the receiver operating characteristic (ROC) analysis were calculated using the MedCalc program (Nomura, 1979). For the HMM models, testing on the 25% test set was performed. While for the associative classification, calculating the support and confidence ratings for the class-association rules and selection of top-ranked rules were conducted.

Authors' contributions

ME envisioned, designed, and helped supervise the study. FF and ME collected the biological background and proposed data, contributed to the overall design, built the classifiers, carried out all the computational work, and drafted the manuscript. ME designed the methodologies, planned the data analysis and revised and edited the manuscript. Both authors read and approved the final manuscript.

Acknowledgments

We thank Prof. Yasser Kadah for his support, Dr. Khaled El-Sayed and Dr. Feryal Farouk for helpful discussions, and Nourhan Abu-Shahba for help in editing the manuscript.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2013.11.010.

References

Allen, J.E., Gardner, S.N., Vitalis, E.A., Slezak, T.R., 2009. Conserved amino acid markers from past influenza pandemic strains. BMC Microbiol. 9, 77.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Attaluri, P.K., Chen, Z., Weerakoon, A., Lu, G., 2009a. Integrating decision tree and Hidden Markov Model (HMM) for subtype prediction of human influenza A virus, Cutting-Edge Research Topics on Multiple Criteria Decision Making: 2009. Chengdu/Jiuzhaigou, China.

Attaluri, P.K., Chen, Z., Lu, G., 2010. Applying neural networks to classify influenza virus antigenic types and hosts. In: Proceedings of the IEEE Symposium onComputational Intelligence in Bioinformatics and Computational Biology (CIBCB), Montreal.

Attaluri, P.K., Zheng, X., Z., Chen, G., Lu, 2009b. Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic. In: Proceedings of the 6th Annual Biotechnology and Bioinformatics Symposium (BIOT-2009). The University of Nebraska-Lincoln, Lincoln, Nebraska.

Auewarakul, P., Suptawiwat, O., Kongchanagul, A., Sangma, C., Suzuki, Y., Ungchusak, K., Louisirirotchanakul, S., Lerdsamran, H., Pooruk, P., Thitithanyanont, A., et al., 2007. An avian influenza H5N1 virus that binds to a human-type receptor. J. Virol. 81 (18), 9950–9955.

Bao, Y.B.P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., Lipman, D., 2008. The influenza virus resource at the National Center for Biotechnology Information. J. Virol. 82 (2), 596–601.

Beare, A.S., Webster, R.G., 1991. Replication of avian influenza viruses in humans. Arch. Virol. 119 (1–2), 37–42.

C4.5. Programs for Machine Learning.

Chander, Y., Jindal, N., Stallknecht, D.E., Sreevatsan, S., Goyal, S.M., 2010. Full length sequencing of all nine subtypes of the neuraminidase gene of influenza A viruses using subtype specific primer sets. J. Virol. Methods 165 (1), 116–120.

Chen, G.W., Shih, S.R., 2009. Genomic signatures of influenza A pandemic (H1N1) 2009 virus. Emerg. Infect. Dis. 15 (12), 1897–1903.

Chen, G.W., Chang, S.C., Mok, C.K., Lo, Y.L., Kung, Y.N., Huang, J.H., Shih, Y.H., Wang, J. Y., Chiang, C., Chen, C.J., et al., 2006. Genomic signatures of human versus avian influenza A viruses. Emerg. Infect. Dis. 12 (9), 1353–1360.

Connor, R.J., Kawaoka, Y., Webster, R.G., Paulson, J.C., 1994. Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates. Virology 205 (1), 17–23.

Deng, T., Sharps, J.L., Brownlee, G.G., 2006. Role of the influenza virus heterotrimeric RNA polymerase complex in the initiation of replication. J. Gen. Virol. 87 (Pt 11), 3373–3377.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics 14 (9), 755–763.

ElHefnawi, M., Alaidi, O., Mohamed, N., Kamar, M., El-Azab, I., Zada, S., Siam, R., 2011. Identification of novel conserved functional motifs across most Influenza A viral strains. Virol. J. 8, 44.

ElHefnawi, M.M., Zada, S., El-Azab, I.A., 2010. Prediction of prognostic biomarkers for interferon-based therapy to hepatitis C virus patients: a meta-analysis of the NS5A protein in subtypes 1a, 1b, and 3a. Virol. J. 7, 130.

ElHefnawi, M., Hassan, N., Kamar, M., Siam, R., Remoli, A.L., El-Azab, I., et al., 2011. The design of optimal therapeutic small interfering RNA moleculestargeting diverse strains of influenza A virus. Bioinformatics 27 (24), 3364–3370.

Fayroz, F., Sherif, Y.K., Mahmoud, ElHefnawi, 2012. Influenza A subtyping and host origin classification using profile hidden Markov models. J. Mech. Biol. Med.

Finkelstein, D.B., Mukatira, S., Mehta, P.K., Obenauer, J.C., Su, X., Webster, R.G., Naeve, C.W., 2007. Persistent host markers in pandemic and H5N1 influenza viruses. J. Virol. 81 (19), 10292–10299.

Gabriel, G., Dauber, B., Wolff, T., Planz, O., Klenk, H.D., Stech, J., 2005. The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. Proc. Natl. Acad. Sci. USA 102 (51), 18590–18595.

Garten, R.J., Davis, C.T., Russell, C.A., Shu, B., Lindstrom, S., Balish, A., Sessions, W.M., Xu, X., Skepner, E., Deyde, V., et al., 2009. Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325 (5937), 197–201.

Gendoo, D.M., El-Hefnawi, M.M., Werner, M., Siam, R., 2008. Correlating novel variable and conserved motifs in the hemagglutinin protein with significant biological functions. Virol. J. 5, 91.

Gewehr, J.E., Szugat, M., Zimmer, R., 2007. BioWeka – extending the Weka framework for bioinformatics. Bioinformatics 23 (5), 651–653.

Haft, D.H., Selengut, J.D., White, O., 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31 (1), 371–373.

Horimoto, T., Kawaoka, Y., 2005. Influenza: lessons from past pandemics, warnings from current incidents. Nat. Rev. Microbiol. 3 (8), 591–600.

Kawaoka, Y., Krauss, S., Webster, R.G., 1989. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J. Virol. 63 (11), 4603–4608.

Kida, H., Ito, T., Yasuda, J., Shimizu, Y., Itakura, C., Shortridge, K.F., Kawaoka, Y., Webster, R.G., 1994. Potential for transmission of avian influenza viruses to pigs. J. Gen. Virol. 75, 2183–2188. (Pt 9).

Klumpp, K., Ruigrok, R.W., Baudin, F., 1997. Roles of the influenza virus polymerase and nucleoprotein in forming a functional RNP structure. EMBO J. 16 (6), 1248–1257.

Korber, B., Myers, G., 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retrovir. 8 (9), 1549–1560.

Murthy, S.K., 1998. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Min. Knowl. Discov. 2 (4), 345–389.

Matrosovich, M.N., Gambaryan, A.S., Teneberg, S., Piskarev, V.E., Yamnikova, S.S., Lvov, D.K., Robertson, J.S., Karlsson, K.A., 1997. Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. Virology 233 (1), 224–234.

Munch, M., Nielsen, L.P., Handberg, K.J., Jorgensen, P.H., 2001. Detection and subtyping (H5 and H7) of avian type A influenza virus by reverse transcription-PCR and PCR-ELISA. Arch. Virol. 146 (1), 87–97.

Nomura, Y., 1979. Significance of the ROC (receiver operating characteristics) curve in diagnostic tests. Nihon Rinsho Suppl, 1402–1404.

Pedersen, J.C., 2008. Hemagglutination-inhibition test for avian influenza virus subtype identification and the detection and quantitation of serum antibodies to the avian influenza virus. Methods Mol. Biol. 436, 53–66.

Rota, P.A., Rocha, E.P., Harmon, M.W., Hinshaw, V.S., Sheerar, M.G., Kawaoka, Y., Cox, N.J., Smith, T.F., 1989. Laboratory characterization of a swine influenza virus isolated from a fatal case of human influenza. J. Clin. Microbiol. 27 (6), 1413–1416.

Scholtissek, C., Rohde, W., Von Hoyningen, V., Rott, R., 1978. On the origin of the human influenza virus subtypes H2N2 and H3N2. Virology 87 (1), 13–20.

Scholtissek, C., Burger, H., Bachmann, P.A., Hannoun, C., 1983. Genetic relatedness of hemagglutinins of the H1 subtype of influenza A viruses isolated from swine and birds. Virology 129 (2), 521–523.

Scholtissek, C., Burger, H., Kistner, O., Shortridge, K.F., 1985. The nucleoprotein as a possible major factor in determining host specificity of influenza H3N2 viruses. Virology 147 (2), 287–294.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P., 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 28 (1), 231–234.

Schultz, U., Fitch, W.M., Ludwig, S., Mandler, J., Scholtissek, C., 1991. Evolution of pig influenza viruses. Virology 183 (1), 61–73.

Schuster-Bockler, B., Bateman, A., 2007. An introduction to hidden Markov models. Curr. Protoc. Bioinformatics, pp. 1–12. (Appendix 3–Appendix 3A).

Sherif, F.F., El Hefnawi, M., Kadah, Y., 2011. Genomic signatures and associative classification of the hemagglutinin protein for Human versus Avian versus Swine Influenza A viruses. In: Proceeding of IEEE 2011, pp. 1–8.

Sherif, F.F., El-Hefnawi, M., Kadah, Y.M., 2012. Influenza A subtyping and host origin classification using profile hidden Markov models. J. Mech. Med. Biol. 12 (2), 1240009.

Shinya, K., Hamm, S., Hatta, M., Ito, H., Ito, T., Kawaoka, Y., 2004. PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice. Virology 320 (2), 258–266.

Starick, E., Romer-Oberdorfer, A., Werner, O., 2000. Type- and subtype-specific RT-PCR assays for avian influenza A viruses (AIV). J. Vet. Med. B Infect. Dis. Vet. Public Health 47 (4), 295–301.

Steel, J., Lowen, A.C., Mubareka, S., Palese, P., 2009. Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N. PLoS Pathog. 5 (1), e1000252.

Suzuki, Y., Nei, M., 2002. Origin and evolution of influenza virus hemagglutinin genes. Mol. Biol. Evol. 19 (4), 501–509.

Triki, H., 1997. Clinical virology laboratory. Arch. Inst. Pasteur Tunis 74 (1–2), 51–55.

Vacic, V.I.L., Radivojac, P., 2006. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22 (12), 1536–1537.

Wiley, D.C., Skehel, J.J., 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annu. Rev. Biochem. 56, 365–394.

Wong, S.S., Yuen, K.Y., 2006. Avian influenza virus infections in humans. Chest 129 (1), 156–168.

Wu, L.C., Horng, J.T., Huang, H.D., Chen, W.L., 2008. Identifying discriminative amino acids within the hemagglutinin of human influenza A H5N1 virus using a decision tree. IEEE Trans. Inf. Technol. Biomed. 12 (6), 689–695.

Zhang, Y., Lin, X., Zhang, F., Wu, J., Tan, W., Bi, S., Zhou, J., Shu, Y., Wang, Y., 2009. Hemagglutinin and neuraminidase matching patterns of two influenza A virus strains related to the 1918 and 2009 global pandemics. Biochem. Biophys. Res. Commun. 387 (2), 405–408.