Journal of Mechanics in Medicine and Biology
Vol. 12, No. 2 (2012) 1240009 (11 pages)

DOI: 10.1142/S021951941240009X

World Scientific
www.worldscientific.com

INFLUENZA A SUBTYPING AND HOST ORIGIN CLASSIFICATION USING PROFILE HIDDEN MARKOV MODELS

FAYROZ F. SHERIF
Bioelectronics Department
Modern University for Technology and Information
Cairo, Egypt
ffs@k-space.org

MAHMOUD EL-HEFNAWI
Informatics and Systems Department
National Research Centre, Giza, Egypt
mahef@aucegypt.edu

YASSER M. KADAH
Biomedical Engineering Department
Cairo University, Giza, Egypt

Received
Revised
Accepted

Influenza is one of the most important emerging and reemerging infectious diseases, causing high morbidity and mortality in communities (epidemic) and worldwide (pandemic). Here, classification of human vs. non-human influenza, and subtyping of human influenza viral strains virus is done based on profile hidden Markov models (HMM). The classical ways of determining influenza viral subtypes depend mainly on antigenic assays, which is time-consuming and not fully accurate. The introduced technique is much cheaper and faster, yet usually can still yield high accuracy. Multiple sequence alignments were done for the 16 HA subtypes and 9 NA subtypes, followed by profile-HMMs models generation, calibration and evaluation using the HMMER suite for each group. Subtyping accuracy of all HA and NA models achieved 100%, while host classification achieved accuracies around 53% and 95.1% according to HA subtype.

Keywords: Bioinformatics; influenza virus; profile hidden Markov model.

1. Introduction

Influenza A viruses belong to the Orthomyxoviridae family of negative sense, single-stranded, segmented RNA viruses. The RNA core consists of 8 gene segments. Immunologically, the most significant surface proteins include Hemagglutinin HA (16 subtypes) and Neuraminidase NA (9 subtypes). Influenza A subtypes are

traditionally identified by their HA and NA proteins.[1,2] The HA and NA proteins are integral membrane proteins and are considered as the major surface antigen of the influenza virus virion. HA is responsible for binding of virions to host cell receptors and for fusion between the virion envelope and the host cell.[3] 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 virus spread.[4] All the 16 subtypes of HA and 9 subtypes of NA are found in avian but the first three subtypes H1, H2, H3 and recently H5, are found in human influenza viruses.[5] The most common strains which infect humans during the annual influenza season are H1N1 and H3N2.[6] Swine influenza is known to be caused by influenza A subtypes H1N1, H1N2, H3N1, and H3N2. Rapid virus subtype identification is critical for accurate diagnosis of human infections, effective response to epidemic outbreaks and global-scale surveillance of highly pathogenic subtypes such as avian influenza H5N1 and H1N1 2009 virus.[7] The classical ways of subtyping influenza A virus for HA segments are hemagglutination inhibition (HI) assay which are capable of distinguishing antigenic differences between influenza even of the same subtype. However, as noted in Ref. 8, when working with uncharacterized viruses or antibody subtypes, the library of reference reagents required for identifying antigenically distinct influenza viruses and/or antibody specificities from multiple lineages of a single HA subtype requires extensive laboratory support for the production and optimization of reagents. Another possible method is the subtyping of HA genes by reverse transcription PCR.[9] Real-time PCR is highly specific. But there are some things to be considered such as cost and time. While the cost of primers is probably manageable, probes are very expensive. There will be a lag time as we will have to obtain all the probes and primers and do validation studies. A common way to find which subtype a genetic sequence belongs to is through the BLAST search.[10] However, there are issues associated with the BLAST algorithm as described in Ref. 11. Most importantly, the BLAST result cannot reveal important mutations that may be functionally related to the structure and function of proteins.

Profile hidden Markov models (HMMs) are statistical models of multiple sequence alignments.[12] They capture position-specific information about how conserved each column of the alignment is, and which residues are likely. Recently related studies have been conducted to classify influenza virus antigenic types and hosts. An Integrated approach of using decision trees and HMM for subtype prediction of human influenza A virus — HA subtypes (H1, H2 and H3) and NA subtypes (N1 and N2) — has been introduced in Ref. 13. They extracted some informative positions from decision tree algorithms in the Weka package, and then modeled into profiles through hidden Markov modeling at nucleotide level, using HMMER with subtype prediction accuracy of 88% for human subtypes. Also, they developed a web system for accurate subtype detection of human influenza virus sequences only. The preliminary experiment showed that this system is easy-to-use but not powerful in identifying human influenza subtypes and there is no facility to use protein sequences. Another study in Ref. 14 applied two machine learning techniques

(decision trees and support vector machines) to identify the origin of latest pandemic outbreak of H1N1 viral strains. Their results have shown that human and swine groups are well distinguishable, with classification accuracy above 95% at prediction. All sequences from HA, M, NA, NP, NS, PA, PB1, and PB2 are classified as swine influenza, which means sequences in these segments are more closely related to Swine strain. Therefore, it was suggested that the latest pandemic viral strain is of swine origin. Finally, the most recently study discussed in Ref. 15 has applied the feed-forward back-propagation neural network for the classification analysis of influenza virus.

Our study aims to generalize and extend influenza subtype and host classification to include all influenza A viral subtypes and host origins, by developing a prediction tool using Profile HMM at protein level, for identifying all influenza viral strains in the different hosts not only human. In this work, the subtype prediction achieved 100% accuracy while host origin identification achieved accuracies around 53% and 95.1% according to HA subtype.

2. Data and Methods

2.1. *Data collection*

All sequences were downloaded from the NCBI's (National Center for Biotechnology Information) Influenza Virus Resources.[16] We ensured the downloaded sequences were non-redundant and the complete isolation of HA and NA segments. Part of the data is used for training and the remaining part is used for testing (Table 1). We used amino acid sequences because they are known to give more reliable results than nucleotide sequences when the sequence divergence is high.[17]

2.2. *Multiple sequence alignment (MSA)*

One of the cornerstones of modern bioinformatics is the comparison or alignment of protein sequences. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment).[18] Each group of training sets found in Table 1 was collectively aligned using Clustal X program, which supports multiple sequence alignment for protein sequences through window graphical user interface and built by adding the sequences sequentially to the growing MSA produced a consensus sequence representing the highly conserved regions from the aligned sequences.[19,20]

2.3. *Modeling using profile HMM*

Profile HMM techniques are among the most powerful methods for protein homology detection scoring them above the noise level.[21] HMM profile includes more flexible information on a given set of sequences than a single sequence.[22] Therefore, database search methods using profiles is more sensitive to remote similarities than those

Table 1. Number of downloaded sequences used for each subtype of HA and NA segments.

HA segment	Group	# of training sequences	# of test sequences
H1	Total H1	971	100
	H1-Human	749	100
	H1-Avian	105	10
	H1-Swine	300	68
H2	Total H2	126	50
	H2-Human	50	13
	H2-Avian	124	30
H3	Total H3	814	100
	H3-Human	550	69
	H3-Avian	263	30
	H3-Swine	100	29
H4	Total (Avian)	200	64
H5	Total H5	1500	256
	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	400	97
	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	1500	205
	N1-Human	600	56
	N1-Avian	830	70
	N1-Swine	100	20
N2	Total N2	1500	561
	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

based on pairwise alignments (e.g., regular BLAST). In particular, profile HMM have generated good results, and are today employed by several databases such as Pfam and Superfamily.[23,24] We divided our analysis into two main steps; profile HMM model building and database searching.

Model building involves converting a multiple alignment of each group of sequences into a probabilistic model, while database searching involves scoring a sequence to the profile HMM. One of the most widely used profile HMM packages is HMMER packages.

2.4. *Model building*

A profile HMM is a probabilistic model of multiple alignments of related proteins. The alignment is modeled using a series of nodes (roughly one per alignment column) each composed of three states: match, insert and delete. Match and insert states emit amino acids with probabilities learned during model estimation while delete states are quiet. Insertions and deletions with respect to the HMM are modeled by insert and delete states and transition probabilities to them.[12] "Hmmbuild" program in HMMER package v2.3.2 was used to build a different HMM profiles for each subtype of HA and NA segments; the input to "Hmmbuild" program were the pre-aligned sequences of each group in Table 1. In order to increase the sensitivity of database search we used "hmmcalibrate" program in HMMER to calculate the E-value. The E-value is quite literally the expected number of false positives at this raw score; the larger the database you search, the greater the number of expected false positives. HMM database has been built by concatenating HMM files that are already built and calibrated.[25]

2.5. *Database searching*

Any sequence can be compared to a model by calculating the probability that the sequence was generated by that model. The negative logarithm of this probability corresponds to the NULL score calculated for a simple HMM. To score a match to HMM we have two algorithms: Viterbi algorithm to give the probability of the most probable alignment with the sequence or Forward algorithm to give the full probability of a sequence aligning to the profile HMM.[26] "Hmmsearch" program in HMMER package searches one or more sequences against HMM profile. The output of the program is the sequence family classification top hits list, ranked by E-value. The scores and E-values here reflect the confidence that this query sequence contains one or more domains belonging to a domain family. "HmmPfam" program Searches an HMM database for matches to a query sequence and get score for each model.[23]

3. Results

Multiple sequence alignments were done for the 16 HA subtypes, 9 NA subtypes and 12 "HA-Host" host specific subtypes, using ClustalX, followed by profile-HMMs models building, calibration and database generation using the HMMER suite for each group.

3.1. *Subtyping classification results*

Subtyping classification was done by scoring the entire test-sets (human) (Table 1), with each HA and NA HMM models, using "HMMPfam" program in HMMER suite.

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

HA subtype	Host	Accuracy	Sensitivity	Specificity
H1	Human	94.4%	93.7%	95.7%
	Avian	89.5%	100%	95.3%
	Swine	84.5%	90.3%	82.9%
H2	Human	95.1%	100%	96%
	Avian	90%	91.7%	87.5%
H3	Human	80.8%	86.9%	71.1%
	Avian	90.9%	82.4%	92.7%
	Swine	78.7%	71.4%	78.8%
H5	Human	53%	95.2%	43.5%
	Avian	63%	58.1%	76.9%
H9	Avian	55%	46.7%	80%
	Swine	90%	80%	93.3%

Matches to the right HA or NA subtype were classified as true hits. Matches to a different subtype were classified as false hits. The accuracies of classification results achieved 100%. These results are encouraging and bear great promise for application to influenza virus classification. Therefore any viral strain like H1N1, H1N2, H2N2, H3N2, H5N1 and H9N2 can by accurately classified using HMMs with 100% accuracy.

3.2. *Host classification results*

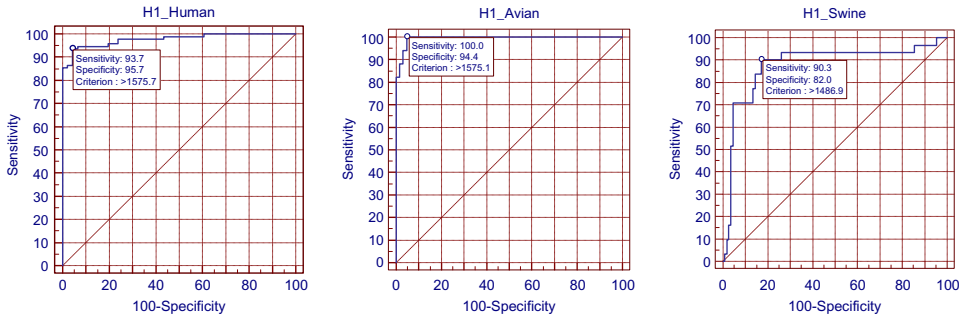
Identifying the origin of viral strains as human avian or swine has been done by scoring the pre-identified HA subtype with the corresponding "HA-Host" HMM models for better matching. "HMMSearch" program in HMMER suite has been used for this classification. The test results details of host classification for different HA subtypes in terms of accuracy, sensitivity and specificity are summarized in Table 2.

3.3. *Model evaluation using ROC analysis*

In a receiver operating characteristic (ROC) curve the true positive rate (sensitivity) is plotted in function of the false positive rate (100-specificity) for different cut-off points. Each point on the ROC plot represents a sensitivity/specificity pair corresponding to a particular decision threshold.[27] The following ROC curves were drawn using MedCalc program.[28] The curves indicate the observed criterion (threshold) values that maximized both sensitivity and specificity values. ROC curves for host identification of different HA subtypes are indicated in Figs 1−5.

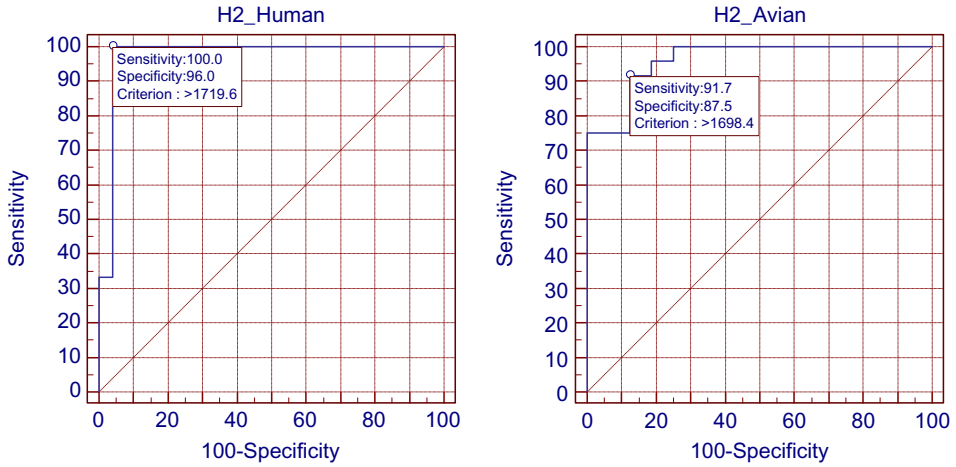
4. Discussion

The obtained results confirm that profile HMM can successfully be used for classifying all influenza A stains hosted in all species in two major steps. First through



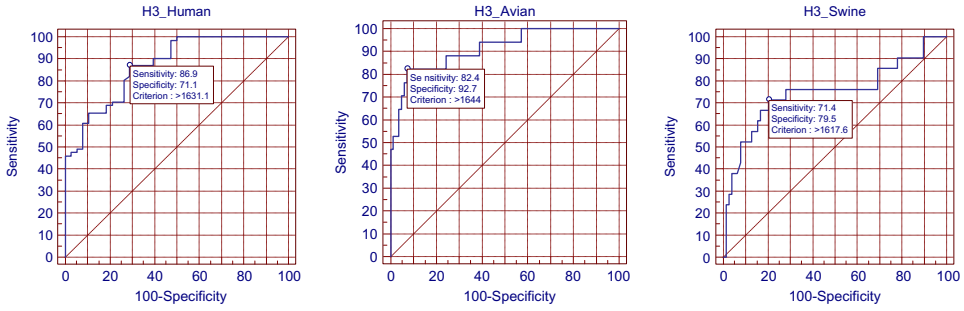
(a) (b) (c)

Fig. 1. ROC curves for host classification results of H1-human, H1-Avian and H1-Swine using HMM.



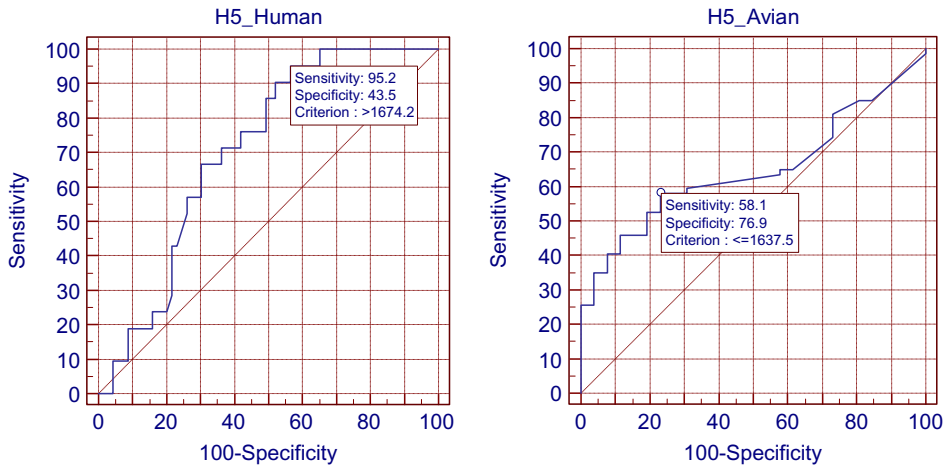
(a) (b)

Fig. 2. ROC curves for host classification results of H2-human and H2-Avian using HMM.



(a) (b) (c)

Fig. 3. ROC curves for host classification results of H3-human, H3-Avian and H3-Swine using HMM.

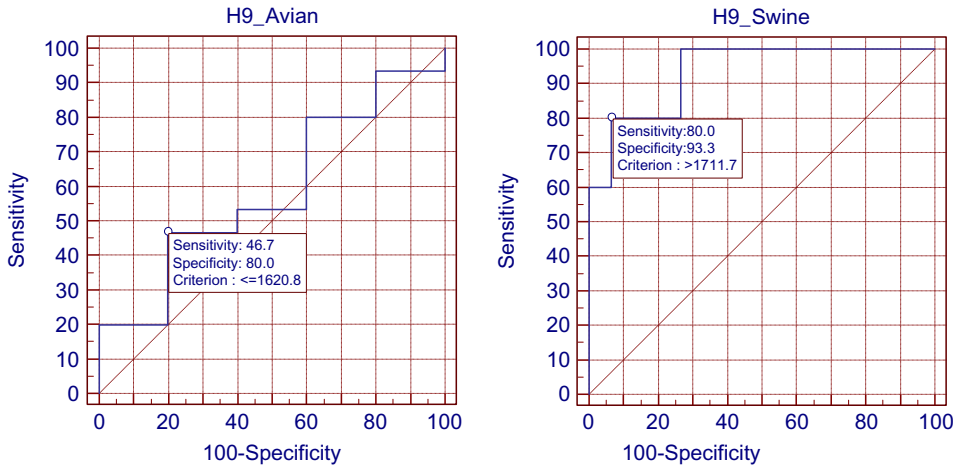


(a) (b)

Fig. 4. ROC curves for host classification results of H5-human and H5-Avian using HMM.

identifying HA and NA subtypes. Second through predicting the host of origin of the pre-identified HA subtypes, by scoring it with the corresponding "HA subtype-Host HMM" models, searching for the best match.

For example, if a query HA sequence has been searched with each HA model separately, and we get the highest score with H1 model for example, then the entire sequence will be farther scored with each H1 specified host separately: H1-human, H1-Swine and H1-Avian models searching for the highest match.



(a) (b)

Fig. 5. ROC curves for host classification results of H9-Avian and H9-Swine using HMM.

All the 16 HA and 9 NA models have sensitivity of 100% and specificity of 100%. Although, there are differences in the criteria used in Attaluri *et al.*'s study and this study, their findings may support our findings that any unknown viral strain of influenza A, can be easily distinguishable as they have an extensive genetic diversity in HA and NA subtypes. Notably, our results achieved higher accuracy over Attaluri *et al.*'s study.[13] On the other hand, host classification of any viral sequence as human, avian or swine varied according to HA subtype. Among HA subtypes, there were some HAs (H1, H2, H3, H5 and H9) that can infect more than one species, through transmission of the whole virus or ever, the reassortment between avian and human viruses. Also, we found that some of those HA subtypes which can infect more than one species; vary greatly between human, swine and avian viruses. While some others vary little so it was difficult to identify their host of origin.

By comparing our results, we found that, H2 HA models have a higher accuracy over H1, H3, H5 and H9 HAs models. These results indicate that H2 viral subtypes have more genetic diversity between human and avian, compared to the other subtypes. In contrast, H5 HA models accuracies were not much higher than 53% for H5-Human and 63% for H5-Avian. This means that, no significant differences can be detected between human and avian H5 viruses using HMM.

These results agree with previous findings in Refs. 29 and 30, that highly pathogenic avian influenza H5N1 virus strains can transmit directly from avian species to humans and cause severe disease. The receptor binding preference of H5N1 viruses can be altered by only a few amino acid substitutions in the HA protein. H1 HA has accuracies of 94.4%, 84.5% and 89.5% for H1-human, H1- Swine and H1-Avian models, respectively. The host classification of H3 HA has the accuracies of 80.8%, 78.7% and 90.9% for H3-human, H3-Swine and H3-Avian models, respectively. These results seem reasonable as cross-species infections usually take place in these subtypes, through reassortment or through whole host shift events. Nevertheless, further improvement may be required in host classification to achieve higher accuracy. The remaining subtypes of HA are found only in avian hosts, so once they are classified by their subtypes as H4, H6, H7, H8, H10-H16, etc. they are also identified as having an avian host specification.

5. Conclusions

Accurate detection of influenza viral origin and subtyping can significantly improve influenza surveillance and vaccine development. In this study, host identification and subtyping of influenza A virus were done based on HMMs for each subtype and major hosts (humans, avian, and swine). This study demonstrated the power of integrating the multiple sequence alignment and profile HMM approaches in classifying influenza A viral stains and their host of origin. In conclusion, our results indicate that influenza A sequences are HA and NA subtype specific and highly sensitive against HMM models (H1-H16), (N1-N9) and can easily be predicted with 100% accuracy. Host classification has accuracies that vary between 53% and 95.1% according to HA subtype.

References

1. Horimoto T, Kawaoka Y, Influenza: Lessons from past pandemics, warnings from current incidents, *Nat Rev Microbiol* **3**:591−600, 2005.
2. Triki H, Clinical virology laboratory, *Arch Inst Pasteur Tunis* **74**:51−55, 1997.
3. Wiley DC, Skehel JJ, The structure and function of the hemagglutinin membrane glycoprotein of influenza virus, *Annu Rev Biochem* **56**:365−394, 1987.
4. Chander Y, Jindal N, Stallknecht DE, Sreevatsan S, Goyal SM, Full length sequencing of all nine subtypes of the neuraminidase gene of influenza A viruses using subtype specific primer sets, *J Virol Methods* **165**:116−120, 2010.
5. Munch M, Nielsen LP, Handberg KJ, Jorgensen PH, Detection and subtyping (H5 and H7) of avian type A influenza virus by reverse transcription-PCR and PCR-ELISA, *Arch Virol* **146**:87−97, 2001.
6. Zhang Y, Lin X, Zhang F, Wu J, Tan W, Bi S, Zhou J, Shu Y, Wang Y, Hemagglutinin and neuraminidase matching patterns of two influenza A virus strains related to the 1918 and 2009 global pandemics, *Biochem Biophys Res Commun* **387**:405−408, 2009.
7. Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, Sessions WM, Xu X, Skepner E, Deyde V, Okomo-Adhiambo M, Finelli L, Bridges CB, Shaw M, Jernigan DB, Uyeki TM, Smith DJ, Klimov AI, Cox NJ, Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans, *Science* **325**:197−201, 2009.
8. Pedersen JC, 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, 2008.
9. Starick E, Romer-Oberdorfer A, Werner O, Type- and subtype-specific RT-PCR assays for avian influenza A viruses (AIV), *J Vet Med B Infect Dis Vet Public Health* **47**:295−301, 2000.
10. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res* **25**:3389−3402, 1997.
11. Lu G, Jiang L, Helikar RM, Rowley TW, Zhang L, Chen X, Moriyama EN, GenomeBlast: A web tool for small genome comparison, *BMC Bioinformatics* **7**(4):18, 2006.
12. Eddy SR, Profile hidden Markov models, *Bioinformatics* **14**:755−763, 1998.
13. Attaluri PK, Chen Z, Weerakoon AM, Lu G, Integrating decision tree and hidden Markov model (HMM) for subtype prediction of human influenza A virus, in *Cutting-Edge Research Topics on Multiple Criteria Decision Making*, Springer, 2009, pp. 52−58.
14. Attaluri PK, Zheng X, Chen Z, Lu G, Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic, in *The 6th Annual Biotechnology and Bioinformatics Symposium (BIOT-2009)*, The University of Nebraska-Lincoln, Lincoln, Nebraska, 2009.
15. Attaluri PK, Chen Z, Lu G, Applying neural networks to classify influenza virus antigenic types and hosts, in *2010 IEEE Symp Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Montreal, Canada, May 2−5, 2010.
16. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, Ostell J, Lipman D, The influenza virus resource at the National Center for Biotechnology Information, *J Virol* **82**:596−601, 2008.
17. Suzuki Y and Nei M, Origin and evolution of influenza virus hemagglutinin genes, *Mol Biol Evol* **19**:501−509, 2002.
18. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD, Multiple sequence alignment with the Clustal series of programs, *Nucleic Acids Res* **31**:3497−3500, 2003.

19. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ, Multiple sequence alignment with Clustal X, *Trends Biochem Sci* **23**:403−405, 1998.
20. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG, Clustal W and Clustal X version 2.0, *Bioinformatics* **23**:2947−2948, 2007.
21. Schuster-Bockler B and Bateman A, An introduction to hidden Markov models, *Curr Protoc Bioinformatics*, Appendix 3, p. Appendix 3A, 2007.
22. Hughey R, Krogh A, Hidden Markov models for sequence analysis: Extension and analysis of the basic method, *Comput Appl Biosci* **12**:95−107, 1996.
23. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A, The Pfam protein families database, *Nucleic Acids Res* **36**:D281−D288, 2008.
24. Wilson D, Madera M, Vogel C, Chothia C, Gough J, The SUPERFAMILY database in 2007: Families and functions, *Nucleic Acids Res* **35**:D308−D313, 2007.
25. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D, Hidden Markov models in computational biology. Applications to protein modeling, *J Mol Biol* **235**:1501−1531, 1994.
26. Barrett C, Hughey R, Karplus K, Scoring hidden Markov models, *Comput Appl Biosci* **13**:191−199, 1997.
27. Nomura Y, Significance of the ROC (receiver operating characteristics) curve in diagnostic tests, *Nippon Rinsho Jpn J Clin Med* 1402−1404, 1979.
28. Mariakerke B, MedCalc Software for Windows, Version 11.3.8 ed, 1993−2010
29. Wong SS, Yuen KY, Avian influenza virus infections in humans, *Chest* **129**:156−168, 2006.
30. Auewarakul P, Suptawiwat O, Kongchanagul A, Sangma C, Suzuki Y, Ungchusak K, Louisirirotchanakul S, Lerdsamran H, Pooruk P, Thitithanyanont A, Pittayawonganon C, Guo CT, Hiramatsu H, Jampangern W, Chunsutthiwat S, Puthavathana P, An avian influenza H5N1 virus that binds to a human-type receptor, *J Virol* **81**:9950−9955, 2007.