# Classification of Human vs. non-Human, and Subtyping of Human Influenza viral strains using Profile Hidden Markov Models

Fayroz F. Sherif, Yasser Kadah
Biomedical Engineering Department
Cairo University
Cairo, Egypt
Fayroz_Farouk@yahoo.com, Ymk@K-space.org

Mahmoud El-Hefnawi
Informatics and Systems Department
National Research Centre
Cairo, Egypt
Mahef@aucegypt.edu

*Abstract*— **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 sub-typing of human influenza viral strains virus is done based on Profile Hidden Markov Models. 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 all human HA subtypes (H1, H2, H3 and H5), and NA subtypes (N1 and N2), 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 (human vs. non-human) has accuracies varied between (55.5% and 97.5%) according to HA subtype.**

***Keywords-component; Bioinformatics; Influenza virus; Profile Hidden Markov Model.***

## I. 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 consider 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]. The first three subtypes H1, H2, H3 and recently H5, are found in human influenza viruses. The most commonly strains which infect humans during annual influenza season are (H1N1 and H3N2) [5]. 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 [6]. 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 [7] , when working with uncharacterized viruses or antibody subtypes, the library of reference reagents required for identifying antigentically distinct influenza viruses and/or antibody specificities from multiple lineages of a single hemagglutinin 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 [8]. 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 [9]. However, there are issues associate with the BLAST algorithm as described in [10]. Most importantly, the BLAST result can not reveal important mutations that may be functionally related to the structure and function of proteins.

Profile HMMs are statistical models of multiple sequence alignments [11]. 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 [12]. Some informative positions are extracted from decision trees and modeled into profiles through hidden Markov modeling, with subtype prediction accuracy of 88%. Another study has applied the feed-forward backpropagation neural network for the classification analysis of influenza virus [13]. Our study aims to Classify human vs. non-human viral strains, and sub-typing of human influenza viral strains using Profile Hidden Markov Model at protein level. With subtype prediction achieved 100% accuracy and host identification (as human or non-human) with accuracies varied between (55.5% and 97.5%) according to HA subtype.

978-1-4244-7000-6/11/$26.00 ©2011 IEEE

TABLE I. COUNT OF SEQUENCES USED FOR EACH SUBTYPE OF HA AND NA SEGMENTS.

| HA Segment | # of training sequences (Human) | # of test sequences | |
|---|---|---|---|
| | | *Human* | *Non-Human* |
| H1 | 1318 | 150 | 150 |
| H2 | 540 | 95 | 95 |
| H3 | 1426 | 220 | 220 |
| H5 | 150 | 25 | 25 |
| **NA segment** | | | |
| N1 | 600 | 90 | 90 |
| N2 | 761 | 150 | 150 |

## II. DATA AND METHODS

### A. *Data collection*

All sequences were downloaded from the National Center for Biotechnology Information NCBI's Influenza Virus Resources) [14]. And force the downloaded sequences to be non redundant and complete isolation of HA and NA segments. Part of the data is used for training and the remaining part is used for testing (Table I). We used amino acid sequences because they are known to give more reliable results than nucleotide sequences when the sequence divergence is high [15].

### B. *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) [16]. Each group of training sets found in Table I was collectively aligned using Clustal X program, which supports multiple sequence alignment for protein sequences through window graphical user interface [17, 18] and built by adding the sequences sequentially to the growing MSA produced a consensus sequence representing the highly conserved regions from the aligned sequences.

### C. *Modeling using Profile HMM*

Profile hidden Markov model (HMM) techniques are among the most powerful methods for protein homology detection scoring them above the noise level [19]. HMM profile includes more flexible information on a given set of sequences than a single sequence [20]. Therefore, database search methods using profiles is more sensitive to remote similarities than those based on pairwise alignments (e.g., regular BLAST). In particular, profile Hidden Markov models have generated good results, and are today employed by several databases such as Pfam [21] and Superfamily [22].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.

### D. *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 [11]. 'Hmmbuild' program in HMMER package v2.3.2 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 I. 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 already built and calibrated [23].

### E. *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 [24]. '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 [21].

## III. RESULTS

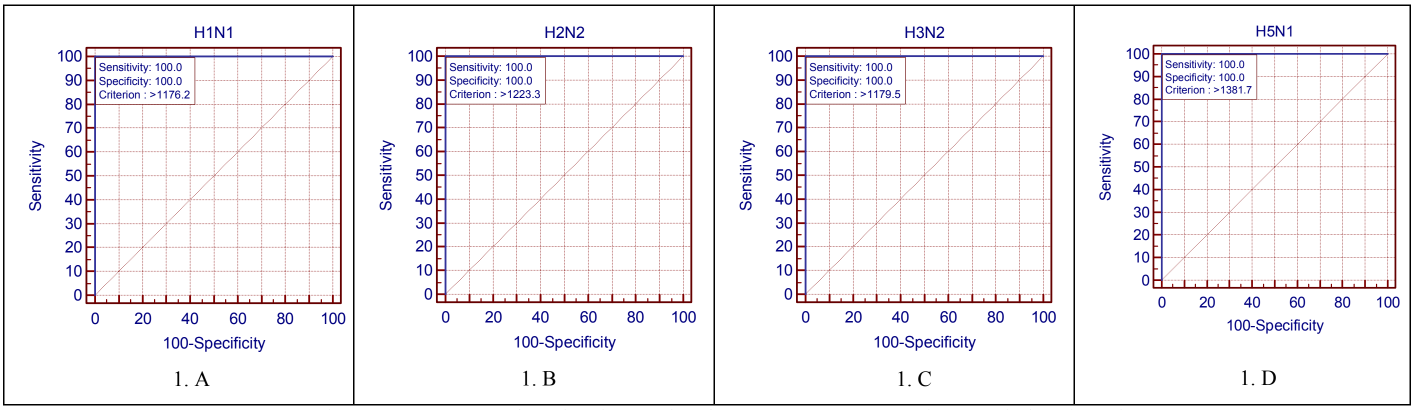
Multiple sequence alignments were done for the four HA subtypes and the two NA subtypes using ClustalX, followed by profile-HMMs models building, calibration and database generation using the HMMER suite for each group.

### A. *Subtyping classification results*

Subtyping classification was done by scoring the entire test-sets (human) (Table I), with each HA and NA HMM models, using 'HMMPfam' program in HMMER suite. 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. The test results details of subtybing classification of different viral subtypes in terms of accuracy, sensitivity and specificity were summarized in Table II.

TABLE II. SUMMARY TABLE OF SUBTYPING AND HOST CLASSIFICATION RESULTS OF INFLUENZA A VIRUS USING HMMS.

| Viral Subtype | Subtyping classification | | | HA subtype | Host classification (Human vs. Non-Human) | | |
|---|---|---|---|---|---|---|---|
| | *Accuracy* | *Sensitivity* | *Specificity* | | *Accuracy* | *Sensitivity* | *Specificity* |
| H1N1 | 100% | 100% | 100% | H1 | 94.4% | 93.7% | 95.7% |
| H2N2 | 100% | 100% | 100% | H2 | 97.5% | 93.3% | 100% |
| H3N2 | 100% | 100% | 100% | H3 | 80.8% | 86.9% | 71.1% |
| H5N1 | 100% | 100% | 100% | H5 | 55.5% | 95.2% | 43.5% |



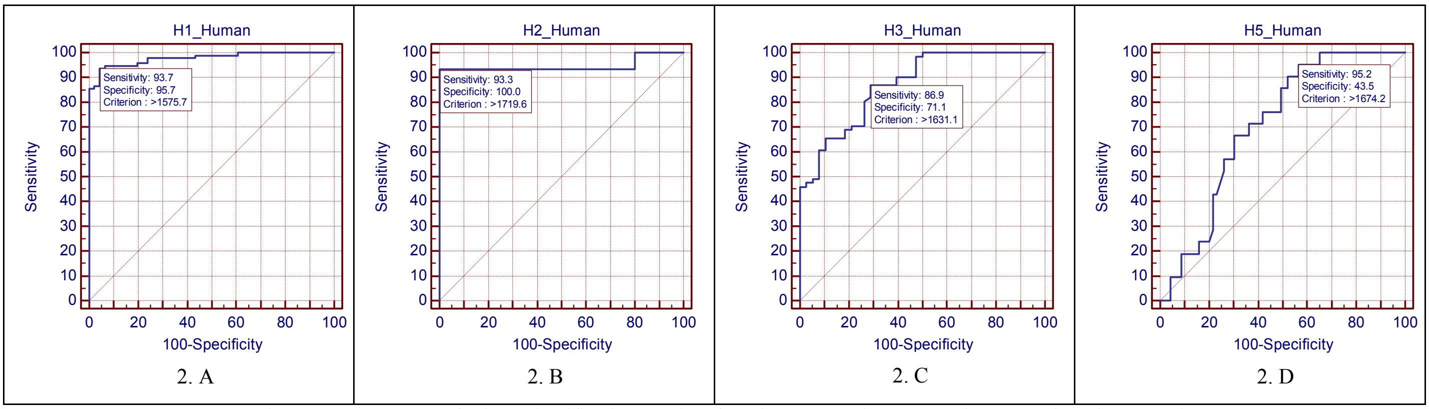
1. A

1. B

1. C

1. D

Figure1. ROC curves for subtyping results of H1N1, H2N2, H3N2 and H5N1 viral strains using HMM.



2. A

2. B

2. C

2. D

Figure2. ROC curves for host classification results of H1-human, H2-human, H3-human and H5-human using HMM.

### B. *Host classification results*

Identifying the origin of viral strains as human or non-human has been done by scoring each human HA model with its corresponding test-set (human and non-human) (Table I). 'HMMSearch' program in HMMER suite has been used for this classification with accuracies varied between 55.5% and 97.5% according to HA subtype (Table II).

### C. *Models evaluation: ROC*

ROC is a classical chart with the number of true positives (TP) versus the Number of false positives (FP). True positives are homologous pairs and false positives are non-homologous pairs with a score above a certain threshold. By varying the threshold score the curve of TP versus FP is traced out [25]. The following ROC curves drawn using MedCalc program [26]. The curves indicate the observed criterion (threshold) values that maximized both of sensitivity and specificity values have. The ROC curves in Figure 1.A, 1.B, 1.C and 1.D show the result of subtyping H1N1, H2N2, H3N2 and H5N1 viral strains respectively, using HMM search program in HMMER package. While the ROC curves in Figure 2 show the result of host identification of HA subtypes using HMMs. Figure 2.A, 2.B, 2.C and 2.D show the ROC curves of H1-human, H2-human, H3-human and H5-human models respectively. The curves indicate the observed criterion (threshold) values that maximized both of sensitivity and specificity values.

## IV. DISSCUSION

Our results confirm that profile HMM can successfully be used for classifying all human influenza A strains through identifying Hemagglutinin (HA) subtypes and Neuraminidase (NA) subtypes with 100% accuracy. All the 4 HA and 2 NA models have sensitivity of 100% and specificity of 100%. Although, differences in the criteria used in Attaluri's study and this study but 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 archived higher accuracy over Attaluri's study [12]. On the other hand, host classification of any viral sequence as human or non-human varied according to HA subtype; Among HA subtypes, there were some HAs (H1, H2, H3 and H5) 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-human model achieved a higher accuracy in host classification over H1, H3 and H5 HAs models, these results indicate that H2 viral subtypes have more genetic diversity between human and non human, comparing to the other subtypes. In contrast, H5-human model accuracy was not much higher than 55.5%, which means that, no significant differences can be detected between human and non-human H5 viruses using HMM. H1-human and H3-human models have accuracies of 94.4% and 80.8% respectively. These results seem reasonable so that cross-species infections are usually taken 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. Furthermore to extent the classification to include all influenza A Viral subtypes and host of origins.

## V. CONCLUSION

Accurate detection of influenza viral origin and subtyping can significantly improve influenza surveillance and vaccine development. In this study, host identification and subtyping of human influenza A virus is done based on hidden markov models. This study demonstrated the power of integrating the multiple sequence alignment and profile hidden Markov models approaches in classifying influenza A viral stains and their host of origin. In conclusion, our results indicate that human influenza A sequences are HA and NA subtype specific and highly sensitive against HMM models (H1-H3 and H5), (N1 and N2) and can easily predicted with 100% accuracy. While human vs. non-human classification has accuracies varied between (55.5% and 97.5%) according to HA subtype.

## REFERENCES

[1] T. Horimoto and Y. Kawaoka, "Influenza: lessons from past pandemics, warnings from current incidents," *Nat Rev Microbiol,* vol. 3, pp. 591-600, Aug 2005.

[2] H. Triki, "[Clinical virology laboratory]," *Arch Inst Pasteur Tunis,* vol. 74, pp. 51-5, Jan-Apr 1997.

[3] D. C. Wiley and J. J. Skehel, "The structure and function of the hemagglutinin membrane glycoprotein of influenza virus," *Annu Rev Biochem,* vol. 56, pp. 365-94, 1987.

[4] Y. Chander, N. Jindal, D. E. Stallknecht, S. Sreevatsan, and S. M. Goyal, "Full length sequencing of all nine subtypes of the neuraminidase gene of influenza A viruses using subtype specific primer sets," *J Virol Methods,* vol. 165, pp. 116-20, Apr.

[5] Y. Zhang, X. Lin, F. Zhang, J. Wu, S. Bi, J. Zhou, Y. Shu, and Y. Wang, "Hemagglutinin and neuraminidase matching patterns of two influenza A virus strains related to the 1918 and 2009 global pandemics," *Biochem Biophys Res Commun,* vol. 387, pp. 405-8, Sep 18 2009.

[6] R. J. Garten, C. T. Davis, C. A. Russell, B. Shu, S. Lindstrom, A. Balish, W. M. Sessions, X. Xu, and N. J. Cox, "Antigenic and genetic characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans," *Science,* vol. 325, pp. 197-201, Jul 10 2009.

[7] J. C. Pedersen, "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,* vol. 436, pp. 53-66, 2008.

[8] E. Starick, A. Romer-Oberdorfer, and O. Werner, "Type- and subtype-specific RT-PCR assays for avian influenza A viruses (AIV)," *J Vet Med B Infect Dis Vet Public Health,* vol. 47, pp. 295-301, May 2000.

[9] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res,* vol. 25, pp. 3389-402, Sep 1 1997.

[10] G. Lu, L. Jiang, R. M. Helikar, T. W. Rowley, L. Zhang, X. Chen, and E. N. Moriyama, "GenomeBlast: a web tool for small genome comparison," *BMC Bioinformatics,* vol. 7 Suppl 4, p. S18, 2006.

[11] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics,* vol. 14, pp. 755-63, 1998.

[12] Y. Shi, S. Wang, Y. Peng, J. Li, Y. Zeng, P. K. Attaluri, Z. Chen, A. M. Weerakoon, and G. Lu, "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*. vol. 35: Springer Berlin Heidelberg, 2009, pp. 52-58.

[13] P. K. Attaluri, Z. Chen, and G. Lu;, "Applying Neural Networks to Classify Influenza Virus Antigenic Types and Hosts," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2010 IEEE Symposium on* Montreal, QC 2010.

[14] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *J Virol,* vol. 82, pp. 596-601, Jan 2008.

[15] Y. Suzuki and M. Nei, "Origin and evolution of influenza virus hemagglutinin genes," *Mol Biol Evol,* vol. 19, pp. 501-9, Apr 2002.

[16] R. Chenna, H. Sugawara, T. Koike, R. Lopez, T. J. Gibson, D. G. Higgins, and J. D. Thompson, "Multiple sequence alignment with the Clustal series of programs," *Nucleic Acids Res,* vol. 31, pp. 3497-500, Jul 1 2003.

[17] F. Jeanmougin, J. D. Thompson, M. Gouy, D. G. Higgins, and T. J. Gibson, "Multiple sequence alignment with Clustal X," *Trends Biochem Sci,* vol. 23, pp. 403-5, Oct 1998.

[18] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics,* vol. 23, pp. 2947-8, Nov 1 2007.

[19] B. Schuster-Bockler and A. Bateman, "An introduction to hidden Markov models," *Curr Protoc Bioinformatics,* vol. Appendix 3, p. Appendix 3A, Jun 2007.

[20] R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: extension and analysis of the basic method," *Comput Appl Biosci,* vol. 12, pp. 95-107, Apr 1996.

[21] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Res,* vol. 36, pp. D281-8, Jan 2008.

[22] D. Wilson, M. Madera, C. Vogel, C. Chothia, and J. Gough, "The SUPERFAMILY database in 2007: families and functions," *Nucleic Acids Res,* vol. 35, pp. D308-13, Jan 2007.

[23] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology. Applications to protein modeling," *J Mol Biol,* vol. 235, pp. 1501-31, Feb 4 1994.

[24] C. Barrett, R. Hughey, and K. Karplus, "Scoring hidden Markov models," *Comput Appl Biosci,* vol. 13, pp. 191-9, Apr 1997.

[25] Y. Nomura, "[Significance of the ROC (receiver operating characteristics) curve in diagnostic tests]," *Nippon Rinsho,* vol. Suppl, pp. 1402-4, Jun 29 1979.

[26] B. Mariakerke, "MedCalc Software for Windows," Version 11.3.8 ed, 1993-2010.