Applying machine learning techniques to classify H1N1 viral strains occurring in 2009 flu pandemic

Pavan K. Attaluri
Department of Computer Science
University of Nebraska at Omaha
Omaha, NE 68182, USA
pattaluri@mail.unomaha.edu

Ximeng Zheng
Department of Computer Science
University of Nebraska at Omaha
Omaha, NE 68182, USA
xzheng@unomaha.edu

Zhengxin Chen
Department of Computer Science
University of Nebraska at Omaha
Omaha, NE 68182, USA
zchen@unomaha.edu

Guoqing Lu
Departments of Biology,
University of Nebraska at Omaha, Omaha, NE 68182, USA
glu3@mail.unomaha.edu

***Abstract*—A phase 6 alert has been declared by the World Health Organization (WHO) in response to the ongoing global spread of the influenza H1N1 virus in humans. Genetic sequence analysis suggests that this pandemic strain evolves from reassortment of swine viruses. The objective of this research is to conduct a series of bioinformatics analyses to characterize currently circulating pandemic influenza viral strains and identify their evolutionary origin. Three groups of sequences (i.e., human, swine, and the latest pandemic human/swine) were used for phylogenetic analysis, decision tree analysis and support vector machine (SVM) analysis. Our results strongly support the finding that the latest pandemic viral strain is of swine origin. To facilitate early detection of human and swine H1N1 viral strains, we have developed a web tool based upon the results obtained in this study and used Hidden Markov Model (HMM) for accurate prediction of influenza H1N1 origin.**

***Index Terms*—H1N1, influenza A virus, machine learning, swine flu**

I. INTRODUCTION

INFLUENZA is one of the most important emerging and reemerging infectious diseases, causing high morbidity and mortality in communities (epidemic) and worldwide (pandemic) [1]. The influenza virus is an RNA virus and comprises three types: A, B, and C based upon their protein consumption. Influenza A is the most virulent human pathogen among the three types and is believed to be responsible for the global outbreaks of 1918, 1957, 1968 and 2009 [2]. Influenza A is subdivided into subtypes based on two surface proteins, HA and NA. Mutations on these proteins may result in different influenza subtypes. So far, there are 16 H and 9 N serotypes found in influenza virus. This genetic drift process often results in different strains of H1N1 and H3N2 circulating in humans during annual influenza seasons. Another process called genetic shift undergoes infrequent and sudden changes of genome segments from different viral strains, which is speculated to be the major cause for influenza pandemics [3].

An influenza pandemic occurs when a new influenza virus appears against which the human population has no immunity. In the past, influenza pandemics have resulted in increased death, disease and great social disruption. Influenza A virus caused three major global epidemics during the 20th century: the Spanish Flu in 1918, Asian Flu in 1957 and Hong Kong Flu in 1968. These pandemics were caused by strains of influenza A virus that had undergone major genetic changes and for which the population did not possess significant immunity [4]. The 2009 flu pandemic is a global outbreak of a new influenza A virus H1N1 strain, identified in April 2009 and commonly referred to as *Swine Flu* [5]. Analysis suggests that the H1N1 strain responsible for the current outbreak first evolve around September 2008 and circulate in the human population for several months before the first cases were identified [6]. Fig. 1 depicts genetic origins of the 2009 swine flu virus.

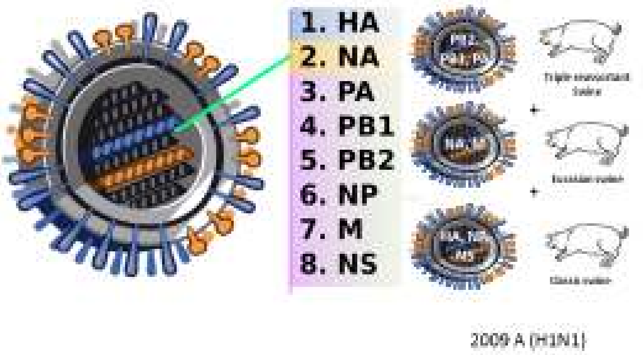


Fig.1. Genetic structure of influenza A virus (left panel) and the origins of the 2009 swine flu virus (right panel).

In order to effectively deal with the problem of identifying origin of pandemic swine flu viral strains, we have explored approaches using machine learning techniques. Machine learning is a subfield of artificial intelligence and is concerned with the development of algorithms that allow computers to learn, and has found many applications in bioinformatics [7]. Machine-learning approaches are suitable for datasets containing large amounts of data and presence of noisy patterns. The idea behind these techniques is to learn the theory automatically from the data, through a process of inference, model fitting or learning from examples. A number of machine learning approaches are applied to identify relationships or associations in biological data, to group similar genetic elements, to analyze and predict diseases. Current research domains where machine learning techniques applied are multiple sequence alignment, structure and function prediction, molecular clustering and classification, and expression analysis.

The classical ways of determining the subtype of influenza virus for HA and NA segments are hemagglutination-inhibition (HI) assay and neuraminidase-inhibition (NI) assay which are capable of distinguishing antigenic differences between influenza even of the same subtype. However, as noted in [8], 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. Using machine learning methods to predict the lineage of virus is much cheaper and faster, yet usually can still yield high accuracy.

In our current study, we have applied two machine learning techniques (decision trees and support vector machines) to identify the origin of latest pandemic outbreak strains. Furthermore, based on Hidden Markov Model (HMM), we have developed a Web system for the prediction of influenza A virus hosts using the informative positions found through decision tree method. Research on the evolution of latest swine flu outbreak will improve the design of vaccines and diagnostic tools.

II. DATA AND METHODS

A. Data

Three groups of influenza A H1N1 sequences were determined, including Human (found only in human), Swine (found only in swine), and Human_Swine (i.e., the latest pandemic influenza viral strains). Sequences were downloaded from the Global Initiative on Sharing All Influenza Data (GISAID) and the National Center for Biotechnology Information (NCBI) [9]-[10]. Sequence comparison showed that the sequence dataset from GISAID is more complete compared with that of NCBI. We therefore used sequences from GISAID in the analysis. When comparing sequences from different groups, we found that a number of sequences appeared in the human group and in the human swine group as well. These sequences were excluded for further analysis.

Table I shows the count of sequences used for each segment and host. We restricted to a maximum of 150 sequences of each host and of each segment. Part of the data is used for training and the remaining part is used for testing. The sequence data from segments HA, NA, PA, NS, PB1, PB2, M, and NP are considered for SVM and decision tree analysis. We used nucleotide sequences for decision tree analysis, and protein sequences for support vector machine analysis.

TABLE I
Influenza H1N1 sequences used in the analysis

Segment	Human	Swine	Human_Swine
HA	150	150	150
M	150	147	150
NA	150	148	150
NP	150	150	137
NS	150	147	111
PA	150	146	85
PB1	150	145	93
PB2	150	146	98

B. Sequence alignment

We used MUSCLE (http://www.drive5.com/muscle/) for multiple sequence alignment. MUSCLE stands for multiple sequence comparison by log-expectation, and is one of the most popular multiple alignment software for protein and nucleotide sequences [11]. MUSCLE can achieve both better average accuracy and better speed compared with several other multiple alignment tools such as CLUSTALW [12]-[13] or T-Coffee [14], by choosing maximum number of iterations and diagonal optimization.

In this study, all H1N1 nucleotide sequences from a specific segment were aligned collectively and then divided into three host groups for training the decision trees. The parameters –*maxiters* and *–maxmb* in MUSCLE were set to 2 and 250 MB, respectively. The first two iterations of the algorithm were performed to compromise between speed and accuracy. As for the user submitted sequences, only one iteration is set for prediction.

C. Phylogenetic tree analysis

Phylogenetic analysis is important to understand the evolution of species and of gene and protein families. Phylogenetic trees define the functional subfamilies within protein or gene families with multiple functions. Neighbor-Joining is a method for reconstructing phylogenies from a set of distances between each pair of sequences by successive clustering [15]. Neighbor-Joining can reconstruct trees with additive edge lengths without making the assumption that the divergence of the sequences occurs at the same constant rate at all points in the tree. The trees are generated using CLUTALW2.0 from EBI.

D. Machine learning techniques

The phylogenetic trees classify the sequences and describe the evolutionary relationships. However, inferring phylogenetic trees is difficult and these trees do not always show a clear picture of different clusters. In order to effectively achieve the goal of identifying evolutionary origin of pandemic influenza viral strains, we employed two machine learning approaches for classification analysis of the influenza A virus hosts. Machine learning techniques such as decision tree are able to find the subtle differences in the classifications associated with the real data. The information of critical sites with sequence data is useful for host prediction of influenza A virus. Specifically, the decision tree approach classifies the sequences of different groups based on the nucleotide information at specific positions, whereas the support vector machine (SVM) approach classifies sequences based on the frequency of amino acids appearing in various sequences. This complementary strategy (one based upon DNA sequences and the other on protein sequences) provides a comprehensive description of influenza A viral classification. As to be described in a later section, results obtained from these two approaches are consistent to each other. Based on the classification results, a Web prediction tool was developed, where Hidden Markov Model (HMM) is used for modeling the informative positions generated from the decision trees. In the rest of this section we provide a brief review of these techniques.

E. *Decision tree method*

A decision tree is a simple but a powerful machine learning algorithm that has been successfully used for classification problems. The decision tree technique employs a supervised approach for classification, where the leaves on the tree represent classifications and the branches represent conjunctions of features that lead to classification. A series of decisions were made when classifying an instance from root to leaf nodes and the instance was classified to the one associated with the leaf node at the end of the traversal. Each internal node is a decision node and a value of given instance is compared to the decision function to decide which branch to follow. A decision tree is built using a training data set so as to reduce the average depth of each path from root to leaf node. Decision tree classification as a standard machine learning technique has been used for a wide range of applications in bioinformatics [16]-[17]. The software package Weka (Waikato Environment for Knowledge Analysis) (http://www.cs.waikato.ac.nz/ml/weka/), consisting of a number of machine learning algorithms, was used for decision tree analysis of aligned sequences [18]-[19]. The decision trees are generated using the C4.5 algorithm; Weka has its own version of C4.5 known as J48.

We used aligned sequences of each host to train decision tree classifiers in the J48 program of Weka, which allows the most informative nucleotide positions to be found. In each of the iteration steps, one or more critical positions, in which different subtypes can be most likely identified, were determined. These positions were collectively utilized to build HMMs for further host prediction. We applied the cross validation technique for testing. The three groups Human, Swine and Human_Swine strains are trained and then classified.

An example of decision tree is shown in Fig. 2, where the positions can be used to classify hosts, Human_Swine and Human. The position C604 is an input variable. The leaf nodes represent target variables Human_Swine and Human.

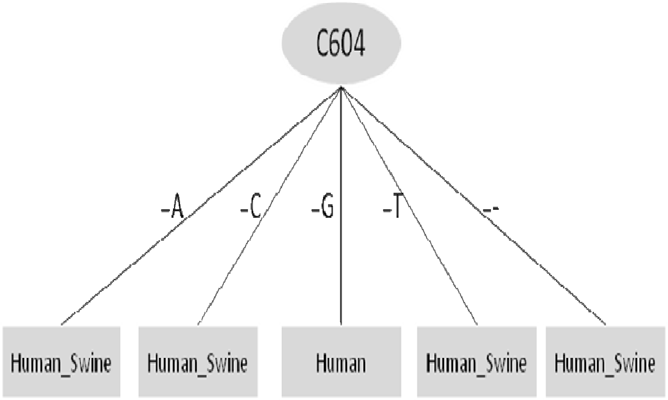


Fig. 2. Decision tree classifying Human and Human_Swine. Nucleotides: A, C, G and T; Gap: -

F. *Support Vector Machine method*

Support Vector Machine (SVM) is an alignment-free method which uses vectors to classify objects. One of the advantages of SVM is that it does not depend on multiple alignment, thus it can avoid errors, if any, in multiple alignment files. Another advantage of SVM is that it can classify sequences with low similarity and/or even with very short lengths.

In order to use SVM for classification, first we need to compute the frequency of each amino acid or a certain length of amino acids group. For example, sequence "GPPAV" can be treated in one letter a time (1-mer): "G" with frequency of 0.2 and "P" with frequency of 0.4, etc. Alternatively, it can be treated as two letters at a time (2-mer): "GP" with frequency of 0.25 and "PP" with frequency of 0.25, etc. These amino acid frequencies were treated as vectors, and the distribution patterns of k-mer (up to 3) amino acids were used for classification analysis.

Data from each segment of different origin were divided into two sets with approximately the same number of sequences: one for training and the other for testing. The sequences used for training are selected randomly by our computer program. The software SVM-light was used for classification analysis. Two experiments were conducted: one is trained and tested with Human and Swine sequences, and the other is trained with Human and Swine sequences, but tested with the Human_Swine strains.

G. *Modeling using HMM*

The Hidden Markov Model (HMM) is used for modeling the informative positions generated from the decision trees. An HMM is a statistical model representing sequences from a gene family. HMMs have a formal probabilistic basis, which is their advantage over other methods [20]. An HMM profile includes more flexible information on a given set of sequences than a single sequence. Therefore, database search methods using profiles is more sensitive to remote similarities than those based on pairwise alignments (e.g., regular BLAST). HMMER, a package that uses Hidden Markov models (HMMs) for sequence database searching, was used to build

HMM models based upon the most informative sites determined by the decision tree method [21]-[22].

To take advantage of most informative sites found by Decision Tree, we built different HMM profiles for the prediction of each influenza A virus host. HMM profiles are statistical representation of a specific group of informative sites. They were built using the program *hmmbuild* in HMMER. These profiles are used in the Web prediction program to determine the host of a viral strain through sequence comparison.

H. *Summary of experiment design*

Our experiment design is summarized in Fig. 3. In brief, the comparison analysis was conducted with three methods, decision tree, phylogenetic trees and support vector machine. The consensus of the three methods is that the Human_Swine sequences are more closely related to swine origin. Decision tree and HMM were used for web prediction of influenza viral origin.

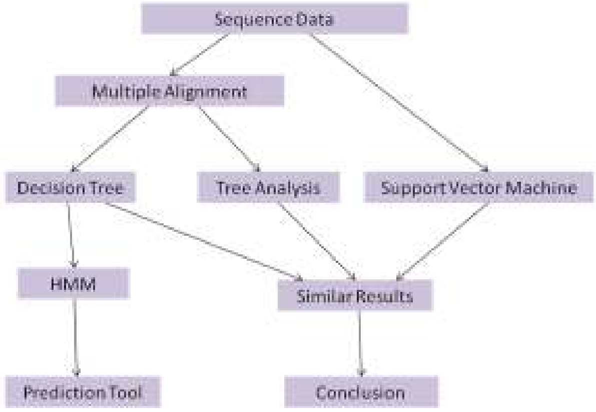


Fig.3. Experiment Design

III. Results & Discussion

A. *Phylogenetic analysis of HA and NA segments*

The tree diagrams (Figures 4 and 5 in the next page) are made by using randomly selected 10 sequences from each group of Human strain, Swine strain and Human_Swine strains. Here, HA and NA segments are chosen as examples.

The sequences in the top right corner of the diagrams (marked by a box) are from Human host. The sequences in the bottom left of the diagram (marked by another box) are from Human_Swine. The rest of the sequences are from Swine host. As presented in the tree diagrams, they indicate that the origin of 2009 H1N1 human influenza A is from swine rather than traditional human H1N1. In order to have a more comprehensive and confident result, alternative approaches using machine learning techniques are needed, as to be described below.

B. *SVM analysis*

The results of the classification experiment showed above 95% accuracy at prediction and the test results details are summarized in TABLE II. Accuracy is defined as the number of correctly classified sequences divided by the total number of testing sequences. The results shown are the classification results when 3-mer is used and its frequency is computed as a vector input for SVM.

TABLE II
SVM Results

Segment	Accuracy	Sensitivity	Specificity
HA	98.00%	97.37%	98.67%
M	97.97%	98.65%	97.33%
NA	95.30%	91.46%	100%
NP	98.67%	98.67%	98.67%
NS	98.65%	97.40%	100%
PA	97.97%	96.15%	100%
PB1	98.64%	100%	97.33%
PB2	99.32%	98.68%	100%

These results have shown that human and swine groups are well distinguishable, which indicate there are significant differences between them. The outcomes of the experiment that trying to predict the lineage of Human_Swine strain are presented in TABLE III, where 3-mer is used for all segments except PA. Since the results of 2-mer and 3-mer for PA segment differ significantly, the average of 2-mer and 3-mer results are presented instead. The percentages in the Table III indicate the percentages of Human_Swine sequences in the testing data sets that are classified as Swine strain. In other words, 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. However, the result for PA segment is inconclusive; it is possibly equally similar to both human strain and swine strain.

TABLE III
Results of predicting Human_Swine sequences against Swine strains

Segment	Accuracy
HA	100.00%
M	100.00%
NA	100.00%
NP	100.00%
NS	100.00%
PA	50.00%*
PB1	100.00%
PB2	100.00%

* classified as Human strains by 2-mer SVM whereas classified as Swine strains by 3-mer SVM

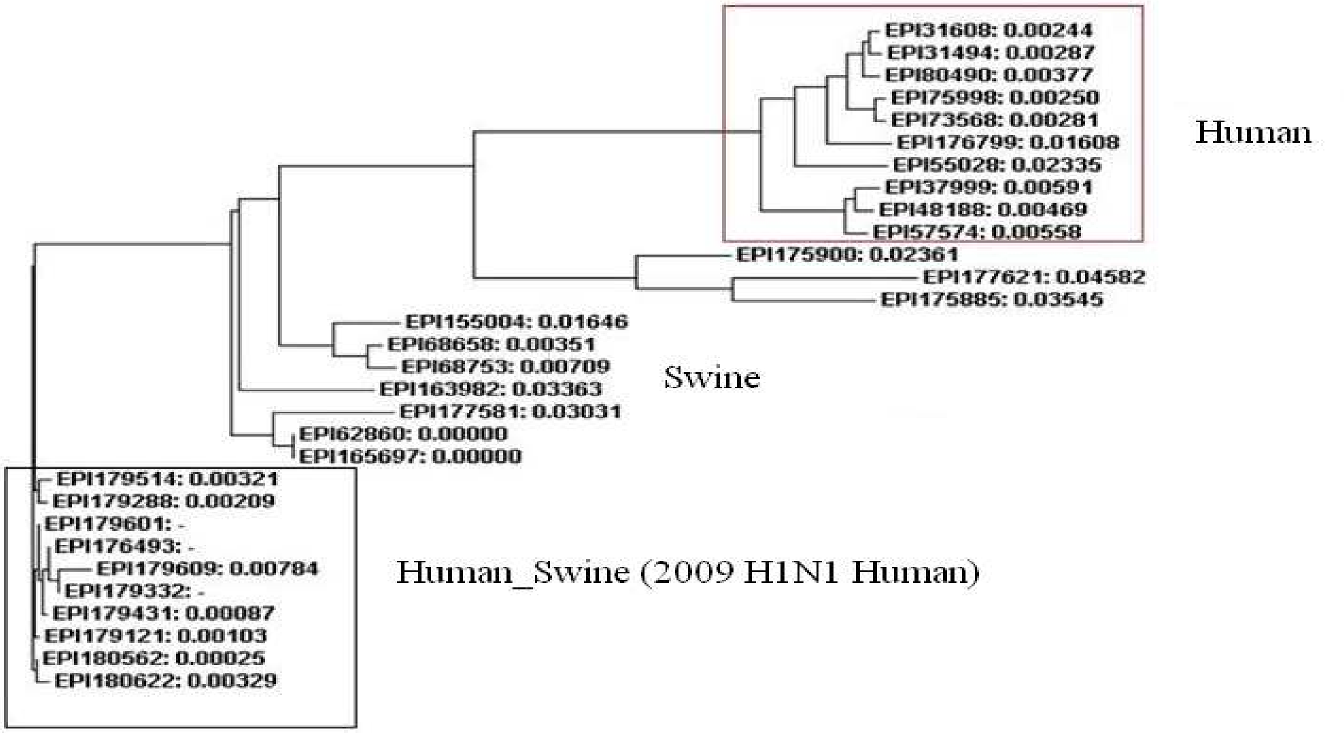


Fig. 4. Neighbor-Joining tree of influenza A viral strains based upon HA sequences

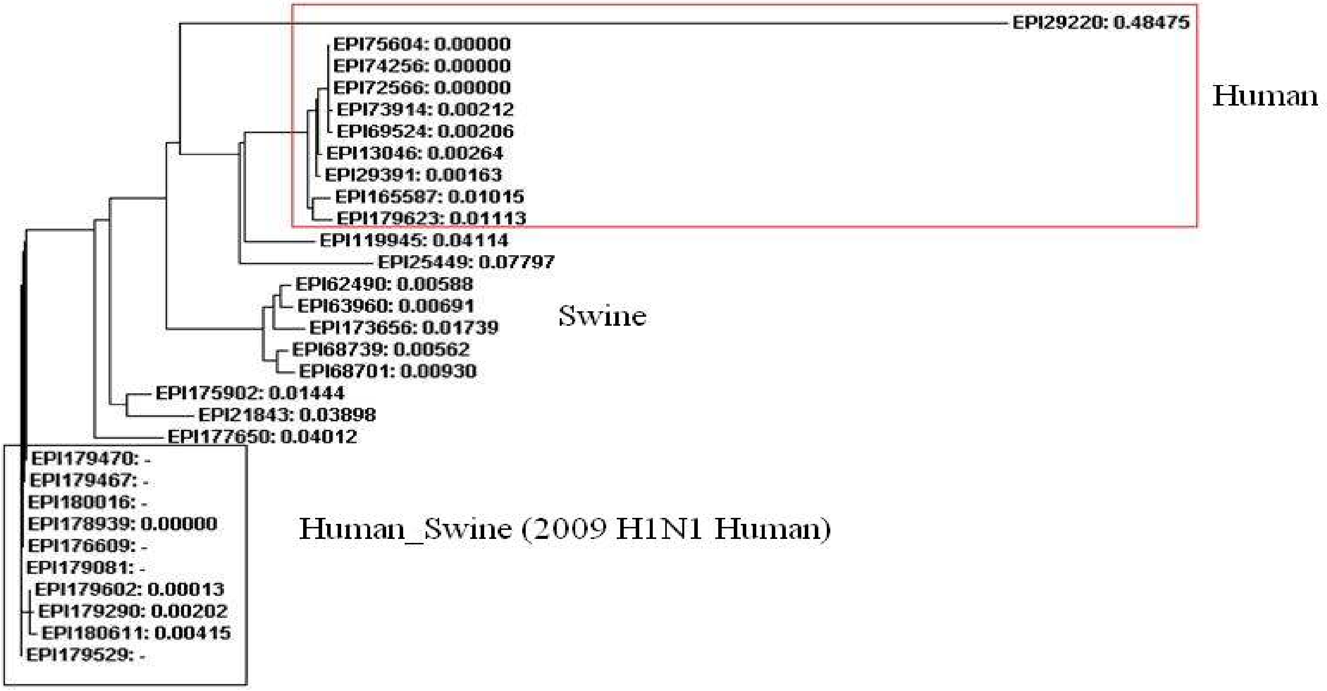


Fig. 5. Neighbor-Joining tree of influenza A viral strains based upon NA sequences

C. *Decision tree analysis*

We have performed 25 iterations for HA, NA, M, NP, PA, NS, PB1, PB2 segments. During each iteration, a decision tree is generated with one or more informative positions which classify the three sets of data (i.e., Human, Swine and Human_Swine). Two thirds of the data is used for training and the remaining one third is used for testing. We applied the cross validation technique which automatically divides the input data into two sets, training and testing. TABLE V summarizes the average accuracy of classification between Human, Swine and Human_Swine strains for each segment of all iterations. A total of 67 nucleotide positions for the HA segment and 64 nucleotide positions for the NA segment have been identified as informative by the decision tree analysis. (We referred the positions determined by decision tree for classifying different groups of data as informative.) The informative positions of HA and NA segments are summarized in TABLE IV. The sequence data from only these positions are collected together to form HMM profiles. All the three groups are classified at a time using decision tree, unlike the case of SVM where every two groups are classified at once.

TABLE IV
Summary of HA & NA segments informative positions for each iteration

Iteration	HA Positions	NA Positions
1	27, 574, 671	47, 681, 728
2	28, 543, 408	48, 478, 421
3	29, 604, 625	124, 289, 837
4	331, 633, 634	281, 439, 793
5	404, 443, 500	267, 354, 721
6	597, 841, 1015	58, 293, 842
7	412, 725, 1036	654, 841, 994
8	590, 881, 944	176, 351, 1014
9	423, 623, 631	78, 112
10	37, 287, 448	731, 927, 1085
11	717, 857	493, 919
12	361, 580, 624	208, 643
13	867, 1111	585, 774, 1183
14	171, 873, 1126	89, 818, 1076
15	421, 556, 795	434, 1008
16	800, 846, 858	717, 982, 1126
17	32, 918, 1058	990, 1073
18	321, 446	256, 542,1045
19	162, 721	92, 548, 978
20	221, 230	144, 662
21	187, 1156	593, 834
22	254, 553	784, 929
23	282, 440	789, 822
24	726, 730, 771	419, 1031
25	316, 364, 473	680, 936

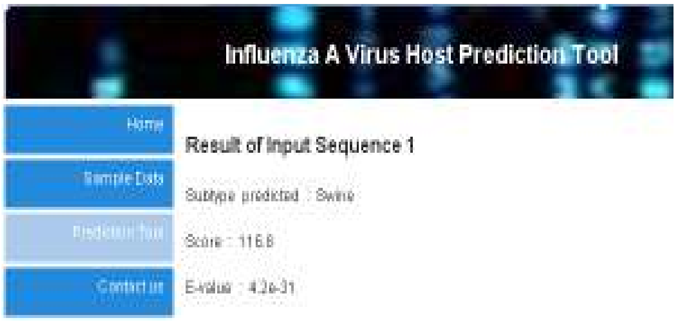
TABLE V
Decision Tree Results

Segment	Accuracy	Sensitivity	Specificity
HA	97.55%	96.18%	97.26%
M	96.38%	97.78%	97.08%
NA	98.12%	99.35%	98.24%
NP	98.28%	97.38%	98.16%
NS	97.22%	98.5%	98.78%
PA	98.13%	99.34%	99.12%
PB1	97.92%	96.88%	97.73%
PB2	96.86%	95.52%	97.84%

D. *A web prediction tool*

In order to assist the task of detecting swine flu origin, we have developed a prediction tool for human influenza A virus hosts, which is now available at http://glee.ist.unomaha.edu/~pattaluri/swine/. The prediction tool has been developed using LAMP technology. We have provided sample data and a simple tutorial on how to use the tool for host prediction. The tool allows users to submit sequences, choose the input segment, and select options to view the result. The given sequence is aligned and the informative positions are extracted to compare with the HMM profiles. The result of prediction will be displayed based on matching scores.

Here we show how to use the web prediction system and what the result page looks like by analyzing a real sequence. We downloaded a Human_swine (outbreak) sequence from NCBI, copied and pasted the sequence in text area, selected nucleotide type, checked selected options, and clicked Submit button. The prediction result is shown as in Fig. 6, where the sequence is predicted correctly as the Swine type, with a score 116.8 and an E value 4.2×e^{-31}.



Result of Input Sequence 1

Subtype predicted : Swine

Score : 116.8

E-value : 4.2e-31

Fig. 6. Influenza A virus host prediction system

The classification results from our experiments indicate that data from Human and Swine origin are easily distinguishable as they have considerable positions for predicting the host. Also the Human_Swine sequences in the testing data sets are classified as Swine strain. This suggests that the outbreak sequences be more closely related to Swine rather than Human viral strains.

IV. Conclusion

Accurate detection of influenza viral origin can significantly improve influenza surveillance and vaccine development. In this research, we applied machine learning techniques to identify the evolutionary origin of the latest human pandemic influenza H1N1 viral strains. Phylogenetic analysis of randomly selected sequences revealed significant differences between human and swine influenza A H1N1 viral strains. Both Support Vector Machine and decision tree methods agreed each other on that the viral strains causing the latest human pandemic are of swine origin. Along with our previous findings [23], this study demonstrated the power of integrating the decision tree and hidden Markov model approaches in classifying influenza A viral subtypes and hosts.

Acknowledgment

This publication was made possible by NIH grant numbers P20 RR016469 from the INBRE Program of the National Center for Research Resources and R01 LM009985-01A1. We are grateful to Dr Ruben Donis of the Centers for Disease Control and Prevention (CDC), US Department of Health and Human Services for constructive discussion. The authors also acknowledge the UCRCA, the University of Nebraska at Omaha, for continuous funding support.

References

[1] G. Lu, K. Buyyani, N. Goty, R. Donis and Z. Chen, “Influenza A Virus Informatics: Genotype-Centered Database and Genotype Annotation,” *Proc. IEEE International Multi-Symposiums on Computer and Computational Sciences (IMSCCS’07)*, 2007, pp. 76-83.

[2] G. Neumann, T.Noda, and Y. Kawaoka. “Emergence and pandemic potential of swine-origin H1N1 influenza virus,” *Nature*, 2009, 459(7249), pp. 931-9.

[3] R. Moellering, Jr. “Avian influenza: the next pandemic?,” *Clinical Microbiology Newsletter*, 28(13), 2006, pp. 97-101

[4] E. D. Kilbourne, “Influenza pandemics of the 20th century,” *Emerg Infect Dis* [serial on the Internet], 2006 Jan. Available from http://www.cdc.gov/ncidod/EID/vol12no01/05-1254.htm

[5] World Health Organization, “Swine Influenza,” 27 April 2009, http://www.who.int/csr/don/2009_04_27/en/index.html

[6] Garten et al, “Antigenic and genetic characterstics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans”, PMID: 19465683.

[7] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez and V. Robles, “Machine learning in bioinformatics,” *Briefings in Bioinformatics,* 7(1), 2006, pp. 86-112.

[8] J. C. Pederson. “Test for Avian Influenza Virus Subtype Identification and the Detection and Quantitation of Serum Antibodies to the Avian Influenza Virus,” *Avian Influenza Virus.* 2008, 436, pp. 53-66.

[9] P. Bogner, I. Capua, D. J. Lipman, and N. J. Cox. “A global initiative on sharing avian flu data” *Nature* 442, 981 (31 August 2006)

[10] Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, and T. Tatusova, “FLAN: a web server for influenza virus genome annotation,” *Nucleic Acids Research,* 35, 2007, W280-W284.

[11] R. C. Edgar, “MUSCLE: multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Research,* 32(5), 2004, pp. 1792-7

[12] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson and D.G. Higgins**, “**Clustal W and Clustal X version 2.0,” *Bioinformatics* 23(21): pp. 2947-8 (2007)

[13] J. D. Thompson. D.G. Higgins, and T. J. Gibson, “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice,” *Nucleic Acids Research,* 22, 1994. pp. 4673-4680.

[14] C. Notredame, D.G. Higgins, and J. Heringa, “T-Coffee: A novel method for fast and accurate multiple sequence alignment,” *J Mol Biol* 302(1), 2008, pp. 205-17 (2000)

[15] N. Saitou, and M. Nei, “The neighbor-joining method; a new method for reconstructing phylogenetic trees,” *Mol. Biol. Evol.* 1987, 4(4):406-25.

[16] J. E. Gewehr, M. Szugat, and R. Zimmer, “BioWeka--extending the Weka framework for bioinformatics,” *Bioinformatics*, 23(5),pp. 651-3

[17] F. Firouzi, M. Rashidi, S. Hashemi, M. Kangavari, A. Bahari, N.E. Daryani, M. M. Emam, N. Naderi, H. M., Shalmani, A. Farnood, and M. Zali, “A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software,” *Eur J Gastroenterol Hepatol,* 19(12), 2007, pp. 1075-81

[18] A. K. Sigurdardottir, H. Jonsdottir, and R. Benediktsson, “Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis,” *Patient Educ Couns,* 67(1-2), 2007, pp. 21-31

[19] E. Frank, M.Hall , L. Trigg , G. Holmes’ and I. H. Witten, “Data mining in bioinformatics using Weka,” *Bioinformatics,* 20(15), 2004, pp. 2479-81

[20] M. Gribskov, A.D. McLachlan, and D. Eisenberg, “Profile analysis: detection of distantly related proteins,” *Proc Natl Acad Sci USA,* 84(13), 1987, pp.4355-8

[21] T. Friedrich, B. Pils, T. Dandekar, J. Schultz and T. Müller, “Modelling interaction sites in protein domains with interaction profile hidden Markov models,” *Bioinformatics,* 22(23), 2006, pp. 2851-7

[22] S. R. Eddy, ”Profile hidden Markov models,” *Bioinformatics,* 14(9), 1998, pp.755-63

[23] 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,” *MCDM Workshop on Mining Text, Semi-structured, Web or Multimedia Data*, 2009.