28th NATIONAL RADIO SCIENCE CONFERENCE
(NRSC 2011)
April 26-28, 2011, National Telecommunication Institute, Egypt

C33

Genomic signatures and associative classification of the Hemagglutinin protein for Human versus Avian versus Swine Influenza A viruses

Fayroz F. Sherif [#1], *Mahmoud El Hefnawi* [*2], *Yasser Kadah* [#3]
[#1,3] *Biomedical Engineering Department, Cairo University*
[#2] *Informatics and Systems Department, National Research Centre*

Abstract

Global outbreaks of human influenza arise from influenza A viruses with novel Hemagglutinin (HA) molecules to which humans have no immunity. So understanding of the origin and evolution of HA genes is of particular importance. Here, genomic signatures of the HA protein in different hosts was identified and associative classification for host-typing was conducted. We therefore conducted multiple-sequence alignment and detecting the most statistically significant differences between human, avian and swine group of sequences using VESPA, then applying class associative rule mining to identify amino acid–conserving positions that are specific to host species, called signatures. We applied strict thresholds to select only markers which are highly preserved in each influenza virus host isolates over time. Also, the two Sample sequence logo server was used to identify and confirm significant variations between the hosts. Host-specific signatures have created from scanning 1500 sequences of HA from human, swine and avian influenza A viruses. A total of 9, 31, 11, 6, 22, and 31 most informative positions of 560 amino acid residues yielded significant differences between Avian vs. Human, Human vs. Avian, Human vs. Swine, Swine vs. Human, Avian vs. Swine, and Swine vs. Avian respectively. Positions 438K, 458N and 286V were associated with avian, human and swine respectively, with support and confidence of (90.7% and 79.5%), (82.8% and 92.9%) and (51.4% and 98%) respectively. Host-specific class association rules aid in the prediction of prognostic biomarkers and improve the accuracy of prognosis.

***Keywords**:* Influenza, signatures and Class association rules.

I. INTRODUCTION

Influenza A viruses belong to the Orthomyxoviridae family of negative sense, single-stranded, segmented RNA viruses. Influenza A viral genome consists of 8 segments, including: HA (hemagglutinin), NA (neuraminidase), NP (nucleoprotein), M (two matrix proteins, M1 and M2), NS (two distinct non-structural proteins, NS1 and NEP), PA (RNA polymerase), PB1 (RNA polymerase and PB1-F2 protein), and PB2 (RNA polymerase) [1, 2]. The hemagglutinin (HA) of influenza A viruses is a major surface glycoprotein that is responsible for attachment of the virus to the cell surface of oligosaccharide receptors. All known subtypes of influenza A viruses are found among wild avian species that serve as primary reservoirs for these agents [2].

In general, an influenza virus infects only a single species; however, whole viruses may occasionally be transmitted from one species to another, and genetic reassortment between viruses from two different hosts can produce a new virus capable of infecting a third host. Avian influenza viruses are not readily introduced into humans [3], possibly because humans do not possess the a(2,3)-sialyllactose (NeuAc-2,3Gal) receptors required for attachment of the viruses to epithelial cells. However, individual viral genes can be transmitted between humans and avian species, as demonstrated by avian human reassortant viruses that caused the 1957 and 1968 influenza pandemics [4, 5]. This finding suggested that an intermediate host may be needed for genetic reassortment of human and avian viruses. Pigs are considered a logical candidate for this role because they can be infected by either avian or human viruses [6, 7] and because they possess both NeuAc-2,3Gal and NeuAc-2,6Gal receptors. In addition, there is good evidence that pigs are more frequently involved in interspecies transmission of influenza A viruses than are other animals [6, 8, 9].

Such studies indicate host range restriction of influenza viruses. The viruses that caused the 1957 and 1968 influenza pandemics were reassortant viruses of human and avian influenza viruses [5]. Avian influenza virus genes were somehow introduced into the human populations, breaking through the host range restriction. To elucidate the mechanism by which pandemic influenza virus strains are generated, we must first understand the molecular basis of host range restriction of influenza virus and how such restriction is breached. The amino acids that make up the receptor-binding site (RBS) are highly conserved, even among the HAs of different subtypes of avian influenza virus; however those of human viruses display distinct variability. In particular, the residues at positions 138, 190, 194, 225, 226, and 228 are highly conserved in the avian RBS, whereas human HAs harbor substitutions at these positions [10] .

In H2 and H3 influenza virus strains, residues at positions 226 and 228 in the HA correlate with the preferential recognition of the SA-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 SAa2,6Gal, whereas those with Gln-226 and Gly-228 (avian and equine viruses) recognize SAa2,3Gal [11].

Although the HA genes play a critical role in host specific infection, but many research efforts have focused specifically on the persistent markers and host specificity markers found only on the more heavily conserved internal proteins. [12] Found sixteen positions in the influenza genome associated with human host specificity. The markers were found on the non-structural protein 1 (NS1), nonstructural protein 2 (NS2), matrix protein 1 (MP1), nucleoprotein (NP), acidic protein (PA), and the basic polymerase 2 (PB2) protein. Recently, large-scale sequence analyses revealed ‘signature’ amino acids at specific positions in viral proteins that distinguish human influenza viruses from avian influenza viruses [13-15]. These host lineage-specific amino acids were mainly found in components of the viral RNA polymerase complex, such as PB2, PA and NP, essential for viral genome replication [16, 17]. It is likely that these amino acids contribute to the host-range restriction of influenza viruses [18, 19]; however, their biological significance remains to be established, 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 [20, 21].

Data mining has the potential to provide the necessary tools for better understanding of gene expression, drug design, and other emerging problems in genomics and proteomics. Association rule mining is an important task in data mining that finds correlations between items in a database. The association algorithm is commonly used to identify large and frequent item sets and mine hidden relationships among items [22]. The concept can be applied in many fields other than market basket analysis. The association algorithm has also been employed to mine gene expression data [23] and medical data [24]. The method is extended here to mine the association rules which are then applied to identify different host classes. The current prediction scheme emphasizes on influenza host typing and signature. The associative classification technique was chosen because it builds more accurate and easily interpretable set of rules than traditional classification approaches [25].

Here, several computational approaches for finding specific genetic signatures characteristic of human, avian and swine influenza A viral genomes in HA gene were used. Detecting the most statistically significant differences between human, avian and swine group of sequences was done using the VESPA [26], available from the HCV database, which gave the most variable positions and their frequencies between human, avian and swine. Two Sample logos server was used to identify and confirm significant variations between the hosts for each subtype and statistical significance assessed [27] . Distinct amino acid residues between human, swine and 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 VESPA. We subsequently validated the robustness of those signatures with human, avian and swine sequences downloaded from Influenza Virus Resources at the National Center for Biotechnology Information (NCBI) [28].

II. MATERIALS AND METHODS

Our workflow for finding markers is composed of sequence collection and sorting, multiple sequence alignments, informative site identification and feature selection by using comparative sequence logos and viral epidemiology signature pattern analysis (VESPA) for positional enumeration of amino acids in each host group followed by generation of class association rules followed by selection of the best set of rules.

1. SEQUENCE COLLECTION AND ANALYSIS

All sequences were downloaded from the NCBI's Influenza Virus Resources (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html), with including sequences from laboratory-adapted viruses and pandemic (H1N1) 2009 sequences within human host. And force the downloaded sequences to be non redundant and complete isolation of HA segments. Part of the data is used for training and the remaining part is used for testing. We used amino acid sequences because they are known to give more reliable results than nucleotide sequences when the sequence divergence is high [29].

To compare the genomic patterns of avian, swine and human influenza viruses with each others, we downloaded 1500 protein sequences of HA from NCBI's Influenza Virus Resources, isolated from various host species (500 for each host). They were grouped according to host type, and cover all the viral subtypes found in that host. Downloaded FASTA format sequences were parsed into each category such as accession number, subtype, gene, host, occurring year, and other parameters. The signatures obtained from analyzing the primary dataset, have been validated and tested using human, avian or swine test data sets.

2. MULTIPLE SEQUENCE ALIGNMENT (MSA)

One of the cornerstones of modern bioinformatics is the comparison or alignment of protein sequences. With the aid of multiple sequence alignments, biologists are able to study the sequence patterns conserved through evolution and the ancestral relationships between different organisms. Sequences can be aligned across their entire length (global alignment) or only in certain regions (local alignment) [30].The most widely used approach to multiple sequence alignments uses a heuristic search known as progressive technique (also known as the hierarchical or tree method), that builds up a final MSA by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related [31]. All progressive alignment methods require two stages: a first stage in which the relationships between the sequences are represented as a tree, called a guide tree, and a second step in which the MSA is the most popular progressive alignment method has been the Clustal series of programs [30]. The rationale behind the development of the Clustal series has been to provide robust, portable programs that are capable of providing good, biologically accurate alignments within a reasonable time limit. Each group of training sets was collectively aligned using Clustal X program which supports multiple sequence alignment for protein sequences through window graphical user interface [32, 33] and built by adding the sequences sequentially to the growing MSA produced a consensus sequence representing the highly conserved regions from the aligned sequences.

3. PATTERN DISCOVERY AND FEATURE SELECTION

Feature selection techniques have become an apparent need in many bioinformatics applications. In addition to the large pool of techniques that have already been developed in the machine learning and data mining fields, specific applications in bioinformatics have led to a wealth of newly proposed techniques. Apart from the basic features that just represent the nucleotide or amino acid at each position in a sequence, as many of them will be irrelevant or redundant, feature selection techniques are then applied to focus on the subset of relevant variables.

VESPA program can be used to quickly detect amino acids which characterize differences between two groups of sequences. It compares two groups of sequences and looks for a "signature" pattern, or the set of amino acids that is conserved among each set, but differing between the sets [26]. It will pick out those distinguishing amino acids, and calculate their frequencies in each set. The sequences should all be of the same length and a threshold for the minimum degree of conservation of signature amino acids in the query set should be adjusted.

Two Sample Logo calculates statistical significance of the relative position-specific symbol frequencies between two sets of aligned sequences [27]. For example, sequences that are known to share a sequence motif may be locally aligned including positions upstream or downstream from the motif. All aligned sequences in both samples are required to be of the same length, so dash characters ("-") should be used to pad the positions in case some sequences are shorter.

4.Class association rules

The association rule model represents rules where a set of items was associated with each other. For instance, a rule could specify a certain product that was frequently bought in combination with other products. The rules were extracted from some large and frequently occurring itemsets. An itemset was regarded as frequent if the possibility of its occurrence exceeded a specified minimal support criterion. The accuracy of the rules is measured by their support and confidence. The support of the rule is the relative frequency of transactions that the rule can be applied to, confidence which is the number of cases in which the rule is correct relative to the number of cases in which it is applicable (and thus is equivalent to an estimate of the conditional probability of the consequent of the rule given its antecedent). To select interesting rules from the set of all possible rules, a minimum support and a minimum confidence are fixed.

In this study, Class association rules were generated for the sites with statistically significant variations between the host groups in both the VESPA and comparative sequence logo, those whose support and confidence are above 40% were retained.

III. RESULTS

Protein multiple sequence alignments (MSAs) for 1500 sequences of HA for different hosts, were performed all together and then sequences of each specific host were extracted separately. Break down the result of alignment into three sets (human, avian and swine) for comparison. Each two groups of sequences were compared at every position looking for a "signature" pattern in the positive (query) set relative to the negative (background) set. Six different comparisons between hosts were generated as human vs. avian, human vs. swine, avian vs. swine and vice versa. Positional variations between different hosts in HA were compared using a number of tools: VESPA and comparative sequence logos (see methods for elaboration). Tables (1-6) in additional file, showed the most informative positions which are relatively variable in one group than the other and their relative frequencies. Informative class association rules with a certain threshold of support and confidence were generated from the VESPA and comparative sequence logos results. The complete set of rules was shown in tables (1-6) in the additional file with their support and confidence. A support threshold value of 80% was set herein to indicate that for each host, the attribute must appear 400 times in 500 instances and the threshold of confidence was 0.4.

A total of 9, 31, 11, 6, 22, and 31 most informative positions of 630 amino acid residues in HA, yielded significant differences between Avian vs. Human, Human vs. Avian, Human vs. Swine, Swine vs. Human, avian vs. swine, and swine vs. avian respectively (tables (1-6) in additional file).

Table1 Positional variations of HA for human, avian and swine with their support and confidence

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

Table 2 list of the most informative rules obtained with highest support and confidence.

Position	Rule	Support	Confidence
438K	K → Human	82.8%	92.9%
458N	N→ Avian	90.7%	79.5%
463E	E→Avian	90%	79.8%
205T	T→Avian	85.2%	80.2%
603L	L→Avian	83.1%	79.3%
286V	V→Swine	51.4%	98%

Three signature residues at positions 331 and 463 of HA (yellow highlight in table1) exhibited a dominant change between the three hosts (human, avian or swine) with strong support in avian and moderate support in human and swine viruses. Avian signature showed the largest number of informative positions, Positions 10, 33,133, 147, 149, 181, 198, 205, 258, 267, 267, 274, 285, 292, 297, 354, 388, 438 and 603 were informative for finding signatures both for human vs. avian and swine vs. avian. These positions found to separate the two classes (human/swine or avian) with support varied between 59% and 99.3% (table 1). While positions 264 and 356 (gray highlight in table1) were informative for finding signatures both for human vs. avian and human vs. swine. These positions found to separate the two classes (human or avian/swine) with support of 82.8% and 82.4% respectively. Positions 286, 287 and 581 (blue highlight in table1) were informative for finding signatures both for swine vs. human and swine vs. avian. These positions found to separate the two classes (swine or avian/human) but their support was not much higher than 55.5%. This finding indicates that the swine strains are less distinct from avian and human strains. Table 2 listed the most informative rules obtained with highest support and confidence.

Rule support and confidence are the two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. For example, a support of 90% for an association rule in position 463 means that 90% of sequences under analysis show that "E" and "Avian" occur together. A confidence of 79.8% means that 79.8% of the sequences that contain the residue "E" at the same position also found in avian. Typically, association rules are considered interesting if they satisfy both minimum support threshold and a minimum confidence threshold. Such threshold can be set by users or domain experts.

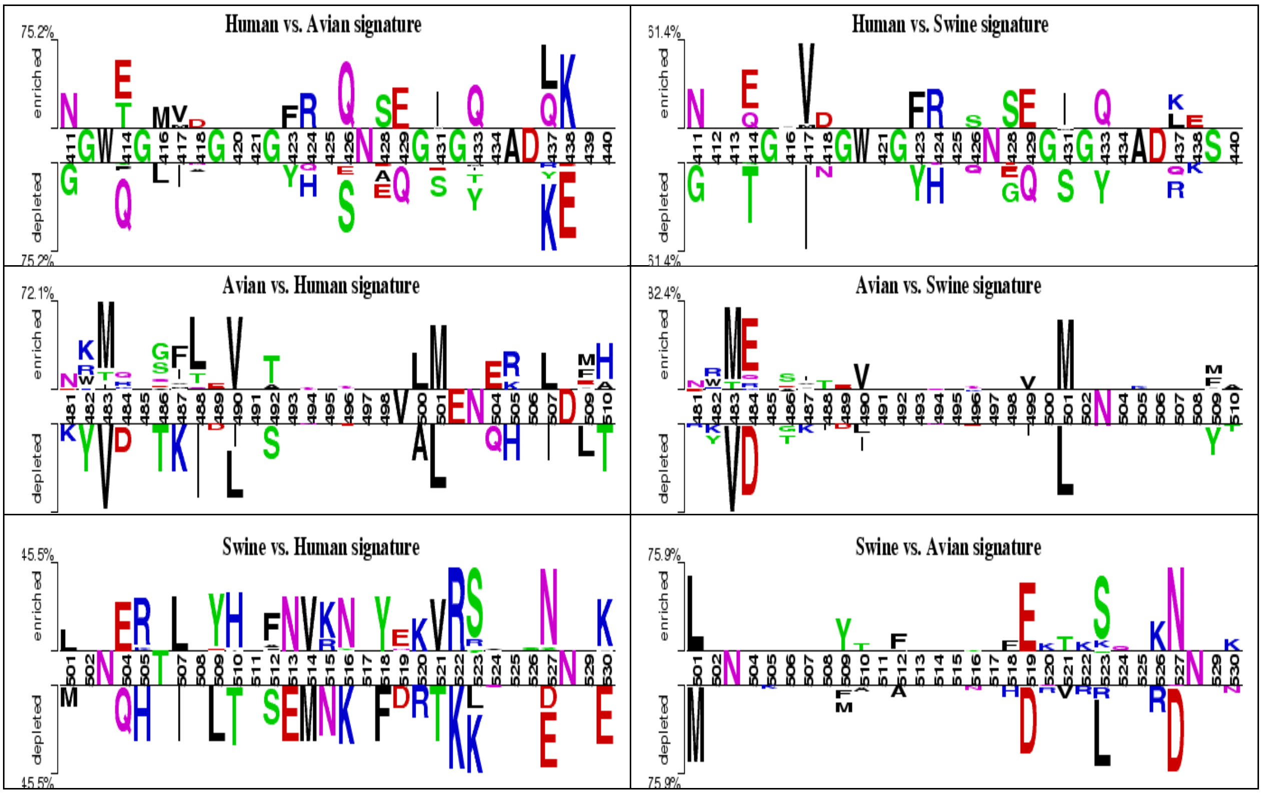

Figure 1 Six comparative sequence logos for human vs. avian, human vs. swine, avian vs. swine and vice versa

The graphical motif representation enables a quick identification of positions that are clearly different by their length, and can therefore be incorporated in the classifier. Figure1 showed six comparative sequence logos for human vs. avian, human vs. swine, avian vs. swine and vice versa. In the figure, the letters in the middle bar represent conserved positions. The totally empty positions represent variations within each group but no considerable variations between the two groups. Comparative sequence logos confirm the results of comparing each two host groups using VESPA (tables 1-6 in additional file). Almost all positions coincided with the comparative sequence logo and signature pattern analysis. However comparative logos added some signature positions such as positions 482,490 and 501were statistically significant in avian vs. human signature.

IV. Discussion

We proposed a computational approach capable of indicating species-associated signatures in studying human, avian and swine influenza viral genomes. This study focused specifically on the persistent markers, and host specificity markers were found only in the surface glycoprotein HA, because of immune pressure and because of the receptor specificity of the HA receptor binding site. This approach using class association rules extracted from the VESPA results and confirmed by comparative sequence logos can help increase the sensitivity and specificity of genetic biomarker discovery in general. These class association rules, which are position and amino acid specific, proved more appropriate and gave high support and confidence.

The approach for defining host specificity markers presented in [12] which predicted positions in the genome associated with human host specificity however, these studies failed to find host markers in the surface glycoproteins HA and NA or in the polymerase protein PB1. All amino acid markers from the HA, NA, and PB1 genes, as well as the alternate transcripts NS2, M2, and PB1-F2, were either poor-quality host discriminators. Thus, it may be that residues in HA, NA, PB1, and PB1-F2 are simply less host differentiating than are other genes. Other recent work in [15] looked more broadly for human markers beyond the pandemic conserved regions but their approach of looking for species-associated signatures by entropy is less useful for HA and NA genes. The genetic diversity that exists in either human or avian viruses for these 2 gene segments can markedly boost their respective entropy to more negative values, thus making it difficult to find residues conserved enough for identifying such signatures .

To identify host specific amino acids that distinguish between human, avian and swine influenza isolates, we first compared HA sequences of human, avian and swine viruses with each other and found potentially important sites that may be related to host tropism and immune responses. These sites may be important for evolutional process in different hosts and host adaptation.

Positions 10, 33,133, 147, 149, 181, 198, 205, 258, 267, 267, 274, 285, 292, 297, 354, 388, and 438 were found to separate the two classes (human/swine or avian) with support varied between 59% and 99.3%, while positions 264 and 356 were found to separate the two classes (human or avian/swine) with support of 82.8% and 82.4% respectively. And Positions 286, 287, 581 and 603 were found to separate the two classes (swine or avian/human) but their support was not much higher than 55.5%. This finding indicates that the swine strains are less distinct from avian and human strains. Notably from table1, that none of these positions could differentiate the three host classes at the same time, except for positions 331 and 463 of HA (yellow highlight in table1) exhibited a dominant change between the three hosts (human, avian or swine) with strong support in avian and moderate support in human and swine viruses. Some markers had little impact on distinguishing the functional classes by themselves; however in combination with other markers they improved class prediction. These findings demonstrate the importance of these residues for receptor specificity and for host range restriction of the virus.

Our methods are designed to identify strictly conserved residues that persist over time and will not capture seasonal changes or even changes between pandemic isolates which is a challenge with HA. For that any classification errors appeared to be due to recent reassortment events that suggest the presence of influenza genomes that are a mix of both human and avian strains. Conner et al. reported 6 amino positions that distinguish human and avian influenza viral sequences [11], none of them were identified in this study because our study mixing all HA subtypes into one class that has been found in human, avian or swine which would substantially alter the reported set of persistent markers. In addition to the data limitations, accurate alignments of HA is hampered by high variability, and despite the care taken in manual editing, false-negative errors may occur due to alignment errors. Although we intended to analyze a comprehensive set of avian versus human influenza A viral genomes, the available sequences are predominated by H5N1 in avian viruses and H3N2 in human viruses. The short supply of sequences other than those 2 subtypes may inevitably cause a certain amount of bias in our results. Nevertheless, these positions may be important for evolutional process in different hosts and host adaptation.

V. CONCLUSIONS

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. One might expect, a priori, to find host markers in the surface glycoprotein HA because of immune pressure and because of the receptor specificity of the HA receptor binding site. Genome wide comparison of human vs. avian, human vs. swine, avian vs. swine and vice versa using VESPA and comparative sequence logos, would show the evolutionary differences between them and thus provide information for studying mechanism of influenza viral infection and replication in different host species. Informative class association rules with a certain threshold of support and confidence were generated to improve prognosis prediction. Pattern and variability analysis on the hole HA sequences are performed, to extract its most important motifs. These host markers has been confirmed and validated by human, avian and swine test sets. Host-specific class association rules for avian gave better support and confidence than human or swine. Positions 205T, 286V, 438K, 458N, 463E and 603L have given the best support and confidence, and would thus aid in the prediction of prognostic biomarkers and improve the accuracy of prognosis.

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] A. S. Beare and R. G. Webster, "Replication of avian influenza viruses in humans," Arch Virol, vol. 119, pp. 37-42, 1991.

[4] Y. Kawaoka, S. Krauss, and R. G. Webster, "Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics," J Virol, vol. 63, pp. 4603-8, Nov 1989.

[5] C. Scholtissek, W. Rohde, V. Von Hoyningen, and R. Rott, "On the origin of the human influenza virus subtypes H2N2 and H3N2," Virology, vol. 87, pp. 13-20, Jun 1 1978.

[6] H. Kida, T. Ito, J. Yasuda, Y. Shimizu, C. Itakura, K. F. Shortridge, Y. Kawaoka, and R. G. Webster, "Potential for transmission of avian influenza viruses to pigs," J Gen Virol, vol. 75 (Pt 9), pp. 2183-8, Sep 1994.

[7] U. Schultz, W. M. Fitch, S. Ludwig, J. Mandler, and C. Scholtissek, "Evolution of pig influenza viruses," Virology, vol. 183, pp. 61-73, Jul 1991.

[8] P. A. Rota, E. P. Rocha, M. W. Harmon, V. S. Hinshaw, M. G. Sheerar, Y. Kawaoka, N. J. Cox, and T. F. Smith, "Laboratory characterization of a swine influenza virus isolated from a fatal case of human influenza," J Clin Microbiol, vol. 27, pp. 1413-6, Jun 1989.

[9] C. Scholtissek, H. Burger, P. A. Bachmann, and C. Hannoun, "Genetic relatedness of hemagglutinins of the H1 subtype of influenza A viruses isolated from swine and birds," Virology, vol. 129, pp. 521-3, Sep 1983.

[10] M. N. Matrosovich, A. S. Gambaryan, S. Teneberg, V. E. Piskarev, S. S. Yamnikova, D. K. Lvov, J. S. Robertson, and K. A. Karlsson, "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, vol. 233, pp. 224-34, Jun 23 1997.

[11] R. J. Connor, Y. Kawaoka, R. G. Webster, and J. C. Paulson, "Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates," Virology, vol. 205, pp. 17-23, Nov 15 1994.

[12] J. E. Allen, S. N. Gardner, E. A. Vitalis, and T. R. Slezak, "Conserved amino acid markers from past influenza pandemic strains," BMC Microbiol, vol. 9, p. 77, 2009.

[13] D. B. Finkelstein, S. Mukatira, P. K. Mehta, J. C. Obenauer, X. Su, R. G. Webster, and C. W. Naeve, "Persistent host markers in pandemic and H5N1 influenza viruses," J Virol, vol. 81, pp. 10292-9, Oct 2007.

[14] M. Shaw, L. Cooper, X. Xu, W. Thompson, S. Krauss, Y. Guan, N. Zhou, A. Klimov, N. Cox, R. Webster, W. Lim, K. Shortridge, and K. Subbarao, "Molecular changes associated with the transmission of avian influenza a H5N1 and H9N2 viruses to humans," J Med Virol, vol. 66, pp. 107-14, Jan 2002.

[15] G. W. Chen, S. C. Chang, C. K. Mok, Y. L. Lo, Y. N. Kung, J. H. Huang, Y. H. Shih, J. Y. Wang, C. Chiang, C. J. Chen, and S. R. Shih, "Genomic signatures of human versus avian influenza A viruses," Emerg Infect Dis, vol. 12, pp. 1353-60, Sep 2006.

[16] T. Deng, J. L. Sharps, and G. G. Brownlee, "Role of the influenza virus heterotrimeric RNA polymerase complex in the initiation of replication," J Gen Virol, vol. 87, pp. 3373-7, Nov 2006.

[17] K. Klumpp, R. W. Ruigrok, and F. Baudin, "Roles of the influenza virus polymerase and nucleoprotein in forming a functional RNP structure," EMBO J, vol. 16, pp. 1248-57, Mar 17 1997.

[18] G. Gabriel, B. Dauber, T. Wolff, O. Planz, H. D. Klenk, and J. Stech, "The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host," Proc Natl Acad Sci U S A, vol. 102, pp. 18590-5, Dec 20 2005.

[19] C. Scholtissek, H. Burger, O. Kistner, and K. F. Shortridge, "The nucleoprotein as a possible major factor in determining host specificity of influenza H3N2 viruses," Virology, vol. 147, pp. 287-94, Dec 1985.

[20] K. Shinya, S. Hamm, M. Hatta, H. Ito, T. Ito, and Y. Kawaoka, "PB2 amino acid at position 627 affects replicative efficiency, but not cell tropism, of Hong Kong H5N1 influenza A viruses in mice," Virology, vol. 320, pp. 258-66, Mar 15 2004.

[21] J. Steel, A. C. Lowen, S. Mubareka, and P. Palese, "Transmission of influenza virus in a mammalian host is increased by PB2 amino acids 627K or 627E/701N," PLoS Pathog, vol. 5, p. e1000252, Jan 2009.

[22] K. S. Leung, K. C. Wong, T. M. Chan, M. H. Wong, K. H. Lee, C. K. Lau, and S. K. Tsui, "Discovering protein-DNA binding sequence patterns using association rule mining," Nucleic Acids Res, Jun 6.

[23] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," Bioinformatics, vol. 19, pp. 79-86, Jan 2003.

[24] S. Doddi, A. Marathe, S. S. Ravi, and D. C. Torney, "Discovery of association rules in medical data," Med Inform Internet Med, vol. 26, pp. 25-33, Jan-Mar 2001.

[25] L. Merschmann and A. Plastino, "A lazy data mining approach for protein classification," IEEE Trans Nanobioscience, vol. 6, pp. 36-42, Mar 2007.

[26] B. Korber and G. Myers, "Signature pattern analysis: a method for assessing viral sequence relatedness," AIDS Res Hum Retroviruses, vol. 8, pp. 1549-60, Sep 1992.

[27] V. Vacic, L. M. Iakoucheva, and P. Radivojac, "Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments," Bioinformatics, vol. 22, pp. 1536-7, Jun 15 2006.

[28] 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.

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

[30] 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.

[31] D. G. Higgins, J. D. Thompson, and T. J. Gibson, "Using CLUSTAL for multiple sequence alignments," Methods Enzymol, vol. 266, pp. 383-402, 1996.

[32] 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.

[33] 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, vol. 23, pp. 2947-8, Nov 1 2007.