J. Biomedical Science and Engineering, 2010, 3, 584-601 **JBiSE**
doi:10.4236/jbise.2010.36081 Published Online June 2010 (http://www.SciRP.org/journal/jbise/).

Novel host markers in the 2009 pandemic H1N1 influenza a virus

Wei Hu

Department of Computer Science Houghton College, Houghton, USA.
Email: wei.hu@houghton.edu

Received 21 March 2010; revised 28 April 2010; accepted 8 May 2010.

ABSTRACT

The winter of 2009 witnessed the concurrent spread of 2009 pandemic H1N1 with 2009 seasonal H1N1. It is clinically important to develop knowledge of the key features of these two different viruses that make them unique. A robust pattern recognition technique, Random Forests, was employed to uncover essential amino acid markers to differentiate the two viruses. Some of these markers were also part of the previously discovered genomic signature that separate avian or swine from human viruses. Much research to date in search of host markers in 2009 pandemic H1N1 has been primarily limited in the context of traditional markers of avian-human or swine-human host shifts. However, many of the molecular markers for adaptation to human hosts or to the emergence of a pandemic virus do not exist in 2009 pandemic H1N1, implying that other previously unrecognized molecular determinants are accountable for its capability to infect humans. The current study aimed to explore novel host markers in the proteins of 2009 pandemic H1N1 that were not present in those classical markers, thus providing fresh and unique insight into the adaptive genetic modifications that could lead to the generation of this new virus. Random Forests were used to find 18 such markers in HA, 15 in NA, 9 in PB2, 11 in PB1, 13 in PA, 10 in NS1, 1 in NS2, 11 in NP, 3 in M1, and 1 in M2. The amino acids at many of these novel sites in 2009 pandemic H1N1 were distinct from those in avian, human, and swine viruses that were identical at these positions, reflecting the uniqueness of these novel sites.

Keywords: 2009 Pandemic H1N1; Host Switch; Influenza; Mutation; Random Forests

1. INTRODUCTION

In addition to the common seasonal H1N1 influenza virus, an antigenically novel swine-origin pandemic H1N1 influenza virus marked the flu season in 2009. It is likely that both 2009 pandemic H1N1 and seasonal influenza will coexist for some time. Elucidation of the characteristics of this new virus has become an important part of the current flu research. The identification of molecular markers for drug resistance, virulence, viral transmission and replication, human adaptation, and evolution can shed new light into the nature of this virus.

There are eight single-stranded RNA segments of the influenza A virus genome. They code 11 proteins: hemagglutinin (HA), neuraminidase (NA), matrix 1 (M1), matrix 2 (M2), nucleoprotein (NP), non-structural protein 1 (NS1), non-structural protein 2 (NS2; also termed nuclear export protein, NEP), polymerase acidic protein (PA), polymerase basic protein 1 (PB1), polymerase basic protein 2 (PB2), and polymerase basic protein 1 – F2 (PB1-F2). Segments 1, 3, 4, 5, and 6 each encode a single protein, *i.e.*, PB2, PA, HA, NP, and NA, respectively, whereas segments 2, 7, and 8 each encode two proteins, *i.e.*, PB1 and PB1-F2, M1 and M2, NS1 and NS2, respectively. The life cycle of influenza virus has the following steps with several proteins involved in each: entry into the host cell (HA, M1 and M2), entry of viral ribonucleoproteins (vRNP) into the nucleus (NP, PA, PB1 and PB2), transcription and replication of the viral genome (PA, PB1, PB2, NS1, and NP), export of the vRNPs from the nucleus (NP, NS2 and M1), and assembly and budding at the host cell plasma membrane (HA, NA, M1 and M2) [1].

Besides mutations, viruses with segmented genomes can generate genetic diversity by exchanging gene segments between different viruses to produce a new virus. Comprehensive phylogenetic analysis suggested that the genes of 2009 pandemic H1N1 were derived from avian (PB2 and PA), human H3N2 (PB1), classical swine (HA, NP and NS), and Eurasian avian-like swine H1N1 (NA and M) lineages [2].

The symptoms of the 2009 pandemic H1N1 flu are

similar to the 2009 seasonal flu with a possibility of additional symptoms such as vomiting and diarrhea [3]. The Center for Disease Control and Prevention (CDC) and Mayo Clinic have developed several molecular tests to detect and discriminate the novel 2009 pandemic H1N1 virus and the 2009 seasonal virus [4]. The matrix (M) gene is highly conserved compared to other gene segments, which makes it an ideal target for RT-PCR assays used to detect the presence of influenza. A mutation in the M gene of the 2009 pandemic virus could invalidate these tests [5].

Sequence survey suggested that there were two distinct evolutionary trends in antigenic drift of H1N1 HAs at two residues 190 and 225. The epidemic H1N1 HAs favor position 190 while the 1918 pandemic and swine HAs favor position 225 [6]. In contrast to these two trends, the 2009 pandemic H1N1 strains are highly conserved at both HA 190 and 225 and possess the signature markers Asp190 and Asp225 that are known to confer specificity to the human α2-6 sialylated glycan receptors [7]. Further analysis indicated the 2009 pandemic H1N1 HAs possess residues that can be positioned to bind to avian α2-3 sialylated glycan receptors as well [7,8]. By homology modeling of the HA structure, the antigenic similarity between the 1918 H1N1 and the pandemic 2009 H1N1 viruses was confirmed, and the future amino acid substitutions on the antigenic sites of 2009 pandemic H1N1 HA were also predicted [9], raising the concerns that these two pandemic H1N1 viruses may share a similar evolutionary path. With informational spectrum method [10], a bioinformatics technique, highly conserved domains and mutations in the 2009 pandemic H1N1 HAs were identified and the contributions of these mutations to the changes of binding specificity of the 2009 pandemic H1N1 HAs were quantified [11-13].

Many 2009 seasonal H1N1 strains carry a NA mutation H275Y that confers high-level resistance to oseltamivir. Although most 2009 pandemic H1N1 strains are susceptible to oseltamivir, the co-circulation of pandemic and seasonal H1N1 viruses might provide opportunities for 2009 pandemic H1N1 to develop oseltamivir resistance through mutations and reassortments between pandemic and seasonal H1N1 viruses. Positive natural selection was detected in the NA proteins of 2009 pandemic H1N1 at codons 275 and 248 and seasonal H1N1 at codon 275, with statistically significant bias of nonsynonymous mutations relative to synonymous mutations [14]. Besides position 275, mutations at other positions in NA such as 116, 117, 119, 136, 150, 151, 199, 223, 275, and 295 could also alter NA inhibitor susceptibility [15,16].

Two recent reports [17,18] revealed three NA variant groups in 2009 pandemic H1N1. The first group had V106 and N248, the second included I106 and N248, and the third contained I106 and D248, highlighting the rapid genetic variation of this surface antigen under host immune pressure and the need for close monitoring. The NA protein of the avian viruses has, in addition to the catalytic site, a separate sialic acid binding site that is not present in human viruses, which could enhance the catalytic efficiency of NA [19]. Although the second binding site was not conserved in swine NA strains, a recent report found the 2009 pandemic H1N1 strain of swine origin appeared to have retained some of the key features of the second binding site. Their data showed possible lowered HA activity for this second site, which might be an important event in the emergence of the 2009 pandemic strain [20].

The interaction of NP and the influenza polymerase, containing the PA, PB1 and PB2 proteins, catalyses viral RNA replication (vRNA→cRNA→vRNA) and transcription (vRNA→mRNA) in the nucleus of infected cells. The PB2 protein of human viruses tend to possess a lysine at position 627 (K627), whereas avian viruses generally have glutamic acid at this position (E627). The mutation E627K allows avian virus to efficiently grow in humans and was identified experimentally as a crucial host range and pathogenicity determinant [21,22]. The 2009 pandemic H1N1 strains could transmit in humans efficiently, but exclusively possess the avian signature E627. Therefore, there might be alternative strategies employed by the novel 2009 H1N1 polymerase to maintain the efficient replication rate. A recent study discovered that serine at position 590 (S590) and arginine at position 591 (R591) might serve as a regulator of polymerase activity that contributes to the increased replication efficiency of 2009 pandemic H1N1. The paired mutations S590 and R591, termed the SR polymorphism, were present in only three of the 2849 PB2 sequences of human viruses before 2009 [23]. Other sites might affect the polymerase activity as well. The mutations at position 504 in PB2 (I504V) and position 550 in PA (I550L) could result in enhanced virulence [24]. A special region (residues 360-374) in the NP protein was found to play a vital role in overcoming species barrier for 2009 pandemic H1N1 [25].

Compared to other proteins in the influenza viruses, PB1-F2 is a newly discovered protein, which is unique in that this protein is coded by a subset of the nucleotides that code for PB1 due to the use of a different reading frame. The PB1-F1 protein has been implicated in pathogenicity and the induction of cell death [26-28]. The 2009 pandemic H1N1 virus has a truncated PB1-F1 protein, because its genome contains three stop codons preventing PB1-F2 expression. Recently, studies found that its function is not universal, but cell type and virus

strain dependent [29], and it plays a critical role in the pathogenicity and tranmissibility of 2009 pandemic H1N1 [30,31].

NS1 is a multifunctional protein that contributes to viral pathogenesis by neutralizing the interferon (IFN) -based defense system of the host cell [32], and serves as a strong inducer of apoptosis in infected human respiratory epithelial cells [33]. The NS1 protein of 2009 pandemic H1N1 is truncated and therefore missing a domain responsible for increased pathogenicity of avian virus [34]. There are two molecular markers that account for the virulence of the highly pathogenic avian H5N1 viruses are not present in 2009 pandemic H1N1. They are a lysine (K) at position 627 of PB2, and glutamic (E) acid in position 92 of NS1 that might increase the replication efficiency and block host inhibition of viral replication, respectively [35,36].

A recent study on M gene identified sites of high selective pressure between human and avian influenza, which were 115, 121, 137 in M1, and 11, 16, 20, 54, 57, 78, 86, and 93 in M2 [37]. The 2009 pandemic H1N1 virus contains the adamantine-resistant mutation S31N in its M2 protein, thus making the NA Inhibitors oseltamivir and zanamivir the only options available to treat the infections caused by the pandemic virus [38].

Several studies focused on determining which amino acid changes best distinguish an avian or swine influenza virus from a human virus. An entropy analysis revealed the human-avian host shift genomic signature of 52 markers in ten proteins of the influenza virus [39]. This signature extended in [40] provided the basis for finding the amino acids of 2009 pandemic H1N1 at the host species-specific positions to illustrate the adaptive mutations of this virus. By comparison of the protein sequences of 2009 pandemic H1N1 with those in the previous pandemics and human, swine, and avian influenza viruses, the mutation trend of the residues at the signature positions was discovered, and the potential roles of the mutated residues in human adaptation and virulence was probed in [41]. With mutual information analysis, the characteristic sites for human-to-human transmission in PB2 of influenza viruses were uncovered [42], and subsequently a catalogue of 68 such sites in eight internal proteins were found to derive adaptation signatures of viral proteomes [43], which included many of the 32 and 34 markers identified in [44,45], respectively.

Many of the molecular determinants associated with adaptation to human hosts or to the emergence of a pandemic virus are not present in 2009 pandemic H1N1, suggesting that other previously unrecognized molecular markers are responsible for its ability to infect humans [17]. Therefore, uncovering new molecular features of 2009 pandemic H1N1 is of prime significance. In this study, we collected all the protein sequences of the 2009 pandemic H1N1, 2009 seasonal H1N1, avian, human, and swine influenza viruses available from the National Center for Biotechnology Information (NCBI). Our objective was to explore the novel host markers in 2009 pandemic H1N1 that were not present in the classical avian-human or swine-human host shift markers, and the top markers that could differentiate 2009 pandemic H1N1 from 2009 seasonal H1N1.

2. MATERIALS AND METHODS

2.1. Sequence Data

All influenza virus protein sequences were retrieved from the Influenza Virus Resource (http://www.ncbi/nlm.nih.giv/genomes/FLU/FLU.html) of the National Center for Biotechnology Information (NCBI). Detailed information about these sequences is in **Table 1**. All the sequences used in the study were aligned with MAFFT [46].

2.2. Random Forests

Random Forest, proposed by Leo Breiman in 1999 [47], is an ensemble classifier based on many decision trees. Each tree is built on a bootstrap sample from the original training set and is unpruned to obtain low-bias trees. The variables used for splitting the tree nodes are a random subset of the whole variable set. The classification decision of a new instance is made by majority voting over all trees. About one-third of the instances are left of the bootstrap sample and not used in the construction of the tree. These instances in the training set are called “out-of-bag” instances and are used to evaluate the performance of the classifier, which can achieve both low bias and low variance with bagging and randomization.

2.3. Feature Selection Using Random Forests

Random Forest calculates several measures of variable importance. The mean decrease in accuracy measure was employed in [48] to rank the importance of the features in prediction. This measure is based on the decrease of classification accuracy when values of a variable in a node of a tree are permuted randomly. In this study, two packages of R, randomForest and varSelRF [48], were utilized to compute the importance of the amino acids in a given protein sequence dataset. The effectiveness and robustness of this technique as a feature selection method has been demonstrated in various studies [49-54].

2.4. Procedure to Find Novel Host Sites in 2009 Pandemic H1N1

Four steps were created to locate the novel sites associated

Table 1. Counts of the influenza protein sequences used in the current study.

Host	Subtype	Protein	Number of Sequences	Years	Host	Subtype	Protein	Number of Sequences	Years
Human	Pandemic H1N1	HA	710	2009	Human	All types	NS1	509	All years
Human	Pandemic H1N1	NA	643	2009	Human	All types	NS2	383	All years
Human	Pandemic H1N1	NP	394	2009	Human	All types	PA	279	All years
Human	Pandemic H1N1	M1	490	2009	Human	All types	PB1	289	All years
Human	Pandemic H1N1	M2	482	2009	Human	All types	PB2	269	All years
Human	Pandemic H1N1	NS1	366	2009	Avian	H1	HA	120	All years
Human	Pandemic H1N1	NS2	358	2009	Avian	N1	NA	1821	All years
Human	Pandemic H1N1	PA	295	2009	Avian	All types	NP	2888	All years
Human	Pandemic H1N1	PB1	311	2009	Avian	All types	M1	4232	All years
Human	Pandemic H1N1	PB2	311	2009	Avian	All types	M2	3182	All years
Human	Seasonal H1N1	HA	128	2009	Avian	All types	NS1	4610	All years
Human	Seasonal H1N1	NA	125	2009	Avian	All types	NS2	3422	All years
Human	Seasonal H1N1	NP	25	2009	Avian	All types	PA	3106	All years
Human	Seasonal H1N1	M1	129	2009	Avian	All types	PB1	2979	All years
Human	Seasonal H1N1	M2	129	2009	Avian	All types	PB2	2643	All years
Human	Seasonal H1N1	NS1	25	2009	Swine	H1	HA	379	All years
Human	Seasonal H1N1	NS2	25	2009	Swine	N1	NA	278	All years
Human	Seasonal H1N1	PA	23	2009	Swine	All types	NP	420	All years
Human	Seasonal H1N1	PB1	25	2009	Swine	All types	M1	516	All years
Human	Seasonal H1N1	PB2	25	2009	Swine	All types	M2	406	All years
Human	H1	HA	640	All years	Swine	All types	NS1	506	All years
Human	N1	NA	1127	All years	Swine	All types	NS2	351	All years
Human	All types	NP	393	All years	Swine	All types	PA	343	All years
Human	All types	M1	1512	All years	Swine	All types	PB1	368	All years
Human	All types	M2	1415	All years	Swine	All types	PB2	327	All years

with host adaptation in 2009 pandemic H1N1.

Step 1: For each protein, the consensus sequence of avian, 2009 pandemic H1N1, human, and swine viruses were calculated separately, and the positions with different amino acids of the four consensus sequences were identified, since the different amino acids at these positions have the potential to contribute to host switches.

Step 2: For each protein, Random Forests were used to identify the top 20 positions that have highest importance in separating avian from human viruses, and swine from human viruses, respectively.

Step 3: Finding the intersection of the top positions with importance larger than 0.005 for separating 2009 pandemic H1N1 from human viruses and the positions with different consensus amino acids found in step one.

Step 4: The positions discovered in step three minus the positions found in step two will be the novel positions important for separating 2009 pandemic H1N1

Table 2. This table contains the Hamming distances of consensus protein sequences of avian, human, 2009 pandemic H1N1, and swine viruses.

Proteins	HA	NA	NP	M1	M2	NS1	NS2	PA	PB1	PB2
Dist(Avian,2009_pandemic)	90	35	28	8	6	35	11	17	23	17
Dist(Human,2009_pandemic)	109	61	42	19	15	42	13	30	21	31
Dist(Swine,2009_pandemic)	43	52	10	11	6	17	7	15	20	17
Dist(Avian,Human)	90	50	31	11	13	23	5	20	4	19
Dist(Avian,Swine)	61	37	21	3	2	22	7	4	6	2
Dist(Human,Swine)	97	52	35	8	11	36	7	19	6	21

from human viruses.

The purpose of steps 2 and 3 was to calculate the adaptation signatures for various virus groups, which were then used in step 4. The amino acids at most of these novel sites in 2009 pandemic H1N1 turned out to be different from those in avian, human, and swine viruses that were the same at these positions.

Random Forests produce non-deterministic outcomes. To compensate this bias, the Random Forests algorithm was run multiple times and then the average of the results was taken. The importance of each residue in the protein sequences was based on the averaged calculations by using the function randomVarImpsRF in varSelRF repeated 5 times.

3. RESULTS

3.1. Comparison of Consensus Protein Sequences of Influenza Viruses

In considering the relationship among the proteins of influenza viruses, the Hamming distance, defined as the number of positions at which the corresponding amino acids of two sequences are different, of any two consensus protein sequences of avian, human, 2009 pandemic H1N1, and swine viruses was calculated. The distance information in **Table 2** provided insight into the sequence similarity between the proteins of 2009 pandemic H1N1 and those of other virus groups. In particular, the distances between 2009 pandemic H1N1 and avian, human, and swine viruses reflected the origin of 2009 pandemic H1N1 [2].

3.2. Novel Host Sites in the Proteins of 2009 Pandemic H1N1

Our analysis discovered a catalogue of novel host markers in the proteins of 2009 pandemic H1N1 that included 18 markers in HA, 15 in NA, 9 in PB2, 11 in PB1, 13 in PA, 10 in NS1, 1 in NS2, 11 in NP, 3 in M1, and 1 in M2. In the following sections, each of the ten proteins of 2009 pandemic H1N1 was compared to that of avian, human, and swine viruses. Random Forests were employed to identify the top important positions in the proteins of influenza that could separate 2009 pandemic H1N1 from avian, human, and swine viruses, and the top positions that could discriminate 2009 pandemic H1N1 and 2009 seasonal H1N1.

The novel host markers in 2009 pandemic H1N1 were uncovered with the procedure outlined in Section 2.3. Some of the markers that could classify 2009 pandemic H1N1 and 2009 seasonal H1N1 were also part of the previously discovered genomic signature that separate avian or swine from human viruses. Because the sequences of 2009 seasonal H1N1 were a subset of those of human viruses, there were common important sites in each protein between the sites in 2009 pandemic versus 2009 seasonal and the sites in 2009 pandemic versus human viruses.

To render a complete picture of host shift markers of different types, the novel sites in each of the ten proteins of 2009 pandemic H1N1 were exhibited along with the avian-human and swine-human sites in a single table. The conservation of residues comprising these sites in each protein as represented by their frequency at these positions was also displayed in the table. The top important sites in each protein for differentiating 2009 pandemic H1N1 from avian, human, and swine viruses were displayed in a single figure, which were used in the procedure to find the novel sites.

Due to high genetic variation of the HA and NA proteins, only the HA protein sequences of H1 subtype and the NA protein sequences of N1 subtype of avian, human, and swine viruses were used to compare those of 2009 pandemic H1N1 in the current analysis. Therefore, the novel markers in HA and NA of 2009 pandemic H1N1 found in this study were subtype-specific. Because all the PB1-F2 proteins of 2009 pandemic H1N1 were truncated and nonfunctional, they were excluded in this study.

3.2.1. HA Protein

As the primary target of host immune responses, the surface protein HA is under high selection pressure, as evidenced by the large number of amino acid substitutions in this protein. There was a clear distinction of amino acids at position 127, where the human HA had a

Table 3. This table contains the consensus amino acids and their frequency at positions in HA that have high importance in separating 2009 pandemic H1N1 from human H1 viruses. The single letter 'a' (for avian) or 's' (for swine) in parenthesis after a position number indicates whether the same position is also important for separating 2009 pandemic H1N1 from avian or swine viruses or both. The novel host sites in this protein are the positions without an 'a' or a 's' or both.

Position	71	84	120(a)	127(a,s)	128(s)	129(s)	130(a,s)	142	168	216	239	250
Avian	L(93.3%)	N(99.2%)	A(96.7%)	E(97.5%)	T(100%)	T(93.3%)	K(94.2%)	S(92.5%)	N(99.2%)	A(94.2%)	T(97.5%)	A(94.2%)
Human	I(92.5%)	N(98.4%)	E(95.0%)	-(100%)	T(95.8%)	V(95.0%)	T(97.8%)	S(88.4%)	N(98.4%)	K(96.1%)	T(99.2%)	A(100%)
2009 H1N1	S(100%)	S(100%)	T(100%)	D(99.7%)	S(97.0%)	N(100%)	K(100%)	K(100%)	D(100%)	I(99.9%)	K(100%)	V(99.9%)
Swine	F(58.6%)	N(91.8%)	A(44.9%)	E(57.3%)	T(81.0%)	N(64.12%)	R(62.0%)	N(66.2%)	N(85.5%)	A(46.7%)	T(80.0%)	V(62.5%)
Position	257	258	260	261	298	302	314	365	374	493	527	
Avian	L(85.8%)	N(95.0%)	G(94.2%)	S(98.3%)	I(93.3%)	E(98.33%)	M(99.2%)	Q(94.2%)	G(100%)	S(96.7%)	L(99.2%)	
Human	L(96.7%)	S(96.7%)	G(98.0%)	F(97.3%)	V(99.5%)	E(100%)	M(99.5%)	Q(98.9%)	G(99.5%)	S(99.2%)	L(99.1%)	
2009 H1N1	M(100%)	E(99.9%)	N(99.0%)	A(99.6%)	I(100%)	K(100%)	L(99.9%)	L(100%)	E(99.4%)	A(100%)	V(99.9%)	
Swine	M(54.1%)	N(41.7%)	G(78.1%)	S(72.0%)	V(73.1%)	E(94.2%)	M(95.3%)	Q(61.7%)	G(90.8%)	S(67.0%)	L(78.1%)	

Table 4. This table contains the consensus amino acids and their frequency at positions in NA that have high importance in separating 2009 pandemic H1N1 from human N1 viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	84	126	149	163(s)	166	189	257	269	285(s)	321
Avian	T(76.4%)	H(98.9%)	V(95.2%)	V(94.1%)	A(99.6%)	S(93.5%)	K(96.8%)	L(99.7%)	A(97.5%)	V(94.6%)
Human	T(92.8%)	H(100%)	V(95.6%)	L(83.8%)	A(99.8%)	G(86.4%)	K(99.7%)	L(100%)	T(85.5%)	V(99.9%)
2009 H1N1	K(100%)	P(100%)	I(100%)	I(100%)	V(99.7%)	N(99.8%)	R(99.8%)	M(99.7%)	S(100%)	I(100%)
Swine	I(55.8%)	H(89.2%)	V(73.0%)	I(84.5%)	A(64.0%)	G(58.3%)	K(90.3%)	L(90.7%)	T(55.04%)	V(71.6%)
Position	331	365(a,s)	369(a,s)	385	389	395	397	398	436	
Avian	G(99.7%)	T(91.4%)	S(99.4%)	S(88.3%)	V(90.4%)	A(99.2%)	T(99.2%)	D(99.5%)	T(99.6%)	
Human	G(98.8%)	N(84.7%)	K(84.5%)	S(99.7%)	V(94.5%)	A(99.7%)	T(99.7%)	D(99.7%)	T(99.5%)	
2009 H1N1	K(100%)	I(99.8%)	N(100%)	N(100%)	I(100%)	G(100%)	N(99.7%)	E(100%)	-(100%)	
Swine	G(68.7%)	I(63.0%)	S(82.0%)	S(69.4%)	V(38.1%)	A(63.3%)	T(87.8%)	D(92.1%)	T(99.6%)	

deletion whereas the other three virus groups had not (**Table 3**). However, as will be seen in the NA protein section below (**Table 4**), the NA protein of 2009 pandemic H1N1 had a deletion at position 436 though the other three virus groups had not. The positions in Table 3 including 71, 84, 130, 257, 258, and 314 had significant effects on the receptor binding specificity of HA of 2009 pandemic H1N1[13]. HA has two functional domains HA1 (residues 1-327) and HA2 (residues 328-549). Evidently, most of the sites in **Table 3** were in HA1, illustrating a much higher selection pressure of HA1 relative to HA2. The HA active site located in a cleft is composed of the residues 91, 150, 152, 180, 187, 191, and 192. The active site cleft of HA is formed by its right edge (131_GVTAA) and left edge (221_RGQAGR) [55]. Four sites 127, 128, 129, and 130 in **Table 3** were near the right edge of the active site (**Table 3**).

3.2.2. NA Protein

In addition to the surface protein HA, the influenza A virus also has NA as another surface protein, and the balanced interplay between them is essential for the life cycle of this virus. Because of its critical role in viral replication and its highly conserved active sites, NA is the main target for drug design against influenza virus. The NA Inhibitors oseltamivir and zanamivir were the only drugs available to treat the infections caused by 2009 pandemic H1N1, because the novel virus had an adamantine-resistant mutation S31N in its M2 protein [38]. As a result, the surveillance of any potential drug-resistant mutations in the NA protein of 2009 pandemic H1N1 received high priority. The mutation H275Y (N1 numbering, H274Y in N2 numbering) in NA is well known for its resistance to NA Inhibitors. There were 123 Ys and 2 Hs in 125 NA sequences of 2009 seasonal H1N1 and 12 Ys and 631 Hs in 643 NA sequences of 2009 pandemic H1N1 used in the current study. Both NAs in 2009 pandemic and 2009 seasonal H1N1 did not have the novel NA mutation Q136K [41] that confers zanamivir resistance.

NA is also is constantly evolving under host immune pressure, and the mutations in Table 4 illustrated its genetic variation. As mentioned in the HA protein section above, the NA of 2009 pandemic H1N1 had a deletion at

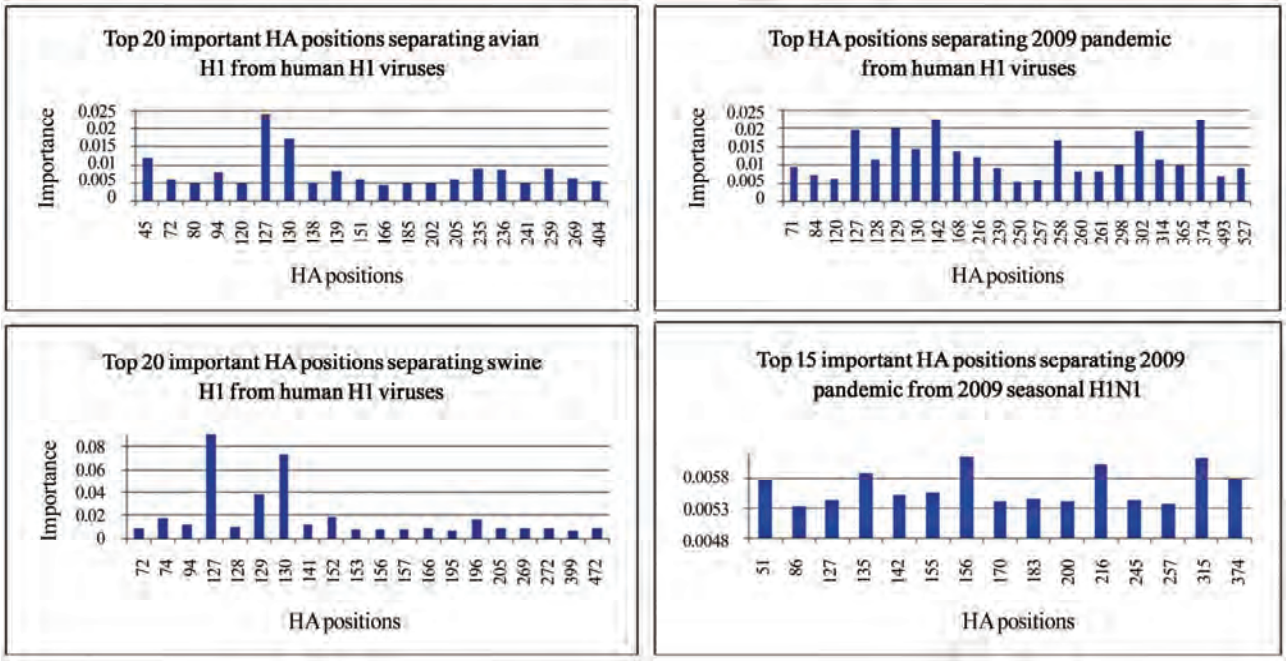

Figure 1. Top important HA positions in distinguishing avian H1, human H1, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine H1 viruses.

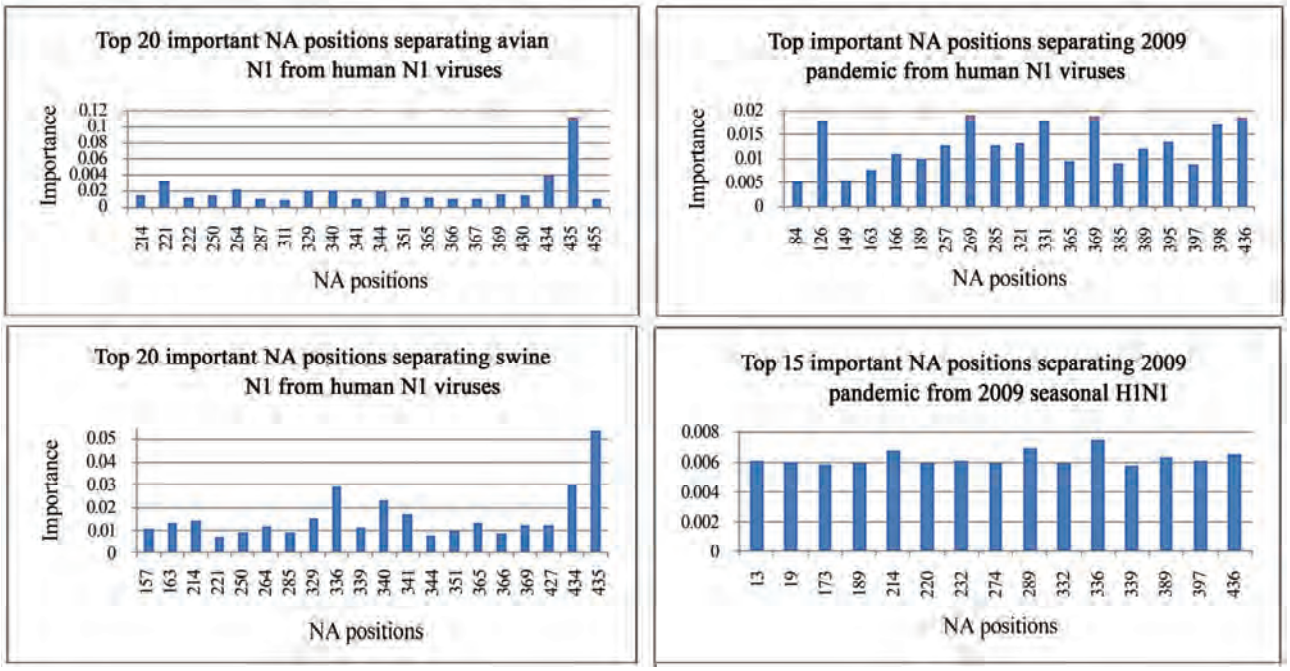

Figure 2. Top important NA positions in distinguishing avian N1, human N1, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine N1 viruses.

position 436 while the other three virus groups had not. However, the HA of human virus had a deletion at position 127 but the other three virus groups had not.

The NA active site is a shallow pocket constructed from conserved residues, some of which contact the substrate directly and participate in catalysis, while others provide a structural framework [56]. According to the numbering in [57], these residues of N1 are 118, 119,

151, 152, 156, 179, 180, 223, 225, 228, 247, 277, 278, 293, 295, 368, and 402. The antigenic sites of N1 are residues 83-143, 156-190, 252-303, 330, 332, 340-345, 368, 370,387-395, 431-435, 448-468. Novel host sites 84, 126, 166, 189, 257, 269, 389, and 395 were at the antigenic sites of N1 (**Table 4**).

Because of a common deletion in the stalk region of the NA proteins of avian viruses, only the residues after position 82 were included in the Random Forest analysis on avian, human, 2009 pandemic H1N1, and swine viruses. However, the whole NA sequences were used in the analysis of 2009 pandemic and 2009 seasonal H1N1.

3.2.3. M1 Protein

M1 protein forms a shell inside the viral envelope to offer strength and rigidity to the viral structure. M1 interacts with HA, NA, M2, and lipid membranes during budding of new virions from the cell surface, and functions in the formation of vRNP complexe and the dissociation of vRNP from the nuclear matrix, and in assembly by recruiting the viral components to the site of assembly. The dissociation of M1 from vRNP is triggered by transport of hydrogen ions across the viral membrane by M2, an early step preceding entry of vRNPs into the cytoplasm of the host cells. M1 also binds to NS2 to facilitate nuclear export of the vRNP [37]. There was a mutation R101K in the M1 protein of 2009 pandemic H1N1 (**Table 5**). It would be of interest to exam the impact of this mutation on viral replication. The basic amino acids 101RKLKR105 of M1 were involved in vRNP binding and nuclear localization. In [58], the functions of 101RKLKR105 were studied by introducing mutations into the M gene of influenza virus A/WSN/33. Individual substitution, R101S or R105S, had a minimal effect on viral replication, but the double mutation R101S-R105S reduced viral replication at a restrictive temperature.

The M1 is a highly conserved protein. Therefore, the changes of M1 may reflect host-specific adaptation. Positions 115, 121, and 137 were identified as avian-human host shift markers in [43]. Our investigation indicated position 218 was as important as position 121. Position 137 was a swine-human marker in [40], but our study also revealed positions 115 and 218 were as important as positions 137 as swine-human markers (**Figure 3**). The novel site 30 was in the membrane binding domain [43], and sites 207 and 209 were in the C-terminal part of M1 (residues 165-252) that binds to vRNP [59].

3.2.4. M2 Protein

This 97 amino acid-long integral membrane protein has three domains, one N-terminal extracellular domain (24 residues) recognized by host immune system, one 19-residue transmembrane domain responsible for ion channel activity, and one 54-residue cytoplasmic tail interacting with M1 and required for genome packing and formation of virus particles [37]. Two M2 inhibitors (adamantine and rimantadine) affect two steps in the replication cycle, viral uncoating and viral maturation. There are five known adamantine-resistanant mutations in M2 (L26F, V27A, A30V, A30T, S31N, and G34E). The 2009 pandemic H1N1 virus contains a mutation S31N. They also contain a mutation L43T in M2 (**Table 6** and **Figure 4**), which is not present in seasonal, triple-reassortant swine or H5N1 influenza viruses [15]. The replacement of the non-polar residue L43 by the polar residue T43 in M2 may influence a nearby functional residue W44, the channel lock and the binding site of rimantadine [60]. Positions 11, 14, 20, 28, 54, 55, 57, 78, and 86 were avian-human host shift sites found in [43]. However, the positions 18, 50, 86, and 93 were as important as these sites in our examination. Positions 57, 86, and 93 were swine-human shift markers in [40], but our analysis also included positions 28, 54, 77, 78, 79, and 89 as swine-human markers with high importance (**Figure 4**). The only novel site in this protein was 13 which was in the extracellular domain (**Table 6**).

3.2.5. NP Protein

The NP protein of the influenza virus binds the RNA genome and functions as an adaptor between the virus and the host cell. The interaction of the NP protein with the viral polymerase is required for viral RNA replication, but not for the synthesis of viral messenger RNAs(transcription). Previous experiments implicated

Table 5. This table contains the consensus amino acids and their frequency at positions in M1 that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	15(a)	30	101(a,s)	115(a,s)	116(s)	121(a,s)	137(a,s)	142(a,s)	166(a,s)	207	209	214(s)	218(a,s)
Avian	I(52.4%)	D(99.9%)	R(52.7%)	V(99.5%)	A(97.5%)	T(96.2%)	T(99.4%)	V(91.6%)	V(53.7%)	S(68.4%)	A(98.9%)	Q(99.2%)	T(99.8%)
Human	V(67.6%)	D(99.7%)	R(93.1%)	I(92.3%)	A(98.7%)	A(92.9%)	A(93.1%)	V(68.2%)	V(93.2%)	S(92.7%)	A(99.9%)	Q(99.8%)	A(84.5%)
2009 H1N1	I(100%)	S (99.8%)	K(100%)	V(99.4%)	S(100%)	T(100%)	T(100%)	A(99.6%)	A(100%)	N(99.8%)	T(100%)	H(100%)	T(100%)
Swine	V(65.3%)	D(77.3%)	R(64.7%)	V(90.9%)	A(68.4%)	A(59.7%)	T(93.2%)	V(77.3%)	V(66.1%)	S(91.9%)	A(78.5%)	Q(67.1%)	T(92.4%)

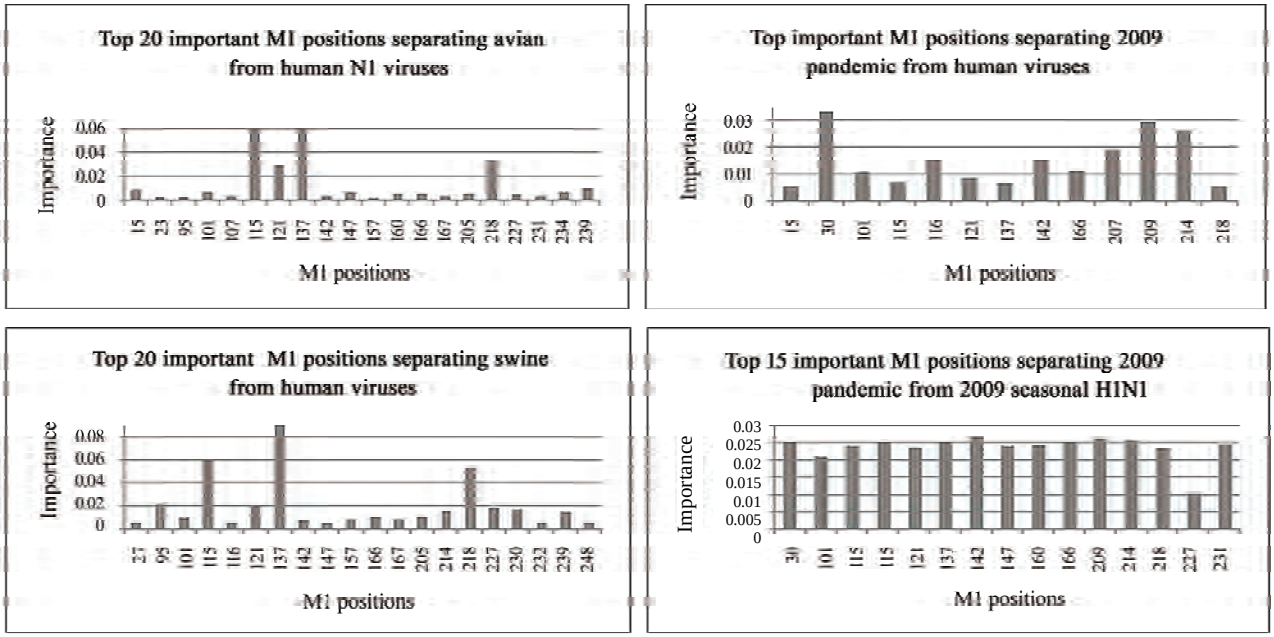

Figure 3. Top important M1 positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

Table 6. This table contains the consensus amino acids and their frequency at positions in M2 that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	11(a,s)	13	16(a,s)	20(a,s)	28(a,s)	31(s)	43(a,s)	77(s)	78(a,s)	86(a,s)
Avian	T(92.0%)	N(91.70%)	E(92.4%)	S(96.2%)	I(54.5%)	S(88.6%)	L(97.0%)	R(98.1%)	Q(99.4%)	V(99.6%)
Human	I(91.6%)	N(99.29%)	G(91.2%)	N(92.2%)	V(96.8%)	S(67.6%)	L(67.7%)	R(99.6%)	K(60.9%)	A(91.7%)
2009 H1N1	T(100%)	S(99.79%)	E(100%)	S(100%)	I(100%)	N(100%)	T(100%)	Q(100%)	Q(100%)	V(100%)
Swine	T(52.7%)	N(72.66%)	E(57.9%)	N(52.7%)	I(38.9%)	S(60.8%)	L(93.8%)	R(63.5%)	Q(95.6%)	V(94.6%)

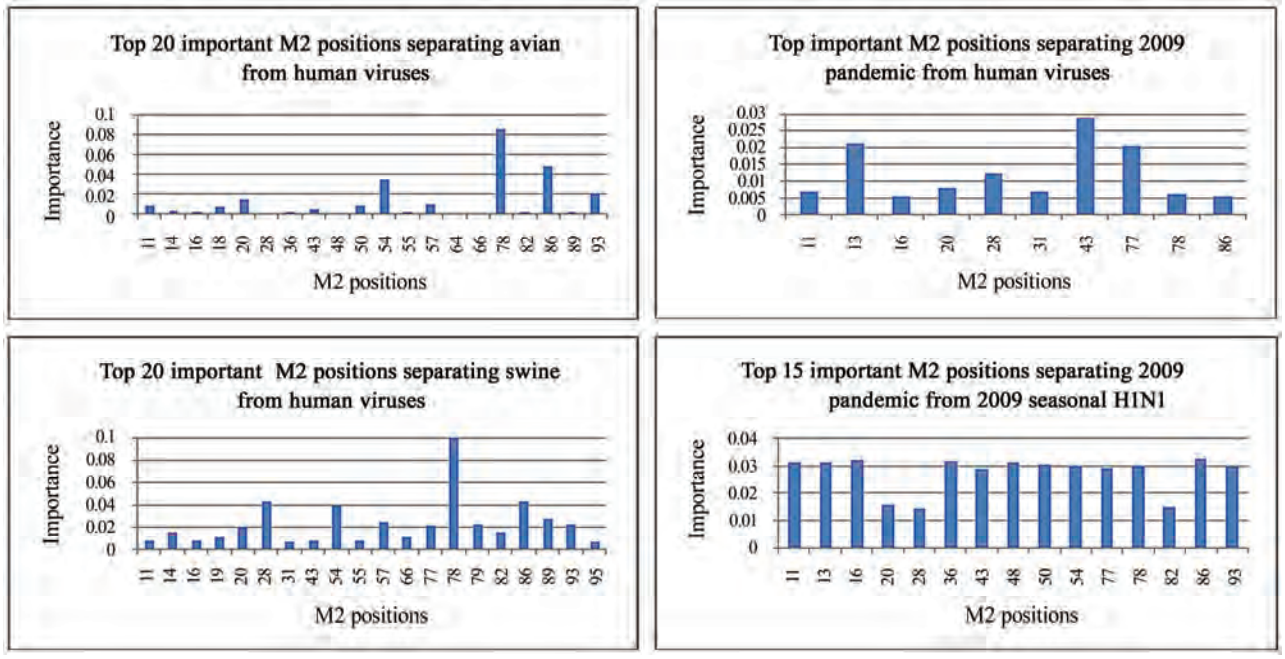

Figure 4. Top important M2 positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

three NP regions (residues 1-160, 256-340 and 340-498) in binding to PB1 and PB2 [61]. Novel sites 21, 53, 119, 316, 353, 371, 377, 433, 444, and 498 were scattered in these three regions (**Table 7**). One region, residues 360-374, in NP of 2009 pandemic H1N1 was deemed extremely important for host range restriction, and is a common feature of pandemic viruses [25]. Two positions 371 and 373 in Table 7 were in this region. Residue 100 was involved in the NP-PB2 interaction [62], and ranked second in separating 2009 pandemic H1N1 from avian viruses. The consensus amino acids at 100 of PB2 proteins of avian, human, 2009 pandemic H1N1, and swine viruses were R, V, I and V respectively. The mutation V100I might contribute to the increased transmissibility or infection of 2009 pandemic H1N1 [41].

Positions 16, 33, 61, 100, 136, 214, 283, 305, 313, 357, 375, and 423 were avian-human host shift markers in [43]. Furthermore, we found positions 31, 217, 373, and 455 significant for discriminating avian and human viruses (**Figure 5**).

3.2.6. NS1 Protein

All of the proteins in influenza virus are structural except for NS1 and PB1-F2. This protein is designated as non-structural because it is synthesized in infected cells, but is not incorporated into virions. NS1 is a multifunctional protein involved in both protein-protein and protein-RNA interactions. Its N-terminal region has an RNA-binding domain (residues 1-73) and its C-terminal region (residues 74-237) contains the effector domain that inhibits the maturation and exportation of the host cellular antiviral mRNAs [63].

Because of a truncation in the NS1 proteins of 2009 pandemic H1N1, only the first 219 residues of the NS1 proteins were included in our analysis. Positions 22, 60, 81, 84, 215, and 227 were avian-human host shift sites in [43], whereas our Random Forests analysis implied positions 79, 81, 114, 171, and 215 were as significant as

Table 7. This table contains the consensus amino acids and their frequency at positions in NP that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	21	31(a,s)	53	119	189(s)	190	217(a)	289(s)	313(a,s)	316
Avian	N(99.20%)	R(99.8%)	E(99.90%)	I(97.65%)	M(99.1%)	V(98.61%)	I(94.8%)	Y(99.2%)	F(99.0%)	I(99.58%)
Human	N(97.5%)	K(65.4%)	E(100%)	I(97.46%)	M(97.5%)	V(97.20%)	S(48.1%)	Y(97.5%)	Y(78.1%)	I(99.75%)
2009 H1N1	D(100%)	R(100%)	D(100%)	V(100%)	I(99.75%)	A(100%)	V(98.2%)	H(99.7%)	V(100%)	M(100%)
Swine	D(61.19%)	R(82.6%)	E(98.10%)	V(57.14%)	I(60.00%)	A(56.42%)	I(76.90%)	H(64.29%)	F(86.66%)	I(95.24%)
Position	350(s)	353	371	373(a)	377	430(s)	433	444	456(s)	498
Avian	T(94.39%)	V(90.30%)	M(94.77%)	T(69.18%)	S(69.67%)	T(94.8%)	T(95.36%)	I(99.00%)	V(98.4%)	N(96.09%)
Human	T(97.20%)	S(52.42%)	M(91.35%)	A(34.35%)	S(80.66%)	T(83.72%)	T(88.04%)	I(98.22%)	V(82.95%)	N(96.18%)
2009 H1N1	K(100%)	I(99.75%)	V(100%)	T(76.40%)	N(100%)	S(100%)	N(100%)	V(100%)	L(100%)	S(99.24%)
Swine	K(64.52%)	V(53.81%)	V(59.29%)	A(57.38%)	S(51.90%)	S(38.57%)	N(60.95%)	I(64.76%)	L(61.19%)	N(69.29%)

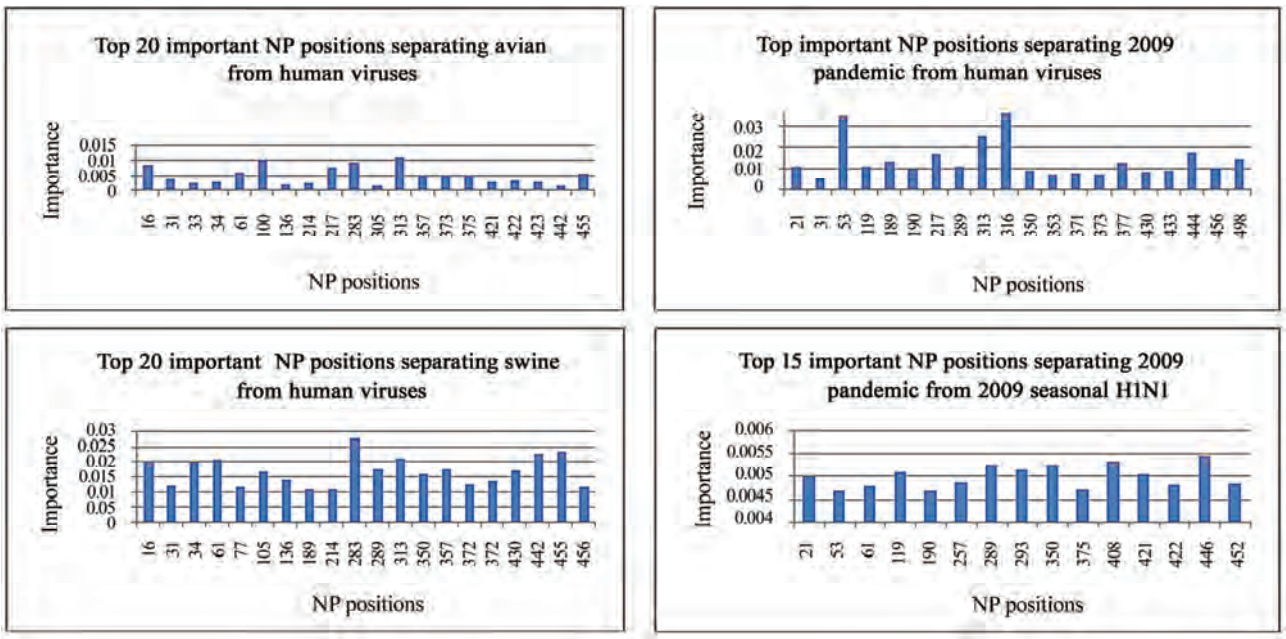

Figure 5. Top important NP positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

Table 8. This table contains the consensus amino acids and their frequency at positions in NS1 that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	6	25(s)	59	67(s)	74	76	91	111	112(a)
Avian	V(77.85%)	Q(79.35%)	R(73.41%)	R(78.85%)	D(97.87%)	A(77.79%)	T(97.33%)	V(79.76%)	A(59.11%)
Human	V(95.48%)	Q(95.09%)	H(45.19%)	R(54.62%)	D(97.64%)	A(97.84%)	T(96.66%)	V(96.07%)	E(55.00%)
2009 H1N1	M(99.73%)	N(99.73%)	L(100%)	W(100%)	S(99.73%)	T(100%)	S(99.73%)	I(100%)	I(99.45%)
Swine	V(95.26%)	N(58.70%)	L(59.49%)	W(59.68%)	D(37.49%)	T(58.70%)	A(60.28%)	V(50.59%)	A(36.96%)
Position	119	129(s)	171(a,s)	198	205	206(s)	207	213(s)	217(s)
Avian	M(99.39%)	I(75.55%)	D(48.87%)	L(52.52%)	S(69.50%)	S(67.66%)	D(59.20%)	P(92.56%)	K(68.00%)
Human	M(85.85%)	M(51.28%)	I(55.20%)	L(84.28%)	S(92.93%)	S(91.55%)	N(78.00%)	P(97.05%)	K(69.16%)
2009 H1N1	L(100%)	V(100%)	Y(100%)	I(100%)	N(100%)	C(100%)	D(100%)	S(100%)	E(100%)
Swine	M(90.12%)	I(64.23%)	D(59.29%)	L(97.04%)	S(64.62%)	R(56.13%)	N(92.09%)	P(51.38%)	E(57.51%)

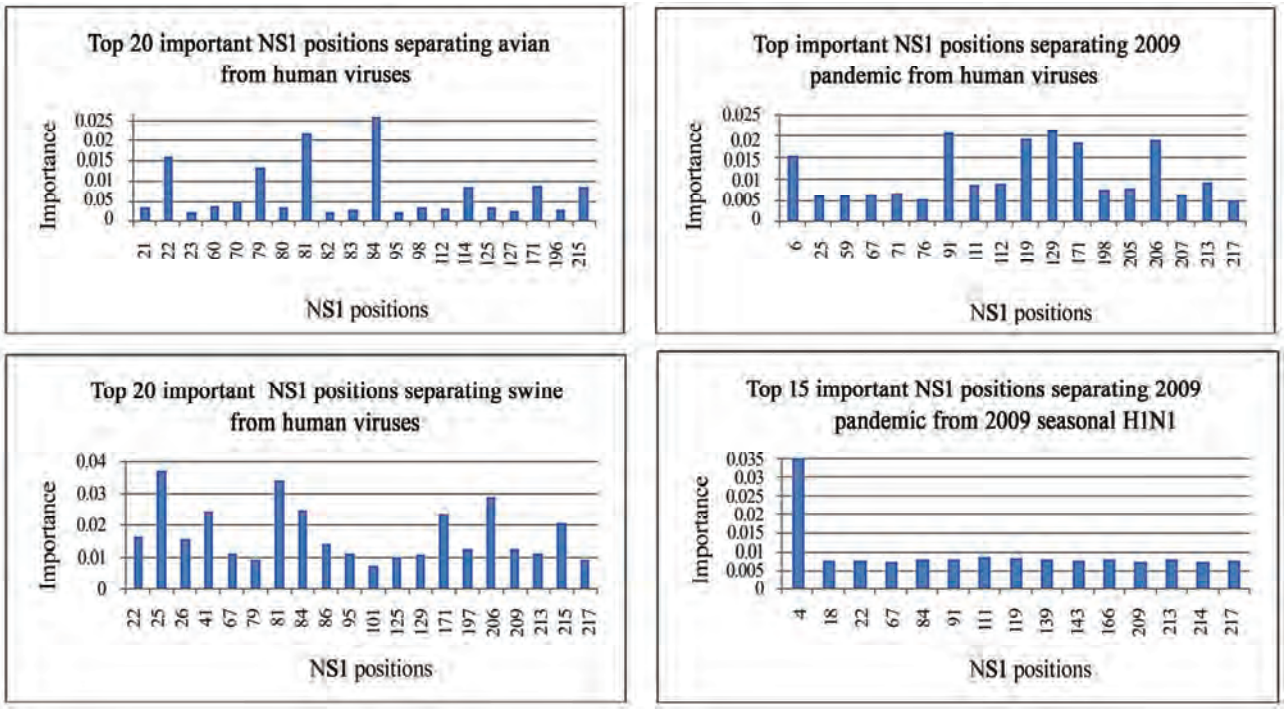

Figure 6. Top important NS1 positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

these sites (**Figure 6**). There were two novel sites 6 and 59 in the RNA-binding domain and the other novel sites in the effector domain (**Table 8**).

3.2.7. NS2 Protein

Influenza virus replicates its RNA genome in the nucleus of infected cells. The NS2 protein mediates the nuclear export of virion RNAs, with help from M1 and NP. A recent report indicated that it also has a role in the regulation of viral transcription and replication [64]. NS2 contains a highly conserved nuclear export signal motif in its amino-terminal region (residues 12-21) [65], and site 14 in **Table 9** was in this region.

Positions 60, 70, and 107 were avian-human host shift markers in [43]. We found position 14 important as a host marker as well. Position 107 was a swine-human host switch marker in [40], but our analysis also pointed to positions 14, 32, 49, and 57 as such sites of high significance (**Figure 7**). The NS2 protein of 2009 pandemic H1N1 contained so many important avian-human or swine-human sites, resulting in only one site as a novel site (**Table 9**).

3.2.8. PA Protein

Compared to the well-defined functions of PB1 and PB1, PA is involved in a diverse range of functions of the polymerase complex, including protein stability, endonuclease activity, and cap binding and promoter binding

[66]. Positions 28, 55, 57, 65, 66, 100, 225, 268, 321, 337, 356, 382, 400, 404, 409, 421, and 552 were avian-human host shift markers in [43]. Additionally, we found positions 241 and 383 equally important as these positions as avian-human markers. Positions 268 and 552 were swine-human markers uncovered in [40]. Our analysis suggested the positions 28, 225, 337, and 400 were equally crucial as these two sites as swine-human markers (**Figure 8**).

The N-terminal domain of PA (residues 1-256) harbors several functional domains, including an endonuclease active site with a putative active site motif, two putative nuclear transport motifs (residues 124-139 (NLS1) and residues 186-247 (NLS2)), and a proteolytic domain that can induce generalized proteolysis of both viral and host proteins. The C-terminal domain of PA (residues 257-716) binds to PB1 for complex formation and nuclear transport [66]. There were three novel sites 186, 204, and 213 within the second putative nuclear localization signals (NLS2), and one novel site 626 within the PB1 binding domain (**Table 10**).

3.2.9. PB1 Protein

The influenza virus polymerase is responsible for replication and transcription of the eight gene segments of the viral RNA genome in the infected host cell. PB1 can interact with PB2, PA, and NP and binds to viral promoter, and is accountable for viral RNA elongation and cap RNA cleavage activities [66,67].

Position 336 was the only avian-human host shift markers in [43]. We found positions 212, 327, 361, 375, 384, 401, 473, and 584 equally significant as position 336 (**Figure 9**). There were one novel site 12 within the PB1-PA binding domain (residues 1-25) and two novel sites 618 and 728 in the PB1-PB2 binding domain (residues 600-757) (**Table 11**) [68].

PB1-PA binding domain (residues 1-25) and two novel

Table 9. This table contains the consensus amino acids and their frequency at positions in NS2 that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	6	14(a,s)	32(s)	34(s)	40(a,s)	48(a)	57(a,s)
Avian	V(79.1%)	M(56.2%)	I(99.0%)	Q(95.6%)	L(71.54%)	A(73.2%)	S(97.6%)
Human	V(98.2%)	L(55.6%)	I(99.0%)	Q(98.7%)	L(61.9%)	A(96.6%)	S(60.3%)
2009 H1N1	M(99.7%)	M(100%)	V(99.7%)	R(100%)	I(100%)	T(100%)	Y(100%)
Swine	V(95.2%)	M(83.2%)	V(68.1%)	Q(53.3%)	I(67.8%)	A(70.1%)	Y(65.8%)
Position	60(a,s)	63(a,s)	83(a)	89(a,s)	107(a,s)	115(a)	
Avian	S(55.7%)	G(75.8%)	V(71.7%)	I(69.8%)	L(99.9%)	T(84.4%)	
Human	N(68.7%)	G(96.3%)	V(98.2%)	T(63.2%)	F(74.9%)	T(89.8%)	
2009 H1N1	S(100%)	E(95.8%)	M(99.7%)	A(97.8%)	L(100%)	A(99.4%)	
Swine	N(61.0%)	E(62.1%)	V(97.4%)	M(32.5%)	L(90.9%)	T(98.3%)	

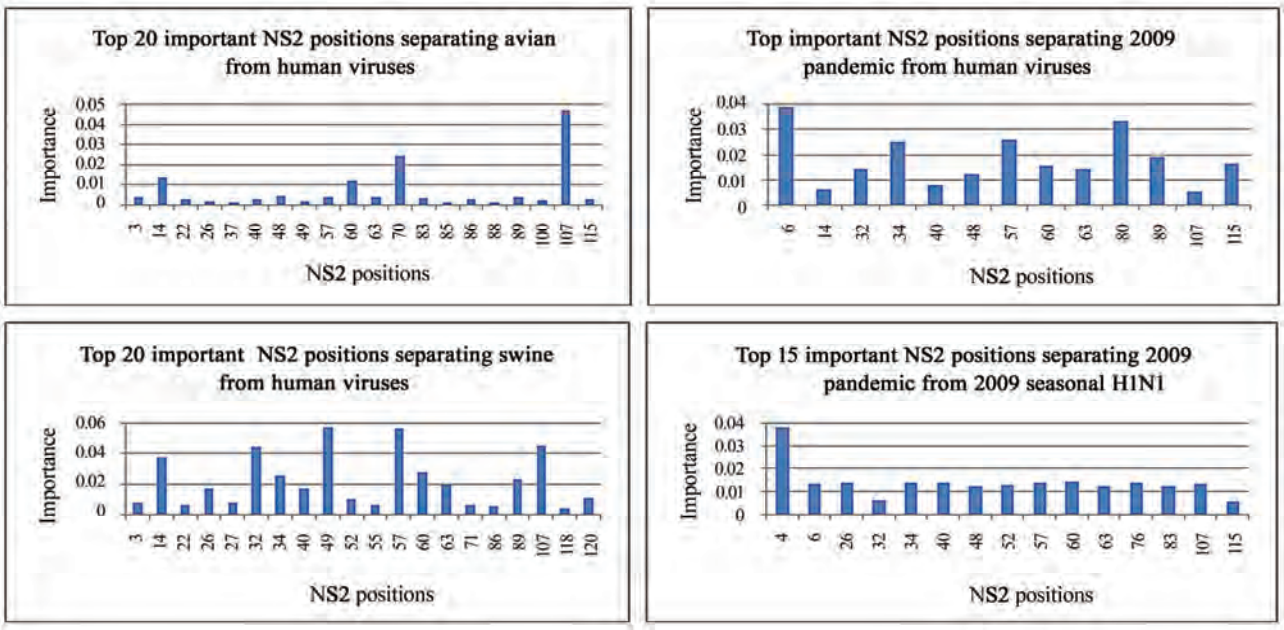

Figure 7. Top important NS2 positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

Table 10. This table contains the consensus amino acids and their frequency at positions in PA that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	28(a,s)	55(a)	85	100(a,s)	186	204	213	256	262	275	277
Avian	P(99.5%)	D(98.9%)	T(96.5%)	V(94.5%)	G(98.0%)	R(80.6%)	R(97.8%)	R(98.6%)	K(96.9%)	P(97.5%)	S(97.1%)
Human	L(70.3%)	N(71.3%)	T(91.0%)	A(70.6%)	G(100%)	R(53.0%)	R(97.5%)	R(54.5%)	K(95.0%)	P(98.9%)	S(35.8%)
2009 H1N1	P(100%)	D(100%)	I(100%)	V(100%)	S(100%)	K(100%)	K(100%)	K(100%)	R(100%)	L(98.6%)	H(100%)
Swine	P(79.9%)	D(53.4%)	T(60.9%)	V(86.6%)	G(96.8%)	R(90.7%)	R(92.1%)	R(63.8%)	K(77.8%)	P(91.5%)	S(53.4%)
Position	336	337(a,s)	356(a)	362	388	400(a,s)	404(a,s)	407	552(a,s)	626	
Avian	L(99.5%)	A(88.7%)	K(98.9%)	K(99.5%)	S(80.1%)	S(40.9%)	A(93.2%)	I(95.9%)	T(99.7%)	K(83.7%)	
Human	L(97.1%)	S(35.8%)	R(69.9%)	K(98.6%)	S(84.2%)	L(79.2%)	S(72.0%)	I(98.6%)	S(71.7%)	K(98.2%)	
2009 H1N1	M(100%)	A(99.0%)	R(99.7%)	R(100%)	G(99.0%)	P(100%)	A(100%)	V(99.3%)	T(99.7%)	R(99.3%)	
Swine	L(95.6%)	A(86.9%)	K(53.9%)	K(71.1%)	S(52.8%)	P(30.6%)	A(86.9%)	I(71.1%)	T(91.5%)	K(96.2%)	

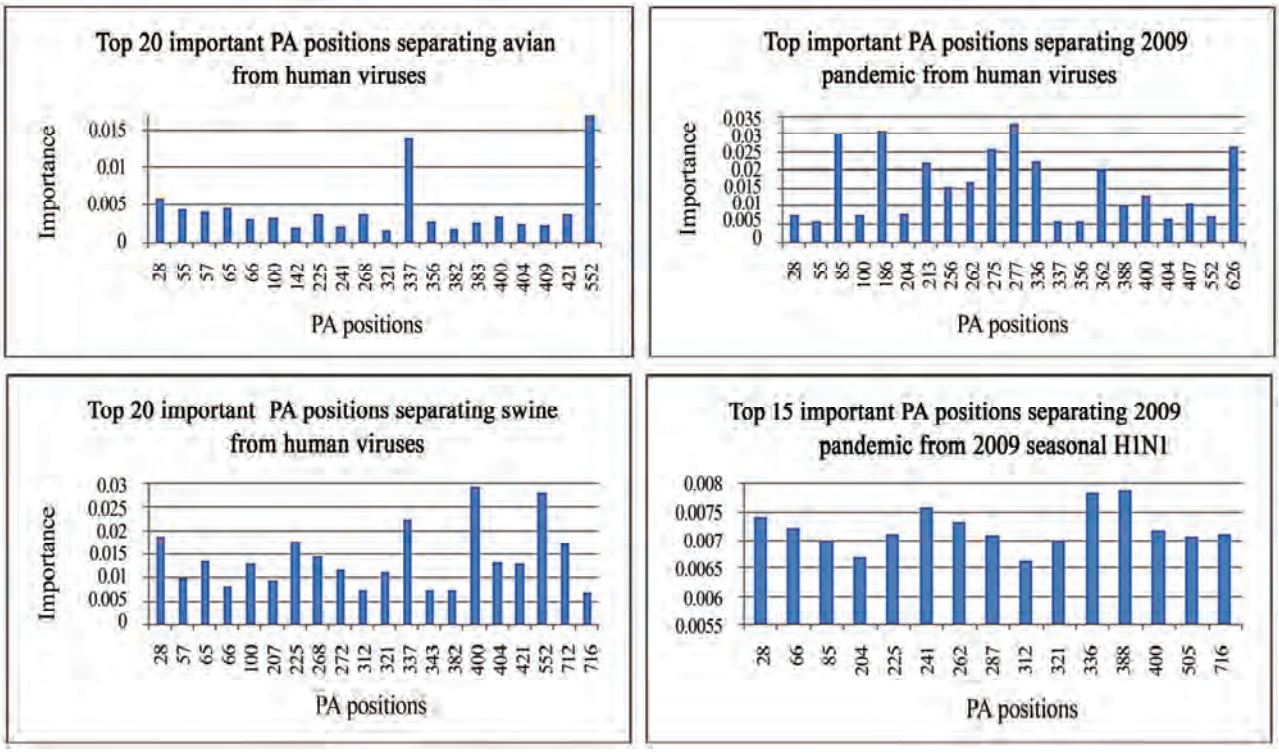

Figure 8. Top important PA positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

sites 618 and 728 in the PB1-PB2 binding domain (residues 600-757) (**Table 11**) [68].

3.2.10. PB2 Protein

PB2 interacts with PB1 and NP, but not PA. Its primary function is binding to cap structures on host cell pre-mRNAs before they are cleaved to provide primers for viral mRNA synthesis [66]. Positions 9, 44, 64, 81, 105, 199, 271, 292, 368, 475, 567, 588, 613, 627, 661, 674, and 702 were avian-human host shift markers in [43]. Positions 108, 197, and 684 were as significant as these sites in our finding. Position 44 was a swine- human marker in [40], but our analysis implied positions 64, 65, 81, 105, 199, 292, 567, 627, 649, 661, and 674 were equally important as position 44 (**Figure 10**). Position 702, an avian-human marker selected in [40,43,69], ranked 21th in our Random Forests analysis, and therefore it was not included in our plot in **Figure 10**. In addition to the SR polymorphism, S590 and R591, found in [23], novel sites in PB2 discovered here provided additional polymorphism that might convey enhanced polymerase activity in human cells. Position 627 in PB2 was considered critical for host shifts in our analysis, a well-known host marker discussed in [21,22], and was located in the PB2-PB1 and PB2-NP binding domains [43].

The PB2-NP binding domain contains residues 1-269 and 580-683, and the PB2-PB1 binding domain contains residues 51-259 and 580-759. There were novel sites 54, 590, 645, and 667 in the PB2-PB1 and PB2-NP binding

Table 11. This table contains the consensus amino acids and their frequency at positions in PB1 that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	12	175	179	216	298	327(a,s)	339(s)	361(a,s)	364	386(s)
Avian	V(99.0%)	D(95.9%)	M(95.9%)	S(95.7%)	L(98.6%)	R(98.7%)	I(98.7%)	S(99.0%)	L(99.5%)	R(57.0%)
Human	V(99.7%)	D(97.2%)	M(69.6%)	S(65.7%)	L(79.2%)	K(53.6%)	I(96.9%)	S(61.9%)	L(99.3%)	R(65.4%)
2009 H1N1	I(99.7%)	N(100%)	I(100%)	G(100%)	I(100%)	R(100%)	M(100%)	R(100%)	I(100%)	K(100%)
Swine	V(95.1%)	D(96.7%)	M(66.0%)	S(60.1%)	L(96.7%)	R(89.7%)	I(44.6%)	N(32.3%)	L(96.5%)	R(95.4%)
Position	435	486	517(s)	584(a,s)	587	618	638(s)	728	741(a,s)	
Avian	T(99.0%)	R(98.6%)	I(99.1%)	R(97.1%)	A(98.4%)	E(97.7%)	E(98.8%)	I(99.1%)	A(96.2%)	
Human	T(99.3%)	R(64.7%)	I(81.3%)	R(63.7%)	A(98.6%)	E(99.7%)	E(98.6%)	I(100%)	A(59.2%)	
2009 H1N1	I(99.4%)	K(100%)	V(100%)	Q(100%)	V(97.4%)	D(100%)	D(100%)	V(100%)	S(100%)	
Swine	T(66.8%)	R(64.1%)	I(75.5%)	R(40.5%)	A(86.7%)	E(61.4%)	E(69.0%)	I(98.4%)	A(59.0%)	

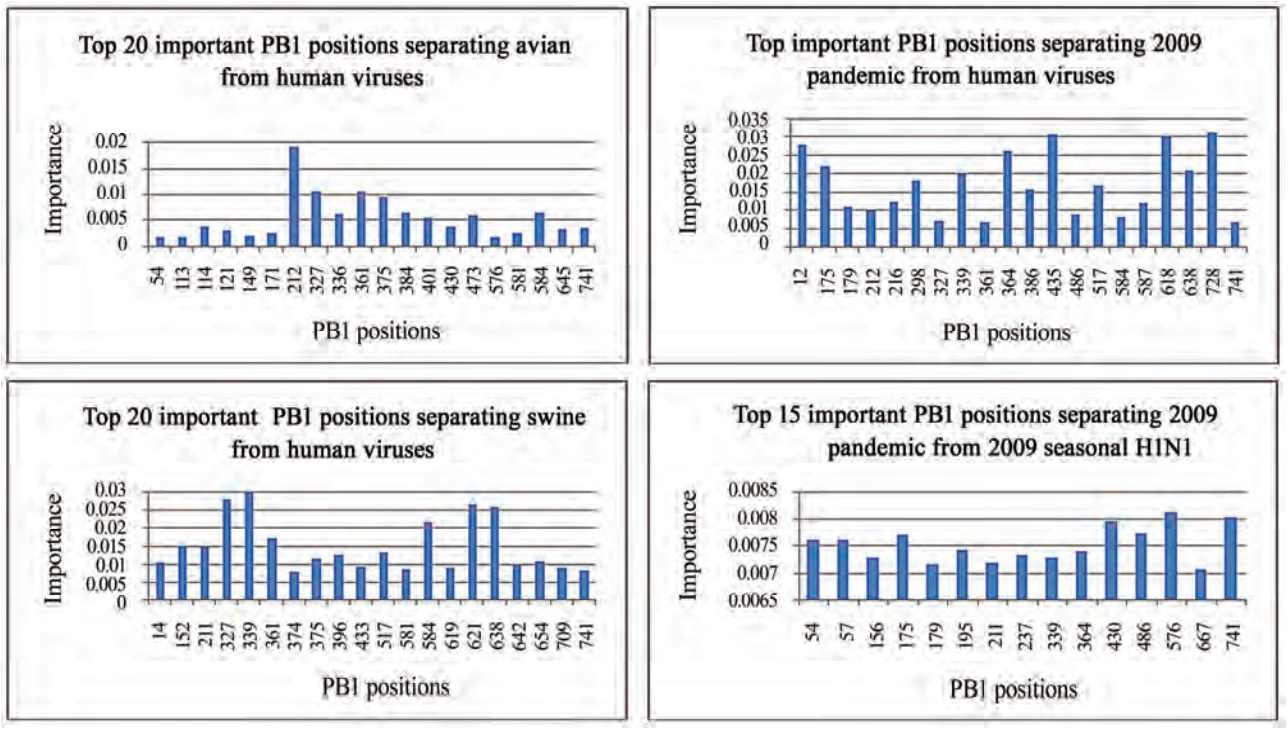

Figure 9. Top important PB1 positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

domains and sites 147 and 225 in the PB2-NP binding domain [43] (**Table 12**).

4. DISCUSSIONS

Extensive research to date provided highly informative knowledge about the origin and genetic lineages of 2009 pandemic H1N1, but the host markers of this new virus remained elusive. Recent studies indicated that human host adaptation is complex and multigenic, and the well-known host shift markers are lacking in this new virus. The hypothesis in the current study was that these markers of 2009 pandemic H1N1 might exist outside of the space of traditional host switch markers. To test this hypothesis in this study, Random Forests were applied to uncover novel important markers in each of the ten proteins of influenza that could differentiate 2009 pandemic H1N1 from human viruses, but were not present in the previous avian-human or swine-human host switch markers.

Our approach naturally led to a systematic discovery of new host markers like the SQ polymorphism found in [23] that could enrich our current knowledge of 2009 pandemic H1N1 and complement the repertoire of existing host shift signatures. Among others, this study revealed the novel host sites 54, 147, 225, 315, 453, 559, 590, 645, and 667 in PB2 of 2009 pandemic H1N1. They provided ample potential sites to investigate experimentally whether they also compensate the lack of

amino acid lysine at residue 627, as the SR polymorphism. Their prospective broader roles in enhancing this new virus's replication and transmission in humans are worthy of further research. In this regard, the three positions 54, 315, and 559 in PB2 were particularly of interest because they had much higher importance than the two positions 590 and 591 associated with the SR polymorphism.

Four proteins are involved and required in the synthesis of influenza virus RNA, which are PB2 and PA of avian lineage, PB1 of human origin, and NP derived from classical swine viruses in 2009 pandemic H1N1. To gain insight into the adaptive strategies employed by these four proteins of different origins to evade restriction in human cells will be a challenge. The novel sites identified in this study provided a starting point for future integrative examination of the interactions of these proteins.

It was expected that 2009 pandemic H1N1 would co-circulate with seasonal H1N1 for some time. Our catalogue of amino acid markers that could effectively separate 2009 pandemic H1N1 from 2009 seasonal H1N1 presented a valuable view of these two viruses that share similar clinical courses but are unique genetically.

Table 12. This table contains the consensus amino acids and their frequency at positions in PB2 that have high importance in separating 2009 pandemic H1N1 from human viruses. The novel host sites in this protein are the positions without an 'a' (for avian) or a 's' (for swine) or both.

Position	9(a)	54	64(a,s)	65(s)	81(a,s)	105(a,s)	147	184(s)	199(a,s)
Avian	D(97.5%)	K(99.7%)	M(74.9%)	E(97.8%)	T(97.3%)	T(90.9%)	I(82.1%)	T(96.3%)	A(99.2%)
Human	N(71.0%)	K(100%)	T(68.0%)	E(98.5%)	M(52.0%)	V(52.4%)	I(87.7%)	T(99.6%)	S(72.9%)
2009 H1N1	D(100%)	R(100%)	M(100%)	D(99.7%)	T(100%)	T(100%)	T(100%)	A(99.0%)	A(100%)
Swine	D(63.9%)	K(98.5%)	M(53.2%)	E(69.4%)	T(84.7%)	T(87.8%)	I(68.2%)	T(57.2%)	A(55.0%)
Position	225	292(a,s)	315	340(s)	453	475(a)	559	567(a,s)	588(a,s)
Avian	S(99.4%)	I(88.6%)	M(95.2%)	R(52.2%)	P(94.7%)	L(99.2%)	T(91.0%)	D(98.0%)	A(95.8%)
Human	S(98.9%)	T(73.6%)	M(99.6%)	R(60.2%)	H(52.0%)	M(70.3%)	T(71.7%)	N(70.3%)	I(68.4%)
2009 H1N1	G(100%)	V(99.4%)	I(100%)	K(97.4%)	S(99.7%)	L(100%)	I(100%)	D(100%)	T(98.4%)
Swine	S(72.2%)	I(56.0%)	M(97.2%)	R(59.9%)	P(57.8%)	L(54.1%)	T(70.9%)	D(90.2%)	A(55.7%)
Position	590	591(s)	613(a,s)	627(a,s)	645	661(a,s)	667	674(a,s)	684(a)
Avian	G(87.6%)	Q(97.9%)	V(98.4%)	E(91.7%)	M(99.4%)	A(96.2%)	V(92.4%)	A(97.1%)	A(96.9%)
Human	G(69.9%)	Q(98.9%)	T(64.3%)	K(80.3%)	M(98.9%)	T(78.4%)	I(62.1%)	T(69.1%)	A(51.3%)
2009 H1N1	S(99.7%)	R(100%)	V(100%)	E(100%)	L(100%)	A(100%)	V(100%)	A(99.7%)	S(100%)
Swine	G(71.3%)	Q(67.3%)	V(74.9%)	E(53.8%)	M(74.9%)	A(48.0%)	V(69.4%)	A(86.5%)	A(74.3%)

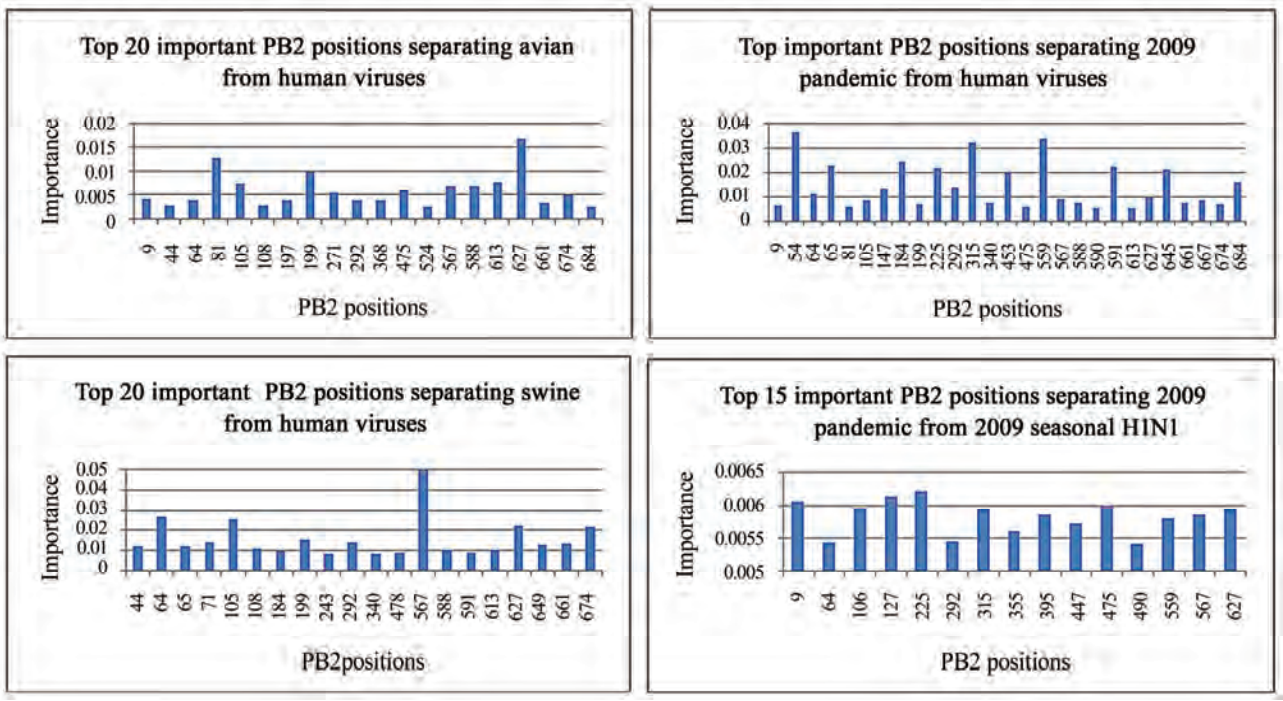

Figure 10. Top important PB2 positions in distinguishing avian, human, 2009 pandemic H1N1, 2009 seasonal H1N1, and swine viruses.

Various computational techniques including entropy [39,40], mutual information [42,43], statistical tests [44], and support vector machines [45] were utilized to discover molecular markers in influenza viruses. To demonstrate the validity of using Random Forests as a feature selection technique in identifying novel host markers in 2009 pandemic H1N1, the top markers found by Random Forests to distinguish the human virus from avian or swine viruses were also included in this report, which contained many known host adoption markers from previous studies. There were fewer novel sites in M1, M2, and NS2 than in the other proteins under this study resulting from many avian-human or swine-human sites among these proteins.

5. CONCLUSIONS

Our findings confirmed that there are novel host sites in the proteins of 2009 pandemic H1N1 that could separate this new virus from human viruses with high confidence. These markers could not be found in the search space of traditional avian-human or swine-human host shift markers, thus offering new potential sites for further experimental verification to elucidate their biological functions.

6. ACKNOWLEDGEMENTS

We thank Houghton College for its financial support.

REFERENCES

[1] Samji, T. (2009) Influenza A: Understanding the viral life cycle. *Yale Journal of Biology Medicine*, **82(4)**, 153-159.

[2] Gavin, J.D., Smith, D.V., Bahl, J., Lycett, S.J., *et al.* (2009) Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**, 1122-1125.

[3] Chang, Y.S., van Hal, S.J., Spencer, P.M, Gosbell, I.B. and Collett, P.W. (2010) Comparison of adult patients hospitalised with pandemic (H1N1) 2009 influenza and seasonal influenza during the PROTECT phase of the pandemic response. *The Medical Journal of Australia*, **192(2)**, 90-93.

[4] Dhiman, N., Mark, J.E., Irish, C., Wright, P., Smith, T.F. and Pritt, B.S. (2010) Mutability in the matrix gene of novel influenza A H1N1 virus detected using a fret probe-based real-time reverse transcriptase PCR assay. *Journal of Clinical Microbiology*, **48(2)**, 677-679.

[5] Zheng, X., Todd, K.M., Yen-Lieberman, B., Kaul, K., Mangold, K. and Shulman, S.T. (2009) Unique finding of a 2009 H1N1 influenza virus—positive clinical sample suggests matrix gene sequence variation. *Journal of Clinical Microbiology*, **48(2)**, 665-666.

[6] Shen, J., Ma, J. and Wang, Q. (2009) Evolutionary trends of A (H1N1) influenza virus hemagglutinin since 1918. *PLoS One*, **4(11)**, e7789.

[7] Soundararajan, V., Tharakaraman, K., Raman, R., Raguram, S., Shriver, Z., Sasisekharan, V. and Sasisekharan, R. (2009) Extrapolating from sequence—the 2009 H1N1 'swine' influenza virus. *Nature Biotechnology*, **27**, 510-513.

[8] Childs, R.A., Palma, A.S., Wharton, S., Matrosovich, T., Liu, Y., Chai, W.G., Campanero-Rhodes, M.A., *et al.* (2009) Receptor-binding specificity of pandemic influenza A (H1N1) 2009 virus determined by carbohydrate microarray. *Nature Biotechnology*, **27**, 797-799.

[9] Igarashi, M., Ito, K., Yoshida, R., Tomabechi, D., Kida, H. and Takada, A. (2009) Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. *PLoS One*, **5(1)**, e8553.

[10] Cosic, I. (1997) The resonant recognition model of macromolecular bioreactivity, theory and application. Birkhauser Verlag, Berlin.

[11] Veljkovic, V., Niman, H.L., Glisic, S., Veljkovic, N., Perovic, V. and Muller, C.P. (2009) Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. *BMC Structural Biology*, **9**, 62.

[12] Hu, W. (2010) Identification of highly conserved domains in hemagglutinin associated with the receptor binding specificity of influenza viruses: 2009 H1N1, avian H5N1 and swine. *Journal of Biomedical Science and Engineering*, **3**, 114-123.

[13] Hu, W. (2010) Quantifying the effects of mutations on receptor binding specificity of influenza viruses. *Journal of Biomedical Science and Engineering*, **3**, 227-240.

[14] Janies, D.A., Voronkin, I.O., Studer, J., Hardman, J., Alexandrov, B.B., Treseder, T.W. and Valson, C. (2010) Selection for resistance to oseltamivir in seasonal and pandemic H1N1 influenza and widespread co-circulation of the lineages. *International Journal of Health Geographics*, **9(1)**, 13.

[15] Deyde, V.M., Sheu, T.G., Trujillo, A.A., Okomo-Adhiambo, M., Garten, R., Klimov, A.I. and Gubareva, L.V. (2010) Detection of molecular markers of drug resistance in 2009 pandemic influenza A (H1N1) viruses by pyrosequencing. *Antimicrob Agents Chemother*, **54(3)**, 1102-1110.

[16] Hurt, A.C., Holien, J.K., Parker, M., Kelso, A. and Barr, I.G. (2009) Zanamivir-resistant influenza viruses with a novel neuraminidase mutation. *The Journal of Virology*, **83(20)**, 10366-10373.

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

[18] Itoh, Y., Shinya, K., Kiso, M., Watanabe, T., Sakoda, Y., Hatta, M., Muramoto, Y., *et al.* (2009) In vitro and in vivo characterization of new swine-origin H1N1 influenza viruses. *Nature*, **460**, 1021-1025.

[19] Uhlendorff, J., Matrosovich, T., Klenk, H.D. and Matrosovich, M. (2009) Functional significance of the hemadsorption activity of influenza virus neuraminidase and its alteration in pandemic viruses. *Archives of Virology*, **154(6)**, 945-957.

[20] Sung, J.C., van Wynsberghe A.W., Amaro, R.E., Li, W.W. and McCammon, J.A. (2010) Role of secondary

sialic acid binding sites in influenza N1 neuraminidase. *Journal of the American Chemistry Society*, **132(9)**, 2883-2885.

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

[22] Subbarao, E.K., London, W. and Murphy, B.R. (1993) A single amino-acid in the Pb2-gene of influenza-A virus is a determinant of host range. *Journal of Virology*, **67**, 1761-1764.

[23] Mehle, A. and Doudna, J.A. (2009) Adaptive strategies of the influenza virus polymerase for replication in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **106(50)**, 21312-21316.

[24] Rolling, T., Koerner, I., Zimmermann, P., Holz, K., Haller, O., Staeheli, P. and Kochs, G. (2009) Adaptive mutations resulting in enhanced polymerase activity contribute to high virulence of influenza A virus in mice. *Journal of Virology*, **83 (13)**, 6673-6680.

[25] Liu, X. and Zhao, Y.P. (2010) Switch region for pathogenic structural change in conformational disease and its prediction. *PLoS One*, **5(1)**, e8441.

[26] Chen, W., Calvo, P.A., Malide, D., Gibbs, J., Schubert, U., Bacik, I., Basta, S., O'Neill, R., Schickli, J., Palese, P., Henklein, P., Bennink, J.R. and Yewdell, J.W. (2001) A novel influenza A virus mitochondrial protein that induces cell death. *Nature Medicine*, **7**, 1306-1312.

[27] Lamb, R.A. and Takeda, M. (2001) Death by influenza virus protein. *Nature Medicine*, **7**, 1286-1288.

[28] Zell, R., Krumbholz, A., Eitner, A., Krieg, R., Halbhuber, K.J. and Wutzler, P. (2007) Prevalence of PB1-F2 of influenza A viruses. *Journal of General Virology*, **88**, 536-546.

[29] McAuley, J.L., Zhang, K. and McCullers, J.A. (2010) The effects of influenza A virus PB1-F2 protein on polymerase activity are strain specific and do not impact pathogenesis. *Journal of Virology*, **84(1)**, 558-564.

[30] Ramakrishnan, M.A., Gramer, M.R., Goyal, S.M. and Sreevatsan, S. (2009) A Serine12Stop mutation in PB1-F2 of the 2009 pandemic (H1N1) influenza A: A possible reason for its enhanced transmission and pathogenicity to humans. *Journal of Veterinary Science*, **10(4)**, 349-351.

[31] Trifonov, V. and Rabadan, R. (2009) The contribution of the pb1-f2 protein to the fitness of influenza a viruses and its recent evolution in the 2009 influenza A (H1N1) pandemic virus. *PLoS Current*: *Influenza*, **21**, RRN1006.

[32] Hale, B.G., Randall, R.E., Ortín, J. and Jackson, D. (2008) The multifunctional NS1 protein of influenza A viruses, *Journal of General Virology*, **89**, 2359-2376.

[33] Zhang, C.F., Yang, Y.T., Zhou, X.W., Liu, X.L., Song, H.B., He, Y.X. and Huang, P.T. (2010). Highly pathogenic avian influenza A virus H5N1 NS1 protein induces caspase-dependent apoptosis in human alveolar basal epithelial cells. *Virology Journal*, **7**, 51.

[34] Jackson, D., Hossain, M.J., Hickman, D., Perez, D.R. and Lamb, R.A. (2008) A new infl uenza virus virulence determinant: The NS1 protein four C-terminal residues modulate pathogenicity. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 4381-4386.

[35] Seo, S.H., Hoffmann, E. and Webster, R.G. (2002) Lethal H5N1 influenza viruses escape host anti-viral cytokine responses. *Nature Medicine*, **8**, 950-954.

[36] Salomon, R., Franks, J., Govorkova, E.A., Ilyushina, N.A., Yen, H.L., Hulse-Post, D.J., Humberd, J., Trichet, M., Rehg, J.E., Webby, R.J., Webster, R.G. and Hoffmann, E. (2006) The polymerase complex genes contribute to the high virulence of the human H5N1 influenza virus isolate A/Vietnam/1203/04, *Journal of Experimental Medicine*, **203(3)**, 689-697.

[37] Furuse, Y., Suzuki, A., Kamigaki, T. and Oshitani, H. (2009) Evolution of the M gene of the influenza A virus in different host species: Large-scale sequence analysis. *Virology Journal*, **6**, 67.

[38] Furuse, Y., Suzuki A. and Oshitani, H. (2009) Large-scale sequence analysis of M gene of influenza A viruses from different species: Mechanisms for emergence and spread of amantadine resistance. *Antimicrobial Agents and Chemotherapy*, **53(10)**, 4457-4463.

[39] Chen, G.W., Chang, S.C., Mok, C.K., Lo, Y.L., Kung, Y.N., *et al.* (2006) Genomic signatures of human versus avian influenza A viruses. *Emerging Infectious Diseases*, **12**, 1353-1360.

[40] Chen, G.W. and Shih, S.R. (2009) Genomic signatures of influenza A pandemic (H1N1) 2009, Virus. *Emerging Infectious Diseases*, **15**, 1897-1903.

[41] Pan, C., Cheung, B., Tan, S., Li, C., Li, L., *et al.* (2010) Genomic signature and mutation trend analysis of pandemic (H1N1) 2009, Influenza A virus. *PLoS One*, **5(3)**, e9549.

[42] Miotto, O., Heiny, A., Tan, T.W., August, J.T., Brusic, V. (2008) Identification of human-to-human transmissibility factors in PB2 proteins of influenza A by large-scale mutual information analysis. *BMC Bioinformatics*, **9**, S18.

[43] Miotto, O., Heiny, A.T., Albrecht, R., García-Sastre, A., Tan, T.W., August, J.T. and Brusic, V. (2010) Complete-proteome mapping of human influenza A adaptive mutations: implications for human transmissibility of zoonotic strains. *PLoS One*, **5(2)**, e9025.

[44] Finkelstein, D.B., Mukatira, S., Mehta, P.K., Obenauer, J.C., Su, X., Webster, R.G. and Naeve, C.W. (2007) Persistent host markers in pandemic and H5N1 influenza viruses. *Journal of Virology*, **81(19)**, 10292-10299.

[45] Allen, J.E., Gardner, S.N., Vitalis, E.A., Slezak, T.R. (2009) Conserved amino acid markers from past influenza pandemic strains. *BMC Microbioloy*, **9**, 77.

[46] Katoh, K., Kuma, K., Toh, H. and Miyata, T. (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, **33**, 511-518.

[47] Breiman, L. (2001) Random Forests. *Machine Learning*, **45(1)**, 5-32.

[48] Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

[49] Archer, K.J. and Kimes, R.V. (2008) Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, **52**, 2249-2260.

[50] Reif, D.M. Motsinger, A.A., McKinney, B.A., Crowe, J.E. and Moore, J.H. (2006) Feature selection using a random forests classifier for the integrated analysis of

multiple data types. *Proceedings of* 2006 *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, Toronto.

[51] Granittoa, P.M., Furlanellob, C., Biasiolia, F. and Gasperia, F. (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, **83**, 83-90.

[52] Menze1, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. and Hamprecht, F.A. (2009) A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, **10**, 213.

[53] Gao, D., Zhang, Y.X. and Zhao, Y.H. (2009) Random forest algorithm for classification of multi-wavelength data. *Research in Astronomy and Astrophysics*, **9(2)**, 220-226.

[54] Hu, W. (2009) Identifying predictive markers of chemosensitivity of breast cancer with random forests. *Journal of Biomedical Science and Engineering*, **3(1)**, 59-64.

[55] KováccaronOVá, A., Ruttkay-Nedecký, G., Karol HaverlíK1, I. and Janecccaronek, S. (2002) Sequence similarities and evolutionary relationships of influenza virus A hemagglutinins. *Virus Genes*, **24**, 57-63.

[56] Colman, P.M., Hoyne, P.A. and Lawrence, M.C. (1993) Sequence and structure alignment of paramyxovirus hemagglutinin-neuraminidase with influenza virus neuraminidase. *Journal of Virology*, **67**, 2972-2980.

[57] Maurer-Stroh, S. Ma, J.M., Lee, R.T.C., Sirota, F.L. and Eisenhaber, F. (2009) Mapping the sequence mutations of the 2009 H1N1 influenza A virus neuraminidase relative to drug and antibody binding sites. *Biology Direct*, **4**, 18.

[58] Liu, T. and Ye, Z.P. (2005) Attenuating mutations of the matrix gene of influenza A/WSN/33 Virus. *Journal of Virology*, **79(3)**, 1918-1923.

[59] Baudin, F., Petit, I., Weissenhorn, W. and Ruigrok, R.W.H. (2001) In vitro dissection of the membrane binding and RNP binding activities of influenza virus M1 protein. *Virology*, **281**, 102-108.

[60] Dua, Q.S., Wang, S.Q., Huang, R.B. and Chou, K.C. (2010) Computational 3D structures of drug-targeting proteins in the 2009-H1N1 influenza A virus. *Chemical Physics Letters*, **485**, 191-195.

[61] Ye Q., Krug R.M. and Tao Y.J. (2006) The mechanism by which influenza A virus nucleoprotein forms oligomers and binds RNA. *Nature*, **444**, 1078-1082.

[62] Biswas, S.K., Boutz, P.L. and Nayak, D.P. (1998) Influenza virus nucleoprotein interacts with influenza virus polymerase proteins. *Journal of Virology*, **72**, 5493-5501.

[63] Lin, D., Lan, J. and Zhang, Z. (2007) Structure and function of the NS1 protein of influenza A virus. *Acta Biochim Biophys Sin* (*Shanghai*), **39(3)**, 155-162.

[64] Robb, N.C., Smith, M., Vreede, F.T. and Fodor, E. (2009) NS2/NEP protein regulates transcription and replication of the influenza virus RNA genome. *Journal of General Virology*, **90**, 1398-1407.

[65] Iwatsuki-Horimoto, K., Horimoto, T., Fujii, Y. and Kawaoka, Y. (2004) Generation of influenza A virus NS2 (NEP) mutants with an altered nuclear export signal sequence. *Journal of Virology*, **78(18)**, 10149-10155.

[66] Yuan, P.W., Bartlam, M., Lou, Z.Y., Chen, S.D., Zhou, J., He, X.J., Lv, Z.Y., Ge, R.W., Li, X.M., Deng, T., Fodor, E., Rao, Z.H. and Liu, Y.F. (2009) Crystal structure of an avian influenza polymerase PAN reveals an endonuclease active site. *Nature*, **458**, 909-913.

[67] Biswas, S.K. and Nayak, D.P. (1994) Mutational analysis of the conserved motifs of influenza A virus polymerase basic protein 1. *Journal of Virology*, **68**, 1819-1826.

[68] Ohtsu, Y., Honda, Y., Sakata, Y., Kato, H. and Toyoda, T. (2002) Fine mapping of the subunit binding sites of influenza virus RNA polymerase. *Microbiology and Immunology*, **46**, 167-175.

[69] Taubenberger, J.K., Reid, A.H., Lourens, R.M., Wang, R., Jin, G. and Fanning, T.G. (2005) Characterization of the 1918 influenza virus polymerase genes. *Nature*, **437(7060)**, 889-893.

