J. Biomedical Science and Engineering, 2009, 2, 550-558 **JBiSE**
doi: 10.4236/jbise.2009.27080 Published Online November 2009 (http://www.SciRP.org/journal/jbise/).

Analysis of correlated mutations, stalk motifs, and phylogenetic relationship of the 2009 influenza A virus neuraminidase sequences

Wei Hu

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

Received 13 October 2009; revised 27 October 2009; accepted 30 October 2009.

ABSTRACT

The 2009 H1N1 influenza pandemic has attracted worldwide attention. The new virus first emerged in Mexico in April, 2009 was identified as a unique combination of a triple-reassortant swine influenza A virus, composed of genetic information from pigs, humans, birds, and a Eurasian swine influenza virus. Several recent studies on the 2009 H1N1 virus utilized small datasets to conduct analysis. With new sequences available up to date, we were able to extend the previous research in three areas. The first was finding two networks of co-mutations that may potentially affect the current flu-drug binding sites on neuraminidase (NA), one of the two surface proteins of flu virus. The second was discovering a special stalk motif, which was dominant in the H5N1 strains in the past, in the 2009 H1N1 strains for the first time. Due to the high virulence of this motif, the second finding is significant in our current research on 2009 H1N1. The third was updating the phylogenetic analysis of current NA sequences of 2009 H1N1 and H5N1, which demonstrated that, in clear contrast to previous findings, the N1 sequences in 2009 are diverse enough to cover different major branches of the phylogenetic tree of those in previous years. As the novel influenza A H1N1 virus continues to spread globally, our results highlighted the importance of performing timely analysis on the 2009 H1N1 virus.

Keywords: Entropy; Co-mutation; Mutation; Mutual Information; Neuraminidase; Phylogenetic Analysis; Random Forest; Stalk Motif; Swine Flu

1. INTRODUCTION

There are three types of flu viruses, types A, B, and C. Type A viruses are the most pathogenic to humans. The influenza has on its surface two glycoproteins, hemagglutinin (HA) and neuraminidase (NA), based on which influenza is classified. There are 16 types of HA proteins and 9 types of NA proteins, which are named H1, H2, H3 and etc. For example, "bird flu" is H5N1 and "swine flu" is H1N1. The HA binds the virus to sialic acid receptors on the host cell surface. The NA protein facilitates the release of virions to infect other cells by removing sialic acid residues from the viral HA during entry and release from cells. The NA protein is a tetramer of four identical polypeptide chains anchored in the membrane of the virus. Its head domain is globular and supported by a long and thin stalk.

The "Spanish" influenza pandemic of 1918–1919 caused about 50 million deaths worldwide and about one-third of the world's population was infected. One unique feature of the 1918 influenza pandemic was the simultaneous infection of humans and swine. Recent studies on the 1918 virus revealed that the genes encoding the HA and NA surface proteins of the 1918 virus were derived from an avian-like influenza virus shortly before the start of the pandemic [1].

In April 2009, a novel strain of the influenza A (H1N1) virus was discovered in patients from Mexico and the United States and it spread across the globe via human-to-human contact within a very short time. Because of the seriousness of this new flu virus, the World Health Organization (WHO) has officially declared the H1N1 virus a global pandemic. Genomic analysis of the 2009 influenza A (H1N1) virus suggested that it is closely related to common reassortant swine influenza A viruses isolated in North America, Europe, and Asia. Its NA sequences have 94.4% similarity at the nucleotide level with European swine influenza A virus strains from 1992 [2].

Flu drugs such as oseltamivir (Tamiflu®) and zanamivir (Relenza®) currently in use only target the NA proteins, and disrupt the capability of the virus to escape infected cells and move elsewhere to infect other healthy cells. Clinical reports suggested that the new virus is susceptible to the two drugs [3]. However, a growing concern is that more drug resistant mutants will emerge under the selection pressure of constant drug use. Re-

searchers from Rensselaer Polytechnic Institute [4] designed a new flu drug by targeting both the HA and NA genes of the virus as an effective way to treat the next mutation of H1N1 swine flu.

Two recent studies [5,6] provided insights into the interactions of flu drugs with NA of the 2009 H1N1 virus. One study [6] developed a 3D structure of 2009 H1N1 NA and compared it with the crystal structure of 2006 H5N1 NA and the structure of 1918 H1N1 NA. It found that the hydrophobic Try347 in H5N1 NA does not match with the hydrophilic carboxyl group of oseltamivir as in the case of H1N1 NA, which explains in part the reason why the H5N1 avian influenza virus is drug-resistant to oseltamivir.

Another study [5] found that the NA sequences of 2009 H1N1 are phylogenetically more closely related to European H1N1 swine flu and H5N1 avian flu rather than to the H1N1 counterparts in the America. It also investigated the sequence variations of 2009 H1N1 NA, using three sequences of NA, A/H1N1 /California/04/2009, A/H5N1/Vietnam/2004, and A/H1N1/Brevig Mission/1/18 (the 1918 Spanish flu). With multiple sequence alignment, they found that among the 387 residues of the NA domain, the 2009 H1N1 NA differs from the other two strains in 21 positions. The novel mutations of NA are mainly located at the protein surface and not near the binding pocket for currently used NA inhibitors. It is natural to explore whether there are any potential mutation sites that may interfere with the active sites in the near future.

In light of the possible emerging of new mutations of 2009 H1N1 that could lead to serious drug resistance, it is imperative to study the potential mutations and co-mutations of the current sequences of 2009 H1N1, because mutations tend to function in concert to achieve some biological purposes. In this study, we employed entropy and mutual information theory [7,8] to investigate this issue.

Because the NA stalk supports the head domain, its length can influence the function of NA. A special NA stalk motif with a 20-amino acid deletion in the 49th to 68th positions of the stalk region was first identified in H5N1 in 2000. There was a gradual increase of this special NA stalk motif in H5N1 isolates from 2000 to 2007, and it was in all 173 H5N1 human isolates from 2004 to 2007. The H5N1 virus carrying this special NA stalk motif has the highest virulence and pathogenicity in chicken and mice [9]. This finding prompted us to search for similar stalk motifs in the current 2009 H1N1 virus strains.

In summary, the goal of our study is to conduct a timely analysis of mutations, co-mutations, stalk motifs, and phylogenetic relationship of the 2009 H1N1 NA sequences available up to date. Such information can be valuable in further efforts to improve drug design and flu treatment.

2. MATERIALS AND METHODS

2.1. Sequence Data

Published NA sequences of 7251 influenza A virus were downloaded from the Influenza Virus Resource (http://www.ncbi/nlm.nih.giv/genomes/FLU/FLU.html) of the National Center for Biotechnology Information (NCBI) on Sept 13, 2009. We were mainly interested in the sequences in 2009, but also needed the sequences in 2008 and 2007 to provide comparison in the study of stalk motifs. There were 283 sequences of H1N1 and H5N1 and 52 of H3N2 in 2009. All the sequences used in the study were aligned with MAFFT [10].

2.2. Entropy and Mutual Information

In information theory [7,8], entropy is a measure of the uncertainty associated with a random variable. Let x be a discrete random variable that has a set of possible values $\{a_1, a_2, a_3,\ldots,a_n\}$ with probabilities $\{p_1, p_2, p_3,\ldots,p_n\}$ where $p(x=a_i)= p_i$. The entropy H of x is

$$H(x) = -\sum_i p_i \log p_i$$

The mutual information of two random variables is a quantity that measures the mutual dependence of the two variables or the average amount of information that x conveys about y, which can be defined as:

$$I(x, y) = H(x) + H(y) - H(x, y)$$

where $H(x)$ is the entropy of x, and $H(x,y)$ is the joint entropy of x and y. $I(x,y)=0$ if and only if x and y are independent random variables.

In current study, each of the n columns in a multiple sequence alignment of a set of NA sequences of N residues is considered as a discrete random variable x_i ($1 \le i \le N$) that takes on one of the 20 (n=20) amino acid types with some probability. $H(x_i)$ has its minimum value 0 if all the residues at position i are the same, and achieves its maximum if all the 20 amino acid types appear with equal probability at position i, which can be verified by the Lagrange multiplier technique. A position of high entropy means that the amino acids are often varied at this position. While $H(x_i)$ measures the genetic diversity at position i in our current study, $I(x_i, y_j)$ measures the correlation between residue substitutions at positions i and j.

Entropy and mutual information were applied to sequence analysis extensively. Mutual information was employed to identify groups of covariant mutation positions in the sequences of HIV-1 protease and to distinguish the correlated residue substitutions resulting from neutral mutations and those induced by multi-drug resistance [11]. Based on entropy a simple informational index was proposed in [12] to characterize the patterns of synonymous codon usage bias. In another

study, sequence data of 1032 complete genomes of influenza A virus (H3N2) during 1968-2006 were used to construct networks of genomic co-occurrence to describe H3N2 virus evolutionary patterns and dynamics. It suggested that amino acid substitutions corresponding to nucleotide co-changes cluster preferentially in known antigenic regions of HA [13]. Further, mutual information was used to construct site transition network based on 4064 HA1 of A/H3N1 sequences from 1968 to 2008, which was able to model the evolutionary path of the influenza virus and to predict seven possible HA mutations for the next antigenic drift in the 2009–2010 season [14]. Recently, entropy and mutual information were also applied to indentify critical positions and co-mutated positions on HA for predicting the antigenic variants [15].

2.3. Mutual Information Evaluation

In order to assess the significance of our mutual information values of residue pairs of NA, it is necessary to show that these values are significantly higher than those based on random sequences. For each residue position of NA, we randomly permuted the amino acids from different sequences at that position and calculated the mutual information of these random sequences. This procedure was repeated 1000 times. The P value was calculated as the percentage of the mutual information values of the permuted sequences that were higher than those of the sequences of NA.

2.4. Random Forest Clustering

Random Forest, proposed by Leo Breiman in 1999 [16], is an ensemble classifier based on many decision trees. The structure of a single tree could be easily altered by a small perturbation of data. Random Forest overcomes this problem by averaging across different decision trees. For many data sets, Random Forest produces a highly accurate classifier for supervised learning, comparable to Support Vector Machine, the state of the art machine-learning algorithm. It computes proximities between cases and this technique can be extended to unlabeled data, leading to unsupervised clustering. In [17] random forest clustering was applied to renal cell carcinoma.

To view the clusters formed by Random Forest, multidimensional scaling [18] was utilized to project high-dimensional data down into a low-dimensional space while preserving the distances between them. First the proximities between cases i and j form a symmetric and positive definite matrix {prox(i,j)}. Then a second positive definite and symmetric matrix {cv(i,j)} is constructed using the entries of {prox(i,j)}. Random Forest extracts a few largest eigenvalues of the cv matrix and their corresponding eigenvectors. The values of $\sqrt{e(i)}v(i)$ are referred to as the ith scaling coordinate, where e(i) and v(i) re the ith eigenvalue and eigenvector of matrix cv [19]. In this study, the first and second scaling coordinates were utilized to visualize the data.

2.5. Important Sites in NA

The N1 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 [12]. According to the numbering in [5], these residue positions 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, and 448–468.

3. RESULTS

3.1. Mutations and Co-mutations

The NA molecule is a homotetramer consisting of four identical polypeptide chains, each of about 470 amino acids. The exact number varies depending on the strain of the virus. The enzymatic domain of the NA is supported from the virus envelope by a polypeptide stalk of variable length. The major molecular determinants that are known to influence the functional activities of the NA protein are the enzyme active site, the stalk length, the sialic acid binding site, and potential glycosylation sites.

In this study, entropy analysis was applied to locate the positions that have elevated likelihood to develop mutations and those that already had mutated. The top 31 positions with high entropy are displayed in **Figure 1**. All except four, 248, 339, 340, and 454, were mutational positions discovered in [5]. These exceptional positions are of interest because they have the potential to mutate and position 248 is close to one active site 247. Notably, there were two clusters of mutations, one near position 286 and another near position 386 (**Figure 1**).

For the sake of overview and comparison, we also plotted in **Figure 1** the distinct entropy distributions of the N1 and N2 sequences deposited at the NCBI web site during 2009 so far. The sequences of N1 and N2 varied the most in the neighborhood of the stalk region between positions 36 and 76. In addition, the N1 sequences in H1N1 and H5N1 are experiencing a much greater genetic change than the N2 sequences in H3N2 this year as illustrated by their entropy (**Figure 1**).

Next, we seek to probe the potential mutations that may affect the drug binding sites from a greater distance. We calculated the mutual information of each possible residue pairs from 469 residues of NA, a total of 109746 pairs. The top 44 pairs (top 0.04% of all pairs) were selected, all with a P value of zero. Because we were interested in the mutations in the NA domain, only those pairs in that region were chosen, which gave us 11 pairs:

(149, 263), (149, 321), (263, 321), (228, 321), (188, 365), (189, 369), (221, 369), (189, 386), (149, 389), (263, 389), and (321, 389). Of these 11 pairs, two networks of co-mutations were uncovered (**Figure 2**), which were mapped to the homology-based 3D structure of N1 built in [5] (**Figures 3** and **4**). These two networks of co-mutations may form interaction chains to connect distant residues to the active sites.

The first network, consisting of positions 149, 263, 321, and 389, has a remarkable property that any one of them is highly correlated to all the other three. Position 149 is near the active site and the bound drug, therefore is of great importance and it is a part of the 150 loop region including positions 147, 148, 150, and 151. Calcium ions are important for the thermo stability and enzyme activity of influenza virus NAs. Three potential metal binding sites in each monomer of the tetramer were observed. The two mutation positions, 321 and 389, are located in the region of one such site at position 470 [20] (**Figure 4**).

The second network has positions 188, 189, 221, 365, and 369. Position 221 is near the three active sites 223, 225, and 228. Positions 365 and 369 are close to the active site 368 and the bound drug; therefore positions 188 and 189 may function together with 365 and 369 to influence the active site 368 and the bound drug (**Figures 5** and **6**). Position 221 is not a mutation site in the three-sequence alignment in [5], but it has high entropy.

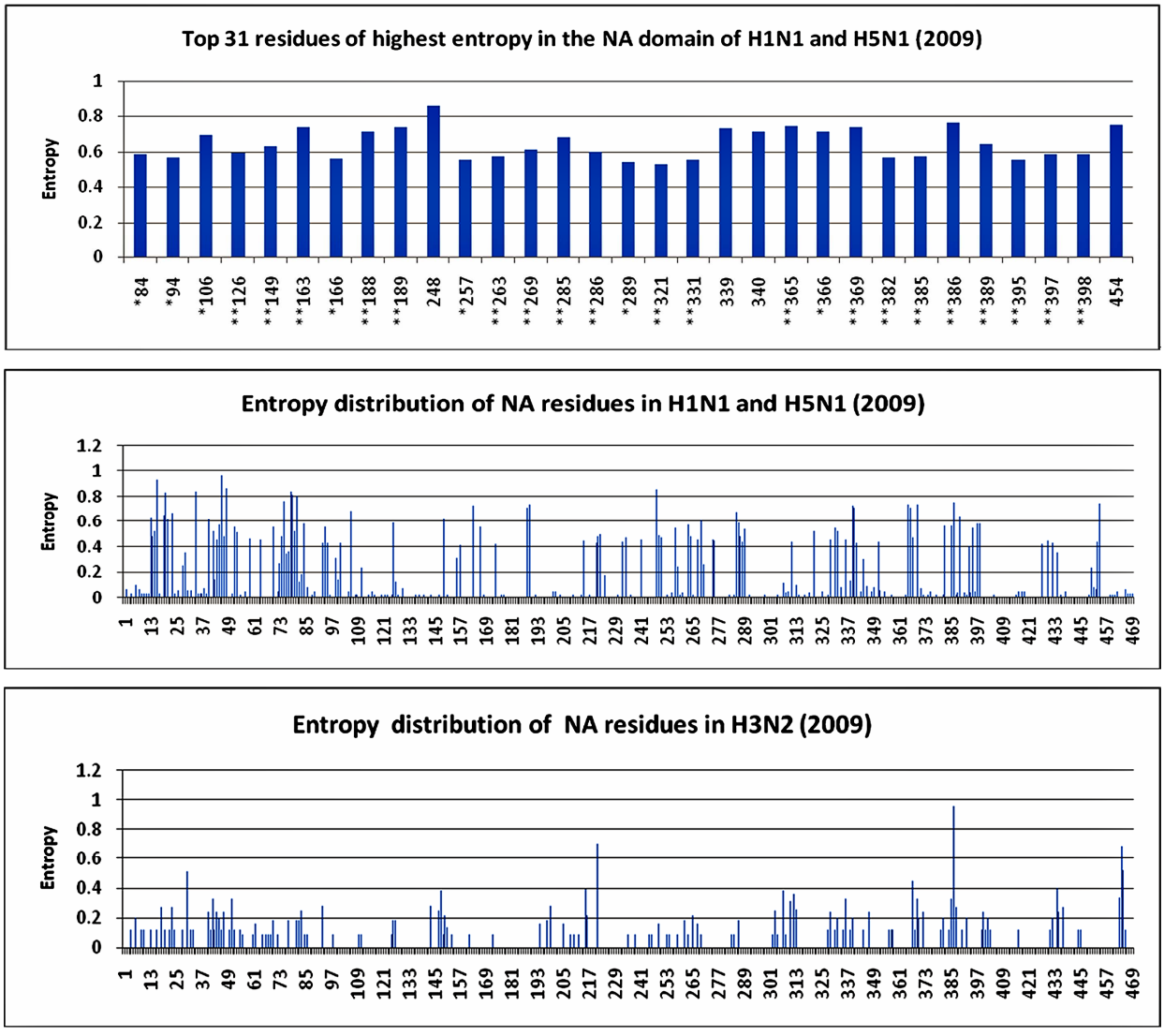

Figure 1. The top plot shows the top 31 residues of highest entropy in the NA domain (83–469) of H1N1 and H5N1 (2009). The residues that had one different amino acid than the two reference strains in [5] are marked with one asterisk, and those that had two different amino acids are marked with two asterisks. The middle and the bottom plots show the entropy of all residues in NA (1 – 469) of H1N1 and H5N1 (2009) and H3N2 (2009) respectively.

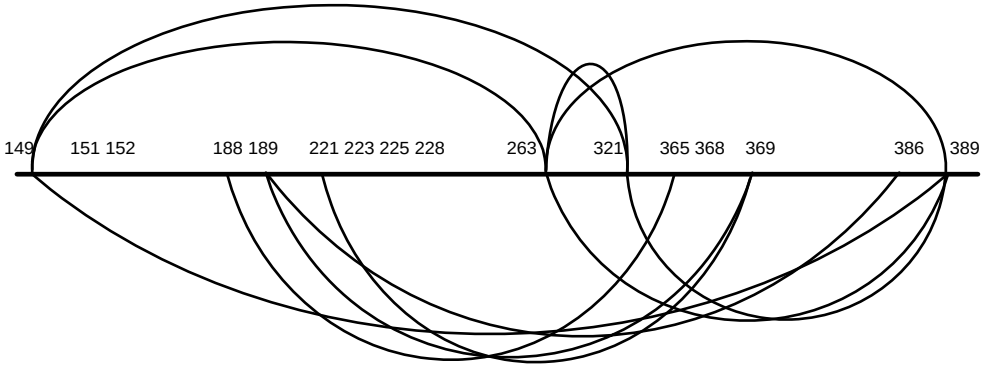

Figure 2. This plot shows the residues involved in the two networks of co-mutations with an arch to indicate the correlation between co-mutations. Three active sites 151, 152 and 368 are displayed next to their closest mutations.

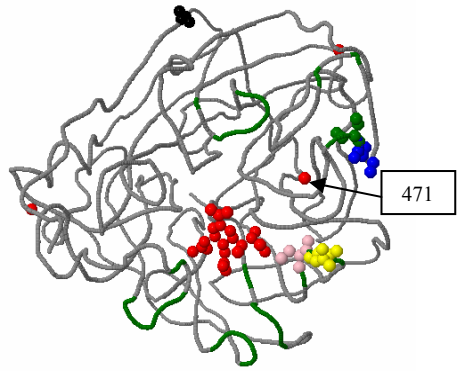

Figure 3. This plot shows in 3D structure four residues, 248, 339, 340, and 454, that have high entropy and have not mutated yet and one active site 247. Residue 248 is very close to active site 247. Calcium ion site 471 is marked to show its closeness to two residues 339 and 340. Residue 247 is in pink, 248 in yellow, 339 in blue, 340 in green, and 454 in black. The backbone of the antibody recognition sites is colored green and the bound drug (zanamivir) and three calcium ions are shown in red.

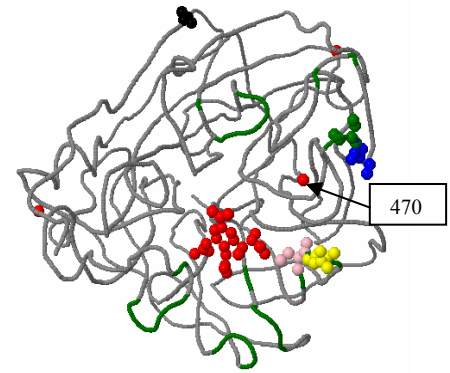

Figure 4. This plot shows in 3D structure the four residues, 149, 263, 321, and 389, in the first network of co-mutations. Residue 149 is in pink, 263 in yellow, 321 in blue, and 389 in green. Calcium ion site 470 is marked to illustrate its close position to residues 321 and 389.

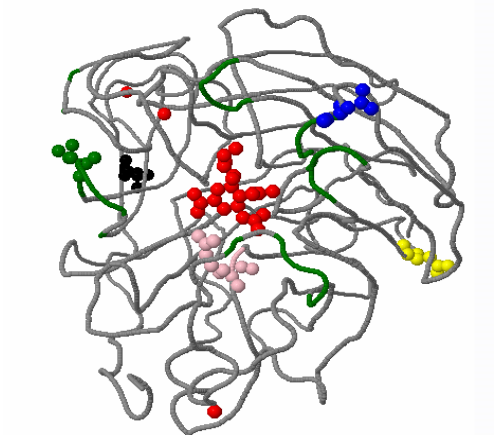

Figure 5. This plot shows in 3D structure the five residues, 188, 189, 221, 365, and 389, in the second networks of co-mutations. Residue 188 is in pink, 189 in yellow, 221 in blue, 365 in green, and 389 in black.

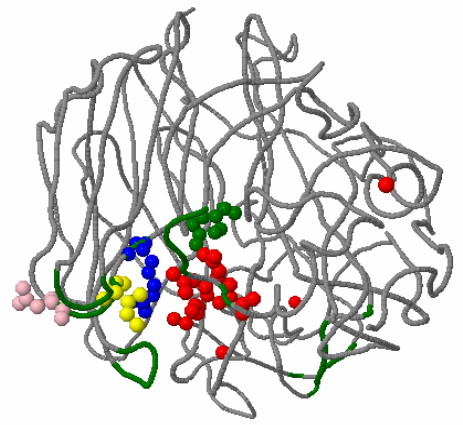

Figure 6. This plot shows in 3D structure the closeness of the four residues 221, 223, 225, and 228. One residue 221 is a part of the second network of co-mutations and the other three residues 223, 225, and 228 are active sites. Residue 221 is in pink, 223 in yellow, 225 in blue, and 228 in green.

This is a position of interest, because it is close to the active sites 223, 225 and 228 and co-mutates with a cluster of mutation positions 365, 366, and 369. The latter cluster also encloses another active site 368 (**Figures 2**, **5** and **6**).

3.2. Stalk Motifs

In this study, our intention was to discover the NA stalk motif patterns of 2009 H1N1and H5N1, which is the focus of this work, and those of 2008 and 2007 to place our findings in the right historical context. As noted in the previous section, the N1 strains in 2009 are experiencing a rapid genetic variation in the stalk region compared with other regions of N1, which could be reflected in the different stalk motifs appearing this year.

We extracted the sub-sequences consisting of positions from 36 to 79 in the NA sequences and discarded those that contained no amino acids. All the different motifs found are displayed in **Table 1**. In 2007, all NA stalk motifs in H1N1 and H5N1 had three different types referred to as types 1, 2, and 3. In 2008, type 3 stalk motif disappeared and type 4 appeared. In 2009, types 1, 2 and 4 persisted. In all these three years, types 1 and 2 persisted. Type 2, referred to as the special stalk motif in [9], was in H5N1 in all three years and type 1 was in H1N1 in all three years.

The important change in 2009 was that type 1, a common stalk motif in H1N1, was more prevalent in H5N1 and type 2, a common stalk motif in H5N1, was in H1N1 for the first time, a dramatic exchange of the two different motifs between these two subtypes of flu viruses. The special stalk motif was dominant in H5N1 in 2007 and 2008, but it was no longer the case in 2009. The new type 4 was more evident in H5N1 in 2009. These new patterns or exchange of different patterns of stalk motifs reminded us again of the fast evolutionary nature of the flu virus and the need for timely analysis of its data. The occurrence of the special stalk motif in 2009 H1N1, which may bring increased virulence to the current swine flu epidemic, is worthy of further attention and surveillance. These alterations in the stalk region of NA in H1N1 and H5N1 could also be reflected in the phylogenetic analysis conducted in the next section.

3.3. Phylogenetic Analysis

Even though the phylogenetic tree of a small number of NA sequences was constructed before [5]. With new H1N1 NA sequences being deposited at the NCBI web site regularly, it is constructive to perform phylogenetic analysis on these new sequences. For easy comparison, software MEGA [21] was used to reproduce the phylogenetic tree in **Figure 3** in [5] with the same NA sequences, which had eight sequences of 2009 H1N1 (numbered from 1 to 8) available at the time of the research (as of April 29th) in [5] and 44 different representative sequences of H1N1 or H5N1(numbered from 9 to 52) in previous years (left plot of **Figure 7**). We employed Random Forest to cluster these sequences to get a different view of their phylogenetic relationship (**Figure 8**), where a number is used to represent a sequence due to the limited space in that plot. The association of these numbers with their sequences can be found in **Figure 7**, where a number is printed before each flu subtype such as "2 H1N1 California 09 2009" and "6 H1N1 Texas 05 2009". Random Forest clustering revealed that all eight 2009 H1N1 sequences, which formed their own single cluster, were similar to others only in the second scaling coordinate, but not in the first. We reasoned this was because all eight sequences were from the same country and the structures of these clusters might be different if the plentiful sequences deposited recently at the NCBI web site were used.

We took note of some minor but interesting differences between the phylogenetic tree in the left plot of **Figure 7** and Random Forest-based clusters of the same sequences in the left plot of **Figure 8**. In **Figure 8**, the eight sequences 1, 2, 3, …, 8 were clustered in one cluster away from all the other sequences. This cluster was close to two groups of sequence numbers in the second scaling coordinate. The first group consisted of sequences

Table 1. Different NA stalk motifs

36th ---------- Stalk region ----------79th	Subtype	Year	Number of strains shared this motif	Motif type number
HS IQTGSQNHTGICNQRIITYENSTWVNHTYVNINNTNVVAG KD	H1N1	2007	196	1
HS IQTGNQHQAEP----------ISNTNFLTE KA	H5N1	2007	189	2
-- ---------NQNQVEP----------ISNTNFLTE KA	H5N1	2007	1	3
HS IQIGSQGYPETCNQSVITYENNTWVNQTYINISNTNLIGG QA	H5N1	2007	2	1
HS IQTGSQNNTGICNQRIITYENSTWVNHTYVNINNTNVVAG ED	H1N1	2008	147	1
-- ----------NHTYVNINNTNVVAG ED	H1N1	2008	1	4
HS IQTGNQCQAEP----------ISNTKFLTE KA	H5N1	2008	70	2
HS IQLGNQNQIETCNQSVITYENNTWVNQTYVNISNTNFAAG QS	H1N1	2009	157	1
HS IQTGNQCQDEP----------ISNTKFLTE KA	H1N1	2009	21	2
----------SVITYENNTWVNQTYVNISNTNFAAG QS	H1N1	2009	39	4
HS INTGNQHQAEP----------ISNANFLTE KA	H5N1	2009	1	2
HS IQLGNQNQIETCNQSVITYENNTWVNQTYVNISNTNFAAG QS	H5N1	2009	20	1

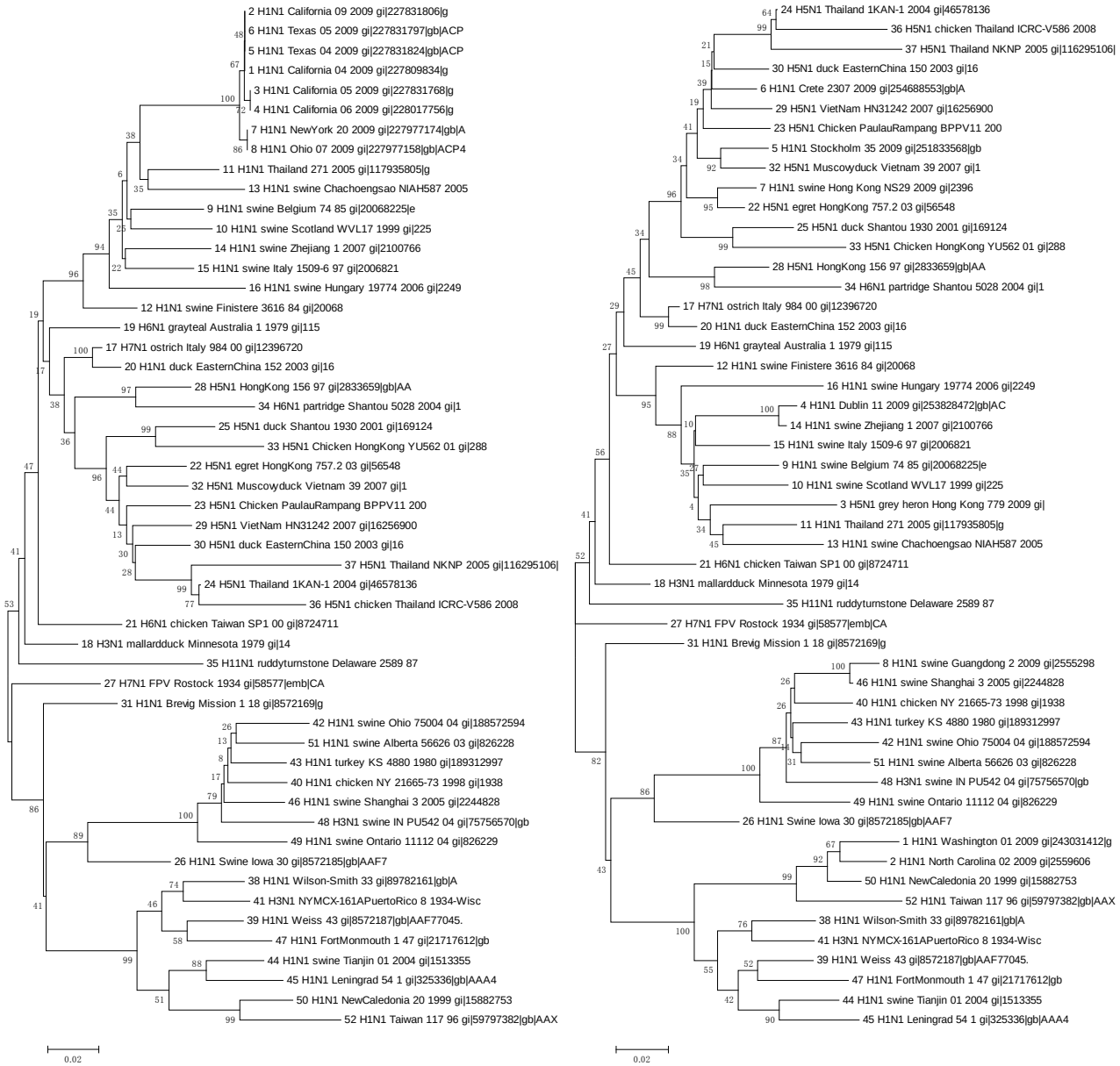

Figure 7. Left plot: a reproduced phylogenetic tree of the NA protein sequences of the N1 subtype family in [5]. Right plot: A phylogenetic tree of the same sequences used in the left plot except that the first eight sequences were replaced with eight new representative sequences from 283 N1 sequences in 2009.

13, 14, 16, 10, and 11, and the second had sequences 38, 35, and 31. The phylogenetic tree displayed the close relationship between the first group and the group of sequences 1, 2,…., 8, but did not do so for the second group. The first group had H1N1 swine Chachoengsao 2005 (13), H1N1 swine Zhejiang 2007 (14), H1N1 swine Hungary 2006 (16), H1N1 swine Scotland 1999 (10), and H1N1 Thailand 2005 (11). The second group had H1N1 Wilson Smith 33 (38), H11N1 Delaware 1987 (35), and H1N1 Brevig Mission 1918 (31). These two groups were all H1N1 subtype and the second group inherited from the ancient strains in the first group, and that is why they were similar and clustered together by Random Forest.

To get a new view of the clusters of the NA sequences of 2009 H1N1 and H5N1 available up to date, eight new representative sequences were selected from 283 sequences in the same year with cd-hit [22] to replace the eight NA sequences, numbered from 1 to 8, used in the left plot of **Figures 7** and **8**. The new sequences were H1N1 Washington 2009 (1), H1N1 North Carolina 2009 (2), H5N1 Hong Kong 2009 (3), H1N1 Dublin 2009 (4), H1N1 Stockholm 2009 (5), H1N1 Crete 2009 (6), H1N1 swine Hong Kong 2009 (7), and H1N1 swine Guang-

dong 2009 (8). In the Random Forest-based clusters in the right plot of **Figure 8**, there were three major branches with the center made of ancient NA sequences. Sequences 1, 2, and 8, sequences 3 and 4, and sequences 5, 6, and 7 were on each branch separately, which was also similarly reflected in the phylogenetic tree in the right plot of **Figure 7**. Sequences 1 and 2 were close to sequences 50 and 52 and sequence 8 was close to sequences 46 and 43. Sequences 3 and 4 were close to sequences 18, 16, 10, 13, 12, 9, 15, 11, and 14, which were mainly European swine strains. Sequences 5 and 6 were close to sequences 30 and 32 and sequence 7 was close to sequences 22, 25, and 33. All eight representative sequences were evenly spread in various clusters in the right plot of **Figure 8** and the phylogenetic tree in the right plot of **Figure 7**, which illustrated that as the 2009 H1N1 strains continue to evolve, diverse genetic makeup of the sequences would develop.

4. CONCLUSIONS

The recent outbreak of the novel 2009 H1N1 influenza has raised global concerns regarding its virulence and pandemic potential. Several recent studies on the 2009 H1N1 virus used small datasets to conduct analysis. With new sequences available up to date, we were able to extend the previous research in three areas using entropy, mutual information, and Random Forest. The first was finding two networks of co-mutations that may potentially affect the current flu-drug binding sites on NA. The second is discovering a special stalk motif, which was dominant in the H5N1 strains in the past, in the 2009 H1N1 strains for the first time. Due to the high virulence of this motif, the second finding is significant in our current research on 2009 H1N1. The third was updating the phylogenetic analysis of the NA sequences of 2009 H1N1 and H5N1, which demonstrated that, in clear contrast to previous findings, the N1 sequences in

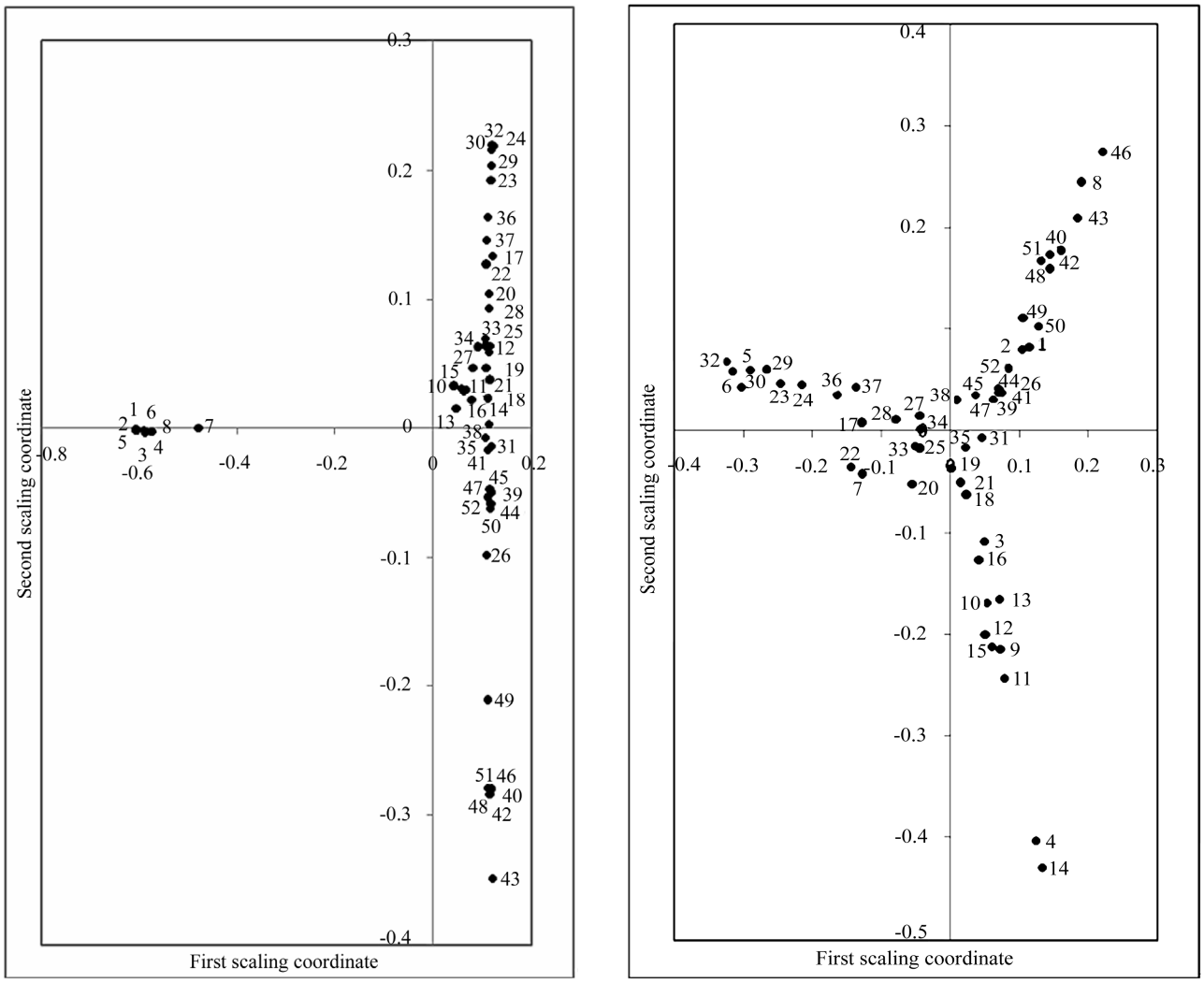

Figure 8. Left plot: Random Forest-based clusters of the same NA sequences used in the left plot of **Figure 7**. Right plot: Random Forest- based clusters of the same NA sequences used in the right plot of **Figure 7**.

2009 are diverse enough to cover different major branches of the phylogenetic tree of those in previous years. As the novel influenza A H1N1 virus continues to spread globally, our results highlighted the value of performing timely analysis on the 2009 H1N1 virus.

5. ACKNOWLEDGMENTS

We thank Houghton College for its financial support.

REFERENCES

[1] Taubenberger, J. K. and Morens, D. M., (2006) 1918 Influenza: the mother of all pandemics, Emerg. Infect. Dis., **12(1)**, 15–22.

[2] Trifonov, V., Khiabanian, H., and Rabadan, R., (2009) Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus, N. Engl. J. Med., **361**, 115–119.

[3] Centers for Disease Control and Prevention (CDC), (2009) Update: Drug susceptibility of swine-origin influenza A (H1N1) viruses, MMWR Morb Mortal Wkly Rep 2009, **58**, 433–435.

[4] Weïwer, M., Chen, C. C., Kemp, M. M., and Linhard, R. J., (2009) Synthesis and biological evaluation of non-hydrolyzable 1,2,3-triazole-linked sialic acid derivatives as neuraminidase inhibitors, European Journal of Organic Chemistry, **16**, 2587.

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

[6] Wang, S. Q., Du, Q. S., Huang, R. B., Zhang, D. W., and Chou, K. C., (2009) Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus, Biochemical and Biophysical Research Communications, **386(3)**, 432–6.

[7] Cover, T. A. and Thomas, J. A., (1991) Elements of information theory, John Wiley and Sons, NewYork.

[8] MacKay, D., (2003) Information theory, inference, and learning algorithms, Cambridge University Press.

[9] Zhou, H. B., Yu, Z. J., Hu, Y., Tu, J. G., Zou, W., Peng, Y. P., Zhu, J. P., Li, Y. T., Zhang, A. D., Yu, Z. N., Ye, Z. P., Chen, H. C., and Jin, M. L., (2009) The special neuraminidase stalk-motif responsible for increased virulence and pathogenesis of H5N1 influenza A virus, PLoS One, **4(7)**, e6277.

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

[11] Liu, Y., Eyal, E., and Bahar, I., (2008) Analysis of correlated mutations in HIV-1 protease using spectral clustering, Bioinformatics, **24(10)**, 1243–1250.

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

[13] Du, X. J., Wang, Z., Wu, A. P., Song, L., Cao, Y., Hang, H. Y., and Jiang, T. J., (2008) Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution, Genome. Res., **18**, 178–187.

[14] Xia, Z., Jin, G. L., Zhu J., and Zhou, R. H., (2009) Using a mutual information-based site transition network to map the genetic evolution of influenza A/H3N2 virus, Bioinformatics, **25(18)**, 2309–2317.

[15] Huang, J. W., King, C. C., and Yang, J. M., (2009) Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses, BMC Bioinformatics, **10(Suppl 1)**, S41.

[16] Breiman, L., (2001) Random forests, Machine Learning, **45(1)**, 5–32.

[17] Shi, T., Seligson, D., Belldegrun, A. S., Palotie, A., and Horvath, S., (2005) Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma, Mod. Pathol., **18(4)**, 547–57.

[18] Cox, T. F. and Cox, M. A. A., (2001), Multidimensional scaling, Chapman and Hall.

[19] http://www.stat.berkeley.edu/~breiman/RandomForests/.

[20] Xu, X. J., Zhu, X. Y., Dwek, R. A., Stevens, J., and Wilson, I. A., (2008) Structural characterization of the 1918 influenza virus H1N1 neuraminidase, Journal of Virology, **82(21)**, 10493–10501.

[21] Kumar, S., Nei, M., Dudley, J., and Tamura, K., (2008) MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences, Brief Bioinformatics, **9**, 299–306.

[22] Li, W. and Godzik, A., (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, **22**, 1658–1659.

