CLUSTERING NEURAMINIDASE INFLUENZA PROTEIN SEQUENCES *

X. LI (1), H. JANKOWSKI (1), S. BOONPATCHARANON (1), V. TRAN (1), X. WANG (1), J. M. HEFFERNAN (1,2,3)

(1): Department of Mathematics and Statistics
(2): Modelling Infection and Immunity Lab
(3): Centre for Disease Modelling
York University, 4700 Keele St.
Toronto, Ontario M3J 1P3, Canada
E-mail: lixuan@mathstat.yorku.ca, hkj@yorku.ca, sawitree.boonpat@gmail.com, vickictran@gmail.com, stevenw@mathstat.yorku.ca, jmheffer@yorku.ca

We seek a better understanding of this evolution over influenza seasons. In a previous work, we studied the evolution (antigenic drift) of the highly variable influenza A H3N2 by focusing on the hemagglutinin (HA) viral glycoprotein. In this update, we include also the neuraminidase (NA) glycoprotein, another protein that contributes to the antigenic drift of influenza. Our method is based on a dimension reduction technique combined with a fully automatic Hamming distance statistical clustering method for categorical data (Zhang et al. JASA, 2006). The new NA results are compared with the previous HA results to provide a more complete picture of flu virus evolution.

1. Introduction

In the northern and southern parts of the world, influenza outbreaks occur mainly in the winter months while in areas around the equator outbreaks may occur at any time of the year[1]. The seasonal pattern of infection in the hemispheres has coined the name 'seasonal influenza'. Seasonal influenza is associated with significant human mortality and morbidity worldwide[1,3]. Much of the seasonal influenza burden is caused by influenza A[1].

Influenza A viruses are classified into subtypes based on antibody responses to their two surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA). There are 16 HA and 9 NA subtypes known, but only HA 1, 2, and 3, and NA 1 and 2 are commonly found in humans[2,3]. Among

*This work is supported by NSERC.

the many subtypes of influenza A viruses, the influenza A H1N1 and H3N2 subtypes currently cause seasonal influenza epidemics, with H3N2 causing the vast majority of infections[3].

Influenza A viruses continuously undergo mutation in the HA and NA surface antigens. This is called antigenic drift. Through antigenic drift, an increasing variety of strains are created. The new strains can then cause seasonal epidemics, since the population can only gain partial immunity from previous infection(s). The HA protein has been identified to the be the major contributor to the antigenic drift seen in influenza A[3]. Changes in the NA protein, however, have also been shown to contribute[3].

Vaccination against influenza is recommended every year[1,4,5]. The continuous change in the circulating influenza strains requires the seasonal influenza vaccine formulation to be considered yearly. The vaccine, however, takes approximately six months to formulate and produce. Throughout this manufacturing period the influenza virus continues to evolve. This, in turn, affects the efficacy of the vaccine in the population. Between the years 1997 to 2007, vaccine efficacy ranged from 18%–90%[6]. Clearly, the methods employed to predict the circulating influenza strains from year to year are not optimal.

A second control strategy against influenza includes the use of antiviral drugs, which can reduce the severity of symptoms and pathogen transmission during influenza infection[7]. There are two classes of antiviral drugs currently in use, neuraminidase inhibitors (oseltamivir and zanamivir) and M2 protein inhibitors (adamantine derivatives). Currently, adamantine is not recommended for treatment of influenza[7]. When influenza is circulating in a community, either oseltamivir or zanamivir are recommended in the treatment of patients that have risk of severe complications from infection, but only if treatment can be initiated within 48 hours of the onset of symptoms[7]. NA mutations that confer resistance to oseltamivir and zanamivir have been identified in seasonal influenza epidemics[7]. Neuraminidase inhibitor efficacy, thus, is affected by changes in the NA glycoprotein[8].

We are interested in quantifying the evolution of the HA and NA glycoproteins. We have developed a formal cluster-based technique that can be used to study the evolution of influenza over time[9]. Previously, we employed our technique to determine families (or clusters) of the H3N2 HA glycoprotein genetic sequence[9]. Our results uncovered important new trends in HA evolution[9]. We now continue our study of seasonal influenza A H3N2 mutation focusing on the NA glycoprotein.

Table 1. Vaccine sequences in the dataset.

Stain Name	Number of sequences	Accession Number
A/Moscow/10/99	2	AY531035, DQ487341
A/Fujian/411/2002	2	CY088483, CY112933
A/California/7/2004	1	CY114373
A/Wisconsin/67/2005	4	CY033646, CY163936 CY114381, EU103823
A/Brisbane/10/2007	3	CY035022, CY039087 EU199366
A/Perth/16/2009	1	GQ293081
A/Victoria/361/2011	1	KC306165
A/Texas/50/2012	2	KC892248, KC892952

2. Data Description and Methodology

2.1. *Data acquisition*

The NA sequences considered in the study were obtained from the publicly available online repository known as the Influenza Research Database[10] (IRD), `www.fludb.org`. The specific sequences used were chosen based on the criteria given in Table 2. The calendar year, country and city of isolation for each sequence is provided in the IRD. We also wanted to make sure that strains used for vaccines (Table 1) were included in the data. Vaccine sequences containing the complete date (year, month, and day) are naturally selected by our search criteria. Some vaccine sequences did not have a complete date, and were added to the data set manually. The criteria yield a total of 2049 sequences with 550 amino acids each, and among these are 12 vaccine sequences.

Table 2. IRD criteria: All other settings kept default or blank.

Option	Criteria
"Data to return":	protein
"Virus type":	A
"Sub type":	H3N2
"Select segments":	NA
"Complete sequences":	Complete Segments Only
"Date range":	1998 to 2012
"Host":	Human
"Geographic grouping":	All
Advanced options	
"Month Range":	Sep 1998 to July 2012
"Remove Duplicate Sequences":	Yes

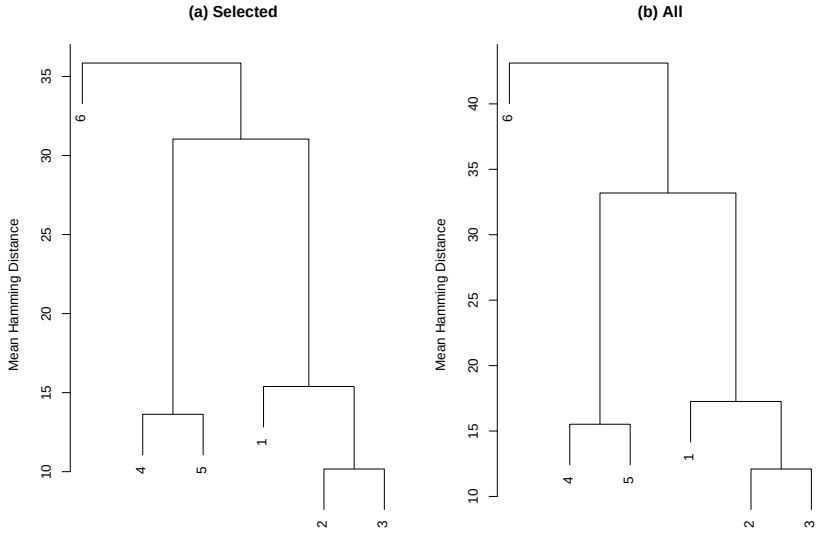


Figure 1. Dendrograms of clusters by mean Hamming distance. This plot is drawn using hierarchical cluster analysis with complete linkage. The left plot is based only on the 75 highest entropy sites, whereas the right plot uses all 550 sites to calculate the Hamming distance.

MEGA 5.2 software[11] was then used to translate the RNA sequences into protein sequences, while the software MUSCLE[12] was used to align the sequences. Perl script was written to order and combine the sequences for processing in Matlab. This procedure resulted in 2049 observations with 550 categorical variables, each containing 21 categorical states (20 for each kind of amino acid and one to represent a gap). The occurrence of gaps may be due to some deletion or transition of a nucleotide, which is highly related to random genetic drift and evolution. Another reason for gaps is the inappropriate alignment of the sequences. Since NA protein sequences are highly conservative and the alignment uses pairwise comparison, the probability of improper alignment should be quite small.

Files containing both the pre- and post-processed data are provided as supplementary material, and are also available online at `www.math.yorku.ca/~hkj/Research`.

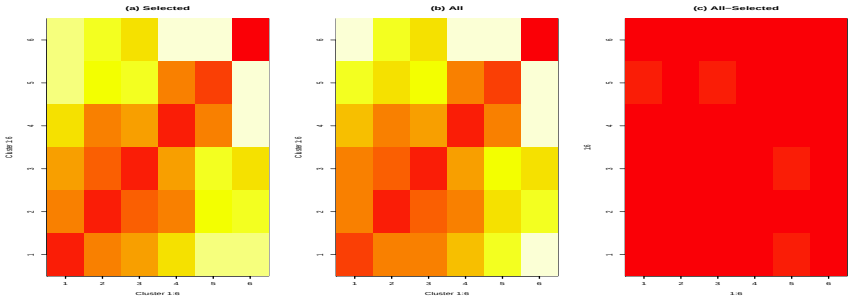


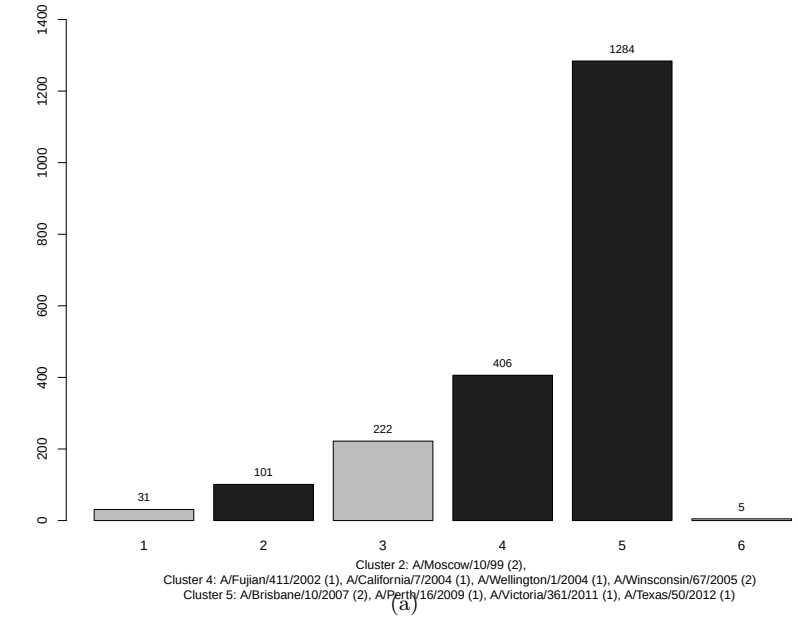
Figure 2. From left to right: mean Hamming distance matrix of 75 selected most varied sites by cluster, mean Hamming distance matrix of the whole sequence with 566 sites by cluster, absolute differences of the two matrices. Both matrices (left and centre) have been standardized by dividing by their corresponding maximum values.

2.2. *Clustering the sequences*

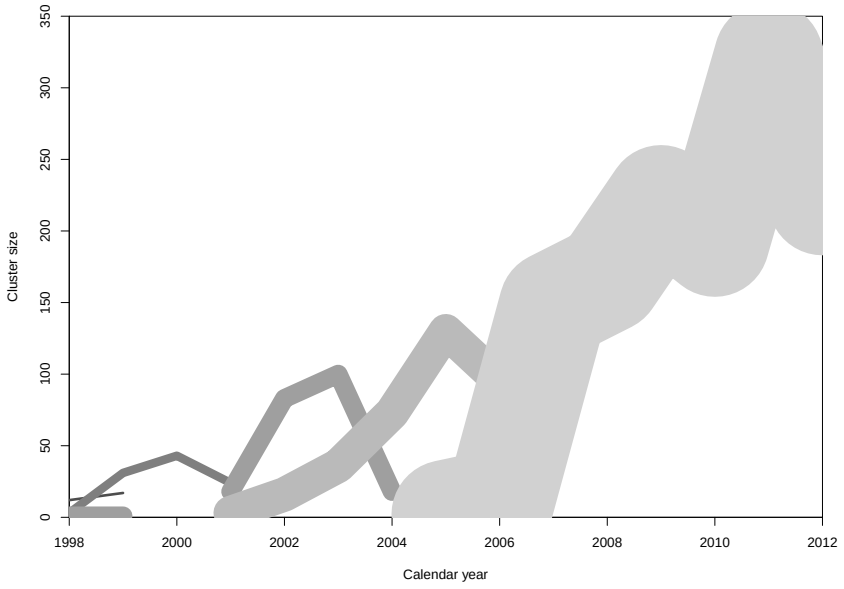
Our goal was to analyze the vaccine and observed strain sequences via clustering. Our methodology is the same as previously employed in Li et al. [9] and comprises two main steps: a dimension reduction step and a clustering step based on Hamming distance.

(1) **Dimension reduction step:** As the original data lives inside a space of dimension 21^{550}, a dimension reduction step is necessary. To do this, the entropy of the empirical distribution on proteins was calculated at each site for the 2049 observed sequences. The sites with no variability or only one varying location were removed from the data (these correspond to very small entropy or zero entropy). The remaining entropies were clustered using a Gaussian mixture model[14,16] implemented in the R software package[15]. The cluster with the highest entropy was then selected for further analysis. This allows us to consider only 75 sites with highest variability for the next step.

(2) **Clustering step:** In the second step we cluster the data, which now lives in a space of maximal size 21^{75}. The clustering method was that of Zhang et al. (JASA, 2006)[13], see also Li et al. (2015) [9] for additional details. We remark that the method is fully automatic, intuitive, and based on the Hamming distance between sequences.



(a)



(b)

Figure 3. (a) Cluster sizes and vaccine locations. For convenience, the clusters have been re-ordered by earliest year of isolation. Clusters containing a vaccine strain are denoted in black. (b) The number of HA protein sequences within each cluster plotted versus calendar year of isolation.

3. Results

We first analyzed the sequence data as described above. The initial dimension reduction (step (1)) yielded 75 sites of "high variability." The clustering step (step (2)) yielded six clusters. Figure 1 shows the dendrograms of the resulting clusters where the distance is based on the mean Hamming distance for (left panel) just the 75 sites of highest entropy/variability, and (right panel) all 550 sites. We can see that there is little difference between these two dendrograms, providing evidence that our dimension reduction step does not lose much (if any) important information. This is confirmed also in Figure 2, where the mean Hamming distances of the six clusters are compared when calculated for the 75 selected sites, and for all 550 sites.

Figure 3 (a) shows histograms of the cluster size, where clusters containing vaccines are identified in black. Figure 3 (b) shows the number of protein sequences within each cluster plotted against the calendar year of virus isolation. The dominant cluster, cluster five, overwhelms these results. For this reason, we added an additional analysis, whereby cluster five was again clustered (or sub-clustered) using the previously described two-step procedure. Details are given in the following section.

3.1. *Sub-clusters of cluster five*

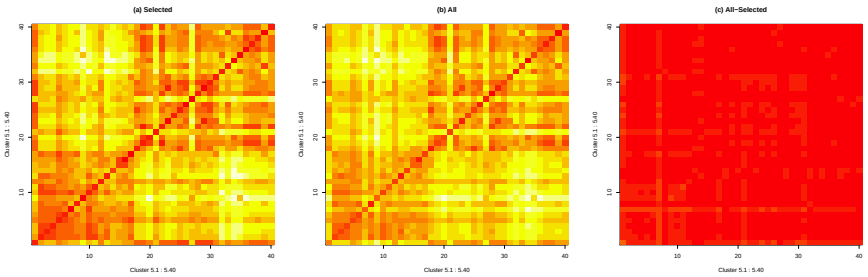
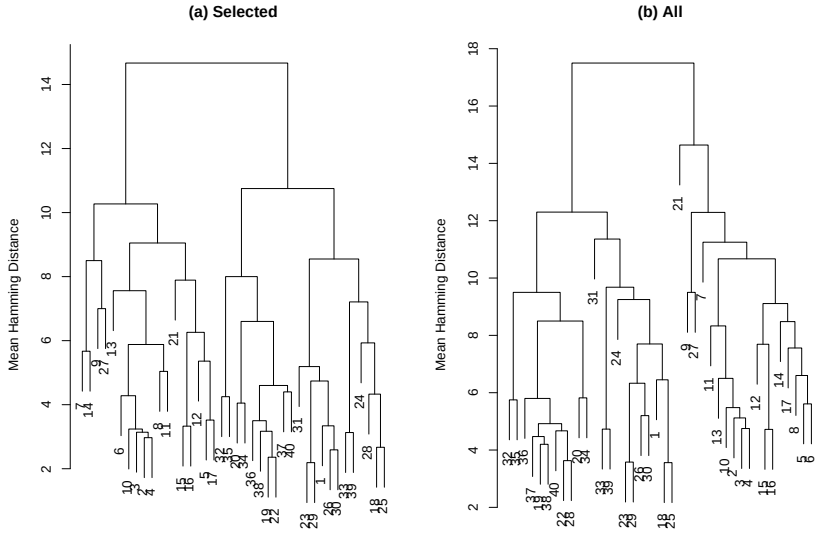
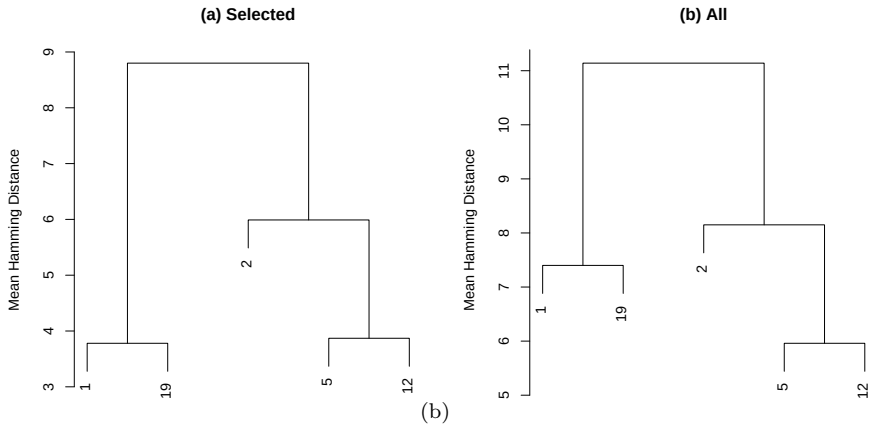


Figure 4. Mean hamming distance is shown for the (a) selected sites, (b) all sites, and (c) the difference between all and the selected sites.

To find sub-clusters of cluster five, we repeated steps (1) and (2) as above. Namely, selecting only the sequences of cluster five, we first performed the dimension reduction step on the entire length of the sequence. This yielded a reduction from 550 to 38 sites. Secondly, we performed the clustering step. This yielded 40 sub-clusters, which we denote as 5.1–5.40.

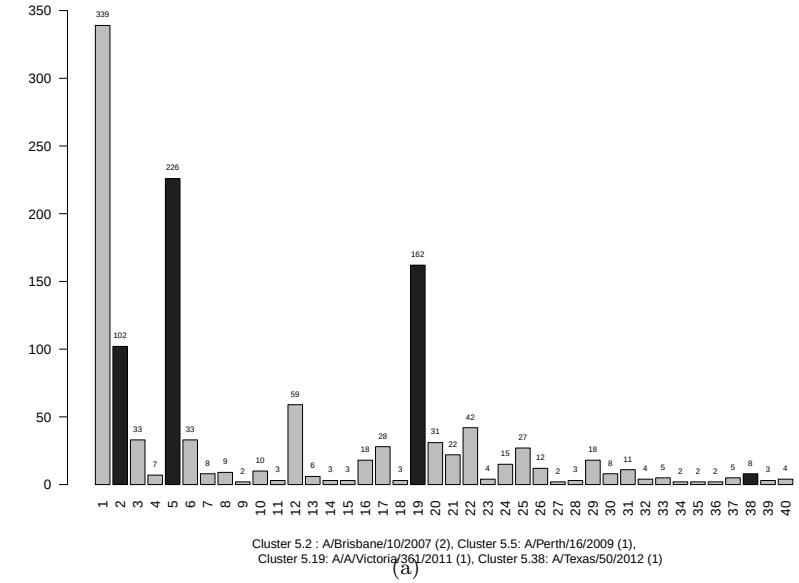


(a)

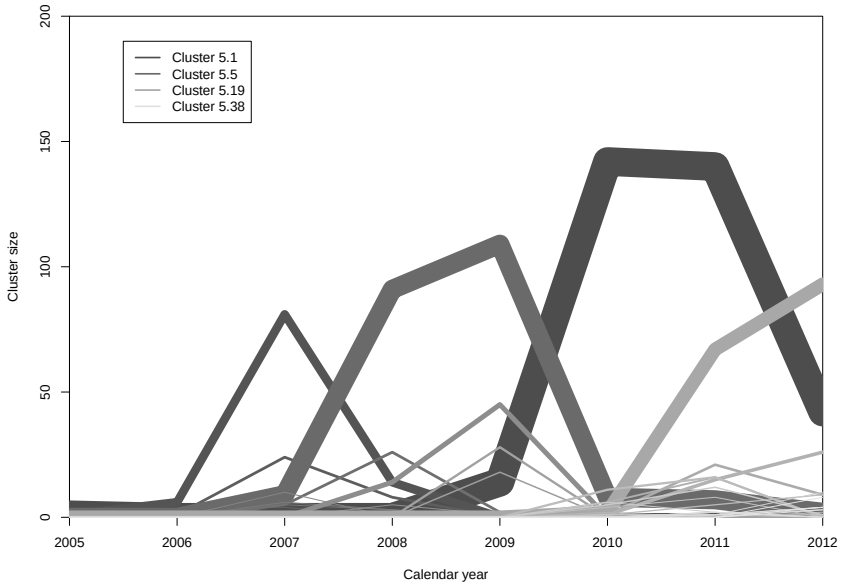


(b)

Figure 5. Dendrograms of clusters by mean Hamming distance. This plot is drawn using hierarchical cluster analysis with complete linkage. The left plot is based only on the 38 highest entropy sites, whereas the right plot uses all 550 sites to calculate the Hamming distance. (a) All sub-cluster are shown. (b) Only sub-clusters with at least 50 sequences are shown.



(a)



(b)

Figure 6. (a) Sub-cluster sizes and vaccine locations for the sub-clusters of the dominant cluster five. The subclusters have been re-ordered by earliest year of isolation. Clusters containing a vaccine strain are denoted in black. (b) The number of HA protein sequences within each sub-cluster of cluster five plotted versus calendar year of isolation.

Heatmaps confirming the validity of the dimension reduction step are given in Figure 4. Dendrograms of the clusters drawn by using the 38 sites and all 550 sites are given in Figure 5. Figure 5 (a) shows all of the sub-clusters, which is difficult to read due to the large number of small sub-clusters. Figure 5 (b) shows only those sub-clusters which contain at least 50 sequences. Here, the dendrograms are quite similar, again justifying the dimension reduction step.

Figure 6 (a) shows histograms of the sub-cluster sizes, where sub-clusters containing vaccines are identified as black. Here, we identify five large (with at least 50 sequences) sub-clusters: 5.1, 5.2, 5.12, 5.19, and 5.38. We see a lot of "variability" here as well though - of the 40 sub-clusters, we only identify several dominant ones. Figure 6 (b) shows the number of protein sequences within each cluster plotted against the calendar year of virus isolation. Here, we again clearly see the dominant sub-clusters. Notice that the largest of these sub-clusters, sub-cluster 5.1, does not include a vaccine strain.

4. Discussion

In this paper we have studied antigenic drift within the NA component of the influenza A H3N2 strain that causes seasonal influenza infections every year. We employed our previously reported method for clustering protein sequences to identify related NA glycoproteins across influenza A H3N2 strains. Analysis of the clusters found that the NA component of influenza A H3N2 is related by year (or 'flu season'), and that clusters appear to replace older clusters over time (Figure 3). In our results, cluster five contains almost half of the NA protein sequences in our data set. We therefore preformed further analysis of cluster five, which resulted in 40 sub-clusters. It is interesting to note that of the sub-clusters, the vaccine strains are mainly located in the "dominant" sub-clusters (largest sub-clusters). These dominant sub-clusters again appear to replace one another every few years. These results may point to a similar trend of genetic drift in the NA glycoprotein to that of the HA glycoprotein: cluster replacement every 2-5 years and evolution of the dominant seasonal strain[9].

The results reported above were observed from two subsequent implementations of our clustering method: once to identify clusters of the NA glycoprotein, and then again to identify sub-clusters of cluster five. Such recursive implementations of our methodology should be seen as an avenue for exploratory data analysis, but the methodology is not statistically rig-

orous. Additional work is required to develop of a formal statistical method that can extract both large scale (cluster-level) and small scale (sub-cluster level) evolutionary trends.

In both our previous and current studies of influenza A antigenic drift, we have observed that dominant clusters of the HA and NA glycoproteins do not always include vaccine strains. For example, sub-cluster 5.1 in the current study does not house a vaccine strain, but sub-cluster 5.2, which is smaller in size, does. Similar observations were made for the HA glycoprotein[9]. These results point to complications in the identification of the dominant strain from year to year.

Our methodology can be improved in several ways. First, as mentioned previously, future work is needed to address both large-scale and small-scale evolution. In addition, it is known that, in viruses, mutations related to immune-escape may occur in combination. The prevalence of epistasis in the evolution of Influenza A surface proteins has been previously studied[17]. We will expand our method to take epistatic mutations into account so that 'hot-spot' combinations of mutations can be identified.

References

1. World Health Organization, http://www.who.int/influenza/en/.
2. E.G. Strauss and J.H. Strauss, *Academic Press* (2007).
3. D.M. Knipe and P.M. Howley, *Philadelphia: Lippincott Williams and Wilkins.* (2007).
4. Centres for Disease Control and Prevention, http://www.cdc.gov/flu/professionals/vaccination/.
5. P.K. Tosh, R.M. Jacobsen and G.A. Poland, *Mayo Clin Proc.* **83**(3), 257-273 (2010).
6. F. Carrat and A. Flahault, *Vaccine* **25**(39-40), 6852 - 6862 (2007).
7. Centre for Disease Control, http://www.cdc.gov/flu/professionals/antivirals/antiviral-drug-resistance.htm .
8. M. Foll, Y.-P. Poh, N. Renzette, et al., *PLOS Genetics* **10**(2): e1004185 (2014).
9. X. Li, H. Jankowski, S. Wang and J.M.Heffernan, *BIOMAT 2014 Conference Proceedings*, in press.
10. Influenza Research Database, www.fludb.org.
11. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar, *Mol. Bio. Evo.* **28**, 2731-2739 (2011).
12. Robert C. Edgar, *Nucleic Acids Res.* **32**(5), 1792 - 1797 (2004)
13. Peng Zhang, Xiaogang Wang and Peter X.-K. Song, *J. Amer. Statist. Assoc.* **101**(473), 355-367 (2006).
14. Chris Fraley and Adrian E Raftery, *J. Amer. Statist. Assoc.* **97**, 611-631

(2002).
15. R Development Core Team, *R Foundation for Statistical Computing*, Vienna, Austria, http://www.R-project.org. (2008).
16. Chris Fraley and Adrian E Raftery, *MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering*, Technical Report No. 504, Department of Statistics, University of Washington, (revised 2009).
17. S. Kryazhimskiy, J. Dushoff, G.A. Bazykin and J.B. Plotkin, *PLOS Genetics* **7**: e1001301 (2011).